Proceedings of the 1998 Geographic Information Systems in Public Health Conference

Contents

Implementation and Operations 287

Disease Surveillance 363

Social and Demographic Analysis 499

Environmental Health Protection

A Study on Environmental Equity in Albuquerque, New Mexico: Executive Summary

Amy Baker (1), Gloria Cruz (supplemental atlas) (2)*
(1) Doctoral Candidate, Economics Department, University of New Mexico, Albuquerque, NM; co-project lead, report author; (2) Data Analyst III, Albuquerque Environmental Health Department, Albuquerque, NM; co-project lead, GIS support and map production

Note: This paper contains the executive summary of the report, *A Study on Environmental Equity in Albuquerque, New Mexico.* For a copy of the full report, contact Gloria Cruz.

Abstract

A study on environmental equity in Albuquerque, New Mexico, analyzed the distribution of sites of environmental concern relative to host community demographic composition. In two phases, the study analyzed environmental equity issues involving the 419 census-defined block groups within the Albuquerque metropolitan area using a geographic information system (GIS) and rigorous statistical modeling. The GIS proved to be a very powerful tool for determining the accuracy of the models and linking the statistical data with spatial analysis. Geographic analysis was used to determine the relationship between environmental sites, demographic data, voter participation, and major transportation corridors. Voter participation and proximity to major transportation corridors were found to be strong indicators of site presence and number. In Phase II, the GIS was used to aggregate block group-level data to the neighborhood level and to calculate the mean populations of the various social groups for the study area. Additional social groupings were included—e.g., children and elderly people in poverty. Areas of special concern were identified based on whether or not they fell above the mean (depending on the criteria); that is, based on how their demographic characteristics related to those of other Albuquerque neighborhoods. Mean difference tests established whether these areas of special concern contained disproportionate numbers of sites. This study involved the creation of a supplemental atlas, which contains over 60 maps including location and bivariate maps of environmental sites overlaid on polygons representing various types of demographic composition. The Albuquerque Geographic Information System, the GIS used in this study, has been online since 1986. The GIS software used by the City of Albuquerque is ARC/INFO (ESRI, Redlands, CA). The hardware components of the system are a Sun S690 server, a Sun Classic workstation, a Hewlett Packard Color LaserJet printer, and a Hewlett Packard DesignJet printer.

Keywords: private, environmental equity, Census

The Albuquerque (New Mexico) City Council allocated funds in its fiscal year 1997 operating budget to the Albuquerque Environmental Health Department (AEHD) for a study identifying the relationships of demographic factors with sites of environmental

*Gloria Cruz, Albuquerque Environmental Health Department, PO Box 1293, Albuquerque, NM 87103 USA; (p) 505-768-2603; (f) 505-768-2617; E-mail: gcruz@cabq.gov

concern ("environmental sites" or "sites"),[1] using appropriate statistical and econometric techniques. AEHD hired an intern, a doctoral candidate aided in an advisory capacity by professors in the University of New Mexico (UNM) Economics Department, to conduct the study as co-project lead in conjunction with the AEHD Environmental Services Division. The study took place between July 1996 and June 1998. This status report describes two phases of the study, its results, and those recommendations for further action that have been attained thus far.

Phase I of the study analyzed environmental equity issues involving the 419 census-defined block groups within the Albuquerque metropolitan area using a geographic information system (GIS) and econometric and statistical techniques (probit and tobit models and mean difference tests). These tools were used to discern whether systematic distributional relationships existed between the 1,387 environmental sites and the differing racial, ethnic, socioeconomic, and collective-choice characteristics of their (block group-level) host communities. "Collective choice" refers to the propensity of residents in an area to engage in collective action, or to the level of political activity within a community. For the purposes of this study, the sites were identified by the federal environmental regulatory statute applied to each site type. Although contamination events were included, actual or potential contamination was not a prerequisite for inclusion in the study.

Racial, ethnic, and socioeconomic information from the 1990 US Census and voter participation results from the Bernalillo County Clerk were used as the demographic information describing each community. Financial data on the disbursement of public funds for remediation of leaking underground storage tanks (LUSTs) were obtained from the New Mexico Underground Storage Tank Bureau (the Bureau) and incorporated into the analysis. AEHD used GIS software called ARC/INFO (ESRI, Redlands, CA) to evaluate the locations, types, and number of environmental sites relative to the block groups. The intern used the statistical package SHAZAM (University of British Columbia, Vancouver, BC, Canada) to model the statistical relationships between site location, public funds distribution, and community social factors. The study relied on available cross-sectional data, which essentially provided a snapshot view of the situation as it *currently exists.* Caution, therefore, is urged in any attempt to infer cause and effect relationships from study results. The full report contains definitions for the environmental sites and other terms.

The sites were divided and grouped by type for analytical purposes. LUST sites were examined separately to facilitate comparison between the distribution of the actual sites and the distribution of Bureau funds among these sites. The remaining sites were grouped into four categories depending on the type of regulation applied to them. The models used in this study then attempted to explain the systematic pattern of distribution of the four categories relative to community demographic composition. In addition, areas of special concern (ASCs) were identified and descriptive statistical

[1] In this study, sites of environmental concern were identified and defined by the type of regulatory statute applied to them. Environmental sites, therefore, included contamination events (regulated by the Comprehensive Environmental Response, Compensation, and Liability Act), hazardous waste generators (Resource Conservation and Recovery Act), and hazardous product facilities (Superfund Amendments and Reauthorization Act, Title III). Leaking underground storage tanks were also classified as environmental sites.

techniques (mean difference tests) were used to compare the distribution of sites within ASCs to distribution of sites in the rest of Albuquerque.[2]

Phase II of the study extended the ASC analysis and used more of the available data to focus on a more diverse set of social factors. The econometric models (probit and tobit) used in Phase I could only analyze a limited amount of information at one time due to modeling problems (specifically, multicollinearity). Study participants were interested in analyzing equity issues involving a more diverse set of social groups. In addition, most participants found block groups to be of little use as units of analysis. Therefore, it was determined that a neighborhood-level analysis would be more amenable to the discussion of policy implications. The data were aggregated to this level, neighborhood ASCs were identified, and statistical analyses were used to determine whether a statistical difference existed between the average number of sites within the ASCs and the average number outside the ASCs.

The study results indicate that on a citywide basis, sites included in the four groups may be distributed disproportionately by levels of community political activity, but not by communities' ethnic or socioeconomic composition. The empirical models incorporated in this study showed no evidence to support a systematic pattern of environmental discrimination against persons of Hispanic origin, low-income communities, or communities with relatively high percentages of persons achieving low educational attainment levels. However, the models did reveal strong patterns in the context of collective choice. The models showed that as the percentage of eligible voters in a block group who actually voted in the 1996 presidential election increases, the probability that an environmental site will be located in that block group and the number of sites in that block group decrease. This result proved to be somewhat robust, appearing throughout most of the estimated models. For the most restricted site grouping (Group 4, which included only the contamination events and Resource Conservation Recovery Act sites that produced high quantities of hazardous waste or were permitted to treat, store, or dispose of hazardous waste), however, only three of the models estimated using this dependent variable yielded evidence of this effect, and that evidence was marginal.

In a more targeted analysis, there was evidence of inequitable site distribution emerging in the context of racial, ethnic, and socioeconomic composition. However, these burdens occurred on both sides of the demographic strata, depending on the type of site being discussed. For instance, primarily non-Hispanic neighborhoods contain on average more hazardous waste facilities within their associated block groups—the opposite of the expected result. These results need to be investigated further to establish their impact, if any, on policy and community action.

The LUST site analysis results imply that these sites *are* distributed inequitably among social groups based on ethnic composition and level of political activity, but that the public funds used to remediate these sites *are not* distributed inequitably in most cases.

Finally, Phase II identified 13 ASCs based on many different social factors. These included the five that were used in Phase I (race/ethnicity, socioeconomic factors,

[2] This targeted analysis was motivated by suggestions from the Albuquerque City Council's original legislation.

educational levels, proximity to the highway, and voter participation), though Phase II defined them differently. In addition, Phase II incorporated elderly people and children living below poverty, unemployment rates, and non-English-speaking communities, among others. The results of the Phase II mean difference tests provide limited evidence that the environmental sites in the study disproportionately burden all but 3 of the 13 identified social groups. The populations living in poverty (particularly, within that category, children and single-female-headed households) are the primary groups impacted by site location. In addition, more sites are found, on average, in neighborhoods with relatively high unemployment rates and relatively politically inactive populations. Finally, the results provide little or no evidence that high migratory rates, high rates of single-male-headed households living in poverty, or high rates of elderly people living in poverty indicate site location or inequitable impact by environmental sites.

Many factors may contribute to the disparities above. One of the main factors is proximity to major transportation corridors, although this effect was considered in most analyses. Other factors that may contribute to the disparate results include the history of residential and industrial growth in the same areas, zoning ordinances, environmental regulations, and property values. However, the study did not try to determine reasons or causes for facility distribution relative to demographics, nor did it try to discern the effects of these factors. Causal relationships were not addressed in this study. The study did not attempt to formally measure potential risks in relation to the environmental sites or the communities in which they exist.[3] These are important issues and naturally follow the subject of the study, but they were beyond the study's scope and budget.

In addition to the report, the study also produced a supplemental atlas, which contains maps that clarify the environmental equity situation as it currently exists in Albuquerque. The maps show locations of the environmental sites, block group demographics, and the ASCs at both the block group and neighborhood levels.

The database used to create this report and the supplemental atlas is readily available for future studies. It is a database of the 1990 US Census data by block group, the 1996 Albuquerque voter results (by precinct, but disaggregated to the block group level), the locations of the environmental facilities addressed in this report, and a detailed summary of public funds disbursed to LUST sites. Its format is GIS-based and its flexibility of form makes it easy to expand on, given available data sources. Reasons for expansion may include extending the project into additional phases and the need to reflect Albuquerque's ever-changing environmental and demographic characteristics.

Data enhancements would facilitate the work to be conducted in subsequent phases of the project. Also helpful would be general coordination within and between city and state agencies, the Albuquerque City Council, local environmental/citizen groups, and the US Environmental Protection Agency. Coordination between agencies has already begun in the form of increased communication between the AEHD and the Bureau. Followup suggestions include expansion of the database to include air quality data and risk measurements. This would allow a risk-based equity study, as opposed to the proximity-based analysis used in this report. Also, public health data could be

[3] A risk ranking, assigned to LUST sites by the Bureau, was incorporated in the analysis of fund distribution among these sites.

incorporated to determine which areas need additional public health intervention and program resources. Further data enhancements would involve updating the 1990 Census data to the 2000 Census data, which would make it possible to compare the two sets of data. This comparison could lead to a "process" equity study (as opposed to this "outcome" equity study), which could begin to explore causal relationships between site location and host community demographic composition.

Taking into account the limitations of the study, the results and their implications should be viewed with caution. Based on the results thus far, promotion of political activity, in the form of increased voting percentage within a community, should be a main concern. The level of political activity within a community is the strongest social indicator of presence and number of environmental sites. Proximity to major transportation corridors is another indicator of site distribution. These two results could have policy implications in the acquisition of Brownfields and Empowerment Zone/Enterprise Community funds. Albuquerque policy makers concerned with environmental justice issues can target areas of relatively low levels of political power and those near major throughways in an effort to obtain funds that can be used in the revitalization.

Given that many of the social groups (e.g., the unemployed, children in poverty) analyzed in Phase II were disproportionately burdened by hazardous materials sites (as defined by the Superfund Amendments and Reauthorization Act), the study participants recommend several policy actions to the City and AEHD. Policy makers should be aware of the locations of ASCs and ensure that new industrial facilities are not located in these areas. In addition, to help alleviate the strain of poverty and unemployment on ASCs, businesses could be given incentives to hire from the communities surrounding them. Finally, all city departments should re-evaluate their programs to ensure that these communities are receiving the services they require.

Health Risk Analysis of the Rio de Janeiro Water Supply Using Geographical Information Systems

Christovam Barcellos,* Kátia Coutinho Barbosa, Maria de Fátima Pina, Mônica MAF Magalhães, Júlio CMD Paola
Dept. of Health Information, Fundação Oswaldo Cruz (DIS/CICT/FIOCRUZ), Rio de Janeiro, Brazil

Abstract

Assessment of health risk in population groups is based on environmental, socioeconomic, and health data. A geographic information system (GIS) was used to simultaneously analyze databases from distinct origins. The case of health risk related to vulnerability of water supply in the city of Rio de Janeiro, Brazil, was examined by using information on the water supply system as well as epidemiological and socioeconomic indicators. The water distribution system covers nearly all city territory and the main treatment plant produces water within the quality guidelines. However, important threats to the population's health remain due to the presence of contamination sources throughout the distribution system and vulnerable small springs. Problems detected involve characteristic city areas, comprising one-third of the total city population.

Keywords: water supply systems, health risk assessment

Introduction

Health risk is the result of a complex interaction between environment and population. Risk factors are comprised of a number of social, environmental, and health variables such as the presence of contamination sources, the geodynamics of contaminants in the environment, population behavior, and the accessibility of exposed groups to education and health care. The integrated analysis of health risks is based on the choice of specific environmental health indicators and their spatial projection (1). Data on each of these factors have different origins and constructive characteristics. The GIS have been used for the gathering, organization, and analysis of large databases on health and environment (2). These systems allow the capture, storage, manipulation, analysis, and display of georeferenced data—i.e., data related to graphic entities representing spatial elements.

The case of water supply in Rio de Janeiro is used here as an example of assembling risk maps from complementary and exchangeable information. Several Brazilian sources of information contain data on the water supply and health conditions. Four information layers were built using demographic and socioeconomic data, as well as data on the water supply system, the distributed water quality, and the number of infant deaths by diarrhea. These information layers were then analyzed.

About 95% of the households in the city of Rio de Janeiro are linked to the public water supply system, the main water treatment plant (WTP) of which (Guandú WTP, Figure 1) produces certified quality water. However, some persistent water supply

* Christovam Barcellos, Dept. of Health Information, Fundação Oswaldo Cruz (DIS/CICT/FIOCRUZ), Avenida Brasil 4365, Rio de Janeiro, RJ 21045-900, Brazil; (p) 0055-21-2901696; (f) 0055-21-2901696; E-mail: xris@dcc001.cict.fiocruz.br

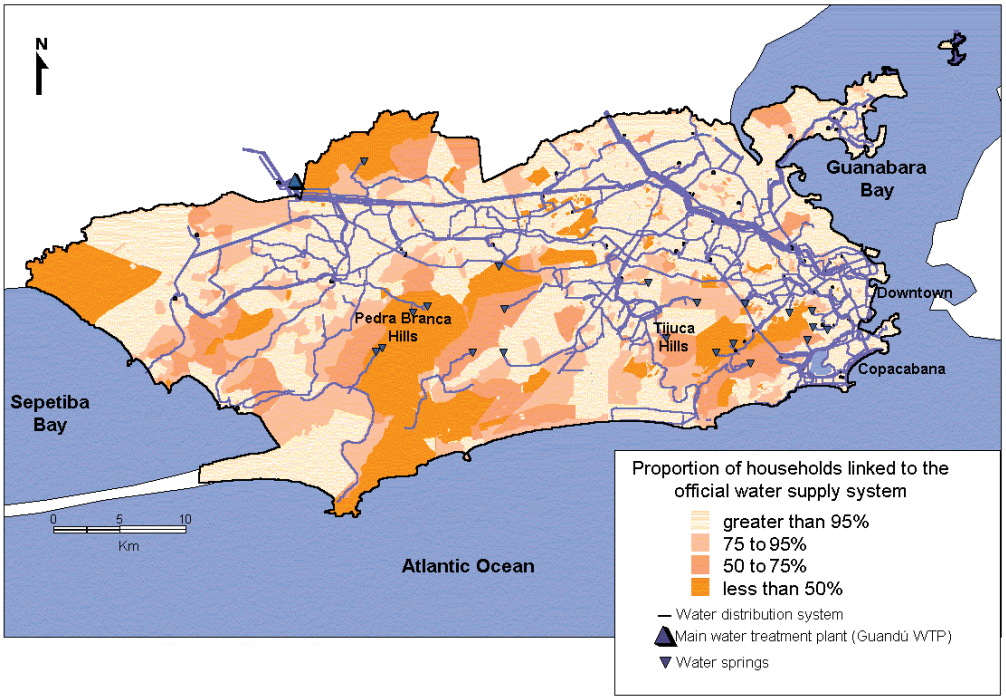


Figure 1 Water distribution system of Rio de Janeiro (location of pipes, Guandú WTP, and small water springs) and the household water supply in census tracts.

issues persist and may lead to negative outcomes in population health. Those issues worth mentioning include water contamination by microorganisms throughout the distribution system; the vulnerability of the system due to the contribution of small and untreated water sources (Figure 1); and, eventually, the absence of this service in some areas. Rio de Janeiro presents an extreme variability in land use for an urban area and a sharp topography, registered in postcards and social statistics. In the city mountainous area, poor settlements ("favelas") and luxury residences share contiguous neighborhoods. A preserved rain forest, protected by a national park, covers a large portion of the hills, where small water springs are localized.

This work aims to identify the main risks related to Rio's water supply as well as who and where the exposed population is. Procedures for assembling and cross-analyzing information layers related to the water supply and epidemiological data are described intending to show different ways of using GIS in ecological health analysis.

Data Acquisition and Management

Each information source constitutes one layer, exhibiting a distinct origin, purpose, and constructive characteristic, enabling spatial operations and population-at-risk calculations in a GIS environment. The city plan of the water distribution system demonstrates the capacity to serve portions of the city located around the pipes. However, the effective provisioning of households relies on other features, such as water pressure, flow regime, and the population's disposition to pay for this service. During the

demographic census, individuals are asked about water origin (official system, wells, or local springs) and domestic provision. The answer to this question should be treated as a manifest about the predominant way the household is provided. Data on water quality are collected by monitoring programs that check water quality indicators (such as the presence of coliform bacteria) that do not imply an immediate health impairment and do not necessarily reflect the overall health risk conditions (3,4). Using epidemiological indicators for assessment of sanitation impact is limited by availability and quality of disease registers. Furthermore, a variation in indicators (e.g., infant mortality or diarrhea incidence rates) cannot be specifically attributed to water contamination because the indicators are also influenced by the subjects' existing educational and economic status and by their degree of health service access (5,6).

Census Tracts Layer

Approximately 6,400 census tract polygons were digitized from analog data obtained from the Brazilian census bureau (FIBGE), in the scale of 1:5,000. Demographic census variables, which include demographic and sanitation information, were associated to a map through census tracts codes.

Epidemiological Data Layer

The neighborhood polygons were obtained from the aggregation of census tracts and stored in a specific layer. Health registries—mortality and live born systems—were georeferenced to 153 neighborhoods.

Water Distribution System Layer

An analog map in scale 1:25,000 locating and identifying reservoirs, water treatment facilities, pumping stations, and main distribution network was obtained from the waterworks agency (CEDAE).

Water Quality Layer

Data on water quality were obtained from the State Environmental Protection Agency (FEEMA) monitoring program. A list of approximately 400 sampling station addresses was digitized and associated to 12,000 sample data referring to concentration of free chlorine, pH, fluoride, color, and turbidity, as well as the presence of total and fecal coliforms in the samples. To georeference sampling stations, the address was matched to a map in scale 1:2,000 for posterior calculation of coordinate pairs.

Identification of Risk Areas

The buffer technique was used to project areas of influence of risk factors. The choice of radius of influence was based on theoretical assumptions. Areas and population affected were identified using the following environmental, epidemiological, and sociodemographic criteria:

- Areas presenting high frequency of coliform contamination, indicating a high risk of water-related diseases transmission. These areas were defined by buffers of 1 kilometer (km) radius around the monitoring points where more than 20% of the samples presented contamination by fecal coliform.

- Areas predominantly served with waters of local origin—i.e., obtained from small springs that are concentrated in Tijuca and Pedra Branca hills (Figure 1). Due to increasing occupation of the hills, local water can be eventually contaminated by human solid and liquid wastes. These areas were defined by buffers of 2 km around small springs.
- Areas distant from the water distribution pipes, defined by buffers of 0.5 km around the arches representing the main distribution network. Large distances between households and the water distribution network can elevate the cost of connecting a house to the pipes, and can incorporate inadequate manipulation practices (7).
- Census tracts where more than 50% of the residents claimed not to be supplied by the official water supply system. This proportion may indicate a collective impairment to accessing the public service.

Table 1 identifies risk populations and areas pointed out according to the four mentioned criteria. Contamination of distributed tap water by coliform (criterion 1) comprises the majority of the population at risk, representing approximately 35% of total municipal population. Use of small local springs (criterion 2) or absence of water distribution network (criterion 3) can also represent risks for a significant portion of the population (10%), which is mainly localized in areas with low population density. A small portion of the population (about 2%) lives in areas where the supply is mostly obtained from alternative water sources (wells and local springs, not explored by the official sanitation company). However, it occupies a considerable area, nearly 16% of city territory. Those are the remaining areas of the city with low population density, where wells (in the case of the western semi-rural region) or springs (mainly in the high areas of the Tijuca Hills) have enough offer of water. These alternatives are impossible in the densely populated eastern areas.

An accumulation of risk factors is verified for some specific socio-spatial groups, which could explain the absence of household connection to the official water supply system. These groups are mainly localized in city areas served by low quality water supply service with incomplete urbanization, presenting both poor "favelas" and luxury residences. The epidemiological impact of the poor sanitation services on each of these groups is divergent. Privileged residents can obtain an alternative water supply, are more informed about water-related diseases, and can be promptly treated in health care facilities.

Table 1 Location, Number of Inhabitants, and Size of Risk Areas According to Different Water Supply Risk Criteria

Risk Criterion	Population (No. of Residents)	Area (km^2)	Location
Water contamination	1,900,000	349	Tijuca Hills northern slope, part of western region
Proximity to local springs	700,000	392	On the city elevated areas
Absence of water distribution pipes	600,000	156	Western region, isolated areas of the northern zone
Use of alternative sources of water	90,000	206	Western region and city, elevated areas

Distances from water sources to sampling points were calculated by using GIS operations. The correlation matrix relating water quality parameters and the distance to water sources is presented in Table 2. Fluoride and color decreases with the distance from the main WTP. Chlorine concentration does not suffer significant decay along the distribution pipes, perhaps reflecting the presence of re-chlorination stations in the distribution network. The proximity to local water springs implies lower mean chlorine concentrations and more frequent coliform contamination. These water sources were thus considered as vulnerability factors to the supply system.

Another risk criterion is the presence of a "sentinel event" in the neighborhood. Diarrhea deaths are dispersed in northern poor areas. In these areas, households are potentially supplied by water but occasional water contamination is observed. Diarrhea events were used to mark risk neighborhoods. According to this risk criterion, about 2.8 million of Rio's inhabitants can be considered as exposed to water-related diseases or suffering serious hampers in the access to health care services. In this case, the sentinel event should activate investigation of possibly related factors of each death: water quality and assistance received in primary health units (8).

Table 2 Correlation Coefficients between the Distance from Water Sources and Water Quality Parameters in the Sampling Stations (Pearson method, n=403)

	Fluoride	Chlorine	Color	Turbidity	Presence of Coliforms
Distance to the main water treatment plant	–.12*	.07	–.11*	–.02	–.09
Distance to local water springs	–.06	.11*	–.02	.05	–.13*

*Statistically significant associations ($\propto$<.05)

Acknowledgments

The author is grateful to Maria de Fatima Pina and Marilia Sá Carvalho for cartographic and epidemiological contributions to this paper.

References

1. Briggs DJ. 1992. Mapping environmental exposure. In: *Geographical and environmental epidemiology: Methods for small-area studies*. Ed. P Elliot, J Cuzick, D English, R Stern. Tokyo: Oxford University Press. 158–76.
2. Loslier L. 1995. Geographical information systems (GIS) from a health perspective. In: *GIS for health and environment*. Org. by P Wijeyaratne. International Development Research Centre, Ottawa.
3. Batallha BHL, Parlatore AC. 1977. *Controle da qualidade da agua para consumo humano. Bases conceituais e operacionais (Control of water quality for human consumption)*. CETESB ediction. São Paulo: Brazil.
4. Barcellos C, Machado JH. 1991. Seleção de indicadores epidemiológicos para o saneamento (Selecting epidemilogical indicators for sanitation). *BIO* 4:37–41.

5. Esrey SA, Potash JB, Roberts L, Shiff C. 1991. Effects of improved water supply and sanitation on ascariasis, diarrhea, dracunculiasis, hookworm infection, schistosomiasis, and trachoma. *Bulletin of the World Health Organization* 69(5):609–21.
6. Heller l. 1997. *Saneamento e saúde (Sanitation and health)*. Panamerican Health Organization, Brasília, Brazil.
7. Drangert JO, Lundguist J. 1990. Household water and health: Issues of quality, quantity, handling and costs. In: *Society, environment and health in low-income countries*. Ed. E. Norberg, D Finer. Karolinska Institutet, Goteborg, Sweden. 71–86.
8. Rutstein DD, Berenberg W, Chalmers TC, Child CG, Fishman AP, Perrin EB. 1976. Measuring the quality of medical care: a clinical method. *New England Journal of Medicine* 294:582–88.

Health Data Mapping in Southeast Toronto: A Collaborative Project

D Buckeridge (1),* L Purdon (2), the South East Toronto Urban Health Research Group (3)

(1) Department of Public Health Sciences, Faculty of Medicine, University of Toronto, Toronto, Canada; (2) Southeast Toronto (SETO) Project, Toronto, Canada; (3) University of Toronto, Toronto, Canada

Abstract

Health data maps and geographic information systems (GIS) are significant resources for health planning and health services delivery, particularly at the local level. The ability to visualize the spatial distribution of health status determinants and indicators can be a powerful resource for mobilizing community action to improve the health of residents. Currently, health data maps and other GIS applications tend to be highly technical and specialized, and are therefore of limited use to community members and organizations providing community-based health services. Developing relevant, accessible, and usable GIS and health data maps for communities and local agencies is an important step toward enabling individuals and communities to improve their health and increase their control over it. This collaborative interdisciplinary project harnesses the energies and skills of community and university partners in the joint design and critical assessment of a relevant and accessible GIS targeted toward respiratory health in an urban community. A respiratory health data model is used to identify appropriate existing data sources and a comprehensive metadata model facilitates assessment of data sources. Community and university partners are collaborating to design and assess the GIS through a series of hands-on workshops. Qualitative methods are employed to examine the nature and effectiveness of the collaborative process. This project has identified, and is attempting to address, a number of technical and collaborative issues. Technical issues include integrating disparate datasets and developing appropriate methods of data depiction for varying levels of users. Collaborative issues include overcoming substantial diversity of user needs, capacities, and perceptions. Lessons learned from this project are applicable to other projects involving health information system design and university-community collaboration.

Keywords: information system design, collaborative research, community health

Introduction

Health data maps and geographic information systems (GIS) are significant resources for health planning and health services delivery, particularly at the local level. More specifically, the ability to visualize the spatial distribution (e.g., by neighborhood or city block) of health status indicators and other health-related information (such as air

* David L Buckeridge, University of Toronto, Dept. of Public Health Sciences, 2-126 Withrow Avenue, Toronto, Ontario M4K 1C9 Canada; (p) 416-469-5543; E-mail: david.buckeridge@utoronto.ca

quality as modeled from traffic and air flow patterns) can be a powerful resource for mobilizing community action to improve the health of residents.

Currently, health data maps and other GIS applications tend to be highly technical and specialized, and are therefore of limited use to community members and organizations providing community-based health services. Developing GIS and health data maps that are relevant, accessible, and useable for communities and local agencies is an important step in realizing the goal of "enabling individuals and communities to increase control over and to improve their health," to cite the 1986 Ottawa Charter for Health Promotion (1).

Background

To develop health data maps and other GIS applications that are relevant, accessible, and useable, it is crucial to work closely with potential user groups in the community. The project described here has evolved from a collaboration between members of the Southeast Toronto Organization (SETO), Toronto, Canada.[1] SETO is a consortium of community members and community agencies providing health and social services in southeast Toronto—a diverse urban area with a population of 120,000.

Before the initiation of the current project, the SETO partners made a number of advances through collaboration. Through multi-faceted community consultation, respiratory health (especially asthma and related conditions) was identified as an area of focus for this project. Further groundwork was laid for this project by exploring the availability and nature of routinely collected health data potentially suitable for mapping; testing the feasibility and utility of developing health data maps (including the sponsorship of a community workshop to test preliminary "health maps" in a mixed audience of professional, scientific, and lay project stakeholders); and development of a project agenda that meets the requirements of all SETO partners.

Objectives

The entire project is based on the premise that the information resource being developed is more likely to be useful for community members and community health agencies if it is designed, from the start, in collaboration with these end-users. In this project, in addition to creating a GIS, SETO seeks to provide an analysis of the development process. The project's objectives, therefore, address the product and process of the collaboration between these partners.

Product

The first goal of this project is to produce an interactive, online GIS that is developed and iteratively refined through active collaboration between SETO partners. The GIS will integrate a wide range of routinely collected information relevant to the determinants and manifestations of respiratory health (particularly asthma and related chronic/recurrent conditions) in the population of southeast Toronto. A fundamental

[1] SETO partners include: Toronto Department of Public Health, Central Neighbourhood House, South Riverdale Community Health Centre, Regent Park Community Health Centre, community residents, University of Toronto, Wellesley Hospital, Central Hospital, St Michael's Hospital.

aspect of the GIS is that it will allow users to access information in a range of formats from collaboratively pre-designed maps to raw data.

Process

The second goal is to document the conceptual, group-process, and logistical/technical problems in the collaborative development of such a GIS for community use, and to describe the solutions found by the project for these problems.

Knowledge Development and Synthesis

The third goal for this project is to analyze and report on potentially generalizable lessons learned concerning both the technical process of GIS development and the social process of collaborating on such a task. In essence, the project seeks to identify factors potentially affecting the success of any collaborative applied research aimed at solving local health problems.

Implementation of Project

Project implementation was preceded by consultation with each SETO member organization to assess their needs and expectations of the project, as well as their current data processing and analysis capability in terms of both hardware/software and personnel. Following the needs assessment, project implementation involves four iterations through two steps: GIS development, then joint critical assessment of the GIS with community-based users.

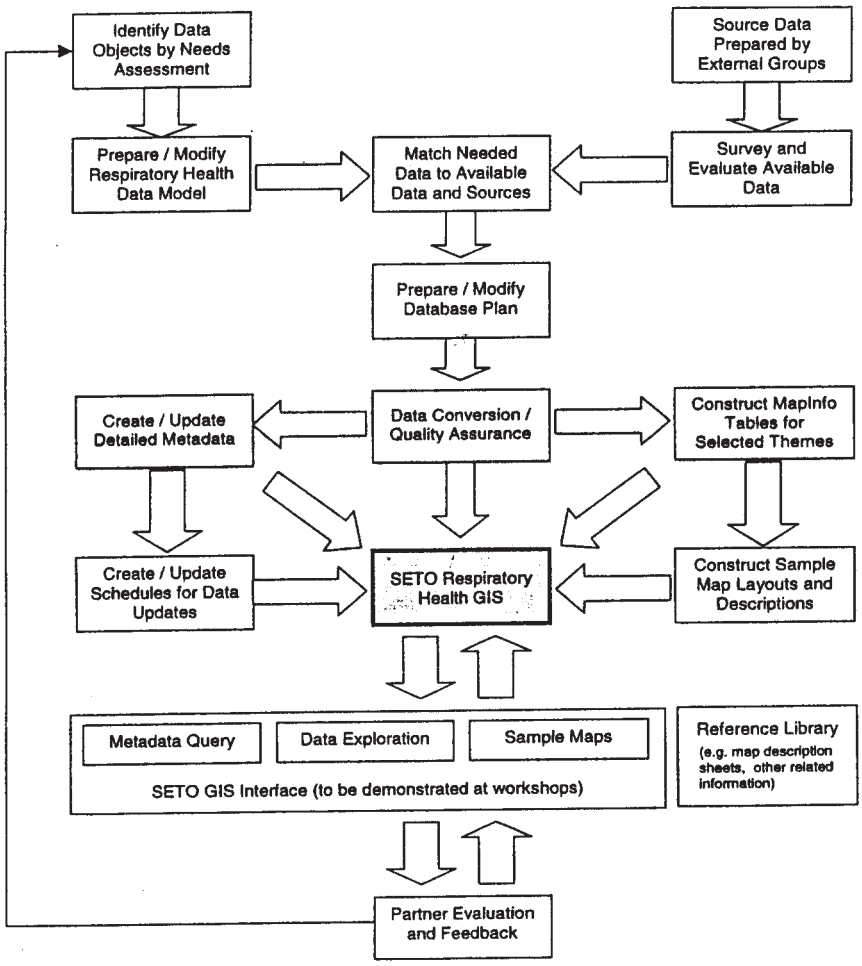
Dividing tasks between a technical team and a collaborative process team facilitates the project implementation. In each iteration, the technical team assesses the needs and capacities of various prospective users among SETO partners, assesses data sources, and produces iterations of GIS prototypes. A user group made up of representatives from SETO member organizations provides feedback on these GIS prototypes. The collaborative process assessment team then analyzes the user group's responses to the prototypes, as well as the technical team's response to that feedback, as obtained through a series of participatory evaluative workshops.

GIS Development

Figure 1 illustrates the iterative process used to develop the respiratory health GIS. This process involves developing data models and assessing candidate data sources, then designing and producing the GIS. The technical team develops a respiratory health data model (Figure 2) to facilitate identification and assessment of candidate data sources. This model attempts to describe the relationship between determinants and indicators of respiratory health.

Criteria for evaluating candidate data sources were developed from a comprehensive metadata model developed for the project. The criteria include the following aspects of a data source: quality, completeness, relevance, ease of integration, potential for misinterpretation, and cost (if any).

The needs assessment discovered a wide range of capabilities among the SETO partners. This finding suggests that the GIS should accommodate this range of capabilities. The technical team therefore decided to base the GIS upon user-friendly, PC-based



Modified from the Department of Geography, University of Buffalo GIS Development Guide. Volume II – Survey of Available Data. http://www.geog.buffalo.edu/ncgia/sara/foursurv.htm

Figure 1 Data flow diagram for iterative design, production, and assessment of SETO respiratory health GIS.

software including MapInfo (MapInfo Corporation, Troy, NY) and Microsoft Excel, and to devise a system of layered access within the GIS.

The system of layered access provides access to information at varying levels of complexity. Users interested in more complex information can access raw data in Excel format or MapInfo tables. Such users can then create maps or perform other analyses with these data. Users interested in a less complex presentation can access collaboratively pre-designed maps (i.e., MapInfo layouts) and basic analyses. Two examples of pre-designed maps are shown in Figure 3.

So far, the technical team has faced a number of challenges in developing the GIS. These include integrating data sources of varying quality, scale, and "ownership";

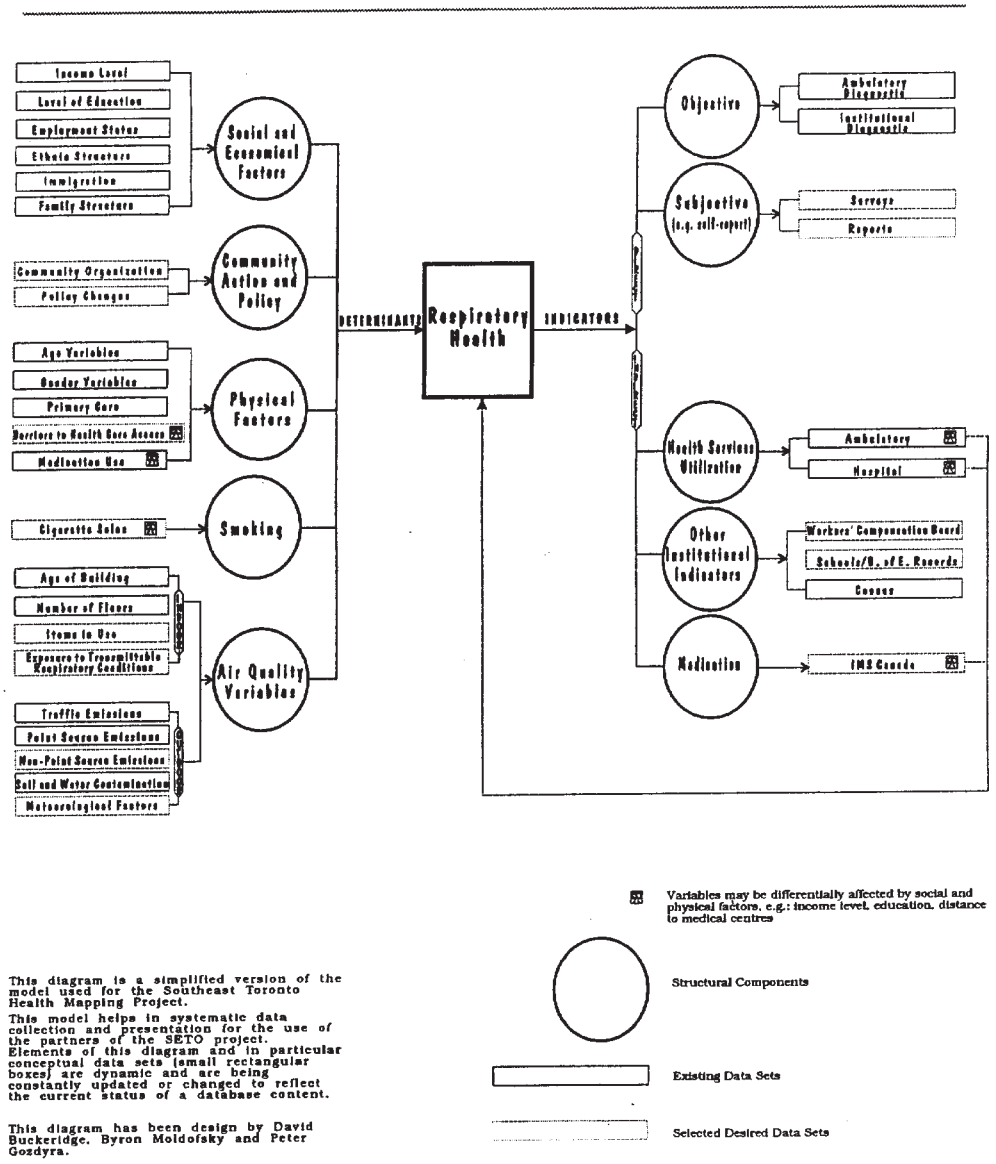


Figure 2 Respiratory health data model.

substantial diversity of user needs, capacities, and perspectives; and the need for data depictions that achieve the essential compromises between various stakeholders' concerns. Depicting qualitative data and integrating them with quantitative data present unique challenges.

Joint GIS Assessment

A joint assessment of the GIS product is conducted through a series of participatory evaluative workshops involving all the project's collaborators. These workshops

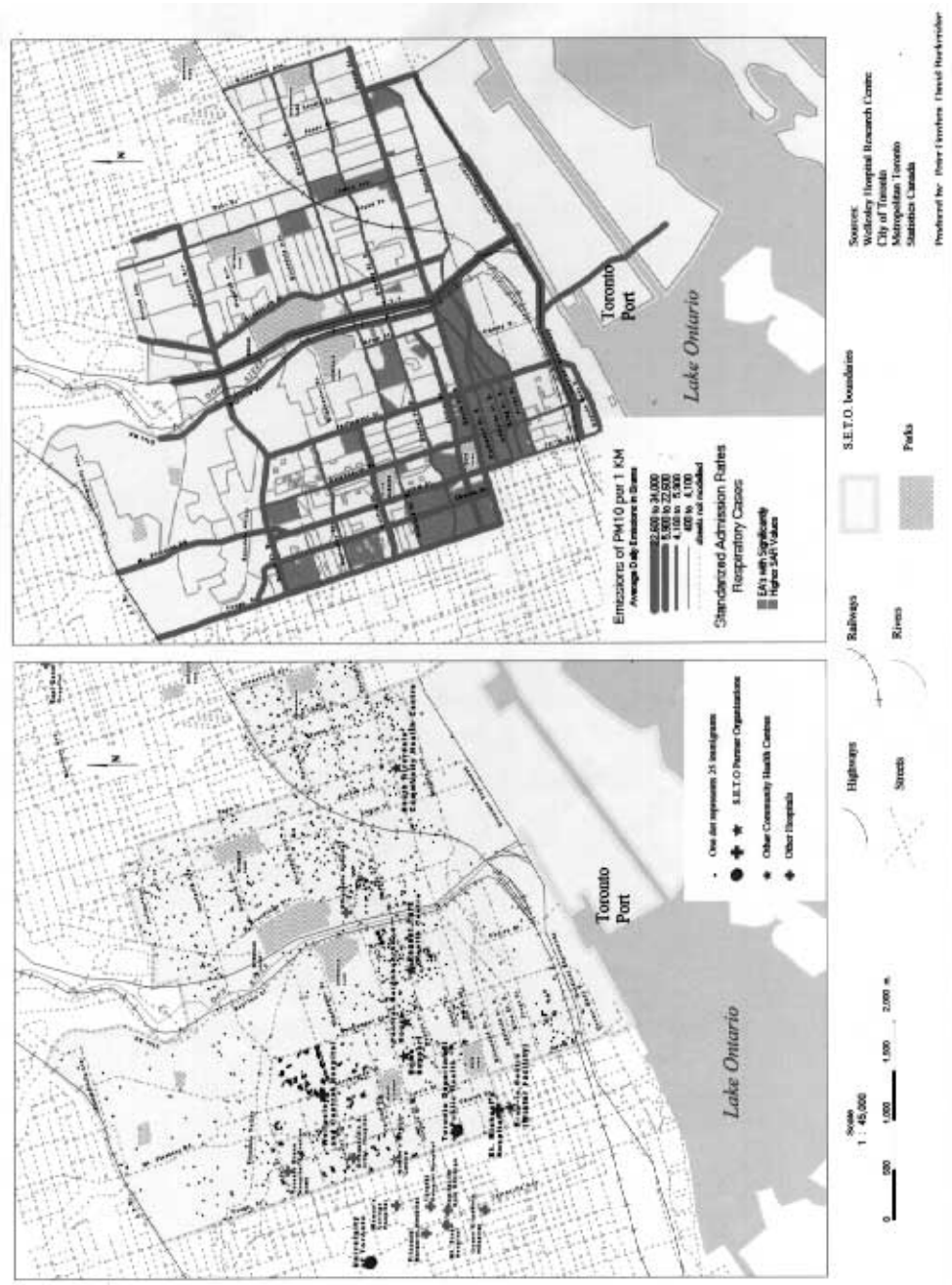


Figure 3 Pre-designed maps, Southeast Toronto Health Mapping Project.

provide feedback to the technical team. They also serve to meet the project's third objective, to assess what factors facilitate or impede collaborative projects of this kind.

The development schedule calls for the technical team to produce, at four points in the project's course, working versions of the GIS for presentation to the user group in day-long workshops. These workshops provide opportunities for hands-on, interactive demonstration and use of the prototype GIS. Feedback from the user group about the quality, relevance, and usability of the GIS is provided to the technical team, which attempts to address the issues raised by revising the GIS for the next iterative loop (i.e., the next workshop) in the collaborative development process.

Immediately following each workshop, the collaborative process assessment team conducts separate focus group debriefing sessions with the technical team and the user group to determine issues of collaboration that arose in the course of the workshop. The sessions are taped and transcribed for qualitative data analysis. Transcripts of the focus group sessions are subjected to a content analysis according to standard qualitative methods. The results of preliminary analysis are presented to a combined meeting of the technical team and the user group and for respondent validation and for discussion. The emerging themes from each set of focus groups are to be carried forward until after the last workshop, at which point the themes will be used to develop a coding scheme that will be applied to transcripts of all focus group sessions.

Discussion of Important Issues

Focus on Public Health Problem Identified by the Community

Community members, community agencies, and researchers have identified respiratory health as an important public health problem in southeast Toronto. Concern among community members in one area of southeast Toronto led to the establishment of an indoor air quality committee some years ago. This committee has held regular workshops and educational events for the surrounding neighborhoods, including participatory theatre performances. In another ethno-culturally diverse area of southeast Toronto, a citizen committee associated with a community health center has conducted a survey of symptoms and other aspects of asthma and its prevention. This survey was prompted by the perception that many new Canadians in southeast Toronto were experiencing respiratory conditions for the first time. Community agencies in southeast Toronto have also worked on the problem of respiratory health for some time in an attempt to achieve a shift in resources from treatment to prevention, including removal of inequities in distribution of risk conditions and access to high-quality care. Finally, epidemiological research done in southeast Toronto has shown a twofold elevated risk of hospital admission for respiratory problems in Regent Park (an area of public housing in southeast Toronto), after controlling for potential confounders.

Community Involvement

Community involvement has always been an essential aspect of this project. As described above, initiatives in the area of respiratory health from concerned community members stimulated the development of this project. Multi-faceted community consultation served to identify health maps and GIS as potentially useful tools for facilitating ongoing activities by community members and community agencies. Furthermore, in

addition to providing definition and scope for the project, community involvement continues to shape the project through the collaborative workshops that are at the heart of the project.

Examination of the Collaborative Process

A decision was made at an early point in project design to examine the effectiveness of the collaborative process rather than the project outcome. This approach was taken because, despite the current enthusiasm for collaborative research, there are virtually no studies that are convincing about the benefits and challenges of this kind of collaboration, or that document and analyze the factors that facilitate or impede collaborative efforts. Collaborative research of this nature requires that people from different backgrounds and institutional cultures, with different demands, incentives, and practices, work together. In so doing, they must overcome conceptual, communication, organizational, and technical differences and challenges. In addition, issues of power and the legitimacy of certain kinds of knowledge arise. In collaborative research, all participants are challenged to ask critical questions about the approaches they take, as well as about the nature, relevance, and efficacy of their respective practices.

Potential to Replicate Project

One of the objectives of this project is to report on potentially generalizable lessons learned concerning both the technical process of GIS development and the social process of collaboration in this task. These lessons should help others to replicate the technical portion of this project. Furthermore, the use of existing datasets, a PC platform, and user-friendly software has kept the cost and expertise required to implement this project to a minimum.

References

1. Health and Welfare Canada. 1986. *Ottawa charter for health promotion.* Presented at the First International Conference on Health Promotion, Ottawa, Canada. November 17–21, 1986. Ottawa: Health and Welfare Canada. http://www.who.dk/policy/ottawa.htm

Exposure Assessment for Trichloroethylene in Drinking Water Using a Geographic Information System

X Chen,* CE Feigley, EM Frank, WA Cooper, Y Huang
School of Public Health, HESC, University of South Carolina, Columbia, SC

Abstract

The Agency for Toxic Substances and Disease Registry's (ATSDR's) baseline survey of its Trichloroethylene (TCE) Subregistry found that there were excess speech and hearing problems among Subregistry children 10 years old and younger compared with national data. In the project described in this paper, drinking water analyses were used to assess TCE exposure of 318 children listed in the Subregistry. Geographic information system (GIS) models were used in conjunction with mathematical interpolation to explore the spatial and temporal variation of TCE exposure. Yearly and cumulative exposures were estimated. The subjects were categorized into subgroups in terms of exposure level, exposure duration, cumulative exposure, and geographic location. The potential confounding exposures were identified and characterized. At one of the six sites in the Subregistry—a site in Rockford, Illinois—GIS detected a contaminant plume spreading in an east-west direction. The estimated cumulative exposures were significantly associated with subject's geographic location and age at the time of remediation. No major changes in the geographic distribution pattern of TCE were observed at the Rockford site over a six-year period, except that the concentration generally increased. The subjects at this site could be categorized into two geographic subgroups: inside the plume area and outside the plume area. It was found that the cumulative exposure ranged from 0 to 59,210 parts per billion (ppb) per year, with a mean of 700 ppb per year. Four subgroups were created based on cumulative exposure levels. The mean exposure duration was 5.5±2.9 years. Of the 198 households in the study, 118 had water with measured TCE concentration exceeding 5 ppb, the EPA maximum allowable level for TCE in drinking water. The TCE exposure was significantly correlated with the potentially confounding exposures at the Rockford site, but not at another site, in Elkhart, Indiana.

Keywords: trichloroethylene, water, exposure

Introduction

Trichloroethylene (TCE) is an organic chemical that has been widely used as a dry cleaning agent, a metal degreaser, a rubber solvent, and an ingredient in printing ink, paper, lacquer, and varnish (1). It also is a frequent contaminant of both untreated and treated drinking water in the United States. Various surveys conducted by the US Environmental Protection Agency (EPA) showed that about 38% of cities sampled had TCE in their drinking water (2). Spills and leaking storage tanks readily contaminate groundwater. The TCE released into soil can also migrate to groundwater, where the chemical remains for months to years (3). High-dose TCE exposure causes central

* Xiaowu Chen, Caliber Associates, 10530 Rosehaven St., Suite 400, Fairfax, VA 22030 USA; (p) 703-385-3200; (f) 703-385-3206; E-mail: chenx@calib.com

nervous system depression, and trigeminal nerve toxicity was among the most often reported adverse effects in case studies (4,5,6).

To facilitate studying the effects of environmental exposure to TCE, the Agency for Toxic Substances and Disease Registry (ATSDR) established a subregistry of people exposed to TCE through their water supplies as part of the National Exposure Registry (7). At the sites covered by the TCE Subregistry, groundwater was contaminated by local industries—wastes containing TCE and other volatile organic compounds were disposed of on site in lagoons or injected into surface soil, resulting in contamination of soil, sediment, and groundwater. To be included in the Subregistry, a person must have used a private well as a water source, and TCE must have been detected in at least one water sample from their home water supply.

ATSDR's baseline survey of the TCE Subregistry revealed excess self-reported speech and hearing problems among the Subregistry population 10 years old and younger compared with National Health Interview Survey (NHIS) data for the same age group. However, these data cannot be used to establish a causal relationship between TCE exposure and impairment of speech and hearing.

To explore the association between speech and hearing impairment and TCE exposure, an ongoing study is evaluating the speech and hearing of the children in the Subregistry who were 10 years old or younger when the Subregistry was established. Results are to be compared with speech and hearing data from control children, matched by age and race, who have not been exposed to TCE or other relevant risk factors. Because a preliminary examination of well water data revealed a wide range of TCE concentrations in the exposed communities, categorizing Subregistry children by exposure level was highly desirable.

TCE has been shown to cause hearing deficits in laboratory animals (8,9). In one study, young animals were much more sensitive than older animals, indicating that the developmental stage may be an important factor in assessing toxicity (8). Estimating exposure as a function of age was therefore an important objective, because it would allow participants to be categorized by exposure during specific age ranges. Thus, an important objective was to assess TCE exposure as a function of participant age, because it would make it possible to investigate the impact of human TCE exposure on speech and hearing during vulnerable developmental stages.

The Subregistry consists of 814 households in six towns—Albion, Michigan; Byron, Illinois; Elkhart, Indiana; Rockford, Illinois; Roscoe, Illinois; and Verona, Michigan. Of the 814 households, the 198 that had children 10 years and younger were selected for the speech and hearing study. The home water TCE measurements available for exposure assessment numbered 248 for the households to be studied and 1,069 for the entire Subregistry (7). Only 20% of the households had more than one sample from their water supplies analyzed (7). This sparseness of data meant that, unless results could be spatially and temporally interpolated, accurate assessment of exposure over a child's life would have been impossible. Thus, the purpose of this effort was to explore new methods for dealing with sparse exposure data, using the data from all the Subregistry households to estimate exposure for the households in the study.

The historical exposure was modeled using a geographic information system (GIS) and a mathematical approach for temporal interpolation. Specific questions addressed were whether spatial and temporal trends exist in the measured TCE concentration data and whether the exposure to potentially confounding environmental contaminants is

correlated with the TCE exposure. Exposure information for each household included household ID, subject ID, street address, city, state, chemical name and concentration in parts per billion (ppb), and sampling date. The data were checked for invalid entries, such as negative concentrations, extremely high values, missing values, or invalid time data. Questionable data were verified from ATSDR documentation.

Methods

Total human exposure to a contaminant is the product of contributions from all environmental media (water, soil, air, and food) that contain the contaminant and all routes of entry to the human body (dermal contact, ingestion, and inhalation) (10). Ingestion and inhalation are thought to be the major routes of exposure for people in these communities who have no occupational contact. McKone and Knezovich (11) investigated the potential inhalation exposure resulting from vaporization of TCE from water in homes. They showed that TCE inhalation exposure is approximately proportional to the chemical's concentration in water; because ingestion exposure is also proportional to the chemical's concentration in water, this means that inhalation exposure to TCE is proportional to ingestion exposure. McKone and Knezovich's experiments revealed that the transfer efficiency of TCE from shower water to air has an arithmetic mean value of 61%. We have assumed that total exposure is proportional to the concentration in well water. Groundwater TCE values were therefore used as an index of total exposure.

Total human exposure by ingestion may be defined by Equation 1:

$$E = \int_{t1}^{tn} C(t)dt \tag{1}$$

where E is the cumulative exposure, *C(t)* is the time-varying concentration of a contaminant (TCE in this case), *t* is time, and t_1 and t_n are the beginning and end times of the period of interest.

An approximate solution for Equation 1 is the application of the trapezoid method, which is illustrated in Equation 2:

$$\int_{t1}^{tn} C(t)dt = \frac{1}{2}(C_0+C_1)\Delta t_1 + \frac{1}{2}(C_1+C_2)\Delta t_2 + \cdots\cdots + \frac{1}{2}(C_{n-1}+C_n)\Delta t_n \tag{2}$$

where Δt_n is the time increment from t_{n-1} to t_n. The TCE concentrations at each time must be determined when solving the above equation.

GIS Modeling

The general functions of the GIS include data input and geocoding, geographic visualization, geographic editing and querying, and spatial analysis. When assessing exposure, one must combine geographic, demographic, and environmental databases to describe temporal and spatial characteristics of exposure. This is very difficult to do using a routine numerical approach. GIS, however, makes it easy to manipulate layered, spatially distributed data, thereby significantly simplifying the database management, visualization, and spatial analysis involved in exposure assessment (12).

This paper uses the Rockford site as an example to describe the methods and results obtained in this study because that site contains more than 50% of the study subjects and almost 50% of the households in the study. The first step in analyzing the Rockford site was to geocode subject households and match them with the digital map extracted from the US Bureau of the Census TIGER/Line census file (13). The Rockford samples were collected in four periods over six years: June to November 1984, January to November 1985, August 1988 to September 1989, and October 1989 to February 1990. Most households, however, had samples for only one of the four periods. For each sampling period, the TCE concentrations for households in the study that were not sampled were estimated by GIS spatial interpolation using data from neighboring households that were in the Subregistry (14).

There are several interpolating methods available in GIS, including inverse distance weighting (IDW), spline, and kriging (15). The two most frequently used methods, IDW and kriging, were applied to interpolate surface data in this study. In IDW interpolation, each input point has a local influence that diminishes with distance; points closer to the location of the estimate are weighted more heavily than those farther away. Kriging regards the statistical surface to be interpolated as a regionalized variable that has a certain degree of continuity. The kriging algorithm minimizes the variance of error.

Temporal Interpolation by Lagrange's Formula

With data from GIS interpolation included, each household in Rockford had TCE values for 1984, 1985, 1989, and 1990. Exposures occurring between 1985 and 1989, however, still needed to be determined to solve Equation 2. To account for these periods without TCE samples, we applied Lagrange's interpolation formula (16). Let $c(t)$ be the polynomial of degree n, which, for values $t_1, t_2, \ldots$, and t_n of the argument t, has the values $c_0, c_1, c_2, \ldots$, and c_n, respectively. Lagrange's formula is shown in Equation 3:

$$c(t) = \frac{(t-t_1)(t-t_2)\ldots(t-t_n)}{(t_0-t_1)(t_0-t_2)\ldots(t_0-t_n)}c_0 + \frac{(t-t_0)(t-t_2)\ldots(t-t_n)}{(t_1-t_0)(t_1-t_2)\ldots(t_1-t_n)}c_1 + \ldots\ldots\ldots + \frac{(t-t_0)(t-t_1)\ldots(t-t_{n-1})}{(t_n-t_0)(t_n-t_2)\ldots(t_n-t_{n-1})}c_n \quad (3)$$

where t_n is the nth time period, c_n is the TCE value corresponding to the nth time period, and $c(t)$ is the TCE value to be interpolated at time t.

Equation 4 was used to account for exposure in utero, under the assumption that subjects were exposed to TCE since their conception:

$$E_0 = E_1 \times 0.75 \quad (4)$$

where E_0 is the in utero exposure level and E_1 is the exposure level for the first year after birth.

According to the time when a subject entered and left the exposed group, an exposure duration in years was computed for each subject. The starting and ending times for each site were different, so the exposure of each subject was calculated individually. The cumulative exposure in ppb per year was estimated by integrating the exposure level over the exposure duration. GIS and repeat measure analysis (taking multiple measurements of the same observational unit) were used to explore temporal trends visually and statistically.

Confounding Exposure

The four other chemicals most frequently found in the water supply of people in the Subregistry—trichloroethane (TCA), dichloroethane (DCA), dichloroethene (DCE), and perchloroethylene (PCE)—were included in an analysis of confounding exposures. Although a literature search did not find any direct evidence that these chemicals cause speech and hearing impairment (17–24), they have central nervous system effects, including depression, irritability, and dementia, at high exposure levels. Therefore, chronic exposure to these chemicals may have some as yet undetected effects on speech and hearing.

Correlation analysis was conducted using SAS software (SAS Institute, Inc., Cary, NC) and GIS to detect the relationship between TCE and the potentially confounding exposures. The subjects were stratified into subgroups according to their levels of potential confounding exposures.

Results

The Rockford Site

There were 330 Subregistry households at the Rockford site, including 98 subject residences. The number of samples collected during each sampling campaign is listed in Table 1. The four maps in Figure 1 were created by interpolating surfaces (using IDW) from samples collected in 1984, 1985, 1989, and 1990, respectively.

Table 1 Number of Samples for Each Sampling Period at the Rockford, Illinois, Site

Sampling Period	Midpoint of Sampling Period	Years of Exposure[a]	Number of Samples
June 1984–November 1984	August 1984	4.67	39
January 1985–November 1985	June 1985	5.50	41
August 1988–September 1989	February 1989	9.17	82
October 1989–February 1990	December 1989	10.00	203

[a] Years of exposure were determined by calculating the time between January 1, 1980 (when exposure at Rockford is assumed to have begun), and the midpoint of the time period.

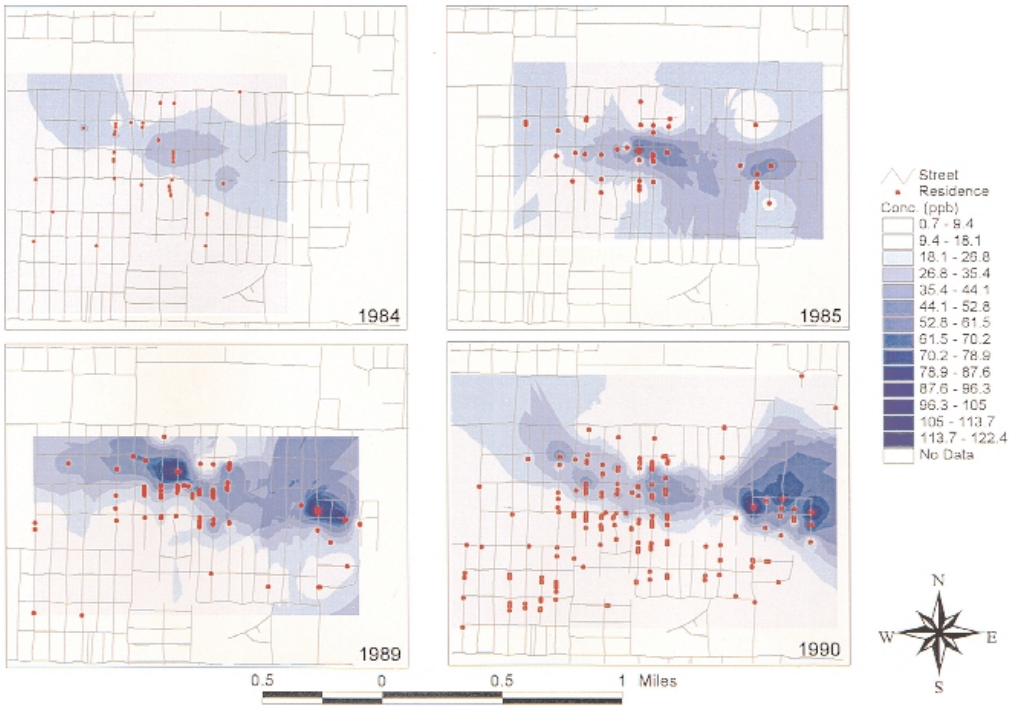


Figure 1 Well water trichloroethylene plume at the Rockford, Illinois, site.

Comparison of IDW and Kriging in Exposure Assessment

Fifty-seven households were covered by both IDW and kriging. The IDW and kriging exposure estimates were compared for the 89 subjects living in these households (Table 2). T-tests did not find a significant difference between the exposures estimated by the two methods (p=.0724). Thus, the results of IDW interpolation were used in later analyses because the IDW algorithm was more stable than kriging for the GIS program used here.

Table 2 Comparison of Inverse Distance Weighting and Kriging by Cumulative Exposure for the Rockford, Illinois, Site

		Cumulative Exposure (parts per billion per year)							
Method	**Number of Subjects**	**Mean ±SD[a]**	**Maximum**	**95%**	**90%**	**75%**	**50%**	**25%**	**Minimum**
IDW	89	173±108	442	383	335	247	166	81	7
Kriging	89	184±102	396	380	309	255	188	98	8

[a] SD=standard deviation

p=.0724 at α=.05

Temporal Trends of TCE Contamination

The surfaces were interpolated from 396 samples collected at the four time periods covering 1984 through 1990. Although TCE levels were not completely the same in the four sampling periods, GIS spatial interpolation indicated that the overall geographic distribution of TCE contamination was similar. A ribbon-shaped contaminant plume extending in a southeast-northwest direction was first found in 1984, consistently appeared in 1985 and 1989, and became most apparent in 1990, when sampling coverage (203 households) was much denser than in previous years. The heavily contaminated area was always located between Road 1 and Road 2, with TCE concentration on the eastern side of the site higher than on the western side. No significant south or north movement of contamination was observed. For example, TCE levels in most areas south of Road 2 were always around 0 to 18 ppb over the six years covered by sampling. The more heavily contaminated area had TCE values ranging from 18 to 139 ppb. On the other hand, repeated measures analyses indicated the existence of time effect (p=.0001). Therefore, even though GIS found no major qualitative changes in the TCE distribution pattern over the six years, comparison of the TCE plumes in Figure 1 suggests a trend of generally increasing concentration with time. Thus, it is unreasonable to assume that the TCE level in households was constant over the six years for which water was sampled.

Geographic Variation of TCE Exposure

The ribbon-shaped plume flowed in an east-west direction and separated the site into three geographic areas: north of the plume, inside the plume, and south of the plume (see the 1990 map in Figure 1). The TCE exposure of the subjects living at these three areas was compared by the SAS General Linear Models (GLM) procedure, which analyzes variance. The model, including independent variables of area and age, was statistically significant (p=.0001). A contrast test among the three areas indicated that the cumulative exposure of the subjects inside the plume area was significantly higher than that of the subjects north or south of the plume (p=.0001). No statistically significant difference between the exposure of the subjects north of the plume and south of the plume was detected (p=.4491). The GLM procedure also found that there was no significant difference between the areas in terms of subjects' age or exposure length (p>.1). With the areas above and below the plume combined, exposure inside the plume was found to be significantly higher (p=.0001, Table 3) than exposures outside the plume. It was concluded that the TCE exposure inside the plume area was higher than exposure in the other areas at the Rockford site.

Table 3 Descriptive Statistics of Trichloroethylene Exposure for the Rockford, Illinois, Site

		Cumulative Exposure (parts per billion per year)							
Area	**Number of Subjects**	**Mean ±SD[a]**	**Maximum**	**95%**	**90%**	**75%**	**50%**	**25%**	**Minimum**
Within plume	64	225±137	768	434	383	276	229	124	7
Outside plume	111	47±56	229	162	137	58	20	9	0

[a] SD=standard deviation

p=.0001 at α=.05

Quantitative Exposure Assessment

The exposure between 1985 and 1989 was interpolated by Lagrange's formula rather than a regression model because each solution has only 4 degrees of freedom. The interpolated curves were developed for each of the 98 subject households in Rockford. Comparison of these curves indicated the following features of Lagrange's interpolation depending on the relationship of the known values.

- $c(t_1)>c(t_2)$ *and* $c(t_3)<c(t_4)$*:* a concave curve was generated, indicating that the interpolated values were lower than the four known values (Figure 2a).
- $c(t_1)<c(t_2)$ *and* $c(t_3)>c(t_4)$*:* a convex curve was generated, indicating that the interpolated values were higher than the four known values (Figure 2b).
- $c(t_1)<c(t_2)$ *and* $c(t_3)<c(t_4)$, or $c(t_1)>c(t_2)$ and $c(t_3)>c(t_4)$: a sinusoidal curve was generated, indicating a trend of increase-decrease-increase or decrease-increase-decrease (Figures 2c and 2d).
- $c(t_1)<c(t_2)<c(t_3)<c(t_4)$: a monotonically increasing curve was generated.

When the slope of a line between two sequential concentration points is very high or very low, Lagrange's interpolation can produce unrealistic concentration estimates. However, this happened only once in this dataset. This single unrealistic estimate—a negative value—was replaced with the average value of the samples from the five nearest neighboring points.

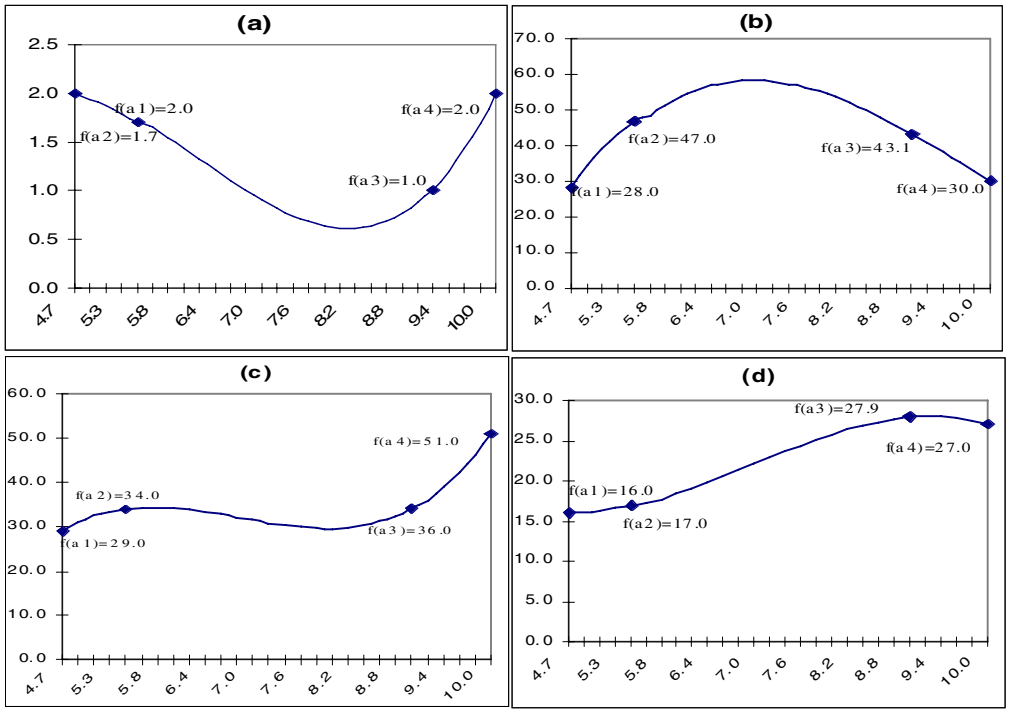


Figure 2 Lagrange's interpolation curve.

Creation of Exposure Subgroups

The subjects were categorized into subgroups to facilitate various types of comparison. The subgroups were generated from the following criteria: geographic location, cumulative exposure, length of exposure, and maximum TCE levels.

Table 4 illustrates the subjects' grouping by cumulative TCE exposure. This scheme considered both level of TCE contamination and length of exposure. The subjects were divided into four subgroups: 0 to 15 ppb per year, 15 to 100 ppb per year, 100 to 550 ppb per year, and more than 550 ppb per year. The mean cumulative exposure was 652 ppb per year.

Table 4 Descriptive Statistics of Cumulative Exposure for All Six Study Sites

		Cumulative Exposure (parts per billion per year)							
TCE Group (parts per billion per year)	Number of Subjects	Mean ±SD[a]	Maximum	95%	90%	75%	50%	25%	Minimum
0–15	101	6±5	15	14	12	10	6	2	0
15–100	86	44±24	99	95	80	56	39	23	6
100–550	98	244±103	527	442	402	307	223	162	101
>550	33	5,900±13,097	59,210	51,570	7,816	4,090	1,566	810	556
Total	318	701±4,524	59,210	1,829	573	229	51	10	0

[a] SD=standard deviation

Table 5 summarizes the length of exposure, including the adjustment for in utero exposure. The subgroups created were 0 to 2.75 years, 2.75 to 5.75 years, 5.75 to 7.75 years, and more than 7.75 years.

Table 5 Descriptive Statistics of Exposure Length for All Six Study Sites

		Exposure Length (years)							
Exposure Length (years)	Number of Subjects	Mean ±SD[a]	Maximum	95%	90%	75%	50%	25%	Minimum
0–2.75	81	1.8±0.9	2.75	2.75	2.75	2.75	1.75	1.00	0
2.75–5.75	108	4.8±0.8	5.75	5.75	5.75	5.75	4.75	3.75	0
5.75–7.75	65	7.3±0.5	7.75	7.75	7.75	7.75	7.75	6.75	0
>7.75	64	9.7±0.8	10.75	10.75	10.75	10.75	9.75	8.75	0
Total	318	5.5±2.9	10.75	10.75	9.75	7.75	5.75	2.75	0

[a] SD=standard deviation

Table 6 illustrates the TCE exposure grouping by site. The Roscoe site had the highest exposure among the six sites, with a mean of 2,128 ppb per year. The next highest exposure occurred at the Elkhart site, which had a mean of 1,564 ppb per year.

Exposures at the Albion and Verona sites were the lowest, with means of less than 100 ppb per year.

Table 6 Descriptive Statistics of Trichloroethylene Exposure by Site

Site	Number of Subjects	Cumulative Exposure (parts per billion per year)							
		Mean ±SD[a]	Maximum	95%	90%	75%	50%	25%	Minimum
Albion	7	74±137	357	357	357	153	0	0	0
Byron	8	682±923	1,970	1,970	1,970	1,698	36	9	4
Elkhart	96	1,564±8,041	59,210	3,048	1,163	366	37	8	0
Rockford	175	112±127	768	352	263	200	53	13	0
Roscoe	22	2,126±2,697	7,816	7,639	6,446	3,732	702	176	1
Verona	10	51±71	195	195	162	124	16	0	0

[a] SD=standard deviation

Factors Affecting Exposure

The variation in TCE exposure might be caused by a variety of factors, including household geographic location, hydrologic features, source of contamination, and years exposed. Each site had a different source of contamination and most sites were contaminated by multiple sources. It was not feasible to characterize the sources of contamination based on the available information. Likewise, hydrologic features and the depth of wells were not included for similar reason. The numbers of men and women at these sites were nearly equal (157/161). Men and women living in the same households received the same exposure for any specific time. Thus, it was reasonable to disregard the effect of gender on TCE exposure. The question that remains is how a subject's age and site affect cumulative exposure.

Subjects were classified into five levels according to their age in 1990: 0 to 2, 3 to 4, 5 to 6, 7 to 8, and 9 to 10. The GLM procedure detected that the full model was statistically significant (p=.0001). The main effects of age and city (i.e., the effects of those two independent variables in the model) were statistically significant (p=.0001). The interaction between age and city (the crossed effect), however, was not significant (p=.2717). It was concluded that levels of cumulative TCE exposure were statistically different among age groups and sites. A contrast test (i.e., a test to determine how the level of one variable influences the affect of another) of the age groups indicated that TCE exposure of the 0-to-2 age group was significantly lower than that of any other age groups. The differences between the 3-to-4, 5-to-6, and 7-to-8 age groups were not statistically significant. The exposures of both the 3-to-4 and 5-to-6 age groups were significantly lower than that of the 9-to-10 age group. No significant difference between the 7-to-8 age group and the 9-to-10 age group was detected.

Analyses of Potentially Confounding Exposure

More than 20 chemicals were found in the groundwater of the contaminated sites, including DCE, DCA, TCA, PCE, benzene, carbon tetrachloride, freon, toluene, xylene, and heavy metals. Other reported secondary contaminants were chemicals and

biological degradation products of TCE and the other chlorinated hydrocarbons (7). The available data, however, indicated that most of the chemicals were not detected often enough to make it possible to characterize exposure to them at any sites. Only the four most frequently detected chemicals, DCE, DCA, TCA, and PCE (whose concentrations varied from 0.7 to 400 ppb), were included in the analyses of potentially confounding exposures.

A correlation analysis was conducted to detect the association between TCE and the four chemicals. TCE concentrations were significantly correlated with the levels of the four chemicals at the Rockford site, but not at the Elkhart site (Table 7). Measurements of the four chemicals above their quantitation limits were too sparse at other sites to allow effective analysis, so only the Rockford and Elkhart sites were considered in the confounding analysis. Confounding contaminant plumes were detected at the Rockford site. These plumes showed geographic distribution patterns similar to that of the TCE plume presented earlier. This provides visual evidence of the correlation (Figure 3).

Table 7 Correlation Analysis between Trichloroethylene and Other Contaminants (Pearson Correlation Coefficients) for the Rockford, Illinois, and Elkhart, Indiana, Sites

Chemical	Number of Subjects	Rockford	Elkhart
TCA	175	.7532 (.0001[a])	.1217 (.5219)
DCE	174	.5785 (.0001[a])	.1877 (.1873)
DCA	175	.7753 (.0001[a])	Sample size small
PCE	172	.3429 (.0197[a])	–.0286 (.8453)

[a] Statistically significant at α=.05.

The cumulative exposure of the four potential confounders was computed for the subjects. The subjects were then stratified into subgroups by level of potential confounding exposure. The three subgroups formed were 0 to 10 ppb per year, 10 to 100 ppb per year, and more than 100 ppb per year for TCA, DCE, and DCA exposure and 0 to 1 ppb per year, 1 to 10 ppb per year, and more than 10 ppb per year for PCE exposure. The subjects within each subgroup were further categorized into sub-subgroups according to their TCE exposure levels.

Discussion

In epidemiological studies and risk assessments, exposure assessment is often the most difficult task because it depends on factors that are hard to estimate and for which there are little data. As was discussed earlier, the environmental data in this study had limitations that increased the complexity of the retrospective exposure assessment. The samples were not taken to quantify exposure over time, but to verify contamination (7). Most of the environmental data consisted of single samples of well water from each household, reflecting exposure during small parts of the subjects' lives. Furthermore, TCE was not distributed uniformly within the study areas; it varied as a function of source characteristics and groundwater flow patterns. It was thus inappropriate to assume that the exposure of Subregistry children was constant over time and use the concentration of TCE in a single sample as the exposure level over the entire exposure

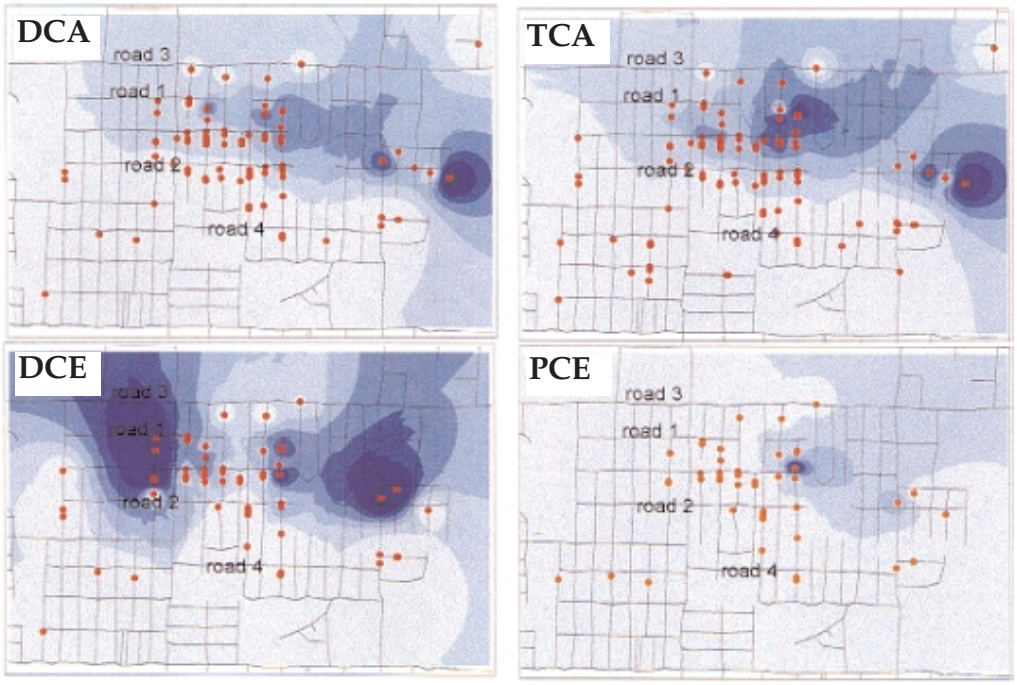


Figure 3 Confounding contaminant plumes in well water at the Rockford, Illinois, site.

period. In addition, we assumed that each subject at each site had been born there and had lived there continuously until the baseline survey was conducted. When each subject's residential history becomes available, the results of this assessment will be adjusted to yield the final estimate of exposure profiles. Groundwater contaminant concentration and other geohydrologic features tend to be spatially correlated; that is, the agreement of contaminant concentration and geohydrologic features at points within an area increases as the distance between the points decreases. This made it possible to estimate the subjects' exposures by using their neighbors' exposures for the time periods when no samples were collected at the subjects' residences.

GIS provided the platform in which spatial and temporal distribution of contaminants could be described, interpreted, and integrated. The strong data manipulation and visualization functions of GIS made exposure assessment rather straightforward and demonstrated the distinct advantage of GIS over routine numerical methods. The spatial interpolation performed by IDW and kriging methods identified a contaminated groundwater plume in Rockford extending in an east-west direction (Figure 1), which was consistent with ATSDR's finding (7). By using the GIS map query tool, a subgroup of 64 subjects living at the heavily contaminated area identified by GIS spatial analysis was created. It was found that their cumulative TCE exposures were significantly higher than those of subjects at the less-contaminated areas, provided that there were no statistical differences in age between the two groups.

GIS was also used to determine the exposure duration for the 103 subjects who lived in four different areas of Elkhart, each of which had a different time period of ex-

posure. The available information only provided a brief description of the exposure time for the areas, and the available geographic demarcation of the areas was not sufficient to permit assignment of a household to one of the four areas based on its address. The geocoding tool of GIS mapped each household to the corresponding location on a digital map. GIS thus made it possible to determine, and show, in which area each household was located.

Statistical and mathematical models, instead of groundwater models, were applied to explore spatial and temporal variation of TCE contamination. Groundwater models are often used to estimate the fate of contaminants that enter groundwater, but these models are usually complex, due to the many physical and chemical processes that can affect transport and transformation of contaminants in groundwater. Using groundwater models requires extensive knowledge of a site's geohydrology and geochemistry. In this retrospective study, many of the elements needed to establish groundwater models, such as sources of contamination and depth of wells, were unavailable. The statistical and mathematical approach presented here was not constrained by these limits. It fully utilized the available measured concentration data and allowed geostatistical inference. The kriging interpolation procedure can provide not only concentration estimates, but also the variance of these estimates. This can be used to establish confidence limits on estimates if needed.

Potential confounding exposure was a critical concern in the exposure assessment. Most of the subjects were exposed to multiple contaminants; more than 20 chemicals were detected in their well water. The typical contaminant sources for the sites were industrial processes that used a variety of common solvents in addition to TCE (7). At the Rockford site, exposure to TCE was significantly correlated with exposure to four other chemicals, TCA, DCA, DCE, and PCE. Spatial analyses found that the area with the highest concentrations of TCA, DCA, and DCE was also located between Road 1 and Road 2, similar to the pattern of TCE distribution. Because of this, caution must be exercised in interpreting results of speech and hearing tests in Rockford. It may be possible to distinguish between the effects of confounders by comparing Rockford results with those from Elkhart, where TCE exposure was not confounded.

Another issue should be kept in mind when interpreting this type of environmental data. Most of the samples were taken one time from households that were exposed over extended periods ranging from 4 to 10 years. These one-time samples might not be representative of the exposure for all of those periods. For instance, weather conditions might affect sample composition. Heavy rain might result in sediment loading to water bodies, which could increase contamination or affect the concentration of other contaminants through adsorption and settling in the water column. Ideally, monitoring should provide a full annual sampling cycle or at least encompass seasonal extremes such as conditions of high water/low water, high recharge/low recharge, and high suspended-solids/clear water (25).

Conclusion

All of the six sites studied were contaminated with TCE. The cumulative exposures were significantly associated with geographic location. No major changes in the geographic distribution pattern of TCE were observed at the Rockford site, except that the TCE levels generally increased over a six-year period. The subjects at this site could be

categorized into two subgroups: within the contaminant plume and outside the plume. It was found that the cumulative exposure ranged from 0 to 59,210 ppb per year, with a mean of 701 ppb per year. The mean exposure length was 5.5±2.9 years. Of the 198 households in the study, 118 had water with measured TCE concentrations exceeding 5 ppb, the EPA maximum allowable level for TCE in drinking water. The TCE exposure was significantly correlated with the potential confounding exposures at the Rockford site. Correlation was not significant at the Elkhart site. The subgroups created by quantitative exposure assessment could be used for the epidemiological investigations of cause-effect and dose-response relationships between TCE exposure and adverse health outcomes. GIS was shown to be a powerful tool for environmental studies involving spatial and temporal data.

References

1. Feldman RG. 1979. Trichloroethylene. In: Intoxications of the Nervous System, Part 1. Ed. PJ Vinken, GW Bruyn. *Handbook of Clinical Neurology* 36:457–64. Amsterdam: Elsevier.
2. US Environmental Protection Agency (EPA). 1985. *Health assessment document for trichloroethylene.* EPA-600/8-82-006F. Research Triangle Park, NC: EPA.
3. Sullivan JB, Krieger GR. 1992. *Hazardous materials toxicology.* Baltimore: Williams & Wilkins.
4. Mitchell ABS, Parsons-Smith BG. 1969. Trichloroethylene neuropathy. *British Medical Journal* 1:422–3.
5. Feldman RG. 1970. Facial nerve latency studies in man: Facts of trichloroethylene exposure in man. *Electromyography* 10:93–100.
6. Barret L, Garrel S, Danel V, Dbru JL. 1987. Chronic trichloroethylene intoxication: A new approach by trigeminal-evoked potentials. *Archives of Environmental Health* 42:297–302.
7. US Department of Health and Human Services (DHHS). 1994. *National Exposure Registry, Trichloroethylene (TCE) Subregistry, baseline technical report* (revised). Atlanta, GA: Agency for Toxic Substances and Disease Registry.
8. Crofton KM, Zhao X. 1993. Mid-frequency hearing loss in rats following inhalation exposure to trichloroethylene: Evidence from reflex modification audiometry. *Neurotoxicology and Teratology* 15:413–23.
9. Rebert CS, Day VL, Matteucci MJ, Pryor GT. 1989. Sensory-evoked potentials in rats chronically exposed to trichloroethylene: Predominant auditory dysfunction. *Neurotoxicology and Teratology* 13:83–90.
10. Aral MM, Maslia ML, Williams RC, Susten A, Heitgerd JL. 1994. Exposure assessment of populations using environmental modeling, demographic analysis, and GIS. *Water Resources Bulletin* 30(6):1025–41.
11. McKone TE, Knezovich JP. 1991. The transfer of trichloroethylene (TCE) from a shower to indoor air: Experimental measurements and their implications. *Journal of the Air and Waste Management Association* 41:832–7.
12. Aral MM, Maslia ML, Radtke TM. 1994. Conducting exposure assessment of populations by integrating environmental transport models, demographic analysis, and geographic information systems. *Proceedings of the International Symposium on Assessing and Managing Health Risks from Drinking Water Contamination: Approaches and Applications.* Rome, Italy: International Association of Hydrological Sciences. September 1994. 221–33.

13. US Census Bureau. 1991. TIGER/Line Census file, 1991. Washington, DC: US Department of Commerce, Bureau of the Census.
14. Environmental Systems Research Institute, Inc. (ESRI). 1996. *ArcView 3.0 manual.* Redlands, CA: ESRI.
15. Isaaks EH, Srivastava RM. 1989. *An introduction to applied geostatistics.* New York: Oxford University Press.
16. Whittaker E, Robinson G. 1924. *The calculus of observations.* 4th Ed. London and Glasgow: Blackie & Son Limited.
17. Bove FJ, Fulcomer MC, Klotz JB, Esmart J, Dufficey EM, Savrin JE. 1995. Public drinking water contamination and birth outcomes. *American Journal of Epidemiology* 141(9):850–62.
18. Cavalleri A, Gobba F, Paltrinieri M, Fantuzzi G, Righi E, Aggazzotti G. 1994. Perchloroethylene exposure can induce color vision loss. *Neuroscience Letters* 179:162–6.
19. Evans EB, Blaster RL. 1993. Inhaled 1,1,1-trichloroethane-produced physical dependence in mice: Effects of drugs and vapors on withdrawal. *Journal of Pharmacology and Experimental Therapeutics* 264(2):726–33.
20. Mochida K, Gomyoda M, Fujita T. 1995. Toxicity of 1,1-dichloroethane and 1,2-dichloroethylene determined using cultured human KB cells. *Bulletin of Environmental Contamination and Toxicology* 55(2):316–9.
21. Rosengren LE, Aurell A, Kjellstrand P, Haglid KG. 1985. Astrogliosis in the cerebral cortex of gerbils after long-term exposure to 1,1,1-trichloroethane. *Scandinavian Journal of Work, Environment, and Health* 11(6):447–55.
22. US Environmental Protection Agency (EPA). 1984. *Health assessment document for 1,1,1-trichloroethane (methyl chloroform).* EPA-600/8-82-003F. Research Triangle Park, NC: EPA. 5.1–5.25.
23. US Environmental Protection Agency (EPA). 1985. Health assessment document for tetrachloroethylene (perchloroethylene). EPA-600/8-82-005F. Research Triangle Park, NC: EPA.
24. White RF, Feldman RG, Travers PH. 1990. Neurobehavioral effects of toxicity due to metals, solvents, and insecticides. *Clinical Neuropharmacology* 13:392–412.
25. Covello VT, Merkhofer MW. 1993. *Risk assessment methods.* New York: Plenum Press.

Environmental Exposure and the Reproductive Health of Hispanic Women in Miami-Dade County, Florida

Alice Clarke (1),* Seemanthini Hariharan (2), Jennifer Fu (3)
(1) Florida International University, Dept. of Environmental Studies, Miami, FL; (2) University of Miami, School of Medicine, Dept. of Obstetrics & Gynecology, Miami, FL; (3) Florida International University, Green Library GIS Laboratory, Miami, FL

Abstract

This project examines potential links between environmental hazards and women's reproductive health, with particular emphasis on fetal health and pregnancy outcomes in Miami-Dade County, Florida. Health data are derived from records of women seen at Jackson Memorial Hospital, University of Miami. The patient pool is predominantly urban and Hispanic and allows us to investigate environmental health issues important to this understudied minority population. This project will eventually examine in detail, through retrospective and prospective studies, the relationship of maternal, fetal, and neonatal (e.g., gestational age, birth weight) outcomes to a variety of point and non-point source environmental exposures. We will also consider confounding socioeconomic variables (e.g., income, health care delivery system used, and cost of medical care). At this initial stage, we are using a geographic information system (GIS) framework and data from EPA's Toxic Release Inventory and the Metro-Dade County Environmental Resources Management Agency to analyze patient exposure risks to point source environmental hazards (e.g., industrial facilities, private wells, petroleum storage sites). Environmental justice issues related to environmental exposure risks are of particular concern in Miami-Dade County at this time. An urban core redevelopment project, "Eastward Ho!," seeks to revitalize and improve quality of life in Southeast Florida's historic, urban areas and simultaneously lessen development pressures and urban sprawl in sensitive environmental lands to the west, including the Everglades ecosystem. This program, however, has also generated concerns about public health effects of utilizing brownfield (contaminated/remediated) sites for urban in-fill. Our project will contribute to this critical discussion by providing information on the relationship between reproductive health and environmental risks within the urban core.

Keywords: environmental justice, reproductive health, Hispanic

Introduction

We are interested in the potential health risks to pregnant women and their fetuses in the predominantly Hispanic, urban population of Miami-Dade County due to exposure to environmental hazards. Patient data will be derived from the patient pool at Jackson Memorial Hospital, a large, metropolitan teaching hospital affiliated with the University of Miami. While obstetrics patients are routinely screened for a variety of common medical/obstetrical and psycho/social risk factors, they undergo no screening for any environmental risk factors. Here we describe the preliminary

* Alice L Clarke, Florida International University, Dept. of Environmental Studies, Miami, FL 33199 USA; (p) 305-348-1693; (f) 305-348-6137; E-mail: clarkea@fiu.edu

planning phase of our study, which will incorporate patient information and environmental data into a geographic information system (GIS) for analysis.

Population at Risk

Miami-Dade County Population

The population of Miami-Dade County is predominantly urban and unique in its high proportion of Hispanic individuals and large number of recent immigrants, originating primarily from the Caribbean but also from a variety of Latin American countries. While the population of the county shares some of the general characteristics of immigrant and Hispanic populations in the United States, demographic and socioeconomic measures vary significantly across Miami-Dade's ethnic groups and within ethnic groups depending in part on the time since arrival in the United States.

Hispanics make up 52% of the county's population, with other minorities (African Americans, Haitians, Native Americans, and Asian Americans) comprising another 22%. Foreign-born and native-born Cuban Americans make up the largest proportion (approximately 50%) of Hispanics in the county. There are also large populations of Puerto Ricans, Colombians, and Nicaraguans. Approximately 60% of the county's population is foreign-born (1).

Although Miami-Dade County is predominately Hispanic, on average, it reflects closely the age and fertility patterns of the US in general. However, average demographic figures may mask significant within-group variation in the county. For example, foreign-born Hispanic women tend to have higher fertility rates and lower educational levels (2). In addition, the relatively large illegal immigrant population in the county may differ in fertility and age structure from the legal population upon which reported demographic measures are based. Average educational achievement level is one variable for which Miami-Dade County falls below the national average (65% and 82% high school graduate or higher, respectively) (1).

Jackson Memorial Hospital Obstetric Pool

As a metropolitan public hospital, Jackson Memorial serves a high-risk population of obstetrics patients. They represent the working poor, and the majority of patients depend on some form of public assistance. Patients are disproportionately ethnic minorities even relative to the overall county population. Patients who tend to be less educated and from lower income groups are less likely to obtain appropriate care prior to the delivery of their infants (3). Poor, young minorities are disproportionately uninsured (3). We feel they are also less likely to be aware of potential environmental health hazards.

Current data (January to June 1998) indicate that, while most mothers delivering at Jackson are between the ages of 20 and 39, 16% are age 13 to 19. Seventy-two percent of the women report themselves to be single parents. The majority of mothers begin prenatal care in the first trimester of pregnancy; however, 5% do not receive care until the third trimester and 10% report no prenatal care before delivery.

Residence within the urban core may lead to higher levels of exposure to environmental hazards and to exposure to different environmental hazards than the population at large. It is now well established that the potential impact of environmental hazards

is not uniform (4). Socioeconomic status and ethnicity are among the factors that make some groups more vulnerable to the adverse health effects of environmental pollutants (5). This is the primary concern of the field of environmental justice, which has found that hazardous waste sites are disproportionately located in minority communities (6), that air pollutants are disproportionately released in minority, especially Hispanic, communities (7), and that average penalties incurred by polluters are substantially lower in minority communities (8). Despite this potential for greater exposure among Miami-Dade's minority, urban core population, identification of environmental risk factors and detection of mother/fetal exposure and consequence are currently unmonitored within the Jackson Memorial obstetric patient pool.

Environmental Hazards to the Fetus

Unfortunately, though the perinatal period is recognized as a sensitive period of life, it is also understudied and there are few data available on the adverse reproductive and developmental effects of most environmental agents. Government policy does not specifically identify the fetus as a potentially vulnerable individual despite the fact that the fetus and the newborn differ biologically from adults.

Although there are many similarities between intra- and extrauterine exposures, there are notable differences as well. The fetus is at risk, first, through the increased basal metabolic rate of pregnant women who have increased minute ventilation and oxygen consumption, which increases their risk of exposure to air pollutants (9). Because of increased basal metabolic rate and accretion of new tissue, pregnant women have an increased caloric requirement, which increases their risk of exposure to pollutants in food and water (9).

Second, the fetus differs from the adult in modes of potential exposure. For instance, the skin of the fetus is underkeratinized, reducing the barrier properties of this tissue. The ability to metabolize various chemicals depends on developmental stage. Perhaps most importantly, fetal organs are in the process of growth and differentiation, which increases their vulnerability to harmful agents (9).

A pregnant mother's environment and her health behaviors are important determinants of fetal exposure. Fetal exposure to lead is dependent on both current and past maternal exposures to the element. Lead accumulates in the bones over time; this accumulate is mobilized from maternal bones during pregnancy and can result in elevated fetal lead levels (10,11). For this reason, lead exposure should be minimized and monitored in young women in order to avoid future fetal exposures.

The primary route of fetal exposure to methyl mercury is parental consumption of contaminated fish. Subsistence fishing with its risk of exposure to methyl mercury, PCBs, and other chemicals is a major environmental justice concern in the state of Florida (12,13). As reported by the Florida Public Interest Research Group in 1998, Florida ranks 12th in the country in mercury emissions from coal- and oil-burning power plants. This mercury then contaminates both inland and coastal bodies of water. Because of bioaccumulation, locally caught fish can contain as much as 100,000 times the concentration of mercury in the surrounding water.

The health consequences of perinatal exposure to dioxin and related compounds have prompted a World Health Organization risk assessment initiative (14). Dioxin-like compounds are known to transfer, although incompletely, across the placental barrier

to the fetus. Developmental abnormalities and neuro-behavioral deficits have been identified in children whose mothers consumed PCB-contaminated cooking oil and organochlorine-contaminated fish during pregnancy (15,16,17,18).

The hormonal effects of pesticides have obvious reproductive and developmental consequences. The in-utero endocrine effects of vinclozolin have been well documented. It acts as an anti-androgen and, as a pesticide residue in food, it may have developmental consequences in fetuses and children who are potentially sensitive to imbalances in hormone levels (19,20). Organochlorines, a group of widely used chemicals, have been implicated in endocrine related events in alligators living in Lake Apopka in Central Florida (21). DDT has been spilled in these areas and the effects are considered to be due to DDE, a potent metabolite of DDT, which leads to an imbalance between androgens and estrogens and causes abnormal sexual development. DDT and DDE, which have a well-documented lactation suppression effect, and vinclozolin are classic examples of endocrine disrupters that can cause abnormal pregnancies, endometriosis, and increased risk of breast and prostate cancer (22).

Environmental Hazards in Miami-Dade County

We have begun to incorporate information and data pertaining to both point source and non-point source reproductive and developmental hazards in Miami-Dade County into a GIS database (zip code and street address linked). Our primary point source database is the US Environmental Protection Agency's annual Toxic Release Inventory (TRI) monitoring database. These data provide information on air and water releases from monitored facilities occurring in Miami-Dade County. Though the TRI data do have recognized limitations, we have used the Environmental Defense Fund's (23) lists of recognized and suspected reproductive and developmental toxins together with existing TRI data to identify facilities within Miami-Dade County that release these substances (Tables 1–3). We can see from preliminary plotting of these TRI sites and 1997 Jackson Memorial obstetrics patient residence by zip code that our patient pool lives in relatively close proximity to these sites (Figure 1). We also have access through the Dade County Environmental Resource Management Agency to data on additional local facility releases (e.g., underground tanks, dry cleaners) that are not included under the TRI system.

Non-point source exposures may be of equal or greater health importance than point source releases. In addition to specific information acquired through our patient survey, we will incorporate into our GIS database information on drinking water sources (public, private commercial, and individual well systems) and age of housing, which can influence the likelihood of exposure to lead through plumbing and paint.

Proposed Study

Our goal is to collect environmental risk data on individual mothers through questionnaires, then link these data to pregnancy outcomes and an environmental hazards database. All pregnant women presenting for prenatal care at the obstetrical service of Jackson Memorial Hospital will be offered a questionnaire that screens for environmental health issues. All interviews will be conducted one-on-one with the patient in a

Table 1 USEPA Toxic Release Inventory Data (1995): Recognized Developmental Toxicants Released in Miami-Dade County, FL (23)

Chemical	Releases to Air (lb)
Toluene	440,919
Arsenic	10

Table 2 USEPA Toxic Release Inventory Data (1995): Suspected Developmental Toxicants Released in Miami-Dade County, FL (23)

Chemical	Releases to Air (lb)
Trichloroethylene	141,495
Styrene	136,295
Tetrachloroethylene	82,000
Phenol	67,004
Xylene (mixed isomers)	8,550
Methyl ethyl ketone	1,827
Methyl methacrylate	985
Glycol ethers	510
Methyl isobutyl ketone	250
Copper	22

Table 3 USEPA Toxic Release Inventory Data (1995): Suspected Reproductive Toxicants Released in Miami-Dade County, FL (23)

Chemical	Releases to Air (lb)
Toluene	440,919
Dichloromethane	371,714
Trichloroethylene	141,495
Tetrachloroethylene	82,000
Xylene (mixed isomers)	8,550
Methyl ethyl ketone	1,827
Methyl methacrylate	985
Glycol ethers	510
Copper	22
Arsenic	10

private setting. The questionnaire will focus on the following aspects of environmental exposure hazards:

- **Water:** Source and consumption patterns.
- **Lead exposure risks:** Age of housing/plumbing.
- **Mercury exposure risks:** Fish consumption patterns, source of fish consumed.
- **Occupational exposure risk:** Occupation information on patient and primary household members.

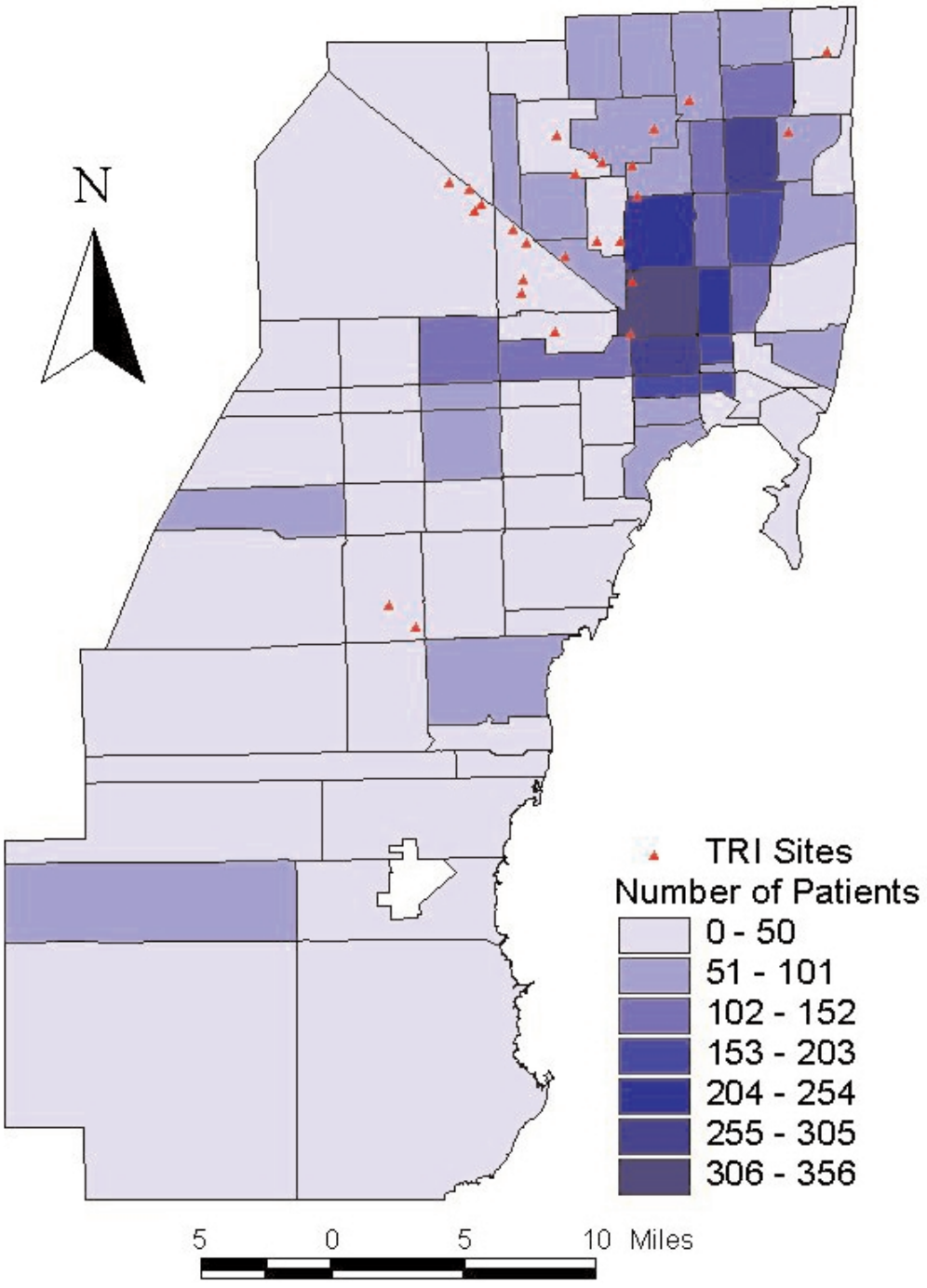


Figure 1 Location of TRI sites releasing known and suspected reproductive and developmental toxins, and 1997 Jackson Memorial Hospital obstetrics patient residence by zip code in Miami-Dade County, FL.

- **Pesticide exposure risk:** Known use of pesticides/herbicides in and around the home.
- **Personal lifestyle risk factors:** e.g., smoking habits of patient and primary household members.

Our assessment of pregnancy outcomes will include a number of variables not readily available in the literature. For each mother, we will record antenatal course, medical risk factors, abnormalities in fetal growth and tests of fetal well-being, gestational age at labor, duration and progress of labor, mode of delivery, indications for operative delivery, hospital course, number of days in hospital, and discharge diagnosis. For each birth, we will also record gestational age at birth, birth weight, number of days in the hospital, stay in normal newborn nursery versus neonatal intensive care unit (days on respirator, treatment modalities), age at discharge from hospital, and discharge diagnosis when other than a normal newborn. In addition, we hope to collect and analyze matched mother/fetal blood samples for assessment of critical risk factors such as lead, mercury, various organochlorines, and indicators of endocrine disruption.

We plan to explore further environmental risk and pregnancy outcome patterns through the development and use of a countywide GIS database. We plan to include the following data:

- EPA TRI annual facility release data
- Superfund site locations
- Other county data on non-TRI release facilities
- Public and private drinking water sources
- Jackson Memorial Hospital obstetrics patient outcome data

This database and the GIS format will allow us to plot potential environmental hazard sources as well as known pregnancy outcomes for surveillance purposes. This approach is critical to beginning to assess spatial distribution of negative pregnancy outcome patterns and their potential relationship with known and suspected reproductive and developmental hazards. This analysis is of particular interest from the perspective of environmental justice in the urban, minority population of Miami-Dade County.

References

1. US Census Bureau. 1994. *US census 1994 estimates*. US Census Bureau Web site. http://www.census.gov.
2. del Pinal J, Singer A. 1997. Generations of diversity: Latinos in the United States. Population Reference Bureau. *Population Bulletin* 52(3):12.
3. Commonwealth Fund, The. 1993. *Survey of women's health.* New York: The Commonwealth Fund.
4. Mohai P, Bryant B. 1992. Environmental racism: Reviewing the evidence. In: *Race and the incidence of environmental hazards*. Boulder, CO: Westview. 163–76.
5. Brown P. 1995. Race, class, and environmental health: A review and systematization of the literature. *Environmental Research* 69(1):15–30.
6. United Church of Christ, Commission for Racial Justice. 1987. *Toxic wastes and race in the United States: A national report on the racial and socio-economic characteristics of communities with hazardous waste sites*. New York: United Church of Christ, Commission for Racial Justice.

7. Wernette DR, Nieves LA. 1992. Breathing polluted air: Minorities are disproportionately exposed. *EPA Journal* 18(1):16–17.
8. Lavelle M, Coyle M. 1992. Unequal protection: The racial divide in environmental law. *National Law Journal* 15 September.
9. Bearer CF. 1995. How are children different from adults? *Environmental Health Perspectives* 103(Supplement 6):7–12.
10. Gulson BL, Jameson CW, Mahaffey KR, Mizon KJ, Korsch MJ, Vimpani G. 1997. Pregnancy increases mobilization of lead from maternal skeleton. *Journal of Laboratory and Clinical Medicine* 130(1):51–62.
11. Gulson BL, Mahaffey KR, Jameson CW, Mizon KJ, Korsch MJ, Cameron MA, Eisman JA. 1998. Mobilization of lead from the skeleton during the post-natal period is larger than during pregnancy. *Journal of Laboratory and Clinical Medicine* 131(4):324–29.
12. May H, Burger J. 1996. Fishing in a polluted estuary: Fishing behavior, fish consumption, and potential risk. *Risk Analysis* 16(4):459–71.
13. Fleming L, Watkins S, Kaderman R, Levin B, Ayyar D, Bizzio M, Stephens D, Bean J. 1995. *Mercury exposure in humans through food consumption from the Everglades of Florida.* Third International Conference on Mercury as a Global Pollutant. 41–48.
14. Lindstrom G, Hopper K, Petreas M, Stephens R, Gilman A. 1995. Workshop on perinatal exposure to dioxin-like compounds: Summary. *Environmental Health Perspectives* 103(Supplement 2):135–42.
15. Rogan WJ, Gladen BC, McKinney JD, Carreras N, Hardy P, Tiglestad J, Tully M. 1986. Neonatal effects of transplacental exposure to PCBs and DDE. *Journal of Pediatrics* 109:335–41.
16. Rogan WJ, Gladen BC, Hung K, Koong SL, Shih LY, Taylor JS, Wu YC, Yang D, Ragan NB, Hsu CC. 1988. Congenital poisoning by poly-chlorinated biphenyls and their contaminates in Taiwan. *Science* 241:334–36.
17. Jacobson JL, Jacobson SW, Humphry HEB. 1990a. Effects of in utero exposure to polychlorinated biphenyls (PCBs) and related contaminants on cognitive functioning in young children. *Journal of Pediatrics* 116:38–45.
18. Jacobson JL, Jacobson SW, Humphry HEB. 1990b. Effects of exposure to PCBs and related compounds on growth and activity in children. *Neurotoxicology Teratology* 12:319–26.
19. Kelce WR, Mononson E, Gamcsik MP, Laws SC, Gray Jr LE. 1994. Environmental hormone disrupters: Evidence that vinclozolin developmental toxicity is mediated by anti-androgenic metabolites. *Toxicology Applied Pharmacology* 126:276–85.
20. Gray Jr LE, Ostby JS, Kelce WR. 1994. Developmental effects of an environmental antiandrogen: The fungicide vinclozolin alters sex-differentiation of the male rat. *Toxicology Applied Pharmacology* 129:46–52.
21. Semenza J. 1997. Reproductive toxins and alligator abnormalities at Lake Apopka, Florida. *Environmental Health Perspectives* 105(10):1030–32.
22. Goldman LR. 1995. Children—unique and vulnerable: Environmental risks facing children and recommendations for response. *Environmental Health Perspectives* 103(Supplement 6):13–18.
23. Environmental Defense Fund. 1998. Scorecard Web site. http://www.scorecard.org.

Using GIS to Create Childhood Lead Poisoning Guidelines in Florida

Christopher M Duclos,* Tammie M Johnson, Trina Thompson
Bureau of Environmental Epidemiology, Florida Department of Health, Tallahassee, FL

Abstract

Over 900,000 children in the United States have blood lead levels high enough to cause health problems that range from learning disabilities to permanent neurological damage. The Florida Department of Health, Bureau of Environmental Epidemiology, was awarded a grant from the US Centers for Disease Control and Prevention (CDC) to conduct childhood lead poisoning surveillance in Florida. The CDC identified older housing stock as the most significant avenue for lead exposure in young children. In 1997, the CDC published screening guidelines that suggested universal screening for all children living in census block groups where 27% or more of the housing was built before 1950. However, dangerous amounts of lead were present in paint until the mid-1970s. Using a geographic information system (GIS), the Bureau of Environmental Epidemiology is developing statewide screening guidelines that are more appropriate to the unique demographics of Florida than the CDC guidelines. With ArcView software, many different variations of the CDC guidelines were examined quickly and easily. The Bureau assigned latitude and longitude coordinates to a table of 1993–1997 lead poisoning cases, then performed a tabular join to link the census housing data to the geographic block group data. ArcView was then used to isolate the housing age by block group in multiple combinations until the best fit with the case addresses was determined. This procedure enables county health departments to use targeted blood lead screening, thus maximizing the number of at-risk children being tested while consuming fewer resources than they would using universal screening.

Keywords: lead, poisoning, housing, block group, screening

Introduction

The United States Centers for Disease Control and Prevention (CDC) has estimated that 900,000 children less than six years of age have blood lead levels higher than 10 micrograms per deciliter. This seemingly low level of lead exposure has been scientifically documented to cause developmental abnormalities such as lower intelligence and reduced stature (1,2). Higher levels of blood lead can cause nervous system dysfunction, reduced blood oxygen capacity, kidney failure, and death.

Contrary to popular belief, lead poisoning is not limited to children of the poor or of minority members. Lead can afflict children regardless of their socioeconomic status and has been used so extensively by industrial society that it is virtually impossible not to consume it. This statement has been proven by the comparative measurement of

* Christopher M Duclos, Bureau of Environmental Epidemiology, Florida Department of Health, 1317 Winewood Blvd., Tallahassee, FL 32399 USA; (p) 850-488-4821; (f) 850-922-8473; E-Mail: Chris_Duclos@doh.state.fl.us

lead in the bones of pre-Columbian New World dwellers to the bones of modern people. The bone lead levels of modern humans are on average 100 to 1,000 times higher than those of pre-Columbian humans (3). While these elevated bone lead levels are generally not high enough to be termed "lead poisoning" under current federal guidelines, the mere presence of lead at comparatively high concentrations in modern humans suggests the inevitability of lead ingestion and hints at the universal problem of lead poisoning.

For several reasons, children are more easily lead-poisoned than adults. First of all, "if the same concentration of lead is present in substances consumed, such as air, food, or water, children ingest or inhale a greater quantity relative to body weight than do adults" (4). This is because children have higher rates of respiration and metabolism than adults. Secondly, when children ingest lead, a greater quantity of that lead is retained in their bodies than in adults' bodies. For example, Ziegler et al. (5) discovered that infants between 14 days and two years old absorbed 42% of ingested lead and retained 32% of ingested lead. In contrast, adults are generally considered to absorb 5 to 10% of ingested lead (although nutritional factors play a significant role in absorption/retention) (4). The third reason for which children are more likely to be lead-poisoned in today's environment is that children, especially those between 0 and 72 months old, are more likely to engage in extensive hand-to-mouth activity. This means they are more likely to ingest lead-contaminated matter such as dust, soil, paint chips, or pottery. For children, the most significant of these sources is leaded paint in older housing. Lead was used extensively in residential paint until the federal government banned its use as an additive in 1978. However, thousands of children continue to be exposed to this deteriorating paint. The most effective methods of preventing childhood lead poisoning are to remove the lead paint from the child's environment or to protect the child from lead exposure. In practice, the elimination of childhood lead poisoning has been a painfully slow process because of a lack of public awareness and because of the sheer enormity of the problem.

The CDC has taken the lead in establishing policy directives to reduce the prevalence of childhood lead poisoning. Recognizing that state and local health agencies are better equipped to deal with the specific lead poisoning issues in their jurisdictions, the CDC has made extensive surveillance and prevention grant monies available to these government agencies. In Florida, the Department of Health, Bureau of Environmental Epidemiology, was awarded a statewide childhood lead poisoning surveillance grant in 1992. In that same year, childhood lead poisoning became a reportable disease in Florida, which means that state and private laboratories were required to report the results of all blood lead tests performed. The cases of childhood lead poisoning used in this study were drawn from the centralized database of the Florida Childhood Lead Poisoning Surveillance Program (CLPSP) for the years 1993 through 1997.

Most of the cases in the CLPSP database were tested by one of the 67 county health departments (CHDs) in Florida. There is a great deal of variability between the level of service provided by the different CHDs. Some counties (e.g., Pinellas, Duval) offer full-service health facilities, while others (e.g., Dade, Broward) have farmed their responsibilities out to private health agencies because of a lack of adequate budget to maintain proper health care services of their own. In the counties that do offer health care programs, most of the children enrolled are Medicaid recipients. Thus, most of the children tested for lead poisoning by the CHDs are Medicaid recipients. When these children are

tested, their blood is sent to the state laboratory in Jacksonville. The state laboratory is then required to send the results to the statewide surveillance database in Tallahassee.

In contrast, private physicians in the counties with little or no CHD services see a mixed bag of children on Medicaid and children with private insurance plans. Unfortunately, the majority of private physicians in Florida do not believe that childhood lead poisoning is a major health concern (6). Testing of blood lead levels by private physicians is sporadic at best, even for the children they see who are on Medicaid. This is significant because Medicaid requires and pays for childhood blood lead screening for all one- and two-year-olds. However, no government agency is enforcing the mandatory blood lead testing required for children on Medicaid.

For children in cost-conscious HMOs, the prospects of a blood test are even less promising. For example, one of the Florida Department of Health employees who administers the statewide lead poisoning database could not get her insurance to pay for a blood lead test for her children because her physician would not approve it. Like many other doctors, he did not believe the infants were in any danger, even though the employee lived in a house built before 1978.

In addition to sponsoring grants for childhood lead poisoning surveillance, the CDC has taken an active role in the delineation of lead poisoning hazards. In 1997, the CDC published a short document entitled *Screening Young Children for Lead Poisoning: Guidance for State and Local Public Health Officials* (7). The purpose of this document was to reiterate the CDC's commitment to the surveillance and prevention of childhood lead poisoning. It recommended a basic targeted screening plan as an interim measure while local data were being reviewed. In other words, local health departments should make a concerted effort to test all children living in areas with greater than or equal to 27% pre-1950 housing (the national percentage). The areas that exceed this national percentage of pre-1950 housing are more likely to contain lead-poisoned children. However, the CDC admits that this definition of what areas to target may not be adequate for all jurisdictions, since a substantial threat remains in housing built between 1950 and 1978. This is precisely the case for the state of Florida, where the building boom did not occur until after World War II.

In comparison to the national situation, Florida's housing does not appear to be as hazardous to young children. Only 7.7% of Florida housing was built before 1950, a percentage that is 47th out of 50 states. However, this percentage represents 472,481 homes, which places Florida 19th out of 50 states in sheer numbers of pre-1950 houses. Furthermore, because the phase-out of leaded paint for residential uses was not complete until 1978, homes built between 1950 and approximately 1970 still represent a significant hazard to children. In Florida, the number of homes built between 1950 and 1970 is 1,708,205, or approximately 3½ times more than in all previous years combined.

Clearly, the CDC recommendation to screen all children living in areas with at least 27% pre-1950 housing may not be ideally suited to Florida, considering the large number of homes built between 1950 and 1970. In other words, using pre-1950 housing alone would not capture enough of the lead poisoning risk. The Bureau of Environmental Epidemiology has taken the initiative in analyzing state childhood lead poisoning data in conjunction with 1990 census data. The ultimate objective of this analysis is to publish a modified screening recommendation to aid in focused screening efforts.

Methods

In order to analyze the geography of childhood lead poisoning in Florida, the statewide database of children with elevated lead levels first had to be geocoded. This process used the residential address from each record in the CLPSP database to assign a latitude and longitude based on where the address fell on the specific street segment. This was possible because nearly every road in Florida has been entered into a geographic information system (GIS) database. This GIS stores the latitude and longitude of every road segment, as well as the address ranges (house numbers) found along it. In this manner, the cases were added to the Florida Department of Health's GIS in order to analyze the case locations at various geographic levels.

The county level was chosen as the most appropriate unit for GIS analysis. The analysis is still ongoing, and only preliminary results are available at this time. For the presentation at the 1998 GIS in Public Health conference, Pinellas County, Florida, was chosen to demonstrate the methodology of analyzing the spatial arrangement of childhood lead poisoning cases in relation to housing data from the 1990 census. Pinellas County is a metropolitan county containing the cities of St. Petersburg and Clearwater. It was one of the earliest counties in Florida to undergo a population explosion (mostly as a result of in-migration) after World War II. As a result, Pinellas County contains a significant number of older homes with deteriorating lead-based paint.

Three years ago, the Pinellas County Health Department (PCHD) was awarded an individual surveillance grant from the CDC. With this federal money, the PCHD administered surveys to determine what areas of the county present the greatest lead poisoning risk to children. Using this knowledge, the PCHD has been able to find more lead-poisoned children than many other counties in Florida. Furthermore, the cases of lead poisoning found by the PCHD are more representative of the overall population of children than are the findings of many counties in Florida. For these reasons, Pinellas County was chosen for the initial GIS analysis of lead poisoning cases in relation to older housing.

Data on the childhood lead poisoning cases for Pinellas County were overlaid on 1990 census block groups. Using GIS, different combinations of older housing could be isolated and then analyzed with the cases. First of all, the CDC recommendation of universal screening in block groups made up of at least 27% pre-1950 housing was tested to see what percentage of cases fall in this defined area. The answer was 65%. A modified standard of at least 58% pre-1970 housing was then used; 84% of the cases fell within this area, for an overall improvement of 19%. The use of this modified recommendation would increase the number of children with elevated lead levels discovered through universal screening in these areas. In the future, more elaborate GIS analyses will be performed using other census variables, such as percent single mothers, percent below poverty line, percent black, etc. This could serve to focus screening efforts to a more sophisticated level than ever before through the use of GIS to analyze childhood lead poisoning cases in relation to demographic data.

Conclusion

In conclusion, childhood lead poisoning remains a problem in the United States despite persistent efforts to reduce its prevalence. Leaded paint in older housing is the single

greatest exposure route for childhood lead poisoning. The CDC has recommended universal screening of children in areas where at least 27% of the housing was built before 1950. However, the CDC is aware of the fact that this recommendation is not ideally suited to all jurisdictions. The Florida Department of Health is using GIS to examine the geography of childhood lead poisoning cases extracted from the statewide CLPSP database. Early indications from the analysis of Pinellas County reveal that a modified CDC recommendation may be better suited to focused screening efforts in the Sunshine State. GIS was central to this conclusion, and it is the hope of this researcher that GIS continues to play a significant role in the effort to eliminate childhood lead poisoning in Florida and around the world.

References

1. Needleman HL, Gatsonis CA. 1990. Low-level lead exposure and the IQ of children. *Journal of the American Medical Association* 263:673–8.
2. Schwartz J, Angle C, Pitcher H. 1986. Relationship between childhood blood lead levels and stature. *Pediatrics* 77:281–8.
3. Fowler BA, Bellinger DC, Bornschein RL, Chisholm JJ, Falk H, Flegal AR, Mahaffey KR, Mushak P, Rosen JF, Schwartz J, Skogerboe RK. 1993. *Measuring lead exposure in infants, children, and other sensitive populations.* Washington, DC: National Academy Press.
4. Mahaffey DR. 1981. Nutritional factors in lead poisoning. *Nutritional Review* 39:353–62.
5. Ziegler EE, Edwards BB, Jensen RL, Mahaffey KR, Fomon SJ. 1978. Absorption and retention of lead by infants. *Pediatric Research* 12:29–34.
6. Hopkins RS, Watkins SM, Quimbo RA. 1995. Elevated blood lead prevalence in Florida two-year-olds. *Journal of the Florida Medical Association* 3:193–7.
7. Centers for Disease Control and Prevention (CDC). 1997. *Screening young children for lead poisoning: Guidance for state and local public health officials.* Atlanta, GA: CDC.

Potential Risk Indexing System (P-RISK Model) Utilizing GIS to Rank Geographic Areas, Industrial Sectors, Facilities, and Other Areas of Concern

Debra L Forman (1), Amy Amina C Wilkins (2),* David West (3)
(1) Waste and Chemicals Management Division, US Environmental Protection Agency, Region III, Philadelphia, PA; (2) Office of Research and Development, National Center for Environmental Assessment, US Environmental Protection Agency, Washington, DC; (3) Office of Policy and Management, US Environmental Protection Agency, Region III, Philadelphia, PA

Abstract

The Potential Risk Indexing System (P-RISK) is a screening methodology and computer-based program that ranks areas of concern (i.e., facilities, industrial sectors, and geographic areas) according to multi-media chemical releases, chemical toxicities, and selected demographics of surrounding populations. The model uses geographic information system (GIS) technologies to display vast quantities of data, assisting users in cumulative risk analysis and other decision-making processes. P-RISK users include risk assessors and managers, US Environmental Protection Agency (EPA) program offices and Regions, state environmental departments, and other communities concerned with environmental targeting, inspection targeting, pollution prevention targeting, resource prioritization, environmental justice analysis, trend analysis, and comparative risk efforts. P-RISK operations currently include five steps. First, release data are retrieved from the EPA's Toxics Release Inventory System, the Aerometric Information Retrieval System, and the Permits Compliance System for reported chemical emissions to air, land, and water. Second, toxicity values are obtained from the Integrated Risk Information System and the Health Effects Assessment Summary for oral carcinogenic and non-carcinogenic effects. A dose is then calculated and weighted to create an index of relative toxicity scores. Third, a visual GIS data layer showing color-coded, indexed areas (based on an 8-mile by 8-mile grid system) ranging from high-release and high-toxicity combinations to low-release and low-toxicity combinations is created and made available for viewing and printing. Fourth, in keeping with EPA environmental justice and children's health protection guidelines, a second GIS color-coded, indexed data layer is generated, using US Census data, to ascertain income and minority status as well as other factors that may influence the relative vulnerability of subpopulations. This approach does not yet imply exposure, which must be assessed using intake parameters adjusted for specific subpopulations. Presently, the vulnerability index created in step 4 characterizes potentially exposed populations, highlighting those that may be more vulnerable. In this way, the screening can include population characteristics without using broad assumptions about exposure conditions. Finally, the chemical/facility index (created in step 3) and the vulnerability index can then be overlaid to match incidence of high-toxicity, large-release combinations with areas having relatively high percentages of vulnerable populations.

* Amy Amina Wilkins, US Environmental Protection Agency, Office of Research and Development, National Center for Environmental Assessment, 401 M St., SW, Mail Code 8623-D, Washington, DC 20460 USA; (p) 202-564-3256; (f) 202-565-0078; E-mail: Wilkins.Amina@epamail.epa.gov

Keywords: facility ranking or index, demographic populations, environmental justice

Introduction: The Potential Risk Indexing System

The Potential Risk Indexing System (P-RISK) is a computer-based screening model that ranks facilities, industrial sectors, or geographic areas according to data on multi-media chemical releases and chemical toxicities, population demographics, and other spatial features. The P-RISK is structured to enable the user to consolidate all available data, ranging from ambient air and water status to watershed health to potential human health risk, in one user-friendly product.

By using geographic information system (GIS) software to display vast quantities of data, the P-RISK assists users in cumulative risk analysis, broadening individual programmatic criteria for targeting enforcement actions in stressed areas, or identifying improved environmental protection in other areas. Expected users of the P-RISK include US Environmental Protection Agency (EPA) risk assessors, EPA risk managers, and state and local environmental regulatory agencies concerned with inspection targeting, pollution prevention targeting, resource prioritization, environmental justice analysis, trend analysis, and comparative risk efforts. Communities that use environmental information in their decision-making may also use the P-RISK.

The P-RISK has resulted from the consolidation of two independent development efforts, the Chemical Indexing System (CIS) (1,2), prepared by EPA Region III, and the Risk-Based Enforcement Strategy (RBES) (3,4), prepared by EPA's National Center for Environmental Assessment at the Office of Research and Development. As part of the consolidated development process, the P-RISK Workgroup, consisting of EPA staff from headquarters and the Regions, has been convened to provide critical input and help develop the P-RISK model as it currently exists. This peer-review package includes input provided by the workgroup during a six-month period from June to November 1998 (5).

Conceptual Model

The P-RISK is intended to address several complex questions pertaining to potential risk and exposure to toxics. It is designed to answer a range of questions from a variety of users using the familiar platform of a personal computer (PC). To do this, the P-RISK takes the following three-step approach:

1. Retrieval of data from EPA databases.
2. Manipulation of these data to calculate a set of indices executed within a series of independent modules.
3. Use of a GIS interface to integrate the data and test scenarios for potential risk, exposure, and compliance.

The modules are arranged to reflect the 1983 National Academy of Sciences risk paradigm (6) and are tooled to permit independent execution of each module. This structure maximizes flexibility for the user and enables a description of uncertainty at each stage of the analysis.

The first step, data retrieval, uses a SAS program (SAS Institute, Inc., Cary, NC) to gather chemical release information from a variety of EPA databases. Currently, data for reported chemical releases to air, land, and water are retrieved from the Toxics Release Inventory (TRI) (7) and the Permit Compliance System (PCS) (8) housed in the ENVIROFACTS data tables (9) on EPA's Valley Unix machine in Research Triangle Park, North Carolina. Data from the Aerometric Information Retrieval System (10), the Resource Conservation and Recovery Information System (11), and the Biennial Reporting System (12) will be added to this system.

The second step in the process generates a suite of indices, described below. (See Figure 1.) Briefly, the Chronic Index, contained in the P-RISK system's P-RISK Module, is based on the volume and toxicity of a chemical release and characterizes a chemical, facility, or Standard Industrial Classification (SIC) of interest to the user. The Vulnerability Index is divided into two components, census and disease. The Vulnerability Index: Census, also contained in the P-RISK Module, describes demographic characteristics that may render a subpopulation more likely than the general population to be harmed by toxic chemical exposures. The Vulnerability Index: Disease describes existing disease incidences and is included in the Exposure Module, along with the Exposure Index. The Compliance Index, contained in the Compliance Module, provides a statutory component to describe current and historical conformance with environmental laws.

To permit the greatest level of flexibility, the model has been developed to calculate each index independently, allowing the user to refine the analysis in a stepwise fashion and gain an appreciation for the data gaps and uncertainties at each stage of the analysis. The P-RISK's modules, some of which are currently under development, are briefly described below.

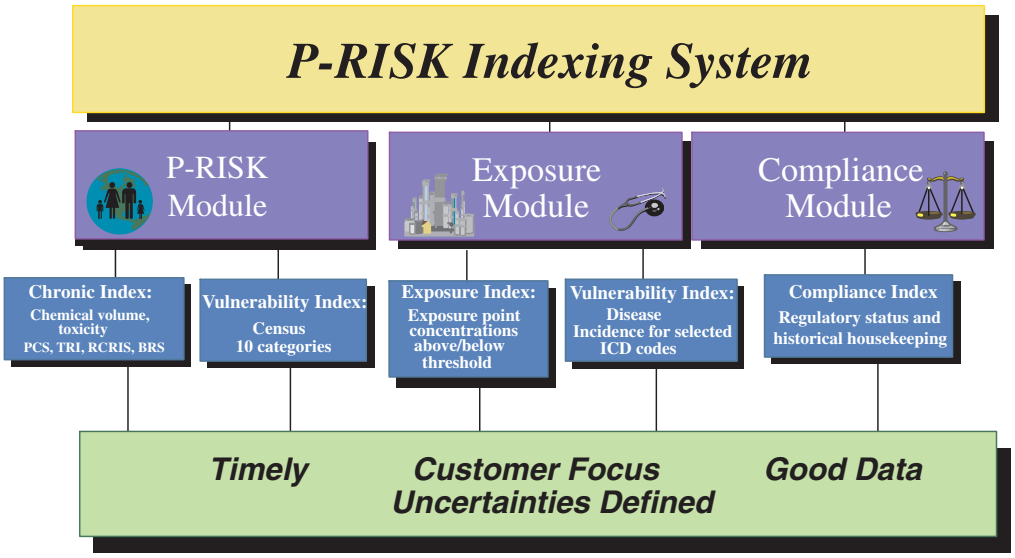


Figure 1 P-RISK structure.

The P-RISK Module has two components:

- The *Chronic Index*, a relative rank using scores derived from the reported volume of each chemical released at a site along with its associated toxicity. Individual Chronic Indices can be ranked or summed according to a given selection of chemical, facility, or SIC. The user has the option of generating these data for any or all of the 50 US states or the 10 EPA Regions. The user also has the option of

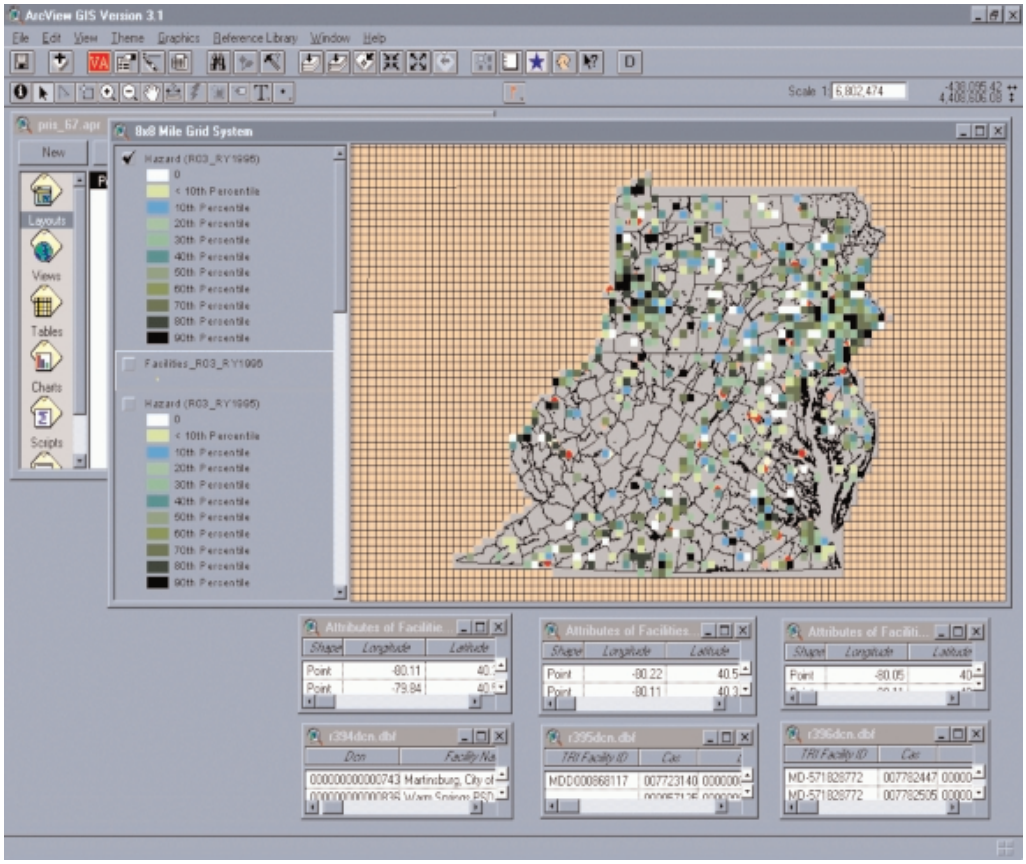


Figure 2 P-RISK interface, Chronic Index view.

creating these data as Lotus (Lotus Development Corporation, Cambridge, MA) spreadsheets or dBASE III (Ashton-Tate Corporation, Torrance, CA) files, which are stored on the user's PC and further processed to create a GIS data layer for use with the ArcView (ESRI, Redlands, CA) GIS software (Figure 2).

- The *Vulnerability Index: Census*, which provides a description of socioeconomic and demographic characteristics that may render a subpopulation more likely than the general population to be harmed by exposure to toxic chemicals. Given that "vulnerability" can mean different things to different people, for this application, the definition of vulnerability is consistent with recommendations of the

P-RISK Workgroup and in keeping with the environmental justice guidelines developed by EPA. Namely, "vulnerability" encompasses both biologically sensitive populations and potentially highly exposed populations. In general, this index is intended to highlight those populations that may be more vulnerable. In keeping with EPA environmental justice (13) and children's health protection guidelines (14), US Census data are available on CD-ROM for all 50 states. The current project uses Census data for the District of Columbia and the five states in Region III: Delaware, Maryland, Virginia, West Virginia, and Pennsylvania. The project contains Census data for the following nine fields: minority status, poverty status, age, pregnancy, female head of household with children and no husband present, educational attainment, unemployment, and age of home.

The first module, P-RISK, purposely does not account for exposure, which must be assessed using intake parameters adjusted for specific populations. Moreover, it is important to note that the term "vulnerable" does not imply causality or a mathematical relationship between a demographic characteristic and a hazard. Rather, the data are intended as an illustration, demonstrating the potential for risk and alerting the user that further analysis may be warranted.

The Exposure Module has two components:

- The *Vulnerability Index: Disease*, which includes a description of disease incidence using county-level data provided by the federal Centers for Disease Control and Prevention. A finite number of International Classification of Diseases (ICD) codes were selected based on the critical effects of priority compounds. Professional judgement should be an integral part of the ICD selection process. Clearly, the manifestation of disease can have many causal factors, including diet, lifestyle, smoking, and environmental insults. Disease occurrence cannot prove a causal relationship with nearby sources; therefore, the purpose of this index is to demonstrate the current disease burden for a selected area and provide a public health context for the P-RISK Module.
- The *Exposure Index,* which provides a method for evaluating potential exposures. This index is the most complex and possesses the greatest uncertainty. Several models are currently under consideration for inclusion in the P-RISK, including RBES (3,4), the Ecological Sensitivity Targeting and Assessment Tool (15), the Cumulative Exposure Project (16), and the Total Risk Integrated Methodology (17). For each of these models, characterization of the uncertainty in the resultant estimates is a key feature.

The Compliance Module has one component:

- The *Compliance Index*, which provides a compliance component for targeting purposes. Several existing compliance rating methodologies are currently under consideration for inclusion in P-RISK, including the Site Index Database (18) rating factors and the Sector Facility Indexing (19) approach to compliance. The Compliance Index takes advantage of compliance history both in terms of issued complaints and the complaint patterns derived from the Emergency Response Notification System database (20) and the Integrated Data for Enforcement Analysis system (21,22). The Compliance Index can serve to highlight areas with a high potential for non-notifiers by focusing on geographic or industry sectors

with large numbers of complaints or poor compliance history. At the same time, the Compliance Index enables the user to identify facilities that have had exceptional environmental and compliance histories and could serve as models for the regulated community.

The final step of the P-RISK model allows users to view and interact with the information contained in the indices, including pertinent geographical datasets that may be available on the user's home system. Using ArcView, the user creates maps, queries the data, and creates new themes on demand according to user-specified criteria. One specific coverage included in the first module of P-RISK is an 8-mile by 8-mile grid of the contiguous United States displaying areas of relatively high volume and high toxicity for TRI and PCS industrial releases for the reporting year of interest (i.e., the Chronic Index). The same module also includes demographic coverages at the census block level to display counts for each of the nine different demographic categories (i.e., the Vulnerability Index: Census).

The project also contains a reference library that gives access to several ecosystem and municipal coverages, including ambient air monitoring data, stream alkalinity, a relative ranking of watershed health, public lands, cities, and zip codes. Each of the coverages in the system is currently available for Region III, but may be revised to include coverages available on the user's home system. Each coverage is documented in the reference library, including a description, a justification for use, and a source citation. With live access to all data in one place, the user can view and query the coverages in a spatial environment, matching incidences of large-volume, highly toxic releases with areas of either complaints, high percentage of vulnerable populations, or vulnerable ecosystems.

ArcView GIS Project

The P-RISK model is delivered in two parts: (1) a user-friendly, interactive ArcView GIS project that can create maps, query data, and create new themes on demand according to user-specified criteria, and (2) an automated SAS program that can be customized to user specifications for risk level, area of concern, and type of output. The objective of P-RISK is to provide the user with real-time access to the data and the discretion to overlay, view, and query the different indices generated in the modules.

Data Analysis

The following steps are taken to produce the data layers or themes for the GIS maps. All coverages in the ArcView project are projected in albers equal-area conic with spheroid GRS 80, central meridian equal to -79, first standard parallel equal to 37, second standard parallel equal to 41, and latitude of projection's origin equal to zero. There is no false easting or northing; both are set to zero. The datum is the North American Datum of 1983, or NAD 83.[1]

[1] For more information on NAD 83, see the Geodetic Glossary (Rockville, MD: National Geodetic Survey, National Ocean Service, National Oceanic and Atmospheric Administration, September 1986) or see questions 2 and 3 at http://www.ngs.noaa.gov/faq.shtml.

Chronic Index Coverage

SAS code is used both to retrieve data from the EPA databases and to calculate the Chronic Index. The final output is provided as four files (one each for chemical, document control number, facility, and SIC code) in either Lotus or dBASE format and is stored on the user's PC. The algorithm also contains a quality assurance check to ensure that the code is executing properly.

P-RISK allows the user the opportunity to map the Chronic Index by facility, SIC code, and chemical. The output files, stored on the user's PC, are retrieved from within ArcView to create a color-coded grid expressing potential risk. This data layer can be manipulated to illustrate the rank of a chemical release in a particular grid, the facilities responsible for the toxic chemical release, and the SIC codes that are primarily responsible for the toxic release. The user has the option to create files either in Lotus or dBASE format. The user's manual provides details on how the program calculates the Chronic Index and aggregates the data by chemical, facility, and SIC code.

Census Block and Demographic Information

Census block group coverages or data layers are available from various electronic sources. For example, the Regional census block group boundary coverages can be found at the following EPA FTP site: ftp://valley.rtpnc.epa.gov. Because these electronic sites may vary from Region to Region, each EPA Regional GIS group should be contacted for specific boundary coverage locations. Once the boundary coverage has been obtained, it can be coupled with demographic information to create additional data layers. In the Region III example, the data for nine census fields (downloaded from the 1990 US Census CD-ROMs) were permanently joined to the block group boundary coverage. For each of the nine census fields, the block groups are sorted from lowest to highest and divided into ten equal percentiles. This ranking permits the user to view any percentile for any census field, such as the top 10% of the Region's census block groups for poverty.

The Fetch Button

The Fetch button is used to process the TRI/PCS facilities database. After clicking on the button, the system prompts the user for the location of the facilities database file to process. The data are then retrieved into ArcView by Fetch and run through checks to see if the proper information is contained in the file. Once the database file passes the checks, a point coverage is created from the available latitude and longitude data. If the latitude and longitude field is not populated, then a point is created using the zip code centroid. Statistics are then run on the new facility point coverage to create 10 percentiles for each of the nine chemical release categories:

- Fugitive air releases
- Stack air releases
- Water releases (containing releases from either PCS or TRI)
- Underground releases
- Land releases
- Onsite releases
- Releases from publicly owned treatment works

- Offsite releases
- Total releases and transfers

The new facility point coverage and its attached percentile database are then added to the view and a legend is created. From the new facility coverage, an 8-mile by 8-mile grid aggregation of the facility points is created. Statistics are also run in the 8-mile by 8-mile aggregated grid to produce percentiles for each of the nine chemical release categories. After adding the grid theme to the view, the system prompts the user for the location of the chemical table. Once the corresponding chemical table is found, it is retrieved into ArcView by Fetch and linked to the facility point coverage. This marks the successful completion of the program, leaving the user with a ranked, color-coded grid coverage and a ranked facility point coverage linked to individually ranked chemical releases. The P-RISK retains up to three years of processed data to enable the user to investigate recent trends in hazard.

Blue Star Query 

From this data universe, the Blue Star button allows the user to select a subset of information contained in any data theme in P-RISK. For instance, the user can use the Blue Star to select the top 10% of the fugitive air combined index to identify and display those facilities releasing chemicals with the highest volumes and toxicities. The user is prompted through a series of help menus to locate the item of interest. Once the query is completed, the resultant subset is linked to both the facility and chemical tables so that the user can determine specific chemicals, their sources, amounts, rank, carcinogenic or non-carcinogenic toxicity, uncertainty, and other attributes.

Red Flag Query 

The Red Flag button allows users to select individual sites from the subset created with the Blue Star. Using this feature, the user can access the row within the data table associated with the particular site. All selected sites can be moved to the top of the table by clicking on the top border of the table and then clicking the “promote” icon.

Reference Library

The Reference Library enables the user to access other available data coverages, consolidating desired information in one place. This structure facilitates place-based analyses and assists in identifying the relative contribution of several different impacts. Online documentation is also available.

References

1. US Environmental Protection Agency (EPA). 1995. *Addendum to Chemical Indexing System Part I: Chronic Index technical guidance manual.* EPA Region III, Air, Radiation and Toxics Division. EPA/903/R-93/002-a. August.
2. US Environmental Protection Agency (EPA). 1993. *Chemical Indexing System Part I: Chronic Index technical guidance manual.* EPA Region III, Air, Radiation and Toxics Division. EPA/903/R-93/002.

3. US Environmental Protection Agency (EPA). 1994. *Risk-Based Enforcement Strategy: Final report, September 30, 1994.* EPA Contract 68-D3-0013, Task Number 2-15. Prepared by Versar, Inc., Springfield, VA.
4. US Environmental Protection Agency (EPA). 1994. *Risk-Based Enforcement Strategy user's guide, September 30, 1994.* EPA Contract 68-D3-0013, Task No. 2-16. Prepared by Versar, Inc., Springfield, VA.
5. US Environmental Protection Agency (EPA). 1998. *Peer review package for the Potential Risk Indexing System (PRIS), September 30, 1998.* Report submitted to EPA by Science Applications International Corporation, McLean, VA. EPA Contract 68-D4-0098, WA IV-5.
6. US Environmental Protection Agency (EPA). 1992. *Guidelines for exposure assessment. Risk Assessment Forum.* EPA/600-Z-92-001.
7. US Environmental Protection Agency (EPA). 1997. *Toxics Release Inventory (TRI) relative risk-based environmental indicators: Methodology, June 1997.* Washington, DC: Office of Pollution Prevention and Toxics, US Environmental Protection Agency. EPA/740-R-97-002.
8. US Environmental Protection Agency (EPA). 1997. *Permit Compliance System (PCS) enforcement action national file* (on diskette). Washington, DC: Office of Wastewater Enforcement and Compliance, US Environmental Protection Agency. NTIS PB97-502314.
9. US Environmental Protection Agency (EPA). 1999. *ENVIROFACTS data warehouse and applications.* Office of Environmental Information, US Environmental Protection Agency. http://www.epa.gov/enviro/.
10. US Environmental Protection Agency (EPA). 1994. *Aerometric Information Retrieval System (AIRS) user's guide.* Washington, DC: National Air Data Branch, Technical Support Division, Office of Air Quality Planning and Standards, US Environmental Protection Agency. EPA-454/B-94-007.
11. US Environmental Protection Agency (EPA). 1998. *Resource Conservation and Recovery Information System (RCRIS) merged database administrator guide: v.7.0.0, February 1998.* Washington, DC: Office of Solid Waste, US Environmental Protection Agency. SCD-0055-033-RL-7051.
12. US Environmental Protection Agency (EPA). 1995. *Biennial Reporting System (BRS) data: Generation and management of hazardous waste, 1995 (final)* (on CD-ROM). NTIS PB98-500077. Washington, DC: Office of Solid Waste, US Environmental Protection Agency.
13. White House. 1994. *Federal actions to address environmental justice in minority populations and low-income populations: Executive Order 12898.* President William J. Clinton. Washington, DC. Executive Order 12898. February 11, 1994.
14. White House. 1997. *Protection of children from environmental health risks and safety risks: Executive Order 13045.* President William J. Clinton. Washington, DC. April 21, 1997.
15. US Environmental Protection Agency (EPA). 1999. *Ecological Sensitivity Targeting and Assessment Tool (ESTAT).* Science Applications International Corporation. http://www.epa.gov/envirofw/html/factsheets/estat.html.
16. US Environmental Protection Agency (EPA). 1999. *Cumulative Exposure Project (CEP).* Office of Policy, US Environmental Protection Agency. http://www.epa.gov/CumulativeExposure/home.htm.
17. US Environmental Protection Agency (EPA). 1999. *Total Risk Integrated Methodology (TRIM) technical support document: External review draft.* Research Triangle Park, NC: Office of Air Quality Planning and Standards, US Environmental Protection Agency. EPA-453/D-99-001. November.

18. Williams S. 1998. Personal communications with Stephen Williams, Delaware Department of Natural Resources and Environmental Control. Dover, DE. June.

19. US Environmental Protection Agency (EPA). 1997. *Science Advisory Board report: Review of the Sector Facility Indexing Project (SFIP).* Washington, DC: EPA Science Advisory Board. EPA-SAB-EEC-97-012.

20. US Environmental Protection Agency (EPA). 1998. *Emergency Response Notification System (ERNS), 1996* (on CD-ROM). Washington, DC: Office of Emergency and Remedial Response, US Environmental Protection Agency. NTIS PB98-500663.

21. US Environmental Protection Agency (EPA). 1996. *Integrated Data for Enforcement Analysis (IDEA) system.* Office of Enforcement and Compliance Assurance, US Environmental Protection Agency. EPA/300/B-95/004D.

22. US Environmental Protection Agency (EPA). 1996. *IDEA basic training course: Student booklet, Integrated Data for Enforcement Analysis, March 1996.* National Technical Information Service (NTIS). PB96-780507.

Strategies for GIS and Public Health

Michael F Goodchild*
National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA; Department of Geography, University of California, Santa Barbara, CA

Abstract

The paper is divided into three sections. The first reviews three broad trends in information technology that will affect geographic information systems (GIS) in the coming years. The second identifies four trends that are specific to GIS, including availability of new data, trends in software, developments in education, and opportunities in new hardware. The third section recommends six potential strategies for advancing applications of GIS to public health. The paper ends by suggesting an analogy between imaging the body and imaging the health of the nation through GIS display and analysis.

Keywords: public health, spatial data infrastructure, future trends, strategies

Introduction

As the conference closed, I was struck by two things: the refreshing diversity among the conference attendees, and the sheer size of this geographic information system (GIS) application domain. Many GIS conferences still have a long way to go before their audiences resemble America, but this one seemed much closer to achieving that goal. The conference showed ample evidence of the vast range of health applications of GIS in the variety of topics among the roughly 130 papers, the posters and demonstrations, and the pre-conference workshops. It showed the power of GIS for mapping, but also for gaining new insight by displaying data in new ways and by organizing data and programs geographically. Yet it seemed to me that the conference demonstrated the potential for a community of health professionals using GIS that would be as much as two orders of magnitude greater in number than the present community, given the potential for improved health as a goal of many, many different types of GIS applications.

The following three sections address the three distinct purposes of this paper:

1. To identify trends in society and in information technology that are larger in scale than GIS, and that will inevitably affect the application of GIS to health problems.
2. To identify trends that are specific to GIS, but still larger than any one application.
3. To suggest some strategies for the community interested in GIS applications to health problems.

The three sections are followed by a brief conclusion.

External Trends

The Mapping Science Committee (MSC) of the National Research Council exists as "a

* Michael F Goodchild, Dept. of Geography, University of California, Ellison 3611, Santa Barbara, CA 93106-4060 USA; (p) 805-893-8049; (f) 805-893-3146; E-mail: good@ncgia.ucsb.edu

focus for external advice to the federal agencies on scientific and technical matters related to spatial data handling and analysis," and has been instrumental in initiating the concept of the National Spatial Data Infrastructure (1). In 1997, the MSC published a report entitled *The Future of Spatial Data and Society* (2), which reported on a workshop sponsored by the committee in 1996. The report reviews trends in society at large that are likely to impact GIS in the next 15 years, and lays out a number of alternative visions for the role of GIS in society in 2010. Below, I review three of those trends, chosen because they seem to be of particular significance to applications of GIS to health problems.

Information Technology

One of the most remarkable trends of the past 15 years has followed the prediction attributed to Gordon Moore, co-founder of Intel, that processor speed would approximately double every 18 months, and that processor costs would stay approximately constant. Two things are remarkable about the prediction: first, the precise way in which the actual performance of the industry has matched it; and second, the fact that it was proposed by one of the key figures in that industry, rather than an independent observer. Over the past 15 years and since the advent of the IBM PC, the speed of a PC's central processor has increased by a factor of roughly 1,000, and costs have stayed roughly the same. Similar improvements in price and performance have occurred in other key areas of the computing industry—memory, hard disks, and CDs.

Over the years, there have been numerous speculations on the accuracy of Moore's Law, and in mid-1997 a number of articles in the popular press announced its demise. But while most speculations have focused on possible under-performance by the industry, the mid-1997 articles predicted over-performance: that speed would begin to rise at an even greater rate due to a series of key breakthroughs in chip design and manufacture. At this point in time, there seems very little reason to be pessimistic about the future of information technology.

Looking back, it is interesting to ask where all the new cycles and bytes made possible by Moore's Law have gone; certainly, there is little evidence that every GIS application is making 1,000 times the number of numerical calculations it was making in 1984. Instead, it seems that much of the new power has gone into areas that in 1984 would have been regarded as somewhat superficial: maintaining a graphic user interface with a specific "look and feel," supporting a connection to the network, and in general making the system easier to use. It seems that every new version of the operating system is friendlier and more visual, but also demands more resources. One might even consider a corollary to Moore's Law: that such non-essential aspects of computing will consume a constant proportion of the increasing resources. Certainly, and despite continued concerns about the difficulties of learning about and making effective use of GIS, the software is in general far easier to use today, and far more productive, than it was in 1984.

The Network

It is amazing to consider that the World Wide Web (WWW), the much-hyped engine of electronic commerce, darling of investors, haven of conspiracy theorists and pornographers, and harbinger of massive changes in our educational system, is only five years old, and was invented only ten years ago by a physicist looking for better ways to share

information with colleagues. Today, it is common to talk about the WWW as a vast information resource, and to suggest that we are drowning in a flood of information partly as a consequence of its power and popularity. Certainly, the WWW is having and will continue to have a powerful impact on the development of GIS that will continue over the next 15 years.

Consider for a moment a comparison between the WWW and the average research library, such as exists on my own campus. A typical research library has on the order of 10^6 books in its collection, each containing perhaps 10^5 words, which could be encoded in perhaps 10^6 bytes. Thus, the text in the library might amount to on the order of 10^{12} bytes, or a terabyte. If we assume that the average person reads at 5 words per second, it would take 3,000 years of reading, or 50 lifetimes, to exhaust the library's information resources. Of course, the volume of information in the library is increasing faster than anyone can read, and these figures ignore everything in the library that is not text. In short, the information in the library drowned us long ago. Perhaps it didn't seem that way, because the library is an ordered space of uniform shelves and rows of well-cataloged books, whereas the WWW is a chaotic space of information that is hard to find and may be unreliable when it is found. The library's information has been edited, proofread, and reviewed—it is quality information—and libraries pay collection specialists to ensure that the information in the library is accurate, meaningful, and useful. Libraries also provide the means to find and retrieve information, in ways that are far more sophisticated than the average WWW search engine.

If libraries are so efficient at providing information, what exactly is the function of the WWW? I suggest that the value of the WWW lies in its power as a source of information *not found in libraries.* Although the library has been an excellent mechanism for disseminating information in the form of books, the WWW is clearly better for information that is:

- Timely, and for which the delays in library dissemination due to the lengthy process of writing, publication, and review (which can often take over a year) are unacceptable.
- Hard to handle in traditional form because of problems with the basic medium (photographs, recordings, or maps), or because of problems of cataloging, and where digitization removes these problems.
- Not of general interest (this might include information of personal or local interest), and therefore of little interest to publishers because the economics of the publishing industry favor production in large numbers.

From this perspective, the WWW is an ideal mechanism for distribution and sharing of geographic data. It solves problems of timeliness because, although much geographic information is static, it is difficult for the normal publishing mechanism to deal with updates, corrections, and immediate need. It solves problems of handling, because a map or image in digital form is in principle no more difficult to handle than a book or manuscript in digital form; both are "bags of bits" to a digital network. The WWW also helps solve the problems of cataloging geographic data, because it can process queries about areas on the Earth's surface, something that is very difficult with a conventional card catalog (3). Finally, many types of geographic data fit the third criterion, because detailed data about a local area are not likely to be of major interest in areas outside the immediate region (4).

In the digital world of the WWW, it is possible for anyone equipped with a simple PC to be a publisher. Moreover, advances in geographic information technology, including GIS and the Global Positioning System (GPS), and massive reductions in cost have made it possible for a large number of people and agencies to begin publishing geographic information, despite the fact that much geographic information is of very limited interest. This is revolutionizing traditional arrangements for production and dissemination, which have emphasized central production at the national level and at public expense. Today, WWW-based projects such as the National Geospatial Data Clearinghouse (http://www.fgdc.gov), the Alexandria Digital Library (http://alexandria.ucsb.edu), and Microsoft's Terraserver (http://www.terraserver.com) offer substantial alternatives to the traditional role of the specialized map library, providing information that is more timely and easier to handle, and possibly of very limited interest.

The impact of the WWW on GIS is much greater than its role as a mechanism for dissemination, however. The WWW is an instance of client-server technology, in which operations are divided between a local client provided by the user and a server provided by the host site. Much of the WWW's genius lies in how operations are divided, and in principle it is possible to make use of the WWW from a client that is extremely simple, perhaps nothing more than a $100 add-on to a home television. In such situations, all of the serious computing is done by the server.

Take the example of geocoding, an important function in GIS applications to health problems. Suppose I have a list of 1,000 addresses of patients, and I want to convert them to coordinates in order to map them, or to analyze them in relation to other data. I have two options in today's computing world. First, I could purchase and install a GIS on my desktop, purchase or obtain the necessary data files, enter the addresses, obtain the coordinates, and construct the map. Alternatively, however, I could send the addresses to a WWW site that offers geocoding services, receive the results, and perhaps send them to another site that offers mapping services. The example illustrates the increasingly important distinction between GISystems, as they have been understood for the past two decades, and GIServices, an important and growing area of the WWW. Today, simple GIServices are typically free, financed by advertising revenues. The Mapquest site (http://www.mapquest.com) is an excellent example. Others are more complex, accurate, and timely, and these are services the user should expect to pay for, by providing a credit card or some other straightforward method of electronic payment. Günther and Müller (5) provide an interesting overview of this rapidly developing area of electronic commerce.

How far will GIServices develop, and how much of GIS computing will be transferred to WWW servers? Five years ago, the question was meaningless, but today it is critically important for the future of GIS. What does the concept of GIServices mean for public health? Should agencies be providing GIServices for their clients, as many agencies in other areas already seem to be doing? And how can a public health agency add value to its information by providing services based on it, rather than providing the information itself in raw form?

Software

One of the effects of Moore's Law, and the use of new computing power to build friendlier user interfaces, has been a vast increase in the population of computer users. In 15

years, we have moved from an era in which computing was the preserve of a small elite to one in which virtually everyone, from children to senior citizens, expects computing to be accessible and able to do something useful for them in their daily lives without a great expenditure of effort in training. The personal computer has empowered everyone to compute, write, calculate, and make maps. Everyone today can install a GIS or go to a WWW site and make maps, and no longer is mapping the preserve of a few trained cartographers. In that sense, GIS is truly destroying cartography, though in other senses it is breathing new life into an old and respected discipline.

Computing has put enormous power in the hands of individuals, and produced massive changes in human behavior. So an all-important question for GIS and public health is: How can we harness this enormous empowerment *to improve public health*? As specialists, we can have much greater impact if we focus on empowering others, rather than on the ways in which GIS helps us do our own jobs.

Trends in GIS

All of the issues discussed in the previous section derived from outside the world of GIS, and yet will affect how GIS develops in the next few years. This section addresses issues and trends that are more specific to GIS. Although the MSC report (2) discusses a large number of these, I have selected four that have significance for GIS and, in particular, its applications in public health.

New Data

Over the next few years, a number of new data sources will be coming online with potential for GIS and public health, and it is important that we take advantage of these opportunities as effectively as possible. First, a new generation of satellites will be producing imagery with a resolution of 1 meter. Remote sensing has already proven useful for detecting conditions of importance to public health, such as breeding areas for disease vectors, but this improvement in spatial resolution will create a host of new opportunities for mapping and monitoring conditions that can only be detected at this level of detail. Many indicators of housing quality become visible, for example, as do other socioeconomic variables such as new housing and other indicators of population.

The GPS is also continuing to have impact on data availability. The development of kinematic GPS and its use in vehicles has already reduced the cost of mapping streets by an order of magnitude. Other types of data may make it possible to develop better indicators of lifetime exposure to environmental risk, and new types of census data may make it possible to maintain much more current perspectives on demographic and socioeconomic conditions than is possible with the current decennial system.

New Software

Within the GIS software industry there is a strong interest in achieving interoperability through the adoption of open standards that allow systems developed by different companies to work together without operator retraining or data reformatting. The Open GIS Consortium (http://www.opengis.org) has spearheaded much of this interest, and has developed a number of critically important specifications. It is already possible to operate GIS within Microsoft's Excel, and to open statistical analysis

packages without leaving ArcView (ESRI, Redlands, CA), and the indications are that this trend to interoperability will accelerate in the next few years.

One of the benefits of such open specifications is that GIS becomes easier to learn and use, because open specifications require that every vendor adopt the same terminology, or make it possible for someone using another vendor's terminology to use a system without retraining. Open systems inevitably lead to more focus on principles, and less on the details and idiosyncrasies of specific systems. Perhaps surprisingly, in the future, there should be less to learn about GIS.

New GIS-Aware Generations

Thus far, much of the leadership in GIS education has come from the four-year universities, and the majority of courses have been offered at the upper-division or graduate levels. Increasingly, however, the education community has begun to address the needs of other sectors, including non-traditional students who are unable to enter full-time college programs. Distance learning is now well established in GIS through such programs as UNIGIS International (http://www.unigis.org) and ESRI's Virtual Campus (http://www.esri.com). Many community colleges now offer GIS programs, and there are several sources of instructional materials on GIS for this sector (see, for example, http://www.ncgia.org/education/ed.html). GIS is being offered in high schools, and there is interest in exploring its use in elementary schools: imagine, for example, being introduced to the concept of measurement not by using a thermometer to measure temperature, but by using a GPS receiver to measure latitude and longitude.

Over the next decade, the GIS and public health community can expect a much higher level of GIS awareness in its new recruits, and much greater accessibility to GIS functions and expertise.

New Hardware

We have grown used to the idea of computing on a desk, either in the office or at home; for years, computers have been tied to sources of power, and now to Internet connections that only exist where there are phone lines. But it is now possible to compute with full desktop functionality using a portable laptop operating on batteries, and downsizing in the industry is now making it possible to compute on palmtop computers and in vehicles. Moreover, the growth of the wireless communication industry has made it possible to connect from anywhere, and we are rapidly entering the world of mobile, ubiquitous computing, in which location is no longer constrained. We can download data into the field, and upload field-collected data to the Internet or to the office.

Field computing is likely to make a major impact on public health, since it will permit a range of new and exciting opportunities:

- The ability to analyze information as it is collected, rather than later, when the field worker returns to the office.
- The ability to download patient records and other background information to onsite interviews.
- The ability to manage emergencies in the field, on site, and yet have full access to the background information that is needed for effective decision-making.

Strategies for GIS and Health

Given these trends, what can we as a community do to advance the use of GIS in solving public health problems, and to improve public health using these remarkable technologies? I would like to make six suggestions.

Education

First, I suggest that we have to do more to prepare the next generation of public health professionals for GIS, and to raise the awareness of the current generation, much of which was educated before the advent of GIS. We can do this by promoting instruction in GIS in public health schools and by developing partnerships in education with other disciplines that already have well-developed GIS programs: geography, computer science, and geomatics, to name a few. We can promote instruction in GIS at all levels of the education system, with workshops for teachers on GIS applications in public health. We can provide similar opportunities for professionals, through informal education, part-time education, distance learning, the WWW, and workshops at conferences like this.

Research

Second, I suggest there are specific research issues that, if addressed, can improve the use of GIS in public health. Public health applications often require a local focus and the use of local data at the individual level rather than the highly aggregated data that have characterized many previous applications of GIS. They often require an approach that is exploratory, visual, and intuitive. (Anselin [6] has made an excellent review of exploratory spatial data analysis in GIS.) Much health information is uncertain, incomplete, or inaccurate (see, for example, the chapters on data quality in the recent compendium by Longley et al. [7]), and analysis often must be conducted at multiple levels of aggregation.

Data

The WWW is a wonderful resource, but it is not by itself the solution to the need for data in GIS public health applications. Effective searching over the WWW requires the creation of catalog information—metadata, in the language of GIS data access—to enable users to find data more easily and, once they find it, to assess the fitness of data for a given application (3). The National Geospatial Data Clearinghouse is a very effective mechanism for finding data at the national level, but similar efforts need to be promoted at state and local levels, and with the specific needs of public health in mind.

Hardware

Advances in GIS have always relied on advances in hardware generally, and tools like the plotter and the tablet digitizer have given the field very effective boosts in the past. Wireless communication and portable devices seem set to provide comparable opportunities in the future, and there will be other advances we have not even thought about. We need to watch for new opportunities in information technology, and think about how they can improve applications of GIS to public health problems.

Software

GIS is a huge application of information technology, responsible for perhaps $10 billion annually in the United States. Public health is one of many applications of GIS; for GIS vendors, it represents a niche market that may grow into a very significant proportion of the overall market, but as yet is relatively small. One GIS size may not fit all applications, and it is already clear that public health applications present particular needs. We need to promote the development of specialized GIS for public health, and to encourage small software developers who may be able to flourish in this niche to provide add-ons to general-purpose software and systems.

Communication

I noted at the beginning that the potential community of people interested in GIS applications in public health was much larger than this conference, perhaps by two orders of magnitude. We need to find ways to reach that wider audience, and a single national conference cannot possibly do that, given the restrictions on travel that most public health workers face. Too many public health workers see GIS as one of many interesting areas competing for their attention, and conferences like this are at best forums for discussion among national-level agencies and specialists (only 15% of the participants are from state agencies and only 15% are from local agencies, though these figures bear no relationship to the real sizes of these sectors). The solution, it seems to me, is to promote regional and local conferences and workshops in addition to this national forum, perhaps through existing state- and regional-level GIS conferences and organizations, and local chapters of national organizations such as the Urban and Regional Information Systems Association. We should encourage the development of specific public health tracks at these conferences, and also make use of other mechanisms like Chuck Croner's excellent GIS newsletter. (To subscribe to or receive a copy of *Public Health GIS News and Information*, a free bimonthly e-mail report, contact Dr. Charles Croner at cmc2@cdc.gov, or call 301-436-7904, ext. 146.)

The Whole Body Metaphor

I would like to conclude with a point that is somewhat abstract, but nevertheless seems important and an appropriate point with which to end. Metaphors seem particularly useful in trying to address the future, because they allow us to reason and obtain insight through parallels with other fields and concepts. An interesting concept in cognitive psychology is the idea that we learn about the world in childhood by extending concepts from our bodies to our surroundings. Linguists might offer evidence of this in language: "the head of the lake," "the finger lakes," or "the heart of Dixie," while Lakoff (8) has suggested that it is one basis for reasoning about space (and see Mark and Frank [9]).

I suggest that the relationship between GIS and the world is somewhat like the relationship between medical imaging and the human body; a state-of-the-environment report is rather like an individual medical checkup, a report on the state of the geographic body if you like. It is important to the individual if a part of the human body is not working, and it is arguably equally important to society if a part of the world is in bad health. Especially important, if media attention is anything to go by, are problems

that are sharply focused in space and time, such as Legionnaire's disease or the Ebola virus.

So my suggestion is that we promote GIS as a tool for exploring the state of the nation's health, with associated diagnostics, policies, and interventions, just as we promote medical imaging as a tool for exploring individual health. The value of a metaphor like this lies, of course, in the thoughts that it provokes: do we need, for example, to develop a profession called "health spatial analyst" that is modeled on the profession of radiology? And what can we learn from the profession of radiology that can help us in imaging the geography of human health?

Acknowledgments

The National Center for Geographic Information and Analysis is supported by the National Science Foundation. Additional support for NCGIA's role in this conference was provided under the Public Health Conference Support Grant Program of the US Department of Health and Human Services.

References

1. Mapping Science Committee. 1993. *Toward a coordinated spatial data infrastructure for the nation.* Washington, DC: National Academy Press.
2. Mapping Science Committee. 1997. *The future of spatial data and society*. Washington, DC: National Academy Press. http://www4.nas.edu/cger/besr.nsf.
3. Goodchild MF. 1998. The geolibrary. In: *Innovations in GIS 5*. Ed. S Carver. London: Taylor and Francis. 59–68.
4. Goodchild MF. 1997. Towards a geography of geographic information in a digital world. *Computers, Environment and Urban Systems* 21(6):377–91.
5. Günther O, Müller R. 1999. From GISystems to GIServices: Spatial computing on the Internet marketplace. In: *Interoperating geographic information systems.* Ed. MF Goodchild, MJ Egenhofer, R Fegeas, CA Kottman. Norwell, MA: Kluwer Academic Publishers. 427–42.
6. Anselin L. 1999. Interactive techniques and exploratory spatial data analysis. In: *Geographical information systems: Principles, techniques, applications and management*. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: Wiley. 253–66.
7. Longley PA, Goodchild MF, Maguire DJ, Rhind DW. 1999. *Geographical information systems: Principles, techniques, applications and management*. New York: Wiley.
8. Lakoff G. 1987. *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.
9. Mark DM, Frank AU, eds. 1991. *Cognitive and linguistic aspects of geographic space.* Dordrecht, The Netherlands: Kluwer.

GIS in a County Environmental Health Agency

Peter J Isaksen, RS,* Margaret M Blanchet, REHS, Todd W Yerkes, RS, Carl Osaki, RS
Seattle-King County Department of Public Health, Seattle, WA

Abstract

A geographic information system (GIS) is an effective tool that local health departments can use in their environmental health programs to perform community health and environmental assessments, improve public access to environmental health information, and increase department effectiveness and efficiency. In the daily work required to protect the public's health, environmental health programs collect large sets of useful data. Most of these datasets have a geographic component. While integral to daily environmental health tasks, these datasets can have many additional applications, particularly as the field of environmental health grows more assessment-oriented. The trend of local government agencies tracking their activities with GIS offers environmental health programs a unique opportunity to share information while serving the public's interest in health and a healthy environment. Implementation of GIS requires several components, including identification of current needs and possible future uses; cooperation with other county agencies; management commitment; budget allocations; and access to technical GIS staff. Some of these components are not readily available in an environmental health program, but can be found in other county agencies. Development of a separate GIS for environmental health purposes is an unnecessary duplication of work. Using GIS to merge data from multiple county agencies is an efficient way to deliver environmental health information. To test this, a traditional environmental health program management task was compared with the same task performed using existing county GIS resources. Use of GIS resulted in increased work efficiency, access to more complete information, and improved public notification.

Keywords: environmental health programs, county health department, hazardous waste site assessments

Environmental Health Programs

Multiple Programs and Locations

The Seattle-King County Department of Public Health (Health Department) is located in King County, Washington, the twelfth most populous county in the United States. It serves a population base of 1.6 million people. The Health Department has over 1,200 employees; its Environmental Health (EH) division has over 160 employees. EH consists of multiple programs in areas such as food protection, living environments, meat/poultry/rabbit/aquatic foods, drinking water, on-site septic, solid waste, site hazard assessment, local hazardous waste management, chemical/physical hazards, vector/nuisance control, and plumbing/gas piping inspections. Most division

* Peter Isaksen, Seattle-King County Dept. of Public Health, 999 Third Ave., Suite 700, Seattle, WA 98104-4039 USA; (p) 206-296-4724; (f) 206-296-4575; E-mail: peter.isaksen@metrokc.gov

programs are administered at four regional locations and/or at a central technical support office.

Interactions with City, County, State, and Federal Agencies to Provide the Community with Better Access to Information

King County includes the cities of Seattle and Bellevue, plus 36 suburban cities and the unincorporated areas of the county. EH routinely interacts with many other county agencies including local planning agencies, building departments, and the county Assessor's Office. In addition, the Health Department is connected through various grant projects and mutual environmental and health functions to the Washington State Department of Health, the Washington State Department of Ecology (Ecology), and to federal agencies such as the National Oceanographic and Atmospheric Administration (NOAA) and the US Environmental Protection Agency (EPA). Numerous opportunities exist to share information with these entities, with the Health Department involved in the health aspects.

The Health Department—like the Assessor's Office, the court system, the Records and Elections Department, and the Regional Parks and Facilities Department—is a regional department that provides services countywide in King County. Other county departments—such as the Department of Development and Environmental Services (DDES), which is the county building department—only have jurisdiction in the unincorporated areas of the county and a few contracted suburban cities. As a regional entity, the Health Department provides services throughout King County, including the cities of Seattle and Bellevue, the suburban cities, and the unincorporated parts of the county.

These interactions put EH in a unique position to function as an intermediary for sharing data between agencies. Local information is often more specific than state and federal information because it is updated more frequently and is more verifiable. The EH geographic information system (GIS) provides a logical location where various data sources can be brought together for use in evaluating health issues, benefiting the community as a whole. This role as a central data location increases the capacity of EH to interact more closely with other divisions within the Health Department and with other local, state, and federal agencies. It also enables EH to provide better assessment capabilities and risk communication to impacted communities.

Countywide GIS Program

Structure

Over the past seven years, several King County agencies have worked cooperatively toward establishing a countywide GIS. Agencies involved include Transportation, the Department of Natural Resources, the Office of Budget and Strategic Planning, the Assessor's Office, and DDES. There has been a concerted effort among these agencies to share data on a central server and to encourage other agencies to participate in the countywide GIS project. DDES uses GIS in conjunction with the computer permitting software, Sierra Permits (Sierra Computer Systems, Inc., Visalia, CA), that the Health Department currently shares with them. Both DDES and EH are in the process of

upgrading their permit systems for Year 2000 compliance and to allow an upgrade of the user interface to a Windows environment.

While EH addresses public health issues, the division's interaction with DDES centers on land development permitting issues such as plumbing, drinking water, and septic systems. Both agencies perform a good portion of work using a parcel-based system and use much of the same geographical information to make respective agency decisions. Inclusion of health data into the building permit process allows health issues to become an integral part of the decision-making process.

With one exception, all King County agencies use ESRI (Redlands, CA) software—ARC/INFO and ArcView. ARC/INFO consists of digitized information that ArcView can access (e.g., parcels, street addresses, floodplains, assessor's maps). ArcView is a user-friendly desktop interface and can be used to combine background geographical information with environmental health data (e.g., permit information, septic-system failures). Other software packages are available, although the decision to go with the county standard seemed an obvious one.

Data Available on System

A large server holds the county GIS data files in a central location. These data files are updated by various GIS programmers employed throughout the county. Some information is updated in real time as the data are entered, but most data tables must be updated manually by the respective agencies on a continuous basis.

Benefits of Using GIS in Environmental Health

GIS Program as Community Assessment Tool

In 1993, Washington State began implementation of the Public Health Improvement Plan (1), which called upon local health agencies to collect and examine data to identify trends of disease and injury; work with communities and decision-makers to target particular issues; and assure services meet community needs. Data collection and analysis are key components of the community assessment process to be performed by local health departments. GIS is an evaluative tool that can be used to examine these datasets spatially. Looking at environmental health data on a map can allow identification of trends and patterns such as failing septic systems in a particular region, an increase or decrease in critical item restaurant violations over time, and so on. GIS can also be used to track environmental health activities in a particular region and to provide communities with site-specific information about these activities. Certain sites may in turn be targeted for specific outreach activities whose outcomes will be monitored and evaluated over time.

Improved Communication

Being a large environmental health division with multiple programs and multiple locations, it is both important and helpful to facilitate communication within the division, with other county agencies, and with the public. Environmental health specialists working in one program are only generally aware of the routine activities taking place in other programs. GIS can be used to link program information together by an address or parcel number and to make the information available to a wide audience. It can be

used as a management tool to evenly distribute inspection workloads and to evaluate the effectiveness of specific programs.

GIS also provides a way to improve public access to environmental health information. Our interaction with the public takes many forms including answering questions about a particular property, drinking-water well, or restaurant, and/or issues affecting a particular community. GIS can assist environmental health staff in answering questions regarding the status of a particular permit or can be used as a tool to evaluate trends in environmental health data.

Increased Effectiveness and Efficiency

Much of the data traditionally collected in paper form on a monthly or yearly basis are now updated automatically in the county GIS. A countywide GIS also brings together data from multiple sources and locations. Data gathering and organization become more efficient as information traditionally collected in duplicate (i.e., multiple databases containing many of the same fields) is compiled in one central location.

Site Hazard Assessment Program

Program Overview

The Site Hazard Assessment Program (SHA) conducted by EH is grant funded by Ecology to investigate, assess, and rank potentially contaminated hazardous waste sites in King County. The Known or Suspected Hazardous Waste Sites List compiled by Ecology under the Washington State Model Toxics Control Act (MTCA), which mandates cleanup of hazardous waste sites, contains numerous sites from several sources within Ecology. Due to the large volume of potential sites on this statewide list, and to the limited staff available at Ecology to conduct site investigations and assessments, local county health departments are funded by a site hazard assessment grant from Ecology to aid in the site ranking process.

A ranking is conducted on sites found to have levels of hazardous waste above state MTCA limits, which set cleanup levels (2) for residential and industrial soils to protect the air, surface, and/or groundwaters of the state. Sites are ranked according to the *Washington Ranking Method (WARM) Scoring Manual* (3). The score estimates the relative risk to the health of people and the environment from a site relative to other ranked sites in the state. The scores range from 1 (highest relative risk) to 5 (lowest relative risk). Sites found to have levels of hazardous wastes below MTCA cleanup levels, or sites inappropriately listed, receive a designation of No Further Action (NFA), which should remove the site from the list. A site may be ranked for any and/or all of three possible exposure routes—surface water, air, and groundwater—depending on the type of hazardous waste and its relative location in the soils and/or groundwater at the site.

Data Sources

Much of the data required to rank a hazardous waste site are geographic in nature. Locations of wells, parks, fisheries resources, local populations and others in the vicinity of the contaminated site are required to assign the overall ranking factors (see Table 1 for a full list of data sources used for the ranking). Traditionally, data sources required to make these determinations were found in various computer database

Table 1 Data Sources for Site Hazard Assessment; King County, WA

Data Sources for Site Hazard Assessment (SHA)	Information Needed for SHA	SHA Routes
Washington State		
Model Toxics Control Act (MTCA): cleanup levels, risk calculations (CLARK II) update	Levels above which SHA is required	SW, GW, A
Washington Ranking Method (WARM) Scoring Manual	Method used to assign human health and/or environmental risk	SW, GW, A
Toxicological database for use in WARM scoring	Values, risks assigned by compound, chemical, etc.	SW, GW, A
Washington State Department of Health public water supply listing (DWAIN)	Wells located by section, township, range, and # of connections by small and large drinking water systems within 2 miles of site	SW, GW
Washington State Department of Ecology water use data: Water Rights Information System (WRIS)	State water rights issued for surface water and wells by section, township, range for irrigation, industry, drinking water, etc. within 2 miles of site	SW, GW
King County		
Sierra Permits: Health and Building Department permit system	Activities of Health and Building Departments related to permits, complaints, etc.	SW, GW, A
Sensitive area map folio for King County	Nearest wetland, stream, floodplain, fisheries resource	SW, GW, A
Situs: Assessor's Office records	Parcel-related information: address, owner's name, parcel size, etc.	SW, GW, A
National		
US Department of Agriculture, Soil Conservation Service, WA Agricultural Station, King County	Surficial soil types listed for western half of county, not including city of Seattle	SW, GW, A
National Weather Service data, WA climate for King County (WSU, College of Agriculture, Cooperative Extension Service)	Precipitation: total annual and November through April (minus evapotranspiration)	SW, GW
National Oceanic and Atmospheric Administration: isopluvials of 2-yr., 24-hr. precipitation, *NOAA Atlas 2, vol. IX*	Maximum precipitation in tenths of an inch	SW
National census data	Population within half-mile radius	A
Other		
Thomas Brothers map	Estimated distance to nearest parks, streams, etc.	SW, GW, A
Various sewer, water company information	Sewer and water service to site, presence of combined sewers for stormwater drainage	SW, GW, A

SW = Surface water
GW = Groundwater
A = Air

printouts, printed lists, and paper maps (Table 2). The process of manually teasing the required elements from these sources was time-consuming, repetitious, tedious, and potentially prone to error.

Table 2 Data Sources for Site Hazard Assessment Prior to GIS Implementation; King County, WA

Data Sources Prior to GIS Implementation	Type of Data	Last Update	Problems Keeping Data Current
Washington State			
Model Toxics Control Act (MTCA): cleanup levels, risk calculations (CLARK II) update	Printed lists, regulations	1996	Updated by state
Washington Ranking Method (WARM) Scoring Manual	Printed document—some data sources	1992	Updated by state
Toxicological database for use in WARM scoring	Printed lists, tables	1992	Updated by state
Washington State Department of Health Public water supply listing (DWAIN)	Computer printout	1994	Parcel #s not included, sources are estimated on GIS; state has a new data base now
Washington State Department of Ecology water use data: Water Rights Information System (WRIS)	Computer printout	1989	Parcel #s not included, sources are estimated on GIS; state has a new data base now (WRATS)
King County			
Sierra Permits: Health and Building Department permit system	Computer permit system (tied to Situs)	Current	New permit system being installed, current one not Y2K compliant
Sensitive area map folio for King County	7 types of sensitive areas—14 maps each	1990	Updated by county; existing maps did not include drainage basin boundaries (needed for surface water route)
Situs: Assessor's Office records	Computer data system (tied to Sierra permits)	Current	Updated by Assessor's office
National			
US Department of Agriculture, Soil Conservation Service, WA Agricultural Station, King County	20 separate maps (not including City of Seattle)	1973	Entire county not on map
National Weather Service data, WA climate for King County (WSU, College of Agriculture, Cooperative Extension Service)	Map showing weather stations, associated precipitation tables	1931–1965 data	Data not based on enough points to show differences due to slopes, valleys, or other changes in geoposition
National Oceanic and Atmospheric Administration: isopluvials of 2-yr., 24-hr. precipitation, *NOAA Atlas 2, vol. IX*	Map of WA state	1970?	NOAA working on more accurate data at this time
National census data	From EPA internet site	1990	Method using 1/4 of population within a one-mile radius of site typically underestimates true population due to Puget Sound, lakes, and other non populated areas falling in sample area
Other			
Thomas Brothers map	45 maps	1998	May have to work on maps from two different pages at once
Various sewer, water company information	Must call each purveyor		May require all utilities to go to GIS, unknown when this will happen

For example, state well locations were printed onto a large stack of computer paper with locations listed by section, township, and range. To locate the wells and the population served within the WARM model, surface water and groundwater routes within a two-mile radius needed to be identified. First a two-mile radius circle was drawn by hand onto a printed diagram of representative sections, townships, and ranges. Then the sections within the circle were listed on a sheet of paper. The wells were found by manually going through two separate printouts, one for the Group A wells (large water systems down to nine connections) and one for the Group B wells (smaller water systems down to two connections). The nearest well to the site would have to be located by address and its distance to the site estimated using a published street guide or similar map. Due to the rough method employed, some wells outside the two-mile radius were inadvertently included, and some wells within the two-mile radius were excluded. Another problem with the dataset used was that it was last updated in 1994. An updated report was not available with any changes or updates to the well list. In fact the computer system that produced the report was no longer available because the state had already upgraded to a new database system.

Data Accumulation Time

The time required for drawing the maps, finding the wells, and writing the lists took anywhere from about 20 to 30 minutes each. That did not include the extra time spent finding a lost printout on a co-worker's desk, various other interruptions, and/or problems due to starting with the wrong information.

The time to complete all required data collection for each site ranking was about one-and-a-half to two hours (Table 3). Each site required similar repetitive tasks, although not all sites were ranked, and, if ranked, some routes were not evaluated (some sites are only ranked on the groundwater route, for example). When considering the number of sites needing evaluation by the SHA program each year (40 or more completed each year at current staffing levels, with a backlog of 275 sites), a significant time-savings could be achieved using GIS to compile, store, and view the data.

Current GIS Program

GIS Simulation

A demonstration of the current GIS shows the ease and quickness with which SHA-required data can be compiled, evaluated, and presented. Once in the system, clicking on the ArcView icon automatically opens a DDES-designed project. This project includes a parcel locator button that automatically can zoom to the site to be ranked once the parcel number, address, owner's name, or permit activity number has been entered. Once the site is chosen, parameters—such as Group B wells (systems serving 2–10 connections), drainage basin name, surface soil type, isopluvial level (a two-year, 24-hour period maximum rainfall), census blocks including population, parks, fisheries resources, floodplains, and sensitive areas themes—can be layered onto the view.

To estimate the population within a half-mile radius, for example, a Select By Theme operation can be performed. The first step is to choose the parcel to be assessed by clicking on it, or by finding it with the parcel locator button. With the Census Theme chosen as the active theme, Select By Theme can be chosen from the Theme pull-down

Table 3 Site Hazard Assessment Data Source Time Study; King County, WA

Data Source	Accumulation Time	Installed on GIS?
Washington State		
Washington State Department of Health public water supply listing (DWAIN)	Draw map = 10 min; look through printout = 10–20 min	yes
Washington State Department of Ecology water use data: Water Rights Information System (WRIS)	Draw map = 10 min; look through printout = 15–25 min	no
King County		
Sierra Permits: Health and Building Department permit system	Open program, get info from address = 4–5 min	yes
Sensitive area map folio for King County	Check all maps = 10–20 min	yes
Situs: Assessor's Office records	Works with Sierra (see above)	yes
National		
US Department of Agriculture, Soil Conservation Service, WA Agricultural Station, King County	Find site on maps, check soil type = 10 min	yes
National Weather Service data, WA climate for King County	Check map, data table = 1–2 min	no
Isopluvials of 2-yr., 24-hr. precipitation; *NOAA Atlas 2, vol. IX*	Check map = 1–2 min	yes
National census data	Contact Web site, request map = 10–15 min; wait 2 hours to overnight for map completion	yes
Other		
Thomas Brothers map	Hand measure for distance = 2–3 min	yes
Various sewer, water company information	Phone calls, may take several calls to get proper info	no

menu. A message box opens and into the first entry box, again using the pull-down menu located there, "Are within distance of" is chosen. In the second box, Parcels is chosen and, then, in the third box the desired distance can be chosen (2,000 feet for this example). After clicking on the New Set button, ArcView sets to work. When finished, the census blocks within about a half-mile will have been highlighted. By opening the Theme Table, clicking the Promote button, clicking on the Population field heading, and then choosing Statistics from the Field pull-down menu, the sum of the population in the chosen census block set can be produced.

Note: We are using census blocks for this calculation.

Time Comparison

Through the use of ARC/INFO and ArcView, multiple databases can be accessed instantaneously by controlling the parameters needed. The actual time required to rank the example site using GIS was clocked at about 20 minutes. In comparison, the time required to rank the example site using traditional methods was between one to two hours. The time saved is in accumulating the required data to perform an SHA ranking, not to mention the fact that the GIS uses the most current data available. Using GIS also saves time previously spent looking for missing printouts, waiting for census maps to

be drawn by the EPA Web site, and other miscellaneous time spent searching the office for the various forms, paper maps, datasets, and so on. However, all of the required data needed for a full SHA had not been added to the GIS at the time this paper was written. Parameters for water rights used to determine nearest surface water uses (for drinking and irrigation uses), Group A wells (wells serving populations of 10 or more connections), private wells, total precipitation, and evapotranspiration totals still need to be entered onto the GIS to be of use for ranking purposes.

Conclusions

Establish Data Linkages

Use of GIS within EH provides an opportunity to establish linkages with GIS programs already in existence. Much of the initial legwork associated with starting a GIS can be avoided by working cooperatively with other established agency GIS. Development of linkages is a wise use of resources wherein each agency develops databases specific to their needs and shares these data to eliminate duplication of effort. All agencies can make their respective decisions based on the best, updated, and most comprehensive information available.

The ongoing tasks of updating and installing new data sources must be recognized as a priority in the move to a fully integrated GIS. This is a necessary commitment of each program and agency, as poor data give inaccurate results and good data accurate results. All users of the GIS must work to integrate and upgrade their own data. Along with enjoying the availability of all of the county's data comes the responsibility to share Health Department data with others in the county, as well as passing along any changes and/or upgrades as required.

Health-Based Decision Making

Inclusion of health data in a countywide GIS provides an opportunity for health information to be considered as a factor in broader decisions made within the county. Due to the very nature of environmental health programs, a wide variety of data is routinely collected. Use of these data in a GIS may facilitate agency and community access to health information.

Community Assessment

GIS is an important tool to help in the community assessment process. It may be used to collect, store, analyze, and communicate public health and other information to the public. As the environmental health field becomes more community assessment oriented, local agencies are exploring new ways to use their data to identify areas of need and improve public access to information. GIS provides a way to accomplish these needs by capitalizing on spatial elements inherent in data routinely collected. Application of GIS within an environmental health agency can be a huge undertaking and seem unrealistic for many local governments. The benefits, however, of using GIS to administer routine environmental health functions can include overall department effectiveness and efficiency.

Implementation of GIS requires an identification of needs and future uses, commitment from management, and room in the budget to cover hardware, software,

training, data input, data updating, and GIS-dedicated technical staff. There is a need for at least one staff member to concentrate only on GIS data management. Trying to keep the GIS progressing is nearly impossible while trying to keep a full-time position workload going. Although this initial investment may seem overwhelming, the benefits to the community as a whole, with the ability to map data geographically, will reward the department on an ongoing basis. The key is sharing data with and between other city, county, state, and federal agencies. By exchanging data with other agencies and using the extensive GIS capabilities already developed by the county GIS through DDES, the community served has gained a valuable assessment tool. The implementation of GIS has added effectiveness, efficiency, and accuracy in an affordable and sustainable way.

References

1. Washington Department of Health. 1994. *Public health improvement plan.* Seattle, WA: Washington Department of Health. November 29.
2. Washington State Department of Ecology. 1996. *The Model Toxics Control Act cleanup regulation.* Chapter 173-340, Washington Administrative Code. Olympia, WA: Washington State Department of Ecology Publications Distribution Office. Publication 94-06. January.
3. Washington State Department of Ecology. 1992. *Washington Ranking Method (WARM) scoring manual.* Prepared by Science Applications International Corporation, Olympia WA, and Parametrix, Inc., Bellevue, WA. Publication 90-14. April.

Disease Cluster Investigation and GIS: A New Paradigm?

Geoffrey M Jacquez, MS, PhD*
BioMedware, Ann Arbor, MI

Abstract

Advances in geographic information system (GIS) and database technologies are introducing a new era of disease control and surveillance. GIS has proven "value added" for targeting public health interventions, identifying study cohorts, mapping disease patterns, and assessing exposures. Nonetheless, it is not entirely clear whether GIS can advance epidemiological science by increasing our understanding of disease etiology. As an enabling technology, the microscope was key in elucidating relationships between pathogens and disease, and made possible fundamental public health advances such as the eradication of smallpox. Does GIS hold equal promise? Can GIS mislead as well as inform us? Can we formulate and test epidemiological hypotheses using GIS? And if we can, what role do disease clustering and other pattern-recognition techniques play? This presentation attempts to place GIS and disease clustering techniques within the context of a systematic approach for formulating and testing epidemiological hypotheses. The elucidation of relationships between disease processes and patterns is identified as an important direction for future research.

Keywords: disease clustering, health surveillance

Introduction

Being one of the last speakers affords me a chance to reflect on the talks and discussions of the last few days. What impresses me the most is how much progress has been made. Three or four years ago many of us were grappling with our first geographic information system (GIS) applications; simply creating a map of exposures and health events justified a presentation. Here I've attended talks that far exceed these tentative first steps. Topics representative of how far and fast we have come include spatial Monte Carlo randomization methods for assessing the significance of spatial patterns, Web-based GIS for dealing with data concurrency and data sharing, and integrated GIS systems for health surveillance and decision-making, to name a few. Indeed, we have come far, but there are flies in the ointment.

Perhaps our biggest weakness is that GIS technology leads the science, and at such a basic level that it determines the very questions we ask of our data. As public health professionals we all know that time is a critical component in all epidemiological processes. Exposure must precede disease outcomes; transmission events require contact in time as well as in space; every disease has a latency period; and so on. Yet time was given little attention in the presentations I've seen at this conference. Why? Because GIS technology leads the science, and time-GIS is not yet commercially available. I think our inability to conduct true space-time queries is one of the greatest

*Geoffrey Jacquez, BioMedware, 516 North State St., Ann Arbor, MI 48104 USA; (p) 734-913-1098; (f) 734-913-2201; E-mail: Jacquez@Biomedware.com

technological deficiencies limiting GIS in public health. It won't be solved until true time-GIS are available. There are other examples of GIS technology leading the science, but I won't go in to them now. What we need is for public health as a science to lead the technology. This will require thinking "outside the box" on our part to identify those epidemiologically valuable functions that are absent from GIS, and incorporation of our input in software development. I believe this is one of the key problems limiting advances in GIS in public health.

Standing here I feel like the Pope preaching to the choir. This conference is attended only by the converted—if you believed GIS were humbug would you be here? Probably not. Is there anyone here who thinks GIS is humbug?[1] Being surrounded by like thinkers can be dangerous. Allow me to play the devil's advocate as I offer some point and counterpoint on statements and observations made over the last few days.

A Dialog with the Devil's Advocate

One of the observations made at this conference is that "all data are spatial," and I think most of us would agree. However, our devil's advocate is a classical epidemiologist, with little or no training in spatial thought. Her counterpoint is, "So what—location is a lousy exposure surrogate." This counterpoint is difficult to parry when we acknowledge that exposure is best measured at the level of the individual.

In the opening plenary session, one of the speakers observed that "the power of GIS is limited only by your imagination," and most of us nodded in agreement. The counterpoint from the devil's advocate is that "it is the expense of GIS, and not its power, that is beyond imagination." And in fact, we all know that establishing a GIS and its data is resource-intensive.

My point is that as GIS enthusiasts we tend not to hear the counterpoint from the devil's advocate. To illustrate: consider two quotes describing different visions of GIS. The first, from David Gerlernter, sees GIS as a powerful representation of our spatial world, depicting the complexity of the ever-changing space in which we live:

> Someday soon you will look into a computer screen and see reality. Some part of your world—the town you live in, the company you work for, your school system, the city hospital—will hang there in a sharp color image, abstract but recognizable, moving subtly in a thousand places. This Mirror World you are looking at is fed by a steady rush of new data pouring in through cables. It is infiltrated by your own software creatures, doing your own business. (1)

This vision is the logical extension of advances now being made in GIS, including self-organizing maps, Web-based GIS, real-time acquisition of Global Positioning System (GPS) data, and open standards allowing access to diverse databases with ready incorporation of "software creatures." In short, Gerlernter envisions GIS as a powerful, enabling technology whose potential in public health is vast and far-reaching. This is consistent with the vision Jack Dangermond presented at lunch yesterday. Contrast this with a second, briefer quote from Marbury, who focuses specifically on the value of GIS in health:

[1] When asked this question, only two in the audience raised their hands.

> For the most part, advances in environmental epidemiology will require carefully designed studies of rigorously defined outcomes combined with good measurements of personal exposure. It would be a shame to be distracted from this effort by the availability of a new tool that affords no new insights. (2)

Marbury recognizes the conundrum facing public health workers: when deciding to undertake one activity, we necessarily commit resources that might have been better spent elsewhere. Such opportunity costs can be substantial for health GIS. Questions such as "Is it wise to spend our health dollars on GIS when we could be vaccinating children?" are powerful illustrations, but of course apply to all public health activities, not just GIS. Here, Marbury is concerned with the opportunity cost of GIS as an epidemiological tool.

Ken Rothman (3) posed similar concerns regarding disease cluster investigations. He observed that cluster investigations usually lead to negative results, are prone to pre-selection bias (the well-known "Texas sharpshooter problem"), and compete for scarce public health resources. These issues become increasingly important as advances in interoperability and data acquisition make integrated health surveillance systems a reality. Health surveillance systems combine GIS and disease clustering software, and raise the possibility of real-time proactive disease clustering.[2] Thus the question of opportunity costs is destined to become even more pressing: is GIS a useful epidemiological tool, drawing on the technological cornucopia envisioned by Gerlernter, or is it simply a convenient way of making maps, one whose applications are ultimately limited? In particular, can the combination of GIS and disease cluster statistics increase our understanding of disease etiology? Or are health surveillance systems technologic flashes in the pan that contribute little to our understanding of human disease?

A Vision of GIS in Public Health

My vision of GIS is of an enabling technology that may lead to fundamental advances in our understanding of relationships between the environment and human health (see reference 4 for more details of this vision). The approach incorporates disease cluster statistics and other tests for spatial patterns, with the objective of generating and testing epidemiological hypotheses. This paradigm is evolving, and its potential is best understood using the water drop lens as an historical analog.

In the 1600s Anton Van Leeuwenhoek glimpsed the first images of microscopic organisms using a water drop lens. These "animalcules" were a curiosity, and no one suspected their role in infection and disease. Improvements in technology led to the compound microscope, which in the 1800s enabled Pasteur and his colleagues to reveal the link between bacteria and infection. This set the stage for major public health successes such as the eradication of smallpox. But it was the application of the technology in the context of a systematic approach that made scientific advances possible.

This analogy suggests that the promise of GIS in public health will not be realized until the technology is applied using a systematic approach such as that proposed by

[2] Reactive clustering responds to possible clusters brought forward by concerned citizens, and hence is prone to pre-selection bias. Proactive clustering surveys health event data as they are collected to identify emerging clusters.

Karl Popper. Using Popper's scientific method, a theory is inferred from observed data and falsifiable predictions are deduced from that theory. The predictions are then evaluated by experiment. When the prediction is falsified the theory is rejected. Theories may be rejected, but not proven, and predictions must be falsifiable by experiment or other means.

A related approach called "strong inference" (5) recognizes that a researcher's knowledge changes as a study progresses. Based on her/his current knowledge of the system, the researcher first formulates a set of alternative hypotheses that could explain the observed data. Systematic experiments are then designed and executed in order to exclude false hypotheses, leaving the remaining hypotheses as the only plausible explanations. During this process the researcher's knowledge base changes, and the set of alternative hypotheses may change too. Strong inference is thus a systematic approach for evaluating hypotheses in an iterative fashion.

Although they are appropriate models for laboratory studies, these systematic approaches are not directly applicable to GIS studies, since they rely on designed experiments. Spatial data typically are observational, and the processes under study often occur on a long time span that precludes experimentation. In addition, spatial systems are usually large and difficult to manipulate. This magnifies, rather than diminishes, the need for a careful and systematic approach. Despite this, we still lack a systematic approach to the application of GIS in public health. As Jacquez (4) pointed out, many health studies are prone to the "Gee Whiz" effect. This is a leap of unsupported inference that begins with the construction of thematic maps. This cartographic exercise is undertaken to visualize spatial patterns—in fact, a dramatic pattern is an important map selection criterion (why present colleagues with a map that doesn't illustrate one's point?). We are then tempted to formulate hypotheses to explain the perceived pattern. The "Gee Whiz" fallacy results: we formulate hypotheses to explain map patterns whose existence has not been demonstrated. Because maps are selected based solely on visual impact, we accept patterns without first demonstrating that they are statistically unusual; finally, hypotheses are formulated to explain patterns that may not even exist.

All of these problems can be ameliorated by making GIS part of a systematic approach that visualizes a spatial disease pattern, evaluates that pattern's statistical significance, and then generates falsifiable hypotheses that might explain the disease processes giving rise to that pattern. Building on the work of Jacquez (see Figure 2.5 in reference 4), the GeoMed project, being conducted by BioMedware and the University of Michigan, is producing a new paradigm for the analysis of spatial disease data (Figure 1). This paradigm is the result of a joint effort by Doctors Leah Estberg, Geoffrey Jacquez, Andy Long, and Mark Wilson, and is detailed in a soon-to-appear joint publication (6).

The boxes in Figure 1 labeled "Disease Data" and "Contextual Data" represent a study's data and setting. Disease data may be locations of cases and controls, disease rates, or case counts. These may or may not have been standardized, and the cases themselves may or may not have been verified, depending on the study's *context.* Contextual data define the study's setting, as do data on the environment, covariates, and confounders. The study's setting includes personnel, institutional, administrative, political, public relations, and other factors that influence how the problem is defined, how the data are collected, how the analysis is conducted, how the results are interpreted, and how interventions are selected and executed.

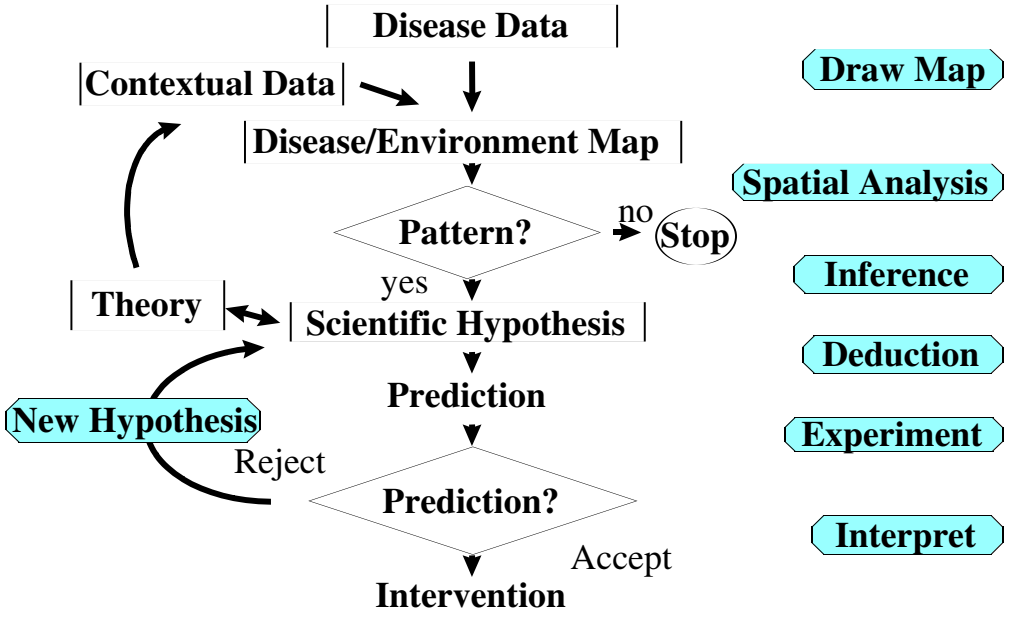


Figure 1 Spatial analysis in public health. A systematic approach for the analysis of spatial disease data.

Disease and environmental data, and information on covariates and confounders, are entered into the GIS, which is then used to construct a thematic map ("Disease/Environment Map," Figure 1) using cartographic functions such as Boolean operations, buffering, interpolation (e.g., kriging), and related techniques ("Draw Map"). The process of drawing a map is iterative and may involve several cycles through collecting data, preparing them for visualization, specifying cartographic parameters, and drawing the map. These first steps leading to map generation can be thought of as the observations in the Popperian paradigm.

Once the map is completed, the process of spatial statistical analysis begins. This process determines whether the spatial disease patterns are statistically unusual or are best explained as a chance aggregation ("Spatial Analysis"). The first step is visual inspection of the map to identify possible patterns to be statistically analyzed. Patterns of interest typically include clusters of health events and spatial associations between disease patterns and environmental variables. Both of these questions must be evaluated against the spatial fabric of disease correlates and confounders. For example, spatial clustering must be evaluated relative to the geographic distribution of the at-risk population, because population density varies from place to place. This is where disease cluster statistics and methods of spatial analysis come into play. Useful techniques include disease clustering methods, methods for analyzing spatial point distributions, adjacency statistics for determining whether classes of areas share common borders, tests for boundary overlap, statistics for evaluating association between two or more spatial variables (7–10), and related techniques for analyzing spatial data. These methods

allow us to determine whether the perceived map pattern is statistically unusual and thus warrants further investigation.

Only then can we justify inferring a theory or hypothesis to explain the spatial relationships, and proceed to the next stage ("Inference"). If the disease pattern is not significant we stop the analysis ("Stop"). This process of map generation and spatial analysis is a form of exploratory spatial data analysis, rather than a more formal approach of statistical inference. Different aspects of a spatial pattern can be explored and, hence, different statistical tests can be applied to the data. This raises issues of multiple testing, and an experiment-wise error approach, or the Bonferroni, Simes, or Holms corrections, may be needed. These techniques adjust p-values to account for repeated tests.

The decision process ("Pattern?") is based on the researcher's knowledge of the disease data and system under study (the contextual data), and *does not* proceed from the statistical results alone. It is of course possible to have significant p-values for disease patterns that are not of public health interest, as occasionally arises for cases of unrelated diseases that lack a plausible common cause or etiology. Similarly, disease patterns are occasionally found to be "not significant" even when there are compelling reasons for proceeding with the analysis.

When the map is deemed significant (e.g., when there is meaningful spatial clustering of cases above and beyond the geographic variation in density of the at-risk population), the researcher formulates a hypothesis or set of hypotheses to explain the spatial disease pattern ("Scientific Hypothesis"). The set of hypotheses may correspond to a larger body of knowledge ("Theory") describing biological mechanisms of disease causation, progression, and propagation. This body of theory contributes to the context in which the study is conducted. As hypotheses are evaluated and rejected, the underlying theory may change, giving rise to new hypotheses. This occurrence is indicated by the double-headed arrow between "Scientific Hypothesis" and "Theory."

Hypotheses in themselves are general statements that are not directly testable; a falsifiable prediction must be deduced from the hypothesis. Our next step, therefore, is to formulate a testable prediction, and design an experiment to test that prediction. As with the strong inference and Popperian approaches, we have the power only to falsify predictions and their corresponding hypotheses.

At least three kinds of experiments seem possible ("Experiment"). We may design an epidemiological study to test a prediction describing disease occurrence in populations. A laboratory study may be designed if the prediction describes disease progression at the organismic level. Finally, another GIS study may be used to evaluate epidemiological predictions that involve a spatial dimension. Notice that the GIS data used to formulate hypotheses cannot be used to test predictions that emerge from those hypotheses. To do so would bias one toward confirming the observed pattern.

Rejection of the prediction may give rise to new hypotheses, with corresponding changes in theory. Acceptance of the prediction may necessitate decisions and actions based on the experimental results. For example, a finding of a significant disease cluster, with a plausible environmental exposure as demonstrated by experiment, may warrant intervention to reduce exposure and to treat the affected population ("Intervention").

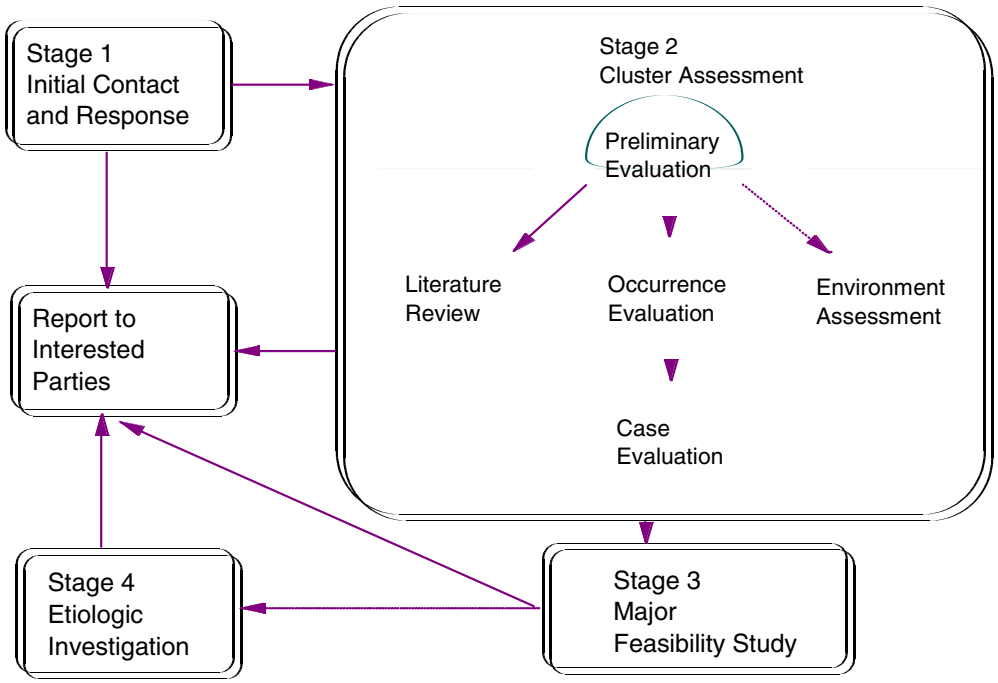


Figure 2 Disease cluster investigation protocol of the Centers for Disease Control. Modified from Centers for Disease Control, 1990 (9).

The Centers for Disease Control and Prevention's Disease Clustering Protocol

How does the schema described above relate to the Centers for Disease Control and Prevention (CDC) guidelines for investigating disease clusters? The CDC guidelines (11) advocate a four-step approach consisting of initial response, cluster assessment, major feasibility study, and finally an etiologic investigation to determine the cause of the disease cluster (Figure 2). The purpose of this protocol is to provide a systematic response to cluster allegations, to maintain good relationships with the community, and finally to conserve public health resources.

Typically, alleged clusters are brought forward by concerned citizens during stage 1, initial contact and response, during which available case data are collected. Disease cluster statistics are used in stage 2, the cluster assessment stage, primarily to determine whether a significant spatial pattern exists. If it does, more resources may be allocated for a feasibility study and etiologic investigation; if not, the investigation may be terminated. In general, disease cluster investigations are reactive and testable hypotheses are not formulated until stage 3, the major feasibility study. Only if the feasibility study is successful will an etiologic investigation take place.

Both the GIS scientific method presented earlier and the CDC guidelines use disease cluster statistics to determine whether the perceived pattern is in some sense unusual and deserving of further explanation. The CDC guidelines are thus a special case of the general protocol for spatial analysis in public health given in Figure 1.

Conclusion

Areas in which GIS technology has made substantial and continuing contributions include exposure assessment, identifying study populations, constructing disease maps and atlases, and disease surveillance to identify the locations of possible outbreaks. Other areas include the geographic placement of health services. Many of these activities yield valuable results without passing through the entire flowchart shown in Figure 1. However, whenever decisions must be made, resources must be allocated, and whenever interventions are needed, this protocol of spatial analysis in public health should be followed. In addition, a systematic approach such as that in Figure 1 must be followed if we are to advance spatial epidemiology as a scientific field by evaluating spatio-epidemiologic theories of disease spread and causation.

The approach in Figure 1 is under development in a collaborative research arrangement between BioMedware and the University of Michigan and is subject to modification. The most pressing need is an improved understanding of the relationships between space-time disease processes and the spatial disease patterns they produce. We do not expect to find a one-to-one mapping of disease process to disease pattern. However, given an observed spatial disease pattern, we do hope to be able to exclude certain disease processes as causal explanations. The 1990s experienced rapid growth in methods of spatial analysis in general and of disease cluster statistics in particular. Our arsenal of spatial analysis tools is robust and will continue to expand. In contrast, our understanding of the relationships between human diseases and their resulting spatial disease patterns is woefully inadequate. The elucidation of these relationships is the salient research need in spatial health analysis.

Acknowledgments

BioMedware and the University of Michigan, with funding from the National Cancer Institute, are preparing Web-based course materials for teaching spatial epidemiology. The graduate-level course "Spatial Epidemiology," offered at the University of Michigan School of Public Health in 1999 and 2000, has the objective of teaching the elucidation of disease processes from spatial disease patterns. The author thanks Doctors Leah Estberg, Andy Long, and Mark Wilson, who are working with the author on that project, for lively discussions on the evolving paradigm of public health surveillance. This research was funded by grants CA65366 and CA64979 from the National Cancer Institute. The views stated in this publication are those of the author, and do not necessarily reflect the perspectives of the National Cancer Institute.

References

1. Gerlernter DH. 1991. Mirror worlds: Or the day software puts the universe in a shoebox . . . how it will happen and what it will mean. Oxford University Press.
2. Marbury M. 1996. GIS: New tool or new toy? *Health and Environment Digest* 9:88–9.
3. Rothman KJ. 1990. A sobering start for the cluster buster's conference. *American Journal of Epidemiology* 132(Supplement No. 1):S6–13.
4. Jacquez, GM. 1998. GIS as an enabling technology. In: *GIS and health.* Ed. A Gattrell, M Loytonen. London: Taylor & Francis. 17–28.

5. Platt JR. 1964. Strong inference. *Science* 146:347–53.
6. Jacquez GM, Estberg L, Long A, Wilson ML. *Project GeoMed: Software and educational modules for spatial analysis in epidemiology.* Presented at the International Conference on the Analysis and Interpretation of Disease Clusters and Ecological Studies. December 16–17, 1999. Conference proceedings to appear in *Journal of the Royal Statistical Society.*
7. Jacquez GM, Waller LA, Grimson R, Wartenberg D. 1996. The analysis of disease clusters, Part I: State of the art. *Infection Control and Hospital Epidemiology* 17:319–27.
8. Jacquez GM, Grimson R, Waller LA, Wartenberg D. 1996. The analysis of disease clusters, Part II: Introduction to techniques. *Infection Control and Hospital Epidemiology* 17:385–97.
9. Haining R. 1998. Spatial statistics and the analysis of health data. In: *GIS and health.* Ed. A Gatrell, M Loytonen. London: Taylor & Francis. 29–48.
10. Kulldorff M. 1998. Selection of statistical methods for the analysis of spatial health data. *GIS and Health.* Ed. A Gatrell, M Loytonen. London: Taylor and Francis. 49–62.
11. Centers for Disease Control. 1990. Guidelines for investigating clusters of health events. *Morbidity and Mortality Weekly Report* 39:1–23.

Perception and Reality: GIS in Environmental Justice through Pollution Prevention

Marion C Kinkade, Jr, MCRP (Cand)*
GIS Coordinator, Lincoln-Lancaster County Health Department, Lincoln, NE

Abstract

Geographical information system (GIS) technology is an increasingly important tool of assessment and technical support for the Lincoln-Lancaster County (Nebraska) Health Department (LLCHD). It is a natural extension of the department's assessment functions and a profound area of innovation in public health information systems that impacts all LLCHD divisions. LLCHD is utilizing public health applications of GIS to enhance community health assessment and disease surveillance, environmental risk assessment, policy development, program evaluation and planning, decision support, public education, and health threat/event response. GIS has many valuable environmental justice applications. Using a recently completed survey titled *Environmental Health Hazard Risks in the Minority Community*, LLCHD used ArcView software to analyze the perceptions of minority populations regarding environmental risks in general and compared them with actual risks believed to be in their home or yard, neighborhood, and work or school. The department mapped these responses and then compared them with known environmental factors/risks from business hazardous chemical inventory (Tier 2 sites), air pollution point sources (Title V sites), Special Waste sites, age and condition of housing in the area (to determine the potential for asbestos or lead-based paint), and the water system (to show where the water comes from and how the system works). The results of this analysis depict the relationships between minority community perceptions and known environmental factors or events. This information enables LLCHD to target specific areas for educational programs, provide a measure of need for community education and risk prevention, better allocate public health resources, assess the effectiveness of community programs, and document public health impacts of land use planning options.

Keywords: minority, environmental justice, community perceptions

Introduction

The concept of environmental justice is relatively new, but the fact that certain neighborhoods bear more than their share of environmental hazards is not. The introduction from the *Environmental Health Hazard Risks in the Minority Community Survey* (1) commissioned by the Lincoln-Lancaster County (Nebraska) Health Department (LLCHD) introduces the concept of environmental justice very well:

> The environmental movement has long been concerned with environmental health hazards—the health risks associated with environmental hazards. Environmental health hazards are substances that have been linked to particu-

* Marion C Kinkade Jr, Lincoln-Lancaster County Health Department, 3140 N St., Lincoln, NE 68510 USA; (p) 402-441-6246; (f) 402-441-8323; E-mail: ckinkade@ci.lincoln.ne.us

lar adverse health effects. Research has generated a great deal of agreement on the health risks posed by materials such as asbestos, lead, PCB's, agricultural chemicals, and smoke stack emissions.

It has been relatively recent that research has begun to focus on the level of risk that different segments of society face. As early as the turn of the century, people recognized that many urban health problems stemmed from the degraded environmental conditions found in the city—particularly the inner city—and that the poor bore a disproportionate share of that burden (Gottlieb 1993). Minority neighborhoods sustained an even greater share of the burden, but it was not until the 1970's that evidence was collected that revealed the correlation between poverty, race, and pollution (Kruvant 1975).

Research has shown the disproportionate risk the minority community bears. While some ethnic communities can be identified with specific health risks associated with traditional employment patterns, such as Hispanics and agricultural chemicals (Traux 1990), others can be identified with geographic distribution that targets minority communities, as in African-Americans and landfill sites (Bullard 1990). In addition, many health risks are associated with poverty and urban decay. Minority communities have historically been located in the poorer and more run down neighborhoods of cities, with their proportionately higher health risks associated with auto emissions, lead paint, and hazardous waste (Gottlieb 1993).

There is a need to understand which came first, the hazard or the minority population (Been 1994). Bullard lays the blame for environmental injustice on racism: "[E]nvironmental inequities do not result solely from differences in social class….[R]ace interpenetrates class and creates special health and environmental vulnerabilities. People of color are exposed to greater environmental hazards in their neighborhoods and on the job than are their white counterparts (Bullard 1993)." He points to studies that find elevated exposure levels by race, even holding social class constant, with respect to distribution of air pollution (Freeman 1971; Gelobter 1988), location of municipal landfills and incinerators (Bullard 1983, 1987), abandoned toxic-waste dumps (UCCCRJ 1987; Mohai and Bryant 1993), and lead poisoning in children (ATSDR 1988). The Commission for Racial Justice (1987) in a comprehensive national study on the demographic patterns associated with location of hazardous waste sites "found race to be the most important factor (i.e., more important than income, home ownership rate, and property value) in the location of hazardous waste sites."

The validity of these findings and the importance of this issue have been recognized at all levels of government. On February 11, 1994, the White House issued an executive order requiring federal agencies to consider environmental justice in all their actions and requiring the US Environmental Protection Agency (EPA) to gather data and issue new regulations concerning the distribution of environmental hazards (GAO 1995).

The challenge faced by our community health providers is the need for a better understanding of the environmental health risks faced by our communities' minority populations, the source of these risks, and how to best respond to this threat.

So in 1996, LLCHD started a three-year grant project funded by the EPA. During the first year of the grant, LLCHD created a survey instrument to survey members of the non-white population to determine their environmental health knowledge base; their environmental exposures at home, in the neighborhood, or at work or school; and their knowledge, beliefs, and practices related to hazardous materials.

In the second year of the grant, LLCHD used its geographical information system (GIS) to map the potential exposure to these populations from known contaminated sites and permitted release sites. In the third year of the grant, the department will use the information gained through the survey and analysis to educate the affected populations, regulators, and permit holders about how to use the Healthy Homes Program and the Technical Assistance Program to reduce the effects of identified potential exposures.

In the first year of the grant, the Environmental Health Hazard Risks in the Minority Community survey was created and a sample of the minority community was surveyed. The survey enabled LLCHD to identify and compare minority community pollution prevention and health risk awareness, attitudes, and behaviors with earlier baseline data established by the LLCHD Community Pollution Prevention Assessment project and the LLCHD Minority Behavior Risk Factor survey. The goals of the project were to develop an environmental health knowledge base; determine perceived environmental exposures at home and work; determine the knowledge level and identify beliefs and practices related to hazardous materials; and obtain a basic understanding of the pollution prevention ethic held by racial/ethnic minority groups in Lincoln.

Methods

The Environmental Health Hazard Risks in the Minority Community survey is the result of a literature review, a continuous exchange of ideas between the research team and LLCHD staff, and feedback from members of the racial/ethnic minority groups.

The survey was developed in three languages: English, Spanish, and Vietnamese. A large portion of the city's minority population is recent immigrants, and language barriers can pose a serious impediment to gathering data. Spanish and Vietnamese were determined to be the languages understood by a majority of the new immigrants who do not understand English or who speak it poorly. The pre-testing was done in all three languages during the first week of October 1996, and revisions were made based on feedback.

A main training session was held in mid-October 1996 at LLCHD. Overall, 28 surveyors were trained to conduct face-to-face interviews. Over half (61%) of the surveyors were bilingual in English and in either Spanish or Vietnamese. The goal was to complete 500 surveys with at least 100 from each racial/ethnic category. The racial/ethnic categories used were: American Indian, African American, Asians, and Hispanic (which conforms to US Census categories and facilitates the comparison of these data with US Census data and other information).

The sample population to interview was randomly selected by census blocks that contained five or more racial/ethnic minorities. All houses in these chosen blocks were visited and qualifying minority members interviewed. Surveys in the chosen sites were carried out during morning, afternoon, and evening hours, and on both weekdays and weekends.

The initial method, however, proved to be less than adequate in obtaining significant numbers of American Indians (the minimum 100 targeted). Therefore, with the American Indian population the "snowball technique" (requesting persons interviewed to help locate other American Indians) and the cooperation of agencies/organizations that service American Indians (e.g., the Indian Center) were used.

A total of 504 persons 18 years of age and older were surveyed the last week of November 1996. One hundred and eight (108) were Native Americans, 124 were African American, and 136 were "Other." One hundred and thirty-nine (139) were Hispanic, which can be of any race, but which, for the most part, fell into the Other category.

Results

The survey reported that 78.1% of minorities in Lincoln and Lancaster County believe that they are at the same or higher risk from environmental hazards than is the majority population. In addition, 80.2% reported that the pollution from neighboring business is either very harmful or somewhat harmful to their health. Nonetheless, it was difficult to distinguish, based on the survey report, a specific injustice originating from business and industry. The survey respondents reported specific hazards in the home, neighborhood, and at work and school that were associated as much with personal, community, and landlord behaviors as they were with business and industry behaviors.

According to the survey report, the top four environmental risk concerns in the home were contaminated water, garbage, carbon monoxide, and tobacco smoke. The greatest actual environmental risks in the home were reported as cockroaches, garbage, chipping paint, and poor indoor air quality. The top four environmental risks in the neighborhood were garbage, air pollution, contaminated water, and hazardous chemicals. The top four environmental risks at work and school were hazardous chemicals, tobacco smoke, air pollution, and asbestos.

Although the local Minority Advisory Committee (MAC) wholeheartedly agreed to refer minority businesses and businesses in minority communities to the pollution prevention technical assistance program, they also reflected a greater concern for household and community environmental health risks and endorsed approaches to reducing these risks. They clearly did not want an exclusive focus on reducing environmental health risks from business and industry but rather a focus on educating the minorities themselves on environmental risks in the home and community. The MAC agreed to and did propose activities that use the pollution prevention ethic to reduce these environmental health risks.

The survey report also noted some important findings:

- Although there is some indication that those who consider themselves more knowledgeable tend to also believe in the potential harm from identified hazards, it is not a consistent correlation. Those who considered themselves "somewhat knowledgeable," were apparently less knowledgeable than the "not knowledgeable."
- Lack of knowledge among respondents about proper hazardous material disposal sites and location of these sites (along with transportation difficulties in

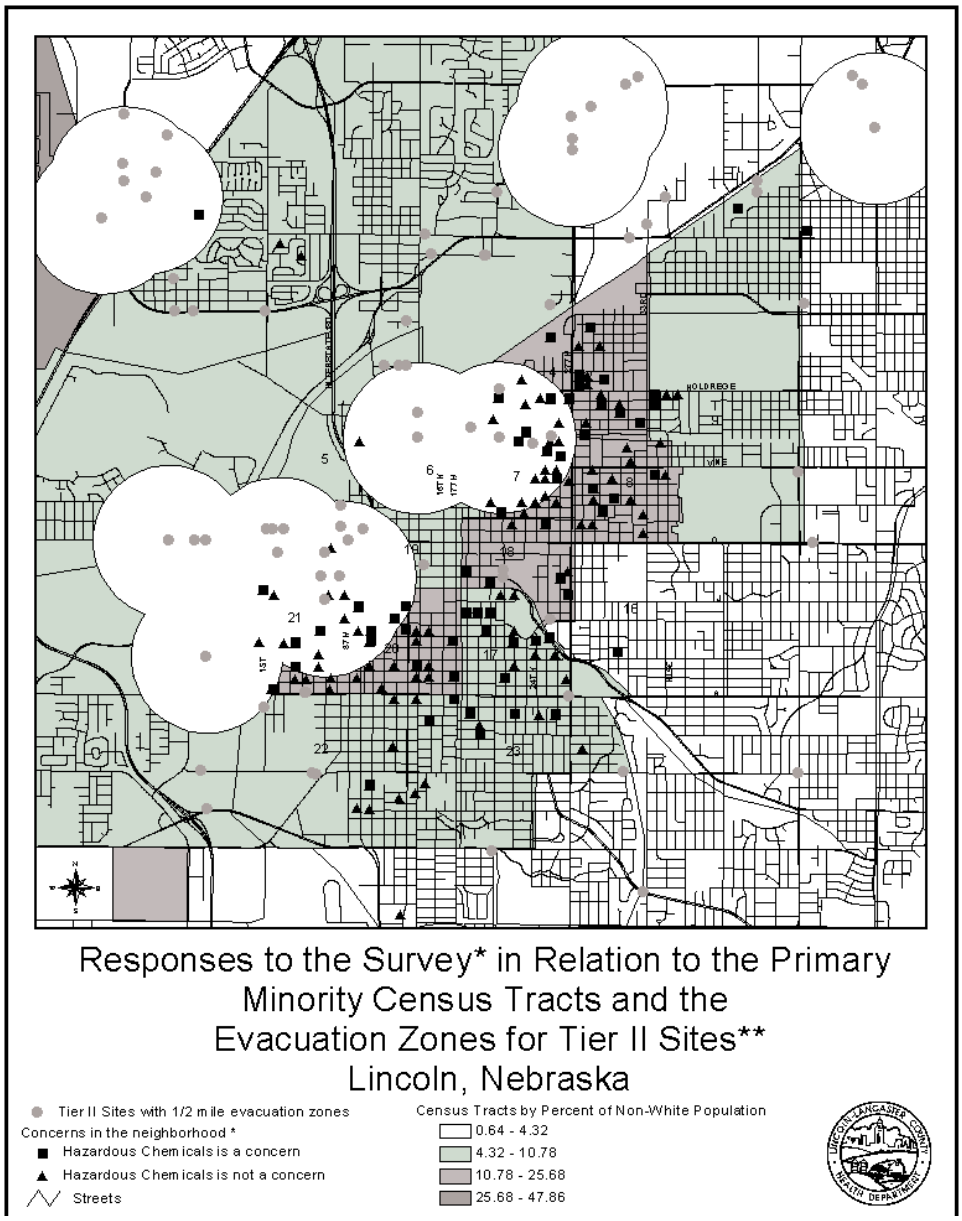


Figure 1 Responses to the minority community survey in relation to the primary minority census tracts and the evacuation zones for Tier II sites, Lincoln, NE, 1996.

getting to them) appears to be a partial but significant component of improper hazardous waste disposal.

- Most identified potentially risky substances as a danger to their health. One out of every four respondents believes asbestos and old car batteries are not

harmful to their health or do not know. About one in four do not believe or do not know that pollution from neighboring business can be harmful to their health.

Discussion

These findings are important to understanding the environmental justice issues in Lincoln and Lancaster County for two reasons. First, the general lack of environmental risk knowledge in the minority community makes specific identification of risk issues difficult. Survey results notwithstanding, at no time in this process has a minority representative suggested to LLCHD that a specific business was presenting excessive risk, nor has anyone indicated that they believe they were at a higher risk than the general community.

Second, the efforts to educate the general community in environmental health and risk identification were not effective in the minority communities. The lack of effectiveness is, in effect, the injustice. The minority communities expressed the need to address basic environmental education issues, specific high priority issues regarding risks in the home and communities, business and industry practices in the neighborhood, and an overall acceptance that pollution prevention would be the guiding set of principles to reduce these risks.

The current process of gaining guidance in these issues from the MAC will continue throughout this proposed grant. Also continuing will be the use of contract funds for the minority community organizations to perform activities that support the goal of using pollution prevention to educate the minority community and reduce environmental health risks.

There were three survey questions concerning environmental hazards that LLCHD concentrated on. The questions asked about people's concern over environmental hazards in their home or yard, neighborhood, and work or school. The environmental hazards we were most concerned about were asbestos, radon, garbage, medical waste, lead paint or dust, hazardous chemicals, contaminated water, air pollution, tobacco smoke, or carbon monoxide. The participants were asked to pick the three about which they were most concerned.

LLCHD was then able to analyze the responses to these questions by race, gender, income level, census block, and age. This technique allowed the department to identify specific areas for educational programs or comparison with known potential environmental hazards.

For example, the Asian community showed a higher than normal number of responses expressing concern about contaminated water. All people in Lincoln drink the same water, so why was one part of the community concerned about the water and the rest not? The conclusion was that many Asians in Lincoln are recent immigrants; because water quality was a concern in their native country, it is a concern here as well. The solution was to take members of the Asian community to the water treatment plant for a tour. This tour allowed them to see how the water was treated and to ask questions.

Applying the Results

The information gathered by census block gave the department the geographical information needed to compare minority perceptions with data on known potential environmental hazards. These data were from Tier II sites (businesses that house hazardous chemicals), Title V sites (businesses that emit air pollutants), and Special Waste sites. Using aerial photos and demographic information, LLCHD was able to analyze the community in a way that could not be done without GIS.

LLCHD used Tier II information to compare the evacuation zones based on chemical types with the residents' concern or lack of concern about hazardous chemicals in their neighborhood as stated in their survey responses. The *North American Emergency Response Guidebook* (2) provided information on the evacuation zones for the various chemicals stored on site at each Tier II facility. This information showed who was potentially in danger if there was a chemical spill. Using symbols to represent responses to the survey question and circular zones for the evacuation zones, the department was able to see who was concerned or not concerned about hazardous chemicals in their neighborhood in relation to Tier II facilities.

With this information, LLCHD was able to identify areas of the community as foci for the Technical Assistance Program (TAP) and the Healthy Homes Program (HHP). These two programs provide services that will make up the activities in the third year of the grant.

Technical Assistance Program

The first approach for reducing environmental risks uses the LLCHD TAP to provide targeted, on-site pollution prevention opportunity assessments to businesses located within minority communities. LLCHD has developed an innovative method for targeting specific businesses in minority communities. The TAP assesses the types of wastes produced by the business, pollution prevention options available for these types of wastes, level of outreach previously conducted, and the relative risk posed to the minority community from the business. The TAP will maintain community involvement in the business selection process by requesting input from minority community residents during MAC meetings. LLCHD believes that community voice is a crucial component in identifying and prioritizing environmental health risks in a community. Once the TAP has developed a list of potential business participants, these businesses will be contacted for appointments to conduct on-site pollution prevention assessments. Arrangements for translation services will be provided if necessary.

The on-site business pollution prevention assessments will serve three purposes. They will:

- Identify pollution prevention and waste reduction opportunities that can reduce risk to minority community residents.
- Further the effort to educate businesses about the pollution prevention waste reduction hierarchy.
- Propose alternative processes and technologies that result in toxicity reduction in minority communities.

To increase the benefit to the community, the TAP will attempt to communicate with businesses that have not participated in community meetings by:

- Using mailings, phone inquiries, information packets, newspapers, trade associations, flyers, and brochures.
- Providing workshop presentations.
- Networking with community or private organizations that interact with the Lancaster County business community.

Healthy Homes Program

The second approach for reducing environmental health risks for minority community residents is via the LLCHD Healthy Homes Program (HHP). This program was established in 1992 as a community integrated service system functioning within LLCHD. Because the program had such a dramatic impact in the minority community, it now receives permanent funding from the Lincoln-Lancaster County Board of Health and the Lincoln City Council.

The HHP improves and/or enhances the health of minority community residents by providing educational outreach services to minority families. This is a one-on-one approach that uses an HHP outreach worker to work with minorities to improve their overall health, the health of their children, and the health of their unborn children. HHP services include:

- Improved access to health care
- Education and parent support
- Healthy behaviors
- Proper nutrition
- Advocacy for families
- Information and referral service
- Early prenatal care
- Exercise
- Dental care and education
- Education and screening for prevention of cardiovascular disease, cancer, and diabetes
- Cultural awareness and opportunities to learn about cultural differences
- Prevention of childhood diseases
  - Immunizations
  - Well child exams
  - Injury prevention

In addition to providing information on the aforementioned benefits, in the third year of the project, the HHP outreach workers will receive training in environmental health hazards and pollution prevention concepts, methodologies, and techniques. This will enable the HHP outreach workers to deliver the pollution prevention ethic directly into minority communities. Teaching minority community residents about pollution prevention will allow them to make more informed decisions about product selection, disposal, housekeeping, and neighborhood standards. For example, teaching them the signal words on products will facilitate choosing less toxic products for use in the home and, therefore, minimize environmental health risks for the homeowner and the community. This knowledge empowers minority community residents to take a day-to-day role in reducing environmental health risks using the pollution prevention ethic. More

importantly, the HHP program encourages minority community residents to become lifelong learners and to become aware of how environment and lifestyle changes affect their overall well being and health.

LLCHD believes that this activity is a logical addition to the third year of the Minority Community Environmental Justice through Pollution Prevention project because the HHP is well established and is highly respected by the minority community. Furthermore, LLCHD believes that sustainability is important in addressing environmental health risks in the minority community. By providing the HHP outreach workers with training in environmental health issues and pollution prevention principles, this program will continue to influence and affect the lives of minority community residents long after the Minority Community Environmental Justice through Pollution Prevention project is finished.

Conclusion

The survey instrument told LLCHD about the knowledge of the minority community in relation to environmental hazards. This information tells not only about the people's knowledge of environmental hazards, but also whether they are being affected or could potentially be affected by environmental hazards in their home or yard, in their neighborhood, at work, or at school. With this information, LLCHD was able to determine whether their concerns were legitimate based on the location of the businesses in the neighborhood, and whether the department needed to visit the businesses in the neighborhood. If their concerns were not legitimate, then LLCHD would need to improve its outreach programs or add an environmental awareness component to the Healthy Homes program, which would educate the community.

GIS is a powerful analytical tool for identifying areas of the community for outreach and technical assistance. GIS gives the ability to analyze demographic information and determine where in the community to concentrate LLCHD outreach efforts. With the demographic information, more specific questions can be asked, like the relationship of industrial businesses to lower income populations and their potential exposure to air pollutants or potential risks from hazardous chemicals. At the same time, this information guides the Technical Assistance Program to work with businesses on pollution prevention. The ability to overlay multiple layers of data, to find patterns in data that the department has previously gathered or obtained from other agencies, gives the department a technique for analyzing the community in a way that was otherwise impossible. Spatially locating businesses, community facilities, responses from surveys (if they have addresses), and demographic information allows LLCHD to identify patterns of disbursement quickly and relatively easily. Also, as modeling techniques become available for use in ArcView GIS (ESRI, Redlands, CA), exposures and studies for specific populations can be done.

References

1. Cantarero Rodrigo, Ramirez Blanca. 1997. *Environmental health hazard risks in the minority community survey.* Lincoln, NE: Lincoln-Lancaster County Health Department.
2. US Department of Transportation. 1996. *North American emergency response guidebook.* Washington, DC: US Department of Transportation.

Using a Geographic Information System to Guide a Community-Based Smoke Detector Campaign

Garry Lapidus, PA-C, MPH (1),* Steve McGee, BS (2), Robert Zavoski, MD, MPH (1), Ellen Cromley, PhD (2), Leonard Banco, MD (1)

(1) Connecticut Childhood Injury Prevention Center, Children's Medical Center and the University of Connecticut School of Medicine, Hartford, CT; (2) Department of Geography, University of Connecticut, Storrs, CT

Abstract

Smoke detectors are proven effective in providing early warning to occupants in residential fires. We used a geographic information system (GIS) to identify areas of greatest need to guide a community smoke detector campaign in Hartford, Connecticut. Computerized fire incident data for all residential fires from 1992 to 1994 were collected from the Hartford Fire Department. PC ArcView was used to geocode street addresses and to categorize census tracts into quartiles by the frequency of house fires occurring within them. Population, income, and housing data by census tract came from the 1990 US census. There were 942 house fires resulting in 41 civilian injuries, 9 civilian deaths, and 282 firefighter injuries. We identified four census tracts with the highest frequency of house fires in homes without functional smoke detectors. In census tracts with a high frequency of house fires a large proportion of those homes either had no smoke detectors or had smoke detectors that were nonfunctioning. Several standard GIS functions were important in the analysis and display of data. We geocoded the street address of over 900 house fires, which allowed us to view the spatial distribution and identify high-risk areas. Each point carried additional data on the characteristics of each fire. We were able to group house fires by census tract and relate them to other geographic information such as population, economic, and housing data. In November 1997, a community fire safety coalition installed more than 75 new smoke detectors, and tested and replaced batteries for existing detectors in one high-risk census tract. This approach is useful for other communities interested in conducting targeted smoke detector campaigns.

Keywords: injury, house fire, smoke detectors

Introduction

Each year in the United States an estimated 5,000 people die and an additional 30,000 are hospitalized due to residential fires (1). Inhalation of carbon monoxide and smoke causes the majority of these deaths (2). Smoke detectors are proven effective in providing early warning to occupants in residential fires (3,4). Many organizations, including local fire and health departments, conduct smoke detector promotion campaigns, often targeting high-risk areas such as low-income neighborhoods with high proportions of children and/or elderly residents (5). Recent studies have noted the increased use of smoke detectors in these areas (6,7) and one program found that most homes in the

* Garry D Lapidus, Connecticut Children's Medical Center, 282 Washington St., Hartford, CT 06106 USA; (p) 860-545-9988; (f) 860-545-9975; E-mail: Glapidu@CCMCKIDS.org

target area already had a detector (8). The purpose of this study was to demonstrate the usefulness of a geographic information system (GIS) in analyzing a variety of data so as to identify areas in greatest need of a community smoke detector campaign to be run in Hartford, Connecticut.

Methods

Computerized fire incident data for all residential fires from 1992 to 1994 were collected from the Hartford Fire Department. Data included temporal characteristics (hour, day, month, year), extent of flame and smoke damage, type of residence (height, construction type), injured persons, method of alarm, response time, street address, cause, room of origin, and smoke detector use (present and working; present and not working; not present; unknown). PC ArcView (a GIS computer program) was used to geocode street addresses and to categorize census tracts into quartiles by the frequency of house fires occurring within them.

The street addresses of schools, firehouses, community centers, and churches were obtained from a standard telephone directory for Hartford, and were geocoded. Population, income, and housing data by census tract came from the 1990 US census.

Results

There were 942 house fires resulting in 41 civilian injuries, 9 civilian deaths, and 282 firefighter injuries. House fires were more likely to occur in more densely populated census tracts (Figure 1). Figure 2 identifies four census tracts with the highest frequency of house fires in homes without smoke detectors. Figure 3 identifies one of the census tracts with the highest frequency of house fires in homes with smoke detectors present but not working. Finally, Figure 4 identifies the location of potential collaborators in a smoke detector promotion campaign within this census tract.

On November 22, 1997, 35 volunteers from across the state gathered at Fire Station House #7 in the north end of Hartford to install smoke detectors in Project Get Alarmed. More than 75 smoke detectors were distributed door to door by firefighters, police officers, teenage volunteers from Explorer posts of both departments, as well as volunteers from Connecticut Children's Medical Center and Connecticut SAFEKIDS. Depending on families' needs, volunteers tested detectors, replaced batteries, or installed new smoke detectors. Along with the families that were affected directly, numerous people were reached through the extensive media coverage the event gained that day. Similar activities are scheduled to reach the other high-risk areas within the city.

Discussion

Interrelated approaches including control of ignition sources, early warning and minimizing losses during fires, and provision of care after fires are required for the reduction of injuries from house fires. Smoke detectors and home sprinklers are proven strategies for early warning and rapid suppression of residential fires, but they have not been universally adopted. Approximately 80% of fire-related deaths occur in homes without working smoke detectors (9).

The results of our project indicate that in census tracts with a high frequency of

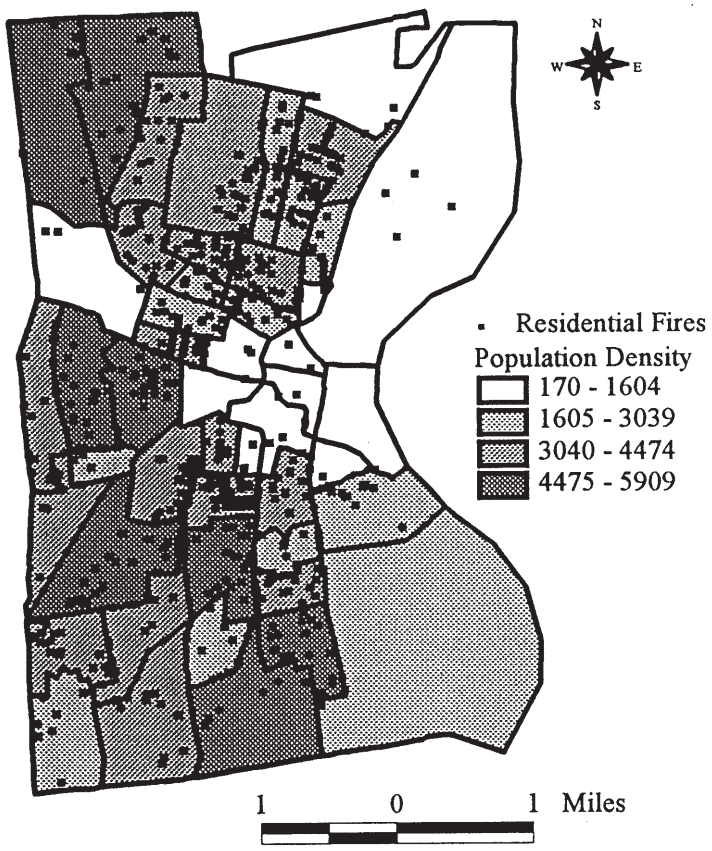


Figure 1 Residential fires, 1992–1994 (n=942), and population density by census tract, Hartford, CT.

house fires, a large proportion of those homes either had no smoke detectors or had smoke detectors that were nonfunctioning (presumably because of an absent or dead battery). Several standard GIS functions were important in the analysis and display of data. For example, we were able to quickly and easily geocode the street addresses of over 900 house fires, which allowed us to view the spatial distribution and identify high-risk areas. Each point on the map (indicating where the house fire occurred) carried additional data on the characteristics of each fire. Another GIS function performed easily was the grouping of house fires by census tracts and relating them to other geographic information such as population, economic, and housing data. Finally, we were able to show specific geographic areas in need of new smoke detectors and/or battery replacements, as well as the location of fire department stations, community agencies, churches, and schools that can be recruited to participate in a fire safety campaign. This approach is useful for other high-risk communities interested in conducting smoke detector campaigns.

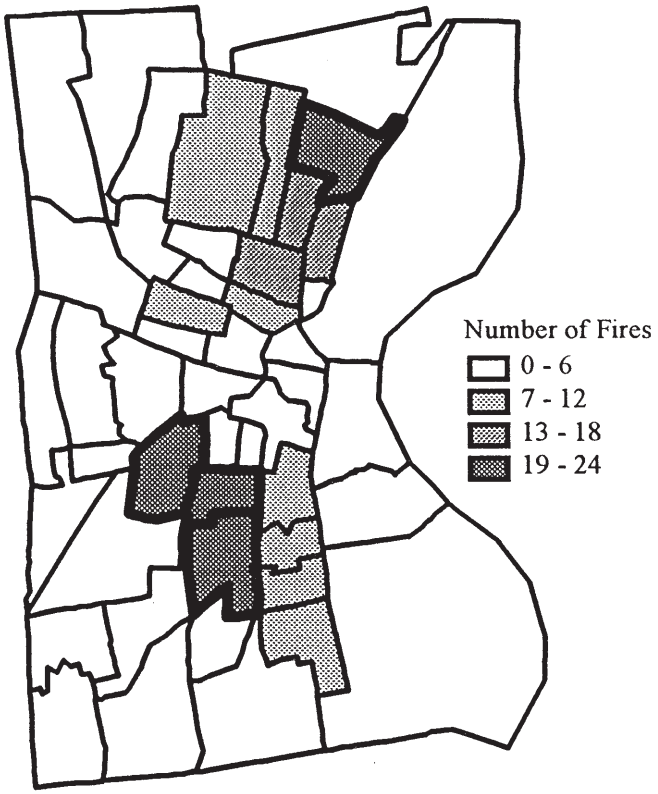


Figure 2 Residential fires, smoke detectors not present (n=286), by frequency and census tract, Hartford, CT.

References

1. Karter MJ. 1994. Fire loss in the United States in 1993. *National Fire Protection Association Journal* 88:57–65.
2. Baker SP, O'Neill B, Ginsburg MJ, Li G. 1992. *The injury fact book.* 2nd Ed. New York: Oxford University Press. 161–73.
3. Runyan CW, Bangdiwala SI, Linzer MA, Sacks JJ, Butt J. 1992. Risk factors for fatal residential fires. *New England Journal of Medicine* 327:859–63.
4. Mallonee S, Istre G, Rosenburg M, Reddish-Douglas M, Jordan F, Silverstein P, Tunell W. 1996. Surveillance and prevention of residential-fire injuries. *New England Journal of Medicine* 335:27–31.
5. Gorman RL, Charney E, Holtzman NA, Roberts KB. 1985. A successful city-wide smoke detector giveaway program. *Pediatrics* 75:14–18.
6. Gallagher SS, Hunter P, Guyer B. 1985. A home injury prevention program for children. *Pediatric Clinics of North America* 32:95–112.
7. Shaw KN, McCormick MC, Kustra SL, Ruddy RM, Casey RD. 1988. Correlates of reported smoke detector usage in an inner-city population: participants in a smoke detector give-away program. *American Journal of Public Health* 78(6):650–3.

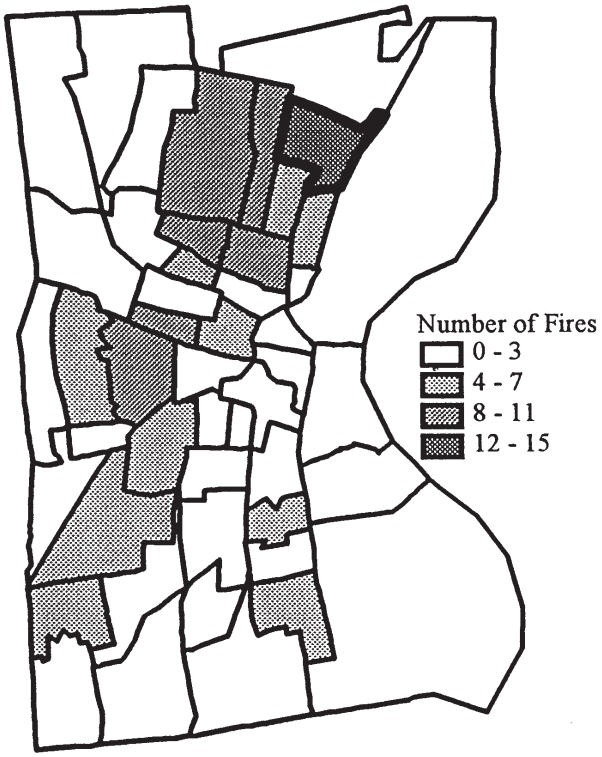


Figure 3 Residential fires, smoke detectors present but not working (n=151), by frequency and census tract, Hartford, CT.

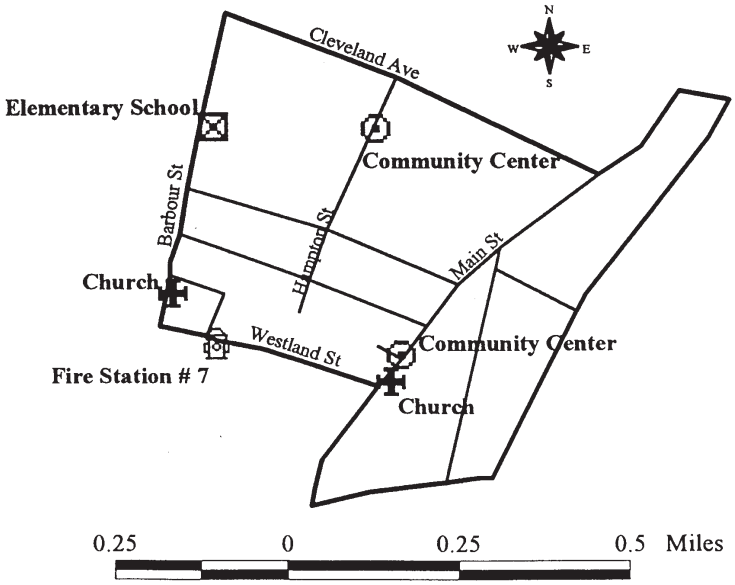


Figure 4 Potential collaborators for a targeted smoke detector campaign, census tract 10.

8. O'Connor MA. 1987. *New Hampshire home injury prevention project: interim report.* Submitted to the New England Network to Prevent Childhood Injuries. 10 September.
9. US Fire Administration. 1993. *Fire in the United States 1983–1990.* 8th Ed. Emmitsburg, MD: US Fire Administration.

Kriging Analysis Applied to Ecological Risk Assessment of Harbor Sediments

Christopher J Leadon*
Environmental Engineer, Environmental Specialist Support Team (ESST), Southwest Division (SWDIV), Naval Facilities Engineering Command, San Diego, CA

Abstract

This study applies the spatial statistical technique, kriging, to the estimation of benthic invertebrate community ecological parameters associated with ecological risks from hazardous contamination in harbor sediments. The types of ecological risk assessment used in the US Navy's Installation Restoration program for cleaning up sites contaminated with hazardous substances are described. Benthic invertebrate community data from contaminated harbor sediments at a California Navy base are presented as an example of ecological community data being used in ecological risk assessment of harbor sediments. The ecological community parameters in the dataset include abundance, diversity, number of species, dominance, evenness, and biomass. The kriging of the benthic ecological community data from the example harbor shows that the total number of sampling stations initially planned may be reduced by 10% without compromising the characterization of the benthic ecological community. Kriging can be a very effective statistical method for limiting the number of samples needed to spatially characterize hotspots while still insuring adequate data quality. Maps of the spatial error variance of a parameter's sample data—the error of estimation—may be used to place additional sampling points or to minimize the number of additional samples needed at a site.

Keywords: kriging, ecological, risk

Introduction

Kriging is a geostatistical method of spatial data interpolation that can be used to limit the number of samples in eco-risk assessments. In 1963 G Matheron named kriging after DG Krige, a South African mining engineer who used the technique to more accurately predict the extent of gold deposits. Kriging is an interpolation method that optimally predicts data values by using data taken at known nearby locations. It can be either two- or three-dimensional. For this paper, ecological data from harbor surface sediments, considered a two-dimensional surface, were kriged.

Kriging is a set of linear regression routines that minimizes estimation variance from a predefined covariance model (1). It is based on the assumption that the parameter being interpolated at a site is a *regionalized* variable. A regionalized variable varies in a continuous manner spatially so that data values from points nearer to one another are better correlated. Data values from widely separated points are statistically independent in kriging.

Estimates of chemical or biological parameters and their associated variances can be predicted at each node of a grid by a kriging model. New proposed sampling

* Christopher Leadon, Southwest Division, Naval Facilities Engineering Command, 1220 Pacific Highway, San Diego, CA 92101-3327 USA; (p) 619-532-2584; (f) 619-532-1195; E-mail: LEADONCJ@efdsw.navfac.navy.mil

locations can be added to a dataset and the reduction in kriging variance can be estimated at each location. The resulting maps of kriging variance with the proposed additional sampling locations can then be used to limit the number of suggested new sampling sites. Only those resulting in significant variance reduction would qualify as actual sampling locations. Alternative new sampling designs can also be assessed with kriging to determine the greatest benefit for the cost of additional sampling sites.

The data from the example harbor in this paper were kriged using the US Department of Defense's Groundwater Modeling System (GMS) package (1). Kriging software is also available in other geostatistics packages, such as Surfer for Windows: Contour and 3-D Surface Mapping (Golden Software, Inc., Golden, CO) and MGE Kriging Modeler (Intergraph Corporation, Huntsville, AL).

Steps in Ecological Risk Assessment

Ecological and human health risk assessments are integral to the scientific investigation of sites contaminated by toxic chemicals. The US Navy conducts cleanups of toxic chemicals found at sites on Navy or Marine bases through its Installation Restoration (IR) program, the Navy's version of Superfund. Evaluation of an IR site includes five steps: scoping, screening, baseline, effectiveness (confirmation sampling), and monitoring eco-risk assessments. Resampling or taking additional spatial samples for hotspot delineation can be very expensive; at Navy and Marine bases, for example, each sampling event of harbor sediments for eco-risk assessments can cost over $1 million. Kriging can reduce the need for resampling if it is used to systematically organize all stages of the assessment effort to reduce spatial error variance at a site. Kriging can also be used to limit the number of sampling stations needed to delineate hotspots.

Data Quality Objectives

Ecological risk assessment follows the seven-step Data Quality Objectives (DQOs) process, as do all sampling investigations in the Navy's IR program. The seven steps in the DQO process are:

1. State the problem
2. Identify the decision
3. Identify inputs to the decision
4. Define the study boundaries
5. Develop a decision rule
6. Specify tolerable limits on decision errors
7. Optimize the design

Kriging should be used to plan the tolerable limits on decision errors in step 6 of the DQO process. The use of error variance maps generated by the kriging of data can reduce the error of spatial estimation in studies such as the one profiled here. The project team has to agree upon the amount of tolerable spatial error of estimation in step 6 before the investigation proceeds to step 7, where the sampling design is optimized. A major consideration in this decision is the cost of sampling. Although kriging may indicate that a certain number of locations should be sampled, it may not be possible to pay for the optimal number of samples in a project. In addition to error from the spatial

variation in data values, other sources of error variance in studies of ecological data are the regular variance about the mean value of the data itself, temporal variation, and error in the reported data from lab tests. In the case of harbor sediments, the temporal variation in the ecological data is due to the movement of sediment and biological changes in response to changing site conditions.

DQO step 7 includes four substeps (2):

1. Review DQO outputs and existing environmental data.
2. Develop general data design alternatives.
3. Formulate the mathematical expressions needed to solve the design problem for each data collection design alternative.
4. Develop and document the sampling strategy.

Kriging can be used in substeps 2, 3, and 4. In substep 2, kriging is useful as a statistical method to determine the appropriate number of samples. In substep 3, it can be used as a statistical model and, in substep 4, to decide on the locations of the sampling stations delineated.

Data and Example Study Areas

The study area for this application of kriging to an eco-risk assessment was a 0.738 acre harbor at a California Navy base. The harbor is approximately 4,500 feet by a maximum of 8,200 feet across (0.85 by 1.55 miles), with an average water depth of 45 feet. The sediments contained a wide range of grain sizes, but they were about 65% fines, meaning the particles were smaller than 62.5 μm in diameter. The sediment in the area near the basin entrance on the east side contained a high percentage of sand-sized particles.

The study included 32 open harbor sampling sites. Only open water sampling stations were included in this analysis. Although samples were also collected from underneath piers around the sides of the example basin, these areas are considered a different ecosystem than the open harbor and the samples were not included in the analysis. Sample volumes of 0.006 m^3 were obtained using a Teflon corer inserted into a box core surface sediment sample from the harbor bottom.

This study used a triad approach to eco-risk assessment of the harbor sediments. The first two components of the triad were chemical contaminant concentrations and bioassay results, including bioaccumulation tests from sediment samples. Some of the hazardous contaminant chemicals found above background concentrations in the example harbor included the metals arsenic, beryllium, cadmium, total chromium, copper, lead, mercury, nickel, silver, and zinc; sulfide; polynuclear aromatic hydrocarbons (PAHs); polychlorinated biphenyls (PCBs) such as Arochlor 1260; and total 4,4'-dichloro-diphenyl-trichloroethane (DDT).

Benthic community data was the third major component of the triad approach. The benthic community found in the open areas of the harbor was dominated by five polychaetes: *Monticellina tesselata, Cossura sp. A, Aphelochaeta multifilis* Type 2, *Chaetozone corona,* and *Paraprionospio pinnata.* The polychaete, *Pseudopolydora paucibranchiata,* and the crustacean, *Amphideutopus oculatus,* were also abundant at several sampling sites in the open harbor.

The benthic invertebrate ecological community parameters calculated from the sediment samples included abundance, the Shannon-Weiner diversity index, evenness,

Margalef species richness, dominance, and biomass. Abundance was reported as the total number of individual specimens collected in each sample. The Shannon-Weiner species diversity index was calculated as:

$$H = -\sum_{i=1}^{s} (p_i)(\ln p_i) \quad (1)$$

where H is the Shannon-Weiner diversity index, s is the number of species, and pi is the proportion of the total sample belonging to the i^{th} species, the abundance of species i/total abundance. Evenness was computed as:

Evenness = (Shannon-Weiner species diversity index H)/ln(number of species) (2)

Margalef's species richness was defined as:

Margalef's species richness = (number of species – 1)/ln(total abundance) (3)

Dominance was calculated as the number of species accounting for 75% of the total abundance. Biomass was the wet weight in grams of all organisms found in a sample (g/[0.006 m^3]).

Variograms

In ordinary kriging, a variogram is first constructed using the dataset from a site. A variogram consists of two parts: an experimental variogram based on the data, and a model variogram. An experimental variogram is constructed by first calculating the variance of each point in a dataset with respect to each of the other points. The experimental variogram consists of the plotted variances versus the distance between each data point at the site. The variance is typically computed as one-half the difference in a data value squared. Several types of experimental variograms can be selected in the GMS software (1).

The following semivariogram equation was used to calculate the variance, $\gamma(h)$, as a function of the distance, h, between data points, for the experimental variogram for this study:

$$\gamma(h) = (1/2n) \sum_{i=1}^{n} (f_{1i} - f_{2i})^2 \quad (4)$$

where n is the number of pairs of points whose separation distance falls within the lag interval and f_{1i} and f_{2i} are the data values at the head and tail of each pair of points (1). In computing the experimental variogram, it is impractical to plot a variance for each data point with respect to every other value in the dataset. Therefore, the variances are averaged for all the data points in donut-shaped areas around each data point called *lags* and plotted on the experimental variogram. The distance between the edges of each adjacent lag area is called the *unit separation distance* (1). Ten lags, the number normally used, was chosen as the number of lags for the kriging of the ecological data in this study. Thus, only ten points are shown on the experimental variogram (Figure 1), corresponding to the average variances in ten lag areas.

The model variogram is a curved line through the experimental variogram points. It represents a simple mathematical function modeling the trend in the points of the experimental variogram. In the GMS package, spherical, exponential, Gaussian, and power model equations can be used to fit a model variogram line to the experimental variogram points. A spherical model equation resulted in a line closely matching that

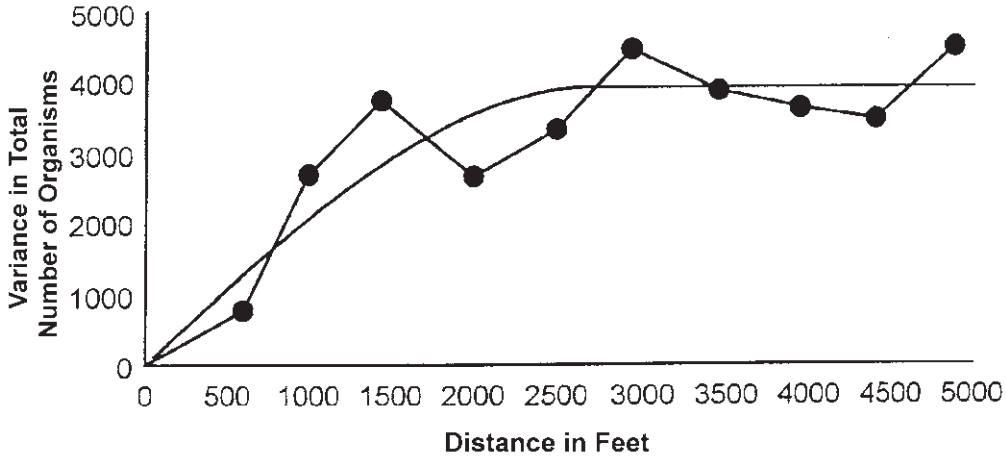


Figure 1 Variogram for benthic invertebrate total abundance in the example harbor.

of the benthic community variance data points on the experimental variogram, and it was used for the model variogram in this study:

$$\gamma(h) = c\,[1.5(h/a) - 0.5(h/a)^3] \quad \text{if } h/a \le a \tag{5}$$

where $\gamma(h)$ is the variance as a function of distance h, a is the *range,* and c is the *contribution* or *sill.*

On a model variogram, a *nugget* is a minimum variance—the point where the model variogram line intercepts the y axis. The sill is the upper flat part of the model variogram line where the variances of far data points have no correlation to the distance from the other data points. The range represents the distance at which there is no longer a correlation between data points (1).

Some datasets exhibit *anisotropy,* meaning the correlation between data points changes with the direction set through the dataset. For isotropic data, the azimuth angle has little effect on the resulting experimental variogram. The benthic ecological community dataset used for this paper was assumed to be isotropic and the azimuth angle was set at zero degrees.

Estimation Error Variance Maps

The variogram in kriging can be used to calculate the expected error of estimation at each target interpolation point because the estimation error is a function of the distance to surrounding data points. The estimation variance can be represented as:

$$s^2_\varepsilon = w_1\,(S(d_{1p}) + w_2 S(d_{2p}) + w_3 S(d_{3p}) + \lambda \tag{6}$$

An estimation standard deviation can also be calculated by taking the square root of the estimation variance. In the kriging module of GMS, a contour map of estimation variance can be generated for a mesh or grid at a site by selecting a simple option button in kriging options.

Data Interpolation

The basic equation used in ordinary kriging is:

$$F(x,y) = \sum_{i=1}^{n} w_i f_i \quad (7)$$

where n is the number of data points in the set, f_i are the values of the data points, and w_i are weights assigned to each data point (1). The weights used in kriging are from the model variogram. To interpolate at a point P, for example, using surrounding points P_1, P_2, and P_3 , the weights w_1, w_2 , and w_3 must be found. The weights are found through the solution of the simultaneous equations:

$$w_1S(d_{11}) + w_2S(d_{12}) + w_3S(d_{13}) = S(d_{1p}) \quad (8)$$

$$w_1S(d_{12}) + w_2S(d_{22}) + w_3S(d_{23}) = S(d_{2p}) \quad (9)$$

$$w_1S(d_{13}) + w_2S(d_{23}) + w_3S(d_{33}) = S(d_{3p}) \quad (10)$$

where $S(d_{ij})$ would be a value from the model variogram evaluated at a distance equal to the distance between points i and j. Because it is necessary that the weights sum to one, a fourth equation is added:

$$w_1 + w_2 + w_3 = 1.0 \quad (11)$$

Because there are now four equations and three unknowns, a slack variable, λ, is added to the equation set:

$$w_1S(d_{11}) + w_2S(d_{12}) + w_3S(d_{13}) + \lambda = S(d_{1p}) \quad (12)$$

$$w_1S(d_{12}) + w_2S(d_{22}) + w_3S(d_{23}) + \lambda = S(d_{2p}) \quad (13)$$

$$w_1S(d_{13}) + w_2S(d_{23}) + w_3S(d_{33}) + \lambda = S(d_{3p}) \quad (14)$$

$$w_1 + w_2 + w_3 = 1.0 \quad (15)$$

These equations are solved for the weights w_1, w_2, and w_3. The f value of the interpolation point is then calculated as:

$$f_p = w_1f_1 + w_2f_2 + w_3f_3 \quad (16)$$

The expected estimation error is minimized in a least squares sense in kriging by using the variogram to compute the weights (1). For this reason, kriging is said to produce the best linear unbiased estimate. In most mapping software manuals, kriging is recommended as the best interpolation method.

Results of Interpolated Data and Variance Mapping

Maps of isopleths for total abundance and other community parameters for benthic invertebrates were computed using the kriging interpolation equations. The maps were first computed and printed for all 32 original sampling stations in the harbor. Figure 2, for example, illustrates isopleths of benthic invertebrate total abundance for all the

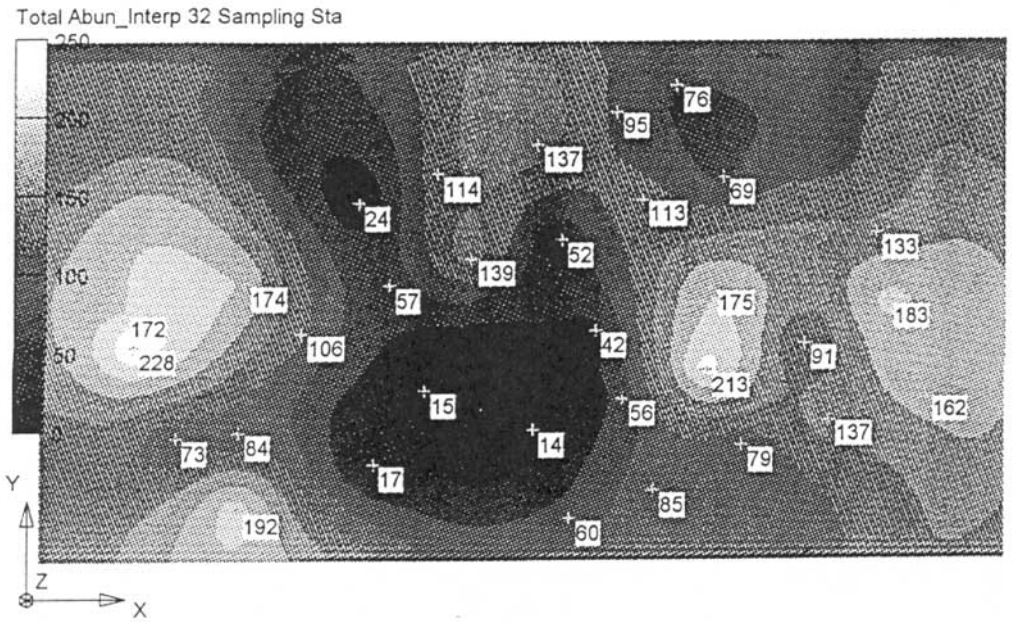


Figure 2 Kriged isopleths for benthic invertebrate total abundance for the original 32 sampling stations in the example harbor.

sampling stations. An estimation error variance map was also computed for each parameter at each station; Figure 3 shows the estimation error variance for total abundance.

Four, and later eight, sampling stations were then removed from the dataset. The eliminated sites were those located nearest to other stations. The purpose of computing interpolated isopleths and estimation error variance maps for datasets with reduced numbers of sampling stations was to see if the isopleths would stay the same or change with fewer sampling stations. Maps of isopleths for total abundance and the other ecological community parameters were then produced using kriging interpolation for reduced datasets with 28 and 24 sampling stations. Figure 4 shows the resulting interpolated isopleths for total abundance for 28 sampling stations. Estimation error variance maps were also computed using 28 and 24 sampling stations.

Conclusions

The positions of the isopleths of the predicted values of total abundance changed very little when four sampling stations were removed from the original dataset (Figures 2 and 4). The positions of the interpolated isopleths did change, however, when the number of stations was reduced from 32 to 24. The positions of the estimation error variance isopleths on maps changed very little when the number of sampling stations was reduced to either 28 or 24 stations. Because the number and position of original sampling stations were arrived at by the best collective judgement of the IR team, it is likely that 10% fewer sampling stations could be used to collect benthic ecological community data for eco-risk assessments of harbor sediments at this site. For a harbor the size of

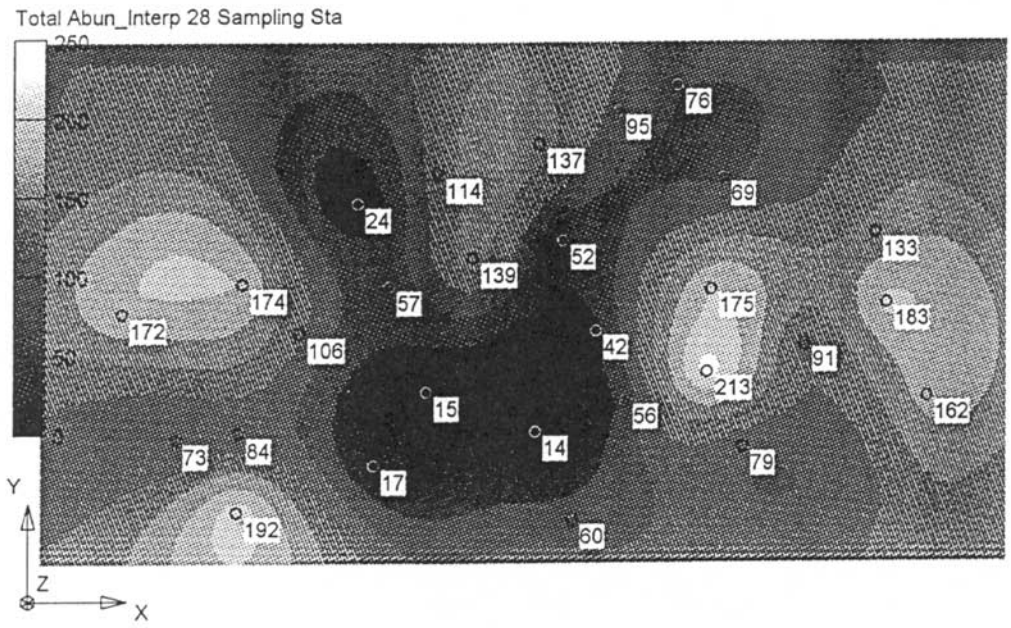


Figure 3 Kriged isopleths for benthic invertebrate total abundance with the number of sampling stations reduced to 28 in the example harbor.

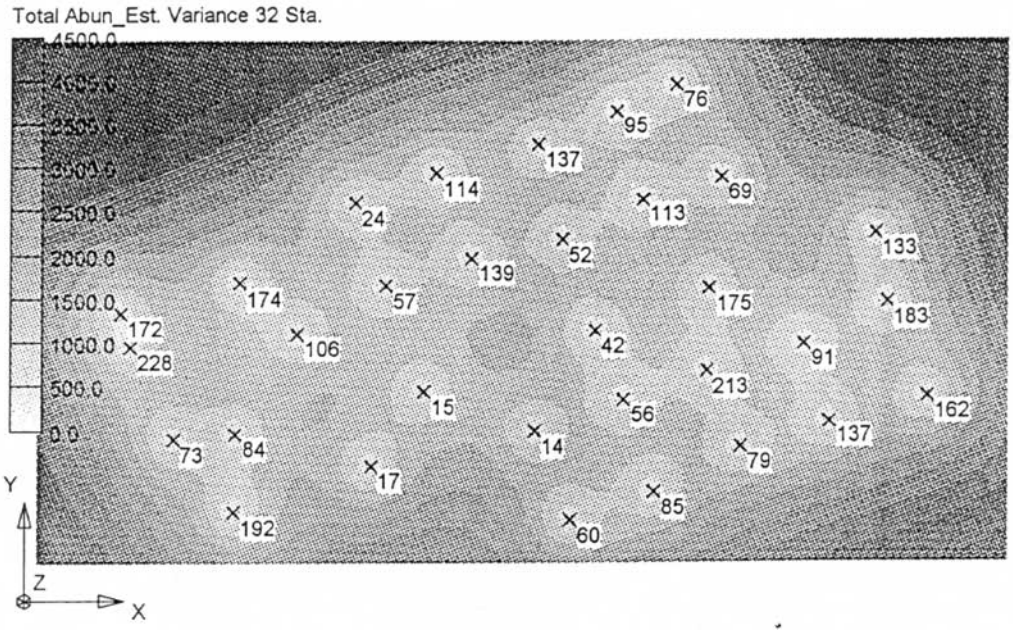


Figure 4 Estimation error variance map for benthic invertebrate total abundance for the original 32 sampling stations in the example harbor.

the one in this study, collecting 10% fewer benthic community samples could result in savings of between $100,000 and $250,000 in 1998 dollars through the three to five major stages of an eco-risk assessment.

Using 28 sampling stations as the number necessary to characterize the spatial distribution of benthic ecological community parameters in a harbor the size of the example harbor, one sampling station is needed per 1,147,995 square feet of sediment. This represents an area 1,071 feet on a side to adequately characterize benthic ecological community parameters. This is probably a larger area of harbor bottom per sampling station than was previously thought adequate to characterize benthic ecology.

Hotspot Definition

Kriging could save the government millions of dollars in sampling costs by reducing the number of samples collected to define the volumes of hotspots—small areas with high contaminate concentrations. Before remediating a hotspot, contractors have been intuitively collecting samples from hundreds of sampling stations to avoid remediating too much soil or sediment and to avoid missing contamination. Using kriging estimation error variance maps to plan the locations of sampling stations in areas with the most estimation variance could reduce the number of stations needed to characterize hotspots. Isopleths of contamination on maps of hotspots could be more accurately predicted by using kriging interpolation.

Recommendations

Kriging should be included in DQO planning for eco-risk assessments of harbor sediments. Through each successive stage of an eco-risk assessment, an effort should be made to build a database by placing sampling station locations in a consistent grid pattern. Kriging estimation error variance maps of preliminary data should be used to plan the size of the grid and optimally place sampling station locations in the areas with the most estimation error variance. The location of each sampling station should be placed randomly inside each grid cell. As additional sampling is planned through the stages of an eco-risk assessment, new stations should always be placed in the areas with the most estimation error variance. Kriging estimation error variance maps and interpolated isopleths of data should be used to plan sampling and define the shape of hotspots.

Acknowledgments

The author would like to thank Jed Costanza at the Naval Facilities Engineering Service Center (NFESC) in Port Hueneme, California, and Walt Kitchin of the Environmental Specialist Support Team, Southwest Division of the Naval Facilities Engineering Command, for their expert assistance in running the kriging software in the GMS computer package.

References

1. US Department of Defense. 1998. *Groundwater modeling system, GMS v2.0, reference manual.* Washington, DC: US Department of Defense.

2. Bilyard GR, Beckert H, Bascietto JJ, Abrams CW, Dyer SA, Haselow LA. 1997. *Using the data quality process during the design and conduct of ecological risk assessments.* Washington, DC: US Department of Energy, Assistant Secretary for Environment, Safety, and Health, Office of Environment.

Automated Process for Accessing Vital Health Information at Census Tract Level

Hsiu-Hua Liao (1),* Paul Laymon (2), Kirk Shull (2)

(1) St. Louis County Department of Planning, Clayton, MO (2) Division of Biostatistics, South Carolina Department of Health and Environmental Control, Columbia, SC

Abstract

Recent emphasis on streamlined government and health care reform encourages community leaders to search for innovative ways to effectively manage their regions of responsibility. Gradually, geographic information system (GIS) technology is becoming a recognizable tool in the public health community for uses from intervention strategies to health care reform. One advantage of implementing GIS is that it can geographically locate personal health data through a geocoding process and allow examination of their spatial patterns. Georeferencing personal health data will greatly enhance decisions made by public health officials; however, it complicates the burden of protecting personal rights to confidentiality. One solution to the dilemma is to aggregate personal identities to a group of data where no identity will be revealed. This paper describes how this process was used to geocode vital health records and aggregate them to the census tract level. Data aggregation was accomplished through the Vital Health and Census Data Integration System (VHCDIS), an ARC/INFO-based GIS automation system. The primary objectives for the process were to promote personal privacy, automate health data aggregation of georeferenced vital records data, and improve national access to spatial health information.

Keywords: geocoding, vital health, automation, census data

Introduction

For centuries, health researchers have been using spatial locations, boundaries, and regions to determine the quality, quantity, and migration of epidemics. Overlaying quantitative graphics upon a map enables the viewer to realize potential information in an extremely clear manner. For example, the famous 1854 London cholera study conducted by Dr. John Snow has been hailed as the geographic benchmark for using maps in epidemiological studies.

Currently, the South Carolina Department of Health and Environmental Control (SCDHEC), Division of Biostatistics, presents spatial health information on the county level. County level data provide a wealth of information. At this macro scale, however, it is difficult for local health officials to adequately identify, analyze, and monitor health problems at a micro scale or community level. Hence, in 1989, the Johnson Wood Foundation authorized a grant for the SCDHEC's Vital Record Geographic Referencing System (VRGRS) and the University of South Carolina's School of Public Health to generate a feasibility study of georeferencing vital records data for the purpose of assisting

* Hsiu-Hua Liao, St. Louis County, Dept. of Planning, 41 S. Central Ave., Clayton, MO 63105 USA; (p) 314-615-3899; (f) 314-615-3729; E-mail: Hsui-Hua_Liao@co.st-louis.mo.us

public health assessments, surveillance, and health hazard evaluations at the community level. The main objectives for VRGRS were:

1. To implement a program that encoded the geographic residential location for births and deaths, and apply a geographic information system (GIS) as part of the statewide vital records system.
2. To demonstrate the application of location data in association with the TIGER (topologically integrated geographic encoding and referencing) system of the federal census of 1990.
3. To design and document the process in a way that facilitated expansion that complemented a statewide GIS for economic development.

The VRGRS project outcome ultimately determined that the processes, scientific techniques, and data were suitable enough to implement an informal GIS program within the Division of Biostatistics. Hence, in 1994 staff and equipment were selected to carry on the objectives of VRGRS and to establish the means to systematically georeference vital health data collected and stored at the Office of Vital Records.

Georeferencing provides an opportunity to examine health data and how they will be distributed over spatial domain; however, this also raises the issue of confidentiality. When the geographic resolution of data is fine enough to identify fewer than four addresses, the data are no longer tools of research, but tools to potentially target and expose individuals (1). The protection from inadvertent disclosure of individuals, households, establishments, or primary sampling units, especially in public use databases, is a concern of government health agencies. Even though confidentiality policies may vary among agencies, they must reflect the laws and regulations imposed upon personal data collection and dissemination activities (2). To date, there is not a minimum national threshold standard defining public or professional access to spatially referenced public health data.

In an attempt to promote spatially referenced public health confidentiality standards, the South Carolina Division of Biostatistics GIS Lab focused on the development of a statewide health information system capable of satisfying the wide range needs of health researchers. To develop such a system, the Biostatistics GIS Lab needed a geocoding system capable of converting large volumes of data with acceptable match rates. After a series of tests that included quality, cost, and turn around times, Geographic Data Technology (GDT) from Lebanon, New Hampshire, was chosen to perform the geocoding process.

Once the vital records health data were converted into individual points, the issue of confidentiality was solved by aggregating the data to the 1990 census tracts. The census tracts were chosen for two reasons. First, census tracts contained a volume of socioeconomic data. Thus, the aggregate vital records attribute information could be combined with the existing socioeconomic census data (e.g., mother's age extracted from the vital records would be stratified into the same categorical breakout as the female populace of the tracts, allowing calculation of statistical rates). Second, geographic boundaries are updated once every decade.

Working with voluminous vital records files proved to be tedious and time consuming. To streamline the process of generating public health data from these records, the Vital Health and Census Data Integration System (VHCDIS) was developed. In designing the system five requirements were determined:

- It must be flexible enough to be continuously improved.
- It must be a time saver.
- It must establish a national precedence for collecting health data.
- It must standardize data output.
- It must accurately aggregate health data to predetermined political boundaries (in this case the census tracts).

In its completed form, the VHCDIS offers national and local programs the ability to join aggregate vital records health data with existing socioeconomic census data as a tool for their respective surveillance and intervention strategies. The remaining point data derived from the geocoding process, which are treated with all the confidentiality of a paper certificate, are stored on a magnetic device for future use in very high resolution studies.

Background

Vital Health Statistics

Vital statistics for the United States are obtained from the official records of live births, deaths, fetal deaths, marriages, divorces, and annulments. These datasets have long been used as statistical measuring devices to identify qualitative and quantitative public health issues. The official recording of these events is the individual responsibility of each state and independent registration areas (District of Columbia, New York City, and territories). The federal government, without expressed constitutional authority to enact national vital statistics legislation, relies upon the states to establish laws and regulations to provide compatible methods of registration and data collection (3).

As public health issues continue to become more and more complex, demand for better vital statistics information increases. For this reason, updating data collecting, recording, and processing techniques to keep aligned with the rapidly evolving need becomes an increasingly important part of the vital statistics program. Improvement began in the 1950s with increased attention placed on improving the quality of vital statistics data to make them more useful and accessible. Interest in vital statistics expanded when the state and federal health and welfare officials began to look for pertinent and reliable statistics on which to base their political decisions. The registration certificates assumed a new role of importance as they were used as a source of credible national vital statistics by all levels of government, institutions, and the general public. The content of the information collected for vital records was expanded and methods to improve its quality and usefulness were added. As health and social issues became more complex, supplemental data sources were developed to augment and enrich the information obtained from the registration system.

Throughout the years the process of producing national vital statistics has shifted several times from one organizational unit of the federal government to another. The National Center for Health Statistics and the National Association for Public Health Statistics and Information System (NAPHSIS) have become recognized for handling health statistics and the associated information systems. NAPHSIS was organized to study and promote all matters relating to the registration of vital statistics. The 1995 revision of the association bylaws states:

> This Association will foster discussion and group action on issues involving public health statistics, public health information systems, and vital records registration. The Association will provide standards and principles for administering public health statistics, public health information systems, and vital records registration. The Association will represent the States and Territories of the United States regarding these issues, and will serve as an advisory group to the Association of State and Territorial Health Officials (3).

With the increasing complexity of public health issues, federal and local health programs need to improve the process of collecting, storing, analyzing, and displaying community level epidemic information. For this reason, future focus on quality spatial vital records data will continue to grow at an exponential rate. Likewise, vital records recording programs tasked to increase the accuracy of vital statistics will continue to explore the development of new technologies, rethinking the use for these valuable resources.

Public Health and GIS

GIS technology has gradually been recognized by public health researchers as a powerful tool for analyzing health data. It provides an opportunity to integrate at least six disciplines (epidemiology, environmental health, geography, cartography, computer sciences, and statistics) for the study of the distribution and possible causes of diseases in population, and the targeting of interventions to improve the health of the population (4). Applications of GIS in the health field vary from the simple automated mapping of epidemiological data (5), to the sophisticated analysis of satellite images to demonstrate vector/environment relationship (6,7,8,9).

The simplified paradigm for implementing GIS technology in public health can be viewed in three phases: data source identification, GIS support system, and health planning (Figure 1). In the data source identification phase, data sources applicable to your cause are selected and converted into digital geography or "coverages." The data sources used in the Division of Biostatistics are environmental health hazards, health services, and socioeconomic and health data. Environmental health hazard data can be defined as any data pertaining to an environmental situation that may have a negative impact on the surrounding population. Health services data are those that identify sources of health correction. And, for the scope of this paper, socioeconomic and health data can be defined as data collected for the purpose of monitoring, tracking, and identifying social and health trends.

Once these data sources are converted into digital coverages, they can be stored, manipulated, analyzed, and displayed in a GIS. This collection of standardized data becomes the foundation for the third phase of health GIS implementation, the health planning phase. In this phase, the GIS becomes the knowledge base for analyzing health outcomes and supporting public health surveillance, where a diverse group of scientific disciplines converge to direct and discover local level health objectives.

> The Centers for Disease Control and Prevention define public health surveillance as [t]he ongoing, systematic collection, analysis, and interpretation of health data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know. The final link of the surveillance chain is the

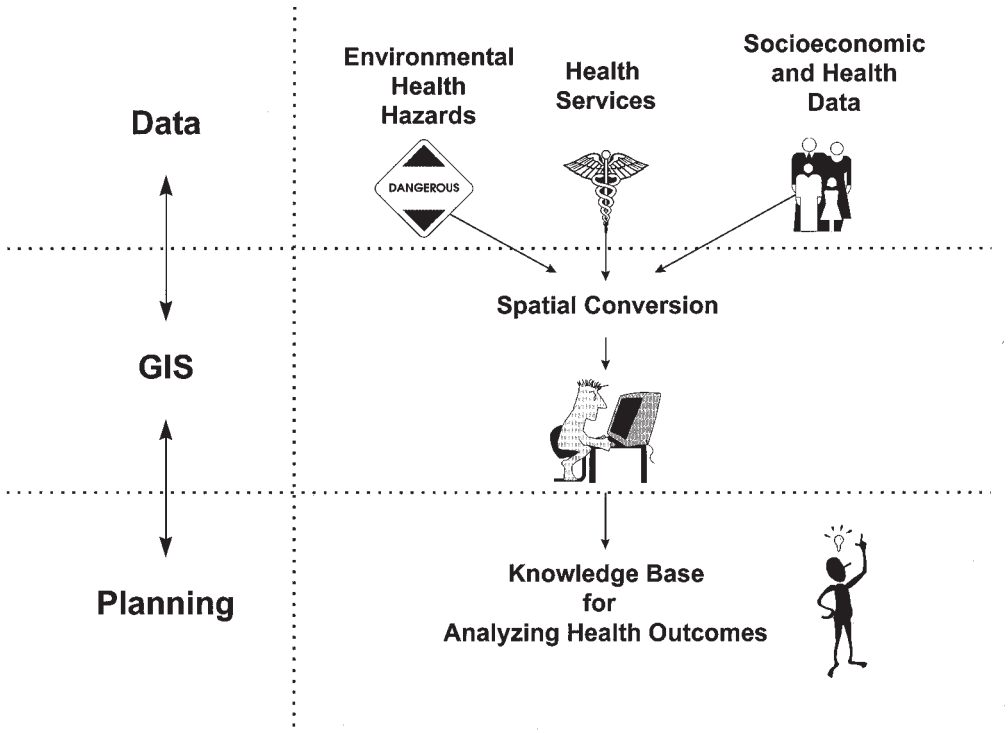


Figure 1 Paradigm for implementing GIS in public health.

application of these data to prevention and control. A surveillance system includes a functional capacity for data collection, analysis, and dissemination linked to public health programs (10).

Public health surveillance evolves with changes in science and technology. With the advent of computers, health surveillance has transformed from a primarily historical function to one that promotes timely analysis of data with appropriate responses to given health outcomes. Historically, the quality and quantity of data over a spatial domain were illusive and difficult to interpret. Today, using GIS technology as a tool we can streamline the processes needed to promote health and protect our environment.

Vital Health Data

In the state of South Carolina, vital health data are collected through official documents filed with the Office of Vital Records and the Public Health Statistics and Information System within the SCDHEC. Each year, the Division of Biostatistics publishes reports on vital statistics data for South Carolina live births, deaths, fetal deaths, marriages, divorces, and annulments that occurred during the previous year. These vital statistics are also available in publicly accessible format for public use. In the case of special requests for data, files and reports are generated and distributed by the Division of Biostatistics to those users who desire analysis different from those that are normally published.

VRGRS justified the use of GIS technology to improve the Division of Biostatistics'

capability of analyzing vital records data at an increased spatial resolution. Ultimately, this new technology functions around the process to geocode temporal vital records residence data of births and deaths in South Carolina. All births by residents were included, regardless of the state of occurrence, while South Carolina occurrences to non-residents were excluded. To support thematic mapping and GIS analysis, an attribute file identifying critical information about the birth and death events was generated and linked to the point by means of a common identifier.

Birth Data

In 1991, South Carolina began using a microcomputer software, Electronic Birth Pages (EBP), to improve the process of generating birth certificates and collecting newborn data for laboratory screening. The main function of the EBP system involves data entry and production of birth certificates. The end product is referred to as an EBC (electronic birth certificate).

To generate spatial information from vital records data, the residential address file is extracted from the mainframe dataset using the statistical software SAS (SAS Institute, Cary, NC). Variables used in geocoding include identification number, residential street address, city, state, zip code, and 4-digit zip code extension. These data undergo quality control measures to identify completeness and accuracy of the address information. To complete the dataset, an attribute record is captured as well. For example, the attribute file used for births includes identification number, county federal information processing standards (FIPS) code, age of mother, attendant at birth, birth weight, education level of the mother, month prenatal care began, number of prenatal care visits, race of the child, race of the mother, sex of the child, and year of birth. These chosen attributes were based on requests made by health districts and the Division of Epidemiology within SCDHEC. The attribute file was imported into ARC/INFO for the data aggregation process. Each variable was aggregated to the census tract level by race group (total, white, black, others, and unknown). Table 1 shows the classification of each variable.

Death Data

Death data were collected through death certificates filed by funeral homes. The funeral director, or person acting as such, is responsible for the completion of the death certificates, including all of the personal information from the family, and the medical portion of the certificate. This certificate is then sent to the county health department where it is screened for completeness. If the certificate is acceptable at the county level, the health department will forward the certificate to the SCDHEC's Office of Vital Records. The certificate is again checked for completeness, then personal data are coded and stored in the database.

For the residential address file, variables used in the geocoding process were synonymous with the birth data. For the attribute file, variables were temporally selected and causes of death were grouped into disease and non-disease type (Table 2). Causes of death are classified for purposes of statistical tabulation according to the *Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death* (11). In this process, only the underlying cause of death was selected for data aggregation.

Table 1 Classification of Birth Data: Live Birth, Low Weight Live Birth, and Very Low Weight Live Birth

A. LIVE BIRTH

RACE
- Total
- White
- Black
- Others
- Unknown

ATTENDANT
- Physician
- Certificated nurse midwife
- Other than physician, midwife, or self
- Self-attended
- Lay midwife, lay midhusband, registered lay midwife
- Nurse-midwife, graduate nurse-midwife
- Nurse, RN, OB nurse practitioner, physician's assistant
- D.O.
- Unreported or unknown

BIRTH WEIGHT
- <500 g
- 500–999 g
- 1,000–1,499 g
- 1,500–1,999 g
- 2,000–2,499 g
- 2,500–2,999 g
- 3,000–3,499 g
- 3,500–3,999 g
- 4,000–4,499 g
- 4,500–4,999 g
- 5,000 g
- Unreported

CHILD'S SEX
- Male
- Female
- Unreported

MOTHER'S AGE
- ≤13
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22–24
- 25–29
- 30–34
- 35–39
- 40–44
- ≥45
- Unreported or unknown

MOTHER'S EDUCATION
- Elementary school
- 1st grade
- 2nd grade
- 3rd grade
- 4th grade
- 5th grade
- 6th grade
- 7th grade
- 8th grade
- 9th grade
- 10th grade
- 11th grade
- 12th grade
- 1st year college
- 2nd year college
- 3rd year college
- 4th year college
- Graduate school
- Technical school
- Unreported

MONTH PRENATAL CARE BEGAN
- No prenatal care
- Began at 1st month
- Began at 2nd month
- Began at 3rd month
- Began at 4th month
- Began at 5th month
- Began at 6th month
- Began at 7th month
- Began at 8th month
- Began at 9th month
- Began at 10th month
- Unreported

NUMBER OF PRENATAL CARE VISITS
- No prenatal care visit
- 1–4 visits
- 5 visits
- 6–10 visits
- 11–15 visits
- 16 visits
- Unreported

B. LOW WEIGHT LIVE BIRTH (Birth Weight <2,500 g)

RACE
- Total
- White
- Black
- Others
- Unknown

MOTHER'S AGE
- ≤13
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22–24
- 25–29
- 30–34
- 35–39
- 40–44
- ≥45
- Unreported or unknown

C. VERY LOW WEIGHT LIVE BIRTH (Birth Weight <1,500 g)

RACE
- Total
- White
- Black
- Others
- Unknown

MOTHER'S AGE
- ≤13
- 13
- 14
- 15
- 16
- 17
- 18
- 19
- 20
- 21
- 22–24
- 25–29
- 30–34
- 35–39
- 40–44
- ≥45
- Unreported or unknown

Table 2 Attributes of Death Data

DEATH

RACE
- Total
- White
- Black
- Others
- Unknown

AGE by race
- 0
- 1–9
- 10–19
- 20–29
- 30–39
- 40–49
- 50–59
- 60–69
- 70–79
- 80

AUTOPSY by race
- Yes
- No

BURIAL DISPOSITION by race
- Burial
- Cremation
- Donation
- Hospital disposition
- Removal
- Other
- Unreported

SEX by race
- Female
- Male

CAUSE OF DEATH by race
- Disease
- Non-disease

DISEASE

Arteriosclerosis

Benign and Unspecified Neoplasms

Cancer
- Bladder
- Brain and other nervous system
- Female breast cancer
- Male breast cancer
- Cervix uteri (female only)
- Colon and rectum
- Corpus uteri (female only)
- Esophagus
- Hodgkin's disease
- Kidney and renal pelvis
- Larynx
- Leukemia
- Liver and intrahepatic bile duct
- Lung and bronchus
- Melanoma of the skin
- Multiple myeloma
- Non-Hodgkin lymphoma
- Oral cavity and pharynx
- Ovary
- Pancreas
- Prostate
- Sarcoma
- Stomach
- Testis (male only)
- Thyroid

Cerebrovascular Disease

Certain Cond. Originating in Perinatal Period

Chronic Liver Disease and Cirrhosis

Chronic Obstructive Pulmonary and Allied Cond.

Congenital Anomalies

Diabetes

Disease of Heart
- Ischemic
- Others

Hernia and Intestinal Obstruction

Hypertension

Infectious and Parasitic
- HIV/AIDS

Meningitis

Nephritis, Nephrotic Syndrome, and Nephrosis

Pneumonia

All Other Diseases

NON-DISEASE

Unintentional Injuries
- Drowning
- Falls
- Fire and flames
- Firearms
- Motor vehicle accidents
- Poisoning by drugs and medicaments
- Railway
- Others

Homicide and Legal Intervention
- Assault by firearms
- Assault by cutting and piercing
- Others

Suicide
- Firearms
- Hanging, strangulation, and suffocation
- Poisoning
- Others

Other External Causes

Geocoding Process

Geocoding is the process of linking a common location identifier such as address, site location, or building to a spatial and geographic database, such as one with census TIGER/Line files, that contains the locations of streets, the ranges of addresses found on each street segment, and the boundaries of political and administrative areas. Because the geographic database contains address ranges defining the beginning and ending address numbers that were assigned to a given street segment, coordinates (i.e., latitude and longitude) for any specific address location can be found through a linear

Table 3 GDT Geocoding Summary Information

Variable	Definition	Notes
GDTPLUS4	4-digit zip code match	
GDTSAD	Street address match	
GDTCITY	City match	
GDTSTATE	State match	
GDTZIP	5-digit zip code match	
GDTSFIPS	State FIPS code match	
GDTCFIPS	County FIPS code match	
GDTTR90	Census tract match	
GDTBG90	Census block group match	
GDTXIN	Centroid type code	0 = Not a centroid (street address level match)
		1 = Zip + 4 centroid
		2 = Zip + 2 centroid
		X = 5 digit zip code centroid
		Blank = No centroid available
GDTSTAT	Match status code	B1 = Batch street address number match
		B2, B3 = Batch street intersection number match
		B5 = Batch matched to a street address on an alternate street name
		B6 = Batch matched to a placeholder
		B7 = Batch matched to a placeholder on an alternate street name
		10 = Not a valid 2-digit alpha state code
		11 = City in not found in the state
		12 = Incomplete or poorly formatted address
		14 = Could not find street name in the city
		15 = Could not match number, directional, or street type
		16 = Multiple addresses found in that range
		17 = Street intersection failure
		18 = City is present in the state but no named or addressed street are present

interpolation of the address number between the starting and ending address numbers assigned to the segment. Once the correct location is assigned, a location identifier is given a map coordinate and becomes a permanent geocode.

At SCDHEC, the statewide geocoding service is conducted by Geographic Data Technology (GDT). Data were address matched to the Dynamap/2000 series, which is a base map used and generated by GDT for the purpose of address matching (12). Table 3 shows the summary information on the geocoding process.

Address Standardization

To improve address accuracy and increase the geocoding match rate, StreetRite

software (Group 1 Software, Lanham, MD), is used to check and correct resident addresses. StreetRite compares the residual addresses against a database of every mailable address in the United States, deciphers inaccurate or incomplete addresses (e.g., misspelled street names and missing zip codes, cities, and states) and replaces them with the correct data. The addresses StreetRite is unable to match are then manually checked, and if an accurate match is found, the address is corrected.

Error Sources of Geocoding

Geocoding is a process of matching an address to a geographic location. The quality of the geocoding process is referred to as the geocoding match rate. An accurate geocoding match process depends on the quality of the address data and the geographic data. There are some errors inherent to the process, and in many cases it is difficult to determine how accurate the results are. It is important to document the potential error sources and understand how they could affect the quality and results of geocoding. The following are factors identified during our geocoding process that could affect the match rate.

Accuracy of Address

In geocoding vital health records data, the initial error is introduced when the individual or family providing information to the medical official are not made aware of the difference between mailing and residence addresses. Mailing addresses quite often are the post office box at the local post office, while the residence address is the street and street number at which the individual resides. New addresses created in the calendar year that do not exist in the current street/road database will also reduce the geocoding match rate.

Address Allocation

The geographic data used in the geocoding process contain a wealth of information about street locations, address ranges, and related information, but they are by no means complete. In urban areas, the percentage of street segments that contain address ranges may be as high as 90% or above. Some rural areas, however, do not contain any address ranges. Therefore, the geocoding matching rate will depend upon the study area.

Assigning Geographic Location

Geocoding is a comparison of each address in an event table with the address ranges in a target address database. When an event address matches the address range of a street segment, an interpolation is performed to locate and assign real-world coordinates to the event. For example, given a line with end point values of 0 and 100 and a street address of 50, the location of the address is estimated at the line's midpoint. The actual street address, however, may not be located at the midpoint of the line segment. During the aggregation process there is the potential for a small percentage of geocoded data to be captured in the wrong polygonal boundary. For instance, in Figure 2, Tract 1 will be assigned an erroneous value.

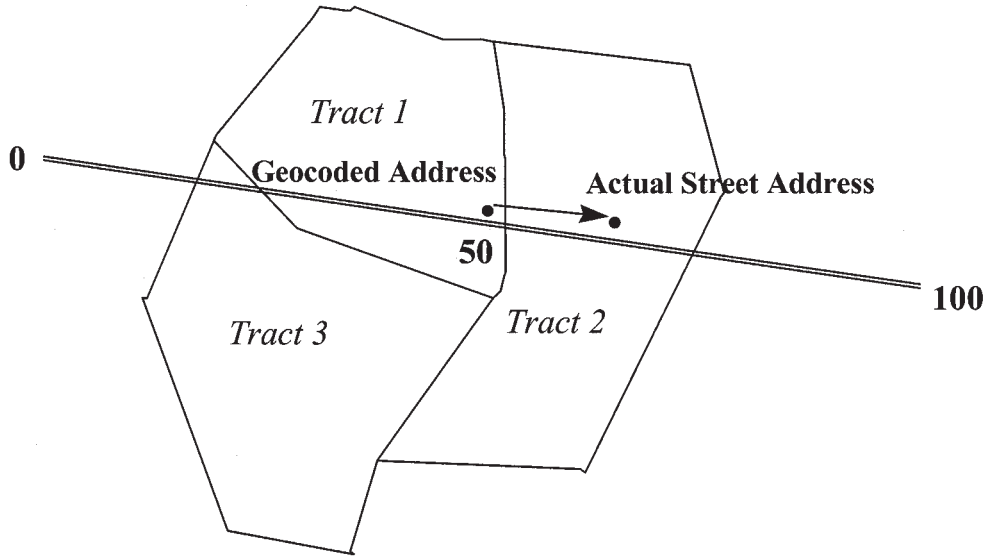


Figure 2 Illustration of potential error from assigning geographic location.

Automation System and Process

System Design

The primary goal of developing the VHCDIS is to aggregate vital health data to census tract level and generate publicly accessible database files. The system is designed to interact with users by generating aggregated information from different public health data. Figure 3 illustrates the general architecture of the VHCDIS. At this point, the system handles only birth and death records. In the future, as the need for geocoding health data increases, more components will be added to the system to handle different health information (such as cancer registry data).

System Resources and Implementation

The VHCDIS was developed using a GIS software, ARC/INFO (ESRI, Redlands, CA), on both Unix and NT platforms. A system supervisor in the form of an X-window graphical user interface (GUI) was used to provide user access to all the various components (birth data and death data) described previously. The GUI provides an interactive environment that facilitates user access to the components, as well as selection and execution of selected options. As described below, user navigation of the entire process is accomplished by appropriate selection from the window menu.

Table 4 summarizes the various steps involved in the automation process. The user first loads the system by opening and running the ARC/INFO software. At this point, the user is looking at the image shown in Figure 4. In this example, we will use the system to generate birth information. Therefore, the user can point and click on the Birth Certificate icon. As shown in Figure 5, twelve options are available for generating aggregated information on live births, low weight live births, and very low weight live births at the census tract level. The classification of each category is shown in Table 1.

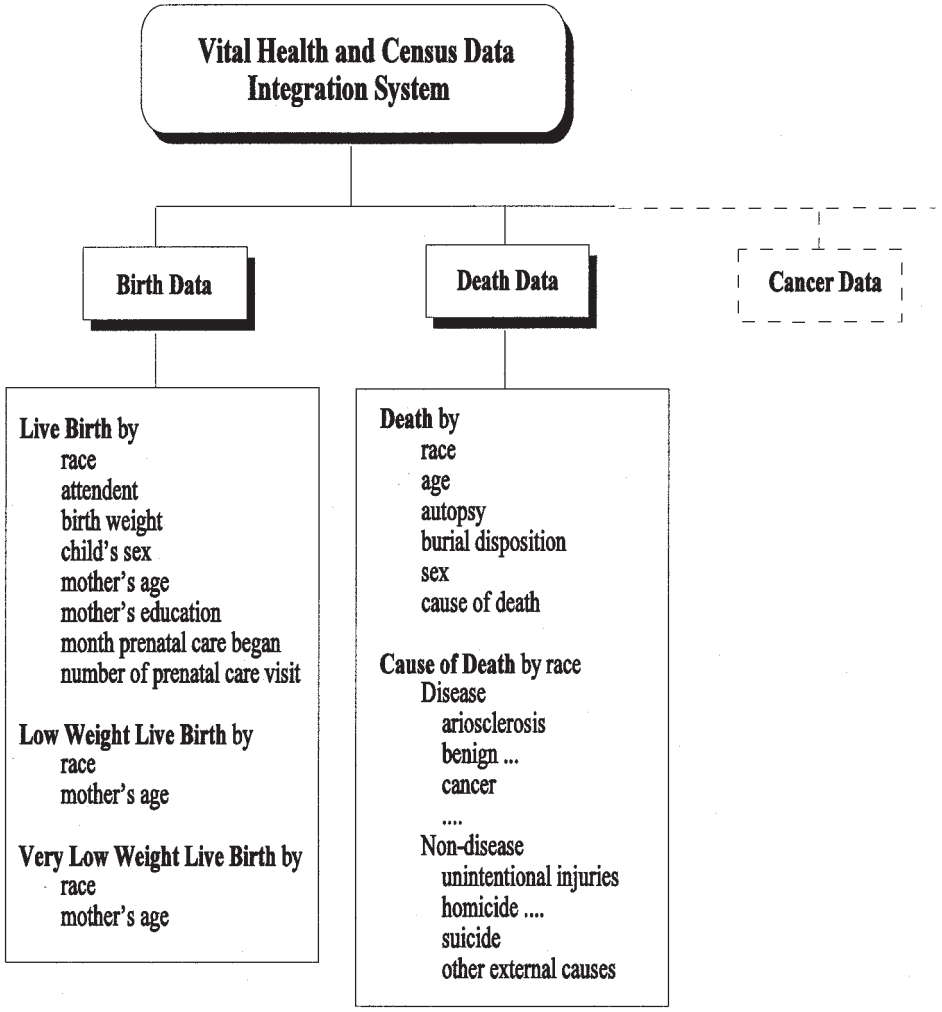


Figure 3 Architecture of the Vital Health and Census Data Integration System.

The user will make selections according to the information needed. For example, to generate live birth information by using the mother's age and race, the user will click the corresponding Select button to continue (Figure 6). In this menu, the user needs to supply two data files: birth data file (with all of the information shown in Table 1 plus the census county-tract number), and census tract data file (with census county-tract number only). When selecting a data file, a pop-up window will appear to help the user make the selection from existing data files.

After specifying the data file, the user can select field item names for each parameter (mother's age and mother's race from birth data; county-tract number from census data). The user should provide an output file name (e.g., lb95race.dbf) and define an item name for each classification shown in this menu. Because it is cumbersome to define the item names one by one, the user can select the USE DEFAULT button to use a

Table 4 Steps in Implementation of the Vital Health and Census Data Integration System

Component 1: Birth Data Aggregation
- Step 1. Select **Birth Certificate** icon
- Step 2. Select following variables accordingly:
  - Live births by mother's race
  - Live births by mother's age
  - Live births by mother's education
  - Live births by child's sex
  - Live births by prenatal care visits
  - Live births by prenatal care began
  - Live births by attendant at birth
  - Live births by birth weight
  - Low weight live births by mother's race
  - Low weight live births by mother's age
  - Very low weight live births by mother's race
  - Very low weight live births by mother's age
  - **Example:** select **live births by mother's age**
- Step 3. Select birth attribute INFO file: **test95.dat**
- Step 4. Select item name for mother's age: **MAGE**
- Step 5. Select item name for mother's race: **MRACE**
- Step 6. Select census tract INFO file: **test_cns.dat**
- Step 7. Select item name for County-Tract number: **CNTR**
- Step 8. Define output DBF file name: **lb95mage.dbf**
- Step 9. Define individual item name or use default setting by selecting **USE DEFAULT**
- Step 10. Select **DONE** to continue the process
- Step 11. Select **CANCEL** to return to previous menu
- Step 12. Select another variable

Component 2: Death Data Aggregation
- Step 1. Select **Death Certificate** icon
- Step 2. Select following variables accordingly:
  - Death by race
  - Death by age
  - Death by autopsy
  - Death by burial disposition
  - Death by sex
  - Death by cause of death
  - **Example:** select **death by age**
- Step 3. Select death attribute INFO file: **dth95.dat**
- Step 4. Select item name for age: **AGE**
- Step 5. Select item name for race: **RACE**
- Step 6. Select census tract INFO file: **test_cns.dat**
- Step 7. Select item name for County-Tract number: **CNTR**
- Step 8. Define output DBF file name: dth95age.dbf
- Step 9. Define individual item name or use default by selecting **USE DEFAULT**
- Step 10. Select **DONE** to continue the process
- Step 11. Select **CANCEL** to return to previous menu
- Step 12. Select another variable

default name for each item. After this, the user can click DONE to continue the generation process. When this process is finished, the user clicks CANCEL to return to the previous menu (Figure 5) to select other parameters for data aggregation. Outputs from the aggregation process are dBASE files (.dbf) that can be imported to other database software or ArcView (ESRI, Redlands, CA) to review.

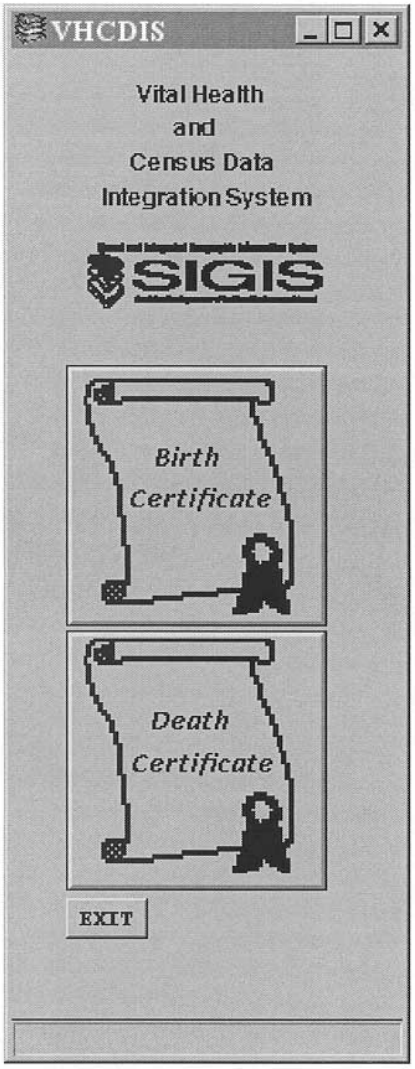


Figure 4 Main menu of the Vital Health and Census Data Integration System.

In this example, an output file (lb95race.dbf) was generated for live birth by mother's age and by race at the census tract level. To display the information, the user first runs the ArcView software, opens a graphic window, and adds the census tract coverage (or shape file) as a new theme. The user then needs to add the output file (lb95race.dbf) as a Table file and open the census tract attribute table. After opening the two tables, the user can perform the spatial join function to join both files and select different classified information to display (Figure 7).

Conclusion and Ongoing Process

GIS technology is emerging as a useful tool in public health studies. The technology

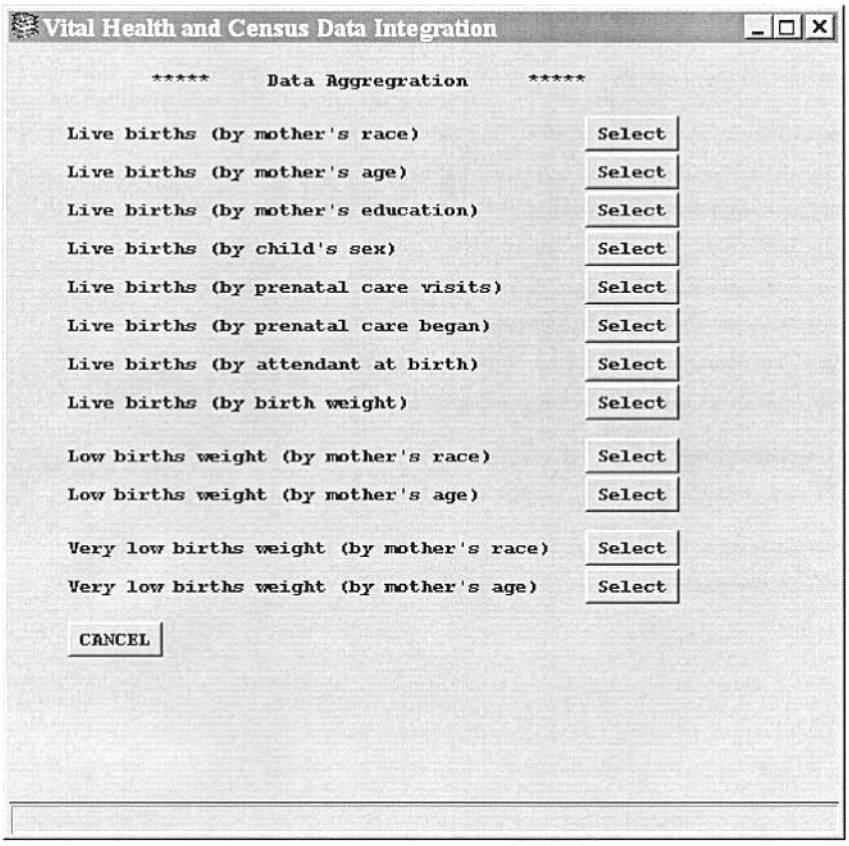


Figure 5 Birth information aggregation menu.

allows for storage and manipulation of large and multi-faceted datasets while maintaining the spatial integrity of each data collection or reporting location. As such, the technology gives rise to intensive investigation of the spatial relationship between variables and outcomes necessary to health risk assessment. Therefore, the key to successful application of GIS technology in the public health field is to understand what GIS functions should be used, what the limitations are, and how we should apply it appropriately to benefit research and assist officials with intervention strategies and health prevention.

In South Carolina, vital health data are collected each year. The need to access spatial health information is increasing as public health officials and researchers see the importance of analyzing spatial patterns of vital health data. Hence, creating a system to assist in standardizing data transformation from individual geocoded confidential health data to non-confidential data is needed.

This paper described an interactive data integration system, the Vital Health and Census Data Integration System (VHCDIS), developed and designed through the use of GIS technology for transforming geocoded confidential health data (birth and death) to non-confidential census health information. Vital health data were linked to census data through the geocoding process. By aggregating geocoded vital health data to

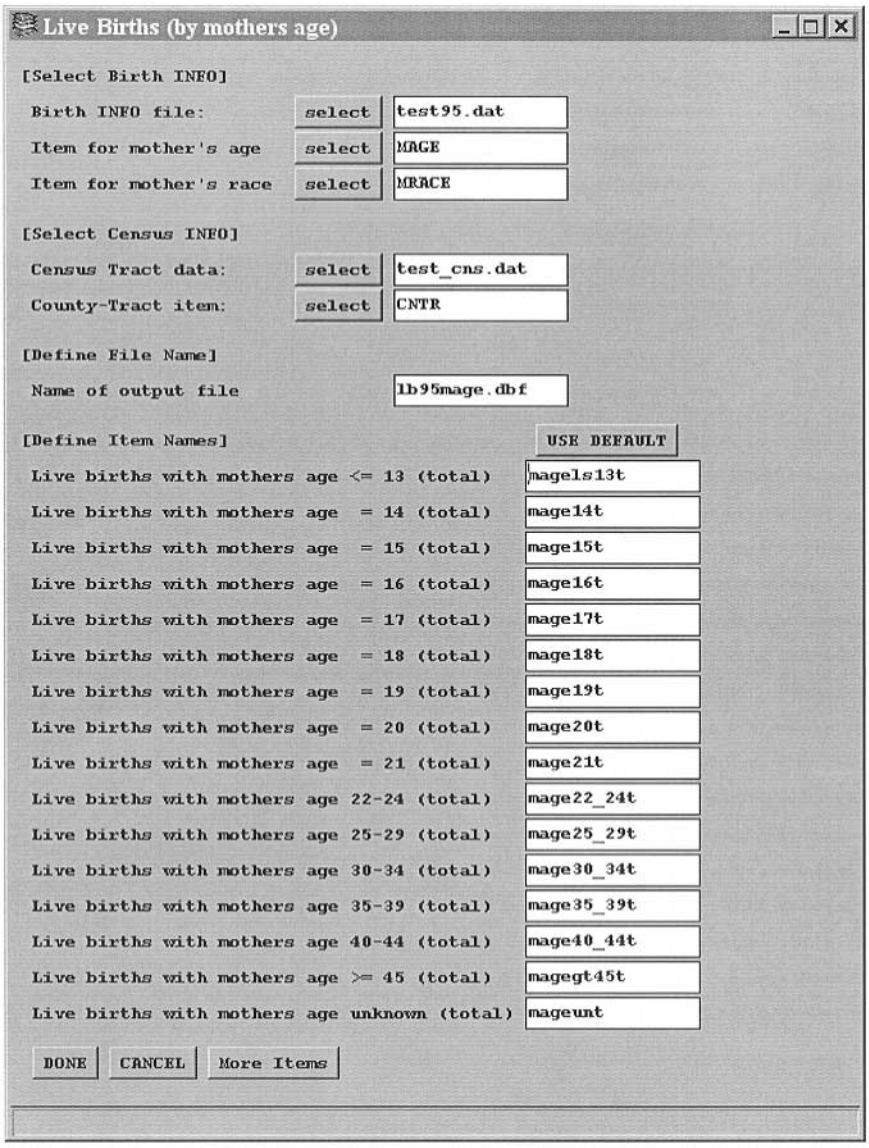


Figure 6 Generating information of live births by mother's age and by race.

census tracts, the output from VHCDIS will be publicly accessible and can be analyzed concurrently with other existing census socioeconomic data.

This report describes a project with the primary goal of developing a system to assist ongoing and systematic collection of health data and disseminate these data to public health officials and researchers for planning, implementing, and evaluating public health practice. Currently, SCDHEC is using the system to develop census health information from South Carolina birth and death data and will continue the process in the future. For the VHCDIS, many improvements and extensions are still underway. For

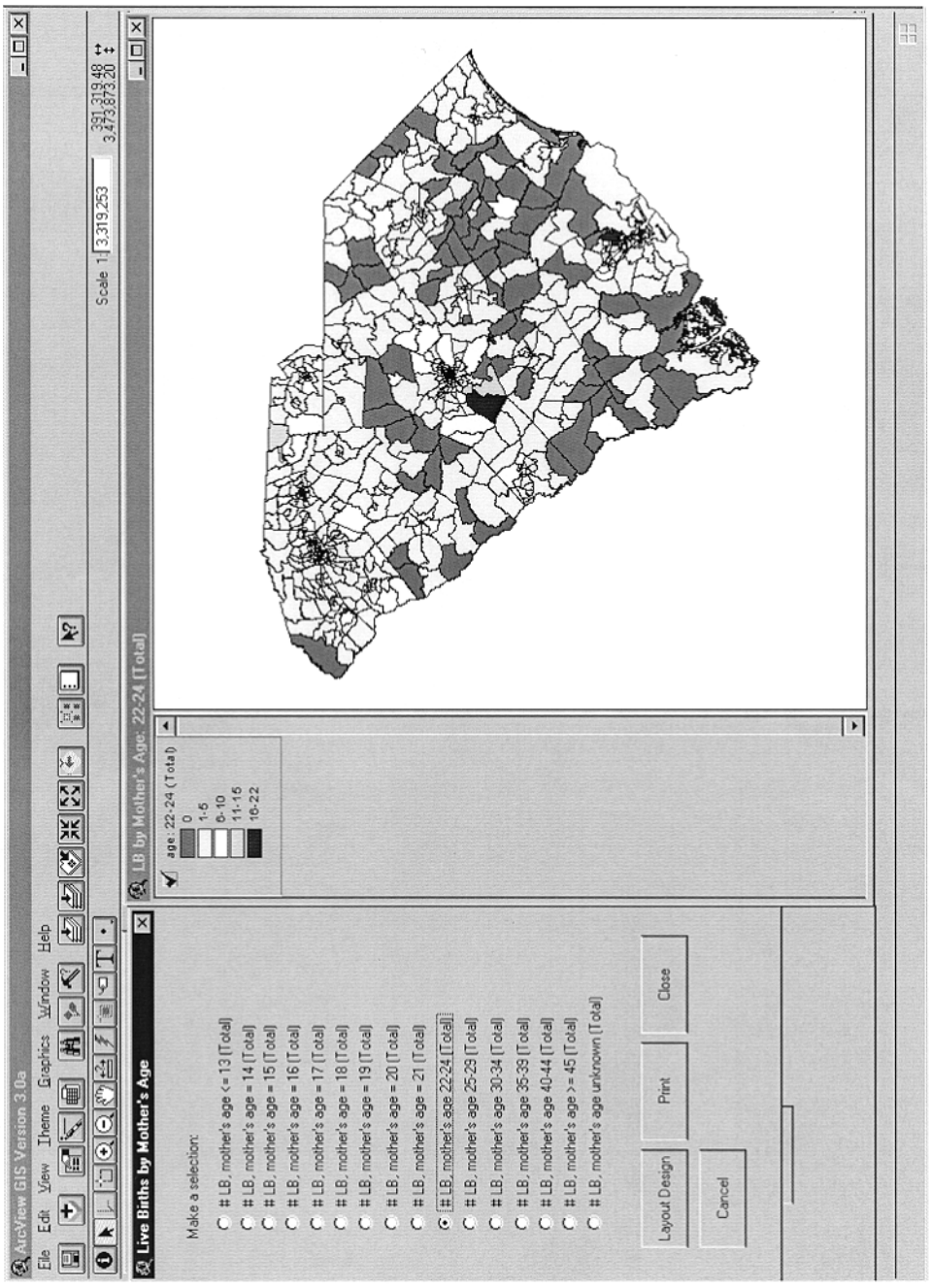


Figure 7 Output data display in ArcView for live births by mother's age and by race.

example, the system is being extended to accommodate infant death data and include cancer registry data next year. Additionally, the system can be extended to census block group levels and possibly to census block levels. In general, the VHCDIS in its present form is a sufficiently realistic demonstration of the flexibility of GIS technology and its ability to aggregate and process large volumes of health data.

References

1. Alpert S, Haynes KE. 1994. Privacy and the intersection of geographical information and intelligent transportation systems. In: *Proceedings of the Conference on Law and Information Policy for Spatial Database.* Tempe, AZ: National Center for Geographic Information and Analysis/Center for the Study of Law, Science, and Technology, Arizona State University College of Law. 198–211.
2. Croner MC, Sperling J, Broome FR. 1996. Geographic information system (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15:1961–77.
3. National Center for Health Statistics. 1995. *US vital statistics system: Major activities and developments, 1950–95*. Hyattsville, MD: US Department of Health and Human Services, Centers for Disease Control and Prevention. DHHS Publication No. (PHS) 97-1003.
4. Feinleib M. 1997. The use of computer mapping in monitoring the nation's health. In: *Proceedings from the International Symposium on Computer Mapping in Epidemiology and Environmental Health*. Geneva: World Health Organization. 1–3.
5. Pyle GF. 1994. Mapping tuberculosis in the Carolinas. *Sistema Terra* 3(1):22–3.
6. Hugh-Jones M. 1997. Applications of remote sensing to the identification of the habitats of parasites and disease vectors. *Parasitology Today* 5(8):244–51.
7. Malone JB, Fegker DP, Loyacano AF, Zukowski SH. 1992. Use of LANDSAT MSS imagery and soil type in geographic information system to assess site-specific risk of fascioliasis on red river basin farms in Louisiana. Reprinted from *Tropical Veterinary Medicine: Current Issues and Perspectives*. In: *Annals of the New York Academy of Sciences* 635:389–97.
8. Perry BD, Kruska R, Lessard P, Norval RAI, Kundert K. 1991. Geographic information systems for the development of tick-bone disease control strategies in Africa. *Preventive Medicine* 11:261–68.
9. Rogers DJ, Rabdikog SE. 1991. Mortality rates and population density of tsetse flies correlated with satellite imagery. *Nature* 351:739–41.
10. Centers for Disease Control and Prevention (CDC). 1988. *CDC surveillance update.* Atlanta: CDC.
11. World Health Organization (WHO). 1977–78. *Manual of the international statistical classification of diseases, injuries, and causes of death: based on the recommendations of the Ninth Revision Conference, 1975, and adopted by the Twenty-Ninth World Health Assembly.* Geneva: WHO.
12. Geographic Data Technology (GDT). 1997. *Dynamap/2000 7.2 user manual*. Lebanon, NH: GDT.

Geographic Information Analysis of Pediatric Lead Poisoning

Florence Lansana Margai, PhD*
Department of Geography, Binghamton University-SUNY, Binghamton, NY

Abstract

This study analyzes the spatial distribution of pediatric lead poisoning cases in relation to environmental indicators of lead, housing, and the demographic attributes of block groups in Binghamton, New York. Primary data on childhood blood lead levels are based on screening records from July 1991 to June 1995. Approximately 17% of all children tested within this period had elevated blood lead levels. A number of geographic information system (GIS) and statistical operations are used to determine (a) whether the distribution of lead poisoning cases reflects a consistent spatial pattern, and (b) the extent to which the pattern is linked to possible sources or pathways of exposure such as lead emitting facilities, major transportation corridors, trace lead in soil and municipal water supply, and housing. The results reveal clearly defined clusters of lead poisoning cases along transportation lines within the urbanized and industrialized zones. Specifically, block groups in the central city that were characterized by old, subdivided, and rented properties and poverty had proportionately higher incidences than others. Nearly six out of every ten cases fell within these clusters. These results demonstrate how comprehensive health and environmental data can serve as input in delineating high-risk areas for lead monitoring and remediation programs.

Keywords: pediatric blood lead levels, lead poisoning, GIS statistics, canonical correlation

Introduction

Lead poisoning is one of the most significant pediatric environmental health hazards in the United States, yet it is one of the most preventable as well. Over the years, several steps have been taken at different fronts to minimize the risks associated with this hazard. The enactment of the Clean Air Act in the 1970s and subsequent federal regulations have led to a more than 90% reduction in atmospheric lead levels. Unfortunately, this metal continues to pose a significant health threat from pre-existing sources such as lead paint and dust in older housing, industrial emissions, and contaminated soils. Several studies conducted over the last two decades have consistently identified neurological and developmental problems in children exposed to lead, even those exposed to levels once considered to be harmless (1,2). Estimates based on the standards set by the Center for Disease Control (CDC) suggest that approximately 10 million children are at risk (3). This is particularly severe in the inner cities, where more than 60% of low income and minority children are believed to have elevated blood lead levels of 10 or more micrograms per deciliter (μg/dL). Explanations for the observed spatial patterns have been linked to historical patterns of urbanization, transportation, and industrialization (4).

* Florence M. Lansana Margai, Department of Geography, Binghamton University-SUNY, Binghamton, NY 13902 USA; (p) 607-777-6731; (f) 607-777-6456; E-mail: Lansana@Binghamton.edu

One of the primary reasons for conducting this study was to explore the use of geographic information systems (GIS) in developing a pediatric lead prevention program for the city of Binghamton, New York. Specifically, the following research questions were examined:

1. What is the spatial distribution of elevated blood lead incidences among children? Is it random or spatially clustered?
2. Is there a significant relationship between the observed distribution of child elevated blood lead incidences and the demographic attributes of residents?
3. Is there a relationship between the observed distribution of child elevated blood lead incidences and significant sources and pathways for environmental lead, such as paint, soil and water quality, transportation corridors, and lead-related businesses and industries?
4. Where are the high-risk areas and what is the density of preschool children within the proximity of these sites?
5. What level of monitoring is necessary to reduce the risks of lead poisoning?

Addressing these questions required the assistance of a GIS and spatial statistics. Using data from different sources, we were able to identify spatial patterns in elevated blood lead incidence cases by block group, inventory the multiple sources of lead contamination, and statistically evaluate the underlying relationships between lead poisoning and various indicators of environmental lead.

Developing a GIS for Pediatric Lead Poisoning Prevention

Recent trends show a growing number of lead monitoring and prevention programs based on guidelines established by the CDC (5,6). Children aged six months through six years are screened regularly at well-child visits. Families of children with elevated lead levels are given prompt medical attention coupled with residential testing to identify and possibly eliminate the source of lead. While these efforts are successful in curbing the rate of lead poisoning, there are still some problems. For example, even though the CDC recommends universal screening, not all children are being tested. Even among those who are tested, no effort is made to educate the parents about the dangers of lead poisoning unless the test results are positive. As Wartenberg (7) argues, such an approach not only hinders the primary prevention efforts but the regional assessment of risks as well. Neighbors may not be readily identified and geographic clusters of highly exposed individuals are likely to be missed.

Children are exposed to lead from multiple sources. Therefore, a major step toward the development of an efficient lead monitoring and prevention plan must start with a comprehensive database that includes not only the screening records of children but ancillary data on industrial emissions, transportation lines, housing characteristics, occupational exposure patterns, and other parameters. The use of a GIS can facilitate this process. A GIS is essentially a collection of computer hardware and software that can be used to capture, store, retrieve, analyze, and visualize various forms of spatial data. It is a very powerful tool for understanding the spatial linkages between multiple layers of human and natural phenomena and for isolating possible cause-and-effect relationships. On matters relating to public health, it can serve as a modeling tool for spatial epidemiological patterns as well

as an analytical tool for testing hypotheses regarding mapped distributions of disease (8).

The methodology for compiling various sources of lead toxicity into a GIS is still in its infancy, however. Few researchers have fully explored this technology as a reliable means of identifying lead exposure patterns and high-risk areas (4,7,9,10). Some of the applications have been exploratory, with limited data used to map exposure patterns at coarse and sometimes inappropriate spatial levels such as zip codes, minor civil divisions (MCDs), or census tracts. It is important, however, to go beyond this exploratory stage and fully utilize the analytical and predictive functions of GIS in discerning potential risk areas.

The Study Area

Binghamton extends across 11 square miles with approximately 53,000 people. This city grew out of an extensive industrial heritage. Some of the businesses with roots in this area include International Business Machines (IBM), Endicott Johnson Shoe Corp., General Electric, Universal Instruments, Ozalid, Link Federal Systems, and Anitec. The rise of these companies brought an industrial boom that required extensive road and rail systems for transportation as well as mass inexpensive housing. Today, Binghamton is criss-crossed by the Canadian-Pacific railway system, the railroad network in northeastern United States, and at least four major highways: US 11, NY 17, Interstate 81, and Interstate 88.

Binghamton is the most urbanized area in the two-county southern tier of New York State. About 88% of the population is white, with minorities constituting the remaining 12%. Like several other northern US cities, the city has experienced population and economic declines. Several industries have closed or downsized their workforce to cope with economic difficulties. There has also been a drop in downtown activities and service functions due to suburbanization and the building of shopping centers and strip malls in the outlying areas. The city today reflects the characteristic patterns of urban decay. Many of the old buildings and homes have been renovated into multiple housing units to serve low-income residents and college students. Within the two-county area, most of the government subsidized housing units are located in the city. Higher income housing and households are located in the outskirts of the city, particularly in the far west and southwestern sides.

Data Collection

The primary data on pediatric blood lead levels consisted of 1,840 records of children tested between 1991 and 1995. The screening tests were typically for children between one and three years, although a small percentage (6%) of the affected children were five or six years old. Previous studies have indicated that lead is prevalent among children in these age groups primarily because of the increased hand to mouth activities.

Initial assessment of the data suggested that at least 17% of the children had high blood levels. Based on the CDC guidelines, almost two-thirds of the children were classified at Level IIa, with a blood lead level concentration of 10–14.9 μg/dL. About 24% of the children are classified as Level IIb, with blood lead levels of 15–19 μg/dL. Another 16% of the cases are categorized as Level III, with blood lead levels of 20–44.9

µg/dL. Fewer than 1% of the children are in Level VI, the highest category observed in the community, with blood lead levels of 45–69 µg/dL. The majority of the children are white, reflecting the racial composition of the population in the city. Based on the addresses, the data were geocoded and exported into a desktop mapping package, MAPINFO, for cartographic analysis.

Data Analysis and Results

The Spatial Distribution of Lead Poisoning Cases

The spatial distribution of lead cases was evaluated by two methods. First, a buffer analysis was performed using the MAPINFO software. The criterion used for detecting spatial clustering was proximity. A lead cluster was inferred in every area where there were four or more confirmed cases of lead poisoning within a 500 foot radius. The results were then validated using cluster analysis within the statistical software, SPSS. Figure 1a shows the location of these clusters at the block group level. About 57% of the lead poisoning cases fell within 13 defined clusters. Most of the clusters were located close to the center of the city and along the transportation lines. The relationships between these lead clusters and the potential sources of exposure were subsequently assessed using a number of statistical procedures.

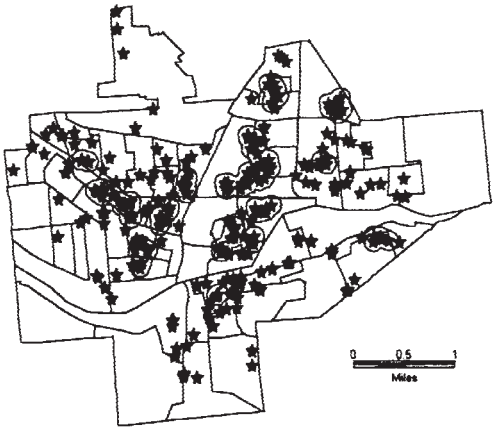


Figure 1a High blood lead occurrences.

Housing and Socioeconomic Correlates of Lead Poisoning

The source of lead largely depends on the characteristics of the community in which a child resides. Previous studies have reported high lead levels in areas of physical deterioration with old, subdivided housing, poverty, and areas of minority concentration. In an attempt to identify such areas in Binghamton, 12 demographic variables were selected using US Census data from STF3A files. A correlation analysis was then performed between the variables and the rate of blood lead incidences within each block group. Among the 12 variables, 9 were significantly related to pediatric lead poisoning cases in Binghamton. The strongest indicators were poverty and housing quality

variables such as block groups with pre-1940s housing, rented property, and subdivided units. Initially, no associations were observed between the lead cases and areas with significant minority and family composition. However, further analysis using only the lead cases that fell within defined clusters suggested a possible link with the proportion of African Americans in the community ($r=0.41$; $p<0.01$).

Environmental Sources of Lead Poisoning

Different sources of data characterizing lead-emitting businesses and industries in the city, automobile-related facilities, and transportation corridors were incorporated into the GIS. First, historical data consisting of all business locations within Binghamton since 1890 were queried. Businesses that qualified for entry into the analysis included those that used lead or lead by-products in their activities. These included factories such as machine shops, foundries, and parts and glass manufacturers. Using the location of these facilities, a 500 foot buffer was established as a reasonable distance over which the airborne effects of lead would be dispersed on land. These buffers covered about 20% of the city's areal extent and about 41% of the confirmed cases fell within the defined buffer (Figure 1b). Further spatial analysis involved subdividing the buffer into smaller polygons with boundaries corresponding to the block group boundaries. The area of each buffered portion was then divided by the total area of the corresponding block group to determine the degree to which each block group was characterized by lead-associated industries or businesses. This newly created variable showed a highly significant relationship with child lead poisoning rates ($r=0.61$; $p<0.05$).

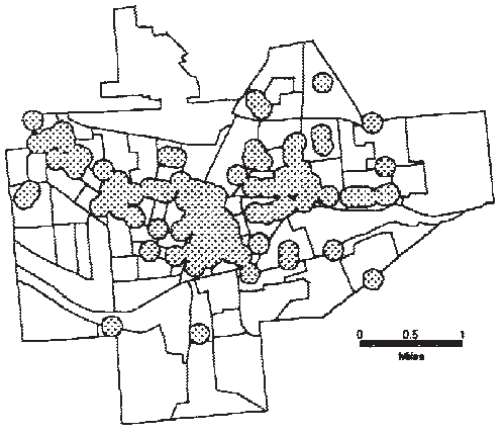


Figure 1b Lead emitting/handling businesses.

Using procedures similar to those explained above, a buffer was established for automobile-related facilities such as gas stations, repair shops, dealerships, and junkyards. About 53% of all confined cases fell within these areas (Figure 1c). The relationship between the buffered variable and lead poisoning cases was also significant ($r=0.5$; $p<0.05$). Buffers were also created around major roads and railways in Binghamton (Figure 1d). The lead cases were significantly related to both the road buffers ($r=0.30$; $p<0.05$) and the railroad buffers ($r=0.46$; $p<0.05$).

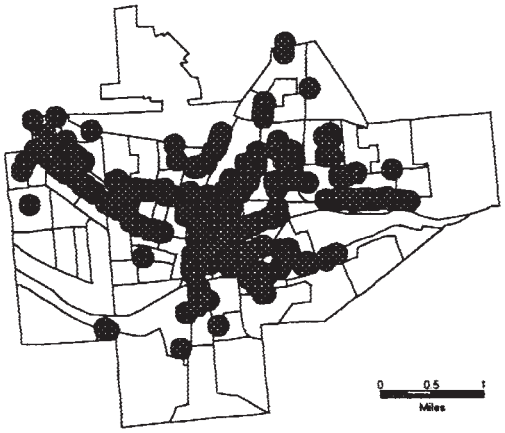


Figure 1c Gas storage/auto-related sites.

Figure 1d Rail corridors.

Environmental Pathways

The visual information provided in Figure 1e reflects the spatial variability of pediatric lead poisoning relative to the industrial locations, automobile-related sites, and rail corridors. However, the use of a GIS in lead monitoring and prevention must include not only the potential exposure sources illustrated in the map, but also the pathways that are likely to directly affect the children. Specifically, contaminated soil and water are major pathways for young children. In Binghamton, the threat of trace lead in the water supply had been minimized by the preventive steps taken earlier by the city to minimize the corrosion from pipes. Certain polyphosphate compounds (with commercial brand names such as Aquamag and Calciquest) that bond to water pipes were first added to the municipal water supply in 1992. The impact of these additives was assessed by examining lead poisoning incidences before and after the changes were implemented. While the total number of reported cases varied from year to year, the mean blood lead levels in children declined consistently over time from 15.86 μg/dL in 1991/1992 to 13.72 μg/dL in 1994/1995.

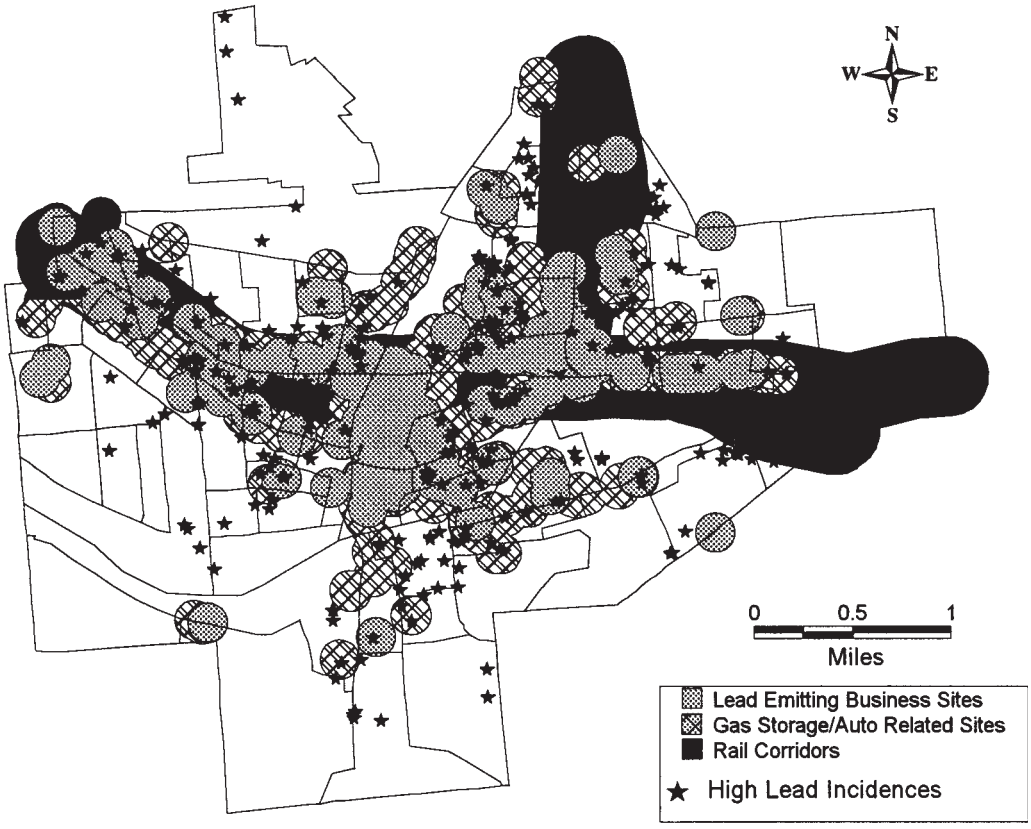


Figure 1e High lead incidences in all three impact zones.

The effect of soil lead on childhood blood lead levels was also evaluated by conducting a detailed soil sampling analysis in July 1996. Soil samples were extracted from the front, rear, and side yards of homes within the lead clusters identified earlier in the study, as well as outside of the clusters. The latter served as the control group. The soil samples were tested in a professional laboratory using the EPA 6010 Method. The results were later integrated with the existing data in the GIS. A query was then performed to identify all sites that exceeded the EPA standards of 500 parts per million. Several of the sites were above the EPA standards for lead in the soil. Lead levels within the clusters found in the central city were two or more times higher than the EPA standards. Statistical analysis of the mean differences confirmed that soil lead levels were significantly higher in the clustered areas than outside of them ($t=3.66$; $p<0.05$). However, the relationship between soil lead and childhood blood lead levels was not significant.

Delineation of Target Communities for Lead Prevention

The final objective in this study was to delineate the high-risk areas based on the comprehensive database described above. This was accomplished by using canonical correlation analysis to quantify the associations between the major lead indicators and

then develop scores that would provide the most explanation for the spatial occurrence of the observed lead clusters. Variable selection for this phase was based on statistical significance from the preceding analyses. Two sets of variables were used. The first set consisted of the three variables that measured the effects of railroads, businesses, and automobile-related sites. The second set included the six demographic variables that were best associated with the location of lead poisoning cases. These variables were entered into the canonical correlation procedure.

At the conclusion of the statistical analysis, the canonical coefficient that maximized the linear relationship between the two sets of variables was selected. This coefficient was very high (r=0.83), implying that at least 80% of the observed lead cases could be jointly explained by these variables. Pairs of canonical correlation scores, representing the aggregate values of the two sets of variables, were also obtained for each block group. Scores with values greater than zero were classified as HIGH and those with values less than zero were classified as LOW. The results were then mapped to visualize the spatial relationships (Figure 2). As expected, block groups with high scores were found mainly in the center of the city and along the rail corridor going westward. Those

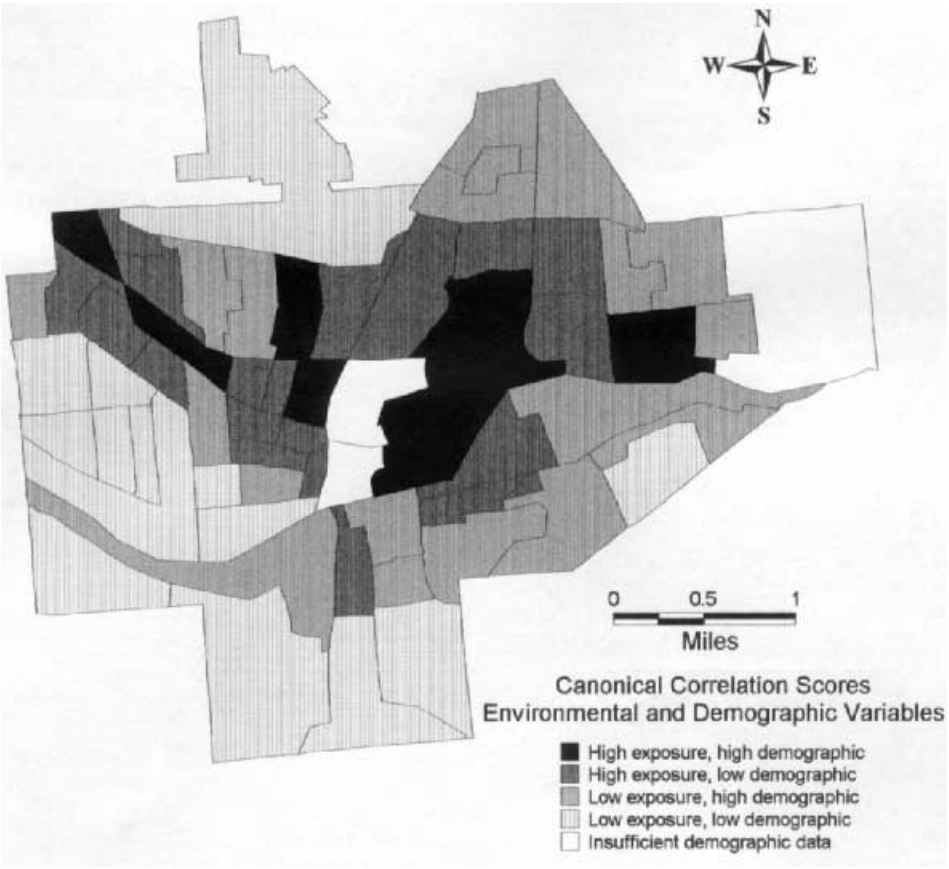


Figure 2 High-risk areas for lead poisoning in Binghamton, NY; canonical correlation results by block groups.

with low scores on both groups of variables were along the outskirts of the city, almost forming a continuous ring. A few block groups showed dissonant canonical correlation scores (high/low or low/high), but overall, the results showed that among the 26 block groups with high canonical scores on both variables, 191 cases of lead poisoning were reported—a rate of 7.3 occurrences per block group. Among the 29 with low scores on both sets of variables, about 85 cases were found with a rate of 2.9 cases per block group. These results confirm that lead poisoning cases are closely grouped in space and not merely random occurrences. Furthermore, these findings demonstrate the strength of two sets of variables in identifying high-risk areas.

Conclusions

In building on the strengths of previous applications, this study has examined the spatial patterns of lead poisoning and, within the context of environmental and demographic variables, isolated the high-risk areas for lead intervention programs. All of these significant steps were made possible through the use of GIS and statistical analysis. Several steps were involved, starting from data collection, storage, analysis, and visualization. Obviously one bottleneck in the process was the constant shift of data back and forth from one software to the other. This is not unusual, however, and future developments in the GIS packages are likely to provide improvements in these statistical procedures. Aside from these minor impediments, the technology provides a good basis for handling spatial epidemiological problems.

Acknowledgments

The contributors to this research were John W Frazier, Steve Walter, and Ron Brink. Assistance with soil sampling analysis was provided by the McNair Scholars Program at Binghamton University. An earlier and more detailed discussion of the study was published in the *Applied Geographic Studies Journal*, Vol.1, No. 4, pp. 253–270.

References

1. Mushak P. 1992. Defining lead as premiere environmental health issue for children in America: Criteria and their quantitative application. *Environmental Research* 59:281–309.
2. Sciarillo WG, Alexander G, Farrell KP. 1992. Lead exposure and child behavior. *American Journal of Public Health* 82:1356–60.
3. Sargent JD, Braun MJ, Freeman JL, Bailey A, Goldman D, Freeman DH. 1995. Childhood lead poisoning in Massachusetts communities: Its association with sociodemographic and housing characteristics. *American Journal of Public Health* 85:528–34.
4. Bailey AJ, Sargent JD, Goodman DC, Freeman J, Brown MJ. 1994. Poisoned landscapes: The epidemiology of environmental lead exposure in Massachusetts children, 1990–1991. *Social Science Medicine* 39:757–66.
5. Agency for Toxic Substances and Disease Registry (ATSDR). 1988. *The nature and extent of lead poisoning in children in the United States: A report to Congress.* Atlanta: US Department of Health and Human Services. Pub. 99–2966.
6. Centers for Disease Control. 1991. *Preventing lead poisoning in young children: A statement by the Center for Disease Control.* Atlanta: US Department of Health and Human Services. October.

7. Wartenberg D. 1992. Screening for lead exposure using Geographic Information System. *Environmental Research* 59:310–17.
8. Cliff AD, Haggett P. 1996. The impact of GIS on epidemiological mapping and modeling. In: *Spatial analysis: Modeling in a GIS environment.* Ed. by P Longley, M Batty. Cambridge, UK: GeoInformation International. 321–43.
9. Guthe WG, Tucker RK, Murphy EA, England R, Stevenson E., Luckhardt JC. 1992. *Environmental Research* 59:318–25.
10. Griffith DA, Doyle PG, Wheeler DC, Johnson JL. 1998. A tale of two swaths: Urban childhood blood-lead levels across Syracuse, New York. *Annals of the Association of American Geographers* 88(4):640–65.

A GIS Analysis of Industrial Pollution in Hartford, Illinois

Richard T Masse, MPH
University of Illinois at Springfield, Springfield, IL

Abstract

This project used geographic information system (GIS) software to create an additional tool for the assessment of industrial pollution by oil refineries near residential areas in southwestern Illinois. The study site was in Hartford, Illinois, which has a population of approximately 1,700. Indoor air, groundwater, and soil gas data from the site were obtained from state agencies and were used to generate point, polygon, and contour coverages of the study area using ArcView 3.0a and ArcView Spatial Analyst software. The resulting coverages allow investigators to assess and monitor a variety of environmental data with a new visual component. Some of the advantages of this geographical tool include corroboration of residential complaints, indication of high-risk areas, assessment of remediation actions, and validation of the need for further testing. Ultimately, this project demonstrates another way that GIS software can be used to enhance the effectiveness of environmental and public health investigations.

Keywords: environmental health, spatial analysis, ArcView 3.0a, industrial pollution

Introduction

In March of 1990 the Illinois Department of Public Health (IDPH) published a preliminary health assessment of an area in southwestern Illinois. IDPH performed this after receiving complaints from residents about the presence of gas fumes and incidences of fires in their houses. Residents voiced additional concerns about symptoms such as breathing difficulties, skin rashes and lesions, bloody noses, headaches, and exhaustion (1).

The IDPH assessment concluded that high levels of petroleum products had contaminated local aquifers, soil, and air quality and that the contamination was caused by three refinery operations in close proximity to residential areas (Figure 1). A local engineering firm, Mathes & Associates, was contracted to further assess contaminated areas and implement a remediation plan for cleaning up the existing contamination and preventing future contamination. IDPH continued surveillance in the area and received complaints again from residents in 1996. Following complaints from residents in Hartford, Illinois, IDPH collected indoor air samples from eight houses four times over a two-year period.

The purpose of this project was to take the IDPH data from the eight houses in the study area, as well as the pre-existing environmental data, and create electronic geographic coverages using geographic information systems (GIS) to aid in further analysis of the study area using a new geographical component.

[1] Richard T Masse, University of Illinois at Springfield, 1052 W. Fayette, Springfield, IL 62704 USA; (p) 217-698-1449; E-mail: rmasse@eosinc.com

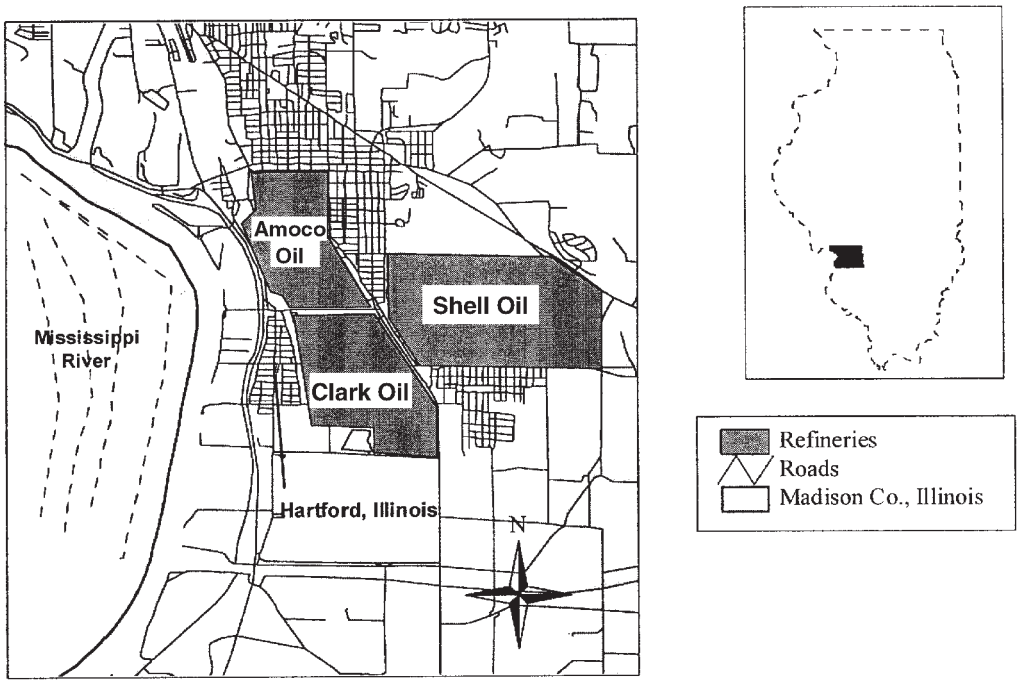


Figure 1 Hartford, IL, study site.

Data and Data Collection

The data selected to be GIS-coded for this project included demographic, air, groundwater, and soil gas data of the study area. These data came from a variety of sources including IDPH, the Illinois Environmental Protection Agency (IEPA), the Illinois Department of Natural Resources (IDNR), and the Mathes & Associates engineering firm in St. Louis, Missouri.

All of the usual demographic data of the study area were obtained, but the demographic data of particular GIS concern were the addresses of Hartford residents who complained to IDPH and had indoor air samples taken in 1996. IDPH would have liked to obtain more samples from the area, but was limited by funding and was also denied access to some buildings by reluctant citizens (2).

As mentioned previously, the indoor air sample data corresponded with complaints made by the residents. Upon receiving complaints from residents, IDPH performed indoor air and environmental sampling in each of the concerned households. Environmental data included temperature, humidity, and carbon dioxide levels. While these data are important to the overall assessment, they were not chosen for GIS conversion. The indoor air data chosen for GIS conversion were obtained by 24-hour sampling with SUMMA cans. These samples were then sent to a private lab, analyzed for over 50 compounds, and summarized in a spreadsheet by IDPH.

Groundwater data were obtained from two sources—IEPA and Mathes & Associates. There were 49 groundwater monitoring wells in the study area around Hartford. The wells were of various depths and included both private and public wells. Groundwater sampling was performed by measuring both water levels and

hydrocarbon levels to assess the thickness of petroleum products on top of the water. The most recent groundwater data were collected from the 49 wells at five different times during 1990 (3).

One set of soil gas data was collected by Mathes & Associates in 1990. A hydraulic probe unit was used to drive and withdraw soil-gas sampling probes at 14 different locations. Samples were collected at a depth of 7 to 34 feet by a vacuum pump used to pull 1 to 5 liters of air from the ground into a collection bulb. A syringe was used to withdraw soil gas that was then injected directly into a gas chromatograph for analysis.

Supplemental data important to this project included existing GIS data for the state of Illinois. These data were contained on a two-CD set distributed by IDNR. Examples of coverages on these CDs include highways, railroad, stream, and county data.

Creation of GIS Coverages

The data for this project were in many forms and of varying thoroughness. It was the goal of this project to take all the data and create foundation GIS files that could be used for analysis, but also supplemented if future data were collected. This involved the conversion of the data from the original sources into a common software. The two software programs used for this project were Quattro Pro and ArcView 3.0a.

The demographic data were first entered into a Quattro Pro database and saved as a text file. This text file was then transferred into ArcView 3.0a using its import function. From the newly created ArcView table containing study site addresses, a point coverage was created using the ArcView function of geocoding. This function adds point locations to the map based on street addresses (4). This is the software's equivalent of pushing pins into a street map on a wall. To locate the study site addresses it was necessary to have a street coverage of Hartford that included address information. This coverage was obtained from IDNR. The final result was a point coverage showing the locations of residents who filed complaints (Figure 2).

The original indoor air data were already in Quattro Pro format but contained information for 50 different compounds. This database was reduced to contain 11 compounds selected by IDPH. The new database then had 11 compound levels for 8 houses in the study area. This database was then imported into ArcView 3.0a and linked with the previously described address point coverage. This was accomplished by using a join function in ArcView 3.0a that allows tables with similar column values to be combined into one table. The final product of the indoor air data conversion is a point coverage similar to the address coverage above but with a different data table containing the compound levels for each house (Figure 3).

Although the previous coverages described originated from a spreadsheet database, this is not always necessary. Coverages can be created directly in ArcView 3.0a by using a mouse to place points or draw lines based on an existing paper map. When a point, line, or polygon is manually placed using a mouse, a table for holding information about that addition is created. The groundwater data were coded for GIS using this technique. First, well locations from a paper map were manually placed on the computer screen and then the accompanying table was filled with data such as well identification numbers and sampling data from the Mathes & Associates reports. The final product was a point coverage of monitoring wells accompanied by a table containing petroleum thickness for each well at different points in time (Figure 4). The soil gas data

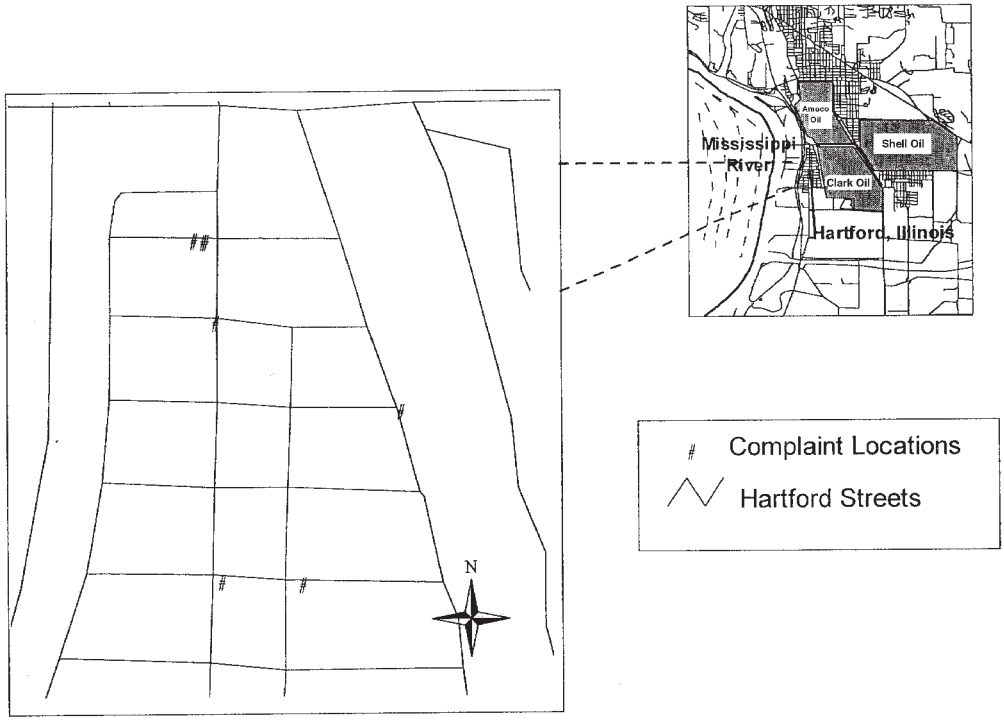
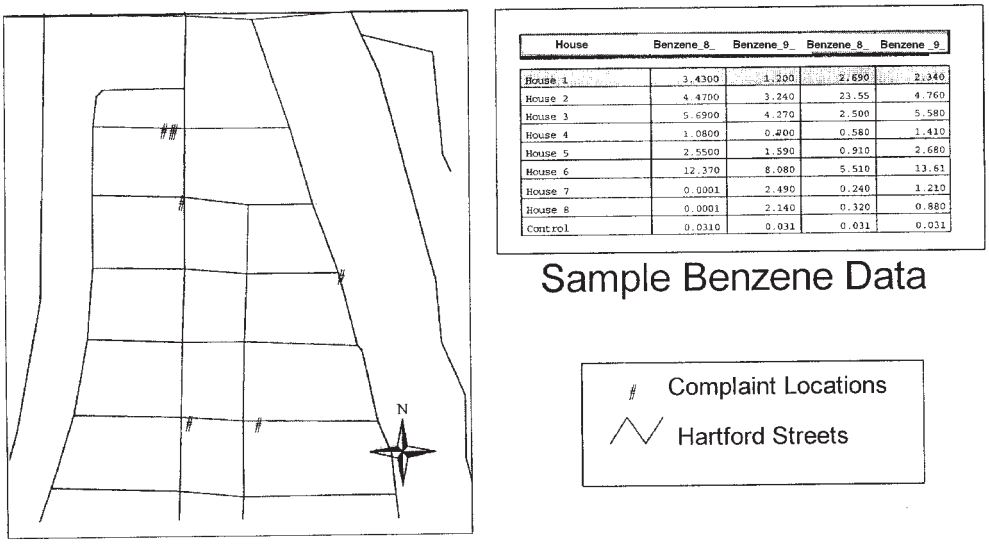


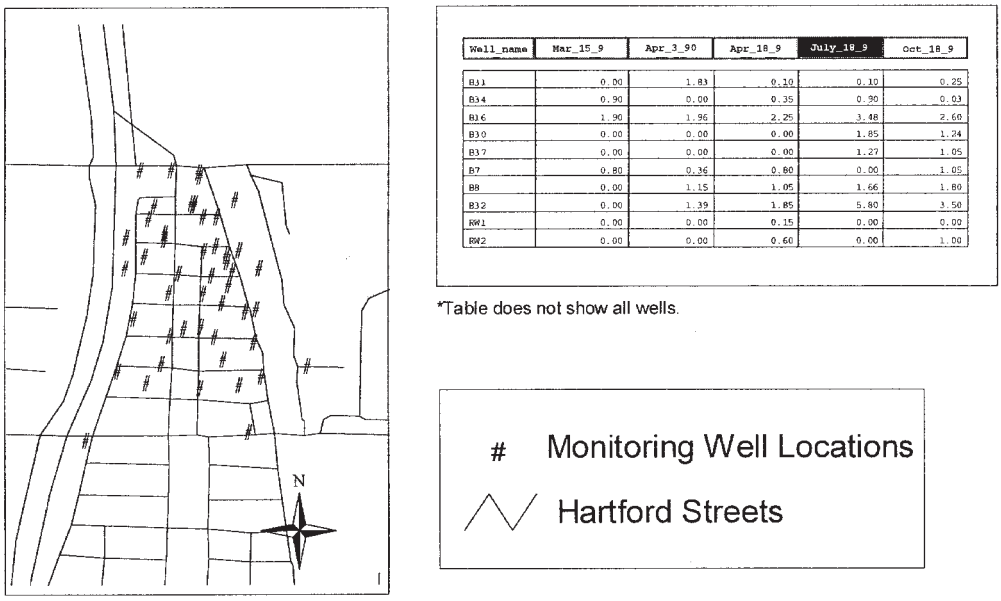
Figure 2 Point coverage of Hartford residents who filed complaints with the IDPH.



House	Benzene_8_	Benzene_9_	Benzene_8_	Benzene_9_
House 1	3.4300	1.200	2.690	2.340
House 2	4.4700	3.240	23.55	4.760
House 3	5.6900	4.270	2.500	5.580
House 4	1.0800	0.400	0.580	1.410
House 5	2.5500	1.590	0.910	2.680
House 6	12.370	8.080	5.510	13.61
House 7	0.0001	2.490	0.240	1.210
House 8	0.0001	2.140	0.320	0.880
Control	0.0310	0.031	0.031	0.031

Sample Benzene Data

Figure 3 Point coverage of indoor air sample locations, with supplemental table containing compound data for each house and the controls.



Well_name	Mar_15_9	Apr_3_90	Apr_18_9	July_18_9	Oct_18_9
B31	0.00	1.83	0.10	0.10	0.25
B34	0.90	0.00	0.35	0.90	0.03
B36	1.90	1.96	2.25	3.48	2.60
B30	0.00	0.00	0.00	1.85	1.24
B37	0.00	0.00	0.00	1.27	1.05
B7	0.80	0.36	0.80	0.00	1.05
B8	0.00	1.15	1.05	1.66	1.80
B32	0.00	1.39	1.85	5.80	3.50
RW1	0.00	0.00	0.15	0.00	0.00
RW2	0.00	0.00	0.60	0.00	1.00

*Table does not show all wells.

Figure 4 Point coverage showing monitoring well location and supplemental table with hydrocarbon thickness data.

were converted using the same technique and resulted in a point coverage of the 14 sampling sites accompanied by a table with the corresponding data.

The point coverages created were useful for logging and searching for data concerning specific spots in the study site, but not for analyzing any spatial relationships between the individual points. A fairly new add-on software called ArcView Spatial Analyst, created for ArcView 3.0a, was used to address these issues. Specific uses to this project included the creation of hydrocarbon plume coverages to overlay indoor air and address point coverages described earlier. This type of coverage allows investigators to see the different thickness layers of petroleum products under the study site. The spatial analyst software creates such coverages by interpolating contours based on point coverage data. This specific software provided four interpolation methods, and the one chosen for this project was Spline interpolation. Spline interpolation is a general purpose interpolation method that fits a minimum-curvature surface through the input points. Conceptually, it is like bending a sheet of rubber to pass through individual points, while minimizing the total curvature of the surface. It fits a mathematical function to a specified number of nearest points. This method is best for gently varying surfaces such as elevation, water table heights, or pollution concentrations and, therefore, was applied to some of the data collected for this project (5).

The indoor air samples collected were in response to community complaints, and as a result, the samples were not appropriate for analyzing the whole study site with this software. The groundwater and soil gas data, however, were collected at strategic locations for the purpose of such analysis and therefore were good candidates for the spatial analyst software. The groundwater and soil gas point coverages described above provided the input points for creating contour (plume) coverages.

From the point coverages, the spatial analyst used Spline interpolation to create surface coverages of hydrocarbon thickness for the groundwater, as well as the gas levels for selected chemicals from the soil gas data. From the surface coverage, a contour coverage was generated but with some negative interpolations. The negative interpolations were removed, as were any contour lines for which no data existed. The resulting coverage was the contours of plumes representing hydrocarbon levels underneath the study site. Eight coverages and tables were created, including those for groundwater and soil gas data (Figure 5).

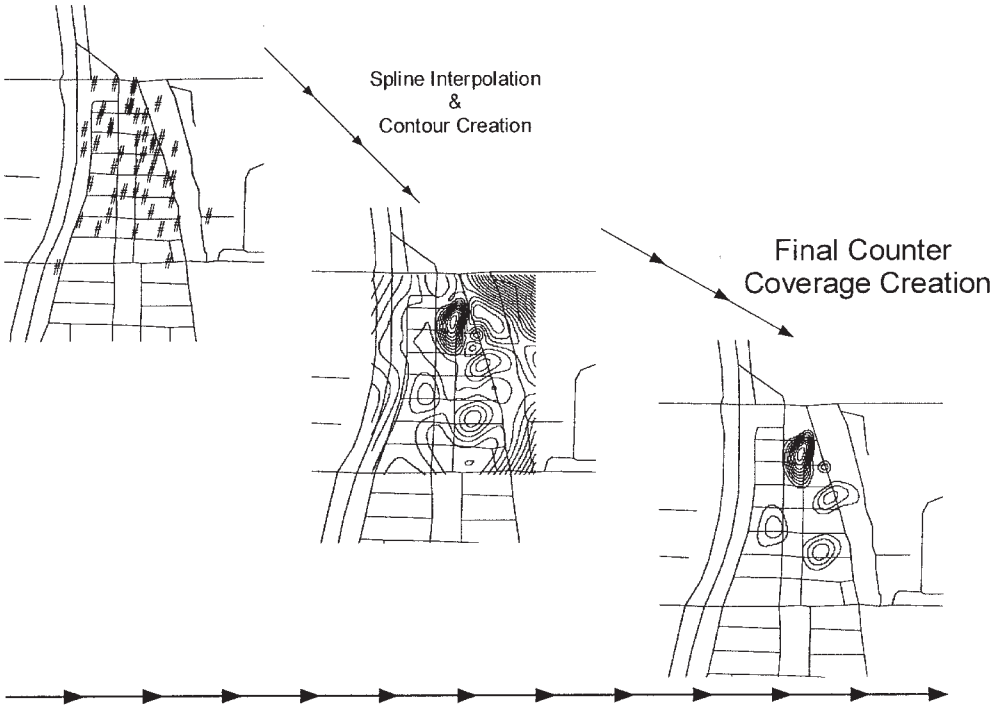


Figure 5 Creation of plume contour coverages.

Lastly, a polygon coverage showing the refinery boundaries was created directly in ArcView 3.0a. This coverage can be seen in Figure 1.

Use of GIS Coverages

Because coverages created for this project can be viewed using most GIS software, a health professional with the appropriate software and coverages can manipulate all of the past data about the study area. For this project, that included demographic data, indoor air data, groundwater data, and soil gas data. For example, an investigator who was new to the study site could use the address coverage to see where prior testing had been performed and what the results were. Also, any new testing sites could be added to the existing point coverages.

The air data could be used to view an individual house's sampling data with a simple click of the mouse; or, an investigator could view how one chemical affected all the houses sampled over a period of time. ArcView 3.0a has graphing built in, so one could view benzene levels, for example, to determine which house or houses may be at higher risk, as well as the time frame for exposure. From a simple graph, an investigator could better decide whether or not to proceed with exposure assessments on household members. This tool could be used in the field as well as in the office.

The most interesting aspect of having all the coverages created for this project was the ability to overlay the different types of data. For example, by overlaying the air sample point coverage and the hydrocarbon plume coverages, an investigator could see if the complaints of residents were corroborated. Figure 6 shows how the plume lies directly below the houses that registered complaints. Because the data sets were taken at different points in time, it was difficult to draw any conclusions using them, but one can see the utility of such coverages if the data were concurrent.

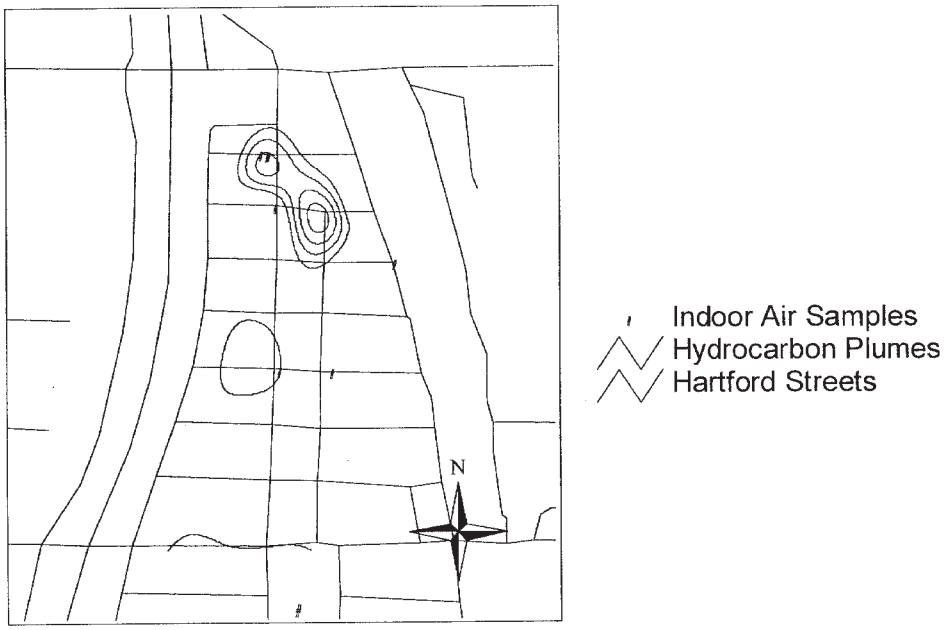


Figure 6 Air sample locations and plume coverage.

From the plume coverages alone, an investigator could learn about any plume movement. Figure 7 shows three plume coverages for three different testing times in 1990. Just by viewing the coverages, an investigator can see plume thickness and movement. This could aid in remediation actions or point out possible exposure points.

Again, all of the coverages and tables created can be manipulated. This means an investigator can add new data points, new data, and create new plume coverages based on such data. The coverages created by this project provided a framework for future GIS work on this particular project site in Hartford.

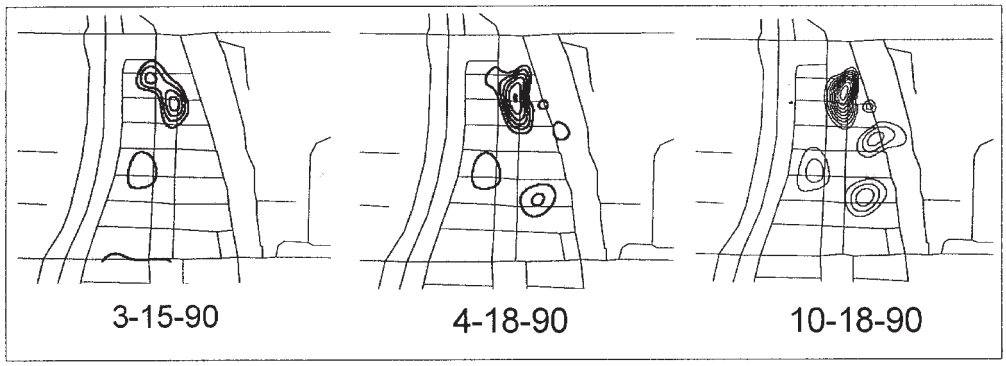


Figure 7 Tracking of 1990 hydrocarbon plume data.

Conclusion

GIS is a powerful public health tool that can only get stronger with time. Some issues that need to be addressed to expedite this growth include easier access to software, more refined software, and increased conversion of environmental and public health data into GIS form.

The software used for this project costs approximately $2,000. While this may not be a lot for some budgets, it will often be considerable for slim public health budgets. Costs can also rise if an organization wishes to run a UNIX-based system. Interagency sharing of data and software can cut back on these kinds of problems. For example, in this project, IDPH had minor GIS capability compared with IDNR or IEPA. It was only with cooperation between these agencies that data were obtained and software was made available. Once GIS's potential is fully recognized in the public health arena, we may see room made in agency budgets for such work.

Another issue that arises when using GIS for public health is the capabilities of the existing software. Because most of the past use of GIS has been by geographical and natural resource-based groups, the software is not tailored for public health investigations. The statistical power of GIS software is weak, and this is limiting for public health users (6). A public health investigator would be better off doing statistical analysis in a different software such as SPSS, SAS, or Excel. One good attribute of the ArcView 3.0a is that data are readily compatible with the aforementioned software. Again, once GIS use in public health increases, the demand for more powerful public health software will increase and so should the reality of such software.

Data are always the key to using GIS. Most projects will only be limited by the type and amount of data they possess. The process of data collection, conversion, and maintenance is a full-time job but is necessary for the success of GIS projects. Again, the time and financial hurdles can be minimized through data sharing and communication between agencies using GIS.

Specific to this project was the fact that there were not enough data or sufficient up-to-date data to make any solid conclusions about the site at Hartford, Illinois. One could, however, compare these data with past studies done on the site using similar data to see if similar trends could be observed. For this software to be useful to this project in the future, more air samples and concurrent groundwater samples are

needed. This will require agency funding and community cooperation, both of which were poor for this site study.

Even though the data sets were limiting for this site, it provided a testing ground and classroom for how GIS coverages can be used as yet another tool in health assessments.

Acknowledgments

Illinois Department of Public Health, Illinois Department of Natural Resources, Illinois Environmental Protection Agency

References

1. Illinois Department of Public Health (IDPH). 1990. *Preliminary health assessment of Hartford, Roxana, South Roxana and Wood River, Illinois*. IDPH. Springfield, IL.
2. Illinois Department of Public Health (IDPH). 1990. File information. Division of Environmental Health, IDPH. Edwardsville, IL.
3. Mathes & Associates, Inc. 1990. *Clark Oil Refining soil gas survey.* File information. Mathes & Associates, Inc. Columbia, IL.
4. Environmental Systems Research Institute, Inc. (ESRI). 1996. ArcView GIS software. ESRI. Redlands, CA.
5. Environmental Systems Research Institute, Inc. (ESRI). 1996. ArcView Spatial Analyst software. ESRI. Redlands, CA.
6. Levine N. 1996. Spatial statistics and GIS: Software tools to quantify spatial patterns. *Journal of the American Planning Association* 62(3):381.
7. Agency for Toxic Substances and Disease Registry (ATSDR). 1991. *GIS applications in public health and risk analysis: An ATSDR workshop.* ATSDR. Atlanta, GA.
8. Environmental Systems Research Institute. 1998. *What is GIS?* www.esri.com.
9. Hazardous Waste Research & Information Center (HWRIC). 1995. *Measurements of indoor toxic VOC concentrations attributed to the residential storage of household products.* HWRIC. Champaign, IL.
10. Mathes & Associates, Inc. 1990. *Recommendations for Hartford plume investigation and remediation pilot study, Hartford, Illinois.* File information. Mathes & Associates, Inc. Columbia, IL.
11. Nyman LW. 1997. GIS emerges in public health. *GIS World* 10:86.
12. Obermeyer NJ, Pinto JK. 1994. *Managing geographic information systems.* New York: The Guilford Press.
13. Petzold R. 1994. Yielding the benefits of GIS. *American City & County* 109(3):56.

Exposure Assessment in Environmental Epidemiology: Application of GIS Technology

John R Nuckols, PhD (1),* Mary H Ward, PhD (2), Stephanie J Weigel, PhD (1)
(1) Department of Environmental Health, Colorado State University, Fort Collins, CO; (2) Occupational Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD

Abstract

Environmental epidemiology evaluates associations between environmental exposures and health outcomes, with the purpose of further understanding the etiology of disease. An important component of such studies is exposure assessment. In many studies, exposure of participants over a relatively long period of time or large geographic region must be reconstructed. Such studies could be improved using technology based on geographic information systems (GIS). The purpose of this paper is to discuss the strengths and caveats of this use of GIS. For example, one strength is the ability to store, process, and analyze exposure data and other information about the study participants with spatial precision. This capability allows the researcher to access information that cannot always be ascertained in a traditional epidemiological study design. A good example of this is a study in which indirect exposure to environmental chemicals is being assessed for persons living in a highly integrated residential and agricultural or industrial landscape. It is unlikely that exposure to contaminants in such an environment could be accurately classified using traditional epidemiological methods such as survey questionnaires. This is because most people living in such landscapes have no knowledge of the chemicals being used and discharged into their environment. Using GIS-based technology, a researcher could locate sources of target compounds and calculate an exposure metric for each participant. Examples of such applications of GIS technology are presented in this paper. The caveats for applying a GIS in an exposure assessment do not differ substantially from other application areas for this technology. The user must be aware of cartographic issues, including scale and resolution. The accuracy of the data, the uncertainties in the analytical process, and the interpretation of the results remain important considerations in all GIS applications. This paper will illustrate the capabilities of a GIS for use in exposure assessment by applying it to an environmental exposure assessment for agricultural chemicals. Recommendations concerning the future of this technology in environmental health sciences will also be discussed.

Keywords: exposure assessment, epidemiology, risk, environmental health, remote sensing

Introduction

Environmental epidemiology evaluates the association between environmental exposures and health outcomes with the purpose of further understanding the etiology of

* John R Nuckols, Department of Environmental Health, 147 Environmental Health Bldg., Colorado State University, Fort Collins, CO 80526-1676 USA; (p) 970-491-7295; (f) 970-491-2940; E-mail: jnuckols@cvmbs.colostate.edu

disease. The term "environmental" implies a spatial component of the exposure metric used in the epidemiological study. In fact, a shortcoming of some environmental epidemiological studies has been that they do not locate subjects in the context of their true environment, which can often bias the exposure assessment.

This deficiency is demonstrated in the early history of health assessments of hazardous waste sites in the United States, conducted after the enactment of federal regulations to control environmental pollution. A National Research Council review of epidemiological studies of such sites concluded that misclassification or poor exposure metrics was a principal source of error (1). A review of the epidemiological studies indicated that, in most cases, exposure was defined as living within a specified distance of a hazardous waste site with little regard for fate and transport mechanisms of the study's target contaminants (2). None of the studies used computer modeling in their exposure assessment.

The advent the use of geographic information systems (GIS) in public health applications has greatly enhanced the capability to examine associations between environmental agents and disease. The purpose of this paper is to describe how GIS can be used in exposure assessment, as well as the strengths and caveats in applying GIS technology in this context.

Methods in Study Design

A GIS is, by definition, a database in which the information is spatially registered. However, a GIS not only maintains spatial registration, but displays the information in a mapped context. This is a major departure from the realm of numerical tables used in traditional epidemiology. A simple example of the power of maps can be demonstrated by visualizing a table composed of a list of travel destinations in the state of Colorado and their respective locations identified by latitude and longitude. Can you imagine trying to plan a vacation based on this type of information? This is the format of data typically used in the planning and implementation of exposure assessment for environmental epidemiological studies. As a result, the exposure assessment "plan" for most epidemiological studies is derived without the benefit of understanding the spatial context of the environment being studied and the ramifications of these data on the outcome of the study.

A good example of these issues can be found in epidemiological studies of agricultural workers. Most such studies are restricted to workers who use agricultural chemicals in their profession. The exposure metrics used in these studies are typically derived from a set of questions asked of the applicator concerning the type, frequency, and duration of chemical use. Many intensive agricultural regions of the country, however, are composed of a highly integrated landscape of agricultural and residential land use. In most cases, the inhabitants of these residences do not work directly in agricultural production. They may also be composed of more vulnerable populations such as children, women and men of child-bearing age, or elderly people. As such, they may be a more valuable population to study than agricultural workers if we want to get a true sense of the association between exposure to agricultural chemicals and certain disease outcomes. A traditional interview-based approach to studying this larger population is most likely doomed to failure. It is highly unlikely that individuals would have any

knowledge of the types of pesticide used on the fields next to their residences or details about their use.

We have recently demonstrated the utility of a GIS in identifying populations possibly exposed to pesticides from agriculture (3). In a feasibility study, we demonstrated that satellite imagery could be used to reconstruct historical crop maps, and that crop type could be used as a surrogate for pesticide exposure. We used historical Farm Service Agency records as a source of ground reference data to classify a late summer 1984 satellite image into crop species in a three-county area in south central Nebraska. Residences from a population-based case-control study of non-Hodgkin's lymphoma were mapped using a GIS. Twenty-two percent (22%) of the residences were within 500 meters of one of the four major crops, an intermediate distance for the range of drift effects from pesticides applied in agriculture (4,5). Using information from pesticide surveys, we identified the crop pesticides that were used most frequently on those crops. This feasibility study demonstrated that a GIS coupled with remote sensing data and historical records on crop location can be used to create historical crop maps. It also showed that probable exposure to crop pesticides near a residence can be estimated when information about crop-specific pesticide use is available.

Exposure, in the purest sense of the word, is the dose of a target substance that reaches the individual being studied. Because measurement or reconstruction of dose is virtually impossible, most environmental epidemiological studies use a surrogate measure of exposure. A useful surrogate of exposure is a variable that is correlated with the true exposure of interest. For example, in the study described previously, the exposure measure is the crop area in proximity to an individual's residence. It is assumed, based on information from other studies, that this variable correlates with exposure to pesticides commonly used on the crops and thus is a useful surrogate for exposure. The surrogate exposure measure could be improved if the type and amount of pesticide actually applied to the crop fields was known. Further improvements could be made in the classification of exposure by taking into account factors such as the application method and usual wind direction and speed at the time of application.

Table 1 Definitions of Cartographic Variables

Cartographic Variable	Definition
Cartographic scale	Relates size of a feature on the ground to size of map feature
Operational scale	Scale at which process of interest occurs
Spatial resolution	Grain or smallest distinguishable unit
Geographic extent	Size of study area

Geographic principles concerning scale and resolution must be considered when using GIS in exposure assessment for epidemiology studies, especially if a surrogate variable is being used to define exposure. Definitions of cartographic variables that could affect the utility of a GIS in exposure assessments are presented in Table 1 from Lam and Quattrochi (10).

Resolution

Resolution is a very important concept in the application of GIS in environmental epidemiology. Suppose, for example, that health data used in the Nebraska study cited above were only available as cancer incidence rates at the census tract level instead of having residence location at the time of diagnosis. The exposure data would also have to be aggregated to the census tract level for the data analysis. This would have greatly compromised the utility of having exposure assessment data at a resolution of less than 500 meters.

The converse of this situation can also occur. In a recent study concerning childhood leukemia and pesticide use by Reynolds (6), the researchers were able to map health outcome data at the residence level. However, the exposure metric used in the study was pesticide use reported at a resolution of 1 square mile (640 acres), which is the reporting unit for the California Pesticide Use Reporting database (PUR). Thus, the exposure metric used in the study was much coarser than the health data because of the difference in resolution of the two datasets.

An example of the effect of resolution on exposure assessment in an agricultural production landscape is presented in Figure 1. In this figure, we demonstrate different methods that could be used for estimating potential exposure to a residence from agricultural chemicals as the spatial resolution of information increases. Plate A in Figure 1 is an example of the data resolution available to the Reynolds study described previously (6). In this case, pesticide use data are reported at the level of resolution of 1 square mile (640 acres). From the PUR, we know that atrazine was applied to 300 acres out of a total of 400 acres of crop land. However, no information is available from the PUR on the location of crops. As a result, even if the residence can be geocoded, the probability of exposure can only be defined as equal for all residences because crop location data would be necessary for further refinement of the exposure metric. Thus, as shown in Plate A, the probability of potential exposure (defined as the ratio of acres where pesticides were applied to total possible acres) is 300/640, or 47%. In Plate B, resolution of exposure data is refined by including crop map data. In this case, probability of exposure can be based on whether a residence is located within an agricultural production area (400 acres). Because the total possible acres is now reduced from 640 to 400 in the study area, the probability of potential exposure for the residence located within the agricultural land use zone is 300/400, or 75%. By this method, the residence located outside the agricultural land use zone is considered unexposed.

Further refinement is achieved in the example in Plate C. In this case, a proximity metric is employed to ascertain potential exposure to a residence. Application of this procedure is described by Ward and Nuckols et al. (3). In their study, the proximity metric was based on the distance of potential drift of pesticides for the type of agriculture used in the study area (4,5). By this method, residences that would be classified as "unexposed" by the method in Plate B could be assigned a probability of potential exposure based on the extent of pesticide use within the designated buffer zone around the residence. Plate D in Figure 1 is an example of how the area of exposure to agricultural chemicals can be further refined using computer-based fate and transport analysis. In this example, a dispersion model for fugitive chemicals migrating from an agricultural field is employed to determine the gradient of concentration in the local environment. Other models that could be used include dispersion models for chemical drift in the atmosphere and hydrologic models for estimating the dispersion of chemicals in

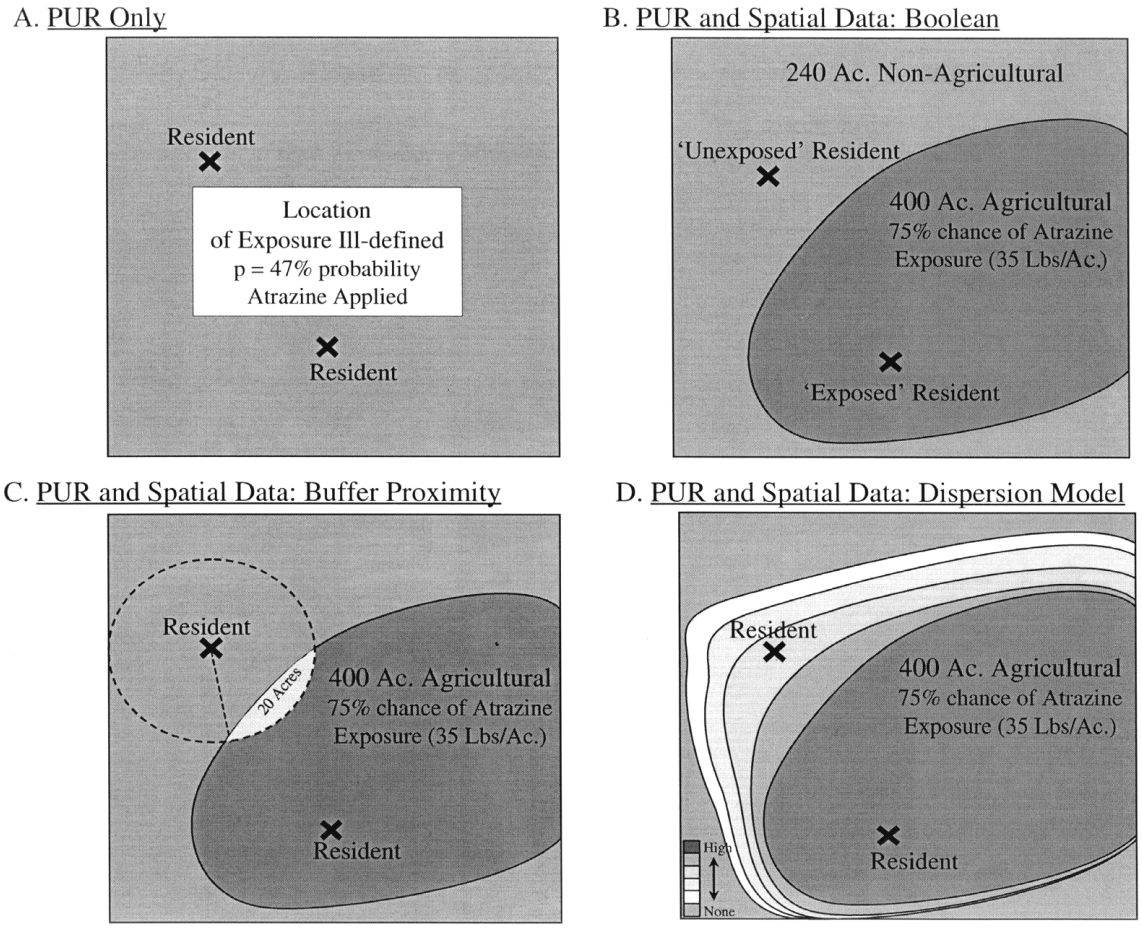


Figure 1 Methods for refining probability of exposure.

groundwater. Application of this technique for primary drift of agricultural chemicals from a spraying operation is described by Miller et al. (7).

Operational Scale

Operational scale is inherently tied to the resolution of the data that one uses because it dictates the resolution at which the exposure metric needs definition. For example, in the Nebraska study we used scientific literature concerning pesticide drift to determine that 500 meters was reasonable as an assumed distance at which drift would occur for the agricultural spraying practices that were prevalent in our study area. Thus, 500 meters became the operational scale at which we needed to be able to detect a change in cropping patterns in the area around each of our target residences. This example also points out the importance of selecting an operational scale based on scientific information about the exposure of interest, not just some arbitrary cutpoint.

Geographic Extent

Geographic extent is another important concept relevant to the application of GIS technology to exposure assessment and environmental epidemiology. Use of a GIS forces the placement of boundaries on the system being studied. Because exposure is in most cases a very dynamic process, the location of these boundaries can significantly affect the outcome of a study and the conclusions one might draw from the results. We demonstrated this in a recent health assessment study concerning a hazardous waste site near Denver, Colorado (8).

In this hazardous waste case, we were charged with determining whether residents living in the community adjacent to the site were being exposed to fugitive contaminants from the site. We concentrated on groundwater as the principal route of exposure. Over a period of several years, we used a series of metrics to classify exposure in this population. Each metric was a refinement of the previous one (i.e., starting with proximity and ending with modeling of contaminants in the water supply). With each refinement, the evidence became more convincing that contaminants that had been identified on the site were indeed present in the environment of the study population. Some of the groundwater modeling data, however, did not confirm the hypothesis that our site was the source. By extending the geographic extent of our GIS by just a few census blocks, we found that the actual source of the contaminant was another hazardous waste site located nearby. [A workshop on this issue, which uses this study as an example of the issues of scale and resolution in the application of GIS technology in exposure assessment, can be viewed at http://ehasl.cvmbs.colostate.edu (9).]

Systems Analysis

GIS is not only a tool with applications to exposure assessment and environmental epidemiology, it is a process. Once the geographic extent of the study area is defined, a GIS can be used to characterize the system of interest in terms of geophysical variables that might affect the study. For our agricultural chemicals example, this might include building layers of data in the GIS that describe the soils, geology, water supply, and meteorological and topographic factors related to pesticide transport phenomena. Sources of the target contaminant(s) can then be located within the system, and fate and transport algorithms can be applied using input data from the GIS. Some functions in GIS software (such as network modeling) can be used to predict fate and transport, but for

the most part, simulation models for the particular transport medium under consideration will be required. Most simulation models have output files that can be imported into a GIS and the results displayed with some programming effort. The resulting maps of results for different media or sources can be overlaid and a composite exposure metric derived using standard functions in most GIS software. In a like manner, demographic and other information concerning the study population can be stored and manipulated in a GIS. The investigator can then test different scenarios using these exposure and other datasets to conduct epidemiological analyses.

It is important, however, to follow a standard scientific protocol in applying GIS as an analytical process. By this we mean that the study hypothesis should be defined and have biological plausibility. Having biological plausibility means that a biological basis for an association between the target substance in the exposure assessment and the disease or health outcome proposed for the study can be demonstrated. This plausibility can be based on evidence from toxicological studies or previous epidemiological studies. The exposure assessment should also take into account other factors that may be correlated with the exposure and health outcome of interest (confounding factors). An example of a confounding factor in a study of an association between a specific pesticide exposure and a disease would be another pesticide that had a similar pattern of use and was also associated with the disease.

Other important considerations in epidemiological studies using GIS are data considerations and validation of the exposure metric. There is a rule of thumb that upwards of two-thirds of the resources in a GIS project can be consumed in database preparation, geocoding, and quality assurance/quality control. Thus, it is critical in the design phase of a study to have a clear understanding of the data that are available and the data collection effort that is required. Validation of the exposure metric can be accomplished by comparing predicted versus simulated exposure variables in a field study. Validation in most cases should be site-specific. That is, the researcher should avoid the assumption that because a simulation model worked in one study area, it works in all study areas.

Discussion

Strengths

The use of GIS in public health applications is in its early stages of development, and there are many considerations that should be taken into account as one attempts to use the technology. There are also a number of research issues that need to be resolved by the scientific community. Our experience indicates that GIS can strengthen an environmental epidemiological study. When appropriately used, the technology allows the investigator to take the subject out of a numerical format and into a mapped database that can be more reflective of the subjects' environment. This can result in better study design and exposure assessment.

Caveats

The multiple databases in a GIS that describe the environment can be used as input to more precise exposure models. However, there are a number of caveats in this application of GIS technology. Perhaps the foremost caveat is that, if after calculating the geo-

graphic extent at which the study needs to be conducted[1] there are insufficient data at the scale and resolution necessary to correlate the exposure metric with the disease outcome, one should be very cautious in using a GIS.

Another important caveat in the use of GIS in exposure assessment and epidemiology is consideration of the uncertainty associated with the data. The power of a GIS is the ability to handle multiple datasets, or layers of data. It should be understood, however, that each of these data layers contains a certain degree of uncertainty. As the user adds more and more layers to the exposure metric, this compounds the uncertainty associated with the final product. How to express this uncertainty in a GIS database and carry it through the analytical process is an important research issue for the GIS community. It is important that every effort to incorporate uncertainty in the metric be made and that this information be provided to the epidemiologist. Misclassification of exposure, when it is nondifferential by disease or exposure status, dilutes the risk estimates and causes associations to be missed. An assessment of the uncertainty in the exposure variable is important so that the effects of misclassification of exposure on the risk estimates can be assessed.

Data interpolation, defined as the estimate of data values between locations of actual measurement, is another important issue in the application of GIS technology to exposure assessment. An example of such error is the interpolation of water quality data from wells across a geographic region. In an epidemiological study, the location of the study participants using wells and the locations of the wells for which there are water quality data may not coincide. One approach for assigning exposure is to apply an interpolation algorithm to the well data, creating isopleths of water quality values for points in between. These derived water quality values are then assigned to the wells of study participants where measured water quality data are not available. A caveat in using such techniques is that the spatial distribution of a substance in groundwater is highly dependent on the geophysical and hydrogeologic characteristics of the aquifer medium. Thus, if the investigator does not include this information in the interpolation procedure, the exposure metric assigned to a subject with missing data can be significantly in error.

Research Issues

There are a number of issues to consider in applying GIS technology to exposure assessments for environmental epidemiological studies. To date, most of the applications of this technology have been "retrofitted" to previously conducted epidemiological studies where a GIS was not considered in the original study design. To truly evaluate the potential use of this technology, its utility in both the design and analytical phases of epidemiological studies should be considered. One means for conducting such research would be to establish a set of long-term research sites such as those established for ecological studies by the National Science Foundation (NSF). Long-term ecological research sites were established by NSF to inventory ecological resources and to understand ecological processes that affect these resources within a specified geographic area. Subsequent research projects can then be conducted on the site to develop tools for understanding changes in the ecology. The technology for these tools can then be transferred for use in other ecological regions. By applying this approach to public health,

[1] This could be a statistical power consideration, or based on the location of a target study population.

GIS-based surveillance and analytical methods—for both long-term and "rapid-response" needs—could be developed that might improve intervention efforts.

Database development is a huge front-end expenditure for a GIS. It stands to reason that such an investment, especially for an application that is still very much in the research stage, should be made in sites that have long-term and multipurpose research potential. In fact, NSF has allocated $1 million to the development of a system of observatories in human-dominated ecosystems where long-term studies of critical ecological processes will be initiated (11). Additional support will be applied to the study of urban communities. Perhaps this is an opportunity for NSF and the National Institutes of Health to collaborate on a GIS-based research initiative.

Finally, there is a need for a support mechanism for strong interdisciplinary collaboration in the field of exposure assessment and environmental epidemiology. GIS provides a powerful platform that can be used to bridge disciplines. There are few opportunities, however, for obtaining funding for research concerning the development of exposure assessment methods with direct application to epidemiological studies. Though most agencies tout interdisciplinary research as a planning objective, little effort has been made to establish programs and proposal review sections that incorporate an interdisciplinary perspective, much less a knowledge of GIS technology.

Conclusion

There have been many advances in public health over the last century. GIS technology has the potential to revolutionize the way we approach exposure assessment in environmental epidemiology, the way we conduct health surveillance programs at the local, state, national and international levels, and the way we report health and environmental data to our citizens. To succeed with such lofty goals, we must be sure that the limitations of the technology are considered and that the use of GIS is based on sound scientific principles.

Acknowledgments

Funding for the projects cited in this paper was provided in part by the National Cancer Institute's Division of Cancer Epidemiology and Genetics (Contract NO2-CP-71100), and the Agency for Toxic Substances and Disease Registry (PA505(HARP)H75/ATH881505). Special acknowledgment to Dr. Stephanie Weigel, Mr. Chris Skinner, Mr. David Ellington, and Mr. Ryan Miller in the Environmental Health Advanced Systems Laboratory at Colorado State University for their contribution to this manuscript. Mr. Chris Skinner produced the graphic in Figure 1.

References

1. National Research Council. 1991. *Environmental epidemiology: Public health and hazardous wastes.* Washington, DC: National Academy Press.
2. Nuckols JR, Berry JK, Stallones L. 1994. Defining populations potentially exposed to chemical waste mixtures using computer-aided mapping and analysis. In: *Toxicology of chemical mixtures: Case studies, mechanisms, and novel approaches.* Ed. RSH Yang. New York: Academic Press. 473–504.

3. Ward MH, Nuckols JR, Weigel SJ, Maxwell SK, Cantor KP, Miller RS. Identifying populations potentailly exposed to agricultural pesticides using remote sensing and a geographic information system. *Environmental Health Perspectives* (In press).
4. Byass JB, Lake JR. 1977. Spray drift from a tractor-powered field sprayer. *Pesticide Science* 8:117–26.
5. Frost KR, Ware GW. 1970. *Pesticide drift from aerial and ground applications.* Agricultural Engineering 51:460–7.
6. Reynolds, P. 1997. *Childhood cancer—etiologic clues using GIS*. Bethesda, MD: National Cancer Institute. RO1 CA71745-02.
7. Miller RS, Nuckols JR, Ward MH. *Estimating historical pesticide exposure using pesticide transport modeling in a GIS.* Fort Collins, CO: Environmental Health Advanced Systems Laboratory, Colorado State University. Unpublished report.
8. Reif JS, Nuckols JR, Ellington D, Weigel SJ, Burch JL, Stone AE. 1999. *Evaluation of priority health conditions at the Rocky Mountain Arsenal: A re-analysis based on geographic information system-derived exposure assessment for trichloroethylene. Final Report.* Atlanta, GA: Agency for Toxic Substances and Disease Registry. PA505(HARP)H75/ATH881505. (In review).
9. Nuckols JR, Weigel SJ. 1998. *Exposure assessment using GIS technology.* Fort Collins, CO: Environmental Health Advanced Systems Laboratory, Colorado State University. http://ehasl.cvmbs.colostate.edu. (See "GIS in Public Health—San Diego, 1998" at this Web site.)
10. Lam NS-N, Quattrochi DA. 1992. On the issues of scale, resolution, and fractal analysis in the mapping sciences. *Professional Geographer* 44(1):88–98.
11. National Science Foundation. 1997. *NSF funds first long-term studies of urban ecology.* NSF PR 97-63. Washington, DC: National Science Foundation.

The Use of GIS in Identifying Risk of Elevated Blood Lead Levels in Australia

Lisel A O'Dwyer, PhD*
National Key Centre for Social Applications of GIS, School of Geography, Population and Environmental Management, Flinders University, Adelaide, South Australia, Australia

Abstract

Unlike in the United States, environmental lead in Australia and the dangers it poses to small children are generally not regarded as a major public health issue. However, the results of a recent national survey, which found that only 7.3% of children under 5 years old had blood lead levels (PbB) over 10 micrograms per deciliter, are questionable due to insufficient attention given during sampling to the geography of urban housing and risk factors. Geographic information systems (GIS) can address such variation, and based on the spatial distributions of known risk factors, can identify areas, streets, and even individual dwellings with a high probability of high environmental lead levels. Predictions can then be validated with atomic absorption spectrometry analysis of blood or dust samples. Preliminary results based on GIS analysis of a metropolitan digital cadastral database and its associated housing data and the spatial distribution of relevant entities, such as childcare centers, suggest that the prevalence rate of elevated PbB in one major Australian city is significantly higher than was reported in the national survey. This analysis will form the basis for a model predicting the presence and risk of environmental lead for any city. It also offers a means of targeting further investigation of lead exposure, using small-area census data to estimate the number of children at risk more accurately, and selecting areas for future sample surveys. Developing a cost-efficient and accurate method of modeling lead exposure risk to children is a task to which GIS is clearly well suited.

Keywords: pediatric blood lead levels, risk, urban, socioeconomic status

Introduction

The issue of elevated blood lead levels (PbB) has not received the detailed attention in Australia that it has in the United States. There are many reasons for this. One is the perception among health authorities and the lay public that environmental lead is not a major concern, especially in relation to other preventable and more manageable childhood illnesses. Despite this perception, it has been suggested that lead poisoning is probably more common than most of the diseases routinely screened for in childhood (1). Another reason is the political difficulty in addressing the issue of lead, especially where the lead industry is important to the local economy, as is the case in South Australia.

While Australia has the same risk factors for elevated PbB as the United States, there may be some differences between the two countries in the relative importance of these factors. In Australia, it has taken eight decades for the lead content in paint to be reduced in increments to its current levels, though the danger was recognized in the

*Lisel O'Dwyer, National Key Centre for Social Applications of GIS, Flinders University School of Geography, Population and Environmental Management, GPO Box 2100, Adelaide, South Australia 5001 AUS; (p) 61-08-8201-2969; (f) 61-08-8201-3521; Email: lisel.odwyer@flinders.edu.au

Australian state of Queensland at the turn of the century. (Most Queensland housing is painted timber, thus maximizing the amount of lead paint in children's environments). The lead content of paint was not reduced in the United States until 1978 (2). On the other hand, the phasing out of leaded gasoline in Australia began in 1986, but its sale was made illegal in the United States in 1976 (3). In 1996, 38% of Australian vehicles still used leaded gasoline, while 45% of vehicles in the state of South Australia still used it. Many households with these vehicles are of low socioeconomic status and low income, and cannot afford to purchase and run later-model cars. People in lower-socioeconomic-status households are also more likely to work in industries involving and handling lead. It should be noted that South Australia's economy has been depressed in relation to the other Australian states for several decades and that South Australia has a strong base in manufacturing and light and heavy industry. The only Australian state that has a higher rate of leaded gasoline usage than South Australia is the economically depressed state of Tasmania. South Australia was also the last state to achieve the recommended lead level of 0.2 milligrams (mg) per liter for "unleaded" gasoline in October 1996 (4).

While there may be differences in the relative importance of the different sources of lead between Australia and the United States, in terms of previous findings and outcomes, there appears to be little difference between the sources of elevated PbB per se.

Lead is both a health problem and a social problem. As an industrial, urbanized country, Australia is no stranger to the known risk factors such as old paint in older housing, residential proximity to industries using lead, and household cleanliness. Many other lead risk factors involve individual behavior, such as hobbies involving lead, a child's tendency toward pica, and nutritional status. However, the literature to date has shown that a great deal of the lead involved in childhood lead poisoning comes from a child's environment. The source and distribution of this lead is thus beyond the control of individual households and parents. The main source of ingested lead in urban residential areas is usually paint in older housing. More than 3.5 million houses in Australia were built before 1971, when paint typically had high levels of lead. In general (though there are variations between the states), Australian paint contained up to 50% lead until 1950, when lead in paint was reduced to about 10%. The concentration was further reduced to about 1% in 1970, to 0.25% in 1992, and then to 0.1% in December 1997.

The literature generally agrees that the age of housing is a good indicator of the presence of old (leaded) paint. It has been found, for example, that Canadian children living in homes built in or before 1945 had an average PbB 62.3% higher than that of children living in homes built since 1975 (5). Australia's National Survey of Lead in Children (NSLIC) (6) also found a relationship between age of housing and PbB, even though the data for house age were based on estimates by interviewers and respondents rather than official sources (3). Even where children do not reside in a dwelling likely to have high levels of environmental lead, it is still important to identify such dwellings. In some cases children's elevated PbB are derived not from their own residence but from other residences in their community (7). Some dwellings are also contaminated by lead paint in adjoining dwellings via airborne and mechanical transport (8).

The status of the lead problem in Australia is difficult to ascertain. No direct comparisons between the United States and Australia are available. Prevalence rates for

PbB in the United States are based on the 1-to-5 age group, while the NSLIC—the only large-scale survey of PbB in Australia—examined children aged 1 to 4. Moreover, the use of mass screening programs in the United States over many years makes it possible to calculate reliable prevalence rates. Australia does not have any ongoing screening programs at all, except in the South Australian lead smelter town of Port Pirie. With these caveats in mind, however, the prevalence of elevated PbB (0.49 micromoles per deciliter [µmol/dL] or 10 micrograms per deciliter [µg/dL]) among 1- to 4-year-olds in Australia was found to be 7.3%. This represents approximately 75,000 children. The prevalence rate for American children in 1994 aged 1 to 5 is 8.9% (9). The finding that the prevalence rate in Australia was only 7.3% meant that the prevalence of elevated PbB was much lower than the Australian target prevalence of 10% by 1998, which was set in 1993. Some Australian health authorities have treated the low prevalence of elevated PbB in Australian children as a sign that there is no need to act to prevent lead exposure. Yet it has been shown that one-quarter of children within a 10-kilometer radius of the Sydney city center had elevated PbB in 1995 (10), as did a quarter of children in a working-class area in Perth, Western Australia, in 1994 (11).

There are clear social processes that result in excessive exposure to lead. The link between gentrification and childhood lead poisoning is supported by several studies that have found that the highest levels of blood lead are among children from the higher social strata, although the prevalence rate is higher among the lower-socioeconomic-status groups (12). The NSLIC found that most children with elevated PbB were socially disadvantaged (3), although an American survey found that 30% of urban infants (aged up to 1 year) with high socioeconomic status had PbB over 10 µg/dL (13). Although other studies agree that lead poisoning is found in all types of communities, it has been found that children in lower-socioeconomic-status areas were 7 to 10 times more likely to have lead poisoning (14). It is also significant that many of the households undertaking renovation of older housing (and who live in older housing either because of the price attraction of older inner-city housing or the investment potential of older housing in more upmarket areas) are young couples, including pregnant mothers and couples with young children. It has been observed that "renovation tends to impact on the most sensitive population" (8).

The fact that lead poisoning continues to occur in Australia shows that it is still a problem. The risk factors for lead poisoning are well known but research identifying the precise locations of these factors has yet to be undertaken in Australia. In the absence of blood testing—which measures only recent ingestion, not long-term exposure—other researchers have attempted to use questionnaires to predict elevated pediatric PbB. Results have shown the questionnaires to be not much better predictors than chance (13,15–18). There is a need to develop an alternative method of identifying children at risk. Because we are unable to identify the locations of individuals, the next best option is to identify precise areas that are likely to have hazardous levels of environmental lead. This is a task to which GIS appears well suited, given its capacity to integrate and query a variety of different datasets on a precise spatial basis. Some studies (19,20) that have used coarse spatial units (namely postcodes or local government areas [LGAs][1]) to examine the spatial distribution of PbB suggest that geographic location is

[1] An LGA is the lowest level of government in Australia, corresponding with the US county.

not of significant predictive value, but the small body of American research using GIS suggests otherwise (21–25).

It is obvious that the use of mass screening programs like those practiced in the United States would be a major expense for the Australian health system. Yet there are variations, spatial patterns, and concentrations of lead risk factors, and thus in elevated PbB, that might make targeted programs worthwhile. Hence the use of geographic information systems (GIS) in examining this public health problem. GIS technology has an important role to play in identifying the areas most likely to have high environmental lead levels—even though the prevalence of elevated PbB may be low in aggregate terms, these areas do exist. Random sample surveys have been useful in identifying the risk factors, and have shown that many of these risk factors have a characteristic spatial distribution or locational element. We may not need GIS to identify low-socioeconomic-status areas, but GIS enables us to identify individual addresses within those areas. It can also identify high-risk dwellings that may be nestled within an apparently low-risk area.

Risk is a two-dimensional concept, involving, first, the possibility of an adverse outcome and second, uncertainty over the occurrence, timing, and magnitude of that adverse outcome (25). If either is absent, then there is no risk. In the case of environmental lead, the presence of known markers or indicators shows there is a possibility of an adverse outcome. Whether lead poisoning actually occurs depends, of course, on the presence of children, while its timing and magnitude depend on a number of less readily measured factors such as the amount of time spent in the hazardous location, nutritional status (particularly iron, zinc, calcium, and fat intake), hand-to-mouth behavior, and the frequency and efficacy of household cleaning and vacuuming. The pathways by which lead enters the human body are undeniably complex, but it is nevertheless useful to consider using the presence of lead indicators and their spatial distribution as one way to estimate risk.

Method

Selection of Case Study Areas

For several reasons, South Australia is a useful area for a case study to demonstrate the use of GIS in identifying areas at risk for high environmental lead. According to the NSLIC, South Australia has the second highest mean PbB in the country (the Northern Territory has the highest level, but this is based on fewer than 20 cases) (6). South Australia's capital city of Adelaide is a relatively small city (1 million people), thereby facilitating travel for fieldwork and minimizing study costs. Another reason to select Adelaide is that networks already exist between Flinders University's Key Centre for Social Applications of GIS and relevant government and university departments. Figure 1 shows the locations of Adelaide and South Australia.

The original aim of this study was to examine the entire city. However, it was found that the database management system simply did not have the capacity to store all of the data required for the whole metropolitan area. Thus it was decided to select two LGAs as pilots with a view to extending the analysis to the whole city, one LGA at a time, after the methodology was established. (The Adelaide metropolitan area has

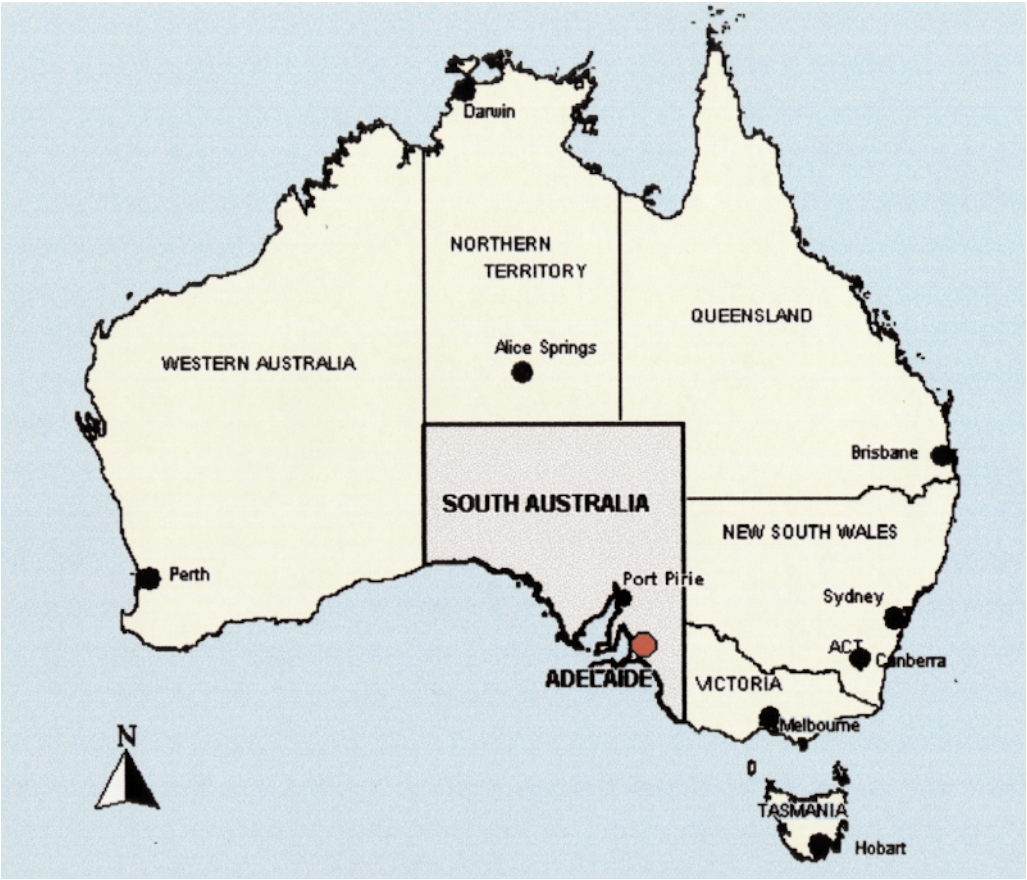


Figure 1 Location of study area in Australia.

approximately 30 LGAs, which vary in their level of heterogeneity with respect to socioeconomic status and land use.)

The two LGAs selected as case studies are Port Adelaide-Enfield LGA and Mitcham LGA (Figure 2). Both areas are located at similar distances from the central business district and have similar urban development histories, but have completely different socioeconomic profiles (Figure 3). They were selected because there is significant socioeconomic and housing type variation within as well as between them, so using them demonstrates the ability of the GIS to target specific small areas.

A high level of both heavy and light industry and a higher proportion of public housing than Adelaide as a whole characterize Port Adelaide-Enfield. Much of the LGA was settled in the first half of the century, although most of the public housing was built in the postwar period. Mitcham was also settled in the first half of the century but is located in a more desirable part of Adelaide. Its land use is mostly residential, with some commercial districts and little light industry.

Although both areas have a relatively old age structure (in relation to Adelaide as a whole), they both have experienced substantial levels of gentrification over the last two decades. This is consistent with the character of the housing in the two areas—

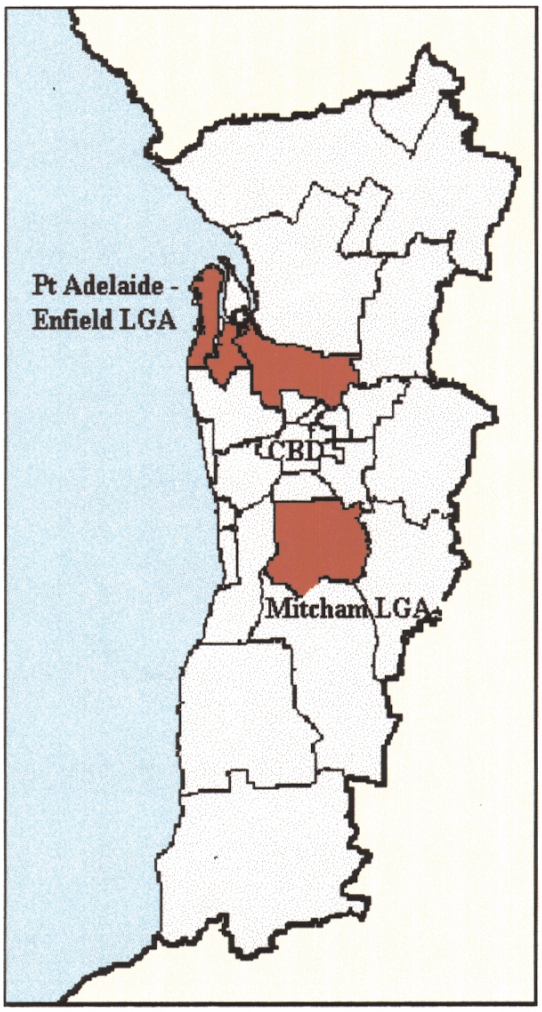


Figure 2 Location of case study areas in Adelaide, South Australia.

older-style cottages and bungalows have become popular and are increasingly in demand, particularly among higher-income professional households. However, the housing in Port Adelaide-Enfield is also low-priced and attractive to many low-income households, such as publicly housed households and some first-time home buyers.

Software and Hardware

ARC/INFO Version 7.1 (ESRI, Redlands, CA) was used on an IBM RS 6000 Unix workstation.

Datasets

Most of the risk factors identified in the literature are already represented in existing

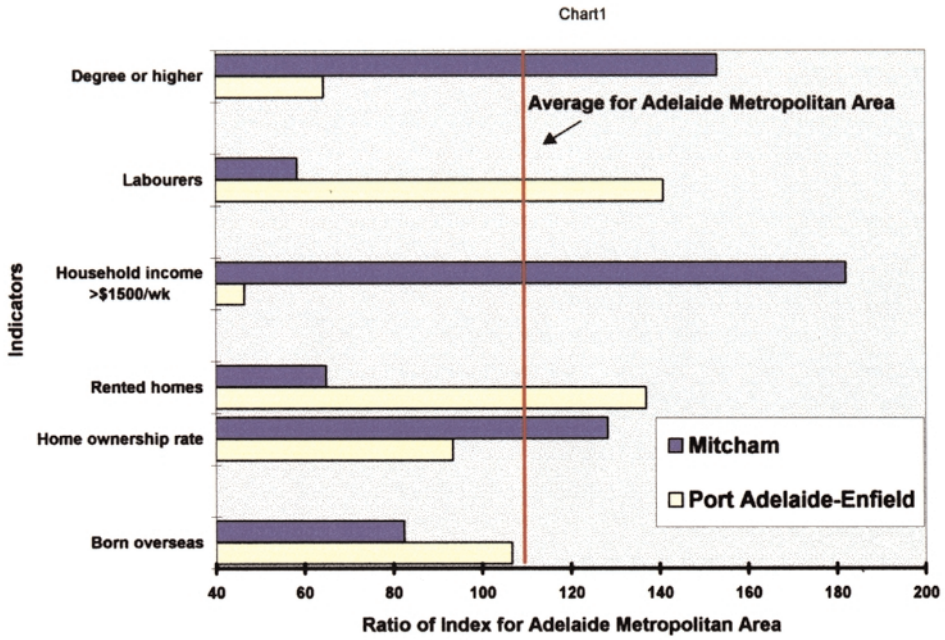


Figure 3 Selected socioeconomic status indicators for Mitcham and Port Adelaide-Enfield, Adelaide, South Australia, 1996.

datasets constructed for various other purposes. All that remains is to bring them together on a spatial basis. In this case, four main datasets were used.

The first is the Digital Cadastral Data Base (DCDB) for South Australia, which is a computer-based map of all land parcels in the state. It comprises approximately 800,000 land parcels, together with their legal identifiers. Associated data include street names and boundaries of administrative regions such as LGAs, wards, suburbs, hundreds, and counties. The South Australian DCDB is operated and maintained by the South Australian Department for Environment, Heritage and Aboriginal Affairs (DEHAA) but the location of the database varies from state to state. For example, there is no centralized DCDB in the state of Victoria—each local government in Victoria is responsible for the cadastre of its area.

The DCDB is widely used by government agencies and utility companies as a basic reference for land administration, local government administration, facilities management, planning, and asset management. It is one of the major core spatial datasets maintained by the government. Each parcel has a unique identifier that can be linked to property valuation assessments for rating and taxing purposes. It is the valuation data that provide a wealth of information pertaining to lead risk, namely:

- The dates on which properties were constructed

- The land use code for each parcel
- The material of dwellings' roofs and walls
- A rating of the condition of each dwelling on a scale of 1 to 9

The two case study LGAs represented 68,000 parcels, which reduced processing time and database management and storage considerably.

A second dataset used was the South Australian subset of the NSLIC. The original plan was to use it to validate the predictions of the GIS. However, the NSLIC proved to be somewhat disappointing for several reasons. One was that data on the ages of dwellings were based on either the interviewer's or the householder's estimates rather than on any reliable basis. Matching the addresses in the NSLIC with the valuation data and the DCDB showed that in 78% of cases, the year the householder or interviewer estimated the dwelling was built was incorrect, by a margin of approximately 5 years on average. Almost 90% of the households that were renting their houses made incorrect estimates of the year their dwelling was built, with an average error of 6 years. Most of these households lived in dwellings built before 1970. The degree of error generally increased with the age of the dwelling. The average size of the error was 10 years for dwellings built before 1920, 9 years for dwellings built between 1940 and 1960, 7 years for dwellings built between 1960 and 1970, 4 years between 1970 and 1980, and only 1.5 years for more recently built dwellings.

Another problem was the small number of cases—only 130 cases for the whole state. Only about half of those were located in Adelaide, even though Adelaide contains three-quarters of the state's population. No cases were located in Mitcham or Port Adelaide-Enfield. In terms of showing any geographical distribution of PbB, the number of cases was too small to draw any valid conclusions. Finally, some of the questions used in the household questionnaire were badly worded, the data on the condition of paint varied according to whether the householder or the interviewer estimated it, and the method of dust sample collection was not always appropriate.

Due to the continuing use of leaded gasoline in Australia, traffic flow rates were seen as an important indicator of lead risk for nearby residences. The South Australian Department of Roads and Transport supplied (at no charge) traffic information it had gathered of the whole state. Somewhat surprisingly, the electronic information they supplied did not include traffic counts associated with roads. Instead, they had hard-copy maps labeled with average daily number of vehicles for main roads. These had to be added to the GIS manually using the ArcEdit module of ARC/INFO. Traffic counts for suburban streets within the two case study LGAs were available from the respective local government engineering departments. The NSLIC defined a heavy traffic flow as 5,000 or more vehicles per day, which is very low. Thus it is not surprising that no correlation between PbB and proximity to roads was found in this study. However, when the same data were reanalyzed by the Victorian health department using traffic flows of greater than 20,000 vehicles per day, there was indeed an association. Consequently, the benchmark used in this study was 20,000 vehicles. Only 11 of the 133 cases in the South Australian subset of the NSLIC had data on traffic counts and only 7 of these were next to roads traveled by more than 20,000 vehicles per day.

The 1991 Census of Population and Housing (26) counts of the number of 0- to 4-year-olds were allocated to residential areas within the DCDB. Here we were forced to use averages to get around the problem of allocating areal data to individual dwellings.

Fortunately, the situation with Australian small-area census data is somewhat better than in the United States. Australia's smallest areal unit is the collector's district (CD), which contains 220 households on average, whereas the smallest American spatial unit, the block group, contains about 400 households. (By comparison, New Zealand's smallest unit is a "mesh block" of approximately 50 households and the United Kingdom's smallest unit contains around 170 households.)

We overlaid the CD boundaries with the DCDB and, for each CD, divided the number of children aged 0 to 4 by the number of residential dwellings shown in the DCDB. The resulting average number of children per dwelling was then allocated to every residential dwelling within the CD. In 1996, at the time of the most recent census, approximately 3,000 children aged 0 to 4 lived in Mitcham, forming 5% of the total population of that area. There were 6,300 children aged 0 to 4 in Port Adelaide-Enfield, forming 6.5% of the population. The percentage for the whole metropolitan area of Adelaide is 6.4%.

Results

Age

The two case study areas had similar proportions of housing built before 1970. Port Adelaide-Enfield had 81%, Mitcham 78%. This compares with 50% for Australia as a whole (27) and 53% for Adelaide (26,28). However, Mitcham has somewhat more housing built before 1952 (34%) than Port Adelaide-Enfield (28%). Clearly, we may expect most children in these LGAs to live in older housing; indeed, it was found that 1,200 Mitcham children under 5 (approximately 30%) and 1,150 Port Adelaide-Enfield children under 5 (approximately 27%) live in houses built before 1952. A further 30% of Mitcham children under 5 lived in houses built between 1952 and 1971, but around half of Port Adelaide-Enfield children under 5 lived in such housing. This is a reflection of the socioeconomic differences between the two LGAs. Figure 4 shows a view of the distribution of dwellings by age in part of Mitcham LGA.

Condition

According to our research, one-quarter of Mitcham dwellings are in bad condition. Approximately 800 children aged 0 to 4 live in these dwellings. However, only a quarter of bad-condition dwellings in Mitcham were built before 1952. This may reflect the level of gentrification in Mitcham—the older houses are very popular and tend to be quite valuable. In Port Adelaide-Enfield, 40% of housing was in bad condition, even though Mitcham and Port Adelaide-Enfield have similar proportions of housing built after 1952. Many of the bad-condition dwellings in Port Adelaide-Enfield were built after 1960; there, an estimated 2,500 children under 5 live in housing built after 1960.

Land Use

All land uses identified as possible lead risks were selected using the ARC/INFO commands reselect and the andselect, which were saved in an ARC Macro Language file (ARC Macro Language, or AML, is the macro language used within ARC/INFO). Based on the literature, "risk land uses" were defined as wholesale trade of petroleum products, service stations (gas stations), printing and allied industries, paint manufacturers,

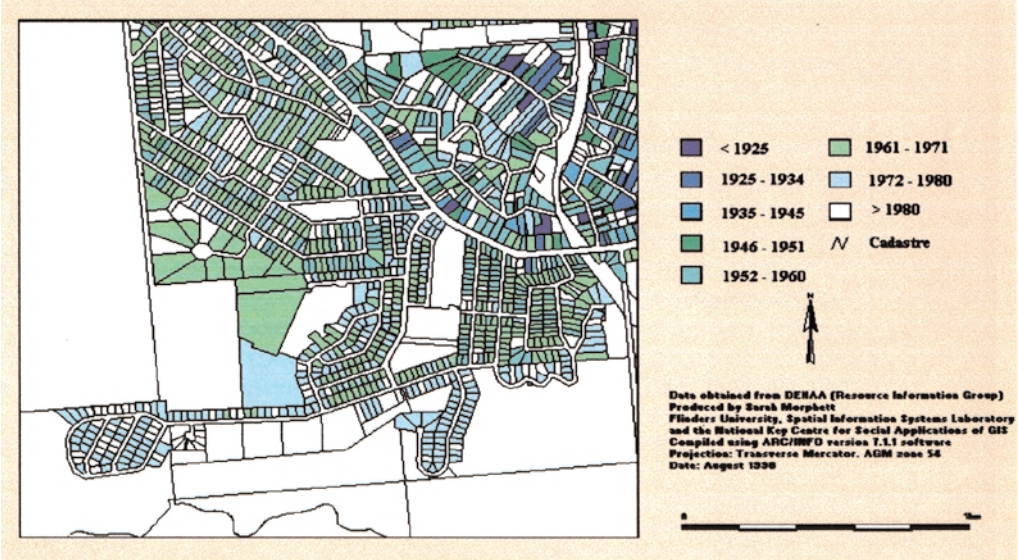


Figure 4 Detailed view of residential parcels by year built, Mitcham LGA.

petroleum refineries, petroleum and coal products, pottery, china and earthenware, iron and steel basic industries, non-ferrous metal industries, industrial waste disposal, active slag dumping and mineral waste disposal, and parking lots.

Port Adelaide-Enfield was found to have 255 risk land uses; there were actually 247 land parcels involved in risk land uses, but 5 of them were engaged in multiple risk land uses. The 247 parcels covered virtually the entire range of land uses identified in the literature, but the most important were service stations, printing industries, wholesale trade of gasolineeum products, and iron and steel basic industries.

Mitcham only had 28 risk land uses, and 23 of these were service stations. We found that approximately 20 children under 5 in Mitcham live within 50 meters of these land uses, while 135 children under 5 live in close proximity to risk land uses in Port Adelaide-Enfield.

Note that the buffer size was deliberately selected to be conservative because there is little discussion of buffers in the relevant literature to date. At this stage, the shape of the buffer is a simple circle with the land use as a point in the center, but we do acknowledge that a rose diagram with an ellipse-shaped buffer taking account of wind strength and direction would refine the buffering technique. It is emphasized that the aim was to keep the procedure as simple as possible with the option of refining the methodology later. Figure 5 shows the number of dwellings in close proximity to risk land uses.

Proximity to Busy Roads

The results from the NSLIC, limited as they are, showed that the average PbB for Adelaide children under 5 within 25 meters of roads with traffic counts of 20,000 or more was 0.35 µmol/dL, compared with 0.27 µmol/dL for those living on quieter streets. However, much previous research shows a strong link between PbB and traffic

Figure 5 Detailed view of residential parcels adjacent to lead land use, Port Adelaide-Enfield LGA.

counts, so it is important to estimate how many children under 5 do live within 25 meters of major roads.

There were 117 children under 5 in Mitcham (4% of children under 5 in that LGA) within 25 meters of roads with average daily vehicle flow of 20,000 vehicles. The 25-meter buffer is measured from the center of the road. Road widths vary; the Department of Transport traffic data do contain information on road widths, so it is possible to increase the size of the buffer according to the width of the road. However, for the sake of simplicity and to construct a basic model, we used a 25-meter buffer on all roads regardless of their width. Note that the NSLIC did not adjust distances from roads according to road widths (which of course reflect traffic flows). This means that the number of children in close proximity to major roads as calculated here is a conservative figure. Figure 6 shows the number of residential parcels in close proximity to busy roads in Mitcham LGA.

Even though the number of children under 5 in Port Adelaide-Enfield is double that of Mitcham, only 151 children under 5 in that LGA (2%; i.e., proportionally half as many) live within 25 meters of roads carrying more than 20,000 vehicles per day. This proportion is smaller for Port Adelaide-Enfield than for Mitcham because there are more industrial and commercial facilities than residential properties along the heavy-traffic roads in Port Adelaide-Enfield.

In sum, it is estimated that more than 4,000 children under 5 (approximately 40% of the population of 0- to 4-year-olds) in the two case study areas are possibly exposed to lead risk factors. The two most common risk factors found together are old housing and poor housing condition; these two risk factors are present for over 90% of dwellings in the two case study areas. The relative importance of each risk factor is similar in both case study areas, although traffic is more important for Mitcham, while risk land uses

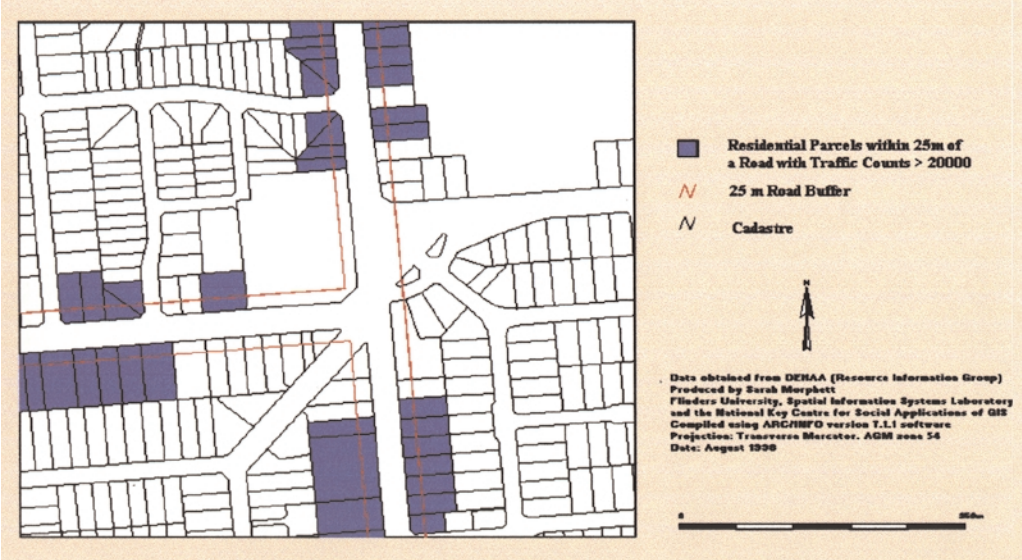


Figure 6 Detailed view of residential parcels adjacent to high traffic counts, Port Adelaide-Enfield LGA.

are more common in Port Adelaide-Enfield. Figure 7 shows a selected portion of Mitcham dwellings identified with at least one risk factor.

Discussion and Conclusion

Validity of Methods

The validity of the GIS' predictions as to which areas and dwellings may be classified as high-risk or low-risk in terms of the presence or absence of lead risk factors is currently being tested using dust sample analysis. Dwellings were divided into three risk categories based on the number of lead risk factors present—none, one or two, or more, corresponding to no, moderate, and high risk. Approximately 100 addresses from each group were randomly selected and letters were sent requesting assistance in the research. Unfortunately, the response rate has been poor, around 20%. Much of this is related to the elderly age structure of the two case study areas—many of the householders felt that the study was not relevant to them, while many others undoubtedly had security concerns. This was not helped by media reports of bogus charity collectors and similar scams, which were prominent at the time the survey was conducted.

Limitations

The most obvious limitation is the lack of data on the actual addresses of the target population (i.e., children under 5). Confidentiality concerns mean that all identifying information collected by the Australian Bureau of Statistics for the Census is destroyed. In addition, these data are only collected once every five years (although this is a distinct advantage over American Census data, collected every 10 years). However, it may be

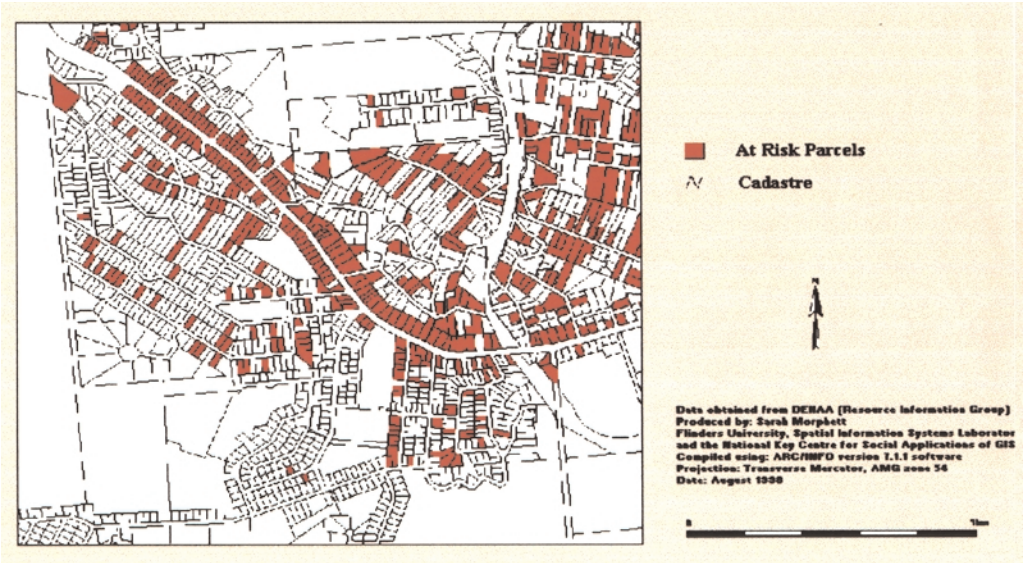


Figure 7 Detailed view of all risk factors, Mitcham LGA.

possible to obtain the addresses of young children via the records for immunization for childhood diseases held by local governments and other organizations.

Another limitation is the use of dust samples in a dwelling as an indicator of PbB. The literature is divided over the utility of dust as an indicator; however, circumstances are such that while the value of blood sampling is appreciated, dust sampling is the only viable alternative. The United States is fortunate in its access to good data on PbB.

The lead risk model presented here is the first step in developing a lead risk model for Adelaide and other urban centers in South Australia. It can be refined by weighting the factors and incorporating other parameters such as wind direction and strength, presence of traffic lights (because the presence of stationary traffic is an important factor in airborne lead [5]), historical land use data, materials of roofs and walls, and road widths. Even in its present form, this model would greatly improve the efficiency of any expenditure on lead as a health problem. The model makes it possible to mail information to specific households, rather than all households. Obviously, we cannot neglect the role of risk factors beyond the domain of GIS, such as the cleanliness of homes and the occupations and hobbies of parents. But the use of GIS in lead prevention programs offers a great deal in terms of targeting of environmental lead factors. This is at least half the battle.

Acknowledgments

I would like to thank Steve Fildes, Sarah Morphett, and David Schroeder of the Spatial Information Systems Laboratory in the School of Geography, Population and Environmental Management at Flinders University for their research assistance and technical assistance with this project.

References

1. Baker NJ de C. 1995. A 13 year review of childhood lead poisoning in Christchurch and Nelson. *New Zealand Medical Journal* 108(1002):249–51.
2. Graef JW. 1997. Foreword. In: *Getting the lead out.* Ed. I Kessel, JT O'Connor. New York: Plenum Press.
3. Donovan J. 1996. No lead is good lead. *Medical Journal of Australia* 64(7):390–1.
4. Nur Ahmed, Australian Institute of Petroleum. 1998. Personal communication (telephone and e-mail). May 11–12.
5. Alder RJ, Dillon JA, Loomer S, Poon HC, Robertson JM. 1993. An analysis of blood lead data in clinical records by external data on lead pipes and age of household. *Journal of Exposure Analysis and Environmental Epidemiology* 3(3):299–314.
6. Donovan J. 1996. *Lead in Australian children: Report on the National Survey of Lead in Children.* Canberra, Australia: Australian Institute of Health and Welfare.
7. Gulson BL, Mizon KJ, Korsch MJ, Howarth D. 1996. Importance of monitoring family members in establishing sources and pathways of lead in blood. *Science of the Total Environment* 188(2–3):173–82.
8. Gulson BL, Davis JJ, Bawden-Smith J. 1995. Paint as a source of recontamination of houses in urban environments and its role in maintaining elevated blood leads in children. *Science of the Total Environment* 164(3):221–35.
9. Brody DJ, Pirkle JL, Kramer RA, Flegal KM, Matte TD, Gunter EW, Paschal DC. 1994. Blood lead levels in the US population. Phase 1 of the Third National Health and Nutrition Examination Survey. *Journal of the American Medical Association* 272(4):277–83.
10. Mira M, Bawden-Smith J, Causer J, Alperstein G, Karr M, Snitch P, Waller G, Fett MJ. 1996. Blood lead concentrations of preschool children in central and southern Sydney. *Medical Journal of Australia* 164(7):399–402.
11. Willis FR, Rossi E, Bulsara M, Slattery MJ. 1995. The Fremantle lead study. *Journal of Paediatrics and Child Health* 31(4):326–31.
12. Rosen JF. 1995. Adverse health effects of lead at low exposure levels: Trends in the management of childhood lead poisoning. *Toxicology* 97(1–3):11–7.
13. Casey R, Wiley C, Rutstein R, Pinto-Martin J. 1994. Prevalence of lead poisoning in an urban cohort of infants with high socioeconomic status. *Clinical Pediatrics* 33(8):480–4.
14. Sargent JD, Brown MJ, Freeman JL, Bailey A, Goodman D, Freeman DH Jr. 1995. Childhood lead poisoning in Massachusetts communities: Its association with sociodemographic and housing characteristics. *American Journal of Public Health* 85(4):528–34.
15. France EK, Gitterman BA, Melinkovich P, Wright RA. 1996. The accuracy of a lead questionnaire in predicting elevated pediatric blood lead levels. *Archives of Pediatrics and Adolescent Medicine* 150(9):958–63.
16. Schaffer SJ, Kincaid MS, Endres N, Weitzman M. 1996. Lead poisoning risk determination in a rural setting. *Pediatrics* 97(1):84–90.
17. Snyder DC, Mohle-Boetani JC, Palla B, Fenstersheib M. 1995. Development of a population-specific risk assessment to predict elevated blood lead levels in Santa Clara County, California. *Pediatrics* 96(4 Part 1):643–8.
18. Threlfall T, Kent N, Garcia-Webb P, Byrnes E, Psaila-Savona P. 1993. Blood lead levels in children in Perth, Western Australia. *Australian Journal of Public Health* 17(4):379–81.

19. Edwards-Bert P, Calder IC, Maynard EJ. 1993. *National review of public exposure to lead in Australia.* Adelaide, South Australia: South Australian Health Commission.
20. Guthe WG, Tucker RK, Murphy EA, England R, Stevenson E, Luckhardt JC. 1992. Reassessment of lead exposure in New Jersey using GIS technology. *Environmental Research* 59(2):318–25.
21. Hanchette C. 1994. *GIS modeling of lead poisoning risk factors.* Paper presented at the Second Comprehensive National Conference on Building a Lead-Safe Future. May 16–18, 1994. Washington, DC.
22. Hanchette C. 1997. A predictive model of lead poisoning risk in North Carolina: Validation and evaluation. In: *Proceedings of the International Symposium on Computer Mapping in Epidemiology and Environmental Health.* Ed. RT Aangeenbrug, PE Leaverton, TJ Mason. Alexandria, VA: World Computer Graphics Foundation.
23. Padgett DA. 1997. Geographic information systems techniques for delineating hot spots of childhood lead-soil exposure sources. In: *Proceedings of the International Symposium on Computer Mapping in Epidemiology and Environmental Health.* Ed. RT Aangeenbrug, PE Leaverton, TJ Mason. Alexandria, VA: World Computer Graphics Foundation.
24. Wartenburg D. 1992. Screening for lead exposure using a GIS. *Environmental Research* 59:310–7.
25. Covello VT, Merkhofer MW. 1993. *Risk assessment methods: Approaches for assessing health and environmental risks.* New York: Plenum Press.
26. Australian Bureau of Statistics. 1991. *1991 census of population and housing.* Belconnen, Australian Capital Territory: Australian Bureau of Statistics.
27. Berry M. 1994. *Reducing lead exposure in Australia.* Commonwealth Department of Human Services and Health. Canberra, Australia: Australian Government Publishing Service.
28. Department of Environment and Natural Resources (DENR). 1997. *1997 digital cadastre database.* Adelaide, Australia: DENR.

Drinking Water Source Protection and GIS in Reno County, Kansas

Daniel L Partridge, RS (1),* Michael Mathews (2)
(1) Reno County Health Department, Hutchinson, KS; (2) Reno County Information Services, Hutchinson, KS

Abstract

Reno County has a population of approximately 60,000 people, 40,000 of whom live in the city of Hutchinson. Hutchinson's water system is dependent entirely upon groundwater for its drinking water supply. The Equus Beds Aquifer, from which groundwater is withdrawn, is a shallow alluvial aquifer formed by the Arkansas River and is overlain by sandy soils. As a result, local groundwater is vulnerable to contamination. Three of Hutchinson's twenty public water supply wells are not currently used due to the presence of volatile organic compounds. In response to this threat, county and city governments formed a wellhead protection committee to develop a drinking water protection plan. The goal of this plan is to limit, as much as possible, future threats to groundwater. The US Environmental Protection Agency describes wellhead protection planning as a five-step process. These steps are: form a planning team; delineate the protection area; inventory the potential risks of contamination within this area; manage the protection area; and plan for the future. Geographic information system (GIS) software was used during the initial consensus-building phase of this project to produce maps for use at public information and planning meetings. GIS mapping was also used to store and display the results of the potential contamination source survey as well as the wellhead protection zone. This protection zone was derived from a computer simulation of three- and five-year zones of groundwater capture for each public water supply well.

Keywords: wellhead protection, non-point source pollution, groundwater, drinking water, source water protection

Program History

The Reno County (Kansas) Health Department, working in cooperation with the city of Hutchinson, Kansas, established a wellhead protection (WHP) steering committee in 1996. The Health Department has funded the WHP program with Local Environmental Protection grants from the Kansas Water Office, along with in-kind services from a variety of local agencies. The steering committee has formed an advisory board whose function is to provide public input into each stage of the plan's development. Volunteers from the Retired Senior Volunteer Program (RSVP) and the community college were trained and used to conduct door-to-door inventories that were then used in this project to evaluate the severity of the potential risks to Hutchinson's water supply.

The first task facing the WHP steering committee was to determine the source of

* Daniel L Partridge, Reno County Health Department, 209 West 2nd Ave., Hutchinson, KS 67501 USA; (p) 316-694-2900; (f) 316-694-2901; E-mail: renocohd01@mindspring.com

the public water supply. The second task was then to describe the risks to water quality within that area. With the exception of chlorination, groundwater is not treated prior to its use as drinking water. Program goals are:

- Identify the surface area having the greatest impact on drinking water quality.
- Identify the threats to groundwater quality within that area.
- Develop and implement a strategy of education and pollution prevention to reduce the impact of current activities on groundwater.
- Encourage groundwater-friendly growth and development in the WHP zone.

Problem

Hutchinson is dependent entirely upon groundwater for its source of drinking water. Three of the 20 municipal wells are not currently used due to the presence of volatile organic compounds. A fourth is not used because a high mineral content gives it an unpleasant taste (Figure 1). The city has invested many of its resources into the development of a series of wells outside the city limits. Currently the water quality in these wells is good. If quality concerns arise in this second wellfield, the alternative would be the construction of a costly water treatment plant and the restructuring of the water distribution system.

Data Development and Analysis

The decision of what information to use, and how it was then tied together spatially, was based on the need to spend as little time creating new data files as possible. Some of the information was already complete before the project began but resided on different systems and file structures. Two of the datasets specifically created for this project were "zones of capture," which identifies the WHP zone, and "potential pollutant inventory," a database created from information gathered in the door-to-door survey. Respondents were asked for information on equipment and practices associated with the potential for groundwater contamination. Table 1 displays the data gathered in that survey.

Once the data were collected and developed, maps were generated that spatially displayed the distribution of each inventoried pollutant through the WHP zone. Fertilizer and pesticide usage were tracked according to their number of applications per year. No attempt was made to determine if these chemicals were overapplied. Querying the inventory database and selecting the associated parcels produced maps showing the distribution of each potential pollutant. Approximately 85% of the 800 parcels in the surveyed area responded to the potential contaminant risk survey. Information on the remaining parcels was not obtained due either to property owner resistance or inability to contact the owner.

Implementation

The results of our efforts so far have shown that the major threats to the northwest wellfield now and for the near future are agricultural and suburban lawn and garden fertilizer usage, and the dependence of homeowners on private septic systems. Best

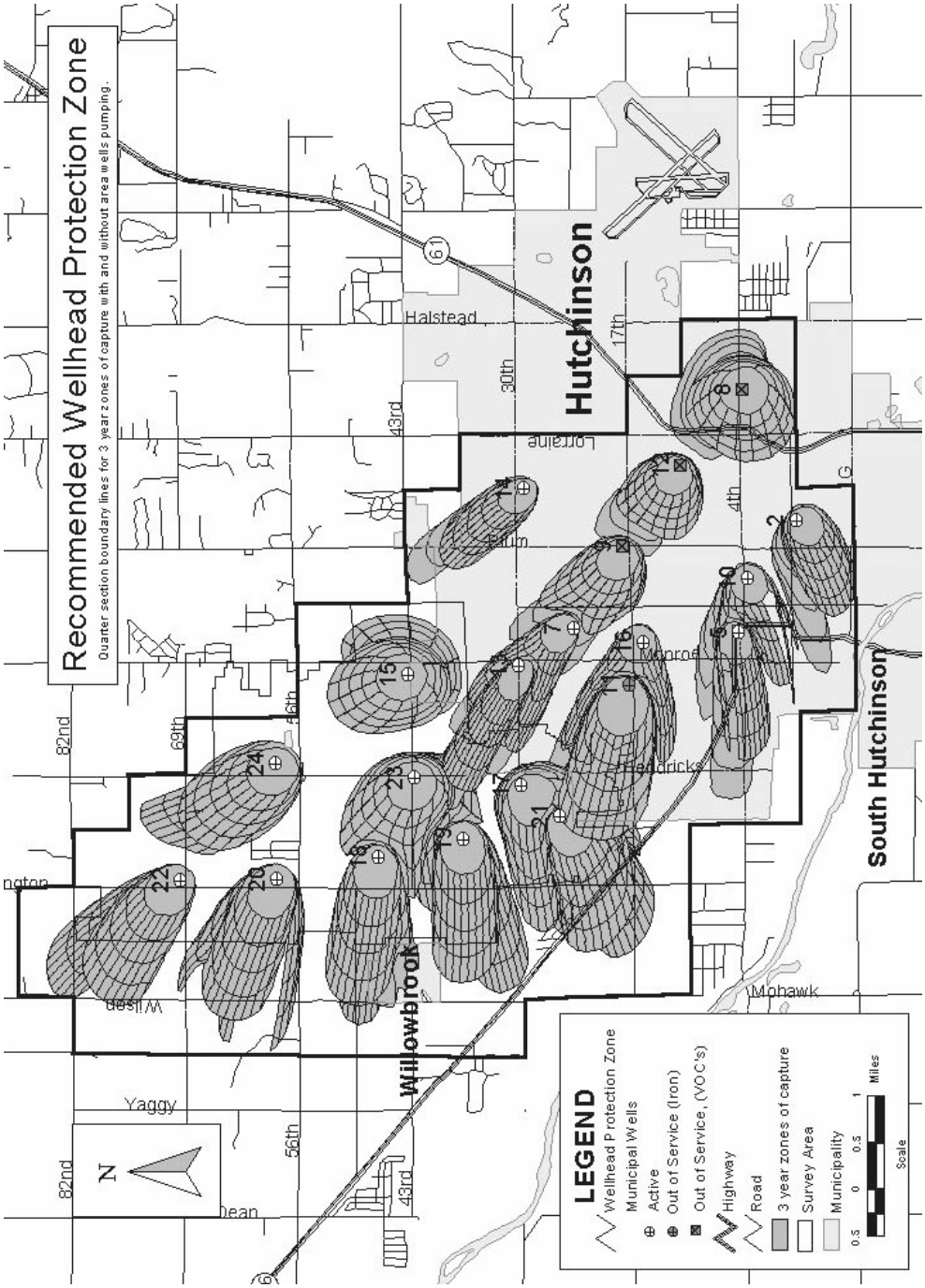


Figure 1 Recommended wellhead protection zone and survey area.

Table 1 Incidence Rates of Potential Sources of Groundwater Contamination, Northwest Wellfield Survey Area, Reno County, KS, 1996

Potential Source of Groundwater Contamination	City Parcels[a] with Potential Contamination Source	Potential Contamination Sources in City Parcels[a]	County Parcels[a] with Potential Contamination Source	Potential Contamination Sources in County Parcels[a]	Total Parcels[a] with Potential Contamination Source	Total Potential Contamination Sources in Parcels[a]	Comments
Fertilizer use	254	1.8[b]	32	1.1[b]	286	1.7[b,c]	Applied to 1,715 acres
Domestic water well	110	117	87	106	197	223	
Other water well	189	191	28	30	217	221	
Pesticide use	130	1.2[b]	9	1[b]	139	1.2[b,c]	Applied to 439 acres
Septic system	12	13	79	86	91	99	
Other	35	N/A	24	N/A	59	60	Primarily pipelines
Above-ground tank	1	4	16	24	17	28	
Abandoned water well	5	7	12	17	17	24	
Stream	1	1	17	17	18	18	
Chemical storage	9	10	0	0	9	10	
Grain storage bin	0	0	4	9	4	9	
Chemical storage facility	2	2	1	1	3	3	
Underground storage tank	1	1	2	2	3	3	
Septage disposal	0	0	2	2	2	2	
Animal feedlot	0	0	2	2	2	2	
Salvage yard	0	0	1	1	1	1	
Private dump	0	0	1	1	1	1	
Airstrip	0	0	1	1	1	1	
Cemetery	0	0	1	1	1	1	
Injection well	0	0	0	0	0	0	
Oil/gas well	0	0	0	0	0	0	
Quarry/sand pit	0	0	0	0	0	0	
Pit privy	0	0	0	0	0	0	
Lagoon	0	0	0	0	0	0	
Chemigation	0	0	0	0	0	0	

[a] Parcels within survey area

[b] Number of applications per year

[c] Weighted average

N/A = Not applicable

management practices (BMPs) for both of these issues have been developed by the steering committee and have been partially implemented by the Health Department and Conservation District. The focus now is to increase the number of landowners implementing these BMPs. This can be accomplished by increasing local funding of voluntary cost share programs, which reimburse property owners for a portion of the costs of implementing BMPs; raising awareness of the problem through education; and amending local codes to raise mandatory minimum standards of construction and operation, to reduce bacterial and nitrate contamination from septic systems.

Education of the public regarding groundwater quantity and quality issues was recognized early on as a key element to the program's success. Toward that goal, a partnership with the Hutchinson School District to provide groundwater education at the elementary and middle school levels began with the 1997–1998 school year. This program is not just for Hutchinson but is serving as a model for the 13 other towns and rural water districts of Reno County. What began as a program to protect Hutchinson's water supply has evolved into a true public health bargain that allows public water suppliers to meet upcoming US Environmental Protection Agency deadlines to meet Safe Drinking Water Act regulations on source water protection.

Acknowledgments

This project was funded in part by a grant from the Kansas Water Office.

Geographic Database for Public Health in Portugal: Public Health National Charter

Ana Patuleia (1),* Marco Painho (2)

(1) Master's student, ISEGI, New University of Lisbon, Lisbon, Portugal; (2) Associate Professor, ISEGI, New University of Lisbon, Lisbon, Portugal

Abstract

Portugal's computerized Public Health National Charter is a dataset including population health parameters, health services, and socioeconomic and environmental factors that directly or indirectly affect human health. It presents data detailed to the administrative level of NUTS IV. The purpose of this article is to illustrate the charter's versatility and strength. It allows users to access a great number of spatially referenced indicators for producing maps and graphs and for evaluating various models. The collection, storage, and manipulation of geographic health data and other health-related information can influence the progress of health surveillance and environmental health assessment, as well as the allocation of health resources, as recognized in the European Charter on Environment and Health. This database is more flexible than traditional ones because it enables the user to select and modify the data presentation options. In addition to offering a set of spatially referenced information, the charter performs a valuable service in collecting and compiling data from myriad sources and making it available from one central repository.

Keywords: public health, spatial analysis, health surveillance, environmental health, epidemiology

Methods

Portugal's Public Health National Charter is based on the same development principles as other geographic information systems. Criteria that were considered in its development include characterizing potential users; determining and collecting information for the database; designing, codifying, and installing the system; and evaluating its operation and use to improve its performance and value to users. The Public Health National Charter presents data separated to administrative level IV of EUROSTAT's NUTS[1] classification system.

Goals underlying the establishment of the charter included:

- Compiling a large and diverse dataset by collecting information from the various entities that are directly or indirectly linked with health.
- Providing access to a wide spectrum of users.
- Integrating GIS into health and health-related issues, not only as a tool for making maps and graphs, performing spatial analysis, and modelling, but also for making and supporting health- and policy-related decisions.

* Ana Patuleia, ISEGI-UNL, Trav. Estevão Pinto–Campolide, 1200 Lisboa, Portugal; (p) 351-1-3870261; (f) 351-1-3872140; E-mail: m318@isegi.unl.pt

[1] NUTS refers to Territorial Units Nomenclature and is divided into different levels starting with NUTS I for countries. NUTS IV is for municipalities. Portugal has 275 municipalities, not including Madeira and Açores.

A driving force behind this effort is the introduction of the charter on the World Wide Web. This will permit different kinds of users, including those directly and indirectly involved with health and environmental issues, to have access to the database.

Data needs were identified through bibliographic searches. Of particular value were those data sources of the World Health Organization (WHO), such as the Health Environment Geographic Information System (HEGIS) project. It quickly became apparent that sources containing data on health and health-related issues are voluminous. In addition, the level of detail—NUTS IV—made the task of data collection more difficult. The enormous quantity of work to be done made it impossible to include all the parameters initially proposed for the database.

The information used in the charter is either published, or publicly available but not published. Data validity was assumed to be the responsibility of the sources; despite the potential for some inaccuracies, it seemed preferable to use the best available information. The spatial data were taken from the *Atlas do Ambiente* at the scale of 1:1,000,000 (1). Some simplifications were made so that municipalities could be represented by a single polygon. Each polygon has a designation (the name of the municipality); a 9-digit code provided by the Instituto Nacional de Estatística (INE), Portugal's national statistical institute, composed of 3 digits for the region, 3 for the district, and 3 for the municipality; and a NUTS IV code from EUROSTAT.

A set of over 700 variables was gathered. The attribute data refer to 1996, whenever possible. Exceptions include figures for mortality, which use data over a five-year period to create a statistical indicator, and figures for diseases with mandatory notification, which used data from 1994 to 1996. The application was developed using ArcView GIS (ESRI, Redlands, CA).

Results

The simple existence of this application has opened many doors for the use of GIS in health. In fact, Portugal's Ministry of Health is currently constructing a geographic database.

This article, which summarizes work presented at the August 1998 Third National Conference on GIS in Public Health, demonstrates some the possibilities of the National Charter database. While not a precise scientific study, it portrays the charter's potential value to health professionals. As an illustration, we include here one of the study profiles: Malta fever, or brucellosis.

Malta fever is one of the diseases with mandatory notification that, for a variety of reasons not discussed here, is underreported. The real numbers for the incidence rate are five to six times higher than those officially recorded.

Malta fever is a bacterial disease with both acute or chronic forms. It is related to contact with cattle and cattle products. Infection results from the ingestion of fresh cheese or infected meat, or from contact with secretions or other products from infected animals. In Portugal, the main sources are cows, goats, and sometimes pigs. Fever, weakness, and pain, especially in the joints, are characteristic symptoms (2).

To prevent this disease, it is essential to treat the milk to be consumed by humans, to vaccinate young animals, and to eliminate all infected animals, including healthy livestock from the same herd (3). For many reasons, but mainly because of economic factors, this prevention program has been difficult to implement.

In the fight against this disease, one of the main objectives is to use GIS to highlight the municipalities where Malta fever is endemic and where the disease frequency has been high over the years (Figure 1) (5). The variation of the incidence rate is shown in Figure 2. The municipalities that over the last three years had an incidence rate higher than 51.55 per 100,000 inhabitants were selected. These are illustrated in Figure 3 and highlighted in Figure 4 (7).

The active population, which is composed of those between the ages of 14 and 65, is the group most susceptible to contracting the disease. Many inhabitants in these areas have cattle, even if they are employed in occupations other than handling or raising cattle. This makes it important to identify the endemic municipalities, and to understand those indices that indicate higher rates of Malta fever. The charter provides ample opportunities for more detailed analysis than the example offered in this summary paper.

Conclusions

The development of the Public Health National Charter is a great opportunity and of immense practical value for all types of health professionals. As with any new database, however, it must be viewed as unfinished work since significant deficiencies remain. Nonetheless, it represents a positive opportunity and a significant step forward in the application of GIS to improving and better understanding health and health-related issues in Portugal.

Acknowledgments

We are grateful to Dr. William D Henriques, without whose support this poster presentation would not have been possible.

References

1. Ministério do Ambiente, Direcção Geral do Ambiente. *Atlas do Ambiente.* www.dga.min-amb.pt/arvore.html.
2. Gonçalves Ferreira FA. 1990. *Moderna saúde pública.* Lisboa: Fundação Calouste Gulbenkian.
3. Patuleia A. 1998. Public health national charter. In: *Proceedings of the Conference on GIS PLANET 98.* September 1998, Lisboa, Portugal: Imersiva. www.imersiva.ch/gisplanet.html.
4. Ministério de Sáude, Direcção Geral de Sáude. 1994–1996.
5. Cliff AD, Haggett P. 1993. *Atlas of disease distributions.* Oxford: Blackwell Publishers.
6. Instituto Nacional de Estatística. 1995.
7. Patuleia A. 1999. *Public health national charter.* Master's thesis. Lisboa, Portugal: ISEGI-UNL.

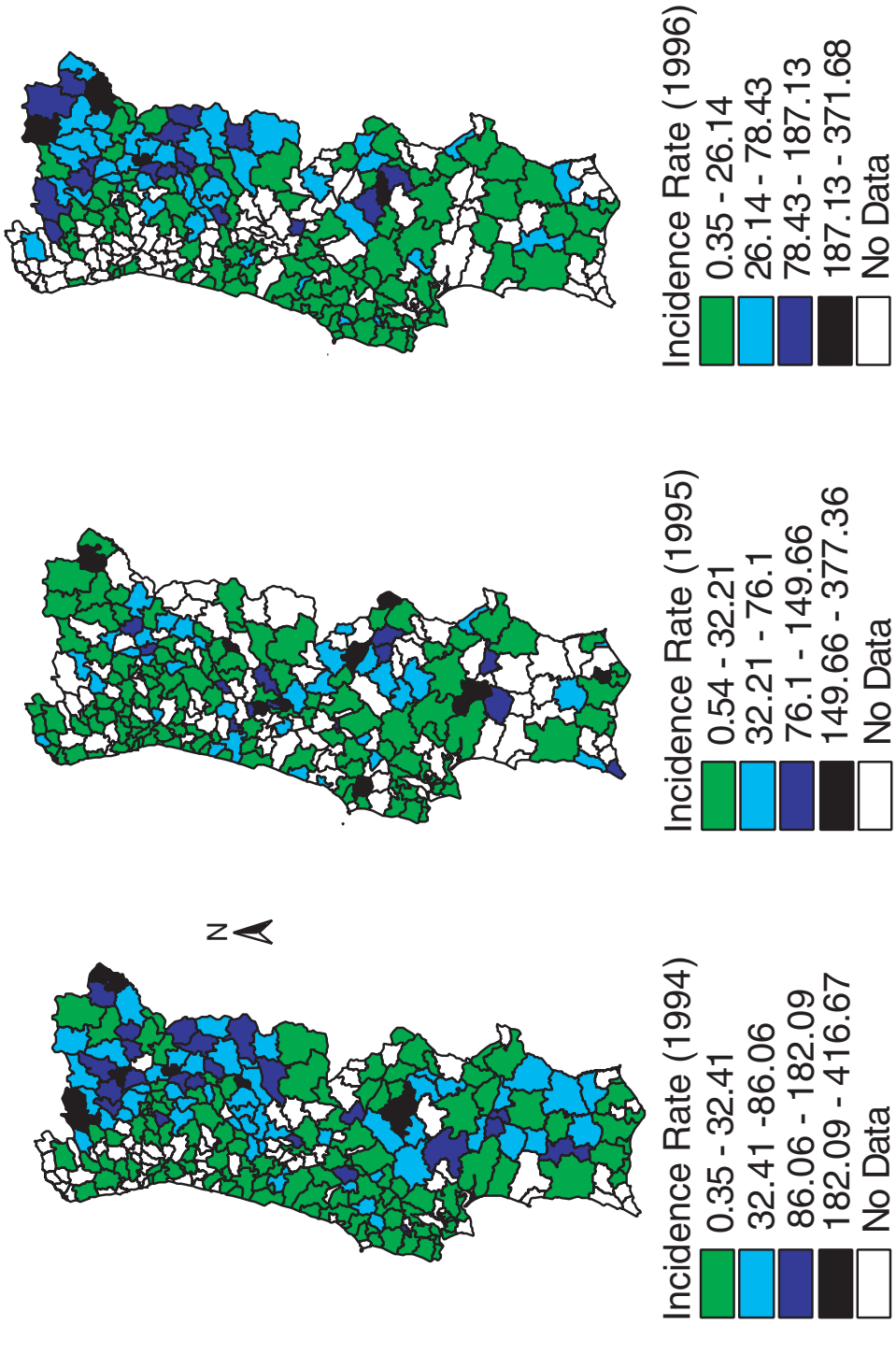


Figure 1 Malta fever reports per 1,000,000 inhabitants, Portugal, 1994–1996. Source: (4).

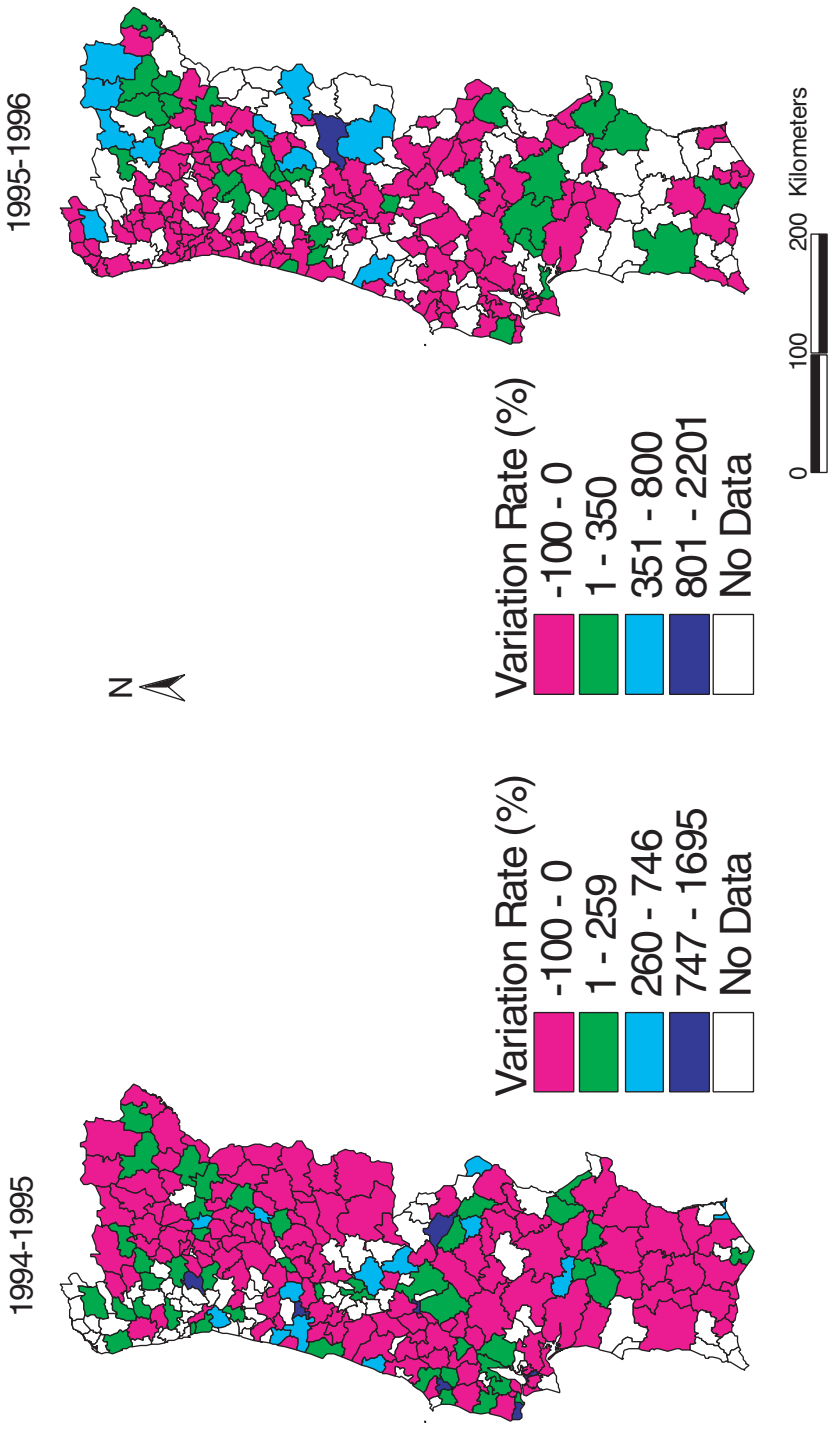


Figure 2 Year-on-year percent variation in Malta fever rate, by municipality, Portugal, 1994–1996. Source: (4).

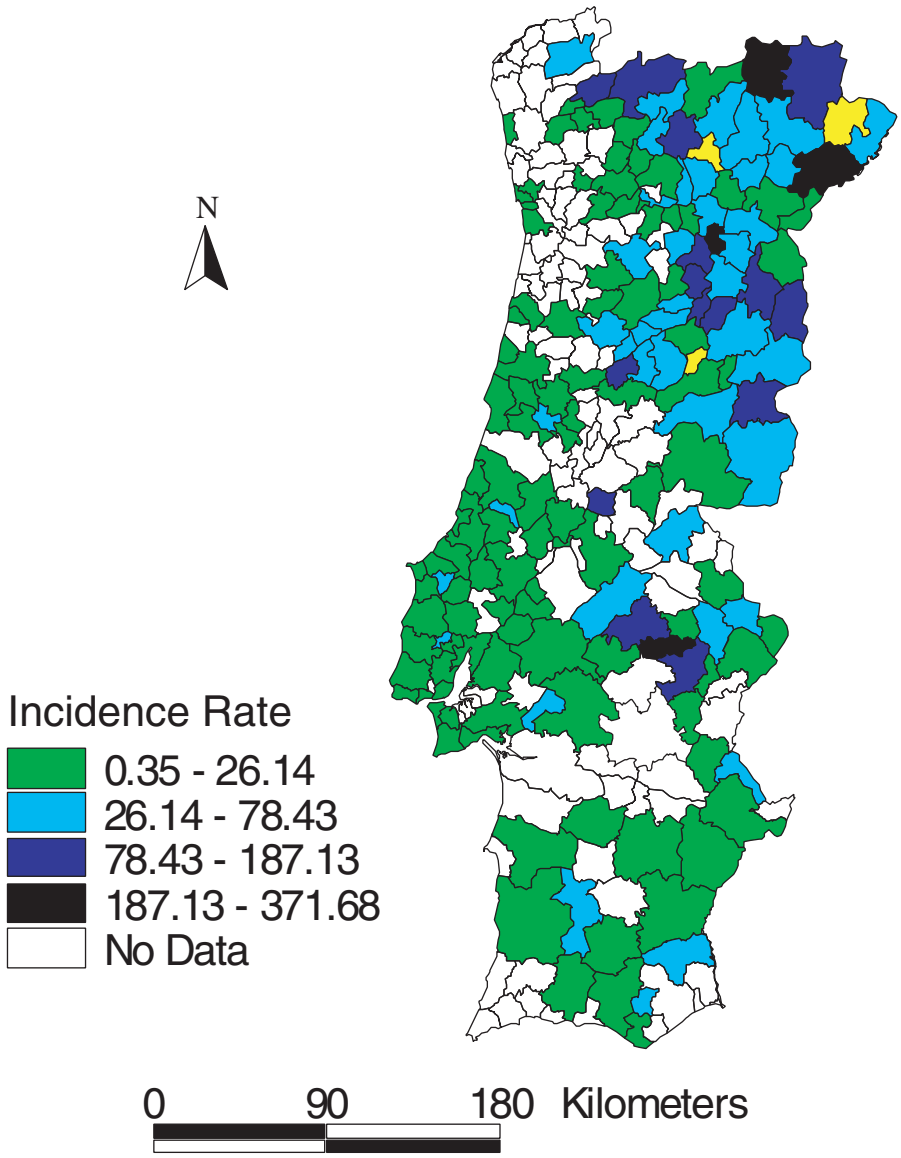


Figure 3 Endemic municipalities for Malta fever, where the incidence rate exceeded 51.55 per 100,000 inhabitants, Portugal, 1994–1996. Source: (4).

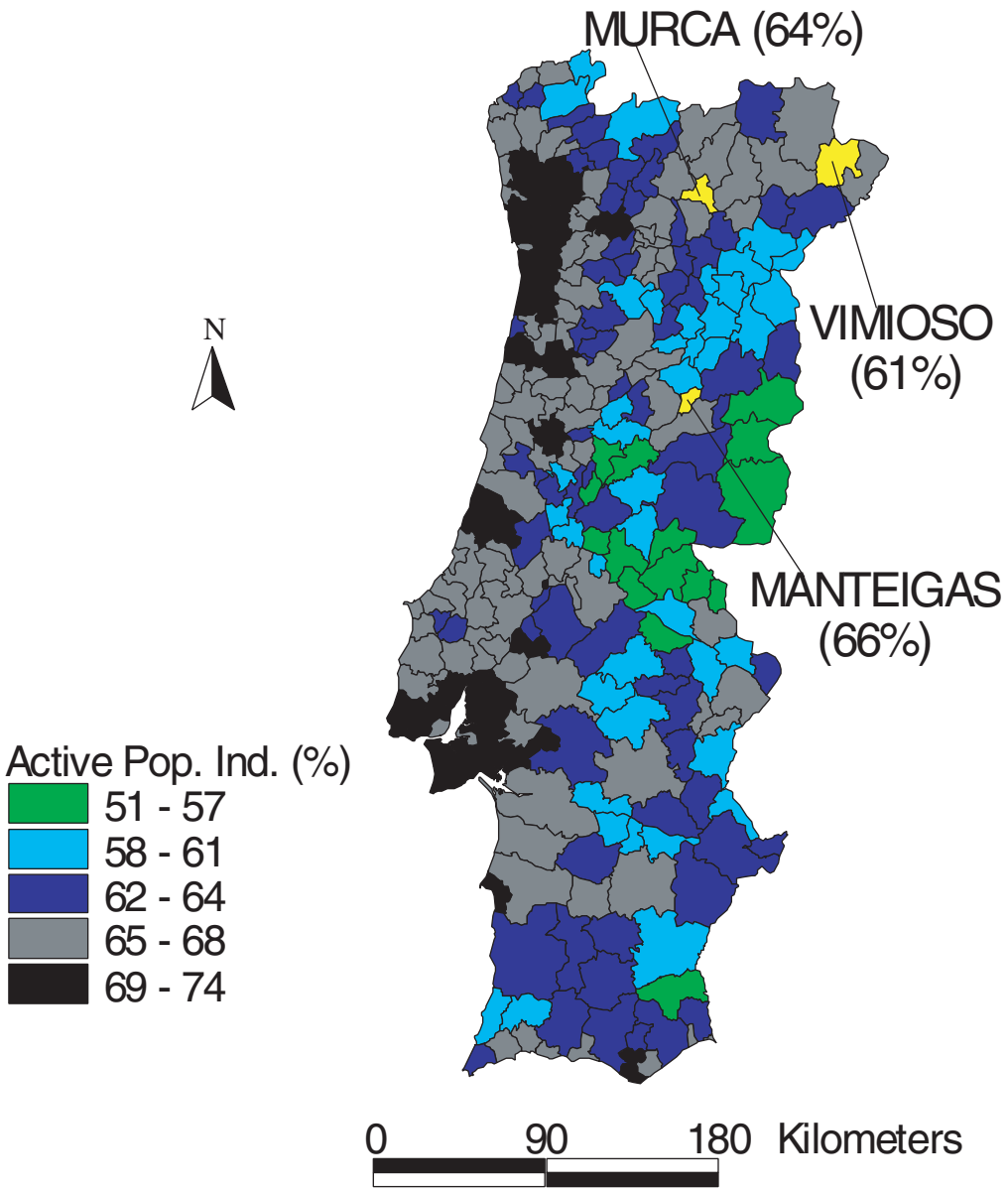


Figure 4 Active population index, 1995. Source (6).

Screening for Childhood Lead Exposure Using a Geographic Information System and Internet Technology

Stephen A Scott (1),* Randy Knippel (2)
(1) Environmental Management Department, Dakota County, Apple Valley, MN; (2) Survey and Land Information Department, Dakota County, Apple Valley, MN

Abstract

Childhood exposure to lead remains a critical health issue to the United States population. The widespread distribution of lead throughout the environment and the profound physiological and cognitive effects it has on children, even at low levels, warrant an aggressive approach toward identifying lead hazards in the environment, determining the population at risk for lead exposure, and developing strategies to prevent exposure. In 1991, the Centers for Disease Control and Prevention (CDC) issued a recommendation supporting near-universal blood lead testing of children under the age of six years. This recommendation is currently under revision, having been criticized as ineffective and unnecessary. Recently, the CDC has issued an updated lead screening guidance document recommending the evaluation of individual and residential exposure risks in order to target screening efforts in areas where lead risks are significant. Dakota County (Minnesota) staff have developed a geographic information system application that spatially evaluates lead sources and returns a screening recommendation based on an individual's risk of exposure to lead. Health professionals can obtain an individual's exposure risk from an Internet Web site simply by entering a residential address. The Web site returns an overall risk value for the specific location by incorporating these lead risk elements. The use of a standard Web browser allows for the cost-effective delivery of accurate and current information. The Internet server also allows the data and application to be updated as needed at a central location without the need to redistribute the data.

Keywords: targeted-screening, lead poisoning, environmental lead

Introduction

Childhood exposure to lead remains a critical health issue confronting the United States population. Although this exposure, as measured by blood lead levels, has fallen dramatically since the 1970s (1,2), significant numbers of children continue to be exposed to toxic levels of lead. Overexposure to lead, however, is not equally distributed in all segments of the population. A 1988 report by staff of the Agency for Toxic Substances and Disease Registry (ATSDR) found that childhood blood lead levels were significantly associated with race, family income, residence inside or outside a metropolitan central city, and, for city residents, the size of the metropolitan area (3). These findings are supported by the results of the recent National Health and Nutrition Examination Surveys (NHANES), which found elevated blood lead levels disproportionately distributed among children who are "poor, non-Hispanic black or Mexican American, living in large metropolitan areas, or living in older housing" (4).

* Stephen A Scott, Dakota County Environmental Management Department, 14955 Galaxie Ave., Apple Valley, MN 55124 USA; (p) 612-891-7537; (f) 612-891-7588; E-mail: steve.scott@co.dakota.mn.us

Within the past two decades, research on childhood lead poisoning has identified profound, adverse physiological and cognitive affects from low-level exposure (5–11). The neurotoxicity of lead is of considerable concern. Several prospective studies have identified significant dose-dependent relationships between lead exposure and impaired neurobehavioral and psychological functioning (12–15). Followup studies of adults with asymptomatic exposure to lead as children have demonstrated that children with early elevated lead exposure were at risk for later educational deficiencies including failure to graduate and poorer reading abilities (16,17). The findings of these and other studies suggest that the deleterious effects of early childhood exposure to low levels of lead result in profound and long-lasting impacts on learning and behavior.

In response to these data, in 1991, the Centers for Disease Control and Prevention (CDC) issued revised guidelines for management of lead poisoning. These new guidelines significantly lowered the levels at which children are considered at risk for lead toxicity (18). The guidelines led to a recommendation to screen virtually all children between the ages of 12 months and 72 months. The CDC also recommended that physicians administer a five-part questionnaire in order to classify children as being at high risk or low risk for lead exposure.

The CDC's recommendation of near-universal lead toxicity screening for children under six has proven to be very controversial. Critics of universal screening have challenged the necessity of the recommendation, citing the dramatic reduction of blood lead levels in the nation's preschool children since 1976. Two additional reasons frequently cited are the low prevalence of elevated blood lead levels in much of the population and the high costs of universal screening (19).

Because of these criticisms, the CDC's recommendation has not been followed by many primary health care providers. Relatively few children in any region of the country have been screened for lead exposure. In a nationwide survey conducted by the CDC in 1994, only about one-fourth of parents reported that their young children had been screened (20). In addition, a survey of New Jersey pediatricians and family practitioners found that only 42% of pediatricians and 24% of family practitioners reported screening the majority of children seen in their practice by the age of two (21).

To address these issues and to further improve the use of screening to identify and prevent childhood lead poisoning, the CDC issued an updated lead screening guidance document in November 1997, entitled Screening Young Children for Lead Poisoning: Guidance for State and Local Public Health Officials (20). The guidance document refers to the use of census tract data, personal risk questionnaires, the socioeconomic status of patients, and other data, to identify and deliver proper services to children at greatest risk for lead exposure.

Dakota County (Minnesota) staff have developed a geographic information system (GIS) application to assist in the identification of populations at high risk for exposure to lead hazards. The application combines information representing risk factors for lead exposure, including the age and distribution of housing, location of lead-emitting sources, distribution of lead-contaminated soils, case mapping of lead poisoning, and demographic data. An individual's overall lead exposure risk can be modeled and classified by incorporating the various contributors to lead overexposure as they relate to the individual's home address. The application is accessed using a standard Web browser, allowing for cost-effective delivery of accurate and current information. Dakota County is in the process of applying for funding from the Minnesota

Department of Health through a grant from the State and Community-Based Childhood Lead Poisoning Prevention Program, which is administered by the CDC. If received, the funding will be used to assess the usefulness of the model and to evaluate the risk factors employed in the model.

Setting

Dakota County is located in Minnesota, on the south side of the Minneapolis/St. Paul seven-county metropolitan area. The county was originally settled in the mid-1850s. Its 1997 population is estimated at 330,000. It is the second-fastest-growing county in the 87-county state, and its population is younger than the statewide and metropolitan averages. Though Dakota County was once largely rural, the second half of this century has seen the intensive suburbanization of the county's northern half. The county is home to 40 regulated solid waste facilities, including two of the state's largest sanitary landfills. The county also licenses and inspects nearly 1,300 hazardous waste generators and facilities, including a large secondary lead smelter. The housing stock is considerably newer than that of the Twin Cities, with approximately 10% of the homes constructed prior to 1950. However, concentrations of older housing are found in cities developed when streetcars were the primary mode of transportation, as well as in communities that historically served as agricultural centers.

Overview of the Lead Exposure Risk Assessment Application

The GIS application, developed in Dakota County, is intended as a tool to help determine when blood lead testing is necessary. The application does not, however, account for all the potential risks that contribute to lead poisoning, and health care providers must still make their own assessments to determine whether or not to perform blood lead testing.

The GIS application was developed in Microsoft Visual Basic and is accessed using a standard Web browser through an Internet Web site. The Web page contains graphics, text, and input fields. A map of the county is depicted as a graphic that can be clicked on with the mouse to perform basic map viewing functions. A child's cumulative risk of exposure to environmental lead can be determined from a residential address.

Elements of a Lead Exposure Risk Model (Lead Sources)

Lead-Based Paint

Lead-based paints are considered the most significant and widespread source of childhood lead exposure. The influence of lead-based paint is most often related to an individual's place of residence. Other places, such as relatives' homes, daycare centers, and schools, can also serve as a source of exposure for children.

The lead content of residential paints can be generally related to their period of production, and it is estimated that approximately three-fourths of the housing built before 1980 contains some lead-based paint. The lead content of paint was largely unregulated before the enactment of voluntary restrictions by the paint industry in the mid-1950s. Prior to these restrictions, paints routinely contained up to 50% lead by dry weight. In

response to increased regulatory pressures, the lead content of paints was gradually reduced. By 1978, lead had been eliminated from paints produced for residential application. For the purposes of the lead risk model, it is assumed that a correlation exists between the year of a building's construction and the presence and concentration of lead in the paints.

For the purposes of tax collection, the county assessor's office collects parcel information that includes the property's address, value, ownership, and building data. Information regarding a home's year of construction can, therefore, be determined from the child's home address.

Lead-Contaminated Soils

Lead-contaminated soils are also important sources of lead exposure, primarily in older, urban residential areas. Their overall influence on lead exposure, however, is considered to be less than the influence of lead-based paint. To reflect this lesser influence, a weighting factor is applied to this risk feature.

Soil can be contaminated by lead from various sources, including weathered lead-based paint and the historical deposition of lead fallout from combustion sources such as incinerators and automobiles fueled with leaded gasoline. Because lead does not decay in the environment, deposits of lead from these sources accumulate over time in the upper 5 cm of undisturbed soils. Urban residential soils are generally more lead-contaminated than rural soils (18). This is because urban areas tend to have a greater concentration and longer operational history of emission sources, as well as higher traffic density and a larger stock of older housing.

The distribution of lead-contaminated soils can be approximated for urban residential areas based on the extent of urban development at a time when the inputs from the various sources are expected to have been the greatest. For the purposes of the model, urban areas developed prior to 1970 are assumed to be associated with greater soil lead contamination than rural or urban areas developed after that time. The cutoff date of 1970 reflects the enactment of regulations that resulted in reduced inputs from lead-based paints and atmospheric fallout from automobile exhaust and industrial sources.

Industrial Point Sources

Soils contaminated by airborne lead emissions from industrial sources can be characterized as a release point with a subsequent downwind zone of influence representing an atmospheric fallout area. A release point would be used to represent known industrial emission sources, such as secondary lead smelters, sewage sludge incinerators, and coal-fired power plants.

The extent of the plume and the degree of overall risk assumed to be associated with a source would be based on the type of the emitting facility, the nature and quantities of its emissions, the prevailing winds, and the facility's history of operation. The areal extent and the level of risk associated with these facilities are approximated, due to the lack of detailed environmental studies of air and soil contamination adjacent to these industrial emission sources.

Industrial and Solid Waste Disposal Sites

Industrial and other solid waste disposal sites can serve as significant sources of

childhood lead exposure. Among Superfund sites, lead is the most frequently identified hazardous substance found in completed exposure pathways. The notable presence of lead at disposal sites is due to its persistence in the environment and its wide use and dissemination in industrialized countries.

Dakota County has created a countywide inventory of known solid and hazardous waste dumpsites. The inventory contains approximately 1,600 sites, ranging from farm dumps to extensive industrial/hazardous waste dumps. It includes sites associated with the operation of the large secondary lead smelter in the county, as well as smelter-related industries. Dumpsites and their surrounding areas are depicted as polygons. Dump attributes include waste characteristics, waste volume, and the potential for direct contact with waste or waste-impacted soils or water. Dumpsites possessing an increased potential for dispersal are assigned a buffer.

Blood Lead Level and Case Mapping

All current and available historic data regarding children's blood lead levels have been geocoded and included in the model. The Minnesota Department of Health provided blood lead testing data for all children in Dakota County screened between 1994 and 1997. These data are used in conjunction with birth record data to allow for the calculation of a blood lead screening "rate" for the at-risk population residing near the target residence. This information is provided to assist physicians in identifying areas where lead-poisoned children are concentrated and where screening efforts are inadequate.

Data Characteristics

Lead risk features are modeled as points and polygons. Points are used to represent the locations of known cases of elevated blood lead levels. The points are used to provide a summary of reported test results within a defined distance from the target location. (To protect confidential patient information, the points themselves are not shown.) Exposure risk features such as lead dumps, residual soil lead, industrial sources, and poverty are depicted as polygons. Parcels are also represented as polygons; the year in which they were built is retrieved to determine degree of risk.

Polygons are also used to represent regions of influence around a feature, either by buffering the feature itself or by representing the boundaries of a risk area such as a neighborhood or park service area. Each feature is modeled independently, to account for unique exposure risks associated with the lead source. Datasets are represented with a unique hatching pattern in which color gradation is used to represent degrees of risk. This method allows overlapping polygons to be visually distinguished and ranked.

Field verification of risk features is very important, because many features may not have been evaluated directly for actual lead exposure characteristics. These unverified exposure-risk features can still be incorporated into the model; efforts must be made, however, to better define their risk attributes or characteristics, as resources permit.

Determining Lead Risk Using the GIS Model

The application was developed to allow easy access, via the Internet, to detailed information regarding a child's risk of lead exposure. A query is initiated by entering a standard street address in the box next to the appropriate prompt. A certain amount of flexibility is allowed for matching misspelled street names. Because the matching

process, or geocoding, uses a street centerline containing segments with assigned address ranges, the address does not need to match an actual property address. This technique establishes a geographic location for any address number that falls within the address range of a given street segment.

The Web page also includes several yes/no questions that were derived from a sample questionnaire developed by the CDC. These questions appear as a series of statements, accompanied by checkboxes. The user answers the questions by checking the corresponding box for each statement that applies. The questions are included to account for potential sources of lead exposure that are not included in the model, such as a recent move, occupational exposure, or home renovation plans. An affirmative answer to the questions automatically returns a recommendation to screen.

Pressing the button labeled "Search" initiates a search for the location of the address entered by the user. If a location is found, the program captures all risk datasets using pre-established techniques for spatially related features. A screening recommendation is calculated by adding the risk factors of these risk features. For example, the risk values shown in Table 1 are associated with the year of construction of a residence. As displayed in Table 1, residing in a home constructed before 1951 is cause for a recommendation to screen. The model also retrieves the average building construction date for all parcels within a 200-foot radius of the target location and returns the count of the structures constructed before 1941. This value is not directly incorporated into the risk calculation, but may be useful in identifying exposure risks from improper paint removal—for example, in cases in which the target residence is a newer home that is surrounded by older homes.

Table 1 Dakota County Screening Recommendations Associated with Year of Home Construction

Year of Construction	Degree of Risk	Screening Recommendation
1950 or earlier	High	Recommended screening
1951–1978	Medium	Screen when associated with other risk features[a]
1979–present	Low	Screen not indicated

[a] Including plans to renovate or location in an area of residual soil lead

Other polygon features including lead dumps, residual soil lead, and industrial lead sources within 400 feet of the target location are identified and incorporated into the overall screen recommendation. These features are assigned a corresponding risk factor, which is applied directly to the total risk as shown in Table 2.

Table 2 Dakota County Screening Recommendations Associated with Environmental Sources of Lead

Environmental Source of Lead	Degree of Risk	Screening Recommendation
Industrial source, lead dump, etc.	High	Recommended Screening
Residual soil lead	Medium	Screen when associated with other risk features[a]

[a] Including residence in home constructed between 1951 and 1978

Residing within the assigned zone of influence of a lead source returns a recommendation to screen. Exposure to multiple medium-risk sources creates an additive risk of exposure, which also can result in a recommendation to screen. When the application completes the analysis, a dot is displayed on the map at the target location and the screening recommendation is displayed. The map extents are changed to be centered on the location at the same scale. The "Zoom to Location" button is provided to change the map extents to include a 200-foot radius around the selected location.

Basic Map Functions

A list of options is provided specifying the operations performed when the mouse is clicked over the map graphic. These include basic functions for zooming and panning. An option is also provided for identifying locations by clicking the mouse over the map. This allows a total risk value to be determined for specific locations, such as street intersections, or for a general area if an exact address is unknown.

An index map is provided to help determine where the current view lies in relation to the entire county. Municipal boundaries are displayed in a small index map that is highlighted with the outline of the approximate extent of the current map view.

Client/Server Application

All map, search, and analysis operations are performed on a centralized Web server using a custom application, which responds directly to the user's requests. This client/server application offers many benefits with respect to data and software upgrades and application enhancements. In addition, the Web server application returns a consistent representation of the model to all users on most computer platforms. The use of Web technology also provides access to many different kinds of information through the same interface. Background research, documentation, and other supporting information can be included or referenced directly through links to other Web sites on the Internet.

Software

The software used in this application consists of a Web server, a custom executable, and a client browser. The executable was developed using Microsoft Visual Basic and runs on Microsoft NT Workstation. MapObjects (ESRI, Redlands, CA) is a set of mapping software components; Visual Basic was used to incorporate basic mapping and querying capabilities from MapObjects into the application. MapObjects Internet Map Server, another set of software components from ESRI, was used to provide communications with the Web server and for image file format conversion.

Any Web browser supporting HTML 2.0 or greater should be compatible with this application. The use of standard HTML maximizes the number of compatible variations of computers, operating systems, and browsers that may be available to users for accessing this application. Minimal configurations, including low-end PCs and Windows 3.1, are assumed to be compatible.

Summary

A Web-based lead exposure risk analysis application provides many benefits by modeling real-world contributors to lead exposure and making this information readily

accessible. Physicians and public health professionals will no longer have to rely solely on an individual's self-reported assessment of environmental lead exposure risks. They can now use objective information to target screening efforts on those individuals who are at greatest risk of exposure.

As additional blood lead results become available, and risk feature attributes are better defined, widespread use of the model will serve as an effective tool in the early identification of lead-poisoned individuals, while improving the cost-effectiveness of lead screening efforts.

The application is easy to use. Users can retrieve the lead exposure risk for an individual simply by typing in an address. In this fashion, localized areas possessing high exposure characteristics can be readily identified. The model accesses parcel-specific data; thus, pockets of elevated risk can be discerned within large geographical units such as census tracts or zip code areas.

Supplying this application on the Internet provides health professionals with near-universal, on-demand access to detailed data that would otherwise not be available. The application makes available previously obtained blood lead results for others in the vicinity of the target residence, thus providing health professionals with feedback regarding the prevalence of screening activity as well as the number of incidences of elevated blood lead in the provider's service area.

Use of Web server technology allows the data elements of the application to be maintained regularly and frequently without burdening the users with file maintenance. The data can also be analyzed and presented to end-users in a consistent manner irrespective of computer resources. Access can be given to data derived from sensitive patient information without compromising privacy. Research, supporting documentation, and links to related Web sites can also be assembled and easily provided.

Screening and early detection of lead exposure are effective means of preventing cases of severe lead poisoning; however, many children exposed to toxic levels of lead are not being identified. The use of GIS and Internet technology can assist in the identification of children at greatest risk of lead exposure and help ensure that exposed children receive the necessary services.

References

1. Brody DJ, Pirkle JL, Kramer RA, Flegal KM, Matte TD, Gunter EW, Paschal DC. 1994. Blood lead levels in the US population: Phase 1 of the Third National Health and Nutrition Examination Surveys (NHANES III, 1988 to 1991). *Journal of the American Medical Association* 272:277–83.
2. Pirkle JL, Brody DJ, Gunter EW, Kramer RA, Paschal DC, Flegal KM, Matte TD. 1994. The decline in blood lead levels in the United States: The National Health and Nutrition Examination Surveys (NHANES). *Journal of the American Medical Association* 272:284–91.
3. Agency for Toxic Substances and Disease Registry. 1988. *The nature and extent of lead poisoning in children in the United States: A report to Congress.* Atlanta, GA: US Department of Health and Human Services.
4. Centers for Disease Control and Prevention. 1997. Update: Blood lead levels—United States, 1991–1994. *Morbidity and Mortality Weekly Report* 46(7):141–55.

5. Landrigan PJ. 1990. Current issues in the epidemiology and toxicology of occupational exposure to lead. *Environmental Health Perspectives* 89:61–6.
6. Goyer RA. 1993. Lead toxicity: Current concerns. Environmental Health Perspectives 100:177–87.
7. Needleman HL, Riess JA, Tobin MJ, Biesecker GE, Greenhouse JB. 1996. Bone lead levels and delinquent behavior. *Journal of the American Medical Association* 275:363–9.
8. Kim R, Rotnitzky A, Sparrow D, Weiss ST, Wager C, Hu H. 1996. A longitudinal study of low-level lead exposure and impairment of renal function. *Journal of the American Medical Association* 275:1177–81.
9. Goldstein GW. 1990. Lead poisoning and brain cell function. *Environmental Health Perspectives* 89:91–4.
10. Ruff HA, Bijur PE, Markowitz M, Yeou-Cheng M, Rosen JF. 1993. Declining blood lead levels and cognitive changes in moderately lead-poisoned children. *Journal of the American Medical Association* 269:1641–6.
11. Lilienthal H, Winneke G, Ewert T. 1990. Effects of lead on neurophysiological and performance measures: Animal and human data. *Environmental Health Perspectives* 89:21–5.
12. Bellinger D, Sloman J, Leviton A, Rabinowitz M, Needleman HL, Waternaux C. 1991. Low-level exposure and children's cognitive function in the preschool years. *Pediatrics* 87:219–27.
13. Dietrich KN, Berger OG, Succop PA. 1993. Lead exposure and the motor developmental status of urban 6-year-old children in the Cincinnati prospective study. *Pediatrics* 91:301–7.
14. Needleman HL, Gunnow C, Leviton A. 1979. Deficits in psychologic and classroom performance of children with elevated dentine lead levels. *New England Journal of Medicine* 300:689–95.
15. McMichael AJ, Baghurst PA, Wigg NR, Vimpani GV, Robertson EF, Roberts RJ. 1988. Port Pirie cohort study: Environmental exposure to lead and children's abilities at four years. *New England Journal of Medicine* 319:468–75.
16. Needleman HL, Schell A, Bellinger D, Leviton A, Allred E. 1990. The long-term effects of exposure to low doses of lead in childhood: An 11-year follow-up report. *New England Journal of Medicine* 322:1037–43.
17. Fergusson DM, Horwood LJ, Lynskey MT. 1997. Early dentine lead levels and educational outcomes at 18 years. *Journal of Child Psychology and Psychiatry and Allied Disciplines* 38:471–8.
18. Centers for Disease Control and Prevention. 1991. *Preventing lead poisoning in young children: A statement by the Centers for Disease Control.* Atlanta, GA: US Department of Health and Human Services.
19. Schaffer SL, Campbell JR. 1994. The new CDC and AAP lead poisoning prevention recommendations: Consensus versus controversy. *Pediatric Annals* 23:592–9.
20. Centers for Disease Control and Prevention. 1997. Screening young children for lead poisoning: Draft guidance for state and local public health officials. Atlanta, GA: US Department of Health and Human Services.
21. Goldman KD, Demissie K, DiStefano D, Ty A, McNally K, Rhoads GG. 1998. Childhood lead screening knowledge and practice. Results of a New Jersey physician survey. *American Journal of Preventive Medicine* 15(3):228–34.

Refined Soil Texture Emission Factors for Estimating PM_{10}

Samuel Soret, PhD (1),* Randall G Mutters, PhD (2)
(1) Geographic Information, Analysis and Technologies Laboratory, Department of Environmental and Occupational Health, School of Public Health, Loma Linda University, Loma Linda, CA; (2) University of California Cooperative Extension, Oroville, CA

Abstract

Airborne particulate matter of less than 10 micrometers, PM_{10}, is a health concern because it can bypass the body's natural defense mechanisms, settle permanently in the lungs, and impair lung function. US Environmental Protection Agency guidelines recommend that the major sources of excess PM_{10} be identified and quantified. In attending to regulatory demands and public health concerns, state and regional agencies are required to develop strategies for improving existing PM_{10} emissions inventories. In some areas of California, the airborne soil originating from agricultural lands and operations is the largest single source of PM_{10} particles. Specifically, land preparation and wind erosion activities are major sources of soil PM_{10}. Predictive factors describing these processes include a soil texture variable. Current procedures rely on one soil texture value for the entire state. Thus, the estimates are error prone because they do not reflect the observed range of soil characteristics. The objectives of this study were to develop a geographic information system (GIS) methodology for generating location-specific soil texture descriptors, and to evaluate GIS-predicted PM_{10} levels with measured values in agricultural areas. Selected attributes describing California soil were extracted from the US Department of Agriculture's (USDA's) STATSGO, a digital soil database, using a GIS. An area-weighted silt content was computed for each soil type. A weighted average was calculated based on the area occupied by each soil map unit in relation to the surface area. Map unit polygons delineated by silt content were intersected with a digital database of irrigated farmland to provide weighted averages relative to planted acreage. A standard USDA wind erosion equation was applied on a grid-wise basis to estimate monthly PM_{10} concentrations in a major agricultural zone during a non-cropping month. Actual PM_{10} measurements from the state air quality monitoring network were used for comparison.

Keywords: airborne particulate matter, PM_{10}, area-weighted soil texture, emission factor, wind erosion

Background

Airborne particulate matter of less than 10 micrometers (μm), PM_{10}, is a health concern because it can bypass the body's natural defense mechanisms and settle permanently in the lungs, impairing respiratory function. PM_{10}'s effects are more evident in chronic heart disease and chronic respiratory disease patients, asthmatics, elderly people, or children, but it ultimately affects everyone. High ambient concentrations of PM_{10} are a

* Sam Soret, Loma Linda University School of Public Health, Loma Linda, CA 92350 USA; (p) 909-478-8750; (f) 909-824-4087; E-mail: ssoret@sph.llu.edu

major California air quality problem that may significantly affect public health in metropolitan areas, where federal standards for PM_{10} are frequently exceeded.

Problem Definition

US Environmental Protection Agency (EPA) guidelines recommend that major sources of excess PM_{10} be identified and quantified. In attending to regulatory demands and public health concerns, state and regional agencies are required to develop strategies to improve existing PM_{10} emission inventories.

In some areas of California, land preparation and wind erosion are major sources of primary PM_{10} emissions. Over 90% of primary PM_{10} emissions may arise from the stationary process category. Agricultural lands and operations (e.g., tillage) may account for more than 50% of those primary PM_{10} emissions (1). Predictive functions, i.e. emission factors, for estimating PM_{10} from agricultural lands include a soil texture variable. Current emission estimation procedures rely on an input variable that defines soil characteristics using one default soil texture value (18%) for the entire state. Estimates are error prone because they do not reflect the range of soil characteristics observed in California's agricultural production regions.

We suggest that estimates using these assumed average values are inadequate for local air quality management districts charged with formulating a state implementation plan. A partial solution for this inadequacy would involve improving PM_{10} emission factor data to increase the accuracy of existing inventories. Location-specific emission factors can be developed through the use of spatially disaggregated soil textural data in California.

Objective

The objective of this study was to develop a geographic information system (GIS) methodology for generating refined agricultural PM_{10} emission factors based on location-specific soil texture descriptors (i.e., silt content). The accuracy of GIS-predicted PM_{10} levels were then evaluated by comparing them with measured levels in an agricultural zone.

Methodology

EPA's AP-42 method (2) for estimating emissions from agricultural tilling, and a standard US Department of Agriculture wind erosion equation (3) were used for this study. Silt content is the required input variable for the tilling emission factor equation, whereas erodibility is the key variable for the wind erosion equation. Erodibility (susceptibility of the soil particles to detachment and transport by an erosive agent, e.g., wind) and silt content attributes of California soils were extracted from the Natural Resources Conservation Service STATSGO digital database (4) using a GIS. From STATSGO data, an area-weighted silt content was calculated for each soil type based on the area of the individual components contained therein (5). The proposed GIS methodology for analyzing these data is illustrated in Figures 1 through 5. For convenience, Kern County, California, was used to graphically illustrate the procedure.

The application of this approach to the agricultural tilling emission factor is

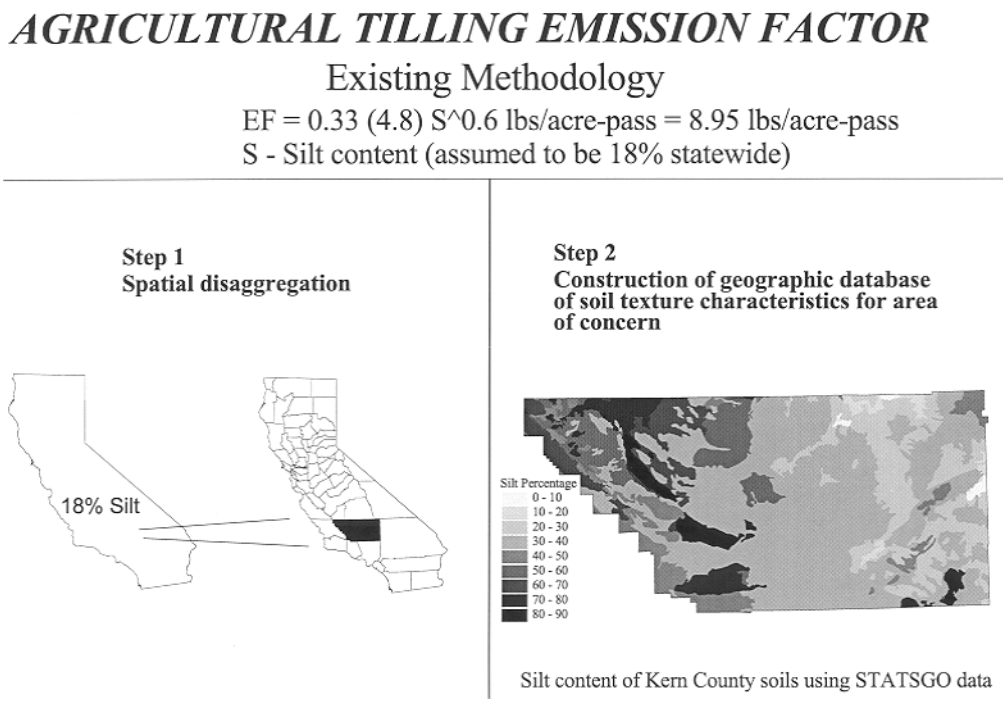


Figure 1 GIS methodology for refining an agricultural tilling PM_{10} emission factor, steps 1 and 2. Spatial disaggregation of soil textural characteristics is graphically illustrated for Kern County, CA.

illustrated in Figures 1 and 2. First, larger areas are disaggregated into smaller geographic zones based on explicit soil characteristics (Figure 1, steps 1 and 2). The level of disaggregation in this case is from state to county. Second, location-specific soil data are developed for use as input in the emission equation. While the vast majority of crops in California are grown under irrigated conditions, irrigated farmland only constitutes a small portion of the total acreage of many counties. Therefore, countywide average soil textures may not accurately describe farmland soil textures. Irrigated farmlands in Kern County, for example, exist almost entirely in the western portion of the county (Figure 2). It follows that agriculturally relevant textural estimates may be improved if restricted to irrigated farmlands. To refine soil texture estimates in the context of farmland location, a silt content coverage was intersected with irrigated farmland at the county level (Figure 2, step 3). The silt parameter is computed as a countywide, area-weighted textural average within irrigated farmlands (Figure 2, step 4). This allows calculation of the refined tilling emission factor (Figure 2, step 5).

Figures 3, 4, and 5 depict the application of similar techniques to the windblown dust emission factor over vegetable crop fields in Kern County. In this instance, spatial disaggregation proceeds from the county level to grid cells of 2 km by 2 km (Figure 3, step 1). Soil attributes for each grid cell are extracted from the STATSGO database, and area-weighted soil erodibility is subsequently computed (Figure 3, steps 2 and 3). Location-specific soil erodibility values are then used to calculate the refined emission factor (Figure 4, step 4). Vegetable crops were assumed to be distributed uniformly

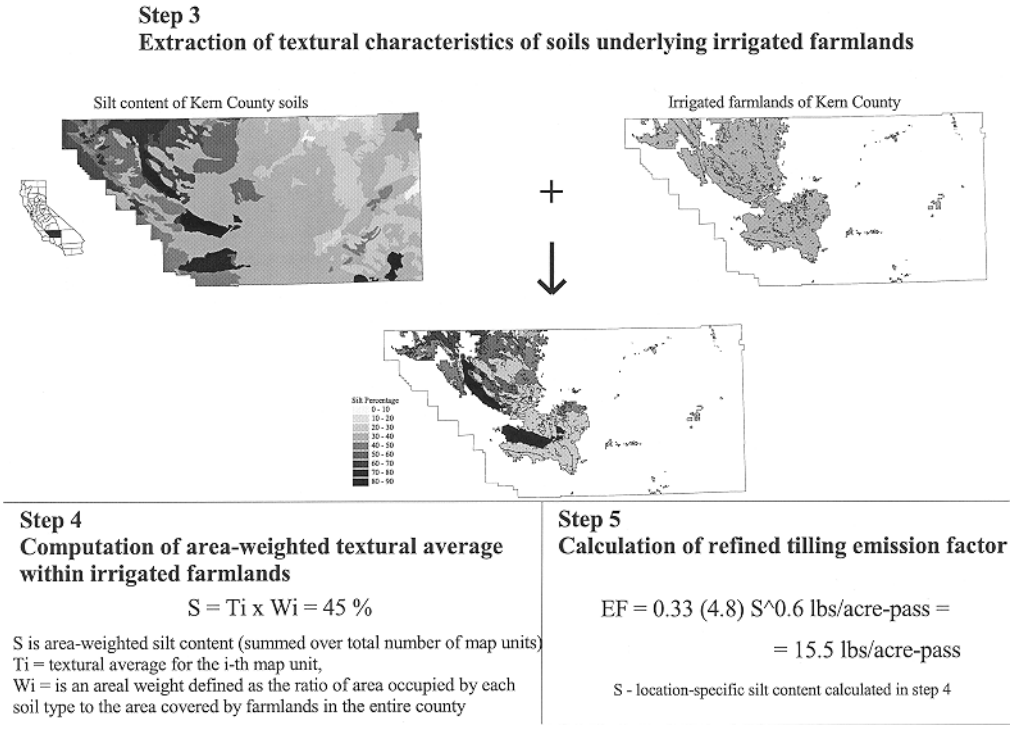


Figure 2 GIS methodology for refining a tilling emission factor, steps 3, 4, and 5. Step 3: Extraction of location-specific textural characteristics. Step 4: Computation of a countywide silt content value within the irrigated farmlands of Kern County. Step 5: Calculation of the refined emission factor.

throughout the irrigated farmlands of Kern County, representing an estimated 12% of acreage for all crops (6). Grid-wise extraction of areal cover corresponding to irrigated farmlands is illustrated in Figure 4, step 5. Emissions are then estimated by multiplying the obtained emission factor by the acres (12% of total farmland acreage) covered by vegetable crop fields within each grid cell.

Results

Area-weighted silt content (46%) for Kern County (Figure 1, step 2) was considerably higher than the value (18%) used in previous estimates of agriculture's contribution to PM_{10}. This translates into nearly a doubling of the emission factor, from 8.95 lb/acre-pass to 15.5 lb/acre-pass. To compare GIS-predicted emission estimates (Figure 5) with actual readings, the estimates were converted into hourly average concentrations of micrograms per cubic meter ($\mu g/m^3$) by assuming a range of inversion layer altitudes (2,000 ft, 1,000 ft, 500 ft), uniform mixing within that volume of air, and by dividing by 720 hr/month. Thus, the calculated volume for comparison equals 2,000 m × 2,000 m × h, where h is inversion layer altitude (meters).

An actual average hourly PM_{10} concentration for the month of October was derived

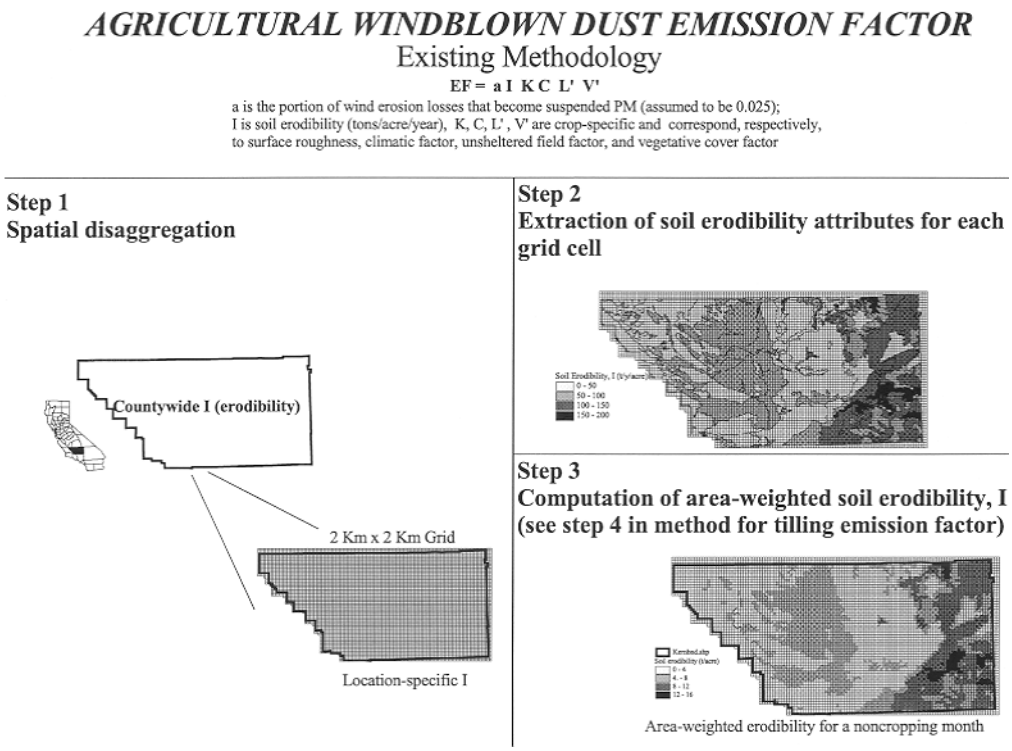


Figure 3 GIS methodology for refining an agricultural windblown dust emission factor, steps 1, 2, and 3. Step 1: Spatial disaggregation. Step 2: Extraction of location-specific soil erodibility attributes. Step 3: Computation of disaggregated soil erodibility values input into wind erosion equation.

from air quality statistics gathered at a California Air Resources Board monitoring station located in an agricultural area near Bakersfield, California. PM_{10} derived from wind erosion of vegetable crop fields after harvest accounted for only a small percentage of the measured particulate mater (Table 1). Wind erosion accounted for a maximum of 3% of the PM_{10}, assuming a 500 ft inversion layer.

Conclusions

The results imply that the use of the 18% default value for silt content in Kern County underestimates agricultural tilling emissions. Therefore, the area-weighted silt content estimated in this study may provide a more realistic description of the magnitude and geographic distribution of PM_{10} emissions from agricultural lands. These results indicate that wind erosion from idle vegetable crop fields accounts for only a small portion of airborne particulate matter present in the southern San Joaquin Valley, California, during a non-production time when there are no crops in the field.

Advantages of using the GIS methodology discussed in this paper include:

- GIS allows calculating and taking into account variations in localized factors,

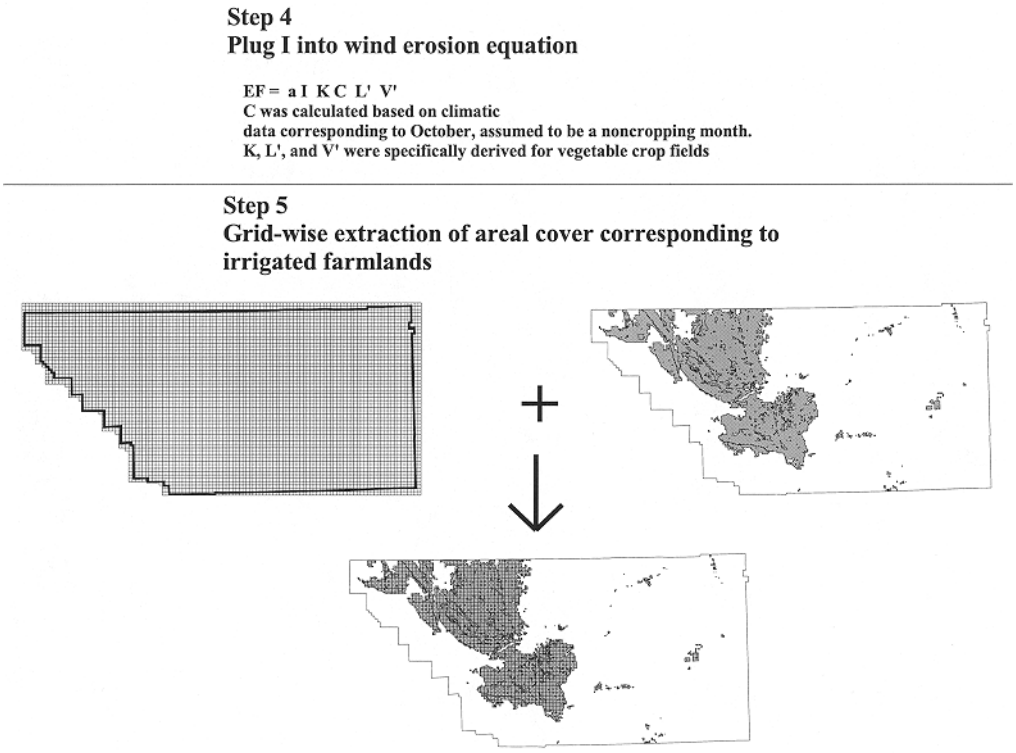


Figure 4 GIS methodology for refining an agricultural windblown dust emission factor, steps 4 and 5. Step 4: Generation of the refined emission factor. Step 5: Grid-wise extraction of acreage corresponding to farmlands.

such as soil texture. This calculation could not be performed without GIS because the computations require the manipulation of multiple layers of geographic data.

- Predictive models can be applied on a grid-wise basis to estimate local PM_{10} concentrations for comparisons with measured levels reported at monitoring stations.
- Estimates of PM_{10} levels in rural non-monitored areas of the state are possible.

Disadvantages of this GIS methodology include:

- The STATSGO digital database contained soil statistics at a resolution adequate at the county level, but it was inappropriate for analysis specific to individual farms.
- This GIS methodology cannot overcome the inherent limitations of standard emission estimation methods.
- There was a lack of digital spatial data describing the geographic location of specific crops.

Because of potential problems with STATSGO data (7), application or extrapolation of weighted textural averages should be adequately qualified to reflect the limitations of

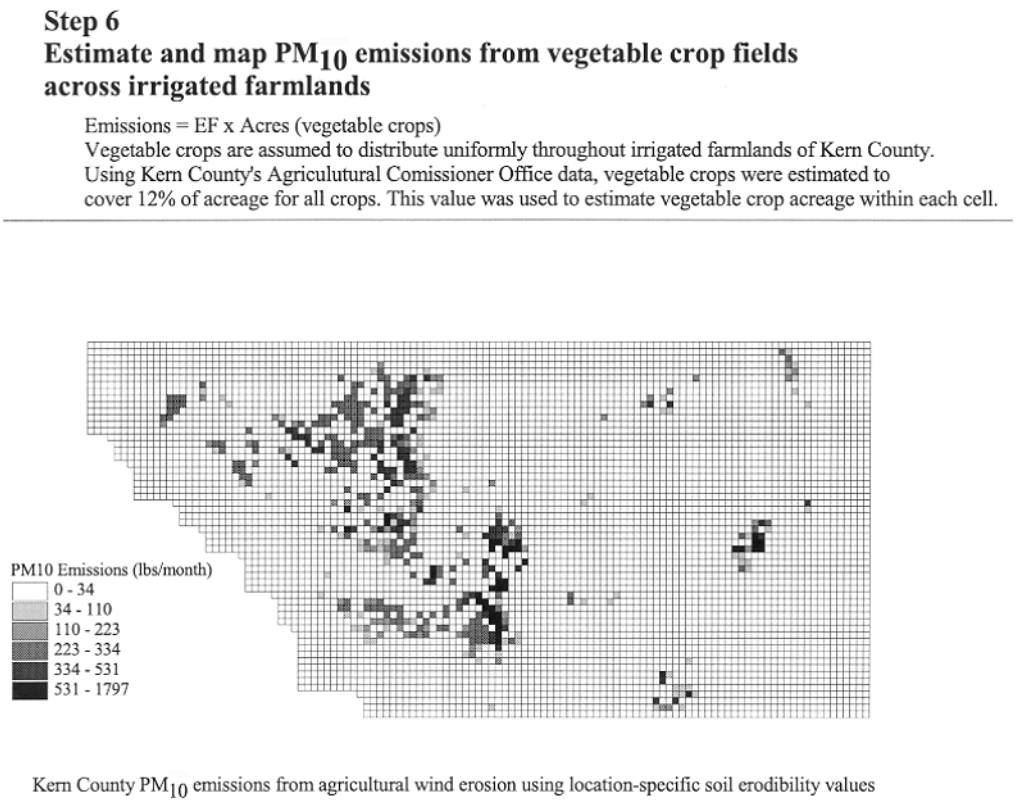


Figure 5 Estimation and mapping of PM_{10} emissions from idle vegetable crop fields across irrigated farmlands.

Table 1 GIS-Estimated Average Hourly PM_{10} Concentrations

Assumed Altitude for Inversion Layer Base (ft)	GIS-Estimated Concentration of Windblown Agricultural PM_{10}[a] (µg/m³)	Percentage of Total PM Measurement (45.6 µg/m³)[b] Accounted for by GIS Estimate (Agricultural PM) (%)
500	1.37	3.0
1,000	0.68	1.49
2,000	0.34	0.74

[a] Estimated average hourly readings for month of October

[b] Average hourly reading for month of October derived from measurements taken at a California Air Resources Board monitoring station near Bakersfield, California. Measurement includes PM from all sources.

PM = Particulate matter

µg/m³ = micrograms per cubic meter

the database before proposing any substantial control measures. This application of GIS to agricultural PM_{10} analysis allowed the generation of location-specific tilling and wind erosion emission factors based on published digital soil data and soil erosion models. The main aim of this study, however, was to illustrate a GIS methodology rather than produce suitable estimates for state implementations plans. Therefore, obtained values should be viewed only as approximations. Nonetheless, we believe that this technique has the potential to become a valuable tool for modeling and estimating air pollution. GIS-enhanced emission factors can lead to future improvements in estimates of particulate matter from agricultural lands in addition to other area source categories, such as road dust or construction operations.

Acknowledgments

This project was supported in part by a grant from the California Air Resources Board.

References

1. Chow JC, Watson JG, Lowenthal DH, Solomon PA, Magliano KL, Ziman SD, Richards LW. 1992. PM_{10} source apportionment in California's San Joaquin Valley. *Atmospheric Environment* 26A:3335–54.
2. Cowheard C, Axetell K, Guenther CM, Jutze GA. 1974. *Development of emission factors for fugitive dust sources.* Research Triangle Park, NC: US Environmental Protection Agency. EPA-450/3-74-037.
3. California Air Resources Board. 1991. *Methods for assessing area source emissions in California.* Sacramento: Emission Inventory Branch, Technical Support Division, California Air Resources Board.
4. Natural Resources Conservation Service. 1993. *State soil geographic (STATSGO) database: Data users guide*. Lincoln, NE: US Department of Agriculture. Misc. Publication 1492.
5. Mutters RG, Soret S. 1995. *Area-weighted soil texture components for estimating agriculture's contribution to airborne particulate matter*. Final report to California Air Resources Board. No. 92-734.
6. Kern County, California crop reports. 1997. Kern County Agricultural Commissioner's Office, Kern County, California. http://www.kernag.com/crp_idx.html.
7. Shimp D, Campbell S. Using a geographic information system to evaluate PM_{10} area source emissions. In: *Proceedings of the Air and Waste Management Associations Emissions Inventory: Programs and Progress Conference.* Research Triangle Park, NC. October 11–13, 1995.

A Public Health Information System for Conducting Community Health Needs Assessment

Elio F Spinello, MPH,* Ronald Fischbach, PhD
Department of Health Sciences, California State University, Northridge, CA

Abstract

Students in the health education program at California State University, Northridge (CSUN), like students in similar programs at other universities, are often trained through the use of practical projects as part of the curriculum. Health education projects typically require that students select a geographically defined community, conduct a community health education needs assessment, and then produce a program plan that addresses the health needs of that community. The purpose of this project was to develop a community health database and mapping system that geographically integrated demographic, housing, morbidity, and mortality data, using specially written software to facilitate analysis and mapping of the data. The creation of this public health information system (PHIS) represents a step toward using geographic information system technology to improve the process of developing public health needs assessments. The PHIS is also a valuable resource for training public health educators, making hard-to-find needs assessment data readily available to them as well as giving them the ability to analyze the data spatially. After pilot testing in fall 1997, the PHIS was implemented for use in community health education classes at California State University. The CSUN Department of Health Sciences intends to find additional applications for the PHIS database. Currently, the PHIS is being tested by a national medical center as part of a program to determine its effectiveness as a decision-support tool. The system also has been made available to the Los Angeles County (California) Department of Health Services for testing in the Border Health program as a needs assessment tool.

Keywords: assessment, community health, health education, morbidity, mortality

Background

A critical step in the development of a health education program is the needs assessment. At the community level, conducting a needs assessment requires collecting a substantial amount of information about the demographic, socioeconomic, and health characteristics of the population. For most communities, this process is a time-consuming, largely manual task. A number of methods are typically used to collect community-level data. These often include focus groups; in-depth interviews; and surveys of residents, key informants, health care providers and community leaders. Indirect methods are often used to gather statistical data about a community. The sources for statistical data can include the US Census Bureau, state and local departments of public health, law enforcement agencies, and health regulatory agencies.

* Elio F Spinello, Dept. of Health Sciences, California State University, Northridge, 18111 Nordhoff St., Northridge, CA 91330-0001 USA; (p) 818-831-7607; (f) 818-831-9078; E-mail: elio.spinello@csun.edu

One of the limitations of statistical data gathered from multiple sources is that the data must be standardized and integrated geographically. This means, for example, that demographic data by census tract would be matched up with housing data by census tract to produce a combined database containing both housing and demographic data. The linkage for combining the information would be the unit of geography—in this case, census tracts. A problem arises when one agency produces data at the census tract level while another produces data at an overall, county, community, zip code, or other level of geography. Various agencies are also often inconsistent with respect to coding variables such as age and ethnicity (1,2). Data collection protocols often differ between agencies; this can cause different agencies to collect data for different time periods, or can mean that some agencies do not collect some needed portion of the data, such as information from a particular geographic region. Many of the difficulties involved in collecting statistical data for communities were noted by Paulu, Ozonoff, Coogan, and Wartenberg (3), when they observed that clinical medicine and public health have, in some ways, become more similar over time. While physicians have become more prevention-oriented, public health professionals are more frequently being called upon to engage in what might be called community diagnosis and treatment. While physicians have patient histories available to them when they conduct individual medical needs assessments, though, there is often no readily available community health history for use by the public health official. Sources of information about a community do not reside in the memory or records of an individual person, but rather in institutional arrangements made by a number of agencies for many different purposes. Given the lack of an integrated public health information resource for Los Angeles County (California), a need was identified for the development and creation of a computerized community-level database that contains demographic, morbidity, and mortality data.

Purpose of the Project

The purpose of this project was to design and implement a public health information system (PHIS) for Los Angeles County, organized by zip code and containing health, housing, and demographic variables. This PHIS was to be used by health education students in the preparation of community health needs assessments. Zip codes were selected as the basic unit of geography because all of the required datasets were believed to be available at the zip code level. Software was designed and developed that would facilitate access to the database by enabling users to rank zip codes, query the data, and produce reports and maps for selected zip codes and census block groups.

Database and System Design

Major foundation courses at California State University, Northridge (CSUN), that prepare professional health educators emphasize the concepts of program planning, implementation, and evaluation as they apply to health education in the community. The primary teaching strategy in these courses involves assigning students a project in which they must select a community, assess the health needs of that community, and then design a health education program to address those needs. Traditionally, students select a community and identify its boundaries, then begin researching the health, demographic, and socioeconomic data available for that area. This process involves

gathering statistical data from local government and private sources, and identifying and interviewing key informants. It also involves other primary and secondary research. Once the research is completed, the student must design a program plan with measurable health education objectives, an action plan for a health intervention, an implementation scheme for conducting the program, and a selection of appropriate evaluative measures. The premise behind the PHIS project was that all of these activities, which are quite time-consuming, could be accomplished much more quickly with the help of a computerized database system.

The concept for the database came from the need for students to be able to assess the health needs of a community. Based on the requirements for undergraduate and graduate students' health education projects, the needs assessment database had to include the following information (4):

- Community backdrop: including demographics, geographic features, transportation systems, and political structure.
- Health care system: hospitals, clinics, health department sites, emergency services, voluntary health agencies, etc.
- Community health status: including morbidity and mortality measures of populations in the community.
- Social assistance system: programs available in the community.

The geographical element common to all of the data sources was a 5-digit zip code. Because the zip code was provided by all of the data sources and was a concept assumed to be easily understood by users, it was adopted as the definition of a community. The demographic data sources were actually found to provide census data for much smaller areas, such as census tracts and block groups. Based on its availability, demographic information in the database was included for block groups, because they represent much smaller geographic areas than zip codes and therefore enable users to conduct deeper analyses. Once the data sources were assembled, a relational database approach was adopted—a design that would allow multiple tables that have a common variable (zip code) to be integrated as if they were a single combined table.

Specific tables used in the database included:

- Hospital discharges for Los Angeles County by major diagnostic categories (MDCs) and *International Classification of Diseases, Ninth Revision* (ICD-9) (5) codes, by zip code.
- Los Angeles County death certificates summarized by MDCs and ICD-9 codes by zip code.
- Hospital point locations for Los Angeles County.
- Locations of cases of chlamydia and gonorrhea treated at Los Angeles County clinics.
- Current-year and projected demographic variables for Los Angeles County zip codes.
- 1990 demographic and housing data for Los Angeles County block groups.

Upon receipt of the various tables mentioned above, the documentation for all of the files was examined to determine the best way to standardize the tables and link them together. Because the coding conventions for the different files varied, coding standardization issues were also identified at this stage. In the death certificate table, for

example, ethnicity was assigned to each case using one of 16 codes, whereas in the hospital discharge file there were only 8 possible codes. The final database design consisted of two levels of summary. The first level applied to the discharge dataset (6). Using an algorithm developed for this project, records were coded as to whether they represented individual or repeat discharges for the same diagnosis. Once coded, the dataset was reduced to one record per individual rather than one record per discharge (4,7,8,9). The second level involved summarizing each discharge and death certificate record down to one record per discharge or death per 5-digit zip code. This reduced the size of the database substantially while still maintaining a reasonable degree of detail for reporting and analysis (see Figures 1 and 2). After the discharge data and death certificate data had been reduced to the zip code level, the processed discharge data, hospital information, death certificates, and demographic tables were assembled into a relational database.

Figure 1 Example of raw data.

Case Number	Zip Code	ICD-9 Diagnosis	Disposition
1	91324	101.1	Normal Discharge
2	91324	101.1	Normal Discharge
3	91324	101.1	Deceased
4	91324	302.0	Normal Discharge
5	91324	302.0	Deceased

An example of raw data. One record is present for each discharge, indicating the zip code, diagnosis, and disposition.

Figure 2 Example of aggregated records.

Zip Code	ICD-9 Diagnosis	Normal Discharges	Deaths
91324	101.1	2	1
91324	302.0	1	1

An example of aggregated data. Raw data are reduced to one record per diagnosis per zip code.

The programming language used to create the system was Microsoft Visual Foxpro 3.0. Visual Foxpro was chosen because it can query large databases quickly and produce reports based on those data. The PHIS was also designed to produce geographic maps of an area depicting the morbidity and mortality data graphically. The mapping system that was integrated into the PHIS software was created with MapObjects LT (ESRI, Redlands, CA).

System Features

The data access software was designed to enable students to select zip codes and then produce tabular reports and maps to describe the communities. With the inclusion of a table of population estimates by zip code, morbidity and mortality rates could be calculated by age cohort and ethnicity category. To facilitate its use, the software was

designed with three major analytical components: a zip code ranking module, zip code query and reporting modules, and an integrated GIS module (10).

The zip code ranking module allowed students to select a particular health problem as defined by a single MDC or ICD-9 code. Additionally, a custom definition could be created by combining ICD-9 codes (e.g., all ICD-9 codes relating to sexually transmitted diseases). Once a pre-defined or custom disease was selected, one of eight ranking measures could then be selected (e.g., the crude death rate, prevalence rate, or total number of cases). The ranking report produced by this module could be used to identify the best-to-worst or worst-to-best zip codes in Los Angeles County based on the selected disease and measure. The results could then be displayed on a theme map.

The query module of the PHIS software was designed to allow users to select one or more zip codes from a list. Users could then press a button that would select different types of available data, and choose tabular reports to be produced for each of the selected zip codes. The zip code reports available included:

- The demographic trends report, which indicated basic demographic characteristics such as average household size and per capita income from 1994 to a 1999 projection.
- The demographic profile report, which contained a breakdown of the population by age and ethnicity as well as households by income category.
- The MDC profile, which summarized the number of cases and other key measures and rates for each of the 27 MDC categories.
- The ICD-9 profile, which summarized the number of cases and other key measures and rates for each of the 909 3-digit ICD-9 categories.
- The hospital profile, which detailed each of the hospitals contained in the selected zip code(s) together with key information such as name, address, number of beds, and relative size of emergency room.
- The morbidity and mortality report, which provided a detailed analysis of the morbidity and mortality of a selected disease in the selected zip code(s), including morbidity and mortality rates by age category and ethnicity.

In addition to producing reports for specific zip codes, students could also produce reports summarizing data for all of Los Angeles County. The county-level reports could then be used as a benchmark when comparing rates for various diseases. Additionally, students could select all of the block groups in a census tract, zip code, or custom-defined area and produce a report that would summarize all of the demographic data for the combined area.

The integrated GIS module was designed to give users the ability to produce a color-coded thematic map of their chosen community for purposes of determining the locations of zip codes, census tracts, block groups, hospitals, and highways. The mapping system also provided the ability to display concentrations of gonorrhea, chlamydia, and tuberculosis cases treated at Los Angeles County clinics. Users could select which types of geographic feature to display on a map and change the map view by zooming in or out and by panning across areas. They could also select individual geographic features such as zip codes, hospitals, or highways. Once they selected a geographic feature, users could view information about it by clicking on the feature with a mouse. The geographic features contained in the mapping system included:

- California county regions
- California 5-digit zip code regions
- California interstate highways
- Los Angeles County census tract regions
- Los Angeles County block group regions
- Los Angeles County chlamydia cases shown as point locations
- Los Angeles County gonorrhea cases shown as point locations
- California hospital point locations

The program and database were documented in a user manual that accompanied the software. The manual provided procedures and a step-by-step process for use of the program. The user manual also contained a number of illustrations of the software's interface; this helped users follow along with the manual's examples. A case study was included as a tutorial exercise. Once written, the manual was converted into a Microsoft Windows help file.

System Integrity and Usability Testing

To identify programming bugs and data errors, both the software and database were subjected to an initial test. This process uncovered a number of problems, including installation problems, errors in calculations, and errors in reports.

Upon completion of the initial test, the software and database were pilot-tested by two undergraduate community health education classes, as well as a graduate community health education class. Students from the three classes attended a 30-minute review session of the database and the software's features conducted by the project investigator. Pilot test feedback was collected using three methods. Users were directly observed by the project investigator, who filled out an observational checklist as they used the software. All users were also provided with test reports that they were asked to complete and return. Finally, focus groups were held with all users.

In general, feedback from all three collection methods was relatively consistent. Overall, users found the system to be a useful tool in completing their projects. The project investigator was already aware of a number of the issues that surfaced, such as a missing manual and bugs affecting two of the reports. A number of other issues that were identified had not been previously considered. These included the need for additional training, the need to train lab assistants, the need to include more specific directions in the manual and online help system, and the existence of bugs affecting the mapping system.

What We Learned

Based on a review of the pilot test results, a number of conclusions were reached with respect to the use of the PHIS at CSUN:

- The need to fully explain the constraints and limitations of the database cannot be dismissed. Although the database methodology has some clear limitations, the ability to quickly and easily develop health status profiles made it tempting to take the results at face value without further research. In reality, the database has significant biases for some diseases. Records of drug and alcohol abuse, for

example, only reflect cases in which individuals died (with drugs or alcohol identified as the cause of death) or cases in which individuals either overdosed or admitted themselves for treatment. The vast majority of untreated cases are not recognized in the database. Many acute and chronic diseases that require hospitalization or tend to have high case fatality rates, however, are better represented by the database, because a larger percentage of those cases will be represented in the discharge or death certificate file.

- Due to the number of tasks that can be accomplished with the PHIS, a number of students felt they needed some initial direction to help them best take advantage of it. The approach most commonly suggested was the use of a tutorial or case study, something that would give students an example to work through.
- The next update to the PHIS database will need to incorporate a more accurate method of estimating repeat discharges by the same individual (7,11,12). Although the method used in the initial database was believed to be at least somewhat effective at reducing the dataset from one record per discharge to one record per individual, a more sophisticated statistical approach will likely produce much more accurate results.
- A more effective method of disseminating the data and analytical system will likely improve the effectiveness of the application. Approaches including Internet access to the database and functionality are being explored (3).

References

1. Andres N, Davis DE. 1995. Linking public health data using geographic information system techniques: Alaskan community characteristics and infant mortality. *Statistics in Medicine* 14(5–7):481–90.
2. Avery C, Zabel D. 1995. *Gathering client data: What works?* Proceedings from the International Conference on TQM and Academic Libraries, Washington, DC, April 20–22, 1994.
3. Paulu C, Ozonoff DM, Coogan P, Wartenberg D. 1995. Making environmental data accessible for public health aims: The Massachusetts Environmental Database Project. *Public Health Reports* 110(6):776–83.
4. Murray SA, Graham, LJC. 1996. Practice based health needs assessment: Use of four methods in a small neighborhood. *British Medical Journal* 310(6992):1443–8.
5. World Health Organization (WHO). 1977–78. *Manual of the international statistical classification of diseases, injuries, and causes of death: Based on the recommendations of the Ninth Revision Conference, 1975, and adopted by the Twenty-Ninth World Health Assembly.* Geneva: WHO.
6. California Office of Statewide Health Planning and Development. 1996. *California healthcare data programs.* http://www.oshpd.cahwnet.gov.
7. Jaro MA. 1995. Probabilistic linkage of large public health data files. *Statistics in Medicine* 14(5–7):491–8.
8. Kirby RS. 1996. Toward congruence between theory and practice in small area analysis and local public health data. *Statistics in Medicine* 15(17–18):1859–66.
9. O'Hara D, Hart W, Robinson M, McDonald I. 1996. Mortality soon after discharge from a major teaching hospital: Linking mortality and morbidity. *Journal of Quality Clinical Practice* 16(1):39–48.

10. Roos LL, Walld R, Bond R, Hartford K. 1996. Record linkage strategies, outpatient procedures, and administrative data. *Medical Care* 34(6):570–82.

11. Huff L, Bogdan G, Burke K, Hayes E, Perry W, Graham L, Lentzner H. 1996. Using hospital discharge data for disease surveillance. *Public Health Reports.* January.

12. Scholten HJ, de Lepper MJ. 1991. The benefits of the application of geographical information systems in public and environmental health. *World Health Statistics Quarterly* 44(3):160–70.

Geographic Information Systems and Ciguatera Fish Poisoning in the Tropical Western Atlantic Region

John F Stinn, BA (1),* Donald P de Sylva, PhD (2), Lora E Fleming, MD, PhD, MPH (3), Eileen Hack, BS (4)

(1) Public Health Practice Program, Centers for Disease Control and Prevention, Atlanta, GA; (2) Rosenstiel School of Marine and Atmospheric Sciences, University of Miami, Miami, FL; (3) NIEHS Marine and Freshwater Biomedical Sciences Center, University of Miami, Miami, FL; (4) Department of Epidemiology and Public Health, University of Miami, Miami, FL

Abstract

Little is known about the epidemiology of ciguatera fish poisoning, the most commonly reported marine toxin disease. In endemic areas and beyond, ciguatera is a seafood-borne illness that affects persons of all ages and socioeconomic groups. Integrating an existing ciguatera database into a geographic information system (GIS) will give researchers new insight into the epidemiology of ciguatera and allow linkage between disparate epidemiological and oceanographic datasets. A voluntary Ciguatera Hotline has collected data from 1977–1998 in the endemic ciguatera area of South Florida. Descriptional statistics and spatial trends of ciguatera cases and the fish sources were examined using ArcView GIS software. A total of 777 cases, 442 on record, with 304 index cases were analyzed from the database. Cases were distributed geographically throughout Miami-Dade County, Florida. A high concordance was shown between the location of ciguatoxic fish and specific coral reef areas in the Caribbean. Using GIS in the future may help prevent disease by pinpointing ciguatera hotspots and facilitating the exploration of possible etiologic relationships between oceanographic and anthropogenic changes in the sources of ciguatera.

Keywords: ciguatera, marine toxin diseases, tropical medicine, Caribbean region

Introduction

Ciguatera is the most frequently reported seafood-related illness in the world, affecting up to 500,000 people per year worldwide (1). The illness is caused by the consumption of coral reef fishes contaminated with a group of natural toxins produced by minute phytoplankton known as dinoflagellates. These toxins are bio-concentrated through the food chain such that humans consuming large reef fish (such as barracuda, grouper, and snapper) are the ultimate predators and receive the highest doses of toxins. The most important of the ciguatera toxins, ciguatoxin, causes a blockage of sodium channels throughout the nervous system; ultimately, this neurological blockade manifests as a multitude of symptoms, affecting numerous bodily functions (2). Ciguatoxin and the other marine toxins are heat/acid-stable; therefore normal food preparation does not detect or eliminate them. Furthermore, these toxins are some of the most highly toxic natural substances; ciguatoxin is toxic to humans in picogram doses (3). The

* John Stinn, Centers for Disease Control and Prevention, Public Health Practice Program Office, 4770 Buford Hwy, MS-39, Atlanta, GA 30341 USA; (p) 770-488-2449; (f) 770-488-2489; E-mail: zjj8@cdc.gov

symptoms of ciguatera may persist in humans from a few days to up to several months or even years, depending on the size of the fish, the size of the contaminated portion, and the seafood consumption history of the victim (4).

Ciguatera is the predominant fish poisoning in the endemic tropical regions of the Pacific and the Caribbean (5). The social and economic impacts of ciguatera in endemic regions are the avoidance of the consumption and sale of seafood. For example, Tahiti, the most populated island of French Polynesia (135,000 inhabitants), loses an estimated US $1 million annually due to banned reef fish sales (6). With increasing international travel and trade as well as increasing fish consumption, ciguatera is being imported to traditionally non-endemic areas (7). The medical costs and lost wages for ciguatera victims can be quite high, especially in non-endemic areas where diagnosis is often delayed due to non-recognition by victims and their healthcare providers. For example, in the non-endemic region of Canada, these costs have been estimated between US $1,850 and US $8,950 per case (8).

The Centers for Disease Control and Prevention (CDC) estimate that fewer than 2–10% of ciguatera cases are actually reported in the United States (9). Many biases contribute to the underreporting of ciguatera. First, there are no easily available inexpensive tests for the ciguatoxic fish or human victims. Second, although ciguatera is a reportable disease, many healthcare providers do not recognize, diagnose, or report ciguatera, especially in non-endemic areas. Third, there is a lack of knowledge in the recreational fishing community about ciguatera and possible prevention measures. Finally, there is a desire in the restaurant and commercial fishing industries to suppress publicity of ciguatera as a threat to seafood consumers (10,11,12). Given the plethora of underreporting issues, large comprehensive ciguatera databases such as the one used in this study are rare (1,9,11,13). Therefore, the epidemiology of ciguatera is still in its infancy.

Expanding ciguatera epidemiological investigations with technology such as geographic information systems (GIS) may help researchers gain new insight. GIS can be used to collect, check, integrate, and analyze information related to the earth's surface, allowing for the integration of non-traditional datasets (14). The digital nature of GIS software allows for the data and analyses to be easily updated, transferred, manipulated, and displayed (15). In the case of ciguatera, GIS could be used to evaluate possible associations between epidemiologic and oceanographic data. By using GIS mapping capabilities, the distribution and trends of ciguatera cases in time and space can be evaluated to increase the knowledge of ciguatera epidemiology. Furthermore, the source of the ciguatera could be traced from the human cases to the contaminated fish and then back to the ciguatoxic coral reef. This in turn could lead to possible primary prevention activities such as ciguatoxic reef postings, thus discouraging fishing and further seafood consumption in known contaminated areas.

The following study is an analysis of a 20-year database of self-reported ciguatera cases in South Florida, using GIS to evaluate both epidemiologic and oceanographic data.

Methods

Since 1977, researchers at the University of Miami's Rosenstiel School of Marine and Atmospheric Science (RSMAS) have attempted to inform the South Florida public, the

seafood industry, and the medical profession about ciguatera (16). Press releases, radio and television interviews, and magazine articles have been used for outreach and education about tropical fish poisoning. A voluntary Ciguatera Telephone Hotline, networked with the local medical and public health communities, was established. A standardized questionnaire was implemented. For over 20 years, investigators at RSMAS have received letters and telephone calls from victims, healthcare providers, and concerned seafood customers, primarily from South Florida and the Caribbean. The resulting database, "Ciguafile" (17), represents one of the largest and oldest collections of ciguatera cases in the world, despite the multitude of underreporting biases.

The Ciguafile database consists of ciguatera cases from 1977 to 1998. Ciguatera victims and healthcare providers voluntarily reported these cases. Case demographics were recorded, including the age, residence, gender, symptomatology, progression of the illness, species and weight of the fish involved in each outbreak, location of the capture, date of the capture, and other pertinent information. Race-ethnic and socioeconomic class data were not collected in this database. Data concerning how many additional people consumed the same fish, if those people became sick, and where the fish was procured were also collected. These data were stored in a Microsoft Excel spreadsheet. The subjects' addresses in Miami-Dade County (Miami, FL) were geocoded and referenced with a South Florida street map using ArcView GIS (ESRI, Redlands, CA).

For those ciguatoxic fish captures with exact data on capture location available, the latitude and longitude were converted to decimal-degree units and displayed over a map of the Caribbean region using ArcView GIS. In addition, ciguatoxic fish captures in the Caribbean documented with sufficient detail in the historical literature were also added to the database. Documented coral reefs of the region were also displayed using historical data on coral reef location (18,19).

Nearest-neighbor analyses were performed on both the residential data and capture location data. The nearest-neighbor statistic (R) is based on the comparison of observed spatial distribution with what one would expect if the distribution were completely random. The statistic has a range of 0 to 2.15. An R-value of zero indicates a completely clustered pattern, a value of 1 indicates a random distribution, and R=2 or greater corresponds with a completely uniform (even) distribution (20).

To evaluate the hypothesis that ciguatera is derived from the consumption of fish associated with coral reefs, the following analyses were performed. First the capture location of each fish was referenced to the distance from the nearest coral reef. Then, based on data for the home ranges of individual fish species, a maximum distance of 1 mile of range from each known coral reef was selected (21). Nearest-neighbor analyses were performed, and the relative spatial density of ciguatoxic fish captures was determined.

Results

There were 442 cases in the Ciguafile database, with a total of 777 reported cases (these included additional cases reportedly sharing the same fish). The mean age of the database cases was 44.7 ±15 years (range 4 months to 87 years); most (53.2%) cases were female. Victims reported a multitude of symptoms, most commonly paresthesia and acute gastrointestinal disorders. A comparison of the symptoms reported in Ciguafile

and other previously published ciguatera case registries can be seen in Table 1 (1,9,22–27).

Many cases were involved in cluster outbreaks, where the sharing of one fish was responsible for multiple cases. Based on the available information, 304 ciguatoxic clusters were recorded in the database. Ciguatera was more likely to be present in disease clusters with an average of 2.36 persons/cluster. Within the 304 cluster outbreaks, there was an attack rate of 87.5% per cluster of those who reportedly consumed a toxic fish and who experienced ciguatera symptoms.

To evaluate the geographic distribution by residence at the time of illness, cases from 1978 to 1981 within Miami-Dade County, a ciguatera endemic region, were analyzed (Figure 1). Of the 304 index cases, 169 occurred in Miami-Dade County, with 102 (60.4% of Miami-Dade County cases) of these cases occurring during the specified time period. A nearest-neighbor analysis was performed in an attempt to show a random distribution of cases in the county. However, despite various attempts to adjust for population density and lack of habitability (e.g., airports, Everglades, and ocean areas), the R-value was 0.10, indicating a strong clustering pattern. Nevertheless, the clustering pattern closely followed densely populated roadways that pass through highly varied race-ethnic neighborhoods in Miami-Dade County.

The causative fish were acquired through individual fishing (31.6% of the cases), buying from fishermen, stores, or restaurants (64.1%), or as gifts (4.3%). Overall, the most frequently implicated culprits in the outbreaks were groupers (47.1%), snappers (30.7%), barracudas (9.6%), kingfish (6.1%), jacks (5.7%), and dolphin fish (4.6%). Because the identification of restaurant-acquired fish type can be faulty (28), only the fish types and weights of the fish reported by individual fishermen were examined. The most commonly reported fishes leading to outbreaks reported from fishermen were barracudas (average size 10.6 lb; range 4–22 lb); kingfish (37.2; 6.5–100); black groupers (51.3; 26–73); amberjacks (26.4; 7–47); red snappers (9.8; 3.5–17.5); and hog snappers (4.6; 2–10).

There were 50 ciguatoxic fish captures with location data; 111 ciguatoxic captures in the Caribbean were also documented in the historical literature in sufficient detail to allow for GIS mapping (4,29). Of the 50 Ciguafile captures, 43 (86%) occurred in the region between the Florida Keys and the Bahamian chain. The historic captures were spread throughout the Caribbean. When examining the 161 confirmed fish captures as a whole, some areas of the Caribbean reported ciguatera outbreaks much more frequently than others, especially Puerto Rico and the neighboring US and British Virgin Islands (Figure 2). Furthermore, the fish captures were strongly clustered, as confirmed by the nearest-neighbor tests with $R<0.02$, using the Caribbean Sea and Gulf of Mexico as the reference region. This clustering pattern closely followed the line of coral reefs adjacent to the small island nations of the Windward and Leeward Islands.

Conclusions

This study is an analysis of a 20-year database of self-reported ciguatera cases in South Florida and uses GIS to evaluate both the epidemiologic and oceanographic data. It illustrates that ciguatera is a disease that occurs in clusters and affects persons of all ages. The data also reflect that ciguatera affects individuals in different ways

Table 1 Reported Frequency of Clinical Symptoms of Ciguatera

	Region of Study Source of Data (Reference #) Number of Cases (n)									
Symptoms (reported by % frequency)	**Caribbean (17) (n=442)**	**Caribbean (24) (n=57)**	**Caribbean (22) (n=47)**	**Caribbean (9) (n=129)**	**Caribbean (25) (n=16)**	**Caribbean (23) (n=80)**	**Caribbean (27) (n=6)**	**South Pacific Islands (26) (n=12,890)**	**Western Pacific (Australia) (1) (n=167)**	**South Pacific Isles (11) (n=3,009)**
Gastrointestinal										
Diarrhea	78.7	77	81	76	56	83	66	72.6	49	70.6
Vomiting	42.5	37	40	68	69	69	66	38.8	50	37.5
Nausea		82				69	100	43.5	50	42.9
Abdominal pain	64.5	58	30		75	74	66	42.5	29	46.3
Neurological										
Arthralgia	78.7	75	34		31	60		85.9	29	85.7
Myalgia	79.0	75	34	86	94	56		85.3	38	81.5
Extremity paresthesia	81.0	79		71	38	36	50	89.0	82	89.2
Circumoral paresthesia	69.5	79	38	54	38	38	33	88.1	82	89.1
Temperature reversal	64.3	77	23		50	48	16	87.2	65	87.6
Headache		56	45	47	50	39		59.6	25	59.2
Dizziness/vertigo	50.0			47	56	33	16			42.3
Weakness		84		30	94	65.4		60.0	70	60.0
Chills/sweating			36	24				59.6		59.0
Other										
Dysuria	25.0				31			12.6		18.7
Pruritus	77.0		66	48	100	45	66	44.0	5	44.9
Dental pain or “looseness”	32.1	23	13		19	11		20.7		24.8
Dyspnea								12.1		16.1
Skin rash	32.1				31					20.5

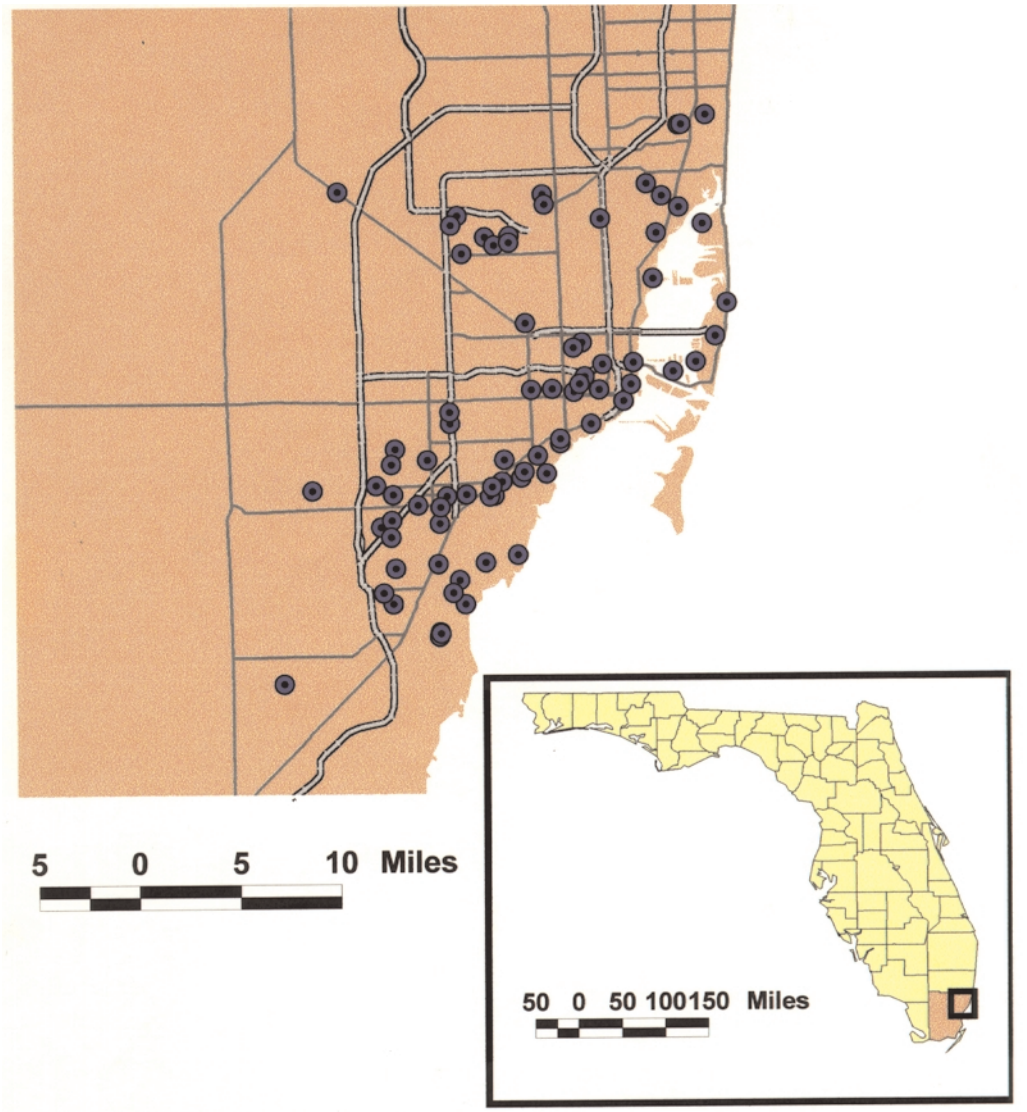


Figure 1 Ciguatera cases, Miami-Dade County, FL, 1978–1981 (17).

symptomatically. Very few people reported all the possible symptoms, making it even more difficult for accurate diagnosis on the health care provider's part.

The Ciguafile is a passive collection database. It relies on the referral and reporting by physicians and ciguatera cases with actual knowledge of the Ciguatera Hotline telephone number. This can lead to obvious reporting bias. As such, no reliable incidence rates can be generated for South Florida from these data. A previous study in Miami-Dade County indicated that the incidence rate of ciguatera was at least 5 cases/10,000 people/year (9). In the neighboring Caribbean and other tropical regions, the rates are even higher, with estimates of over 100 cases/10,000 people/year on some tropical islands (30).

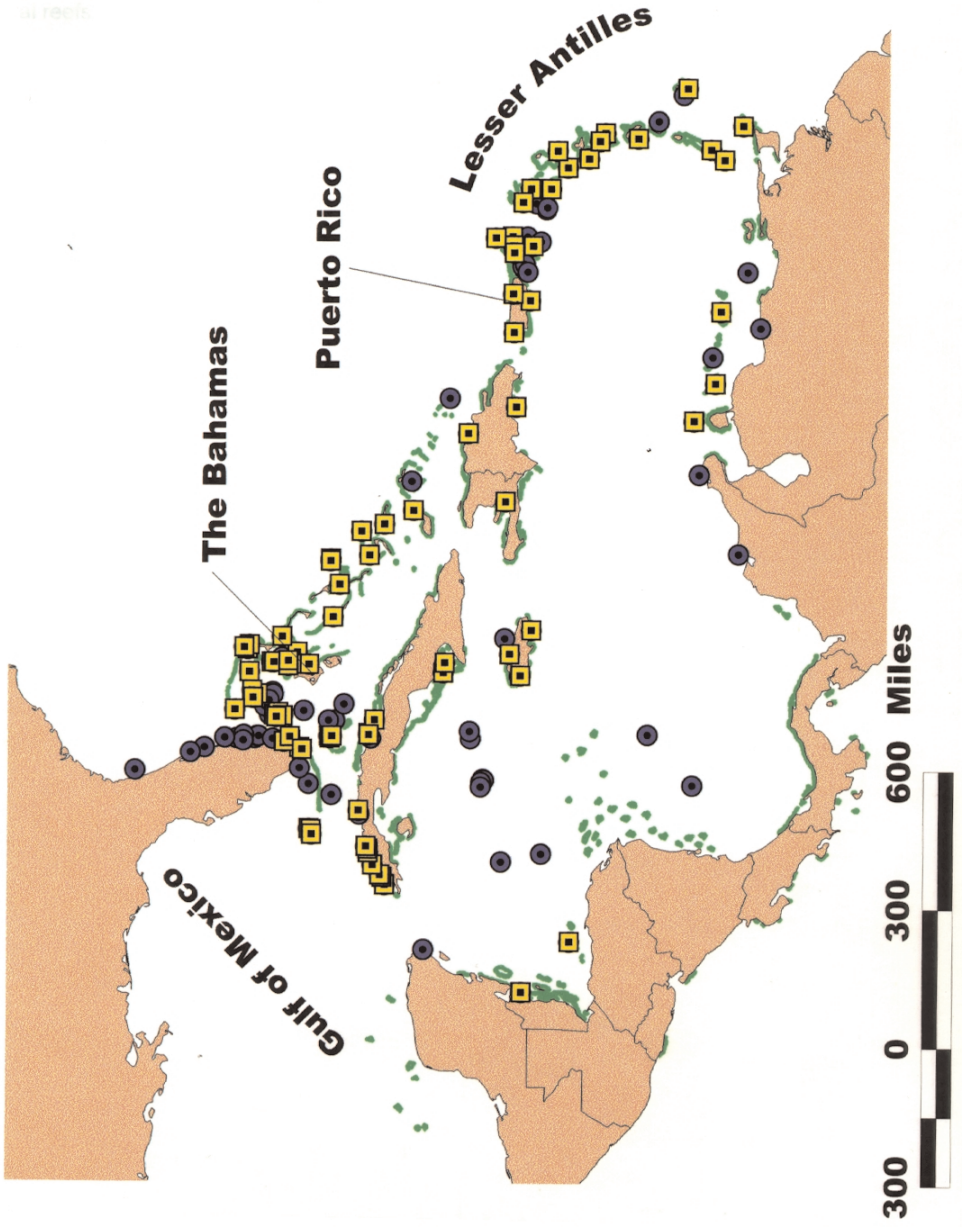


Figure 2 Ciguatoxic fish captures, greater Caribbean, 1900–1998 (4,17,29).

The clustering pattern of residences noted in the geographic mapping of cases in Miami-Dade County occurred along a major roadway throughout the county that crosses race-ethnic and socioeconomic lines in its course. Although precise data were not collected, this suggests that ciguatera affects all persons regardless of race-ethnic group and socioeconomic class. In addition, the non-random spatial clustering in Miami-Dade County may reflect the proximity to and locations of seafood restaurants or fish markets because the most implicated fish were acquired in markets and restaurants.

Another issue is that the capture location data reported from Ciguafile were neither accurate nor precise. This is an issue for the use of GIS (15). Many of the subjects reported incidents weeks or even months after consumption. Aside from two Ciguafile cluster reports with precise global positioning system (GPS) coordinates, fish capture locations were reported in vague terms such as "two miles west of Great Isaac's Light" or "just off the northeast point of Grand Bahama Island." Therefore, the data from this study are good for identifying general ciguatera hotspots, not citing specific individual reefs that may or not be safe. In the future, more accurate capture data could allow for the identification and posting of individual coral reefs. The measure could lead not only to primary prevention of ciguatera (important due to the lack of quick and inexpensive testing) (31), but possibly to ecological relief for over-fished coral reefs (4).

Analysis of the fish capture locations showed an association with specific coral reefs. Changes in the reef environment, however, may inhibit accurate analysis. Overfishing may temporarily eliminate the possibility of ciguatera. It cannot be determined if the environment is not conducive to the disease or if it is simply because the reefs have been overfished. Some biologists feel the fish that are captured in certain overfished or overexploited waters are usually too young and small to be contaminated with potent amounts of ciguatoxin (32). Should the reef communities rebound, the disease may manifest itself again.

Spatial density analysis revealed hotspots near Puerto Rico and the Bahamas, indicating the potential to identify ciguatoxic reefs in these areas (Figure 3). There are, however, severe problems with this type of analysis. Spatial density does not account for disparity in captures over time. Given the multitude of underreporting concerns against ciguatera, gaining these data may be difficult. Spatial density analysis ignores the theoretical possibility of migrating fish from a toxic reef to a safe reef. Nevertheless, the trend of ciguatoxic captures occurring along the reefs, and in certain areas more than others, cannot be ignored and must be investigated.

State and county health departments are increasingly adding GIS to their disease reporting and surveillance systems, and are collaborating with environmental departments when analyzing exposures to dangerous substances. GIS allows for linkages and analysis of different databases, such as oceanographic and epidemiologic, to explore complicated environmental diseases such as ciguatera. Also, the real-time editing capabilities and transferability characteristics of GIS databases may allow for better education and awareness of ciguatera, particularly among health officials in non-endemic regions. This may help overcome many of the historic educational and diagnostic biases against ciguatera. Furthermore, with the advent and affordability of GPS technology, most commercial and recreational fishermen can more accurately record fish capture locations. True ciguatoxic hotspots may soon be found and primary prevention initiated.

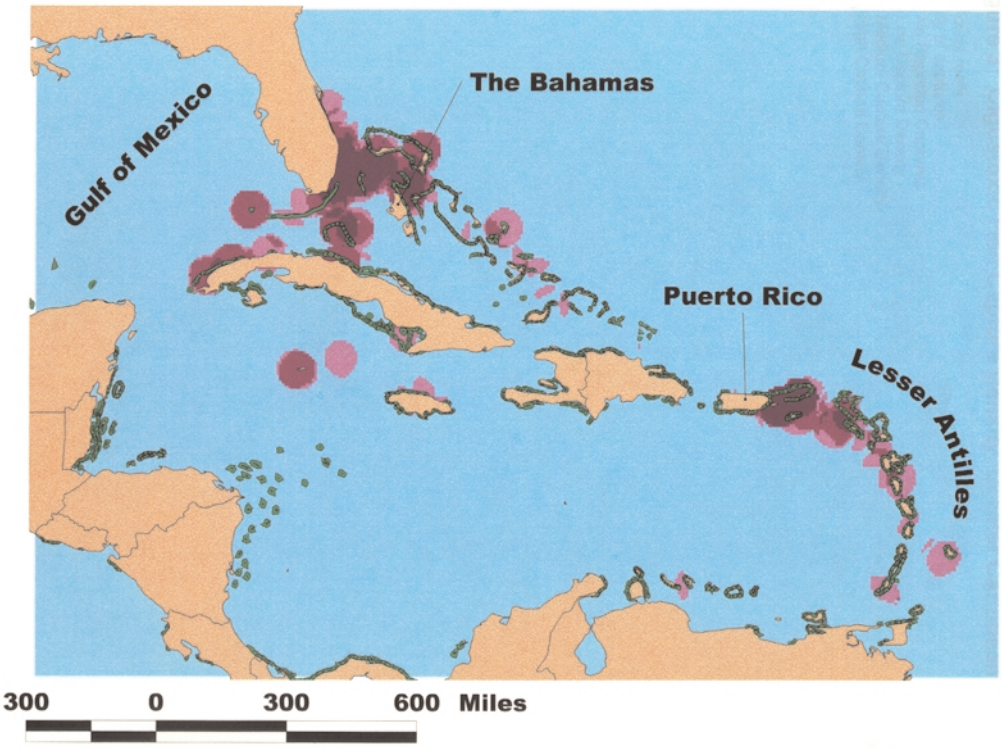


Figure 3 Ciguatoxic fish capture density, greater Caribbean, 1900–1998 (4,17,29).

In the future, GIS could be applied to the micro-marine environment, allowing scientists a new means of studying the ecology involved in ciguatera. The fishes associated with ciguatera have been known for decades (33). In 1980, researchers identified *Gambierdiscus toxicus* as the dinoflagellate most responsible for producing ciguatoxins (34). Much of the life history of *G. toxicus* has been described (35). Despite these facts, researchers continue to be baffled as to why different fishes from the same area may or may not be ciguatoxic and why certain species are poisonous on one reef but not another (36). Furthermore, it is possible that global change, coral bleaching, and anthropogenic effects on coral reef ecology may lead to further changes in the incidence of ciguatoxic reefs. The modeling and statistical capabilities of GIS may allow the biotic and abiotic attributes of contaminated reefs to be investigated in ways that were too expensive or difficult to conceptualize in prior research.

Acknowledgments

This work was supported in part by the Jefferson Lee Ford III Foundation and the National Institute of Environmental Health Sciences (NIEHS) Shannon Award (NIEHS Grant #1 RO1 ES08122). An internal review board did not review this project because no personal identifiers were used. Mr. Carlos Rivero, Director, Geocore GIS Research Facility, RSMAS, Ms. Maria Villanueva, Senior Research Associate, Division of Marine Affairs, RSMAS, and Dr. John Gifford, Associate Professor, Division of Marine Affairs,

RSMAS, aided tremendously in the GIS instruction and display of the data. Dr. Bin Li assisted in the initial GIS mapping. Dr. Phillip Kramer, Division of Marine Geology and Geophysics, provided the coral reef coverages for analysis. Ms. Jana Easom and Ms. Giavanni Washington reviewed the manuscript.

References

1. Quod JP, Turquet J. 1996. Ciguatera in Reunion Island (SW Indian Ocean): Epidemiology and clinical patterns. *Toxicon* 34(7):779–85.
2. Levin DZ. 1995. Ciguatera: Current concepts. *Journal of the American Osteopathic Association* 95(3):193–8.
3. Baden DG, Fleming LE, Bean JB. 1995. Marine toxins. *Handbook of clinical neurology* 21(65):141–75.
4. de Sylva DP. 1994. Distribution and ecology of ciguatera fish poisoning in Florida with emphasis on the Florida Keys. *Bulletin of Marine Science* 54(3):944–54.
5. Lewis RJ. 1992. Socioeconomic impacts and management of ciguatera in the Pacific. *Bulletin de la Societe de Pathologie Exotique* 85(5 Pt 2):427–34.
6. Bagnis R. 1992. Public health, epidemiological and socioeconomic patterns of Ciguatera in Tahiti. In: *Proceedings of the Third International Conference on Ciguatera Fish Poisoning.* April 30–May 5, 1990, La Parguera, Puerto Rico. Quebec: Polyscience Publications. 157–68.
7. Glaziou P, Legrand AM. 1994. The epidemiology of ciguatera fish poisoning. *Toxicon* 32(8):863–73.
8. Todd, E. 1985. Ciguatera in Canada. In: *Proceedings of the Third International Conference on Dinoflagellates.* Elsevier Science Publishing. 505–10.
9. Lawrence DN, Enriquez MB, Lumish RM, Maceo A. 1980. Ciguatera fish poisoning in Miami. *Journal of the American Medical Association* 244(3):254–8.
10. Fleming LE, Baden DG, Bean JA, Weisman R, Blythe DG. 1998. Seafood toxin diseases: Issues in epidemiology and community outreach. In: *Proceedings of the VIII International Conference on Harmful Algae.* June 25–29, 1997. Vigo, Spain. Santiago de Compostela (Spain): Xunta de Galicia and Intergovernmental Oceanographic Commission of UNESCO. 245–8.
11. Bagnis R, Kuberski T, Laugier S. 1979. Clinical observations on 3,009 cases of ciguatera (fish poisoning) in the South Pacific. *American Journal of Tropical Medical Hygiene* 28(6): 1067–73.
12. Fleming LE, Bean JA, Baden DG. 1995. Epidemiology of toxic marine phytoplankton. In: *UNESCO-IOC Manual on harmful marine phytoplankton #33.* Ed. GM Hallegraeff, DAN Anderson, AD Cembella. Paris: UNESCO.
13. Swift EB, Swift TR. 1993. Ciguatera. *Clinical Toxicology* 31(1):1–29.
14. Rhind DW. 1998. A GIS research agenda. *International Journal of Geographical Information Systems* (2):23–8.
15. Star J, Estes JE. 1990. *Geographic information systems.* Englewood Cliffs, NJ: Prentice Hall.
16. Poli M. 1982. *A review of ciguatera, with special reference to the Caribbean, and an investigation into the significance and incidence in Florida.* Master's thesis. Miami, FL: University of Miami.
17. Ciguafile database. 1977–1998. University of Miami Rosenstiel School of Marine and Atmospheric Science, Miami, FL. Database of ciguatera cases.
18. Joubin L. 1912. Bancs et recifs de coreaux (Mandre-pores). *Annals de L'institut Oceanographie* 4(2).

19. Kramer P, University of Miami, Miami, FL. 1998. Oral communication with author. December.
20. West N. 1996. *Applied statistics for marine affairs professionals.* Westport, CT: Praeger Publishers.
21. Bohnsack J, National Marine Fisheries Service. 1998. Oral communication with author. December.
22. Engleberg NC, Morris JG Jr, Lewis J, McMillan JP, Pollard RA, Blake PA. 1983. Ciguatera fish poisoning: A major common-source outbreak in the US Virgin Islands. *Annals of Internal Medicine* 98(3):336–7.
23. Escalona de Motta G, Felix JF, Izquierdo A. 1986. Identification and epidemiological analysis of ciguatera cases in Puerto Rico. *Marine Fisheries Review* 48(4):14–18.
24. Frenette C, MacLean JD, Gyorkos TW. 1988. A large common-source outbreak of ciguatera fish poisoning. *Journal of Infectious Diseases* 158(5):1128–31.
25. Geller RJ, Olson KR, Senecal PE. 1991. Ciguatera fish poisoning in San Francisco, California caused by imported barracuda. *Western Journal of Medicine* 155(6):639–42.
26. Swift EB, Swift TR. 1993. Ciguatera. *Clinical Toxicology* 31(1):1–29.
27. Poli MA, Lewis RJ, Dickey RW, Musser SM, Buckner CA, Carpenter LG. 1997. Identification of Caribbean ciguatoxins as the cause of an outbreak of fish poisoning among US soldiers in Haiti. *Toxicon* 35(5):733–41.
28. de Sylva DP, University of Miami, RSMAS. 1998. Oral communication with author. May.
29. de Sylva DP. 1970. *Systematics and life history of the Great Barracuda, Sphyraena barracuda.* 2nd ed. Coral Gables, FL: University of Miami Press.
30. Lange WR. 1987. Ciguatera toxicity. *American Family Physician* 35(4):177–82.
31. de Sylva DP, Higman JB. 1979. A plan to reduce ciguatera in the tropical western Atlantic region. In: *Proceedings of the Gulf and Caribbean Fisheries Institute No. 32.* November. Miami Beach, FL. Miami Beach: Gulf and Caribbean Fisheries Institute.
32. Szmant A, University of Miami, RSMAS. 1998. Oral communication with author. February.
33. Halstead BW. 1978. *Poisonous and venomous marine animals of the world.* 2nd ed. Princeton, NJ: Darwin Press.
34. Bagnis R. 1980. Origins of ciguatera fish poisoning: A new dinoflagellate Gambierdiscus toxicus. *Toxicon* 18(2):199–208.
35. Bomber J. 1987. *Ecology, genetic variability, and physiology of the ciguatera-causing dinoflagellate, Gambierdiscus toxicus.* Dissertation. Melbourne, FL: Florida Institute of Technology.
36. Kaly UL, Jones GP. 1994. Test of the effects of disturbances of ciguatoxin in Tuvalu. *Memoirs of the Queensland Museum* 34(3):523–32.

Web-Based Access and Visualization of Hazardous Air Pollutants

Jürgen Symanzik (1),* David Wong (2), Jingfang Wang (2), Daniel B Carr (2), Tracey J Woodruff (3), Daniel A Axelrad (3)
(1) Department of Mathematics and Statistics, Utah State University, Logan, UT (2) George Mason University, Fairfax, VA; (3) Office of Policy, US Environmental Protection Agency, Washington, DC

Abstract

This paper reports on a project to provide Web-based access to the US Environmental Protection Agency's (EPA's) extensive model-based summaries of hazardous air pollutants (HAPs). As part of EPA's Cumulative Exposure Project, long-term cumulative concentrations of 148 HAPs for the 60,803 census tracts in the 48 contiguous states have been modeled for 1990. The model results include estimates and confidence bounds that assess the estimated uncertainties for each of the HAPs in each census tract. The project challenge was to concisely display 148×60,803 (8,998,844) estimates along with uncertainty bounds. The project goal was to make these statistical summaries accessible to the public as statistical tables and graphs. The Web provides an easy way to make this information electronically accessible. One big challenge is to make the summaries conceptually accessible. The most difficult part of this is to communicate an understanding of the underlying data limitations, the modeling process, and how to interpret the model results. The easier part of ensuring conceptual accessibility is facilitating navigation through the summaries and consideration of values within a large context. Our approach allows the user to select a HAP and "drill down" through the levels of a geopolitical hierarchy. The hierarchy consists of states within the United States, counties within states, and census tracts within counties. Our Web-based approach also attended to the design of tables and graphics with the intent to make them more readable and useful. For tables, our approach focused on perceptual details such as rounding and foreground-background contrast. For graphics, our approach provided spatial context through the use of recently developed templates called linked micromap plots. Both tables and micromaps provide a hierarchically clickable drill-down to finer details. This provides fast answers to questions about the air quality in any given region in the contiguous United States.

Keywords: Cumulative Exposure Project, HAPs, linked micromap plots, micromaps, Graphics Production Library

Introduction

Over the last few years, researchers have developed many improvements that make statistical graphics more accessible to the general public. These improvements include making statistical summaries more visual and providing more information at one time. Research in this area involved exploring the conversion of statistical tables into plots (1), new ways to display geographically referenced data (2), and, in particular, the

* Jürgen Symanzik, Utah State University, Dept. of Mathematics and Statistics, 3900 Old Main Hill, Logan, UT 84322-3900 USA; (p) 435-797-0696; (f) 435-797-1822; E-mail: symanzik@sunfs.math.usu.edu

development of linked micromap (LM) plots, often simply called micromaps (3–5). Another recent development is the Java-based Graphics Production Library (GPL) (Bureau of Labor Statistics, Washington, DC) for the Web-based distribution of interactive statistical graphics (6). The GPL can be used to distribute federal statistical summaries such as the description of hazardous air pollutants (HAPs).

Staff at the US Environmental Protection Agency (EPA) and contractors modeled 1990 long-term cumulative concentrations for 148 HAPs at the census tract level. Because there are 60,803 census tracts in the contiguous United States, this resulted in 148×60,803 (8,998,844) estimates, along with upper and lower confidence bounds for each estimate. The main goal of this project was to provide Web-based access to these HAP data. We focused on providing a concise display that offered easy access to the data. Another goal was to make the model results easily understandable to an audience not familiar with statistics. To achieve our goals, we developed interactive tables and extended the GPL by adding micromaps. The micromaps serve as a geographic navigational "drill-down" tool as well as being meaningful statistical overviews in their own right. The ability to drill down from the national overview showing states, to a state overview showing counties, and then to a county view showing census tracts, provides rapid access to the fine-grained detail in this substantial dataset.

In the next section of this paper we describe EPA's Cumulative Exposure Project (CEP). In the section entitled "Graphical Statistical Components," we describe components, i.e., GPL and micromaps, that were used to construct the CEP Web site. The section following that provides deeper insights into the CEP Web site and the user's point of view. We finish with a discussion of work that has been done to date and that which is still to come. Additional details and a different set of micromap displays and screenshots from the CEP Web site can be found in Symanzik et al. (7).

EPA's Cumulative Exposure Project

Much of the characterization of air pollution has focused on air pollutants designated as "criteria pollutants" in the Clean Air Act, such as particulate matter (PM), ozone, and lead (8). This is largely due to the obvious health effects demonstrated by major pollution episodes, such as those in Donora, Pennsylvania, and London, England (9,10), and the extensive availability of monitoring data to use in assessing health effects. Relatively little is known about the potential health effects of other air pollutants, a number of which are designated as HAPs in the Clean Air Act. HAPs have been associated (mostly through occupational and animal studies) with a variety of adverse health outcomes, including cancer effects and noncancer neurological, reproductive, and developmental effects (11).

Past analyses have relied on limited emissions and monitoring data and some modeling to assess public health impacts of air toxics. Some studies have attempted to assess differential impacts of air toxics on communities of color using emissions estimates, mostly from the EPA's Toxics Release Inventory (TRI), which contains emissions estimates from major manufacturers in the United States (12,13). Other analyses have attempted to characterize the potential public health impacts of air toxics (13–17). One set of studies evaluates potential noncancer health risk by using monitoring data and concentrations estimated by dispersion modeling of emissions from a subset of commercial and industrial facilities (14–16). These studies found that outdoor

concentrations were often greater than benchmarks representing thresholds for potential public health impacts. However, these studies were not comprehensive in their scope, due either to lack of monitoring data (as described in reference 18) or to lack of emissions data.

A recent analysis by Woodruff et al. (19), as part of the EPA's CEP, has assessed the potential public health implications of air toxics across the United States for 1990. The analysis by Woodruff et al. uses modeled outdoor concentrations of air toxics across the contiguous United States (20) to help compensate for the lack of monitoring data on outdoor concentrations. Emissions data from stationary and mobile sources are used as inputs into a dispersion model that estimates 1990 average outdoor concentrations of 148 air toxics for every census tract in the contiguous United States. The estimated outdoor concentrations from the analysis are used as a reasonable proxy for potential exposure when making relative comparisons of hazard and performing screening-level analysis. The analysis by Woodruff et al. found that many estimated concentrations are above previously defined benchmark concentrations representing thresholds of concern for potential adverse public health impacts (19,21).

Estimating 1990 Outdoor Concentrations of Hazardous Air Pollutants

Outdoor concentrations of HAPs were estimated using a Gaussian dispersion model (20,22). This model—the Assessment System for Population Exposure Nationwide (ASPEN)—is a modified version of EPA's Human Exposure Model (22), a standard tool designed to model long-term concentrations over large spatial scales. Long-term average concentrations of HAPs were calculated at the census tract level[1] based on emissions rates of the HAPs and frequencies of various meteorological conditions, including wind speed, wind direction, and atmospheric stability. In addition, the model used in this analysis incorporates simplified treatment of atmospheric processes such as decay, secondary formation, and deposition.

The choice of pollutants for modeling was based on the list of 189 HAPs in section 112 of the 1990 Clean Air Act Amendments. A baseline year of 1990 was selected for modeling. Available emissions data were reviewed and appropriate data were identified for 148 HAPs.

A national inventory of HAP emissions was developed for this study as a required input to the dispersion model. For large manufacturing sources, emissions data contained in EPA's TRI were used (23). Emissions estimates were developed for other sources, such as large combustion sources, automobiles, and dry cleaners, using EPA's extensive national inventories of 1990 emissions of total volatile organic compounds (VOCs) and PM (24,25). HAP emissions were derived from VOC and PM emissions estimates by applying industry-specific and process-specific estimates of the presence of particular HAPs in particular VOC or PM emissions streams (20). Alaska and Hawaii are not included in this study because the national VOC and PM emissions inventories do not include data for these states.

The dispersion model accounted for long-term concentrations of HAPs attributable to current (i.e., 1990) anthropogenic emissions within 50 kilometers of each census tract centroid. For 28 HAPs, estimated outdoor concentrations also included a "background"

[1] The 60,803 census tracts in the contiguous United States vary in physical size but typically have approximately 4,000 residents.

component attributable to long-range transport, re-suspension of historical emissions, and natural sources derived from measurements taken at "clean air" locations remote from the impact of local anthropogenic sources (20). Compared to available air toxics monitoring data for 1990, 1990 modeled concentrations are typically of the correct magnitude, with a general tendency to underestimate the measured ambient concentrations.

EPA's CEP Web site (http://www.epa.gov/CumulativeExposure) was designed specifically to provide further insight into statistical methods and methodologies and answer questions related to air toxics. In addition to providing explanatory texts, documents, and external links that relate to the material described in this section, one of the main goals for the CEP Web site was to provide information about the estimated air toxics data. An essential part of assessing the data is the option to evaluate them visually. The remainder of this paper describes the work done to present the data and the development of the CEP Web site.

Graphical Statistical Components

This section introduces the main graphical statistical components that are used for the Web-based access and visualization of HAPs through EPA's CEP Web site.

The Graphics Production Library

The GPL is a Java class library of graphics routines that make it possible (and convenient) to create and modify statistical graphics on the Web (see http://www.monumental.com/dan_rope/gpl/) (6). The GPL was initially developed within the Bureau of Labor Statistics (BLS) to facilitate the Web-based distribution of the Bureau's statistical summaries. It has interactive features such as dragging and dropping data columns onto each other to allow easier comparisons of the data, reordering and rescaling of panels, and panning and zooming. Thus, it considerably extends the static features but otherwise closely follows the row-labeled plots of Carr (1). Recent recommendations on statistical graphics, as given in Cleveland (26,27) have also been followed during the design of the GPL. Moreover, the GPL makes it possible to add metadata—i.e., add links to articles associated with the data or include warning flags within a data display. The GPL currently supports three types of graphics displays: bar plots, dot plots, and time series graphics. Other types of graphics have been planned but have not been implemented so far by BLS. Unfortunately, the GPL cannot be used to draw maps and link statistical data to them; however, this is one of the features that have been planned.

Micromaps

Linked micromap (LM) plots, often simply called micromaps, provide a new statistical paradigm for the viewing of spatially referenced statistical summaries in their spatial context. Full details on LM plots can be found in Carr et al. (3–5). Using LM plots for the 50 states within the United States provides an alternative to displaying all statistical information on a single choropleth map. Instead, several small maps (ten maps in our 50-state example) are drawn. The associated statistical data are arranged according to a particular criterion (e.g., from highest to lowest or in alphabetical order by corresponding geographical region). Next, the five highest values of the data are drawn in a statistical plot (e.g., dot plot, bar plot, box plot, plot with confidence bounds, time series plot) on the right side of the first small map. For each data point, a different color

is used in the statistical plot. The corresponding regions (in this case five states) are highlighted in the same colors on the first small map. All other states remain uncolored in this map. The same is done for the next five highest remaining data points in the second small map and associated statistical display. This process continues until all data points/regions have been plotted/highlighted.

This splitting into several maps makes obvious the locations of the high, middle, or low observations. It becomes possible to judge if there are any geographic clusters or if the underlying measurements are randomly spread over the area under consideration. LM plots can display multiple statistical variables at a time. Examples of micromaps and S-PLUS (MathSoft, Seattle, WA) code written to create them can be found at ftp://galaxy.gmu.edu/pub/dcarr/newsletter/micromap/.

Figure 1 shows a sample LM plot created using S-PLUS. This county-level micromap of Pennsylvania generally follows the design principle described above. However, we had to find an (almost) symmetric display for 67 counties. We ended up with 16 maps, 13 of them with four counties and 3 of them with five counties. Moreover, the layout has been split into four quarters. In addition to simply highlighting the individual four or five counties from the associated statistical display in each map, all 16 or 17 counties that fall into the corresponding quarter are highlighted on this map. This makes it easier to understand the spatial structure. For example, the viewer can immediately grasp that the highest benzene concentrations have been modeled for counties surrounding the major cities Philadelphia (e.g., Philadelphia County, Delaware, Montgomery, and Bucks), Harrisburg (e.g., Dauphin and Cumberland), and Pittsburgh (e.g., Allegheny), while the lowest benzene concentrations have been modeled for counties in the sparsely populated Pennsylvania-New York border region (northern border of map).

The CEP Web Site

The User's Point of View

In addition to providing access to explanatory texts and entire CEP-related documents, the main purpose of the CEP Web site is to provide fast and easy access to data on the 148 HAPs at different spatial resolutions ranging from the US level (top) to the census tract level (bottom).

Three mechanisms have been designed for selecting main features of the CEP Web site and for maneuvering from the US down to the census tract level and up again. Standard menus in the upper-right part of the Web page allow the user to select the representation (data tables, micromaps, or raw data), one of the 148 HAPs, a state and, based on this selection, a county for which the data should be displayed (see Figure 2). This navigation and selection menu remains permanently visible, independently from the current statistical display.

The second tool that has been implemented to drill down through the levels of a geopolitical hierarchy makes use of interactive tables that display statistical data and serve as navigational tools at the same time. Figure 2 shows such a table at the county level for Pennsylvania. The user can mouse-click on any of the listed counties and a new table will appear, displaying data (including a 90% confidence interval) for all census tracts within the selected county. Small arrows pointing upward and downward allow

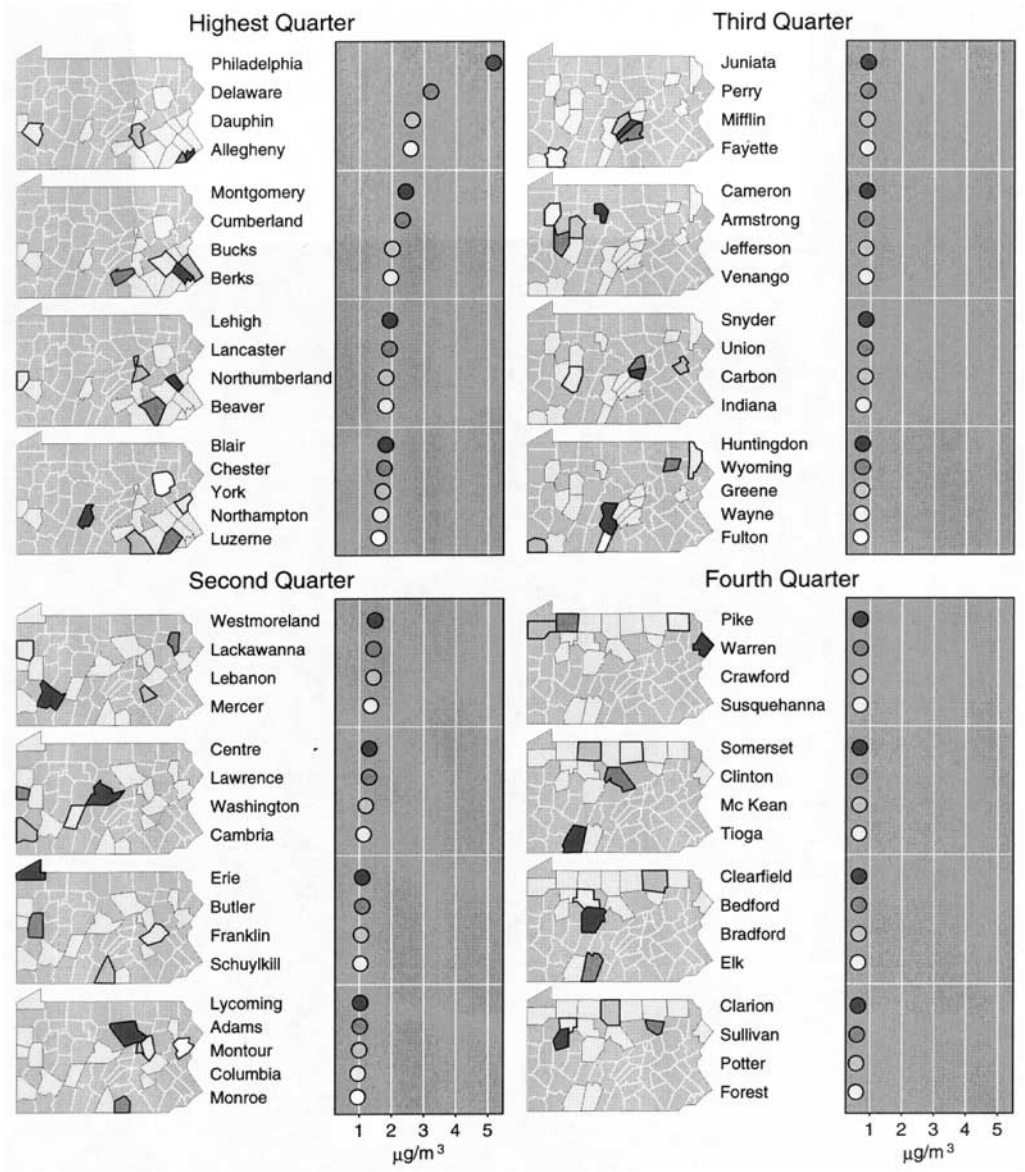
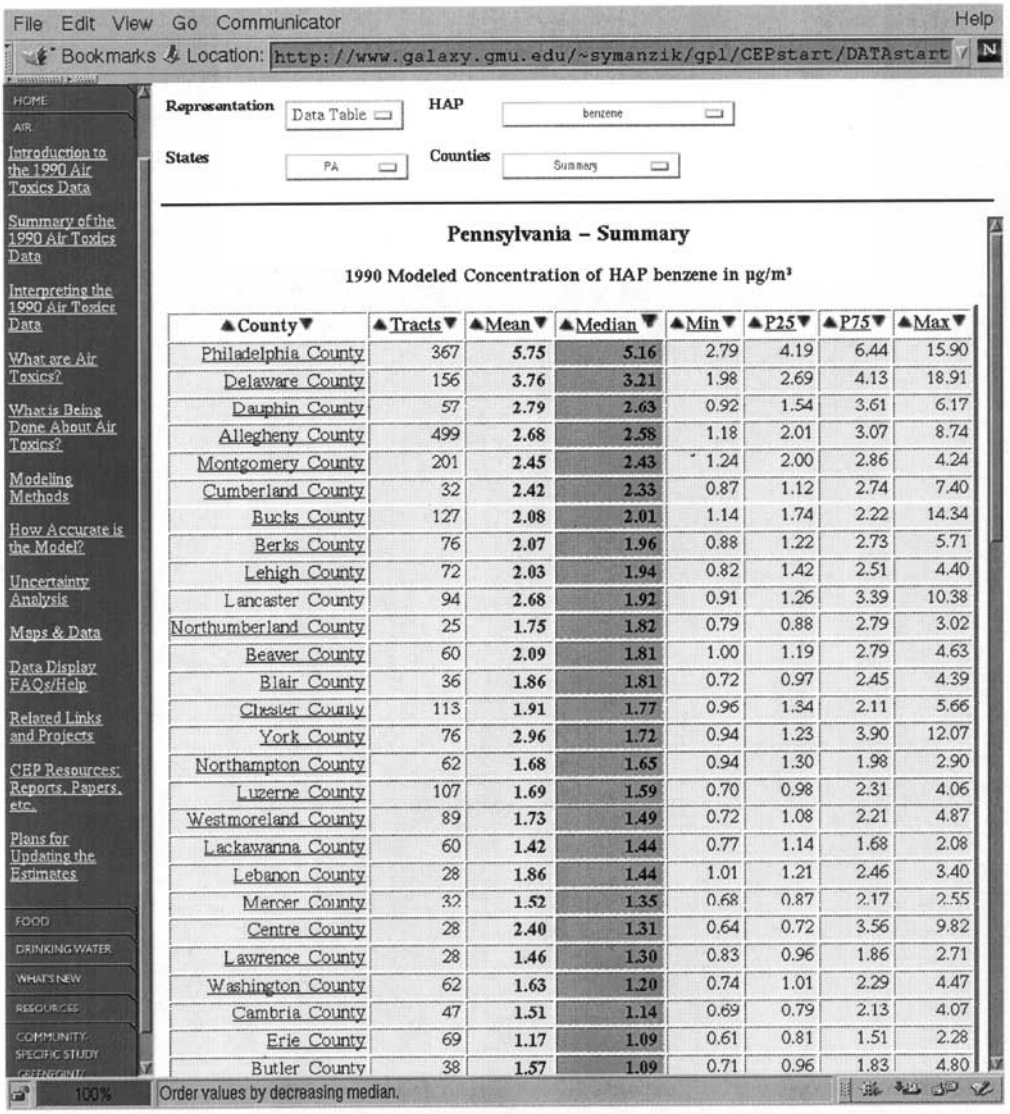


Figure 1 Micromap display at the state level for Pennsylvania, showing all of its 67 counties. Displayed is the median (with respect to the census tract estimates within each county) of the modeled 1990 benzene concentration in micrograms per cubic meter. Counties are ordered from highest to lowest median benzene concentration.

the user to rearrange the rows of the table in increasing or decreasing order as defined by the selected criterion. In the current view, the data are ordered from highest to lowest median benzene concentration. This is indicated through the larger downward arrow and the different-colored background of this data column. Through this interactivity, valuable information can be found in a larger table within seconds instead of



County	Tracts	Mean	Median	Min	P25	P75	Max
Philadelphia County	367	5.75	5.16	2.79	4.19	6.44	15.90
Delaware County	156	3.76	3.21	1.98	2.69	4.13	18.91
Dauphin County	57	2.79	2.63	0.92	1.54	3.61	6.17
Allegheny County	499	2.68	2.58	1.18	2.01	3.07	8.74
Montgomery County	201	2.45	2.43	1.24	2.00	2.86	4.24
Cumberland County	32	2.42	2.33	0.87	1.12	2.74	7.40
Bucks County	127	2.08	2.01	1.14	1.74	2.22	14.34
Berks County	76	2.07	1.96	0.88	1.22	2.73	5.71
Lehigh County	72	2.03	1.94	0.82	1.42	2.51	4.40
Lancaster County	94	2.68	1.92	0.91	1.26	3.39	10.38
Northumberland County	25	1.75	1.82	0.79	0.88	2.79	3.02
Beaver County	60	2.09	1.81	1.00	1.19	2.79	4.63
Blair County	36	1.86	1.81	0.72	0.97	2.45	4.39
Chester County	113	1.91	1.77	0.96	1.34	2.11	5.66
York County	76	2.96	1.72	0.94	1.23	3.90	12.07
Northampton County	62	1.68	1.65	0.94	1.30	1.98	2.90
Luzerne County	107	1.69	1.59	0.70	0.98	2.31	4.06
Westmoreland County	89	1.73	1.49	0.72	1.08	2.21	4.87
Lackawanna County	60	1.42	1.44	0.77	1.14	1.68	2.08
Lebanon County	28	1.86	1.44	1.01	1.21	2.46	3.40
Mercer County	32	1.52	1.35	0.68	0.87	2.17	2.55
Centre County	28	2.40	1.31	0.64	0.72	3.56	9.82
Lawrence County	28	1.46	1.30	0.83	0.96	1.86	2.71
Washington County	62	1.63	1.20	0.74	1.01	2.29	4.47
Cambria County	47	1.51	1.14	0.69	0.79	2.13	4.07
Erie County	69	1.17	1.09	0.61	0.81	1.51	2.28
Butler County	38	1.57	1.09	0.71	0.96	1.83	4.80

Figure 2 Tabular display at the state level for Pennsylvania showing the modeled 1990 benzene concentration in micrograms per cubic meter. Displayed are the number of census tracts and summary statistics (e.g., mean and median) with respect to the census tract estimates within each county. Counties are ordered from highest to lowest median benzene concentration. While only 27 of Pennsylvania's 67 counties are visible in this figure, the remaining 40 counties are accessible through the right scrollbar when looking at these data on the Web.

having to search sequentially through each table column to find largest, smallest, and—even more complicated—center values.

It should be noted that, when designing our tabular display, we have paid particular attention to recommendations from the cognitive sciences. Numbers have been rounded to two significant digits (with respect to the smallest number in any given

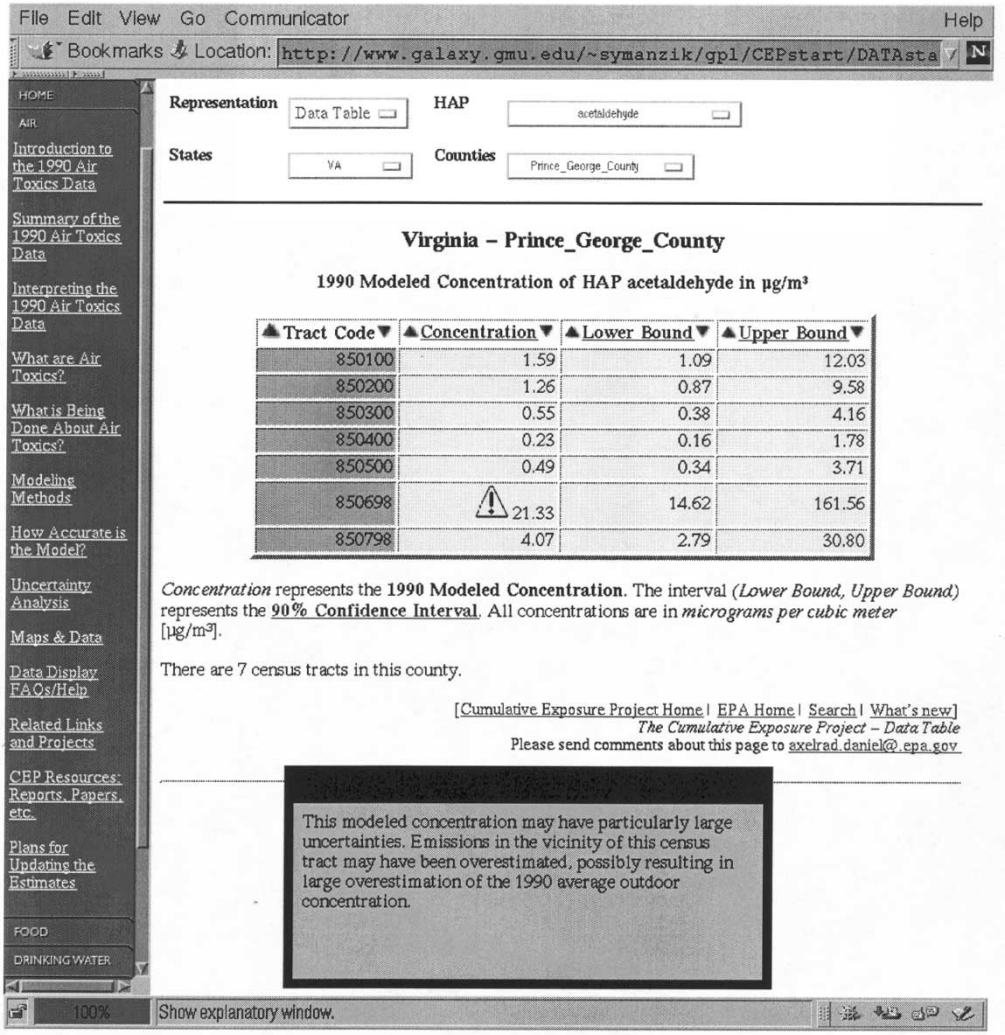


Figure 3 Tabular display at the county level for Prince George County, Virginia, showing an unusual modeled 1990 acetaldehyde concentration. This observation has been marked as an overestimate after EPA's inspection of the modeled 1990 concentrations. After clicking on the "!" icon, the small explanatory message window displayed at the bottom of this figure pops up.

display) and colors have been selected to produce a pleasant visual effect. Also, icons have been incorporated that warn users of suspect numerical values. These are usually particular census tract/HAP combinations for which EPA assumes that the modeled 1990 concentrations are considerably overestimated (Figure 3). In addition to rounded data, the CEP Web site makes raw data available. Raw data are most useful for users who want to conduct additional statistical analyses.

The third navigational tool is based on so-called hierarchical clickable micromaps in the GPL environment. This approach combines LM plots and the GPL and extends their joint features to allow dynamic access to complex, geographically referenced data.

The user can mouse-click on a region (state, county) in the map and the display changes with respect to the selected region. Thus, micromaps serve as a navigational tool, but each individual LM plot is a sophisticated statistical display by itself, as described in the "Micromaps" section of this paper. The idea of using micromaps simultaneously as a statistical display and for navigational purposes, first considered for EPA's CEP Web site, may be of benefit in innumerable future applications.

In the CEP Web site, after the user selects one state in the top US micromap (or tabular display), a new micromap (or tabular display) appears—this time at the county level for the selected state. This time, the user can select a county and reach the lowest level in this hierarchy—a graphical display (or tabular display) at the census tract level. This geographic selection is an easy way to maneuver through the 60,803 census tracts, even for inexperienced users of the Web. At the higher levels (i.e., the US or state level), statistical summaries such as means, medians, minima, maxima, and quartiles are displayed for all census tracts in each region (Figure 2). At the census tract level, uncertainty bounds are displayed (Figure 3).

Creation of Micromaps

Before we can implement LM plots in the GPL environment, generalized maps that form the basis of micromaps have to be created. Previous LM plots described in the literature (3–5) use hand-created generalized maps, for example, of the United States or the countries belonging to the Organisation for Economic Cooperation and Development. The creation of such a generalized map by hand typically requires several hours of work—an option that is clearly not feasible if we are interested in generalized maps for all 50 states or all counties within the United States.

We are not aware of any existing generalized map of the 50 US states or of any of the more than 3,000 counties in the United States that satisfy our specific needs—i.e., that is available in electronic format and provides the required level of generalization. Unfortunately, regular maps are unsuitable for use in LM plots, mostly for two reasons.

First, micromaps in printed form or on a computer display typically are smaller than 2 inches by 2 inches. In this scale, a small region such as Washington, DC, would be invisible on a US map if drawn to scale. Therefore, micromaps require an exaggeration of small regions so that these regions become visible and can be color-coded.

Second, in an interactive environment such as the Web, the number of line segments determines how fast a new display is drawn and an area is filled with color. The fewer edges a map has, the faster the boundary information is passed from the Web server to the client's computer and the faster the entire graphical display is drawn.

Therefore, it is necessary to develop procedures for creating generalized maps with little user interaction. These procedures have to extract selected regions from a larger file; they also have to smooth and simplify boundaries by removing details, but keep the topological integrity of a real unit. Micromaps should not end up with holes, and neighboring regions in the original map should remain neighboring regions in the generalized map.

The creation of generalized maps for use in the CEP Web site starts with boundary files describing geographical regions and with attribute data describing the characteristics of the regions. These data, the boundary files and attribute data, are often stored in geographic information systems (GIS). In our case, we make use of ArcView (ESRI, Redlands, CA), a desktop GIS package, which is one of the most popular of GIS soft-

ware. For future use of the developed procedures for map generalization, users need to have access to ArcView and the geographical data must be available in shapefile format. Obviously, the generalized maps only have to be created once—users of the CEP Web site do not have to deal with this issue.

When creating generalized maps of the United States at the sub-state level, we start with a shapefile of the entire United States at the preferred level of geography (county, township in the New England states, census tract, or even block group). We assume that in the attribute table each record or areal unit includes a state identifier indicating to which state the areal unit belongs. A procedure written in ArcView's Avenue scripting language allows users to extract the boundary of the selected geographical level (for instance, county) by states to create a shapefile for each state. Thus, each state can be individually displayed.

Most boundary files of the United States have a relatively high level of resolution, exceeding the required resolution level for micromaps. As explained earlier, high resolution and detailed data inhibit the fast processing and display of maps on the Web. Therefore, there is a need to smooth, or simplify, the boundary by removing details but preserving the topological integrity of the areal unit.

A set of Avenue scripts based upon the Douglas-Peucker line generalization algorithm (28) has been developed to generalize boundaries to expedite the processing and display of maps on the CEP Web site. The Douglas-Peucker algorithm has been implemented in many GIS packages (including ESRI's ARC/INFO) and has been used on numerous occasions. However, the algorithm was designed to generalize linear features such as rivers and roads. It was not intended to generalize polygonal features, which is what is required for this project. The major challenge in using the Douglas-Peucker algorithm to generalize polygonal features is to maintain the topological integrity among polygons. This means that neighboring relationships among polygons have to be maintained even after the polygon boundaries have been generalized. The algorithm designed for this project can generalize polygon boundaries and, at the same time, retain the topological relationships among polygons. The detail of the algorithm is beyond the scope of this paper, but will be described and published elsewhere.

The generalization process could be performed before individual maps (by states or by counties) are extracted by the first algorithm or after each state or county file has been created. However, it is desirable to generalize maps by individual states or counties instead of the entire country because boundaries of different states have different levels of cartographic complexity. Thus, different parameter values for the generalization process have to be used to yield desirable generalization results for different states or counties. After boundaries have been generalized at the state or county level, the boundaries, in ArcView shapefile format, are converted into ASCII data in a simple format: polygons depicted by a set of points in latitude and longitude. These coordinates are later used for the micromap displays on the CEP Web site.

Implementation Issues

The CEP Web site has been designed using numerous common Web formats and styles. While the explanatory pages are mostly based on HTML files, GIF images, and PDF files for larger documents, the pages that provide access to the data are based on Java and C code, accessible through Common Gateway Interface (CGI) scripts, and automatically created HTML and JavaScript documents. Each user interaction that

results in a new display (e.g., selecting a new state or a different HAP) invokes such a C-CGI script. This script is a C program that first reads the current parameter settings (representation type, HAP, state, county) from the active Web page and then creates a new HTML/JavaScript Web page. This new page is linked once again to the same C-CGI script, but with different parameters active.

Other than two C-CGI scripts that are responsible for the top-right menu and the lower-right data display on the CEP Web site, there exists no hard-coded document that is used for the data display. Each newly visible Web page is created on the fly through the C-CGI scripts.

Developing these two C-CGI scripts required several weeks of programming time. Valuable references during the implementation process were the books by Graham (29) for HTML, Hoque for JavaScript (30), and Eckel (31) for Java. In addition, books by Weinman et al. (32,33) have been a very good source for general design issues of Web pages. The CGI used for the CEP Web site is based on code developed by Thomas Boutell and freely available on the Web at http://www.boutell.com/cgic/.

Two non-statistical Web sites have significantly influenced the design and some of the interactive features of the CEP Web site:

- http://www.usnews.com/usnews/edu/college/corank.htm; in particular, http://www.usnews.com/usnews/edu/college/rankings/natunivs/natu_a.htm, which allows users to sort university rankings according to different criteria.
- http://www.sport1.de; first click on "Fussball," then on "Bundesliga," then on "Tabelle." This is a good example of how to organize frames and update soccer standings according to different criteria (by round, home or away, etc.).

It should be noted that the CEP Web site does not use any commercial database program to access the HAP data. Because the data originate directly from a statistical package that has been used for the modeling, they have not been fed into a database program first. Instead, a three-layer tree-shaped directory structure is used; from this database, an individual data file can be directly accessed based on its state, county, and census tract federal information processing standard code.[2] Data files have been kept as small as possible—no file contains more information than the Web site needs for each newly visible Web page. This makes it unnecessary to search in files, accelerating access to the data. Also, all summary statistics at higher levels (US, county) have been precalculated to speed up access to minima, maxima, means, medians, and lower and upper quartiles.

Discussion

A first version of the CEP Web site went online in March 1998, providing general information and documents related to the project. A major remodeling of the site, providing more extensive descriptions of the air toxics data, took place in November 1998. The release of the data through the CEP Web site was scheduled to begin with interactive tables going online in December 1998. The micromap displays were expected to be posted in the following months. However, EPA ultimately decided not to make the modeled air

[2] For more information on federal information processing standard (FIPS) codes, visit http://www.epa.gov/enviro/html/codes/state.html.

toxics concentrations available through its public Web site, due to concerns that the estimates for 1990 may not be representative of current conditions. EPA has made files of the air toxics concentrations available to the public by request; data from the study have been more widely disseminated, through the Web as well as other mechanisms, by state and local environmental agencies and by other organizations. Recently, the Environmental Defense Fund added data and results based on the CEP to their http://scorecard.org Web site.

While most work on the CEP Web site was completed as originally planned, micromaps have not been fully integrated into the GPL yet. At the current stage, it seems to be advisable to revise the approach for joining the GPL with micromaps. The BLS GPL was originally developed in 1996, and a new (commercial) version of the GPL has been recently developed by an up-and-coming software company. This new GPL is currently in its alpha testing phase and it has been scheduled for release in the spring of 2000. The new GPL already contains many of the features that were intended for the BLS GPL but were never included in the old version. The inclusion of maps and micromaps based on the generalized maps developed for the CEP Web site appears to be straightforward with the new GPL. Although the BLS GPL is still a useful tool, in particular because its source code can be obtained for free from BLS, it seems to be advisable to use the new GPL for larger future applications. Several employees in federal agencies such as the federal Centers for Disease Control and Prevention and BLS have already expressed interest in investigating the use of the new GPL. Even if concerns about data timeliness mean that the CEP Web site never becomes publicly accessible, it seems to be likely that the idea of hierarchical clickable micromaps in the (new) GPL environment might be used in another federal project in the near future.

Acknowledgments

EPA funded the majority of the work behind this paper under contract Nos. 8W-0662-NAEX and 8W-1712-NTEX and cooperative agreement number CR825564-01-0. Additional federal agencies, BLS and the National Center for Health Statistics, supported some facets of this work. The article has not been subject to review by any of these agencies; it does not necessarily reflect their views, and no official endorsement should be inferred. The conclusions and opinions are solely those of the authors and are not necessarily the views of the agencies.

References

1. Carr DB. 1994. *Converting tables to plots.* Technical Report 101. Center for Computational Statistics, George Mason University, Fairfax, VA.
2. Carr DB, Olsen AR, White D. 1992. Hexagon mosaic maps for displays of univariate and bivariate geographical data. *Cartography and Geographic Information Systems* 19(4):228–36, 271.
3. Carr DB, Pierson SM. 1996. Emphasizing statistical summaries and showing spatial context with micromaps. *Statistical Computing and Statistical Graphics Newsletter* 7(3):16–23.
4. Carr DB, Olsen AR, Courbois JP, Pierson SM, Carr DA. 1998. Linked micromap plots: Named and described. *Statistical Computing and Statistical Graphics Newsletter* 9(1):24–32.

5. Carr DB, Olsen AR, Pierson SM, Courbois JP. Using linked micromap plots to characterize Omernik ecoregions. *Data Mining and Knowledge Discovery.* To appear.
6. Carr DB, Valliant R, Rope D. 1996. Plot interpretation and information webs: A time-series example from the Bureau of Labor Statistics. *Statistical Computing and Statistical Graphics Newsletter* 7(2):19–26.
7. Symanzik J, Axelrad DA, Carr DB, Wang J, Wong D, Woodruff TJ. 1999. *HAPs, micromaps and GPL—Visualization of geographically referenced statistical summaries on the World Wide Web.* American Congress on Surveying and Mapping, Annual Proceedings of the ACSM-WFPS-PLSO-LSAW 1999 Conference. March 13–17, 1999. CD-ROM.
8. US Environmental Protection Agency. 1998. *National air quality and emissions trends report, 1997.* Research Triangle Park, NC: US Environmental Protection Agency. EPA 454/R-98-016.
9. Ministry of Public Health. 1954. *Mortality and morbidity during the London fog of December 1952.* Reports on Public Health and Medical Subjects, No. 95. London, England: Her Majesty's Stationery Office.
10. Schrenk H, Heimann J, Clayton G, Gafafer W, Wexler H. 1949. *Air pollution in Donora,* PA. Public Health Service Bulletin No. 36. Washington, DC: Public Health Service.
11. US Environmental Protection Agency. 1994. *Technical background document to support rulemaking pursuant to Clean Air Act Section 112(g): Ranking of pollutants with respect to human health.* Research Triangle Park, NC: US Environmental Protection Agency. EPA-450/3-92-010.
12. Glickman T, Hersh R. 1995. *Evaluating environmental equity: The impacts of industrial hazards on selected social groups in Allegheny County, Pennsylvania.* Discussion Paper 95-13. Washington, DC: Resources for the Future.
13. Perlin SA, Setzer RW, Creason J, Sexton K. 1995. Distribution of industrial air emissions by income and race in the United States: An approach using the Toxic Release Inventory. *Environmental Science and Technology* 25:69–80.
14. Cote I, Vandenberg J. 1994. Overview of health effects and risk-assessment issues associated with air pollution. In: *The vulnerable brain and environmental risks. Vol. 3: Toxins in air and water.* Ed. R Isaacson, K Jensen. New York: Plenum Press.
15. Cupitt L, Cote I, Lewtas J, Lahre T, Jones J. 1995. *EPA's Urban Area Source Research Program: A status report on preliminary research.* Washington, DC: US Environmental Protection Agency. 600/R-95/027.
16. Hassett-Sipple B, Cote I, Vandenberg J. 1991. Toxic air pollutants and non-cancer health risks—United States and a Midwestern urban county. *Morbidity and Mortality Weekly Report* 40:278–80.
17. Office of Air Quality Planning and Standards. 1990. *Cancer risk from outdoor exposure to air toxics.* Research Triangle Park, NC: US Environmental Protection Agency. EPA-450/1-90/004a.
18. Kelly T, Mukund R, Spicer C, Polack A. 1994. Concentrations and transformations of hazardous air pollutants. *Environmental Science and Technology* 28:378A–87A.
19. Woodruff T, Axelrad D, Caldwell J, Morello-Frosch R, Rosenbaum A. 1998. Public health implications of 1990 air toxics concentrations across the United States. *Environmental Health Perspectives* 106:245–51.
20. Rosenbaum A, Ligocki M, Wei Y. 1999. *Modeling cumulative outdoor concentrations of hazardous air pollutants.* Revised Final Report. San Rafael, CA: Systems Applications International, Inc. http://www.epa.gov/CumulativeExposure.
21. Caldwell J, Woodruff T, Morello-Frosch R, Axelrad D. 1998. Application of hazard identification information for pollutants modeled in EPA's Cumulative Exposure Project. *Toxicology and Industrial Health* 14:429–54.

22. Anderson G. 1983. *Human exposure to atmospheric concentrations of selected chemicals.* Vol. 1. US Environmental Protection Agency. NTIS PB84-102540.
23. US Environmental Protection Agency. 1991. *Toxics Release Inventory 1987–1990.* Electronic Version on CD-ROM. Washington, DC: US Environmental Protection Agency.
24. US Environmental Protection Agency. 1993. *Regional interim emission inventories (1987–1991), volume 1: Development methodologies.* Research Triangle Park, NC: US Environmental Protection Agency. EPA-454/R-93-021a.
25. EH Pechan and Associates. 1994. *Emissions inventory for the National Particulate Matter Study.* EPA Contract 68-D3005. Springfield, VA: EH Pechan and Associates.
26. Cleveland WS. 1993. *Visualizing data.* Summit, NJ: Hobart Press.
27. Cleveland WS. 1994. *The elements of graphing data.* Summit, NJ: Hobart Press.
28. Douglas DH, Peucker TK. 1973. Algorithms for reduction of number of points required to represent a digitized line or its character. *The Canadian Cartographer* 10:112–22.
29. Graham IS. 1998. *HTML 4.0 sourcebook.* 4th Ed. New York, NY: John Wiley and Sons.
30. Hoque R. 1997. *Practical JavaScript programming.* New York, NY: M&T Books.
31. Eckel B. 1998. *Thinking in Java.* Upper Saddle River, NJ: Prentice Hall.
32. Weinman L. 1996. *Designing Web Graphics.2.* Indianapolis, IN: New Riders.
33. Weinman L, Lentz JW. 1998. *Deconstructing Web Graphics.2.* Indianapolis, IN: New Riders.

Geographic Analysis of Childhood Lead Exposure in New York State

Thomas O Talbot, MSPH,* Steven P Forand, MA, Valerie B Haley, MS
Geographic Research and Analysis Section, Bureau of Environmental and Occupational Epidemiology, New York State Department of Health, Troy, NY

Abstract

This study examines the geographic variation in the blood lead levels (BLLs) of New York State children using spatial filtering, contour mapping, and regression techniques. Data for 364,917 children tested for BLLs prior to age two were extracted from New York State's electronic blood lead reporting system. Spatial filtering methods were used to determine which areas of the state had the highest prevalence of children with elevated BLLs (BLLs ≥10 µg/dL). The method used a variable filter size to allow for the simultaneous evaluation of urban and rural areas of the state. The results showed that several upstate urban areas had the highest proportion of children with elevated BLLs. Screening rates were also found to be higher in areas with a high proportion of children with elevated BLLs, indicating that areas with a high risk of lead exposure were well screened. Multiple regression analysis, using areas made up of merged zip code regions as the units of observation, was conducted to describe the relationship between the prevalence of children with elevated BLLs and community characteristics. High prevalence of elevated BLLs was predicted in areas with older housing stock, a smaller proportion of high school graduates, and a larger proportion of black births. Separate models were developed for New York City and the rest of the state, since the effect of the variables was lower in New York City.

Keywords: lead poisoning, disease surveillance, socioeconomic status, New York State, children

Introduction

Lead poisoning is considered to be one of the most prevalent and preventable childhood health problems in New York State (1). Blood lead levels (BLLs) as low as 10 micrograms per deciliter (µg/dL) are associated with adverse effects on learning, behavior, and growth. Higher BLLs can lead to anemia, severe central nervous system damage, and even death (2). Young children are at a heightened risk for elevated BLLs because lead intake as a proportion of body mass and metabolic uptake rates are higher in children than in adults. In addition, the central nervous systems of children are more vulnerable during early childhood development (3). Normal mouthing activity may also result in the ingestion of contaminated dust and soil.

A major source of lead exposure for children is lead-based paint in older houses. Indoor house dust may be lead-contaminated when the paint is chipped, peeling, deteriorating, or spread during renovation. Children in New York State are at particular risk because the state has the largest proportion (47%) and largest number of housing units built before 1950 (3.4 million units) of any state. Other sources of lead exposure

* Thomas O Talbot, New York State Department of Health, 547 River Street, Room 200, Troy, NY 12180-2216 USA; (p) 518-402-7950; (f) 518-458-7959; E-mail: tot01@health.state.ny.us

include lead in soil and dust from external paint, industrial emissions, and gasoline; lead in water pipes with lead solder; and lead brought into the home from occupational exposures, hobbies, and ceramic ware.

The prevalence of children with elevated BLLs is not evenly distributed across New York State. Locating areas with a high prevalence of children with elevated BLLs is important for identifying possible exposures in those areas. Two methods were used to examine the geographic variation in the prevalence of elevated BLLs across New York. One method used spatial filtering techniques to identify and display these areas. The other method involved using multiple regression techniques to identify community characteristics associated with areas with high prevalence of elevated lead values.

Displaying rates of disease over a large geographic area has always been problematic. State health departments traditionally display health outcome rates at the county level. These maps may not be very informative; it is difficult to identify high-incidence areas, which may either be localized within part of a county or cross county boundaries. Mapping health outcome data at smaller geographic levels, such as census tract or zip code, often produces rates of disease that vary widely due to chance alone, especially when the health outcome under analysis is rare. In addition, data regarding the underlying population size are often obscured when displayed in this manner.

Population density varies widely across the state. As expected, the distribution of health outcomes follows the population distribution. It therefore becomes necessary to take into account varying population densities when displaying the data on a statewide basis. To accomplish this we used spatial filtering techniques that control the size of the population for which disease rates are estimated. Stabilized rates were then obtained and displayed as a continuous distribution across the state. The geographic patterns in BLLs and blood lead screening were mapped. These maps are useful for identifying areas with high prevalence of exposed children and areas where blood lead screening rates are low.

We also developed a regression model to predict prevalence of elevated BLLs in areas for which insufficient screening data exist. Regression analyses can also identify areas with higher rates of elevated BLLs than we would expect from our model. These areas can be examined more carefully to better characterize the factors that contribute to lead exposure in communities.

Elevated childhood BLLs have been associated with housing and sociodemographic characteristics including older housing stock, a higher proportion of children living below the poverty level, and a lower proportion of high school graduates (4–8). They have also been associated with a higher proportion of households headed by a female, a higher percentage of minority births, and higher population density. Studies have also shown that children's lead levels tend to be higher in the summer months (9–11). Multiple regression techniques were used to examine the relationship between housing and sociodemographic characteristics and children's BLLs across the state. The results from the regression analysis were also examined geographically and compared with the results from the spatial filtering methods.

The objectives of these analyses were to:

- Identify geographic areas of New York State with high prevalence of elevated BLLs in children.
- Identify areas with low blood lead screening rates in children.

- Develop a model describing the prevalence of high BLLs as a function of community characteristics using multiple regression analysis.
- Identify communities that could be targeted with additional educational, screening, and remediation programs.

Materials and Methods

Data

New York State requires universal screening for lead in children under the age of six. The New York State Department of Health currently receives approximately 870,000 blood lead reports a year for both children and adults. Blood lead reports include the name and address of the child, the name and telephone number of a parent or guardian, the blood lead value, and the method by which the blood sample was obtained. A total of 537,704 records for children who were born in 1994 and 1995, resided in New York State, and were screened for blood lead at least once prior to age two were extracted from the lead reporting system (12). This cohort represents 69.3% of the births in New York State in 1994 and 1995.

For children who were screened more than once prior to age two, the highest BLL measure taken by venipuncture was used because these samples are less susceptible to environmental contamination than finger stick samples (13). Finger stick measures were used if no venipuncture measures were available for the child. BLLs were categorized into two groups: less than 10 µg/dL, and 10 µg/dL and above. This is the intervention level recommended by the federal Centers for Disease Control and Prevention. The data were then aggregated at the zip code level. In cases where the child had a missing or invalid zip code, address-matching software (14) was used to assign the correct zip code. To obtain valid zip codes in cases where no street or town information was available for the child, the parent or guardian's phone number was matched to digital phone directories (15). Valid zip codes could not be obtained for 3.8% of the children screened. The remaining dataset contained 364,917 children with blood lead tests, which represented 66.7% of children under two years old born in 1994 and 1995.

To obtain a denominator for screening rates, data on all births in New York State in 1994 and 1995 were obtained from the New York State Bureau of Vital Statistics (16). The data were aggregated at the zip code level and the percent of children tested and percent of black births were calculated. The dataset was then merged with housing and demographic data from the 1990 Census (17) for use in the regression analysis.

Spatial Filtering Methods

In this analysis, it was not practical to map the lead prevalence of the birth cohorts at the zip code level because many of the zip codes had few blood lead tests. Of the 1,601 zip codes, almost half had fewer than 50 children tested for blood lead. Spatial filtering techniques were developed to overcome some of the difficulties in mapping the data by small geographic area. These methods are a variation on the techniques described by Openshaw et al. (18,19), Turnbull et al. (20), and Rushton and Lolonis (21). In the first step, a layer of grid points 1 kilometer apart was created covering the entire state. The zip code file containing the number of births, the total number of lead tests, and the number of lead tests with results higher than 10 µg/dL was mapped, and the zip code

centroids were overlaid on the grid point file. Because there are large fluctuations in population density between the urban and rural portions of the state, a stable population denominator was chosen on which to base the screening rates. In this case, rates were based on a minimum of 200 children screened.

The following procedure was then carried out for each grid point. First, the nearest zip code centroid to the grid point was located. The number of children screened for lead and the number of children with elevated lead values were then tabulated. If the number of children screened was less than 200, the next nearest zip code centroid was located and the number of children screened and the number of children with high lead values were added to the values from the previous zip code. This process was continued until the total number of children screened reached 200, at which point the percent of children with elevated lead values was calculated for that grid point. This process was then carried out on each grid point until all grid points in the state had been assigned prevalence rates. Because the grid points were spaced more closely together than the zip code centroids, there was a significant overlap in the area sampled around one grid point and that of its neighboring grid point. To ensure that all information was used, a minimum radius of 0.75 times the spacing of the grid (in this case, 0.75×1.0 km=0.75 km), was used at each grid point. This radius was then extended, when required, to ensure that the minimum number of observations was captured at each grid point.

Contour modeling software (22) in conjunction with desktop mapping software (23) was used to create a contour model based on the percentage of children with elevated lead values at each grid point calculated in the previous step. In this way, it was possible to represent the data continuously as a moving average.

The percentage of children screened for lead was analyzed using a similar methodology. The New York State vital statistics file provided the number of children born in 1994 and 1995 in each zip code. The percentage of children screened at each grid point was calculated based on a minimum of 200 live births.

Regression Analysis Methods

Least squares regression was used to examine the relationship between children's BLLs and community characteristics. Housing variables examined were percent of houses built before 1940, percent of houses built before 1950, and percent of houses vacant. Socioeconomic variables examined were percent of adults age 25 and older who graduated from high school, percent of children under 5 years living below the poverty level, percent Hispanic, percent black births, percent of population that rents a home, and population density. An additional variable, the percent of children in each zip code who were tested in summer and fall (June to November), was also examined.

The percentage of children with elevated BLLs (i.e., $\geq$10 µg/dL) in each zip code, rather than the mean or geometric mean of the BLLs, was chosen as the dependent variable. This was necessary because labs report different detection limits of BLL and extreme results may be due to child-specific traits, such as pica, that could not be controlled for in this type of analysis. The dependent variable was log-transformed to normalize the distribution.

Zip code areas were used as the level of analysis. The distribution of the prevalence of high BLLs in the 1,601 zip codes was found to be bimodal due to a large number of zip codes having no elevated BLLs. This occurred primarily in areas of low population.

To adjust for this, zip codes with demographically similar adjacent neighbors were manually merged to create zip code groups with a minimum number of children screened. All zip codes with fewer than 100 children tested were merged; merging beyond this point provided greatly diminished returns in correlation. The estimates of the rates of elevated BLLs improved and the zero values changed to small values as the data were aggregated. The distribution of the BLLs also changed from a bimodal distribution to a normal distribution. The final number of zip code groups in the dataset used in the regression analysis was 740.

The regression models were developed using SAS software (SAS Institute, Cary, NC) (24). First, the shape and effect of the bivariate associations of each variable with the dependent variable for New York State were examined. Because the effects of the explanatory variables in New York City were muted compared with the effects in the rest of the state, and no available variables explained this difference, two separate models were developed: one for New York City and one (called the Upstate model) for the rest of the state. The maximum R^2 improvement technique was used to find the "best" one-variable model, two-variable model, etc. A parsimonious model was chosen in which the addition of variables would not add significantly more information to the model. The importance of interactions and curvilinearity were then investigated. Diagnostic methods were used to detect influential observations and multicollinearity, verify the linearity of the regression function, and verify the constant variance and normality of the error terms.

The regression residuals were mapped to more rigorously compare the observed and expected rates. The residuals were standardized for the New York City and Upstate models separately so that the variation would be on the same scale for the combined map. The Moran's I statistic was calculated to test for spatial autocorrelation between the regression residuals of neighboring zip codes.

Results

Spatial Filtering Results

The geographic distribution of the prevalence of elevated BLLs is mapped in Figure 1. The map shows that several urban areas of upstate New York have communities with a large portion of children with elevated lead levels; the cities of Buffalo, Rochester, Syracuse, Schenectady, and Albany all had areas in which more than 25% of the children tested had elevated BLLs. The city of Newburgh in Orange County also had a large area in which 20–25% percent of the children tested had elevated BLLs. The high-prevalence areas are small compared with the size of the counties that contain them, but the population densities of these areas are higher than those of their surrounding counties. In no area of New York City did more than 20% of the children have elevated BLLs. There was one small area of Brooklyn in which 15–20% of the children screened had elevated BLLs. The rest of the city had prevalence rates of elevated BLLs that were low compared with upstate urban areas.

The percent of children screened for lead is mapped in Figure 2. In New York State, the areas with the highest prevalence of elevated BLLs (≥25% of children screened with BLLs ≥10 µg/dL) also had the highest screening rates, with more than 80% tested. With the exception of the city of Schenectady, all of the major upstate cities with areas hav-

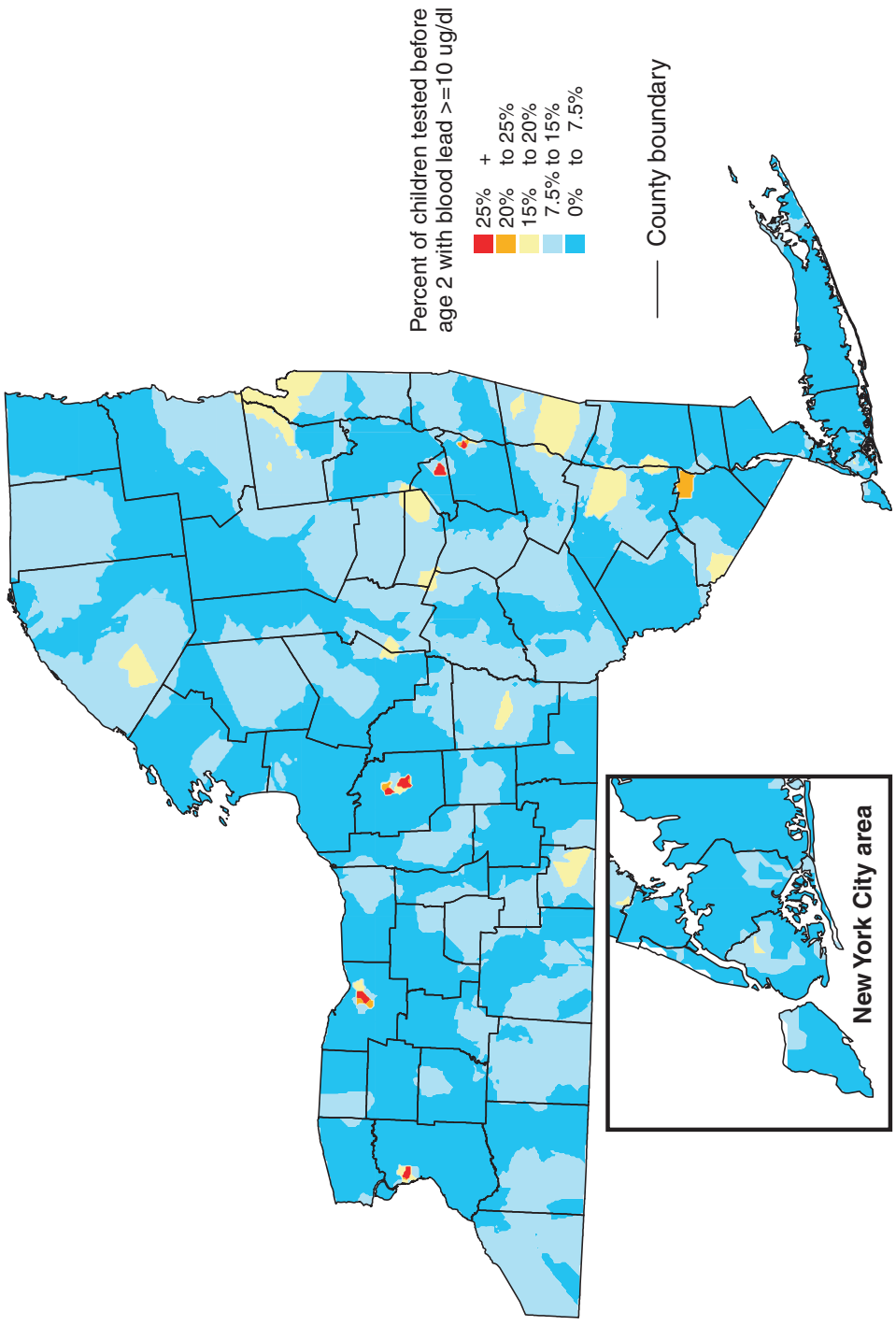


Figure 1 Geographic distribution of blood lead levels in New York State children based on spatial filter method. Rates are based on grid points capturing a minimum of 200 children screened.

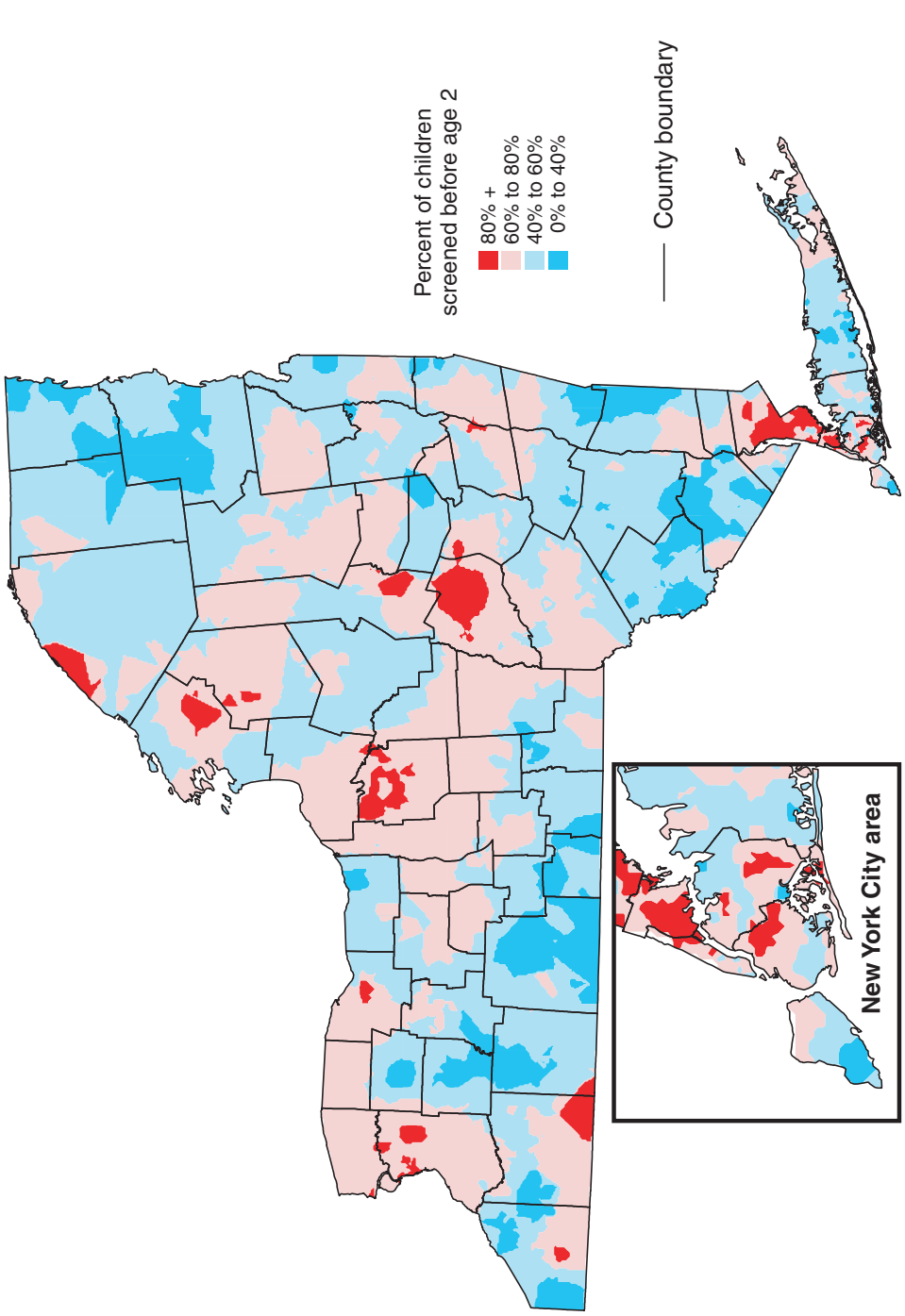


Figure 2 Percent of children screened for blood lead using spatial filter method. Rates based on grid points capturing a minimum of 200 live births.

ing a high prevalence of elevated BLLs also had high screening rates. This trend was also apparent in New York City. More than 80% of the children were screened in the area of Brooklyn that had the highest prevalence of children with elevated BLLs. In addition, high screening rates were also observed in areas of the South Bronx and upper Manhattan.

Regression Analysis Results

Age of housing, race, and education level were the most significant variables in explaining variation in BLLs (Table 1). The variable, percent of housing built before 1940, was selected for inclusion in the model because it was a slightly better predictor of BLL than percent of housing built before 1950. The poverty and education variables were highly correlated (R=–0.8), so including both in the model would have been problematic. Education was chosen for inclusion in the final model because it explained more variation. Other demographic variables only explained a small amount of the variation in BLLs, and were not included in the final model.

The model assumptions of linearity of the regression function and constant variance and normality of the error terms were valid. There were no influential observations, and multicollinearity was small.

The observed prevalence of elevated BLLs in the 740 zip code groups is mapped in Figure 3, and the prevalence predicted by the model is shown in Figure 4. The maps show similar patterns.

The regression residuals for the 740 zip code groups are mapped in Figure 5. Areas where more children have elevated BLLs than the model predicts have positive residuals. These areas appear to be clustered in Brooklyn, eastern upstate New York, and eastern Long Island. The Moran's I test statistics show that zip code groups with common boundaries were positively correlated after adjusting for the education, age of housing, and race variables used in the regression model ($p<0.001$). Spatial autocorrelation was also found to be inversely proportional to the distance between zip code group centroids up to a distance of 40 miles ($p<0.001$).

Figure 6 shows the association between the percent of children screened and the percent of children screened with elevated BLLs in the 740 zip code groups. The percent of children screened increases with the percent of children with elevated BLLs. The results were similar to those observed with the spatial filtering method (see Figures 1 and 2). In the areas of the state, excluding New York City, with the highest prevalence of elevated BLLs, 89% of the children had been screened. This effect was also seen in New York City.

Because the independent variable was log-transformed, the meanings of the regression coefficients are more difficult to interpret. Figure 7 contains conditional effect plots to facilitate interpretation of the results. These plots show the predicted value of elevated lead levels versus each of variables used in the model while holding the other variables at their means. The effect of the three major variables (age of housing, education, and race) is stronger in the Upstate model than in New York City model.

Discussion

Spatial filtering techniques were used to identify areas of the state with the highest prevalence of elevated BLLs in children (Figure 1). In addition, we mapped statewide

Table 1 Bivariate and Multivariate Regression Models for Blood Lead Level Analysis in New York City and Upstate New York

	NYC (n=165)[a]			Upstate (n=575)[b]		
	Bivariate Coefficient (adjusted R^2)	**Multivariate Coefficient (standard error)**	**P value**	**Bivariate Coefficient (adjusted R^2)**	**Multivariate Coefficient (standard error)**	**P value**
Intercept		1.6345 (0.1245)			2.3742 (0.2503)	
% housing units built before 1940	0.0088 (0.20)	0.0104 (0.0010)	<0.0001	0.0304 (0.56)	0.0238 (0.0012)	<0.0001
% high school graduates	–0.0123 (0.19)	–0.0064 (0.0015)	<0.0001	–0.0549 (0.36)	–0.0191 (0.0028)	<0.0001
% black births 1994–1995	0.0068 (0.26)	0.0071 (0.0007)	<0.0001	0.0209 (0.18)	0.0093 (0.0013)	<0.0001
% housing units built before 1950	0.0081 (0.19)			0.0028 (0.52)		
% children under five in poverty	0.0082 (0.16)			0.0385 (0.43)		
% renter-occupied housing units	0.0041 (0.06)			0.0268 (0.30)		
% children screened	0.0141 (0.27)			0.0111 (0.06)		
% vacant housing units	–0.0098 (<0.01)			0.0165 (0.06)		
Population density	1.3E–6 (0.04)			1.9E–5 (0.06)		
% Hispanic	0.0033 (0.03)			0.0062 (<0.01)		
% screened (July–November)	–0.0380 (<0.01)			1.6337 (<0.01)		
Adjusted total R^2		0.60			0.64	

[a] NYC model:
ln(% elevated BLL+1)=1.6345+0.0104×(% pre-1940 houses)+.0071×(% black)-0.0064×(% high school grads)

[b] Upstate model:
ln(% elevated BLL+1)=2.3742+0.0238×(% pre-1940 houses)+0.0093×(% black)-0.0191×(% high school grads)

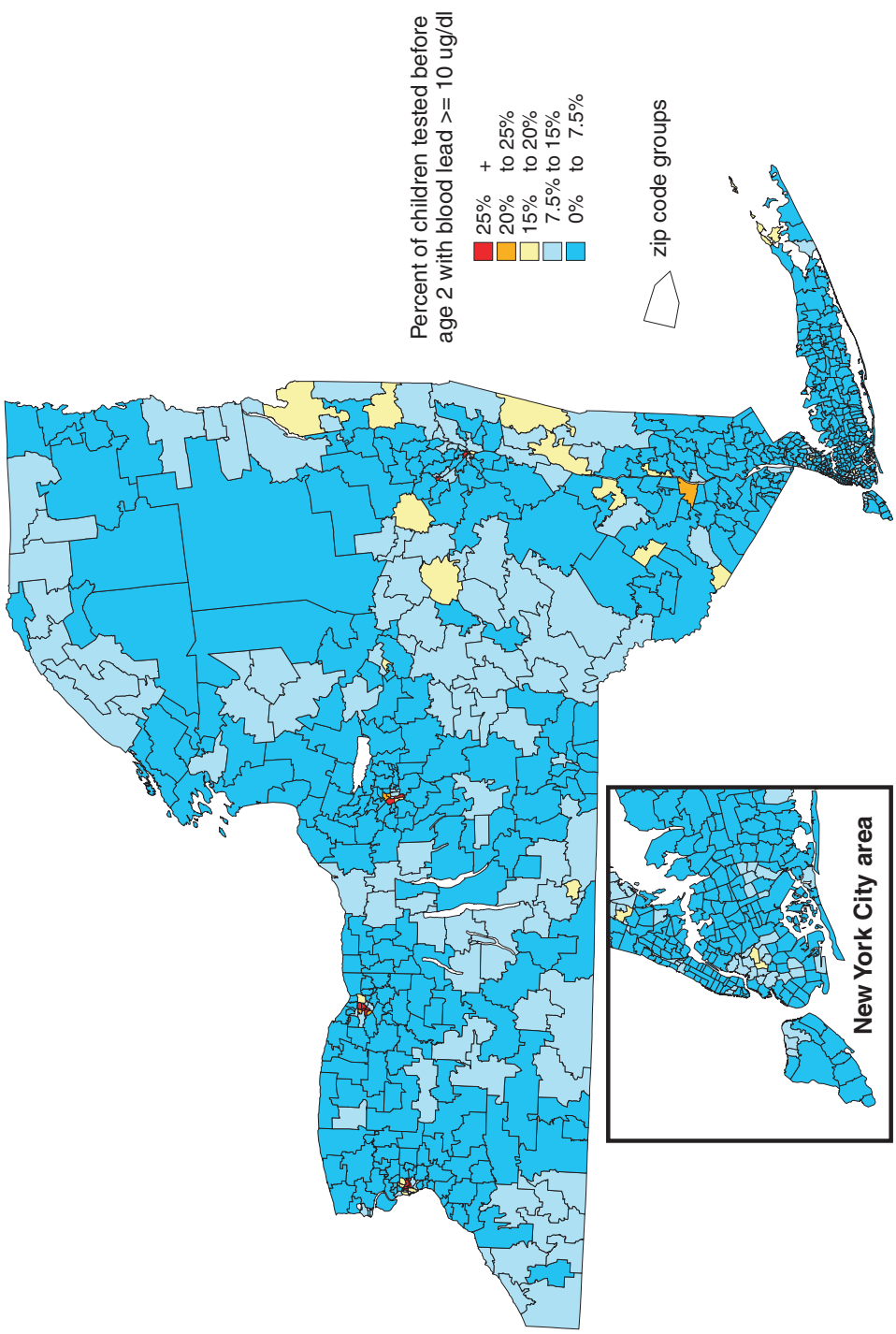


Figure 3 Distribution of elevated blood lead levels in New York State children by zip code group. Areas represent zip code groups with a minimum of 100 children screened.

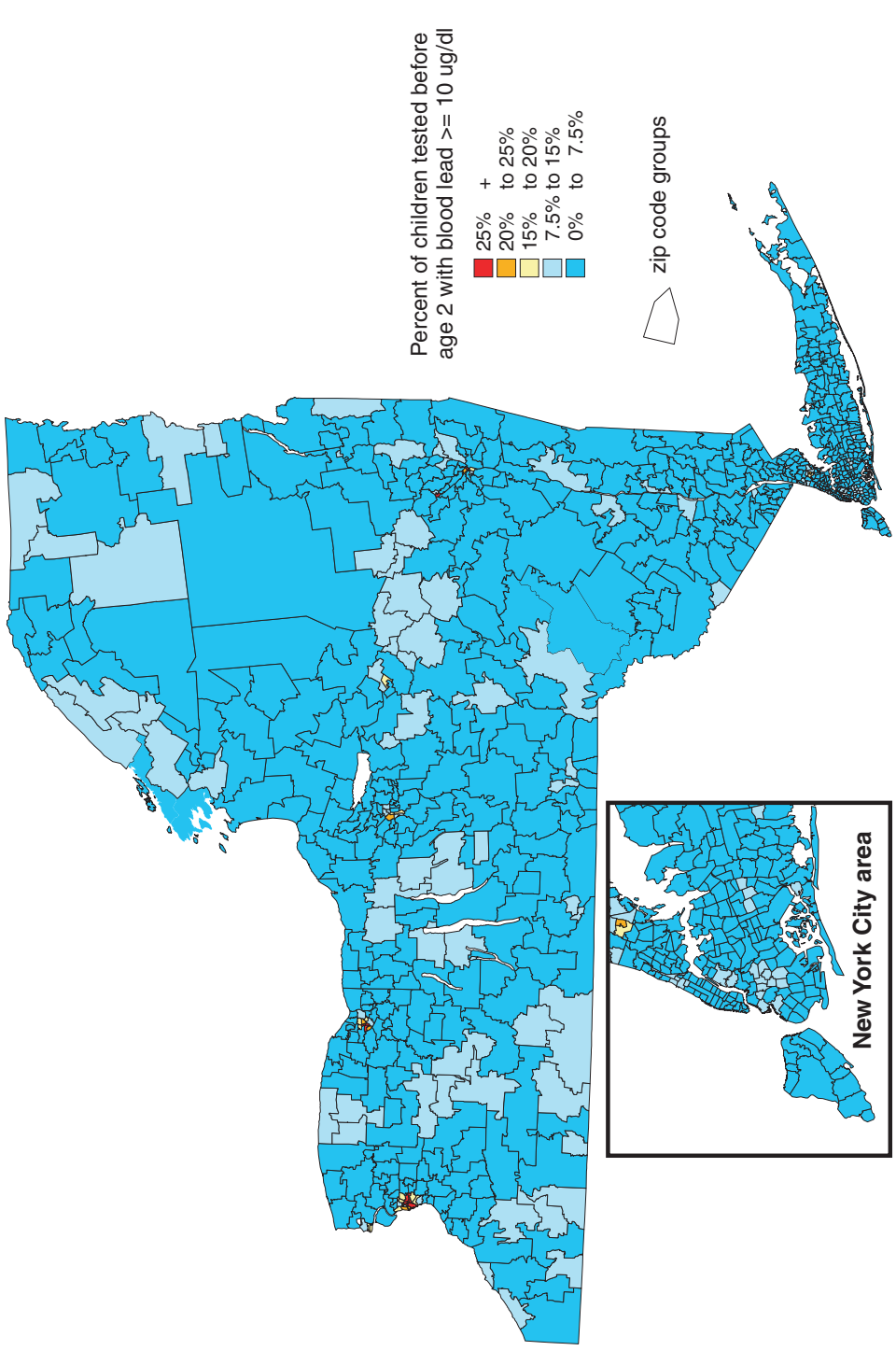


Figure 4 Predicted distribution of children with elevated blood lead levels. Areas represent zip code groups with a minimum of 100 children screened.

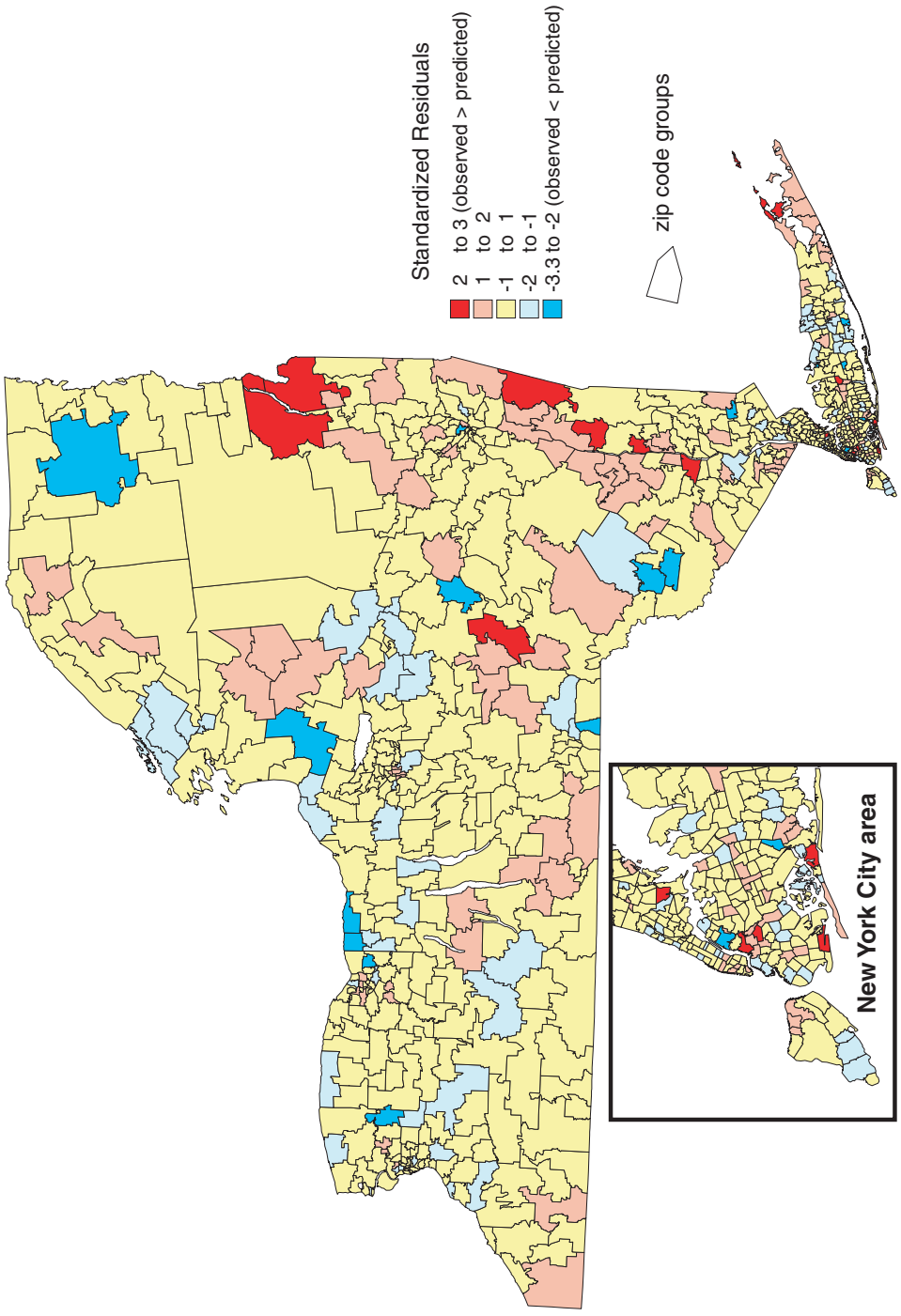


Figure 5 Distribution of regression residuals from the Upstate and New York City models. Areas represent zip code groups with a minimum of 100 children screened.

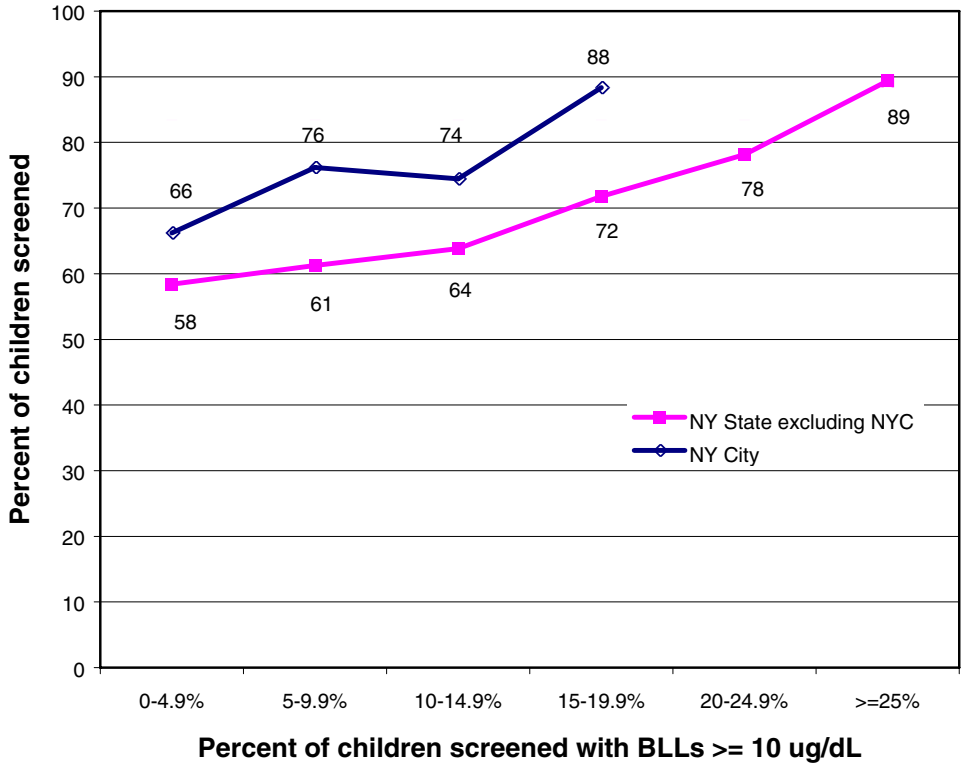


Figure 6 Lead screening and elevated lead rates in children born 1994–1995 and screened prior to age two. Data based on 740 New York State zip code groups. Separate plots are displayed for New York City and the rest of New York State.

screening rates (Figure 2). Screening rates were found to be highest in areas with high proportions of children with elevated BLLs, indicating that areas with a high risk of lead exposure were well screened.

Spatial filtering techniques were used to overcome some of the problems typically seen when mapping disease rates for small areas. Using a spatial filter based on population size rather than a fixed geographic size ensures that enough births were selected at each grid point to calculate a stable rate. One advantage of this method is that we were able to examine the geographic variation in BLLs throughout New York State, which has both rural and urban areas. In areas of high population density the process may capture data from only the nearest zip code. The rate would be stable because the population of that zip code would be large. In areas of low population density it was necessary for each grid point to capture data not only from the nearest zip code, but also from one or more of its neighboring zip codes in order to obtain a stable rate. Because of the varying geographic size of the filter it is possible to display and analyze the data for the entire state using a single summary level.

In addition, we were able to display the information as a continuous distribution

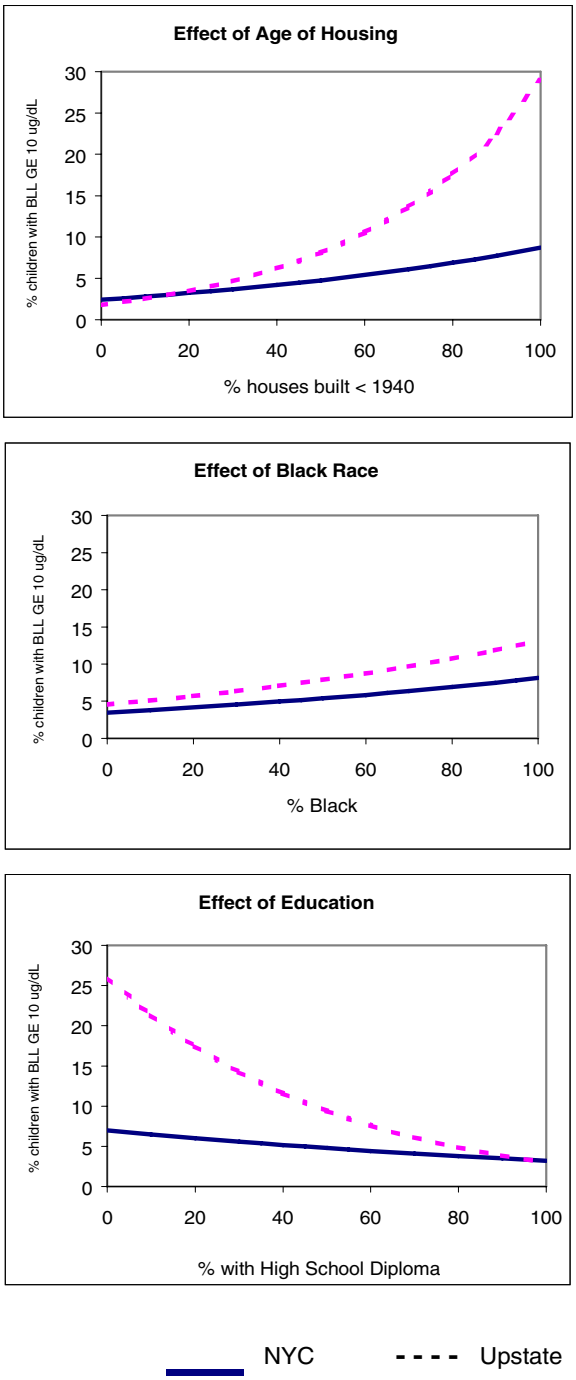


Figure 7 Conditional effect plots for the New York City and Upstate models. Mean values for percent of homes built before 1940=34%, percent black=12%, and percent high school graduates=77%.

rather than one limited to predefined political boundaries. It is illogical to assume that the distribution of a health outcome changes precisely at the border of a county, census tract, or some other predefined boundary. Using grid points spaced much more closely than the unit of analysis ensures that disease rates will be displayed more accurately as a moving average across political boundaries. The effects of the smoothing are critical in areas of low population density.

Multiple regression analysis identified age of housing, education, and race as the community characteristics associated with elevated BLLs. This is consistent with previous studies examining the socioeconomic characteristics associated with elevated BLLs (5–7). This analysis was carried out at the zip code area level. The smaller the units of analysis, the more homogenous the groups will likely be, and the more accurate the estimate of effect. However, the disadvantage of using small groups is that the estimates of disease rates are unstable. As the zip code areas were merged, the resulting correlation coefficient also increased. The increase in the percent-explained variation must be weighed against loss of information from using more heterogeneous areas.

Figure 3 shows the observed prevalence of elevated BLLs in the 740 zip code areas. This map shows a similar pattern of elevated BLLs in the upstate cities to that shown in Figure 1, which was produced using the spatial filtering method. The spatial filtering method is much faster than the manual merging of zip codes. Using it, one can easily plot maps based on the capture of different numbers of children and compare the results. The spatial filtering method can also be used to map individual-level data for which geocoded data are available.

The lower prevalence of elevated BLLs in New York City than in upstate areas is surprising considering that New York City is among the oldest and most densely populated areas of the state. New York City has a higher percentage of older housing stock and minority births and a lower percentage of high school graduates than does the rest of New York State. Based on these characteristics, New York City would be expected to have one of the highest prevalence rates of elevated BLLs in the state. This was not the case, however, and none of the variables examined in this analysis could explain the findings. One possible explanation for this difference is that New York City was one of the first localities to ban lead-based paint, prohibiting residential use beginning in 1960. Lead paint was not banned in the rest of the state until the federally imposed ban took effect in 1978 (26). Further research is needed to determine what factors have contributed to the city's lower prevalence of elevated BLLs.

Spatial autocorrelation of lead prevalence rates suggests that similarity among nearby areas was not completely accounted for by the variables in the regression model. Neighboring zip code groups may be similar because neighborhoods may cross zip code group boundaries, and because children may visit or use services in nearby areas. Environmental characteristics, housing characteristics, secondary occupational exposures, and behavioral factors may also be similar in neighboring zip code groups. Several areas in the state had higher rates of elevated BLLs than we would expect from our model. Further investigation would help us better characterize the risk factors that may be associated with these differences.

This study has several advantages over previously published geographic analyses of lead exposure. These include a larger number of children screened in the study, a higher screening rate, and the use of direct blood lead measurements rather than indicators such as erythrocyte protoporphyrin. In addition, improved mapping techniques,

the inclusion of a wide spectrum of population density, and restriction of the data analysis to children under the age of two allowed a more rigorous analysis of the geographic distribution of elevated BLLs in children.

References

1. New York State Lead Poisoning Prevention Advisory Council. 1994. *Annual report.* Albany, NY: New York State Department of Health.
2. Centers for Disease Control and Prevention (CDC). 1997. *Screening young children for lead poisoning: Guidance for state and local public health officials.* Atlanta: CDC.
3. Mushak P. 1992. Defining lead as the premiere environmental health issue for children in America: Criteria and their quantitative application. *Environmental Research* 59:281–309.
4. Bailey A, Sargent JD, Goodman D. 1994. Poisoned landscapes: The epidemiology of environmental lead exposure in Massachusetts children 1990–1991. *Social Science and Medicine* 39(6):757–66.
5. Sargent JD, Bailey A, Simon P, Blake M, Dalton MA. 1997. Census tract analysis of lead exposure in Rhode Island children. *Environmental Research* 74:159–68.
6. Sargent JD, Brown MJ, Freeman JL, Bailey A, Goodman D, Freeman DH. 1995. Childhood lead poisoning in Massachusetts communities: Its association with sociodemographic and housing characteristics. *American Journal of Public Health* 85(4):528–34.
7. Lanphear B, Byrd R. 1998. Community characteristics associated with elevated blood lead levels in children. *Pediatrics* 101(2):264–71.
8. McDade K. 1994. *Factors that relate to children at high risk for elevated blood lead levels in Syracuse, New York.* Thesis. SUNY College of Environmental Science and Forestry, Syracuse, NY.
9. Hunter JM. 1978. The summer disease, some field evidence on seasonality in childhood lead poisoning. *Social Science and Medicine* 12:85–94.
10. Stark AD, Quah RF, Meigs JW, De Louise ER. 1980. Season as a factor in variability of blood-lead levels in children. *Connecticut Medicine* 44(7):410–3.
11. US Environmental Protection Agency (EPA). 1995. *Seasonal rhythms of blood lead levels: Boston 1979–1983.* Washington, DC: EPA. EPA 747-R-94-003.
12. New York State Department of Health, Bureau of Occupational Health. 1994–1997. *Lead reporting system.* Albany, NY: New York State Department of Health.
13. Parsons PJ, Reilly AJ, Esernio-Jenssen D. 1997. Screening children exposed to lead: An assessment of capillary blood lead fingerstick test. *Clinical Chemistry* 43(2):302–11.
14. MapInfo Corporation. 1997, 1998. *MapMarker.* Versions 3.0 and 3.5. Troy, NY: MapInfo Corporation.
15. Digital Directory Assistance. 1996. *Phone disc New York and New England.* October. Bethesda, MD: Digital Directory Assistance.
16. New York State Department of Health, Bureau of Vital Statistics. 1994–1995. *Birth statistical files.* Albany, NY: New York State Department of Health.
17. US Census Bureau. 1993. *1990 decennial census of population and housing, summary tape file 3B (STF3B).* Washington DC: US Government Printing Office. July.
18. Openshaw S, Charlton M, Wymer C, Craft A. 1987. A Mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographic Information Systems* 1:335–58.

19. Openshaw S, Charlton M, Craft AW, Birch JM. 1988. Investigation of leukaemia clusters by use of a geographical analysis machine. *The Lancet* 272–3.
20. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. 1990. Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 132:S136–43.
21. Rushton G, Lolonis P. 1996. Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine* 15:717–26.
22. Northwood Geoscience. 1996. *Vertical Mapper.* Version 1.5. Nepean, Ontario, Canada.
23. MapInfo Corporation. 1997. *MapInfo Professional.* Version 4.5. Troy, NY: MapInfo Corporation.
24. SAS Institute, Inc. 1997. *SAS.* Version 6.12. Cary, NC: SAS Institute, Inc.
25. Cliff AD, Ord JK. 1975. *Spatial autocorrelation.* London: Pion.
26. New York State Legislative Commission on Toxic Substances and Hazardous Wastes. 1992. *Hearing report.* March.

Integration of Particulate Air Modeling with a GIS: An Exposure Assessment of Emissions from Two Phosphate-Processing Plants

Gregory V Ulirsch,* Debra L Gable, Virginia Lee
Agency for Toxic Substances and Disease Registry, US Public Health Service, Atlanta, GA

Abstract

Particulate air emissions from two phosphate-processing plants, which constitute the Eastern Michaud Flats Superfund site in Pocatello, Idaho, will be modeled as part of an ongoing public health assessment being conducted by the federal Agency for Toxic Substances and Disease Registry (ATSDR). The results of the air dispersion modeling will be integrated into a geographic information system (GIS) as a separate coverage. Based on current health-based guidelines for particulate air exposures, the results of the air dispersion modeling will be overlaid on base coverages provided by the TIGER/Line files, integrated with applicable demographic information from the US Census Bureau summary tapes. Demographic information on the population from the census block groups that are located completely within areas exposed at levels of health concern, will be abstracted from this overlay analysis. Other techniques (e.g., simple population density spreading or the kernel density method) will be employed to determine the demographic information from census block groups that are partially exposed at levels of health concern. Maps will be produced that display the areal concentration isopleths of the modeled particulate emissions and show which census block groups are affected. Demographic information from the created attribute tables will be queried and summarized to determine the total population exposed at levels of health concern, age structure, socioeconomic status, and other parameters. The results of the exposure assessment will be used as a basis for a separate study, an ecologic health study of mortality in the population exposed at levels of health concern. This study will be conducted at the University of North Carolina at Chapel Hill.

Keywords: exposure assessment, epidemiologic health studies, particulate matter, air modeling

Study Background

In 1996, the Shoshone-Bannock Tribes in Fort Hall, Idaho, contacted the federal Agency for Toxic Substances and Disease Registry (ATSDR) requesting an evaluation of current and historical exposures to particulate matter (PM) and other air emissions from two nearby phosphate-processing plants (one owned by the FMC Corporation [FMC] and one owned by the JR Simplot Corporation [Simplot]). Together, these corporations constitute the US Environmental Protection Agency's (EPA's) Eastern Michaud Flats Contamination (EMF) National Priorities List (NPL) site. Tribal and non-tribal community members have consistently expressed concern regarding the occurrence of

* Gregory Ulirsch, Centers for Disease Control and Prevention, 1600 Clifton Rd. (E-32), Atlanta, GA 30333 USA; (p) 404-639-0624; (f) 404-639-0653; E-mail: gru1@cdc.gov

asthma and upper respiratory infections that, they believe, are related to exposure to air pollutants emanating from the EMF site. In 1995, ATSDR completed a health study of persons living on the Fort Hall Indian Reservation to investigate concerns related to respiratory and renal disorders being treated by the Indian Health Service clinic. This study concluded that the prevalence of pneumonia and chronic bronchitis was statistically significantly elevated among Fort Hall participants as compared to participants at another Native American reservation. Testing of pulmonary function in the Fort Hall sample showed decreased air flow, but none of these differences were statistically significant. Biological monitoring for cadmium, chromium, and fluoride values in urine samples from both reservations were within normally defined values, and no differences between the two reservations were found (1). ATSDR is also currently investigating the potential for human exposures (past, present, future) to groundwater, surface water and sediment, surface soil, and biota in relation to the EMF site.

Goals and Purpose of Study

The major limitation of the previous ATSDR health study of the residents of Fort Hall was the uncertainty in assigning exposure levels to contaminants emanating from the two phosphate-processing plants (1). In addition, the study recognized that most of the highest exposures to air contaminants may have occurred in the past and that the study methodology could not identify historically exposed persons (1). The current ATSDR exposure assessment will attempt to determine a population that is currently and historically exposed to air emissions (particularly PM) from the two phosphate-processing plants and other potential sources. Using the results of ATSDR's exposure assessment, the University of North Carolina at Chapel Hill (UNC) School of Public Health will conduct an ecologic health study of respiratory and cardiopulmonary mortality in areas where persons have been exposed to PM at the level of health concern.

Study Area and Site Background

The EMF NPL site is made up of the FMC Elemental Phosphorous Plant and the Simplot Don Plant. The nearest major population areas, Pocatello and Chubbuck, Idaho, are located east-southeast and east-northeast, respectively, of the FMC/Simplot plants (Figure 1). The facilities are about 2.5 miles from Pocatello. The FMC plant is located on Fort Hall Reservation land (in the southern part of the reservation) and the Simplot plant is on state land. The Town of Fort Hall is located about 8 miles north-northeast of the facilities.

The FMC plant covers an estimated 1,189 acres and adjoins the western boundary of the Simplot plant (2). Elemental phosphorus production at the facility has changed little since the plant operations began in 1949. Phosphorus-bearing shale is shipped to FMC via the Union Pacific Railroad during the summer months. Ore cannot be shipped in the winter because it would freeze in the rail cars; therefore, the ore is stockpiled at the facility during the winter months. Ore from the stockpiles is processed in four electric arc furnaces. The furnaces' reaction yields gaseous elemental phosphorus and byproducts. Some of the byproducts contain radioactive components. The elemental phosphorus is subsequently condensed to a liquid state and eventually shipped off site. About 1.5 million tons of ore are processed at the plant each year. The disposal of

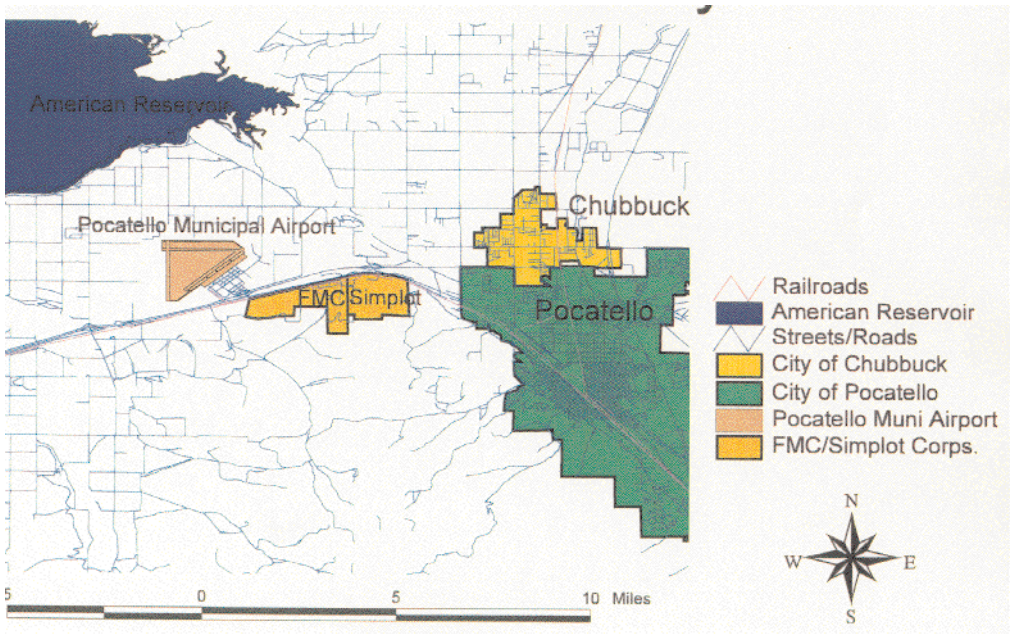


Figure 1 ATSDR/UNC air study area, Pocatello, Idaho.

byproduct waste material at and around the facility has produced slag piles that cover large areas of land (2).

The Simplot plant covers about 745 acres and adjoins the eastern property boundary of the FMC facility (2). The plant began production of single superphosphate fertilizer in 1944. In 1954, the facility began producing phosphoric acid. The phosphoric acid is now produced using an aqueous process. Formerly, phosphate ore was transported from the mines to the facility via rail. As of September 1991, the Simplot plant receives phosphate ore through a slurry pipeline. The phosphate ore slurry is processed at the Simplot plant in phosphoric acid reactors and then further processed into a variety of solid and liquid fertilizers. The plant produces 12 principal products, including five grades of solid fertilizers and four grades of liquid fertilizers (2).

Epidemiologic Studies of Particulate Matter Exposures

"Particulate matter" is the term used for a mixture of solid particles and liquid droplets found in the air. Coarse particles (larger than 2.5 micrometers [μm] in diameter) come from a variety of sources, including windblown dust and grinding operations. Fine particles (smaller than 2.5 μm) result from fuel combustion (from motor vehicles, power generation, and industrial facilities), residential fireplaces, and wood stoves. Before 1987, EPA's standards regulated larger particles (also know as total suspended particles [TSP]). By 1987, research had shown that the particles of greatest health concern were those 10 μm in diameter or smaller, which can penetrate into sensitive regions of the respiratory tract. At that time EPA and the states took action to monitor and regulate PM that was 10 μm and smaller (PM_{10}). In the years since the previous standard was

enacted, hundreds of studies have been published on the health effects of PM. These studies suggest that adverse health effects in children and other sensitive populations have been associated with exposure to PM concentrations well below that allowed by the 1987 PM_{10} standard (3).

Moreover, these studies have indicated that the fine particles ($PM_{2.5}$), which penetrate more deeply into the lungs, are more likely than coarse particles to contribute to adverse health effects. Some of the health effects associated with $PM_{2.5}$ exposures are (3):

- Premature death
- Respiratory-related hospital admissions and emergency room visits
- Aggravated asthma
- Acute respiratory symptoms, including aggravated coughing and difficult or painful breathing
- Chronic bronchitis
- Decreased lung function that can be experienced as shortness of breath
- Work and school absences

These studies indicate that the elderly, individuals with pre-existing heart or lung disease, children, and asthmatics are at the most risk for adverse health effects from exposure to $PM_{2.5}$. For these reasons, on July 17, 1997, EPA revised its PM standards to include a primary (health-based) annual average $PM_{2.5}$ standard of 15 micrograms per cubic meter ($\mu g/m^3$) and a 24-hour $PM_{2.5}$ standard of 65 $\mu g/m^3$ (3).

Study Area Topography and Meteorology

The local terrain in the Pocatello area is classified as meteorologically "complex." East of Pocatello, the Pocatello Mountain Range rises from about 4,400 feet to about 6,500 feet above mean sea level. Southeast of the FMC and Simplot facilities is the city of Pocatello, which lies in the funnel-shaped Portneuf Valley. The valley virtually closes at the southern boundary of the city of Pocatello. The northern end of the Bannock Range is immediately south of the FMC and Simplot facilities. This range tapers down to a north-pointing wedge shape just east of the Simplot facility and forms one side of the Simplot gypsum stacks. The ridge just southeast of Simplot rises from Simplot's base elevation of 4,449 feet to about 5,700 feet. The terrain south of the facilities (between them and the Bannock Range) gives way to the Michaud Flats of the Snake River drainage to the north, and to the Arbon Valley to the west. From the southwest, clockwise to north-northeast of the facilities, the terrain is generally flat for several miles (4).

Long-term meteorological data in the study area are primarily obtained from the National Weather Service (NWS) station at the Pocatello Airport, located west-northwest of the FMC and Simplot facilities (Figure 1). The wind rose (Figure 2) shows a marked preference for west to south winds with a prevailing wind direction from the southwest. A secondary preference for wind direction is indicated from the northeast. The lowest frequency of wind direction is from the east to south-southeast. Wind data from Simplot's meteorological station, located north of the Simplot facility, indicate a prevailing wind direction from the southwest to west-southwest, with a strong second predominant wind direction from the southeast to east-southeast. This secondary flow is clearly out of the Portneuf Valley and is a nighttime drainage wind flow (4).

Emissions from the phosphate plants and area topography both contribute to local

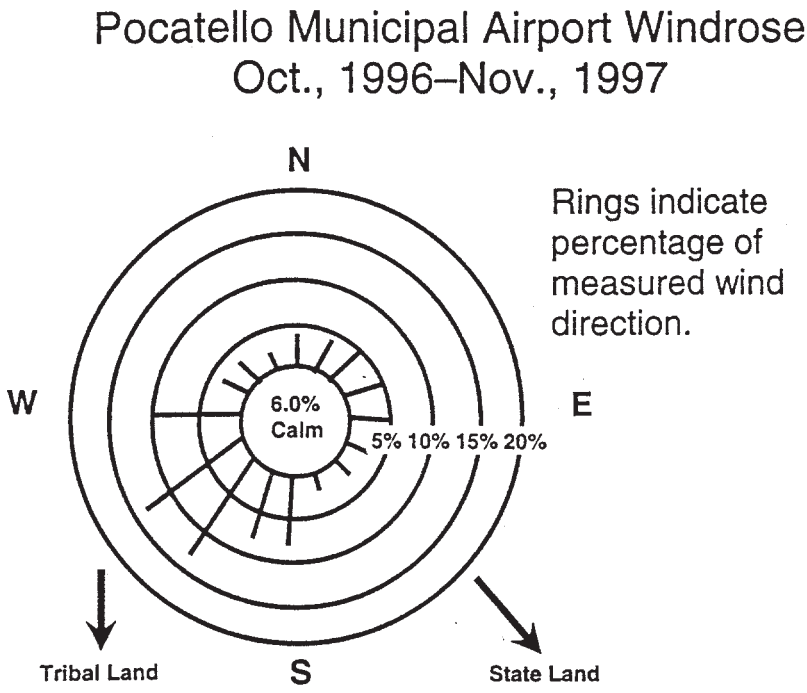


Figure 2 Pocatello Municipal Airport wind rose, October 1996 through November 1997. Source: Idaho Department of Environmental Quality.

air pollution. Particulates, phosphate pentoxide, metals, and radionuclides escape into the atmosphere from the stacks during production of phosphorus products. Fugitive dust from the waste ponds, ore stacks, and waste piles on the site is also a concern. The effect of these industrial emissions on air quality is compounded by the complex local topography and climatic conditions. Winds from the southwest sector carry pollutants from these plants toward population areas in northern Pocatello, Chubbuck, and the Fort Hall Reservation. In the study area, temperature inversions, caused by high-pressure subsidence and radiative cooling, can occur year-round. During these inversions, emissions from the industrial plants might become trapped and form a dense brown cloud about 1,000 feet above the ground, extending 4 to 5 miles in length and 32 miles in width, or the emissions might stay at ground level (1).

Area Monitoring Network and Historical PM Levels

The Idaho Department of Environmental Quality (IDEQ) began monitoring the air in the study area for TSP in 1975. Basing its decision on TSP data collected from 1975 to 1977, EPA designated the area as in nonattainment status (5). The original air monitoring network consisted of three stations: the sewage treatment plant (STP), the Chubbuck School (CS), and Idaho State University (ISU) (Figure 3). Monitoring data for the maximum 24-hour and average annual TSP concentrations are available from these three monitoring stations for 1977 through 1987. EPA's primary health-based average annual TSP standard at that time was 75 $\mu g/m^3$. During this 11-year period, this

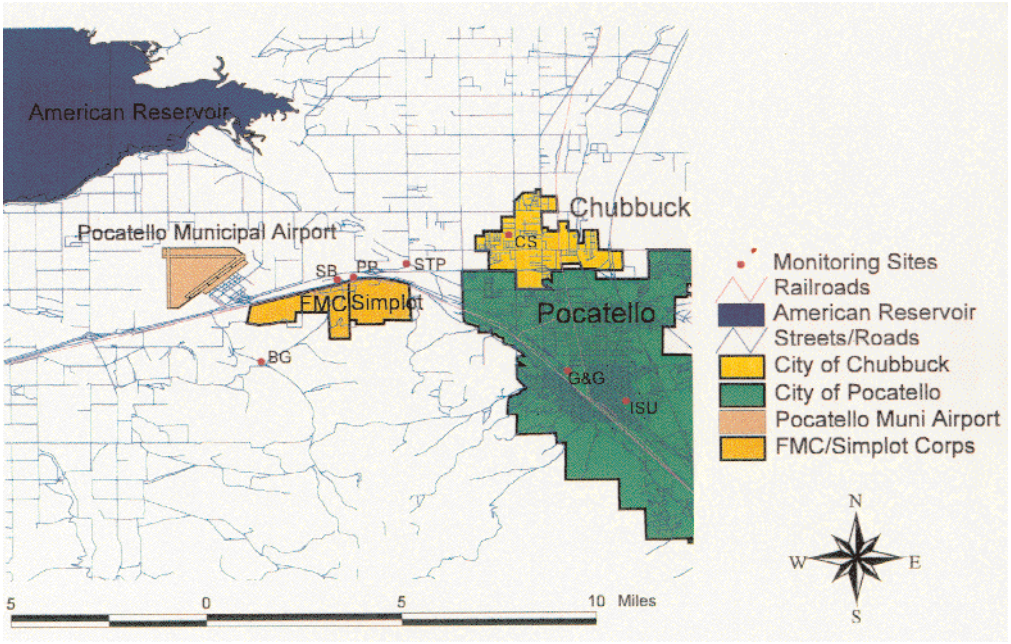


Figure 3 Particulate monitoring sites in study area.

standard was exceeded 10 times at the STP site, 4 times at the CS site, and 0 times at the ISU site. Average annual and 24-hour maximum PM_{10} monitoring has been ongoing at the STP site since 1986, at the CS and ISU sites since 1988, and at a new site (Garrett and Gould [G&G]) since 1990 (Figure 3). EPA's health-based annual average PM_{10} standard (established in 1987) is 50 $\mu g/m^3$. During the time period of PM_{10} monitoring, the standard has been violated 4 out of the 12 times at which samples were taken at STP, 0 out of 9 times at ISU, 0 out of 10 times at CS, and 0 out of 7 times at G&G. Since 1996, the Shoshone-Bannock Tribes and EPA have operated three air monitoring sites: two located just north of the FMC facility (Primary Particulate [PP] and Shoshone-Bannock [SB]) and one background site (BG) located about 4 kilometers west-southwest of the PP and SB monitoring sites (Figure 3) (6). Monitoring at these sites is primarily for PM_{10}; however, every third day, samples from a dichotomous sampler, located at the PP site, provide data for both the PM_{10} and the $PM_{2.5}$ fraction (6). Although average annual monitoring data are not currently available for the PP and SB sites, EPA's 24-hour PM_{10} standard of 150 $\mu g/m^3$ (established in 1987) has been exceeded at least 44 times between October 1996 and September 1997.

Included in EPA's 1997 revision of the PM standards were regulations that called for implementation of a monitoring network for $PM_{2.5}$. The IDEQ is currently in the process of implementing this monitoring network within the study area. Except for the more recent $PM_{2.5}$ data available from the PP monitoring site and several other seasonal studies that included monitoring for $PM_{2.5}$, historical $PM_{2.5}$ data are very sparse. PM_{10} monitoring has also not produced a complete database dating back to 1977. To obtain a better understanding of the historical $PM_{2.5}$ and PM_{10} levels in the study area, several site-specific ratios were used to estimate these levels where monitoring data were not

available. Long-term data indicate an average PM_{10}/TSP ratio of about 0.5 (7). Three different studies have calculated ratios of $PM_{2.5}$ to PM_{10} of 0.5, 0.66, and 0.76 (8). Although the 1978–1979 study ($PM_{2.5}/PM_{10}$=0.76) was considered the best study with respect to the number of samples taken (9), for the sake of this evaluation, the middle value of 0.66 was chosen. Based on this ratio, actual and estimated historical values for PM_{10} and $PM_{2.5}$ were calculated and plotted (Figures 4 and 5). From these data, it can readily be observed that PM levels at these monitoring sites have dramatically declined since 1992. It is not known whether this decline is due to Simplot's switch to a wet process or due to other measures to reduce other sources of PM in the study area, or due to both. From a public health standpoint, these data are also very illustrative. The data indicate that, since 1977, the only monitoring station at which the PM_{10} standard has been exceeded has been the STP site, which is located in a relatively sparsely populated area. Because the CS, ISU, and G&G sites are all located in more densely populated areas, one could conclude that the ambient air levels of PM, based on PM_{10} levels alone, do not indicate a large public health impact. As previously indicated, however, studies have shown that $PM_{2.5}$ is the more important PM fraction from a public health standpoint. As shown in Figure 5, since 1977, the health-based $PM_{2.5}$ standard of 15 $\mu g/m^3$ may have been exceeded frequently at all three monitoring stations located in populated areas. These data provide evidence for public health concern that persons in more densely populated areas have been exposed to $PM_{2.5}$ at levels that may result in adverse health effects.

Design of Air Study

ATSDR is currently designing an exposure assessment methodology that will not only address community concerns regarding past and present exposures to PM and other

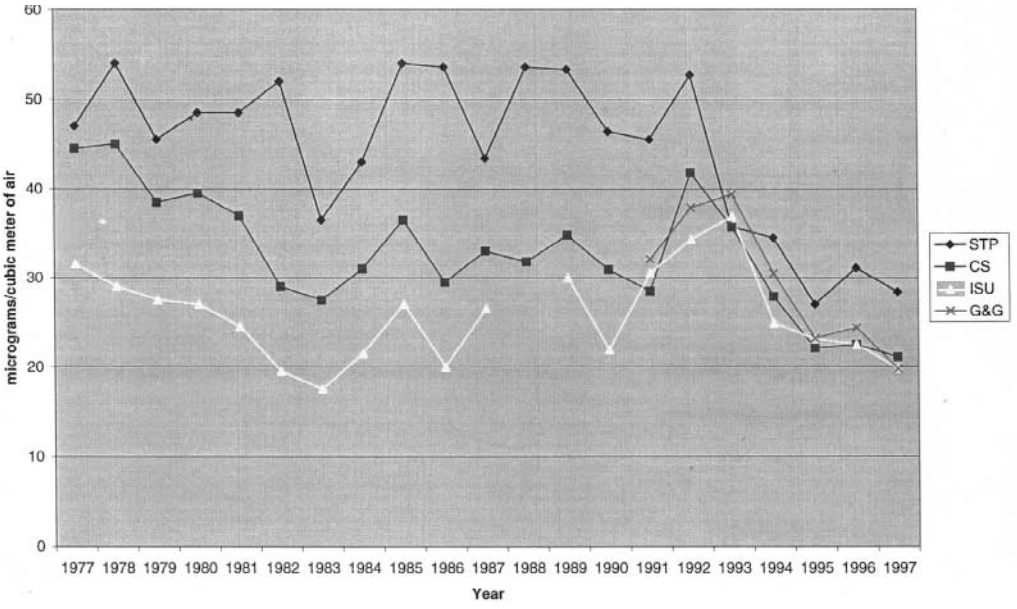


Figure 4 Estimated and actual annual average PM_{10} concentrations in study area.

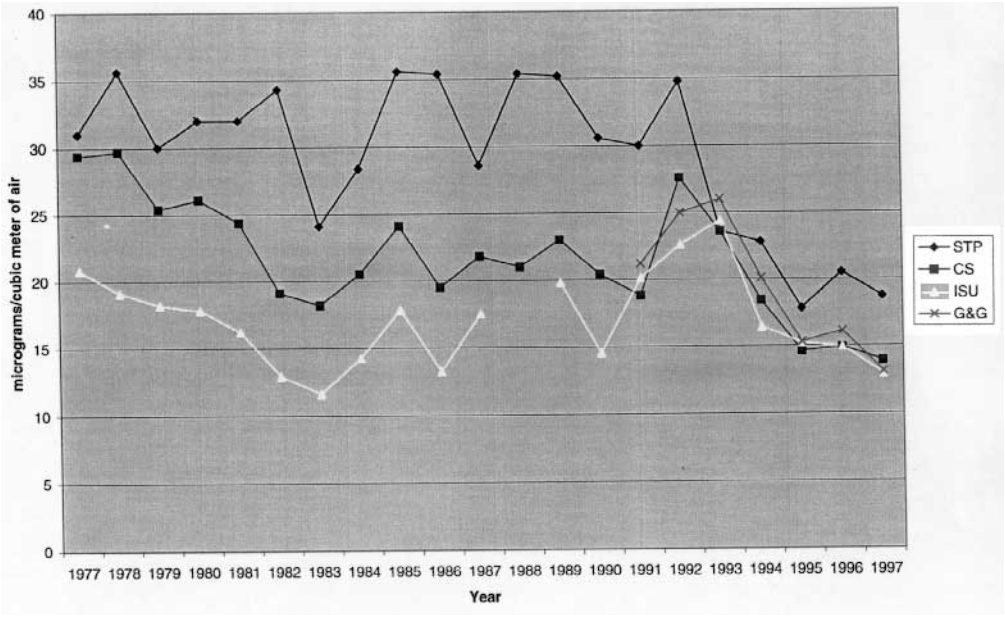


Figure 5 Estimated annual average $PM_{2.5}$ concentrations in study area.

contaminants emanating from FMC/Simplot facilitates, but will also be scientifically defensible. The second phase of the planned study, an ecologic health study of respiratory and cardiopulmonary mortality, to be conducted by the UNC's School of Public Health, is also under design. The basic designs of the exposure assessment and ecologic health study have already been conceptualized. Some of the methodological issues that have been and will be evaluated to determine the final design approach are discussed in the next section.

A geographic information system (GIS) will be a main feature of the design of the studies because (1) a GIS environment is an excellent platform for bringing together disparate databases, (2) a GIS can be used to manipulate data to uncover the underlying spatial associations between various data layers (making it a "value-added" product), and (3) the "value-added" data produced within a GIS environment can be linked with statistical packages outside a GIS.

The first step in the basic design of the exposure assessment will be to use an air dispersion model, like EPA's Industrial Source Complex Model, to determine concentration contours within which persons have been exposed to levels of $PM_{2.5}$ above the health-based standard of 15 $\mu g/m^3$. The results of the air dispersion model will be imported into a GIS and will then be edited to produce concentration polygons. Within the GIS, an analysis will be performed to overlay these $PM_{2.5}$ concentration polygons onto the TIGER/Line files integrated with the 1990 or 1980 US Census Bureau summary tape data. This overlay analysis will "clip" out the demographic information of persons who have been exposed to $PM_{2.5}$ above 15 $\mu g/m^3$, as predicted by air dispersion modeling. Demographic information about the total population exposed, total susceptible populations exposed (e.g., persons over 65 years old and children under 17 years old), and

socioeconomic status of persons exposed will be obtained. This analysis will be performed in five-year periods beginning with 1977 and ending with 1996.

Mortality data for respiratory and cardiopulmonary deaths (*International Classification of Diseases, Ninth Revision* [ICD-9] [10], codes 400–440 and 485–496) along with lung cancer (ICD-9 code 162), will be obtained from the Idaho State Department of Health for the counties that encompass the study area (Power and Bannock Counties) for the years 1977 through 1996. The mortality data will be grouped by the same five-year periods used in the exposure assessment. The addresses for these mortalities will be matched and geocoded to street addresses from the TIGER/Line files. A point-in-polygon analysis will then be performed to determine if the geocoded addresses for the cardiopulmonary and lung cancer deaths are located within or outside each of the polygons defining a geographic area where persons have been exposed at levels of health concern. Data on the total number of mortality cases and the total exposed population for each of the five-year periods will be used to calculate the crude mortality rates. These rates and the aggregated age, sex, and socioeconomic status data for the exposed areas will be used to control for ecologic confounding. The rates will be compared with the rates for the state of Idaho. The analytical technique to be used in the health study to control for ecologic confounding is still being evaluated (see the discussion below).

ATSDR Exposure Assessment Methodological Issues

The first basic issue in designing any exposure assessment that uses GIS is the choice of the analytic method to use to define the areal extent of exposure and the advantages and limitations of available methods. As indicated above, ATSDR has chosen air dispersion modeling to define the areal extent of exposure to $PM_{2.5}$.

There are other impediments and potential data fallacies that need to be addressed before designing any study using GIS or interpreting any result of a GIS analysis (11). Three of the major impediments to the use of GIS within ATSDR's exposure assessment are the areal interpolation problem, the fallacy of the homogeneous polygon, and the fallacy related to fuzzy boundaries (11). The areal interpolation problem arises when data obtained from one reporting unit must be combined with data from a different reporting unit; questions arise about the correct way to interpolate the overlays of these units as well as about what assumptions are to be made. The fallacy of the homogeneous polygon arises when it is assumed that a polygon delineates an area as homogeneous when indeed the phenomenon being represented by the polygon is not evenly distributed across the area. The other fallacy, related to fuzzy boundaries, occurs when it is assumed that the boundary between two polygons is discrete when it actually represents some sort of gradient (11). These methodological issues, as they relate to ATSDR's exposure assessment, are discussed below.

Exposure and risk assessment in relation to environment and health is essentially an attempt to estimate the level of exposure to specified pollutants, either for individuals or particular population groups. Direct measurements of exposure are rare, and usually only exist for a relatively small sample of people. Instead, levels of exposure commonly have to be estimated by indirect means. Two main approaches available are spatial interpolation from measurements of ambient pollution levels and modeling based on data on emission levels and sources (12). Air dispersion models are usually based on Gaussian plume dispersion equations and take into account emission source

and meteorological and terrain effects. The use of dispersion modeling to assess exposure has a number of advantages. These include the fact that dispersion modeling does not rely on measured concentrations, meaning that it can be extended to areas and pollutants for which no monitored data are available. Dispersion modeling also allows for knowledge about dispersion processes and can take into account the pollution surface and local factors that influence these processes (e.g., terrain, weather) (12).

Dispersion models suffer from three major limitations. The first is the demand for data inputs to the models, which require various large databases that can be of questionable quality. The second limitation is that these models only work for areas relatively close to the emission source, where Gaussian dispersion processes can be assumed. The final limitation is that these models, for the most part, are designed to work under simple conditions; that is, they operate best with limited sources of emissions and under relatively simple terrain and weather condition (12,13). Very few attempts at linking air dispersion models within a GIS are available in the literature (12). However, there are a few examples in the published (14,15) and unpublished (16) literature that have incorporated the results of air or (more often) groundwater modeling with GIS in order to perform a risk or exposure assessment. Many of these examples are from applications of GIS to help local, state, and federal agencies evaluate risk from or exposure to contaminants (17).

As indicated above, another method available for evaluating exposure is spatial interpolation—a way to estimate pollutant levels at unsampled sites. In this type of analysis, GIS can provide a range of interpolation techniques that can be used to produce a concentration surface. However, spatial interpolation is not without its limitations and considerations. The first consideration is the method of interpolation used (e.g., inverse distance weighting, kriging, spline). The performance of different interpolation methods depends upon a number of factors, including the nature of the underlying spatial variation in the phenomenon under consideration and the sample density and distribution (12).

In exposure and risk assessment, it is vitally important to assess the link between the environment and health, especially if the results of the analysis will be used for further epidemiologic studies. Misclassification of exposure and misclassification of disease play a major role in the degree of confidence one has in the results of environmental epidemiologic investigations (18,19). Moreover, in contrast to the relatively well-defined exposure characterization variables in an occupational setting, some of the variables in environmental settings are not well defined or are not defined at all. Understanding exposures in the residential setting is even more complicated. These exposures are strongly influenced by seasonal or even daily lifestyle preferences, travel and excursion habits, and indoor/outdoor concentration differences, as well as complexities involved in the estimation of exposures in general (13). The most common approach to determining the populations at risk (exposed) has been to assume that people living near the point source are at greater risk than those who live farther away. Various summary articles (18,19) contain many examples that illustrate the use of GIS to construct a buffer around a point, line, or area source. Buffers are, however, inevitably crude indicators of risk or exposure, and without an understanding of the dispersion processes and pathways involved, it is easy to use buffers of an inappropriate size and/or shape. GIS clearly allows for modeling (either within the GIS or imported from

an outside model) of more complex search areas, taking into account dispersion patterns and other effects, where data permit (12).

The above discussion clearly provides a strong argument for the use of air dispersion modeling as the preferred approach for the ATSDR exposure assessment. The use of modeling to define an exposed population can help overcome some of the impediments and data fallacies mentioned above. For example, if a radial buffer were used in the analysis, the defined zone would represent an area assumed to have homogenous exposure; however, in reality, a buffer-defined area could represent very low to very high exposures. Moreover, air dispersion modeling versus the use of buffers allows a researcher to define polygons that represent various exposure gradients, thus alleviating the problem of fuzzy boundaries. However, as previously indicated, Gaussian plume models must be used with care. The validity of using air dispersion modeling for this exposure assessment could be questioned because the FMC and Simplot sites contain well over 100 different point, line, and area PM sources, the topographic and meteorological conditions of the study area are not simple, and most of the exposed population is not near the primary sources of PM emissions. Many of these drawbacks can be overcome by incorporating corrective equations that calibrate the model with available monitoring data. For this reason, historical exposures before 1977, when air monitoring began in the study area, will not modeled.

The problems of areal interpolation and the fallacy of the homogeneous polygon must also be considered carefully when evaluating the method used to determine the demographics of the exposed population defined by the air dispersion model. The polygons that define the various exposure levels predicted by the air dispersion model will not correspond to the US Census Bureau's reporting units (e.g., block groups). Furthermore, the populations within the census units are not evenly distributed. Therefore, an overlay analysis method that does not provide some estimate of the population densities within each census unit will likely produce much exposure misclassification. ATSDR uses an area proportion program (a script written in Avenue, the programming language of ArcView GIS [ESRI, Redlands, CA]) that is easy to use and is good for many applications; however, it assumes that a population within a given census reporting unit is evenly distributed. This method may provide reasonable estimates of an exposed population in a completely urban setting; however, for this study, it is likely that the exposed population resides in urban, suburban, and rural areas. Because the results of the exposure assessment will be used as the basis for an epidemiologic study of the population exposed at levels of health concern, it is of vital importance that an accurate estimate of the "truly" exposed population be obtained. As previously mentioned, misclassification of exposure or disease in environmental epidemiologic investigations is a primary source of error. For these reasons, other methods are being evaluated that provide better estimates of population densities within the census reporting units. The two methods currently being evaluated are the kernel density method (20) and the census control population method (21). Both of these methods use techniques that "disaggregate" the census reporting units, helping to alleviate the areal interpolation problem and helping to avoid the fallacy of the homogeneous polygon.

UNC Ecologic Health Study Methodological Issues

Ecologic studies have been featured prominently in environmental epidemiology

because exposures are often already measured at the group level or because the limited resources of some studies prohibit collection of individual-level data (22). Because of the various methods that can be used within and outside a GIS environment to define an exposed population, the choice of an ecologic study design is common when using GIS in the analysis. Researchers using such a study design with GIS must address unique methodological issues beyond those they might encounter using other epidemiologic study designs (e.g., cohort or case-control studies). First, researchers must be careful in interpreting and conveying the results of an ecologic health study. Although ecologic studies can provide valid information on associations of exposure and outcome as related to a defined exposed group, they do not provide reliable information on individual-level risk. That is, ecologic bias (fallacy) can arise when inferences are drawn about associations at the individual level based on analyses conducted at the group level (22). Moreover, in addition to the usual sources of bias that threaten individual-level analyses, using ecologic analyses to estimate biological effects has an underlying problem: reflecting the heterogeneity of exposure level and covariate levels within groups. This heterogeneity is not fully captured with ecologic data because of missing information on the joint distribution of exposure, disease, and covariates (22). Ecologic fallacy can be easily avoided by not making any assertions regarding individual risks from the results of an ecologic study. The UNC study follows this admonition in that it attempts to determine the association between PM exposures to the community (a geographically defined exposed group) and cardiopulmonary mortality rates as compared with statewide rates for Idaho (another geographically defined group). The only data that will be collected on an individual level will be the mortality data.

The quality of an exposure assessment will determine the validity of an environmental epidemiology study. Furthermore, errors in measurement of exposure can introduce both bias and imprecision into the estimates of health effects (22). The methodological issues related to the analysis techniques used in the ATSDR exposure assessment have been discussed above. Disease misclassifications can also be a major source of bias within any environmental epidemiologic study. The use of GIS in an analysis does add methodological considerations related to disease misclassification beyond those that can be found in other epidemiology studies (e.g., increased disease diagnosis and systematic over-reporting of a disease). For example, within the current air study design, inaccurate address-matching or low address-matching rates, using a GIS or other software package, can provide incorrect or missing classification of a disease case during the point-in-polygon analysis. The conditions for confounding differ for individual-level versus group-level or ecologic analyses, and some types of confounding cannot be controlled in ecologic analyses (22). Even when all variables are accurately measured for all groups, adjustment for extraneous risk factors may not reduce the ecologic bias produced by these risk factors. In fact, it is possible for such ecologic adjustment to increase bias (22). There are two methods currently being evaluated to control for ecologic confounders.

Conclusions

It is apparent that GIS is an excellent platform for bringing together disparate databases and, through spatial analysis, assessing the exposure and demographics of an area.

It must not be assumed, however, that the results of a sophisticated analysis or well-laid-out map actually are valid representations of the world they are trying to model.

Methods of estimating areas of health concern in exposure assessments (e.g., Gaussian air dispersion models), either outside or within a GIS environment, must be used with an understanding of their limitations and the site-specific factors that may affect the validity of their results.

The use of air dispersion modeling in a study of air exposures can help to alleviate some of the data issues related to the fallacy of the homogeneous polygon and the problem of fuzzy boundaries.

Methods of estimating the demographics captured by modeling techniques that define an area of health concern must be used with care so that the problem of areal interpolation and the fallacy of the homogeneous polygon can be alleviated. Analytical techniques like the area proportion method are excellent for some applications; however, if exposure assessment information is to be used in an epidemiologic study of an exposed population, the distribution of population densities within demographic geographic units must be evaluated before one can feel confident that exposure misclassification has been reduced to acceptable levels.

An ecologic study design based on a GIS analysis carries with it unique methodological issues beyond those that may be encountered in other epidemiologic designs. Ecologic fallacy, disease and exposure misclassification, and control for confounding must be carefully considered when designing an ecologic study and in interpreting its results.

The design of ATSDR's exposure assessment and UNC's ecologic health study will be refined to estimate the association of PM exposures with cardiopulmonary mortality with as much validity and precision as an ecologic approach allows.

References

1. Agency for Toxic Substances and Disease Registry. 1995. *Fort Hall air emissions study—Fort Hall Indian Reservation, Fort Hall, Idaho.* Atlanta: US Department of Health and Human Services. November.
2. Bechtel Environmental, Inc. 1996. *Remedial investigation/feasibility study report for the EMF site.* San Francisco: Bechtel Environmental, Inc. August.
3. US Environmental Protection Agency, Office of Air and Radiation and Office of Air Quality Planning and Standards. Web site.
4. Bechtel Environmental, Inc. 1993. *Remedial investigation/feasibility study: 1993 air data interpretation report for the EMF site.* San Francisco: Bechtel Environmental, Inc. January.
5. Idaho Department of Environmental Quality. 1988. *Idaho air quality annual report—1987*. Boise: Idaho Department of Environmental Quality. July.
6. Air Resources Specialists, Inc. 1997. *Draft monitoring and quality assurance plan for the Shoshone-Bannock/EPA particulate monitoring program, Pocatello, Idaho.* Fort Collins, CO: Air Resources Specialists, Inc. January.
7. Personal communication. 1998. Steven Body, US Environmental Protection Agency, Region 10, Seattle. July 8.
8. IDEQ data report. No date.

9. Personal communication. 1998. Tom Edwards, Idaho Department of Environmental Quality, Pocatello, ID. July 8.
10. World Health Organization (WHO). 1977–78. *Manual of the international statistical classification of diseases, injuries, and causes of death: Based on the recommendations of the Ninth Revision Conference, 1975, and adopted by the Twenty-Ninth World Health Assembly.* Geneva: WHO.
11. National Environmental Health Association. Web site.
12. Briggs DJ, Elliott P. 1995. The use of geographical information systems in studies on environment and health. *World Health Statistics Quarterly* 48(2):85–94.
13. Esmen NA, Marsh GM. 1996. Applications and limitations of air dispersion modeling in environmental epidemiology. *Journal of Exposure Analysis and Environmental Epidemiology* 104(4):414–20.
14. Maslia ML, Aral MM, Williams RC, Susten AS, Heitgard JL. 1994. Exposure assessment of population using environmental modeling, demographic analysis, and GIS. *Water Resources Bulletin* 30(6):1025–41.
15. Von Braun M. 1993. The use of GIS in assessing exposure and remedial alternatives at Superfund sites. In: *Environmental modeling with GIS.* Ed. M Goodchild, B Parks, L Steyaert. New York: Oxford Press. 339–47.
16. Koontz MD, Zarus GM, Stunder MJ, Nagda NL. 1991. Air toxics risk assessment. Unpublished paper. Germantown, MD: GEOMET Technologies, Inc.
17. Croner CM, Sperling J, Broome FR. 1996. Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15:1961–77.
18. Vine MF, Degnan D, Hanchette C. 1997. Geographical information systems: Their use in environmental epidemiology. *Environmental Health Perspectives* 105(6):598–605.
19. Nuckols JR, Berry JK, Stallones L. 1994. Defining populations potentially exposed to chemical waste mixtures using computer-aided mapping and analysis. In: *Toxicology of chemical mixtures: Case studies, mechanisms, and novel approaches.* Ed. RSH Yang. Ann Arbor, MI: Academic Press. 473–503.
20. Page PH. 1996. *A variable kernel density estimation procedure for generating raster GIS population data layers from US Census data*. Master's degree thesis. Chapel Hill, NC: University of North Carolina at Chapel Hill, Department of Geography.
21. Knapp MB, Archambault GV. 1998. Census control population methodology. Unpublished paper. Connecticut Department of Public Health.
22. Rothman KJ, Greenland S. 1998. *Modern epidemiology.* 2d Ed. Philadelphia: Lippincott-Raven Press.

Prenatal Health Behaviors and Birth Outcomes

Jane E Warga (1),* Tracy Benzies-Styka (2), Matthew Stefanak (3), Kimberly Vaughn (4)
(1) Director, Health Education & Assessment, District Board of Health of Mahoning County, Youngstown, OH; (2) Community Health Education Specialist, District Board of Health of Mahoning County, Youngstown, OH; (3) Health Commissioner, District Board of Health of Mahoning County, Youngstown, OH; (4) GIS Administrator, Mahoning County Planning Commission, Youngstown, OH

Abstract

In the early 1990s, the District Board of Health of Mahoning County (Youngstown, Ohio) used pregnancy-related behaviors to assist the community in planning prenatal and birth outcome interventions. The first mappings of prenatal behaviors and birth outcomes were crude small-area-analysis maps by census tract and zip code, prepared by hand. Since 1994, the District Board of Health and the Mahoning County Planning Commission have prepared geographic information system (GIS) mapping of pregnancy behaviors and birth outcomes. These maps have assisted agencies and community collaboratives by highlighting high-risk census tracts indicating a need for public health interventions. With birth certificate data received from the Ohio Department of Health, frequency counts of key health indicators are computed. Prenatal and birth indicators include percentages of low birth weight infants, tobacco usage during pregnancy, trimester of entry into prenatal care, and number of births to teens ages 15 to 17. Census tract mappings of these indicators are made for the entire county, including all cities and villages.

Keywords: birth outcomes, birth weight, prenatal, smoking, teen births

Introduction

In the early 1990s, the Healthy Outcomes of Pregnancy Consortium for Mahoning County was established. The consortium goal was to reduce infant mortality in the community. Health behavioral indicators such as tobacco use during pregnancy, the trimester of entry into prenatal care, and the percentage of low birth weight infants were targeted for review and analysis. Small area analysis of these indicators by census tract as reported on birth certificate data was completed. Small area analysis maps by census tract and zip code were prepared by hand. Birth certificate information was gathered from lengthy printout sheets from the Ohio Department of Health, Vital Statistics Division. Various risk indicators were then hand-counted by census tract. The maps of the percentage or rate occurrence by census tract or zip code were hacked with pencil and ruler.

Methods

In 1996, the District Board of Health of Mahoning County and the Mahoning County Planning Commission combined expertise to create geographic information system (GIS) mappings of public health concerns in the county. On receipt of birth certificate

* Jane E Warga, District Board of Health of Mahoning County, 50 Westchester Dr., Youngstown, OH 44515 USA; (p) 330-270-2855 x131; (f) 330-270-0625; E-mail: Agraw@msn.com

data from the Ohio Department of Health, the Mahoning County Planning Commission assigns census tracts to all the records by first geocoding the home address of the child. Once the records are assigned census tracts, the District Board of Health analyzes them for specific public health indicators. The District Board of Health completes the birth certificate analysis using Statistical Analysis Software (SAS). Major prenatal and birth outcome indicators, such as percentages of low birth weight infants, tobacco usage during pregnancy, trimester of entry into prenatal care, and number of births to teens ages 15 to 17, are computed by census tracts. Census tract mappings of these indicators are then created through Atlas Mapping Software for the entire county, including cities and villages.

While the community's infant mortality rate and percentage of low birth weight infants have decreased, and the percentage of women entering prenatal care in their first trimester has increased, Mahoning County continues to have a high overall infant mortality rate and high percentage of low birth weight infants. This situation creates an ongoing need for analysis of prenatal and birth outcomes. The use of the more accurate GIS maps enables the community to better decide how to spend its dollars on the costly health behaviors of smoking, late or no prenatal care, and teen births.

Data

The percentage of women using tobacco during pregnancy for the entire county has decreased from 26% smokers in 1991 to 20% smokers in 1996. Figure 1 illustrates the disparity of smoking rates throughout the various census tracts in the county. Several census tracts have rates above 30%, while many are below 10%.

The fluctuation of the percentage of low birth weight infants (less than or equal to 2,500 grams) by census tract is visible in Figure 2. The percentage range of this map extends from 0% to greater than 20%. The maps of the various health behaviors and birth

Figure 1 Tobacco use during pregnancy (1996 birth certificate data).

Figure 2 Percentage of low birth weight infants (less than or equal to 2,500 grams) (1996 birth certificate data).

outcomes are illustrative of the disparity in risk factors and birth outcomes between census tracts.

An additional disparity between census tracts is apparent in Figure 3, which identifies the communities in need of an intervention to encourage entry into prenatal care as early as possible during pregnancy.

The GIS mapping of these public health indicators helps the community to target census tracts for public health interventions. Targeting interventions this way can save time, money, and other valuable resources. An initiative of the Mahoning County

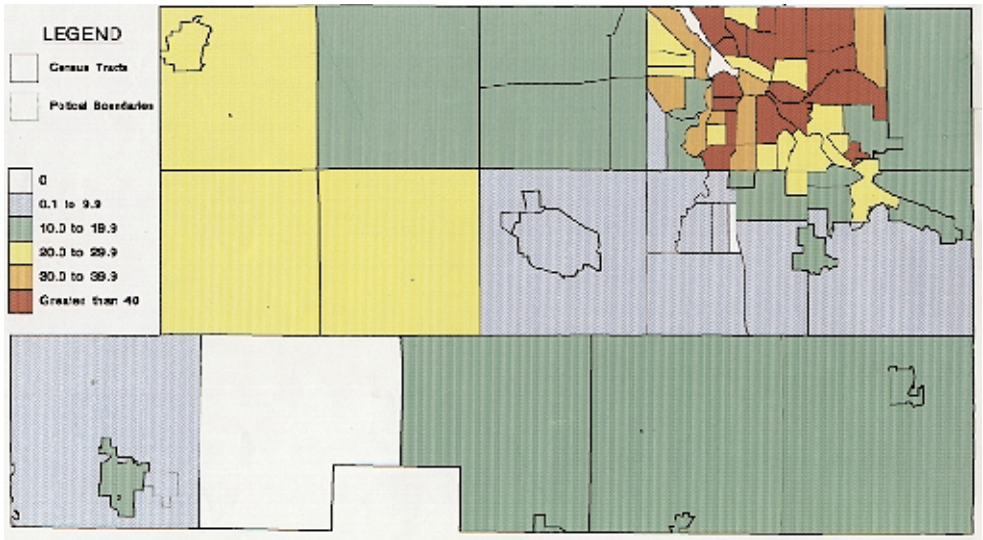


Figure 3 Percentage of mothers receiving late or no prenatal care (1996 birth certificate data).

Family First Council's Wellness Block Grant, Subcommittee for Healthy Children—the "Reducing Teen Pregnancy by Building Assets" project—is one example of the usefulness of a GIS map. The project used the mapping of 1995 teen pregnancy rates to target students grades, 5 through 9, in census tracts with inordinately high teen birth rates for after school programming and formal education in the corresponding school system.

The evaluation built into the grant includes the census tract mapping of teen pregnancy rates in subsequent years (Figures 4a, 4b). Targeted and untargeted census tracts will be analyzed for changes in teen birth rates.

Figure 4a Rate of births to school age teens, 15 to 17 years of age (1995 birth certificate data).

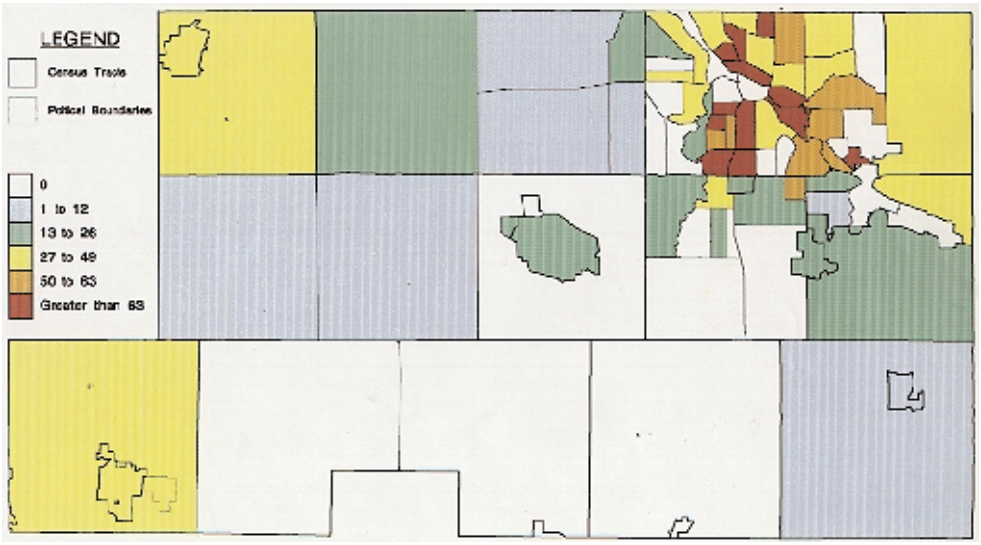


Figure 4b Rate of births to school age teens, 15 to 17 years of age (1996 birth certificate data).

Besides using the data on prenatal behaviors and birth outcome to help with decisions on targeting high-risk communities, the District Board of Health also uses the data to prepare a "Health Report Card" for all villages, cities, and townships within its jurisdiction. The report card includes prenatal health behavior and birth outcome rates for the particular census tract or tracts of villages, cities, or townships. This information affords an opportunity to compare a particular area with the total county, the complete health district, and/or the major city of Youngstown. It also acts as a catalyst for community health-related interventions in schools, churches, and community organizations.

The District Board of Health and the Mahoning County Planning Commission cooperate to produce the following GIS mapping projects, as well: location of landfills; prenatal patient distribution; physician participation in our children's health insurance program; well child patient distribution; rabid raccoon reports; animal bites; tuberculosis cases; and maps for the lead-based paint hazard control program.

Conclusion

The cost involved with developing a good analysis and mapping program includes computers, training, and software. The District Board of Health costs include the yearly SAS contract, the monthly birth certificate data disks, computers able to handle the software, staff training and time, the billing by the Planning Commission for the census coding of the birth data, and the preparation of the maps. The Mahoning County Planning Commission costs include computers, staff training and time, and the Atlas GIS Software.

In summary, GIS mapping of pregnancy behaviors and birth outcomes in Mahoning County has been successful in assisting agencies and community collaboratives to visualize high-risk census tracts needing public health intervention. By illustrating areas of various health-related risk behaviors, the maps have enabled these agencies to prepare specific interventions that meet their communities' needs.

Implementation and Operations

Exploring the Demographic and Socioeconomic Determinants of Health along the US-Mexico Border: An Online Interactive Application

Deborah L Balk (1),* Meredith L Golden (1), Maria Iwaniec (2)
(1) Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY; (2) Saginaw County, Saginaw, MI

Abstract

This paper demonstrates the usefulness of an interactive online geographic information system (GIS) tool—a demographic data viewer—used for the exploration of the determinants of health along the United States-Mexico border region. This tool facilitates access to and analysis of data for users with no GIS experience. The application, DDViewer 3.0, is an innovative Java-based interactive mapping application, freely accessible through the World Wide Web. It integrates many types of demographic and socioeconomic data and visualizes interdisciplinary spatial data online in real time. Users may create customized maps and undertake simple statistical analysis. Thus, DDViewer can play an important role in the dialogue between many different stakeholders in a public health system whose issues span large, heterogeneous, bi-national communities the way they do along the US-Mexico border.

Keywords: social demography, socioeconomic characteristics, health, US-Mexico border, data visualization

Introduction

When public health issues span large, heterogeneous communities as they do along the United States-Mexico border, policymakers and researchers need analytical tools that are neutral, simple, and versatile. This paper showcases the Demographic Data Viewer (DDViewer), an interactive online tool for exploring mortality and its socioeconomic and demographic determinants. DDViewer allows users with no geographic information system (GIS) experience to create customized maps and generate descriptive statistics for any region of interest within the United States.

DDViewer 3.0

DDViewer is an interactive mapping tool that currently makes available 225 demographic and socioeconomic variables from the 1990 US Census at the state, county, and tract level (1). The database is being extended to include US birth and death statistics at the county and state levels, Mexican socioeconomic and demographic data, and birth and death statistics for Mexico. At the time that this paper was written, only demographic and socioeconomic data from the counties on the US side were prepared for use. Future online applications will encompass data for the counties on both sides of the US-Mexico border.

* Deborah L Balk, CIESIN, Columbia University, PO Box 1000, 61 Route 9W, Palisades, NY 10964 USA; (p) 914-365-8965; (f) 914-365-8922; E-mail: dbalk@ciesin.org

DDViewer 3.0 is an innovative Java-based interactive mapping application, accessible through the World Wide Web. At present, it is an excellent data exploration tool. In the future, it will integrate additional analytic capabilities. It was developed by the Center for International Earth Science Information Network (CIESIN) at Columbia University in 1996. In 1997, a version using Java technology was released. Both Java and non-Java versions are available to meet the computer requirements of users with varying technological capabilities. Access to DDViewer is available through CIESIN's homepage (www.ciesin.org) or directly at its own universal resource locator (URL), http://plue.sedac.ciesin.org/plue/ddviewer/.

Applying DDViewer

DDViewer visualizes interdisciplinary spatial data online in real time. Using a map and listbox graphical user interface, users may select regions and socioeconomic and demographic variables of interest. Figure 1 shows that the states along the US-Mexico border can be selected by clicking on each state. Alternatively, an entire area may be chosen via the listbox. This feature is very flexible. Regions need not be contiguous and can be added or deleted easily. Once a state has been selected, lower levels—counties or census tract—may be selected.

The next step is to select variables. There are six categories from which to choose: population, income, education, employment, housing, and a miscellaneous category. Figure 2 shows a sample of the variables in the population category. Once variables are

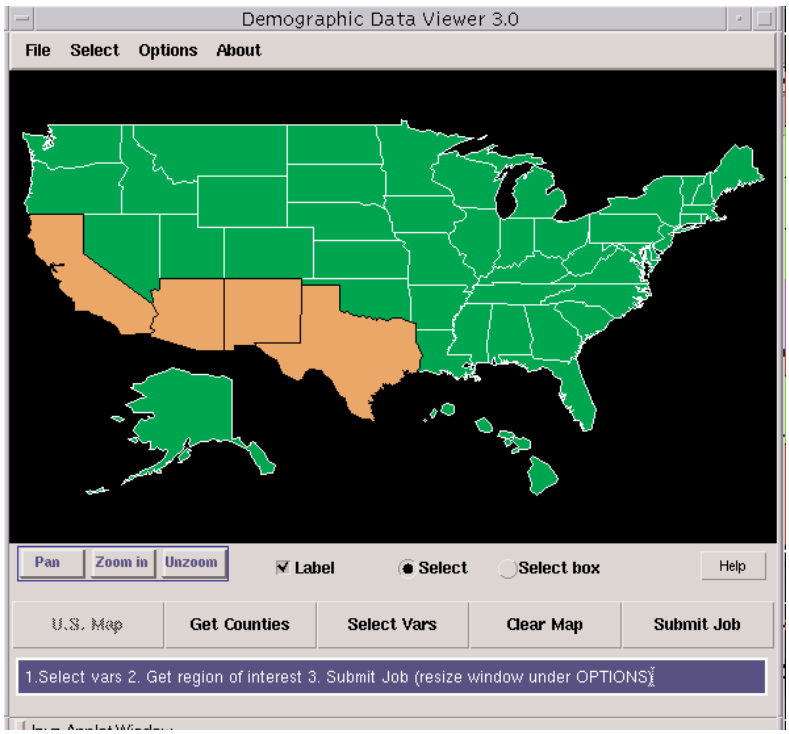


Figure 1 DDViewer display of selected states along the US-Mexico border.

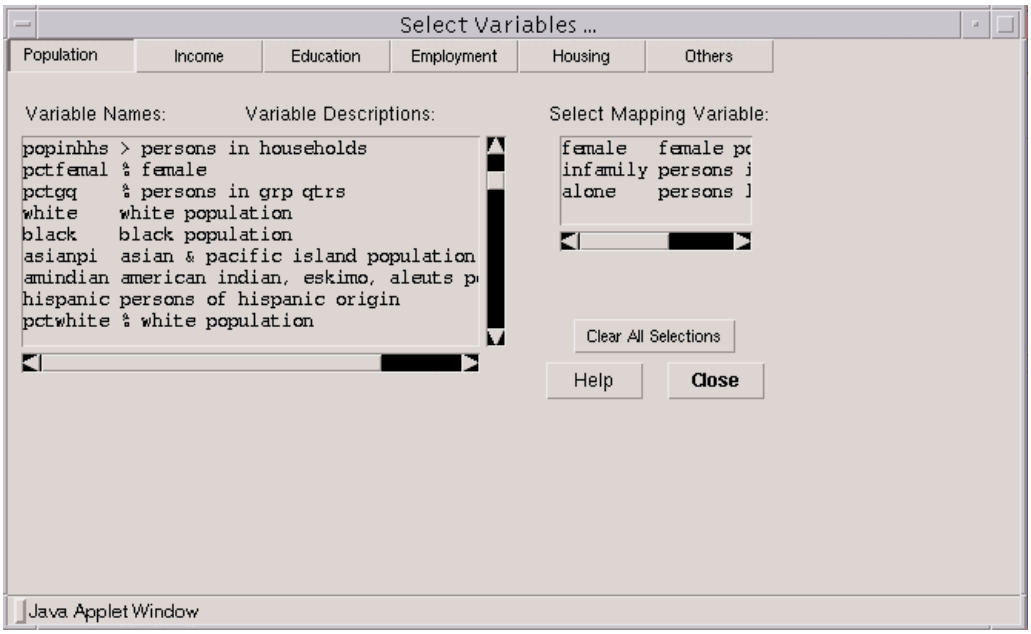


Figure 2 Sample of variables in the DDViewer population category.

selected, it is possible to construct new variables as a function of the original ones. Of course, the time it takes to process the selection depends on the size of the geographic area and number of units of interest, as well as the number of attributes. However, it is typically quick. To produce a county-level map of the distribution of the Hispanic population of the United States, for example, takes less than five minutes. Once the boundary and attribute data have been selected, all data processing takes place within the user's desktop client, and processing tends to be quite rapid. Each time a new region or additional variable is selected, CIESIN's server again transmits data directly to the user's computer.

Users can manipulate and customize their maps. For example, they may add titles and legends, alter the colors, zoom in on smaller regions, and select the manner in which the variables are displayed. To illustrate these capabilities, this paper includes a series of county-level maps generated by DDViewer showing the distribution of the Hispanic population of the United States. This variable was chosen because theory suggests that mortality and other health outcomes are dependent on access to health care, and that access to health care depends in part on the ethnic composition of an area (2).

Results

The first map, Figure 3, displays the distribution of the Hispanic population as quintiles (five categories that each contain one-fifth of the total population), though quartiles (four even categories) are the standard display unit of DDViewer and of the breakdown in which the descriptive statistics are given (see Figure 4). At the 25th percentile (i.e., one-quarter of the counties in the United States), the county-level population is 0.4% Hispanic. Even at the 75th percentile, only 2.4% of the county population is Hispanic.

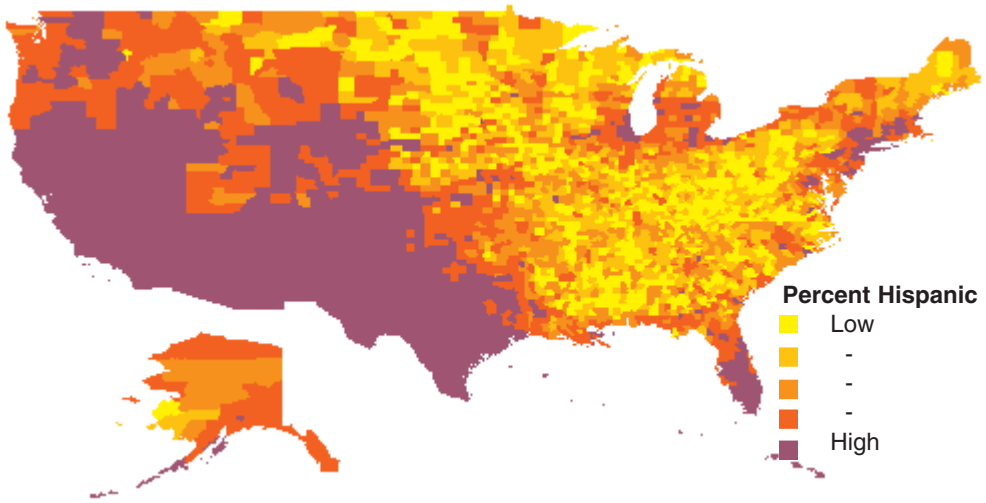


Figure 3 County-level distribution of US Hispanic population, 1990. Percentages displayed as quintiles.

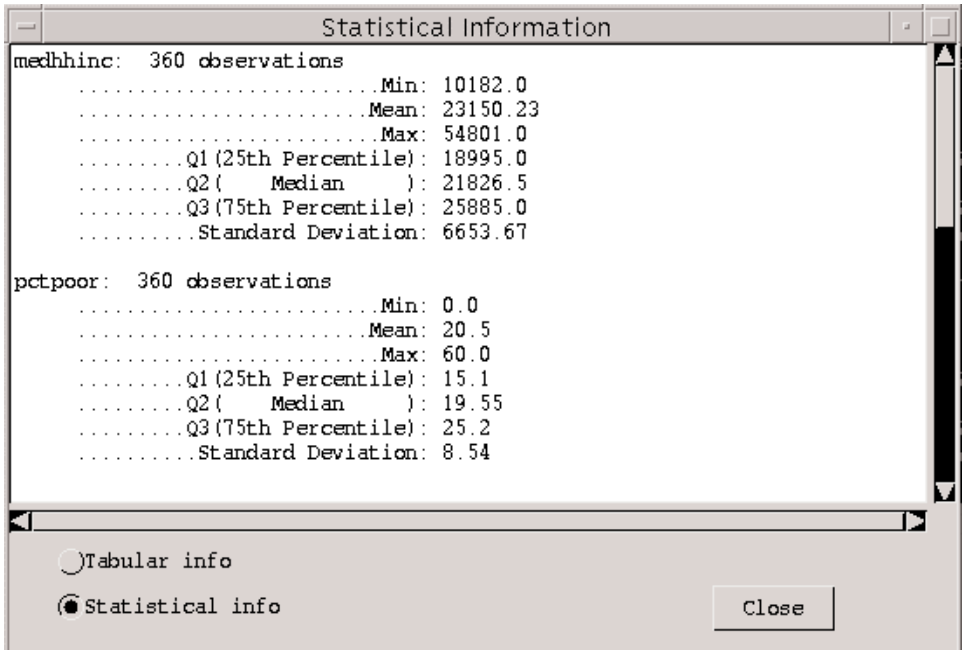


Figure 4 Descriptive statistics of quartiles.

The quintile breakdowns shown in Figure 3 would not be substantially different from those drawn for quartiles.

Although the color schema of the map in Figure 3 is accurate in indicating which areas of the United States have relatively small or large Hispanic populations, it does not describe the absolute distribution. The next map, Figure 5, shows the proportion of

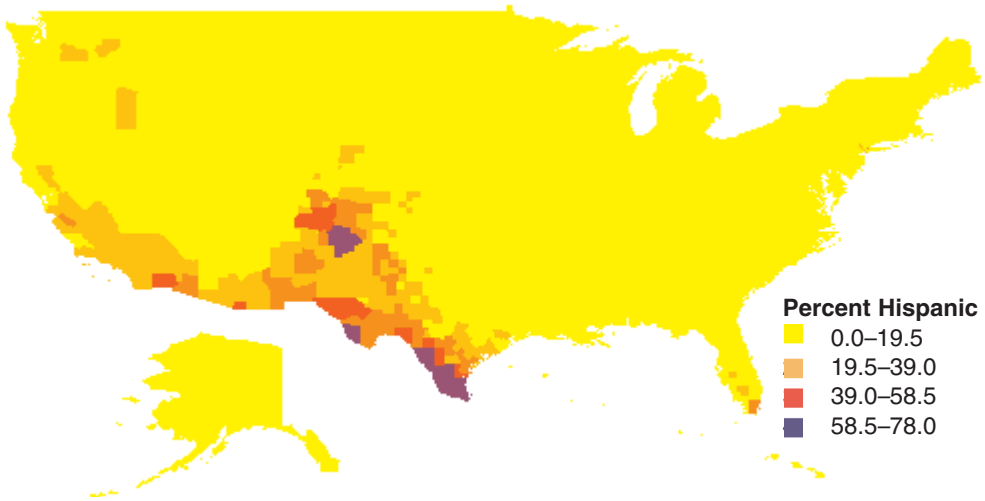


Figure 5 US Hispanic population, county level, 1990. Distribution displayed as five equal intervals.

Hispanic population to total population for each county in terms of five even intervals of the values of the distribution. It indicates that the vast majority of US counties have Hispanic populations of less than 20%. A few, mostly in Texas along the US-Mexico border, have Hispanic populations of 79% or greater.

Figure 6 plots specific proportions of the US Hispanic population based on customized distribution intervals. For example, the provision of Spanish language health services might be based on some critical proportion of the total population being Hispanic. Health planners could decide on the threshold level and then categorize areas in terms of having or not having a sufficient Hispanic population to warrant this service. The red in Figure 6 highlights areas where the Hispanic population is between 20 and 50%. The purple identifies counties with majority Hispanic populations.

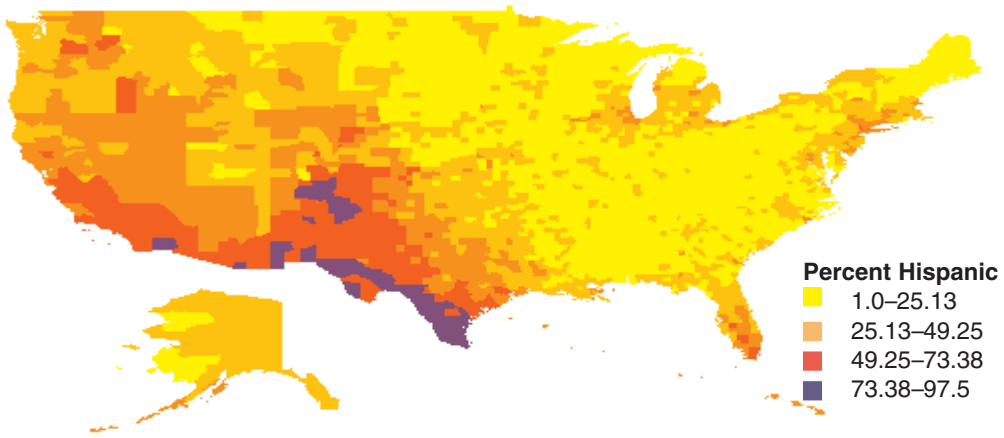


Figure 6 US Hispanic population, county level, 1990. Based on customized distribution intervals.

In addition, health planners might be interested in knowing whether certain counties are relatively underserved because they are predominantly Hispanic. Ideally, one would want to correlate and overlay data of different types, such as the proportion of the population that is Hispanic with the level of health care service provisions. Such developments are currently underway and will facilitate creating useful GIS for many areas of interest. It is important to note that the three maps presented here can be created in only a matter of minutes.

Now for a quick look at the border states, in particular, and some of the other capabilities of DDViewer. The map in Figure 7 shows the Hispanic population of the counties in California, Arizona, New Mexico, and Texas. The county border lines are visible. Figure 8, showing median household income, displays an inverse pattern to that of the Hispanic distribution. Figure 9 shows that the proportion of persons whose education ended at the elementary level corresponds to the Hispanic distribution and is inverse to the pattern of median household income. Figures 10 and 11 describe the extent of poverty in the region. While there are variations within most of the states, the stronger variation is seen from east to west.

The data for these counties can quickly and easily be downloaded into other software and additional analysis can be done. Figure 12 presents the correlation coefficients for these variables. As suspected, they confirm the visual conclusion. All relationships are statistically significant.

Additional Considerations

Users may query the map by pointing to (i.e., placing their mouse on) any polygon to identify underlying values or polygon labels—that is, place names—right on their computer screens. They may retrieve a data listing, as shown in Figure 13. It is also easy to create and re-code derivative variables. The descriptive data can be downloaded simply by cutting and pasting. It can then be read into any other software package for

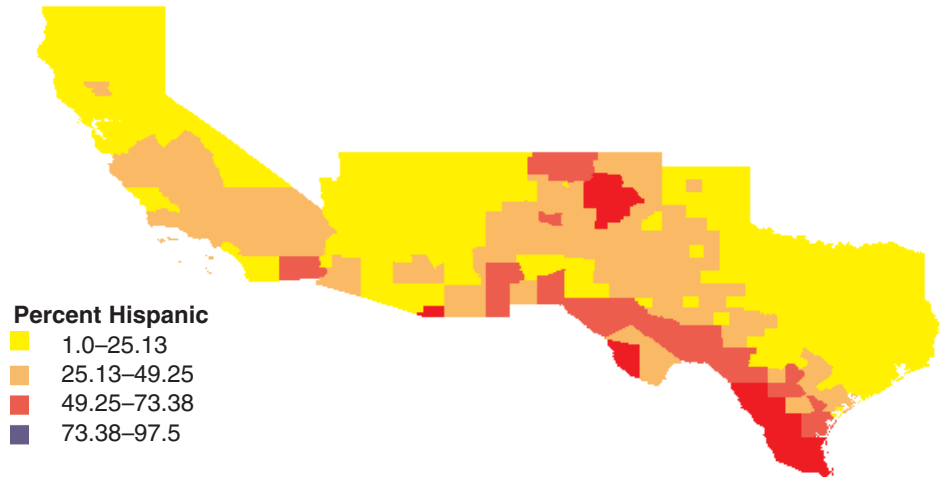


Figure 7 Percentage of Hispanic population for border states AZ, CA, NM, and TX; county level, 1990.

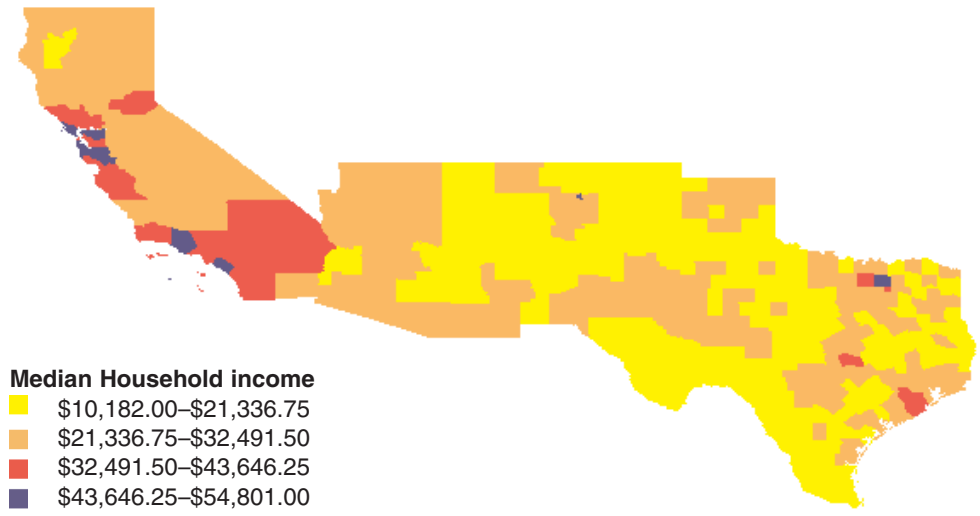


Figure 8 Median household income for border states AZ, CA, NM, and TX; county level, 1990.

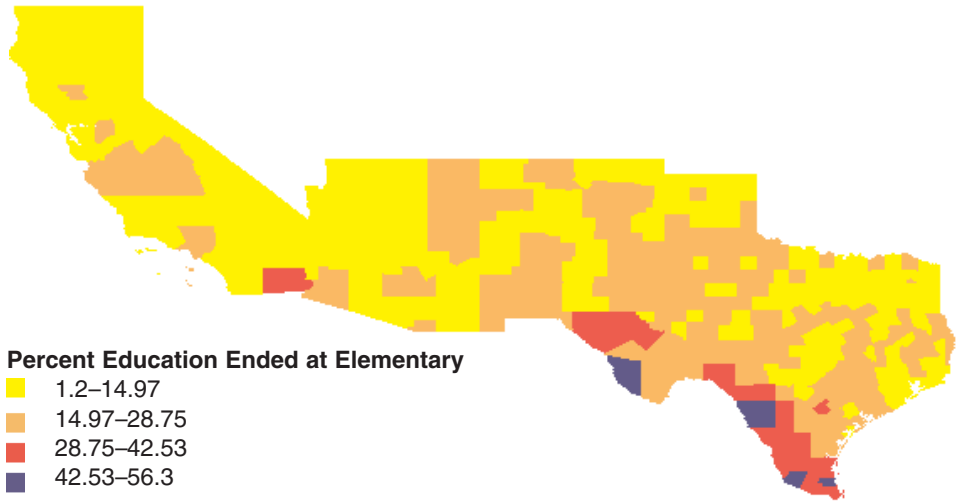


Figure 9 Percentage of persons whose education ended at the elementary level in border states AZ, CA, NM, and TX; county level, 1990.

further analysis. If a user is interested in data for a very large number of counties or tracts, such as the entire country, the data can be downloaded through another tool, the Demographic Data Cartogram, DDCarto, also available from CIESIN's homepage. The direct URL for DDCarto is http://plue.seda.ciesin.org/plue/ddcarto.

The basic GIS for Mexico's states, municipios, and islands has been completed. A

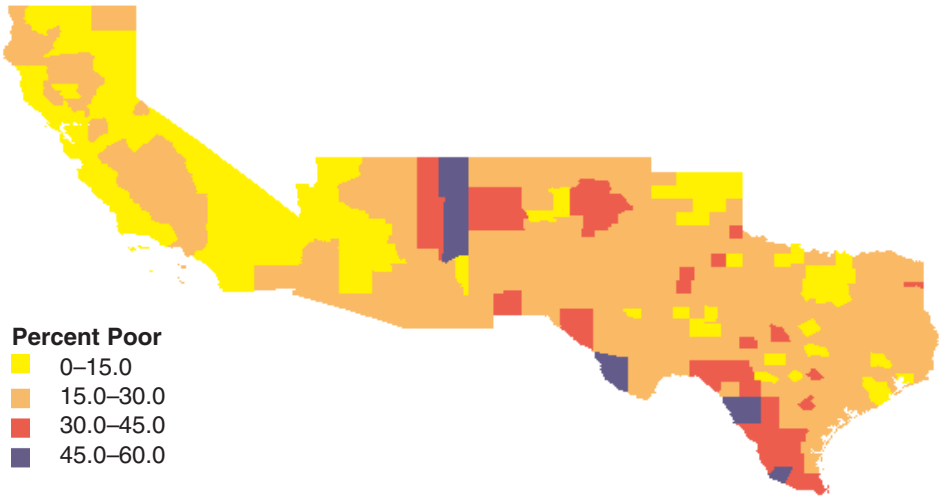


Figure 10 Percentage of persons living below the poverty line in border states AZ, CA, NM, and TX; county level, 1990.

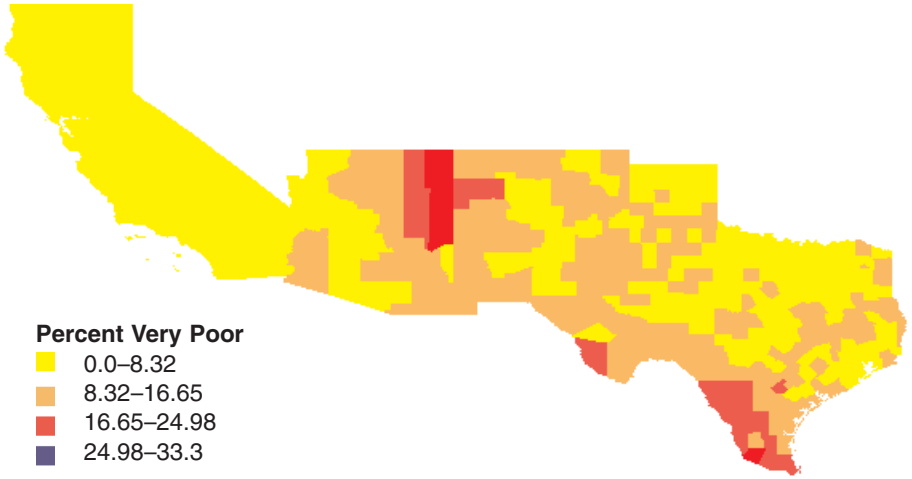


Figure 11 Percentage of persons living 50% below the poverty line in border states AZ, CA, NM, and TX; county level, 1990.

beta test version of it is available on CD-ROM. For anyone interested, copies are available from CIESIN and comments are welcome. The construction of the Mexico GIS required re-mapping the US-Mexican border so that the US and Mexican maps would align perfectly. The boundaries were taken from TIGER 95 for this correction.

Data for the Mexican states are being incorporated into DDViewer. The border states will be done first because of the special interest in border-related issues and the clear need for such information. Health data are also currently being added as part of DDViewer for both the United States and Mexico. The health data will first be introduced for the states along the US-Mexican border. The new applications will be released soon in both English and Spanish.

	Variable labels						
	Fipscode	*pcthisp*	*pct0_19*	*Medhhinc*	*pctpoor*	*pctveryp*	*pcthsch2*
Variable labels							
Fipscode	1.0000						
pcthisp	0.5677	1.0000					
pct0_19	0.3570	0.9205	1.0000				
medhhinc	-0.7080	-0.7154	-0.5428	1.0000			
pctpoor	0.6331	0.8641	0.7506	-0.8729	1.0000		
pctveryp	0.5453	0.7815	0.7021	-0.7806	0.9467	1.0000	
pcthsch2	-0.2161	-0.5123	-0.4054	0.3729	-0.5732	-0.5433	1.0000

Figure 12 Estimated correlation coefficients between exemplary variables.

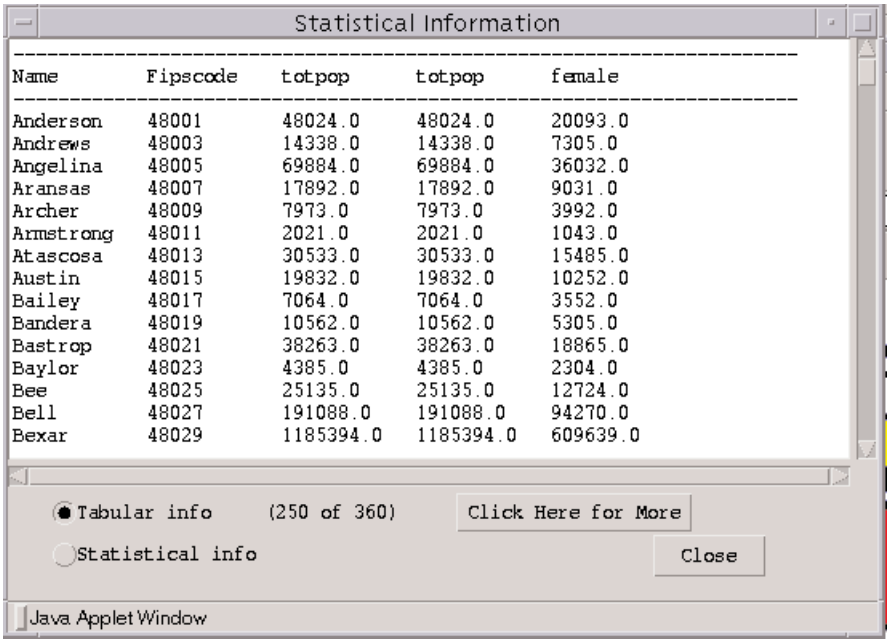


Figure 13 DDViewer data listing: selected variables by county name.

Future plans include moving beyond a strictly demographic data viewer to a spatial data viewer that allows for the integration of many more types and units of data and analysis. In addition, there are other areas for development, such as the inclusion of historical demographic data, hospital-level data, and environmental data.

Conclusion

This paper has focused on the power of DDViewer to display county-level data, but DDViewer can be an equally powerful tool at an even smaller level of analysis. On the

United States side, census-tract data provide that level of detail; however, on the Mexican side, the data are presently limited to the municipio. It is especially important to look closer at the health of the communities along the US-Mexico border. The border communities are affected by one another's general standard of living, political and administrative regimes, and shared environmental and geographic conditions. DDViewer shows how variations occur from east to west along the United States side of the border. Once the Mexican data are incorporated, differences from north to south and from east to west can be compared. The many communities that traverse the border can make use of DDViewer's data and visualization techniques without requiring extensive investment into their own GIS. Users' suggestions along with data from collaborators will determine the future developments of DDViewer in making local-level data widely accessible to researchers, policymakers, health care providers, and the general public.

Acknowledgments

Funding for this study was provided by the National Aeronautics and Space Administration (contracts #NAS5-32632 and #NAS5-98162).

References

1. Sovik N. 1997. *DDViewer 3.0: Interactive visualization of demographic data using Java technology.* Paper prepared for the Conference on Scientific and Technical Data Exchange and Integration. Bethesda, MD. Sponsored by the US National Committee for CODATA. 15–17 December.
2. Albrecht SL, Clarke LL, Miller MK, Farmer FL. 1996. Predictors of differential birth outcomes among Hispanic subgroups in the United States: The role of maternal risk characteristics and medical care. *Social Science Quarterly* 77(2):407–33.

Using GIS as a Management Tool for Health Care Assessment and Planning

Audra Eason, U Sunday Tim*
Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, IA

Abstract

One of the primary goals of public health, and of many health care providers, is to maximize the impact and effectiveness of limited resources in improving health care. Of particular interest are the availability and accessibility of health maintenance organizations (HMOs) to sections of the population that have the greatest need. Studies have shown that elderly populations benefit most from such health care services because many of them live alone, have limited incomes, and have high medical costs due to poor health. In the area of public health and health care management, geographic information system (GIS) technology has emerged as a powerful tool for integrating and communicating information, a tool that offers significant advantages over traditional methods for health surveillance. In this study, GIS techniques were used to determine the spatial distribution of health care facilities and analyze patterns in that distribution with respect to elderly populations in the state of Iowa. The purpose of this study was to demonstrate the role of GIS in health care and to develop a spatial analysis and modeling support system for forecasting future health care needs and planning health management programs.

Keywords: health care, facilities siting, decision support system

Introduction

Several studies have used geographic information systems (GIS) to address issues in health care planning and to examine the accessibility of health care centers to particular sectors of the population. GIS has been used in risk assessments, site selections, health surveillance, epidemiological studies, and other areas in the health care sector (1). Although the specific approaches have varied, GIS has proven to be a powerful system for spatial analysis and decision-making.

Demographic Changes in the Elderly Population

In the United States, the elderly (age 65 or older) population has grown significantly faster than the nation's total population. During the 20th century, the total population in the United States tripled while the elderly population increased by a factor of 11, going from 3.1 million in 1900 to approximately 33.5 million in 1995. The states with the highest proportion of elderly people in 1993 were Florida, Pennsylvania, and states in the Midwest. The states with the highest proportion of people 85 years old and older were all in the Midwest, with Iowa ranking the highest, followed by North Dakota, South Dakota, Nebraska, and Kansas. Currently, the elderly population is not rising, but this is expected to change after the year 2000. According to the US Census Bureau's

* Sunday Tim, Iowa State University, 215 Davidson Hall, Ames, IA 50011 USA; (p) 515-294-0466; (f) 515-294-2552; E-mail: tim@iastate.edu

middle series projections,[1] the elderly population is expected to reach 39.4 million in 2010, 79 million in 2030, and 80 million by the year 2050 (2). One of the primary reasons for such a dramatic rise is the aging of the baby boom generation. The baby boom generation consists of the 75 million people born between 1946 and 1964. Iowa is one of the states expected to have the greatest proportion (between 2.5% and 3.8%) of persons age 85 and over during the period between 2010 and 2030 (3). The aging of the baby boom generation will increase Iowa's need for health care services.

National Health Care Issues

Healthy People 2000, a comprehensive federal agenda for national health care, has recognized the need to identify issues that pose a threat to health and to attempt to address those issues in the public and private sectors. One of the overall objectives of Healthy People 2000 is to increase people's access to preventive health care services. Accessibility of health care facilities is important because it indicates how effectively health care facilities are serving the community. A group's level of access to health care facilities also reflects its mobility, as well as its spatial separation from certain destinations (4).

Spatial separation from health care facilities may be particularly significant for the elderly population. Health care facilities tend to be distant from elderly populations, despite the fact that the elderly have a greater need for health care services because of their increased chronic illness and morbidity. Research suggests that the locations of health care providers are not always based on health needs. The distribution of physicians, for example, is often clustered in the inner city around large hospitals (5). A study published by the New England Journal of Medicine concluded that communities with fewer than 180,000 people may be too small to support effective competition among HMOs (6). Thus, the location of HMO providers may cause problems for smaller populations; HMOs have prospered in areas with larger populations and persons must live within an HMO's service area to receive coverage from it.

Studies also indicate that some elderly persons rely on hospital emergency rooms for primary health care because few of them have personal physicians. In 1995, older persons accounted for 38% of all hospital stays and 48% of all days of care in hospitals; persons under 65 had an average hospital stay of 4.5 days while persons over 65 had an average stay of 6.8 days (7). With the limited numbers of health care providers available in any one area, health care resources become competitive between the elderly and people with fewer financial and physical limitations. These issues can create serious implications for the future of health care, especially considering the expected demographic changes.

Our analysis was guided by several questions related to the expected growth of the nation's elderly population. First, how are the locations of health care facilities distributed in relation to elderly populations with the greatest need? Second, are the health care facilities near enough to those elderly populations? Third, will the locations of health care facilities be sufficient given the projected demographic changes in elderly populations? The focus of this study is to assist in the targeted delivery of health services by identifying areas of need. The emphasis is not on explanation, but on producing

[1] The US Census Bureau's middle series (as opposed to high or low series) projections are made based on the assumption that past and current trends will continue.

an estimate of the variation in health care needs across an area, allowing for the variations in factors such as population age structure.

Methodology

The study was done using the ArcView GIS software package (ESRI, Redlands, CA). GIS techniques were used to analyze the spatial distribution of health care facilities by mapping their locations, then evaluating the distance between the facilities and areas with high concentrations of elderly persons. Demographic characteristics of the elderly populations for 1995 as well as projected trends for the year 2000 were examined.

The area chosen for analysis was the entire state of Iowa, which has 99 counties. The state's total population in 1995 was 2.8 million.

Data Collection and Preparation

Several considerations were made in deciding on the data to be included in the study. One of the problems addressed was how to locate elderly populations in Iowa. The unit of geographic analysis chosen was the census block group. As of 1995 and 2000 (projected), Iowa has 2,939 block groups. Of those, Polk County has approximately 317 block groups and seven hospitals, Johnson County has close to 63 block groups and four major hospitals, and Sioux County has 27 block groups and three major hospitals. Available attributes of the block groups include the percentages of men and women in different age groups, with people 85 and older making up the oldest group. Block group census data from the years 1995 and 2000 (projected) were linked to create one table of information.

Another consideration was the choice of health care providers to include in the study. There is a wide range of HMO providers: medical groups, individual physicians, physician-staffed health centers, and hospitals. In this study, hospitals were chosen as the primary source of HMO providers. The dataset of hospital locations for Iowa listed 137 hospitals in total. Data on Iowa hospitals include each facility's address, county, number of beds, and accreditation.

Another factor in the selection of data was the question of how to measure the accessibility of hospitals to elderly populations. The approach chosen was to measure distances between hospitals and selected groups of elderly populations. The analysis also included a map of major highways in Iowa. The coverage of highways contains a database of 1,821 major highways in Iowa.

Maps of hospital locations, block group boundary files, and major highways were all obtained from the Iowa Department of Natural Resources and imported into ArcView GIS. These parameters were used to measure accessibility of Iowa hospitals to elderly populations and were linked to a map of Iowa counties for analysis.

Measuring Accessibility

At the beginning of the analysis, a group of populations was selected for evaluation. The selection was based on block groups that had the highest percentages of elderly men and women in 1995 and 2000 (projected). Because elderly women significantly outnumber elderly men, different criteria were used for selecting target block groups. The block groups targeted were those in which the elderly population in a block group exceeded 25.6% for men and 38.6% percent for women (Figures 1–4). Of the 2,939 block

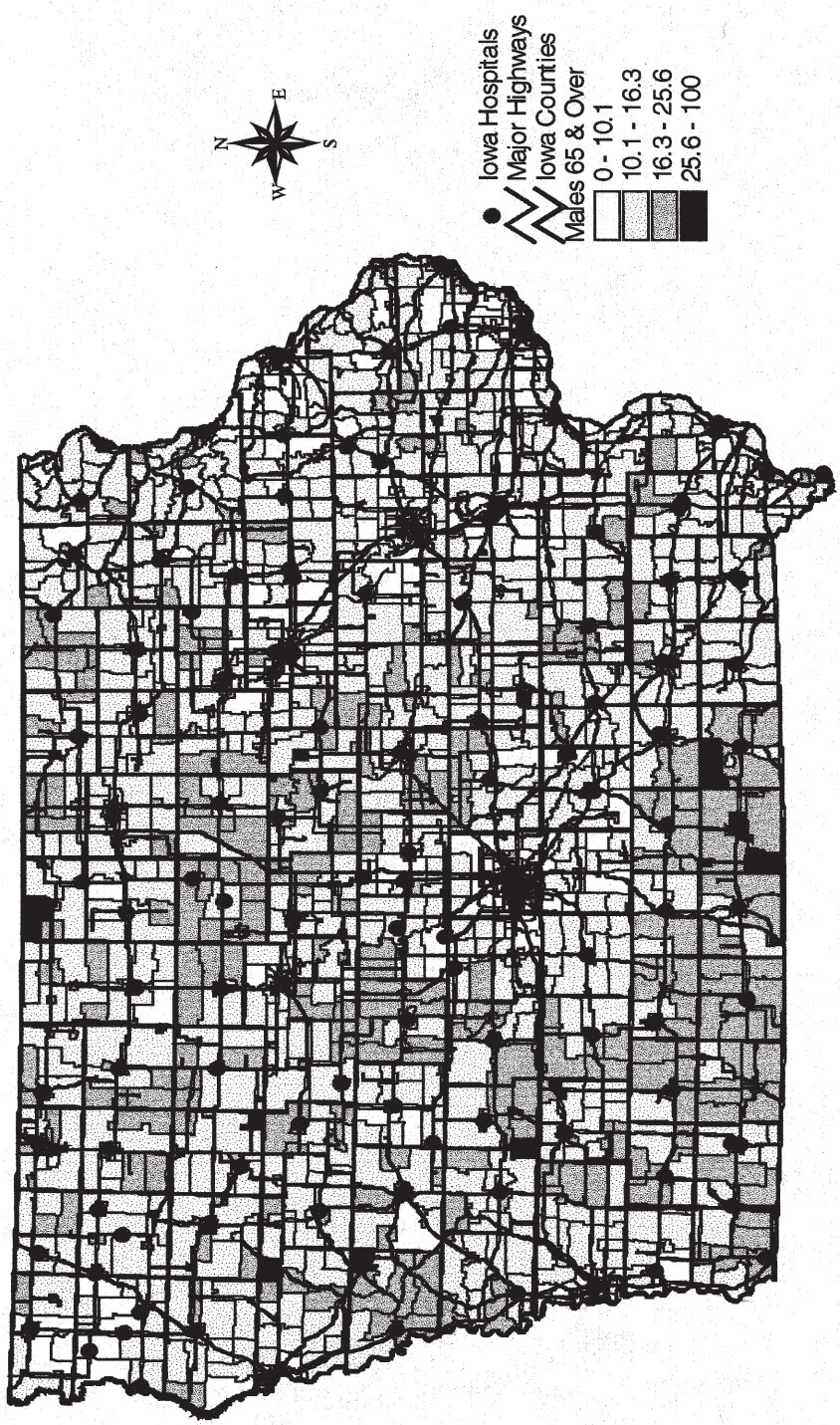


Figure 1 Elderly men and accessibility of major Iowa hospitals in 1995.

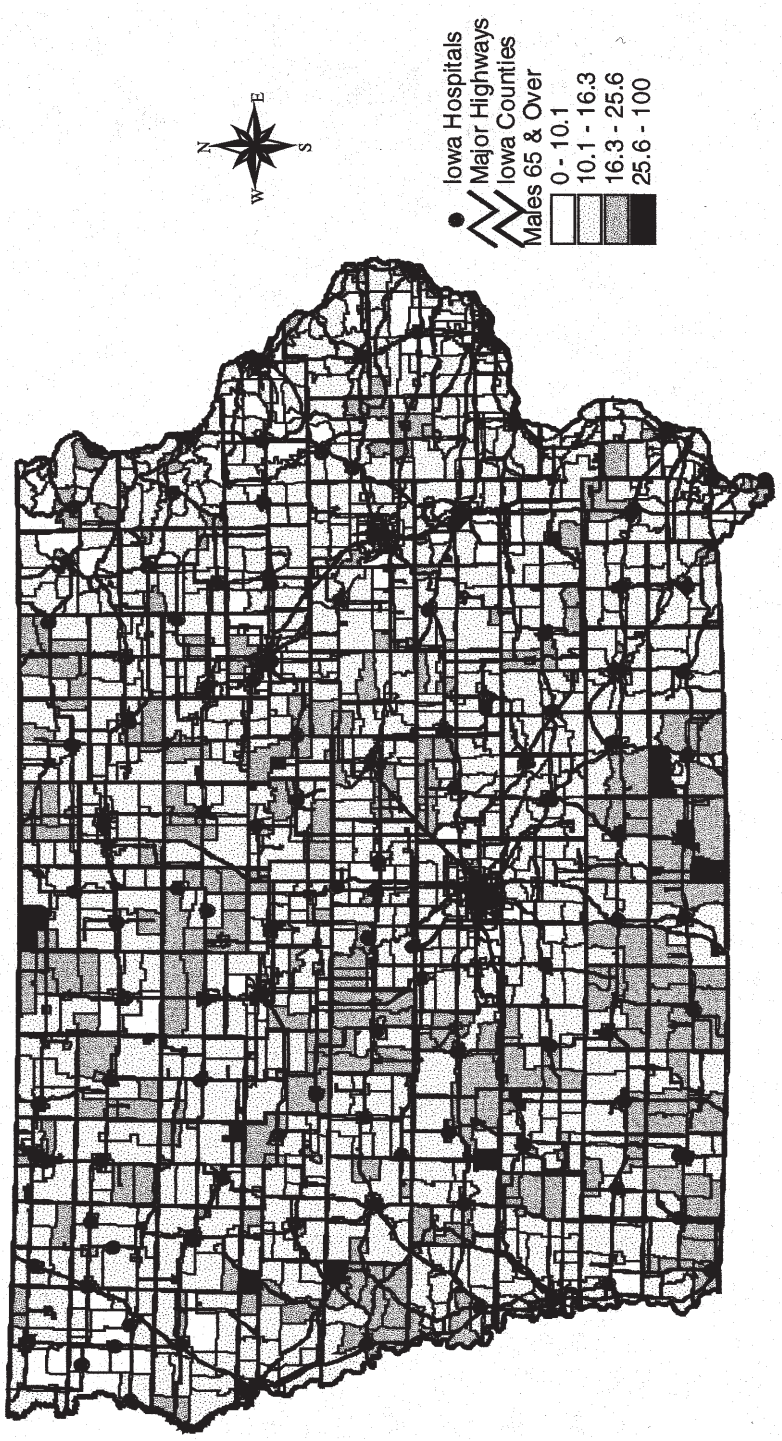


Figure 2 Elderly men and accessibility of major Iowa hospitals in 2000.

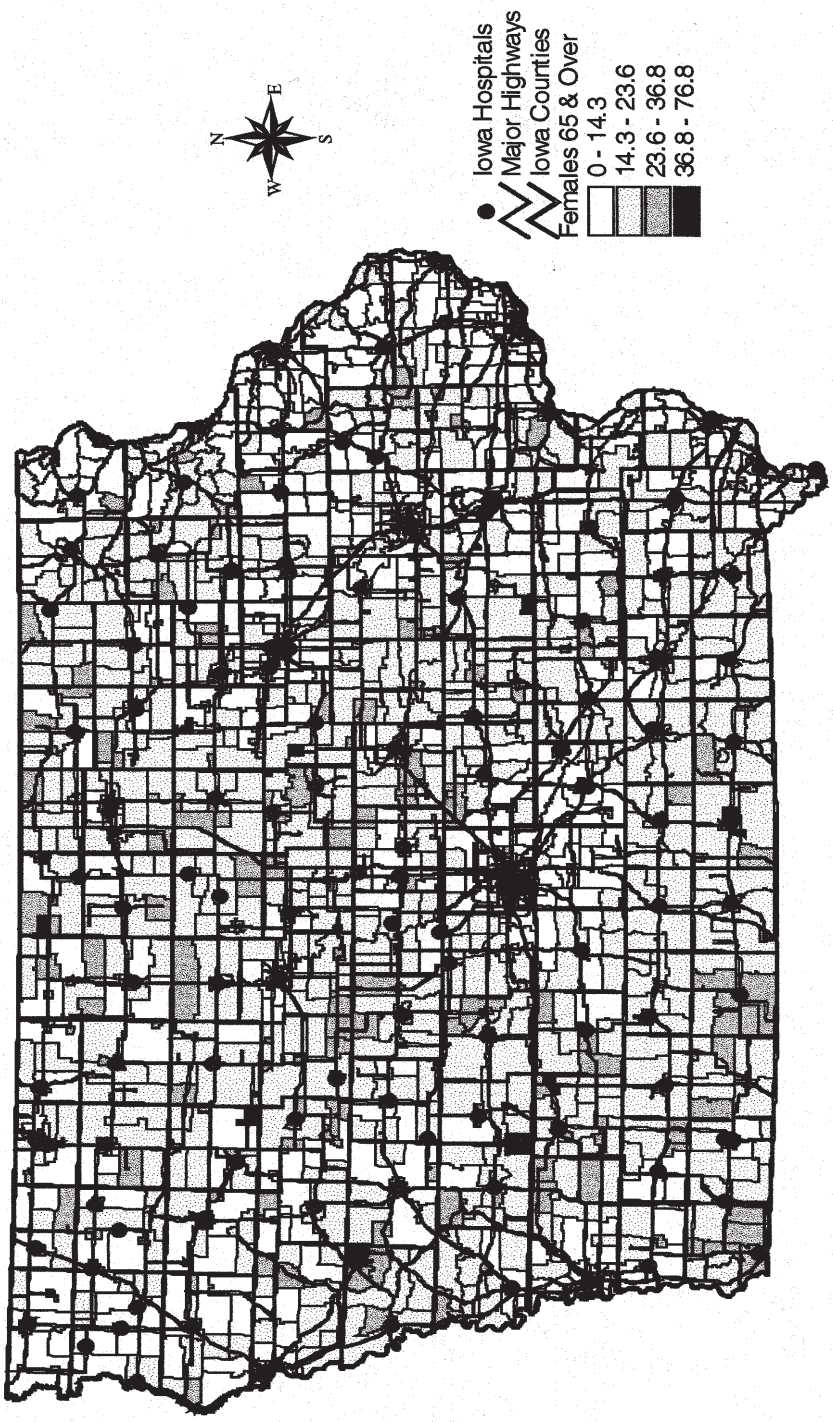


Figure 3 Elderly women and accessibility of major Iowa hospitals in 1995.

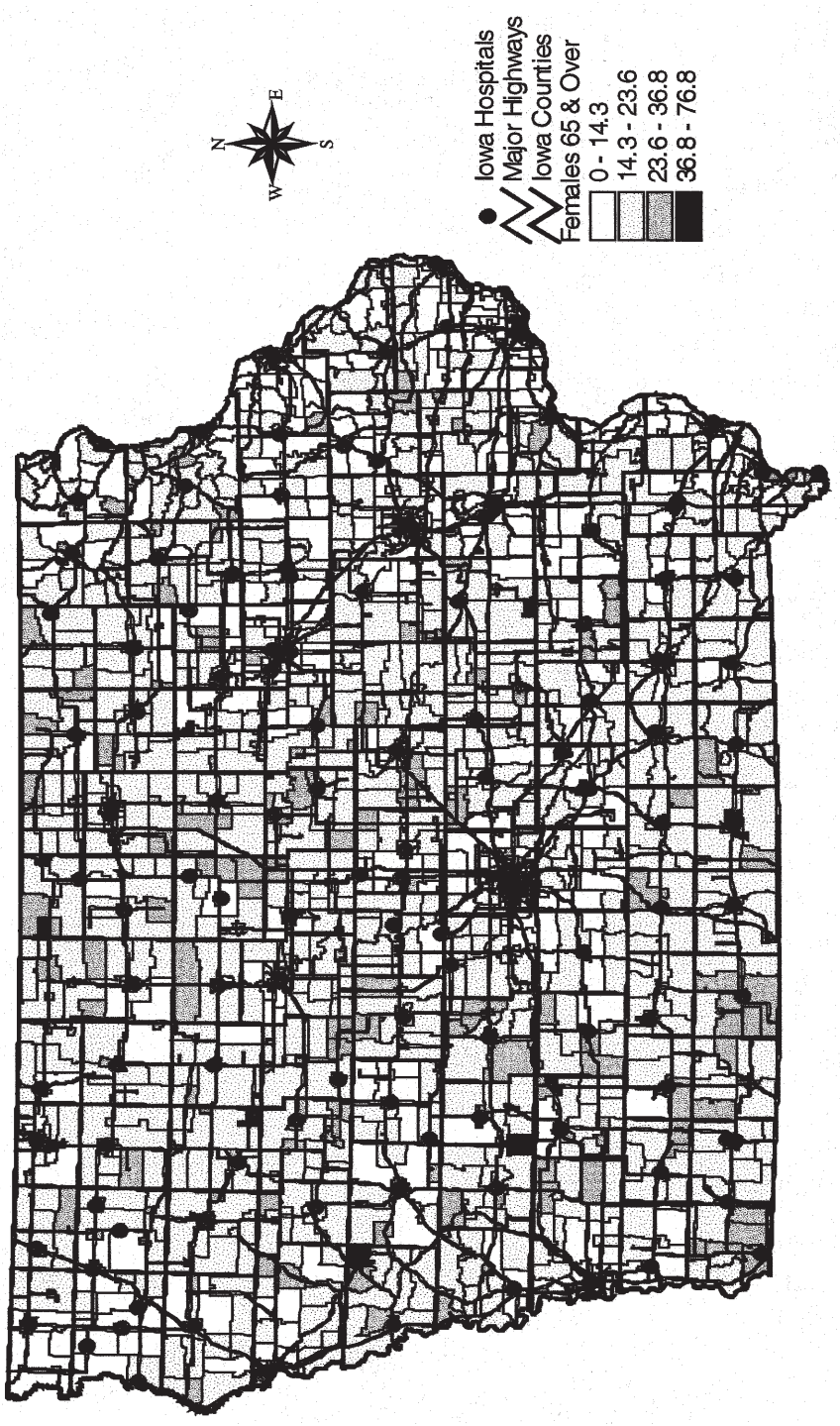


Figure 4 Elderly women and accessibility of major Iowa hospitals in 2000.

groups in Iowa, 3.33% of the block groups were selected for men (98 targeted groups) and 3.84% were selected for women (113 targeted groups). The number of block groups selected for analysis was the same for 1995 and 2000 (Table 1).

Table 1 Percentage of Block Groups in Iowa Identified for Evaluation

Block Groups Evaluated	Evaluated Block Groups as a Percentage of All Block Groups in State
38.6% or more elderly women in 1995	3.84%
25.6% or more elderly men in 1995	3.33%
38.6% or more elderly women in 2000	3.84%
25.6% or more elderly men in 2000	3.33%

Most of the analysis for the selected block groups was done using the Spatial Analyst extension in ArcView GIS. The distance mapping capabilities of ArcView GIS were used to create a grid coverage of distances to all hospitals. For the state analysis, uniform circular buffers of 10 miles were created around all hospitals and targeted groups (Figure 5). It was assumed that 10 miles would be the maximum traveling distance desired by persons in need of services at nearby hospitals. The proximity mapping function of ArcView GIS was used to display hospital service areas and identify the hospitals nearest to the elderly populations that do not have access to nearby facilities. A county analysis was done for Polk, Johnson, and Sioux Counties, which all have a large number of health care facilities compared with other counties in Iowa. The county analysis was done because, due to the large number of health care facilities in Polk, Johnson, and Sioux Counties, some populations may need to travel to these counties for special services (Figure 6). Uniform circular buffers of 50 miles were used for the county analysis.

Results

Iowa's elderly population, male and female, did not grow significantly between 1995 and 2000. Based on the classification of the elderly (Figures 5 and 6), the percentages of men and women within block groups remained about the same, although there were changes in population numbers. This was as expected; the percentage of elderly persons was not predicted to change dramatically through the year 2000. However, a greater change would have been seen if demographic data were projected through the year 2030 because this is the period expected to have a rapid increase in elderly populations. The results of the analysis for the accessibility of health care facilities to targeted populations are provided in Table 2.

Percentages of block groups within 10 miles of health care facilities were the same in 1995 and 2000 for both men and women. For men, the number of targeted block groups within 10 miles of a hospital was 79.59% (78 out of 98 targeted groups); 20.41% of the block groups were more than 10 miles away from a hospital. For women, the number of targeted block groups within 10 miles of a hospital was 84.96% (96 out of 113 targeted groups); 15.04% of the block groups were more than 10 miles from a hospital. For both men and women, the block groups that were more than 10 miles away from

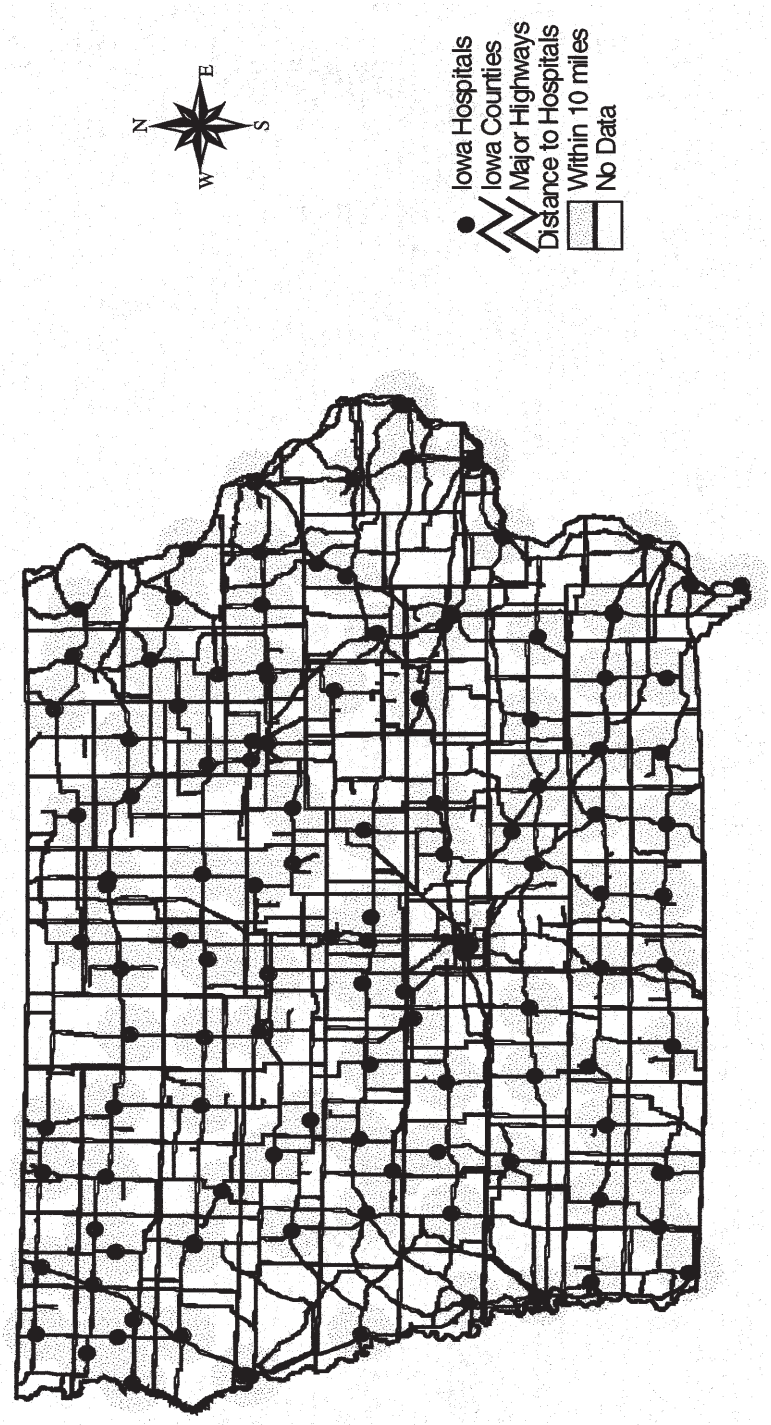


Figure 5 State analysis—10-mile buffers of major hospitals in Iowa.

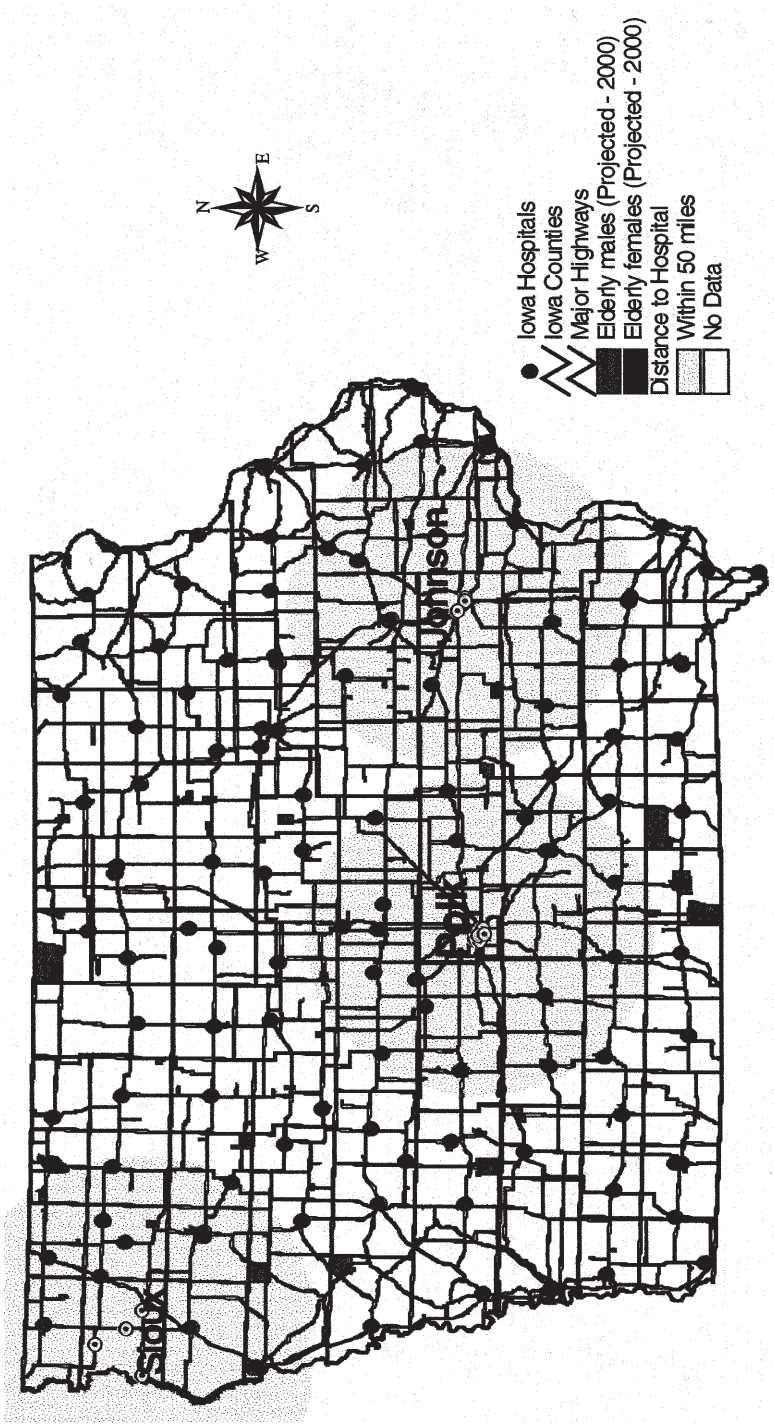


Figure 6 County analysis—50-mile buffers of major hospitals in Polk, Johnson, and Sioux Counties, Iowa.

Table 2 Percentages of Targeted Populations (Defined as Block Groups) within the 10-Mile Buffers of Major Hospitals in Iowa

	1995		2000	
	Women	**Men**	**Women**	**Men**
Percentage within 10 miles of a hospital	84.96%	79.59%	84.96%	79.59%

the nearest hospital were located mostly in the northwest portion of the state. Seven counties in northwest Iowa contained block groups that were targeted for both men and women and were more than 10 miles from the nearest hospital. These counties are Ida, Monona, Sac, Palo Alto, Pocahontas, Calhoun, Webster, Carroll, and Guthrie. Based on the analysis, these areas in northwest Iowa have the greatest need for additional health care facilities.

Polk County has seven hospitals and Johnson and Sioux Counties have four hospitals each. All other counties have one or two major hospitals. Polk County had approximately 18 targeted block groups within 50 miles of one of the seven hospitals. Three of these block groups did not have access to a hospital within 10 miles. There were 17 other health care facilities located in nearby counties. For Johnson County, six targeted block groups were within 50 miles of at least one of the four facilities in Johnson County. Only one of these block groups was not within 10 miles of a hospital. Sioux County had four targeted block groups within 50 miles of its four hospitals. One of the four targeted block groups was more than 10 miles away from a hospital. There are about 12 other hospitals in nearby counties.

Summary and Conclusions

The objective of this research was to examine the spatial distribution of hospitals and elderly populations in Iowa. For this purpose, GIS techniques were used to analyze the distances of hospitals from block group populations with the highest percentages of elderly men and women, and determine the areas that have the greatest need for health care services. In Iowa, the proportion of elderly persons is expected to increase significantly, which may affect the availability of health care resources in the state. Thus, the results of the analyses provide important information for the assessment and planning of health care facilities.

The main findings of this study are that, in Iowa, there were not significant changes in the number of elderly persons between 1995 and 2000, although the same is not expected after the year 2000; at least 15% of elderly block group populations are more than 10 miles from the nearest hospital; and, most of the block group populations that are more than 10 miles away from a hospital are located in northwest Iowa, which is accordingly identified as the area with the greatest need for additional health care facilities.

The results of the analyses are useful in that they show that distance is a good indicator of accessibility. GIS was a useful tool for identifying elderly populations with the greatest need for health care services. The information obtained could serve as a guide for addressing health care needs in the elderly population at the local, state, and

national levels. Using the information would allow health care officials to plan for future health care facilities and meet national health care goals. These findings could also be a basis for further GIS analysis to examine areas with the greatest need for health care services through the year 2050, evaluate other indicators of accessibility such as distance between hospitals and public bus stops, or analyze the accessibility of nursing facilities to the elderly. Further studies would provide a more detailed look at the needs of the elderly and help state agencies evaluate current needs and plan for future population changes.

References

1. Stern RM. 1995. The added value of geographical information systems in public and environmental health. In: *Environment and health data in Europe as a tool for risk management: Needs, uses and strategies.* Stern RM. Boston: Kluwer Academic Publishers. 3–24.
2. US Census Bureau. 1996. *65+ in the United States.* Current Population Reports, Special Studies, P23-190. Washington, DC: US Government Printing Office.
3. US Census Bureau. 1994. *1993 from state age-sex population estimates consistent with Census Advisory CB94-43; 2020 from population projections for states, by age, sex, race, and Hispanic origin: 1993 to 2020.* Current Population Reports, P25-1111. Washington, DC: US Government Printing Office.
4. Robson P. 1982. Patterns of activity and mobility among the elderly. In: *Geographical perspectives on the elderly.* Ed. AM Warnes. New York: John Wiley & Sons. 265–80.
5. Bohland JR, Frech P. 1982. Spatial aspects of primary healthcare for the elderly. In: *Geographical perspectives on the elderly.* Ed. AM Warnes. New York: John Wiley & Sons. 339–53.
6. Ganske G. 1995. HMOs currently lack the capacity to serve the entire nation. In: *Medicare reform: Short-term tourniquets and long-term cures.* G Ganske. San Antonio, TX: National Organization of Physicians Who Care. http://www.pwc.org/ganske/index.html.
7. Administration on Aging. 1997. *Transportation and the elderly. A profile of older Americans.* Washington, DC: US Department of Health and Human Services.

Health Service Sites Access Analysis Using Internet GIS

Yongmei Lu*
Department of Geography, State University of New York at Buffalo, Buffalo, NY

Abstract

Health service sites access assessment is critical for patients looking to get timely and proper service. Also, access analysis results can be helpful when recruiting health care providers in underserved areas, or when referring patients to nearby practitioners. Health service sites access assessment is a typical spatial analysis, which can be greatly improved by using a geographic information system (GIS). However, given the hardware and software requirements, a GIS package itself is not always easily accessible to the public. In addition, lack of adequate training acts as a major barrier blocking ordinary people from exploring the power of GIS. Thus, we are in critical need of a bridge to bring GIS to ordinary people who need to solve spatial analysis problems such as health service sites access analysis. The development of distributed technology today makes it possible for data and software service providers to offer access to their services to anyone connected to the Internet, and for users to benefit from being able to share resources regardless of their physical location. Given the popular access to and general familiarity with the World Wide Web, the Internet could be an ideal vehicle to carry both GIS functions and health service information to people who need them. That will definitely enhance health services' capability for improving people's quality of life. In addition to proposing that health service sites access analysis be available on the Internet, this paper also discusses three architectures and general models using Internet GIS for performing the access analysis.

Keywords: access analysis, common gateway interface (CGI), common object request broker architecture (CORBA), Internet GIS, public health service

Introduction

One of the most widely agreed-upon statements on the role of public health is that it is an essential service of public health to ensure that all members of the community have access to health services (1). Access to a health service site is critical to the increased efficiency and effectiveness of health care delivery. It is an important issue for both service consumer and service provider. Health service consumers have the need to identify or be referred to the most suitable, as well as nearest, service site. Thus, having access to and being able to perform analysis on information about health service sites is essential for consumers if they are to access health services more conveniently and efficiently. For health service providers, their operations are highly related to their service capabilities and market shares. Providers need to define their service area and population. They need to identify the underserved and/or high-risk populations in order to give full play to their services. Hence, information and analysis are necessary

* Yongmei Lu, Dept. of Geography, State University of New York at Buffalo, 105 Wilkeson Quadrangle, Buffalo, NY 14261 USA; (p) 716-645-2722; (f) 716-645-2329; E-mail: yonglu@geog.buffalo.edu

for health providers to get full and accurate knowledge about access to their services for the public. They also need to process the related information to stipulate and evaluate a development strategy for their services. Furthermore, from the standpoint of public health agencies and government, access assessment is a basic method for evaluating and ensuring universal access to health services. The analysis is essential for making decisions on health service recruiting, planning, and other development strategies and policies.

Obviously, the ability to access and process information about health services is the first and foremost step for all parties to improve access analysis and delivery of health services. Actually, many organizations and researchers have already taken steps to implement new information technology in the health care industry. Recent efforts and practices in building and testing both a community care network (CCN) and community health information network (CHIN) are good examples. Both networks have adopted the idea of improving information availability and exchange. A CCN, which emphasizes information exchange and broad collaboration, includes as its vision "making the system more understandable and user-friendly for patients and enrollees" and "continually improving the continuity and quality of health care services" (2). And a CHIN, although still in the innovation stage, has agreement on minimum elements including that "computer-based information systems and networks form the base technologies of CHIN," and "the major domain of CHIN data transferred is health and the provision of health services" (3). In the mean time, many public health agencies take advantage of Internet technology and post certain information (e.g., locations, contact information) about public health services on their Web pages to provide better service for their customers. However, the majority of these existing efforts to improve information availability are of a traditional and rudimentary status—they basically provide text information with little clue about actual spatial accessibility, and information is general rather than locally specific.

Access assessment is a typical spatial analysis problem that can be easily handled by GIS technology. In addition, GIS' usual products of a map and/or report are very easy for GIS laymen to understand and communicate. Nevertheless, GIS, being a technology having special hardware, software, and personnel training requirements, is not readily available to the public. We need to find a way to perform GIS functions for access analysis at a simpler level or relieve the special requirements that block the general public from using the technology. Fortunately, the recent development of information technology in general and distributed computing and Internet technology in specific make it possible for ordinary people to use GIS for health service access analysis while avoiding the almost formidable expenses in terms of both time and money.

The rest of this paper will be devoted to exploring the potential of modern information technology for improving the general public's ability to access and process health service information. Based on the discussion about the potential of distributed computing technology and Internet GIS, three architectures applying these technologies to processing health service information are proposed. The three architectures are a server-side application, a client-side application, and an Internet-savvy application. The purpose and pros and cons of each of these approaches are discussed. Finally, the general implications of these development models for the future of the public health industry are addressed.

Data for Health Service Access Analysis

"Access to health services" refers to how easy it is for consumers to get the necessary service and for providers to deliver their services. Usually, it is measured by the impact of geographical distance on the convergence of health service providers and consumers. Broadly speaking, though, access analysis also concerns the impact of factors related to the consumer's as well as the provider's social, economic, cultural, and even language situation. For example, a non-English speaking patient may prefer to travel longer just to get service from a certain site where the patient feels it is easier to communicate. Hence, data needed for a useful and effective access analysis is far beyond just a simple map from which physical distance between locations can be measured easily. Social, economic, demographic, and environmental data are all necessary.

Generally speaking, the data for health service access analysis have the following properties: First, data volume and categories are large and data are usually generated by various parties. Some data (e.g., census data and TIGER street maps) are more general and easy to obtain; others (e.g., location information about service sites) could be very specific both geographically and thematically and need to be created from scratch. Second, related data are maintained and updated separately and independently by different parties. Consequently, data concurrency and integrity must be checked carefully. Because of the spatial nature of access analysis and the large size of the databases, GIS is undoubtedly the most powerful tool.

GIS is far from reaching its full potential in health service access analysis. Even though GIS is the best tool for maintaining and analyzing large spatial databases, difficulties in physically obtaining, owning, and maintaining all related data as well as a GIS package restrict GIS to the hands of selective researchers and health service professionals who have the resources and expertise. This does not excuse us, however, from attempting to use GIS technology to enhance the effectiveness and efficiency of our health service delivery. The hope lies in new information technologies—distributed computing and Internet GIS.

A New Vision of Distributed Computing Technology

Distributed computing technology is the technology that integrates a collection of distributed operating systems and/or distributed database systems by a communication subnet. The communication subnet may be a widely geographically dispersed collection of communication processors or a local area network. The emergence of distributed computing technology is significant for the development of information technology such as GIS. "It is now possible for parts of a database to be stored and maintained at different locations, for users to take advantage of economical or specialized processing at remote sites, for decision makers to collaborate across computer networks to make decisions, and for large archives to offer access to their data to anyone connected to the Internet" (4).

Distributed computing has certain advantages. First, costs are significantly lowered, while access to information technology is greatly improved. Multiple organizations and users can now share both expensive special processors and mutually desired datasets. Tremendous money, time, and personnel could be saved, and more people can access processor resources and databases. Second, geographical restrictions to accessing

resources and data are totally removed. Distributed computing technology makes it possible to remotely access both resources and data. It gives people more geographical flexibility. Finally, greater flexibility and reliability of a distributed system make facilities and data easier to share and maintain. All elements of the distributed system are connected to each other while relatively independent. Deleting, adding, updating, or even failure of a certain element does not affect the availability of others at all.

When combining distributed computing technology with GIS, we can envision the following benefits for our health service access analysis: First, spatial data generated by different parties could be easily integrated for certain applications; in this case, it would be health services access assessment. Database redundancy and computer facility waste from keeping multiple copies of data are minimized. Second, data maintenance is distributed and data are relatively easy to update. Updating "master copy" data will ensure that every user can get timely data all the time. Spatial analysis results of GIS are more reliable, valuable, and consistent across applications. Third, GIS hardware and software, although sophisticated, can be shared by geographically dispersed users. In sum, the development of distributed computing technology, by enabling the sharing of related data and GIS resources, makes it much easier for people to perform health service access analysis. It greatly reduces the requirements on the public in terms of GIS data, facilities, and experience. So, the problem now is, how do we combine traditional GIS with distributed computing technology?

Internet GIS Technology as a Valuable Direction

"Internet GIS is a network-centric GIS tool that uses the Internet as a primary means of providing access to the functionality (i.e., analysis tools, mapping capability) of GIS and to the spatial data and other data needed for various applications" (5). According to this definition, Internet GIS could be an ideal way to improve the public's assessment of health services access by enabling them to share data and GIS functions via Internet technology.

Although there has been basic research advocating the combining of GIS technology with the Internet (6), most of the development to date has been through the experimental development of both prototype and production services by different organizations. It is commonly acknowledged, however, that the Internet is a good vehicle, with the help of distributed computing technology, to empower more people with the capability to access and handle large spatial databases.

The Internet has certain characteristics that make it especially suitable as a means for combining distributed computing and GIS technology. First, the Internet (e.g., the World Wide Web) is very popular today and highly accessible to ordinary people. New content added to the Internet has good potential for being accessed and accepted by the general population. Next, the user-friendly aspect of the Internet and its browsers' capability for transferring and displaying text, graphics, and image files make the Internet a good medium for communication of spatial information and process results. Lastly, the Internet can provide interactive communication between users and data, and this is the foundation for the interactivity between Internet GIS users and spatial data. This characteristic of the Internet allows for the possibility of sending a dynamic map/image to a browser in addition to ordinary static ones.

Nevertheless, the World Wide Web and its hypertext markup language (HTML)

cannot directly recognize spatial analysis requests; neither can they handle spatial data in the way GIS software does. Therefore, in addition to a Web browser and Web server, special GIS tools need to be added in to understand and respond correctly to spatial data handling requests. Different solutions for adding GIS functions to build real Internet GIS result in different frameworks.

The first solution is purely distributed computing in technology and simple and consistent in logic. It is a server-side Internet GIS solution and is classified as a typical heavy-server and thin-client framework (Figure 1). The most important property of this

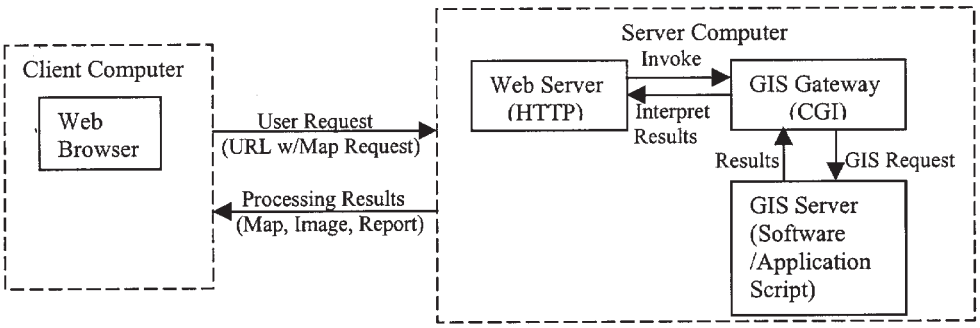


Figure 1 Framework of server-side Internet GIS.

framework is that GIS requests are always handled by the remote server(s). Common gateway interface (CGI) is widely used on the server's side to link Web server with GIS server. All parameters of GIS requests are interpreted by CGI for the GIS server to process; the results from the GIS server are interpreted and passed back to the Web server by CGI as well. In this framework, both spatial data and GIS resources are totally shared by multi-clients. The only task for the client machine (Web browser) is to receive a request from the user, send it to the server, accept processing results from a server, and display results. This solution is good because it could provide a whole set of complicated GIS functions to a user and could deal with a large database. Also, from the server's point of view, it is easy to maintain both database and GIS resources, and easy to control the access for different user groups. Nonetheless, it lacks flexibility and interactivity for the user. The user can only work with data in a limited way and receive static data processing results. Moreover, every request from the user, even though it could be quite simple, has to be sent back and forth through the Internet to the server, which causes high Internet traffic and a heavy workload for the server.

The second solution aims at giving the user more flexibility by doing GIS data processing and analysis on the client side. This client-side Internet GIS is classified as a light-server and thick-client framework (Figure 2). The hallmark of this solution is that GIS tools are located on the client's machine. The functions and powerfulness of GIS tools on the browser's side could be significantly different though, depending on the GIS package designed for the application. However, the user does not necessarily have his or her own GIS package and expertise. This framework supports the user's request for GIS tools from the server and the "installation" of the tools on the client's machine.

There are two general models of GIS tools for the client to request from the server: GIS plug-ins and GIS applets, both of which can efficiently add GIS functions to the

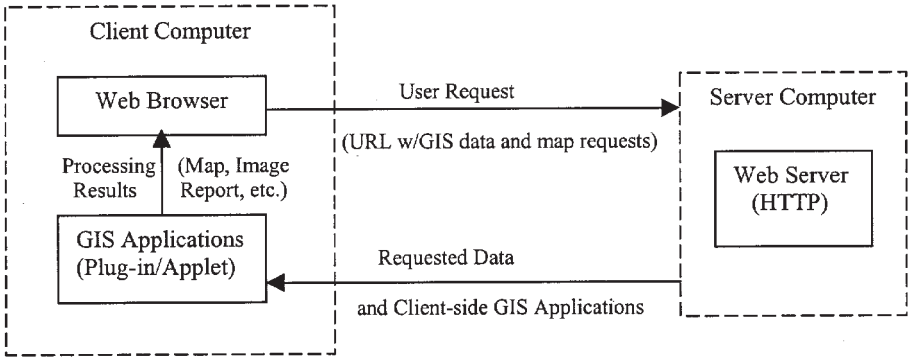


Figure 2 Framework of client-side Internet GIS.

client side. A GIS plug-in needs the client to download and install it as a normal program, which takes up not only time before the Internet GIS can be used but also permanent space on the client's machine. A GIS applet could be written in either Java or ActiveX. It could be designed to download automatically before the user issues a GIS request. An applet takes up space too, but the space will be released when the user leaves the Web site. Compared with server-side Internet GIS, the second solution provides the user with more flexibility and interactivity, because the user can process GIS data as he or she is using stand-alone GIS software. Because the GIS analysis is performed on the client side, not much Internet traffic is generated. Due to the requirement on client-side hardware resources, however, the framework is weak in handling large databases and performing complicated GIS functions.

In addition to the two above-described basic solutions, there are other frameworks, most of which are combinations of server-side and client-side solutions. Every solution has its own advantages and disadvantages, and thus is suitable for different situations. Many aspects of a GIS application, such as size of related spatial database, expected request frequency and data volume, the complexity of most-often-generated GIS queries, hardware capability available for general users, and even the maintenance of the server(s), can all influence the choice of Internet GIS solution.

Three Architectures for Health Service Access Assessment Internet GIS

It is clear now from previous discussion that data for health service access analysis are miscellaneous and huge, and usually generated and maintained by various parties. Also, health service access assessment is important for multiple groups, the three most important of which are health service consumers, health service providers, and public health agencies and organizations. With the assist of distributed computing technology, Internet GIS can be employed to integrate a variety of data and to empower people from different groups to process information. However, different user groups of the Internet GIS health service access application have different expectations and requirements. In order to provide all users with effective Internet GIS analysis power while keeping the application as efficient as possible and avoiding waste of resources, different development strategies should be implemented for different user purposes and analysis levels. The three Internet GIS architectures proposed here for these different

strategies are static access query and display, interactive access information query, and comprehensive and intelligent access analysis.

Static Access Query and Display

The static access query and display (SAQD) model is designed for the general population (health service consumers and patient referrers) to perform a simple and static query on access to health services. This application can perform such queries as location(s) of nearest health service site(s), route and directions to the site(s), and some other general information about the service of the site(s) (e.g., category of service(s), service capability and quality, contacting information). The query results will be a combination of map image and text. To enhance the display quality of the map, functions such as pan and zoom are also available for users. To give the user more flexibility, further development could include more features, such as allowing users to build more complicated and restricted queries according to their special requirements and concerns.

The functions and features of SAQD are designed based on the characteristics of potential users and their queries. This group of users and their usage of the Internet GIS have the following features:

- Queries are usually simple, and a static map plus text information can satisfy their general requests.
- Queries should be in relatively high volume, which is important when considering the amount of requests sent to the server.
- This application requires relatively quick response from the server.
- Users usually do not have access to a complicated GIS package; neither do they necessarily have access to terminals/client computers of high power.

Based on the above-mentioned features, this Internet GIS application is a typical thin-client one. The server does all the data integrating, analyzing, and map generating, while the client machine only accepts and delivers requests, receives results from the server(s), and displays the results (Figure 3). Although the server does most of the job, it does not necessarily result in heavy duties for the server, because the query is very simple and routine. The server does not perform sophisticated processing at all. Not-too-complicated scripts, an efficient search engine, plus well-organized databases could be specially packed together for this application. Actually, there are similar applications developed and used on the Internet. VISA ATM Locator is a good example (http://www.visa.com). The basic approaches are similar, although the static access

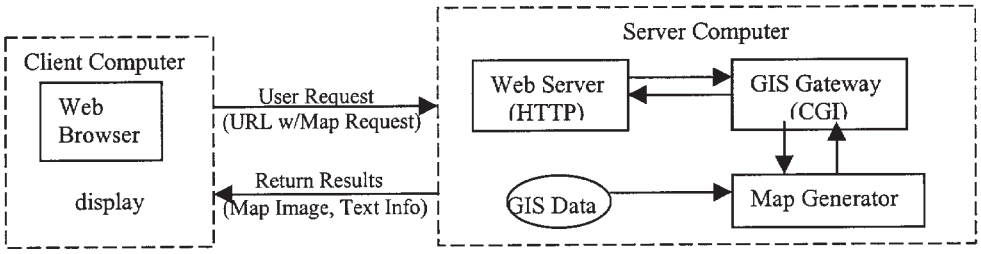


Figure 3 Architecture of static access query and display.

analysis Internet GIS for health service certainly contains more features, including pan and zoom buttons.

This approach has certain advantages. First, it does not require any special GIS or other resources on either client side or server side. Thus, it is easy to install. In addition, access to the application is almost ubiquitous, given the popular access to the Internet. Finally, server-side architecture makes it easy to maintain and update both data and a data analyzing, map generating package. The major drawback of this application is that it lacks the interactivity between the user and the spatial data. Also, the map that is transferred from server to client creates a relatively large volume of cyber traffic.

Interactive Access Information Query

The interactive access information query (IAIQ) Internet GIS application is built to satisfy the needs of relatively higher level queries. Major potential users are health service providers and other people who are concerned not only about the location of and route to health service sites but also about service area, population, and other socioeconomic attributes of a service area. Similar to the first group of Internet GIS users, this group also expects a relatively quick response to its requests. This group is somewhat smaller than the first one; hence, query volume should be relatively low. Also different from the first group, this group tends to generate higher level queries that request more data categories and relatively sophisticated data analyzing. Query format and content could be more interactive and complicated; dynamic maps, interactive queries, and even simple simulations may be required. Obviously, real GIS functions must be implemented. A higher level of Internet GIS service is required. IAIQ Internet GIS is the model able to satisfy these criteria. According to this architecture, maps and text information, as well as basic functions such as pan and zoom, will definitely be retained. New features such as identifying map features, interactively generating an object attribute report, turning on and off the display of multi-layers, and even performing a simple dynamic simulation, will be added in.

IAIQ is a typical client-side Internet GIS solution, conducting data analysis and map generating on the client's computer. Requests about health services, their service areas, and/or the basic situation of their major clients, and so on, are received by the client computer. Next, related data needed to process the requests as well as necessary GIS applications/scripts are passed back to the client computer from the server(s). Finally, GIS tools process related data on the client machine and the results are passed to the browser for display (Figure 4). The significant feature in this model is that the client machine gets both data and specially designed and packed GIS tools from the server. Flexibility of processing data using the GIS tools is guaranteed. The speed of dealing with further requests from the browser is relatively high.

As discussed earlier, there are two basic structures of the GIS tools: plug-in and Java/ActiveX applet. The preferred approach depends on the characteristics of the users. Compared with the first architecture of Internet GIS proposed, this architecture is supposed to serve the needs of people who are interested in analyzing health service access regularly, as well as users who are usually concerned about certain aspects of health service consistently. It is reasonable to assume that they have access to powerful computers and are willing to download and install the plug-in as a stable GIS tool kit

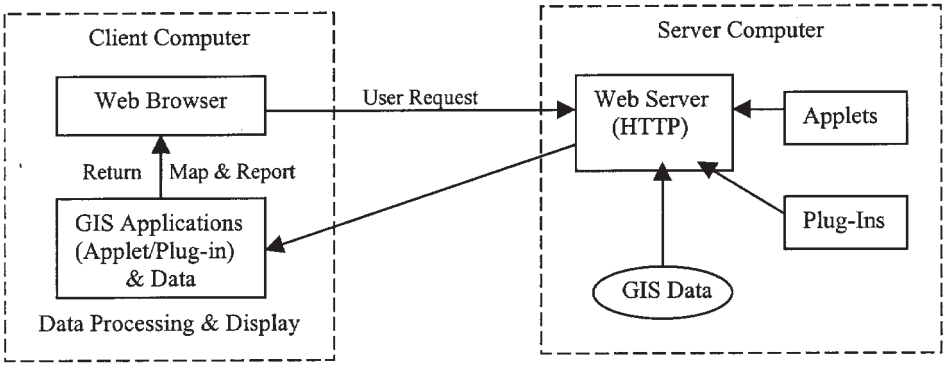


Figure 4 Architecture of interactive access information query.

to assist their analyses. As a matter of fact, plug-ins are usually more powerful than applets and more stable as well.

This approach is good for relatively high level GIS queries and interactive communication between users and data. Also, processing GIS requests on the client side is relatively quick and can avoid the need to send every request to and from the server. One disadvantage is that time is needed to download and install the GIS tool kit. Moreover, users need to receive timely information about the upgrade of the GIS tool kit, especially when a plug-in solution is adopted.

Comprehensive and Intelligent Access Analysis

A comprehensive and intelligent access analysis (CIAA) architecture of an Internet GIS application is on the highest level both functionally and technologically. Almost the whole set of GIS functions that can be found in normal GIS software is provided. A distributed computing technology that can handle real time data sharing is required, though it is actually not mature enough to date. The vision of this architecture is that elements of the Internet GIS, including clients, data providers, and GIS server(s), are distributed across the Internet while being able to communicate with each other and share data and resources in a timely manner. It is referred to as a "Net-savvy" GIS application by Plewe (7). The core point of this solution is the complete distribution of almost every element.

For a GIS analysis on health service access, the CIAA solution makes it possible for users to directly use census data from the Census Bureau, a street map from Etak, data about sites and services from public health service agencies and offices, and even newly released data about the most recent toxic substance leaking accidents. This supports health service access analysis on both the current situation and potential requirements. It can also enhance the evaluation of planning and development strategies and policies of health service systems through simulation of future scenarios. This is an architecture meeting the highest requirements of Internet GIS' function and power. It can satisfy the needs of health service administrator, planner, development policy stipulator and evaluator, and other specialists interested in evaluating the efficiency and effectiveness of a health service delivery system.

From a technical point of view, this approach is a severe heavy-client one (Figure 5). All the powerful GIS analysis is performed on the client side. The client software is a

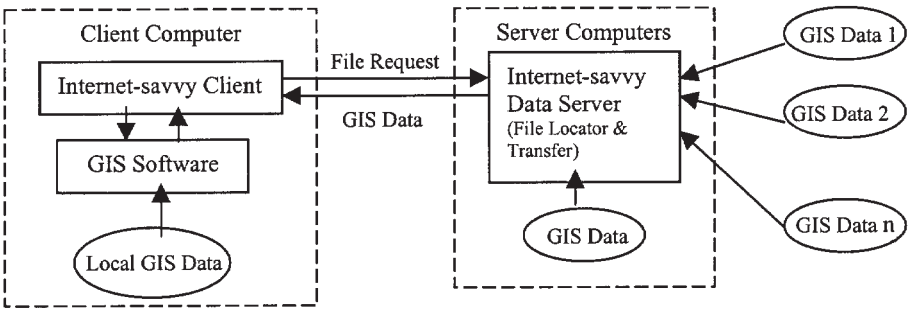


Figure 5 Architecture of comprehensive and intelligent access analysis.

standard GIS package having the ability to use real time data added in. The server is more a facility to receive data requests from the client machine and locate and deliver the required data to the client in real time. The key part of this technology is that the GIS package must be Internet-aware, which means that the GIS program on the client side can read remote data as it does on a local disk and can understand data of various formats. To read remote data as a local disk does, protocols such as Network File System (NFS) for UNIX systems and distributed File System (dFS) for Windows systems need to be followed. Enabling the communication between different data formats is within the working goal of some organizations, including Open GIS. A further-empowered model will need more functions on both client and server side, which will probably request better communication among elements of the whole system. Common object request broker architecture (CORBA) can be introduced into the system to enhance communication, although progress in this area is limited.

Although the CIAA Internet GIS application is the highest and most powerful solution, some related technologies still warrant further exploration. Moreover, the requirements on the client side for both hardware/software and GIS expertise are strict. An efficient system is necessary to provide timely knowledge of data locator and transfer server about the location, maintenance, and availability of multiple data sources.

Conclusion and Discussion

The Internet is a good vehicle for implementing the combination of distributed computing technology with traditional GIS technology. Internet GIS, being a solution that can provide easy access to both information and GIS functions and allow various parties to share information and resources efficiently, is a valuable tool for enhancing health service access analysis so as to improve our health service delivery system. The three architectures proposed in this paper represent three basic solutions for employing Internet GIS in health service access analysis. They differ in that each solution is designed to satisfy different user needs and project objectives. For the majority of people who are more interested in getting general health service information such as locations, directions, and basic service categories, SAQD Internet GIS is efficient and effective enough. For users who are concerned about more details of the location factors and access situation of certain health service sites, IAIQ Internet GIS can provide more powerful service. This architecture, though, requires a relatively more powerful client machine as well as more patience and knowledge from the user in order to download

the GIS tool kit and to formulate queries. CIAA Internet GIS is the highest level and most powerful solution, which is sufficient for health service analysts, planners, development policy and strategy stipulators, and other public health organizations and agencies to analyze, plan, and evaluate the access situation of health service sites. Generally speaking, the more powerful the solution, the more resources both technically and financially are involved, and the more requirements are put on users for their special skills and techniques. As to which solution is the best, it depends on the specific situation of the project in terms of major users, available resources, and the main purpose of implementing the project.

As mentioned, some technical parts for CIAA Internet GIS are still immature. Actually, most efforts so far in trying to link traditional GIS with the Internet and distributed computing technology are in the experimental stage of developing both prototype and production service. Nevertheless, the development of Internet GIS will definitely benefit analysis of health service access greatly.

There are many other problems awaiting a solution before a real time Internet GIS is implemented. These problems could be institutional (both intra-institutional and inter-institutional), financial (investment for project development), and ethical and legal (data privacy and information accountability). But it is quite sure that by helping people from various social, economic, and cultural strata in their ability to access and process spatial information, Internet GIS could be an efficient way to alleviate the potential polarization between the information rich and poor in the coming information society (6).

Acknowledgments

This research is based on work supported by the National Science Foundation under Award No. SBR 9600465, "Project Varenius: NCGIA's Project to Advance the Science of Geographic Information." Support by NSF is gratefully acknowledged.

References

1. Gebbie KM. 1997. Community-based health care: An introduction. In: *Information networks for community health.* Ed. PF Brennan, SJ Schneider, E Tornquist. New York: Springer.
2. Bogue R, Hall Jr CH, ed. 1997. *Health network innovations: How 20 communities are improving their systems through collaboration*. Chicago: American Hospital Publishing, Inc.
3. Dowling AF. 1997. CHINs—The current state. In: *Information networks for community health.* Ed. PF Brennan, SJ Schneider, E Tornquist. New York: Springer.
4. University Consortium of Geographical Information Science (UCGIS). 1997. *Distributed computing.* White Paper. UCGIS, Leesburg, VA.
5. Peng Z, Beimborn E. 1998. Internet GIS and its applications in transportation. *Transportation Research News* March/April.
6. Lu Y, Lin G. 1996. Improving accessibility to spatial information on the Internet. In: *National Center for Geographic Information and Analysis technical report 96-10.* Compiled by H Couclelis.
7. Plewe B. 1997. *GIS online: Information retrieval, mapping, and the Internet*. Santa Fe, NM: OnWord Press.

GIS Analysis of Brain Cancer Incidence near National Priorities List Sites in New Jersey

Oleg I Muravov, MD, PhD,* Wendy E Kaye, PhD, C Virginia Lee, MD, MPH, Paul A Calame, MS, Kevin S Liske, MA
Agency for Toxic Substances and Disease Registry, US Department of Health and Human Services, Atlanta, GA

Abstract

This analysis was done to test the hypothesis that living close to hazardous waste sites included on the US Environmental Protection Agency's National Priorities List (NPL) might be associated with an increased risk for brain cancer. A total of 2,556 cases of primary brain cancer (code 191 of the *International Classification of Diseases* [ICD-9]) received from the New Jersey State Cancer Registry for the years 1986 through 1990, were geocoded using the Matchmaker 2000 address-matching program from Geographic Data Technology. Of the 2,556 cases, 178 (6.96%) could not be geocoded, 1 (0.04%) was found to be in another state, and 226 cases (8.84%) reported from death certificates only were excluded, leaving 2,151 cases (84.15%). The NPL sites in the state were mapped using a geographic information system (GIS), and 1-mile buffers were created around each of them. These areas were analyzed for excess brain cancer. Also, the average distance between cases and the nearest NPL site was determined. There were 177 cases (8.23%) within 1 mile of an NPL site. Using total population data from 112 NPL sites in New Jersey, there were 1,031,504 (13%) persons living within 1 mile of an NPL site. No elevated cancer incidence rates were found in the analyzed areas. Also, the sites were classified according to known off-site contamination. No statistically significant differences were found among either cases' age or distance from the nearest site in relation to the primary site contaminant. This analysis can be useful as a tool for developing more in-depth environmental health hypotheses.

Keywords: brain cancer, cancer epidemiology, National Priorities List (NPL) sites, spatial analysis, surveillance system

Introduction

The federal Agency for Toxic Substances and Disease Registry (ATSDR) is developing a surveillance system using cancer registry data from states to identify potential patterns between the occurrence of brain cancers in those states with US Environmental Protection Agency National Priorities List (NPL) sites and possible exposures to hazardous substances (1). Selected cancers is one of ATSDR's priority health conditions (2). Exposures to several chemicals that have been associated with an increased incidence of primary brain cancer (3) might be occurring at hazardous waste sites included on the NPL and a concern has been expressed about the rates of primary brain cancer around some of these sites. Initially, six states—California, Florida, Massachusetts, New York, Pennsylvania, and Virginia—were included in the project. Virginia was dropped,

* Oleg I Muravov, MD, PhD, Agency for Toxic Substances and Disease Registry, US Department of Health and Human Services, 1600 Clifton Rd., MS E-31, Atlanta, GA 30333 USA; (p) 404-639-5131, (f) 404-639-6219; E-mail: oim0@cdc.gov

however, because address data were not available, and New Jersey was added in response to a reported cluster of childhood brain cancer and leukemia in the Tom's River area. New Jersey brain cancer incidence data from 1986 through 1990 are used for this analysis. Census data for 1990 were used to obtain denominator data for census tracts and counties. Residence location at the time of diagnosis was used for the geographic analysis of cases. The analysis used year, age, residence at diagnosis, type of tumor, race, and sex of the cases.

Goal

The goal of the analysis was to test the hypothesis that living close to hazardous waste sites included on the NPL might be associated with an increased risk for brain cancer.

Objectives

The objectives of the analysis were:

- To compare the incidence rate of brain cancer among residents living within 1 mile of an NPL site with that of all New Jersey residents.
- To compare brain cancer incidences according to the off-site contamination of the nearest NPL site to find out whether there is an association between environmental contamination and brain cancer occurrence.

Data

Case Data Source

The Cancer Registry Program of the New Jersey Department of Health and Senior Services provided for analysis the street addresses of 2,556 cases of primary brain cancer (code 191 of the *International Classification of Diseases* [4]) diagnosed in New Jersey from 1986 through 1990.

NPL Sites in New Jersey

The information on NPL sites in New Jersey was obtained from the HazDat database on ATSDR's Internet home page (http://atsdr1.atsdr.cdc.gov:8080/hazdat.html). In particular, the HazDat Sensitive State Map (http://atsdr1.atsdr.cdc.gov:8080/hazusa1.html) was used to obtain the details of site location, chemical content, and extent of on- and off-site contamination. Where NPL sites were found to be within 1 mile of any of the primary brain cancers included in the analysis, the sites were classified by whether there was known off-site contamination.

Methods

Outline

For the purpose of this analysis, the incidence rate of brain cancer among New Jersey residents living within 1 mile of an NPL site was compared with that of all New Jersey residents. The observed number of cases was compared with the expected

number by the standardized incidence ratio (SIR) and its 95% confidence interval. Also, an additional analysis was implemented using cancer cases grouped by the nearest NPL site's off-site contamination to find out whether there was an association between environmental contamination and brain cancer occurrence. Student's t-test was used to assess the statistical significance of the difference between the above-mentioned groups of cases.

Geocoding

The file obtained from the New Jersey Cancer Registry Program was cleaned to ensure consistency of town and street names. The plus-4 codes were added to the zip codes of streets from the US Postal Service's *Zip+4 State Directory* for New Jersey (5). Street addresses were geocoded using the Matchmaker 2000 address-matching program from Geographic Data Technology (Lebanon, NH). Cases whose addresses were missing or were just post office box numbers were removed from the file before geocoding. The geocoded file was exported into a dBASE format to use in ARC/INFO (Environmental Systems Research Institute, Redlands, CA). The dBASE file was converted to an Info file and projected onto a coverage of New Jersey. Half-mile and 1-mile buffers were created around the NPL sites in the state (6). Cases falling within those buffers were extracted and a new file containing the sites within buffers was created. In addition, the NEAR command in ARC/INFO was run to determine the average distance between cases and the nearest NPL site. A data file was created with the cases, the nearest NPL site for each case, and the distance to that site.

Expected Number of Cancer Cases

The expected numbers of cancer cases were calculated for the 1990 population using the stratum-specific incidence rates observed in New Jersey (the standard population) during the period 1986 through 1990. The expected number of cases was calculated for each stratum first, as a product of the New Jersey incidence rate and the size of the stratum within 1 mile of an NPL site, and then summed over the strata. The number of cases expected annually from 1986 through 1989 was assumed to be equal to that estimated for 1990.

The Standardized Incidence Ratios Estimate

Standardized incidence ratios (SIRs) were used for quantitative analysis of brain cancer incidence in the 1-mile areas around the NPL sites. An SIR is calculated by dividing the observed number of cases by an expected number for the investigated population over the time period reviewed. The observed number of cancer cases for this analysis was provided by the New Jersey State Cancer Registry. The expected number of cancer cases was calculated using average annual State of New Jersey age- and sex-specific incidence rates from 1986 through 1990. The comparison rates were provided to ATSDR by the New Jersey State Cancer Registry. The lower and upper limits of the 95% confidence interval were calculated for each SIR using the Poisson distribution (7).

Results

There were 2,556 cases in the original data set with one duplicate case found. Of that total, 145 (5.7%) did not include any address information and 33 (1.3%) were post office

box numbers. Those 178 cases (7.0%) were removed from the file before address matching was done. An additional 83 cases had no street address and 3 cases had unidentifiable street addresses. After the address matching, there were 2,114 matches to street address (82.7%), 37 (1.5%) matched to the zip+4 centroid, 30 (1.17%) matched to the zip+2 centroid, and 195 (7.6%) matched to the zip code centroid. One of the zip code matches (0.04%) was removed when it was found to be in New York. Thus, 2,377 cases were in the final match for a match rate of 93%. In addition, 226 cases (8.84%) reported from death certificates only were excluded, leaving 2,151 cases (84.15%). Exclusion of the cases reported from death certificates only was based on the assumption that they were not primary brain cancer cases but likely were metastases. These 226 cases had an unspecified histological code 8000/3 ("malignant neoplasm," *International Classification of Diseases for Oncology, Morphology* [8]). In addition, 74% of these cases were 45 years of age or older, while 62% of them were 55 years of age or older. Exclusion of these cases from this analysis did not influence the findings because just 3 (1.32%) of them were within a half-mile of an NPL site. Another 22 cases (9.74%) were within 1 mile of an NPL site but were further than a half-mile, while 110 cases (47.8%) were further than 3 miles from an NPL site.

There were 177 cases (8.23%) within 1 mile of an NPL site and 54 cases (2.51%) within a half-mile of an NPL site. Using the total population data from 112 NPL sites in New Jersey, there were 1,031,504 persons living within 1 mile of an NPL site in the state. The total population of New Jersey was 7,730,188, so 13% of the total population lived within 1 mile of one of those 112 sites. The average distance between the nearest NPL site and any of the cases was 6,265.55±4,324.48 meters (3.89±2.69 miles). Of the 112 NPL sites in New Jersey, 58 (51.79%) were found to be within 1 mile of at least one brain cancer case included in the analysis.

The most frequent histologic types of cancer among the cases within 1 mile of an NPL site were glioblastoma multiforme and astrocytoma (39.53% and 37.85%, respectively). The rarest types were ependymoma (0.56%), medulloblastoma (1.13%), and meningioma (1.69%). No nerve sheath tumors were diagnosed in this population during the period studied (Table 1).

Table 1 Histologic Types of Brain Cancers among New Jersey Residents Living within 1 Mile of an NPL Site, by Year of Diagnosis

Histologic Type	ICD-O[a] Codes	1986	1987	1988	1989	1990	Total
Astrocytoma	(9400–9421)	9	10	16	13	19	67
Glioblastoma multiforme	(9440–9442)	12	11	17	12	18	70
Oligodendroglioma	(9450–9460)	1	—	1	2	—	4
Medulloblastoma	(9470–9472)	—	2	—	—	—	2
Ependymoma	(9391–9394)	1	—	—	—	—	1
Other gliomas	(9380–9383)	1	10	2	5	5	23
	(9422–9430)	—	—	1	1	1	3
Meningioma	(9530–9539)	2	—	1	2	2	7
Other brain cancers	—						
Total		26	33	38	35	45	177

[a] International Classification of Diseases for Oncology, Morphology (8)

The highest number of brain cancers within 1 mile of an NPL site was diagnosed among Caucasian males (61.02%), while the lowest was found among non-Caucasian females and males (1.13% and 3.95%, respectively).

Table 2 presents by age group the observed and expected numbers of cases, SIRs, and lower and upper limits of the 95% Poisson confidence interval within 1 mile of an NPL site. The expected numbers in this table were based on the incidence rates observed in New Jersey from 1986 through 1990. The values for each age group, as well as for all ages combined, were smaller than expected.

Table 2 Standardized Incidence Rates (SIRs) and 95% Confidence Intervals (CI) for Brain Cancer, 1 Mile from NPL Sites, New Jersey, 1986–1990

	Number of Cases			
Age Groups	**Observed**	**Expected**	**SIR**	**95% CI Lower–Upper**
0–14	19	25	0.76	0.46–1.20
15–44	41	60	0.68	0.49–0.93
45–64	50	83	0.60	0.45–0.75
65+	67	109	0.62	0.48–0.78
Total	177	277	0.64	0.55–0.74

Table 3 shows distribution of the cases within 1 mile of an NPL site according to the primary contamination at the closest NPL site. A total of 51 (88%) NPL sites had known off-site contamination. Of those, 30 sites (52%) were contaminated by volatile organic compounds (VOCs) and 21 sites (36%) were contaminated by metals, polychlorinated biphenyls (PCBs), or radiation. A total of 143 cases (81%) were found within 1 mile of an NPL site with known off-site contamination. Of those, 74 cases (42%) were found in proximity to VOC-contaminated sites and 69 (39%) were in proximity to sites characterized by other contaminants.

Table 3 Brain Cancer Cases within 1 Mile of NPL Sites, by Primary Off-Site Contamination, New Jersey, 1986–1990

	Known Off-Site Contamination				
	VOCs	**Metals**	**PCBs**	**Radiation**	**No Known Off-Site Contamination**
Cancer cases	74	26	15	28	34
NPL sites	30	13	2	6	7

Discussion

This analysis did not indicate that residence near an NPL site in New Jersey at time of diagnosis increased the incidence of brain cancer. The observed numbers of brain cancer within a 1-mile radius of an NPL site were lower than expected in total and in each age category. Also, histologic types of the brain cancers diagnosed in these residents

and their age, sex, or racial distributions did not differ from those of other New Jersey residents. However, there are many limitations in this type of analysis that should be considered when interpreting these results. One of the major limitations in projects involving GIS methods is the quality of the geocoding of cases. Typical address matching rates range from 20% up to 95% for rural states (9). This particular investigation had an extraordinary geocoding rate of 93%. At the same time, however, a group of 178 cases (6.96%) whose addresses were missing or just post office box addresses were removed from the dataset prior to analysis. In addition, addresses of 261 cases (10%) were incomplete, so they were geocoded to either zip+ or zip code centroids. Given the relatively small number of cases, this could have had an impact on the findings. In particular, there is a possibility that some of these cases were within a 1-mile buffer zone around an NPL site but were excluded because of geocoding errors. Another limitation in this analysis was the use of addresses available only at the time of diagnosis. Such information might not have reflected where a person got his or her exposure due to a latency period in the development of cancers and the high mobility in the US population. Should these issues be resolved, an association between living close to an NPL site and brain cancer occurrence could be either stronger or weaker than was found.

Also, the sites were classified and analyzed by known off-site contamination. A limited number of the brain cancer cases lived within 1 mile of an NPL site in New Jersey (177 cases, 8.23%), making it impossible to look for associations with specific chemicals or agents, such as ionizing radiation, and forcing investigators to group them into VOC and non-VOC cases with near equal numbers of cases in each of the groups. No statistically significant differences were found among cases' age, histological type of tumor, or distance from the nearest site in relation to the primary site contamination. No differences were found either when comparing the cases within the area of sites with known off-site contamination versus those with unknown off-site contamination.

It should be noted that the overall impact of residential proximity to NPL sites is unknown. No clear association has been found between health effects in humans and hazardous waste sites either (10,11,12). Overall, small sample size, lack of individual exposure data, poor hazardous site selection for analysis, and inappropriate health effects for the toxic substances being studied could have led to negative findings in some cases, as well as possible erroneous positive findings (11). The 1-mile radius buffer zone was chosen for this analysis as the smallest geographic area (with the shortest proximity to possible sources of exposure) in which the number of cases was large enough to provide measurable statistical power for analysis of such a rare health event as brain cancer. At the same time, estimation of relative risk (brain cancer incidence rates within a half-mile versus 1 mile) could be useful and prove a valuable addition, as could comparison of local rates to the state and national cancer rates. The small number of brain cancers diagnosed within a half-mile of NPL sites in New Jersey (54 cases, 2.5%) made it impossible to implement this approach for this particular analysis. However, it should be considered for future investigation when brain cancer incidence data from several states are available.

References

1. Shy C, Greenberg R, Winn D. 1994. Sentinel health events of environmental contamination: A consensus statement. *Environmental Health Perspectives* 102(3):316–7.

2. Lybarger JA, Spengler RF, DeRosa CT. 1993. *Priority health conditions: An integrated strategy to evaluate the relationship between illness and exposure to hazardous substances.* US Department of Health and Human Services. July. NTIS publication #PB93-203529.
3. Inskip PD, Linet MS, Heineman EF. 1995. Etiology of brain tumors in adults. *Epidemiologic Reviews* 17(2):382–414.
4. World Health Organization (WHO). 1977–78. *Manual of the international statistical classification of diseases, injuries, and causes of death: Based on the recommendations of the Ninth Revision Conference, 1975, and adopted by the Twenty-Ninth World Health Assembly*. Geneva: WHO.
5. US Postal Service. 1994. *Zip+4 state directory.* Washington, DC: US Postal Service.
6. Briggs DJ, Elliott P. 1995. The use of geographical information systems in studies on environment and health. *World Health Statistics Quarterly* 48:85–94.
7. Breslow NE, Day NE. 1987. *Statistical methods in cancer research: Vol. II—the design and analysis of cohort studies.* Ed. E Heseltine. Lyon, France: International Agency for Research on Cancer. IARC Scientific Publication #82.
8. Berg J, Maguin P, Muir C, et al. 1987. *International classification of diseases for oncology, morphology.* Ed. C Persy, VV Holten. Field Trial Edition. Lyon, France: International Agency for Research on Cancer.
9. Vine MF, Degnan D, Hanchette C. Geographic information systems: Their use in environmental epidemiologic research. *Environmental Health Perspectives.* In press.
10. Dayal H, Gupta S, Trieff N, Maierson D, Reigh D. 1995. Symptom clusters in a community with chronic exposure to chemicals in two Superfund sites. *Archives in Environmental Health* 50(2):108–11.
11. Johnson BL. 1995. Nature, extent, and impact of Superfund hazardous waste sites. *Chemosphere* 31(1):2415–28.
12. Najem GR, Strunck T, Feuerman M. 1994. Health effects of a Superfund hazardous chemical waste disposal site. *American Journal of Preventive Medicine* 10(3):151–5.

Remote Imaging Applied to Schistosomiasis Control: The Anning River Project

Edmund Y Seto (1),* Don R Maszle (1), Robert C Spear (1), Peng Gong (1), Byron Wood (2)
(1) University of California, Berkeley, CA; (2) NASA Ames Research Center, Moffett Field, CA

Abstract

Landsat Thematic Mapper (TM) imagery was used to identify habitat suitable for *Oncomelania hupensis*, the snail vector for schistosomiasis in the Anning River Valley in Sichuan, China. The location of 55 snail habitat sites and 48 non-habitat sites were determined by GPS measurements. Landsat TM data were found to be quite variable for both snail and non-snail sites. Because of this, supervised maximum likelihood classification produced poor accuracy. It was hypothesized that the variability was due to the existence of multiple microenvironments, each with distinct spectral properties and each suitable as snail habitat. A two-tiered classification approach was developed in which an unsupervised classification was first performed for the snail and non-snail habitat data to generate five snail and five non-snail clusters. The signatures of the 10 clusters were then used to perform maximum likelihood classification. Using this approach, 90.3% of the snail habitat and 86.6% of the non-habitat were correctly identified. These results suggest that remote sensing may be an effective tool for identifying the habitat of the schistosomiasis vector in China. If so, this provides a surveillance method for studying the area affected by the new Three Gorges Dam, where profound ecological change will occur and schistosomiasis is predicted to become a major problem.

Keywords: schistosomiasis, remote sensing, Landsat, China, *Oncomelania hupensis*

Introduction

The use of satellite imaging to remotely detect areas of high risk for transmission of infectious disease is an appealing prospect for large-scale disease monitoring. The detection of large-scale environmental determinants of disease risk, often called landscape epidemiology, has been motivated by several authors (1,2). The basic notion is that large-scale factors such as population density, air temperature, hydrological conditions, soil type, and vegetation can determine in a coarse fashion the local conditions contributing to disease vector abundance and human contact with disease agents. These large-scale factors can often be remotely detected by sensors or cameras mounted on satellite or aircraft platforms and can thus be used in a predictive model to mark high-risk areas of transmission and to target control or monitoring efforts. A review of satellite technologies for this purpose was recently presented by Washino and Wood (3), Hay (4), and Hay et al. (5).

In China, there is currently concern about the establishment and spread of infectious diseases, including malaria and schistosomiasis, in the area along the Yangtze

* Edmund Y Seto, University of California-Berkeley, EHS, School of Public Health, 140 Warren Hall, University of California, Berkeley, CA 94720 USA; (p) 510-649-8152; E-mail: edmund@sparky.berkeley.edu

upstream of the Three Gorges Dam, which is now under construction. Our group has been working with parasitologists from the Sichuan Institute of Parasitic Disease (SIPD) responsible for schistosomiasis monitoring and control in the area of the dam. The profound ecological and social changes that will take place as the dam is being constructed and after its completion may create new habitat for the snail species central to the cycling of the disease, as well as new relationships between humans, domestic animals, and the aquatic environment. The size of the lake that will be created behind the dam and the difficulty of access to this mountainous area make remote sensing technology an attractive adjunct to land-based surveillance of these changes as the lake fills and the dam goes into operation.

To explore the possible use of remote sensing in schistosomiasis control prior to the completion of the Three Gorges Dam, we have been studying a region where the disease is endemic, where ground-based data sets on disease prevalence and snail habitat exist, and which is of a scale suitable for study using remote sensing. With the assistance of our colleagues in the SIPD, we have focused on the area along the Anning River in the Daliang mountainous area of southwestern Sichuan Province. This region includes villages studied in our earlier work.

Remote sensing has been demonstrated to be a viable means of identifying habitat for vectors of other diseases. The potential efficacy for using remote sensing to determine high-risk areas of malaria transmission was recently illustrated (6,7). Two types of Anopheline mosquito habitat—unmanaged pastures and transitional swamps—were shown to be detectable based on classification of Landsat Thematic Mapper (TM) data. That research was an extension of previous work that focused on the identification of high and low Anopheline-producing rice fields (8). Landsat TM data have also been used to map land cover to study landscape correlates of Lyme disease (9). In that study, disease data and landscape classifications were overlaid to look for land cover correlates to disease risk.

Several studies have implied that remote sensing could be a useful tool for schistosomiasis monitoring. Cross and Bailey (10) and Cross et al. (11) showed a correlation between local temperature variation and prevalence rate. Malone et al. (12) showed that historical prevalence data correlated well with remotely detectable geographic features. Both of these studies took a different approach from the Anopheline studies in that they demonstrated a correlation between disease and ecological factors, whereas the malaria vector studies by Beck et al. (6,7) and Wood et al. (8), use remote sensing to identify habitat correlated with the presence of the disease vector.

In the current study, we ask if the second approach is applicable to detecting spatial variations in the vector population that transmits the parasite causing schistosomiasis japonicum, the Asian form of schistosomiasis. The disease vector, or, more appropriately, the intermediate host for schistosomiasis japonicum, is an amphibious snail, *Oncomelania hupensis*. A recent preliminary study by the SIPD used Advanced Very High Resolution Radiometer (AVHRR) data to identify snail habitat (13). In the current analysis we use higher resolution Landsat TM data to look for correlations with detailed ground-based snail ecology surveys. If surveyed snail habitats correlate with the satellite data, there is the potential to use remote sensing to monitor large and remote areas in the region of the dam, and to identify areas at high risk of transmission.

The current problem is different from that of detecting malaria vectors. The vector habitat for *O. hupensis* is usually a microenvironment that is itself not detectable using

most remote sensing data because of their coarse spatial resolution. However, microenvironmental conditions may be affected by larger-scale factors including local vegetation type and surrounding crops, fertilizer usage, and water and temperature patterns. These factors will cause local changes in the environment, which in turn will influence the remote sensing signal. Further, the other two schistosomiasis studies found correlations between large-scale phenomena and disease rates, implying that something can be seen at this scale. The question addressed at present is whether remote sensing data of local areas can be accurately classified, based on large-scale environmental factors, so as to identify habitats that are suitable for these vector snails, and thus at high risk for transmission.

Methods

To address this issue, our group conducted a study in the Anning River Valley in southwestern Sichuan Province. The Anning River Valley is a high mountain valley at an elevation of about 1,500 meters (m). This is primarily an agricultural area with irrigated farming of rice, corn, wheat, a variety of vegetables, and some export crops. The valley is also a highly endemic area for schistosomiasis japonica. The remote sensing data used were from the Landsat TM sensor. The ground data indicating suitable snail habitat were point observations from one environment type and were classified as habitat or non-habitat. Suitability was determined by the presence or absence of young or reproducing snails. Few locations are found with only adult snails present, presumably because snails leave unsuitable locations or die.

A large-scale snail monitoring effort was conducted in 1994 by the Xichang County Anti-Endemic Station (XCAS). The station is responsible for monitoring and controlling human schistosomiasis infection and vector snail ecology in the 17-township middle section of the Anning River Valley. Snail surveys were performed throughout the area in townships where the human incidence exceeded 10%. Snail surveillance was done in June. We chose this section of the river valley as our study area to take advantage of these existing surveillance data. The study area extends from Lizhou Township in the north to Hexi Township in the south, and covers about 45 km of the river valley around Xichang City. A map of the area showing these reference points is shown in Figure 1.

Two Landsat TM scenes (one spring, April 7, 1994, and one fall, October 16, 1994) were obtained for the region. Both images were free of cloud cover over the area of interest, and each represents a distinct agricultural season. The major crops during these times are rice and corn in summer-fall and wheat and beans in the winter-spring season.

Ground data on the locations of snail colonies were obtained from the XCAS's 1994 snail surveys (this is being supplemented with density information). For 10 days in the middle of June 1997, our group, with the help of the local authorities and the head of the XCAS, visited townships and recorded the geographic locations of the 1994 surveillance data. Collection sites were located with a Trimble Pro XL global positioning system (GPS) to allow for correlation with the remote sensing data. Three base stations were established and positioned with respect to a known surveyed control point at the peak of the Lushan mountain southeast of Xichang City. All data points were differentially corrected to the base station locations to provide positioning accuracy in the 1 to 5 m range.

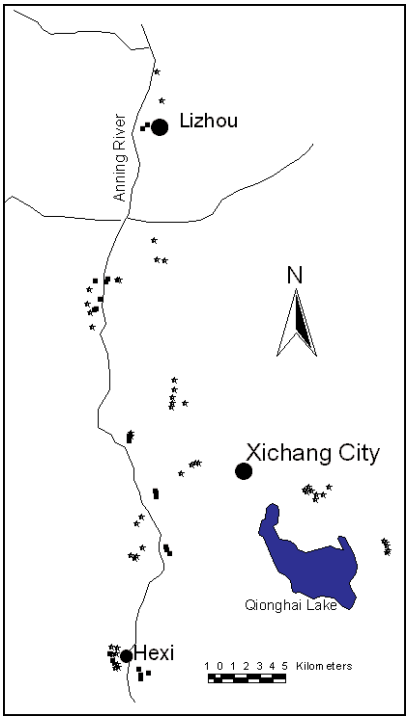


Figure 1 Map of the Anning River Valley study area. Points represent snail habitat and non-habitat sites distributed in the valley from Lizhou Township in the north to Hexi Township in the south.

Collection sites were located in 14 townships throughout the study area. Townships were chosen based on availability of 1994 data or if there was historical knowledge of apparently stable snail habitat or non-habitat. Three environment types exist in the study area: irrigated farming in the river plain, terraced rice culture at the base of the hills, and mountain stream areas higher in the mountains. The three habitat types are structurally different with distinct local ecologies. In light of this, the study was limited to one type of environment, irrigated farming areas in the river plain, for which there was an abundance of ground/field data (and travel was more convenient). Snail habitat in the river plain area is limited to irrigation and drainage ditches and the boundaries of fields. This resulted in a total of 103 data points (55 classified as habitat and 48 as non-habitat).

Image processing was performed using PCIWORKS image processing software. Before data analysis, the images were geometrically corrected and registered using 11 ground control points taken throughout the river valley. Points used for referencing the image to a world coordinate system were large structures easily seen on the image, such as the corners of the Xichang airport runway, large intersections, and an isolated paved village compound. The 103 ground/field data points were located on the image. Each snail habitat and non-habitat site was specified as a 3- by 3-pixel area surrounding the site location as determined in the field by GPS measurements.

After geographic correction, a preliminary supervised maximum likelihood classification was performed using all TM channels from both dates. The 55 habitat and 48

non-habitat areas were used both to train the classification algorithm and assess the accuracy of the classification. The results of this accuracy assessment are presented in the next section.

Realizing that the accuracy of our preliminary classification was inadequate, we next employed a two-tiered analysis approach. The first step of this approach employed an unsupervised classification method called Isodata clustering to break up snail habitat and non-habitat classes into subclasses. The Isodata algorithm is an iterative process whereby the pixels of the image are grouped into clusters based on an examination of their multispectral brightness values. Pixels grouped into the same cluster have similar spectral properties. The Isodata algorithm was first applied to those pixels corresponding to snail habitat sites. The algorithm was used to categorize the pixels into five separate clusters. These five snail habitat clusters may correspond to different microhabitats that are all suitable for snails. The Isodata algorithm was then run using the non-habitat sites to produce five non-habitat clusters. The spectral distributions for each of these 10 clusters were determined and used to perform the second part (i.e., the supervised maximum likelihood classification) of this two-tiered analysis.

Results

The results of the preliminary supervised classification using all TM bands from the spring and fall images are presented in Table 1. For the 55 snail habitat sites, there was good classification accuracy, with 89.3% of the pixels being classified correctly. However, for the non-habitat sites there were many misclassified pixels, with only 52.3% of the pixels being accurately classified as non-habitat. Among unclassified pixels, 3.4% of them corresponded to snail habitat sites and 8.8% of them corresponded to non-habitat sites.

Table 1 Results of Preliminary Maximum Likelihood Classification of Snail Habitat and Non-Habitat Sites

	Total # Pixels	% Unclassified Pixels	% Classified as Snail Habitat	% Classified as Non-Habitat
48 Non-habitat sites	432	8.8	38.9	52.3
55 Snail habitat sites	495	3.4	89.3	7.3

Table 2 shows the result of the two-tiered classification. For the pixels corresponding to the 55 snail habitat sites, 3.6% were unclassified. Of the remaining 96.4%, 90.3% of the pixels were correctly classified as snail habitat. For the pixels corresponding to the 48 non-habitat sites, 4.2% were unclassified. Of the remaining 95.8%, 86.6% of the pixels were correctly classified as non-habitat. Table 3 presents a classification matrix showing the percentages of each cluster for both types of habitat.

The resulting classification for the Anning Valley is shown in Figure 2. A 5- by 5-pixel mode filter was applied to the image for presentation. The mode filter is primarily used to clean up thematic maps for presentation purposes by grouping together areas that are predominantly snail habitat or non-habitat. More specifically, for each 5- by 5-pixel area, the predominant class is assigned to all pixels in the area.

Table 2 Results of Two-Tiered Analysis Using Isodata and Maximum Likelihood Classification Algorithms

	Total # Pixels	% Unclassified Pixels	% Classified within Snail Habitat Clusters	% Classified within Non-Habitat Clusters
48 Non-habitat sites	432	4.2	12.6	83
55 Snail habitat sites	495	3.6	87.1	9.2

Table 3 Percentage of Pixels Classified by Cluster for Snail Habitat and Non-Habitat Sites

			Snail Habitat Clusters					Non-Habitat Clusters				
	Total # Pixels	% Unclass-ified Pixels	% c1	% c2	% c3	% c4	% c5	% c6	% c7	% c8	% c9	% c10
48 Non-habitat sites	432	4.2	6.9	3.7	0.2	1.6	0.2	28.9	9.0	18.3	11.8	15.0
55 Snail habitat sites	495	3.6	31.3	23.6	8.5	17.6	6.1	4.8	0.0	4.0	0.0	0.4

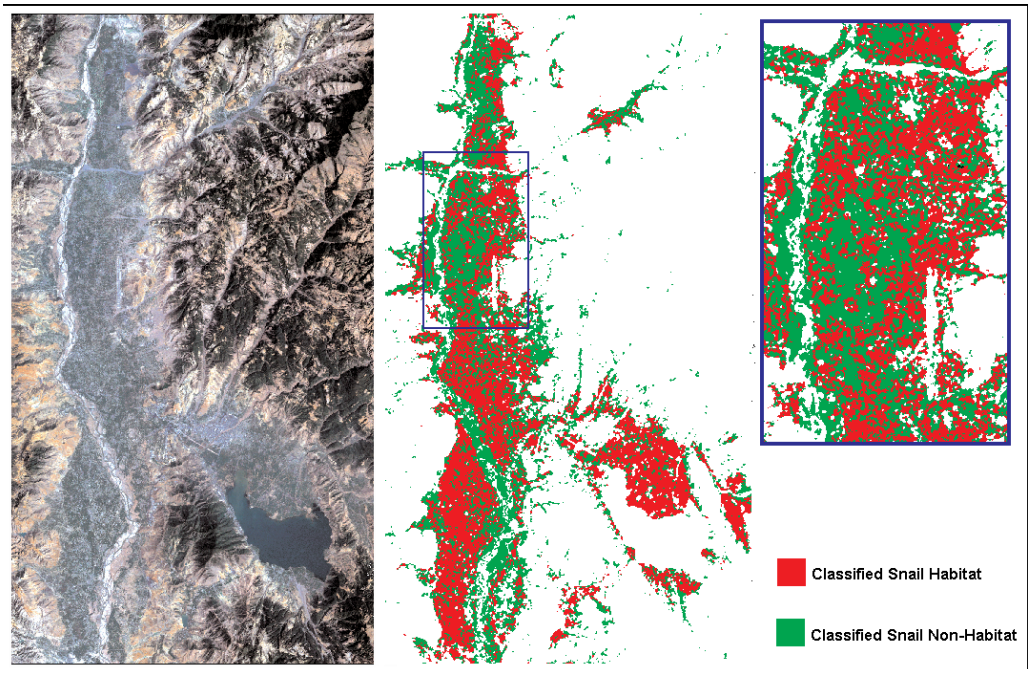


Figure 2 Three panels showing (from left) (a) Landsat TM of Anning river valley, (b) classification of habitat using Isodata and maximum likelihood algorithms; (c) enlargement of valley floor showing mixed habitat.

Discussion

Despite the fact that we limited our analysis to only those sites that were in the irrigated farming areas located in the river plain, there was a great deal of variability within the snail habitat and non-habitat sites. This was observed visually in the field as well as in the distributions of the spectral data. Our preliminary classifications ignored this variability by lumping all of the habitat sites together and all of the non-habitat sites together to train the classification. As a result, the snail habitat class included many of the non-snail sites, while the non-habitat class did not classify enough of the non-snail sites. This poor classification may be due to the existence within the irrigated farming environment of multiple microenvironments/habitats that each have distinct spectral properties. Hence, the terms "snail habitat" and "non-habitat" encompass distinctly different microenvironments that support or do not support snails, respectively. Therefore, when either snail habitat or non-habitat is considered as a whole, it appears to be quite variable.

In the two-tiered approach, we solved the problem of multiple microenvironments by using the Isodata algorithm to effectively separate the highly variable habitats into relatively 'pure,' less variable clusters before performing supervised classification. This was not based on field observation; rather, the spectral data were used to create these clusters. The choice to create five habitat clusters and five non-habitat clusters is not explained in detail because these numbers were chosen somewhat arbitrarily. The high classification accuracy, however, indicates that such numbers are not unreasonable. It will not be hard to fine-tune the number of clusters by looking at the variability and separability between signatures.

In addition to refining the number of clusters, we are also working on reducing the number of bands used to include only those that add information to the classification. Once we have reduced the classification down to the key bands, we hope to develop an understanding of what the clusters correspond to in the field.

Future Work

Although our results thus far are quite promising, we acknowledge that they are still very much preliminary in nature. There is considerable work to be done before the methods described here can be applied to the field in the form of disease surveillance. One of our first aims is to develop a better sense of the accuracy—and, thus, limitations—of remote sensing classifications applied to schistosomiasis.

In our current work we assessed the accuracy of the classification only at the locations of the training sites. Using the same data for the training and validation of the classification may have resulted in artificially high accuracies. We plan to revisit the Anning River Valley to validate our two-tiered analysis with a rigorous field study. For this field study we first intend to obtain spring and fall Landsat TM images from a more recent year than 1994 to repeat the two-tiered classification. This more recent classification would then be validated in the field. We intend to sample pixels from this more recent classification and visit the corresponding locations in the field. True snail status would be assessed for each location and a better assessment of classification accuracy would be produced.

Another study we plan to conduct will assess the degree to which additional ground data might improve the classification accuracy. According to SIPD (14), the

ecological correlates of *O. hupensis* snail habitat in Sichuan include the existence of certain vegetation types; size and density of irrigation ditches; proximity of agricultural field edges; wet lowland areas; soil moisture, type and quality; and, local temperature. In addition, it is our hypothesis that to some degree the chemical properties of the soil and water condition the existence of snails at a particular site. Some information, such as temperature, soil type, and vegetation type and coverage are available at a coarse scale for the Anning Valley, while other data, particularly the soil and water chemistry data, will have to be measured in the field during our randomized field validation study.

Several issues will have to be addressed when dealing with such multivariate data that are measured on several different scales and with varying reliabilities. For example, soil type data is a nominal variable and percent vegetation coverage, a bounded, interval variable. Because traditional remote sensing image analysis algorithms such as the maximum likelihood classifier cannot be used to process nominal and ordinal data, we will analyze these additional ground data using several non-traditional techniques: CART (15), logit regression (16), evidential reasoning (17,18), and artificial neural algorithms (19). Each of these algorithms can handle all the different levels of measurements and have proven useful in classification tasks where similar issues existed.

These multivariate approaches can be used to develop a classification for snail habitat based on ecology. The accuracy of this ecological classification can be compared with that of our remote sensing classification algorithm to gauge the added importance of incorporating ground ecology measurements in our classification of snail habitat. Furthermore, the ecological classification will help in developing an ecological interpretation of the remote sensing classification algorithm, which is central to being able to extrapolate the use of the algorithm to different areas and to different snail subspecies. Of particular interest is the determination of whether the distinct habitat clusters identified in our remote sensing classification correspond to distinct ecological conditions in our ecological classification, and later, how both the remote sensing and ecological classifications change between different schistosomiasis-prevalent regions in China.

The work described thus far has focused on locating snail habitat. Although the existence of snails is a necessary criterion for disease transmission, it does not serve as an accurate indication of disease prevalence since, within areas where snails exist, the extent of human and animal infection vary considerably. Moreover, in some locations where snails exist, no disease transmission occurs at all. It is clear, however, that on a local scale, infection intensity and disease prevalence are related to the relationships between people, animals, and snails, as they may be mediated by landscape features. Alongside our remote sensing work, we have been working with mathematical models as a way to better understand such site-specific factors at the local level and their impact on the dynamics of disease transmission.

From a remote sensing standpoint, however, many of the landscape features that are related to infection intensity—including the nature and density of irrigation in villages, and the proximity and density of settlements—can be identified and quantified using remote sensing technologies. In addition, topographical features such as slope and aspect determine the flow of water channels, which may influence the transmission of disease. Therefore, it is of considerable interest to determine if topographical or landscape features that can be determined remotely are correlates of transmission. Such

information would further inform remote surveillance programs for prioritizing locations within the Three Gorges region for intensive ground investigation. To investigate these questions, higher resolution images than those from Landsat TM would be necessary. A future study will look at aerial photographs and/or higher resolution satellite images, such as those from SPOT HRV-PAN and IRA-1D, which are available now, and from Space Imaging and Earth Watch, which might become available soon.

Acknowledgments

This project has been partially supported by a NASA-Ames Joint Research Interchange (NCC2-5102) and a University of California Pacific Rim Research Grant.

References

1. Pavlovsky E. 1966. *The natural nidality of transmissible disease.* Urbana, IL: University of Illinois Press.
2. Meade M, Florin J, Gesler W. 1988. *Medical geography.* New York: The Guilford Press.
3. Washino R, Wood B. 1994. Application of remote sensing to arthropod vector surveillance and control. *American Journal of Tropical Medicine and Hygiene* 50(6):134–44.
4. Hay S. 1997. Remote sensing and disease control: Past, present and future. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 91:105–6.
5. Hay S, Parker M, Rogers D. 1997. The impact of remote sensing on the study and control of invertebrate intermediate hosts and vectors of disease. *International Journal of Remote Sensing* 18(14):2899–2930.
6. Beck L, Rodriguez S, Dister A, Rodriguez A, Washino R, Roberts D, Spanner M. 1997. Assessment of a remote sensing based model for predicting malaria transmission risk in villages of Chiapas, Mexico. *American Journal of Tropical Medicine and Hygiene* 56(1):99–106.
7. Beck L, Rodriguez M, Dister S, Rodriguez A, Rejmankova E, Ulloa A, Meza RA, Roberts D, Paris J, Spanner M, Washino R, Hacker C, Legters L. 1994. Remote sensing as a landscape epidemiologic tool to identify villages at high risk for malaria transmission. *American Journal of Tropical Medicine and Hygiene* 51(3):271–80.
8. Wood B, Washino R, Beck L, Hibbard K, Pitcairn M, Roberts D, Rejmankova E, Paris J, Hacker C, Salute J, Sebesta P, Legters L. 1991. Distinguishing high and low anopheline-producing rice fields using remote sensing and GIS technologies. *Preventive Medicine* 11:277–88.
9. Dister S, Beck L, Wood B, Falco R, Fish D. 1993. *The use of GIS and remote sensing technologies in a landscape approach to the study of Lyme disease transmission risk.* GIS '93 Symposium, Vancouver, BC.
10. Cross ER, Bailey RC. 1984. Prediction of areas endemic for schistosomiasis through use of discriminant analysis of environmental data. *Military Medicine* 149(1):28–30.
11. Cross ER, Sheffield C, Perrine R, Pazzaglia G. 1984. Predicting areas endemic for schistosomiasis using weather variables and a Landsat data base. *Military Medicine* 149(10):542–4.
12. Malone J, Huh O, Fehler D, Wilson P, Wilensky D, Holmes R, Elmagdoub A. 1994. Temperature data from satellite imagery and the distribution of schistosomiasis in Egypt. *American Journal of Tropical Medicine and Hygiene* 50(6):714–22.
13. Li Z, Yuan P, Yin R, He S, Gu X, Zhao W, Xu F. 1990. Identification of distribution area of *Oncomelania* by remote sensing technique. *Acta Scientiae Circumstantiae* 10(2).

14. Gu X. 1995. *Report of national key project of the 8th five-year plan.* Chengdu, China, Sichuan Institute of Parasitic Disease.

15. Breiman L, Friedman J, Olshen R, Stone C. 1984. *Classification and regression trees.* Monterey: Wadsworth and Brooks/Cole.

16. Chung C, Moon W. 1991. Combination rules of spatial geoscience data for mineral exploration. *Geoinformatics* 2:159–69.

17. Wang Y, Civco D. 1994. Evidential reasoning-based classification of multisource spatial data for improved land cover mapping. *Canadian Journal of Remote Sensing* 20(4):381–95.

18. Gong P. 1996. Integrated analysis of spatial data from multiple sources: using evidential reasoning and an artificial neural network for geological mapping. *Photogrammetric Engineering and Remote Sensing* 62(5):513–23.

19. Gong P, Pu R, Chen J. 1996. Mapping ecological land systems and classification uncertainty from digital elevation and forest cover data using neural networks. *Photogrammetric Engineering and Remote Sensing* 62(11):1249–60.

A Conceptual Model of the Spread of Rabies That Integrates Computer Simulation and Geographic Information Systems

Lorinda L Sheeler-Gordon,* Kenneth R Dixon
The Institute of Environmental and Human Health, Texas Tech University, Lubbock, TX

Abstract

Rabies, a viral infection of the central nervous system, is transmitted by direct contact with an infectious individual and is considered to be a predominately zoonotic disease. Epidemic models have focused on the spatial spread of rabies, and emphasized the importance of understanding the transmission and spread of the disease. This conceptual model concentrates on the geographic spread and transmission of rabies in raccoon populations. A stochastic, individual-based model that incorporates probabilities of contact between groups of the population will be integrated into a geographic information system (GIS) using the ARC/INFO macro language. We anticipate that the integration of computer simulation and GIS will assist in the development of an epidemic model predicting the geographic spread of rabies. The model was designed to investigate disease control scenarios, such as optimizing the placement of rabies oral vaccine to impede the further spread of disease.

Keywords: modeling, simulation, rabies, raccoon, epidemiology

Introduction

Rabies, a viral infection of the central nervous system, is transmitted by direct contact with an infectious individual and is considered to be predominantly a zoonotic disease. Rabies in wildlife such as bats, foxes, skunks, and raccoons occurs when the population of animals reaches a threshold density; hence, transmission is achieved by direct contact. If a human has contact with a rabid animal, rabies becomes a disease that affects humans. Although the incidence of rabies in humans is rare (with only a few deaths per year in the United States), it is a horrifying disease for which there is no known case of recovery after the onset of clinical symptoms (1).

The raccoon, *Procyon lotor*, is considered a major wildlife reservoir of rabies in the eastern United States and is currently spreading its distribution as a vector of the disease. In the 1950s, the first outbreak of raccoon rabies in the United States occurred in Florida, and the number of reported rabid raccoons is continuing to increase (2).

Raccoons are considered a solitary species. Home ranges of adult females overlap broadly and there is no evidence of territoriality, whereas home ranges of adult males overlap less than 10% with adjacent adult males. Males of neighboring home ranges were found to have a separation of at least 2 kilometers (km) (3,4). Barash (5) reports data on captured raccoons from neighboring areas and from widely separated areas. The interactions of the captured raccoons indicate some degree of neighbor recognition. The raccoons from widely separated areas exhibit hostile behavior toward each

* Lorinda L Sheeler-Gordon, Texas Tech University, The Institute of Environmental and Human Health, PO Box 41163, Lubbock, TX 79409-11 USA; (p) 806-885-4567; E-mail: lsheeler@ttu.edu

other, while neighboring individuals show tolerance for each other. These data provide evidence that males show territoriality only with other males (3,4,5). Adult males move around more than adult females, and males generally have a larger home range (3,4). Home range diameters have been reported as measuring 1 to 3 km, with suburban populations having smaller home ranges of 0.3 to 0.7 km (3). The home ranges of adult males and females overlap, but individuals usually remain apart by mutual avoidance, except during the mating period. The only groups having been reported together are family groups, communal winter dens, and those inhabiting areas of abundant food (3,6).

A low density of raccoons is considered to be 5 individuals per square kilometer (individuals/km^2). High densities have been reported up to 20 individuals/km^2. In suburban areas, densities of as many as 68.7 individuals/km^2 have been reported. Ellis (7) radiotracked seven individual raccoons and concluded that when densities were high, raccoons seemed to move less and had smaller home ranges. In areas of high densities, raccoons were distributed evenly throughout all habitats. As densities decreased, however, the distribution favored particular habitats (4).

Raccoon long-range movements can be observed when juveniles disperse, or when environmental conditions and food availability are unfavorable. Reported data have much variability. Distances of 121 km (8), 266 km (9), and 254 km (10) have been recorded for adult raccoons. Butterfield (11) reported a maximum distance of 1.6 km, but with an average of 0.6 km. Distances for juvenile dispersal ranged from a few kilometers up to a maximum of 20 km.

Raccoons are classified as highly susceptible to rabies infection based on an intramuscular injection of the LD50 value (12). The LD50 for raccoon intraspecies transmission is 3.9 virus per inoculum (13). Because the virus concentration in saliva varies among individuals, the dosage transmitted when in contact with another individual varies, influencing the course of infection. An experimental inoculation of the LD50 value in a single dose by intramuscular injection, along with an exposure site of the neck or hind leg, is the best estimate of natural infection. It is noted that experimental inoculation cannot capture all the possibilities of natural exposure.

Exposure can induce the production of antibodies. Not all animals naturally exposed to rabies die of the disease. Field investigations of raccoons collected during an epizootic and observed for the following two years resulted in eight of ten raccoons surviving (14). Virus-neutralizing substances have been induced experimentally and have conferred resistance to further, massive inoculations of the rabies virus (13). Immunity or resistance to the rabies virus seems to play a role in naturally occurring raccoon populations.

Raccoons have been found during the day in residential areas and found with pet animals, resulting in increased human contact. Consequently, every year many people receive post-exposure treatment for exposure to wild animals. To minimize the number of human exposures and control and prevent the further spread of rabies, epidemic rabies models have focused on the spatial spread of the disease. Previous rabies models have focused on the fox as the main vector of the disease. The model presented here focuses on the raccoon as the main rabies vector.

The long incubation period of rabies in raccoons is a major factor in its transmission cycle. Therefore, animal movement was explored as a factor of transmission. The transmission cycle has two possible pathways. The first pathway is the migration of the

healthy raccoon that carries rabies to an uninfected geographic region. After the incubation period, the raccoon becomes infectious, experiences clinical symptoms, and spreads the disease to this new region. The second pathway is the abnormal behavior of the infectious raccoon whose confused movements result in the raccoon wandering and having contact with neighboring individuals. Both pathways are geographic in nature, and both transmission pathways are examined in this model. One objective of the model is to determine which mode of transmission has a larger role in the spread of rabies. A second objective is to assist in the development of rabies control strategies, thus reducing human exposures.

The Conceptual Model

A raccoon population in a rabies cycle can be divided into four categories or groups, depending on their disease status: susceptible raccoons; exposed raccoons (infected, but not infectious); immune raccoons; and infectious rabid raccoons (Figure 1). The susceptibles are those animals that previously have not been exposed to the virus or who have lost their immunity. A susceptible can only move into the exposed group. The exposed are those animals that have been exposed to an infectious, rabid animal. There is a variable incubation period of 39 to 79 days (13) during which rabies cannot be transmitted. The exposed animal can enter either the immune group or the infectious group. The category that the exposed animal moves to depends on the amount of virus to which the individual has been exposed. The immune group is composed of those animals that have been exposed to the rabies virus and have produced rabies-neutralizing antibodies. These animals are in this category for a variable length of time. Assuming that immunity in all animals is lost over time at the same rate, the rate of conversion from the immune group back to the susceptible group is treated the same for all animals that

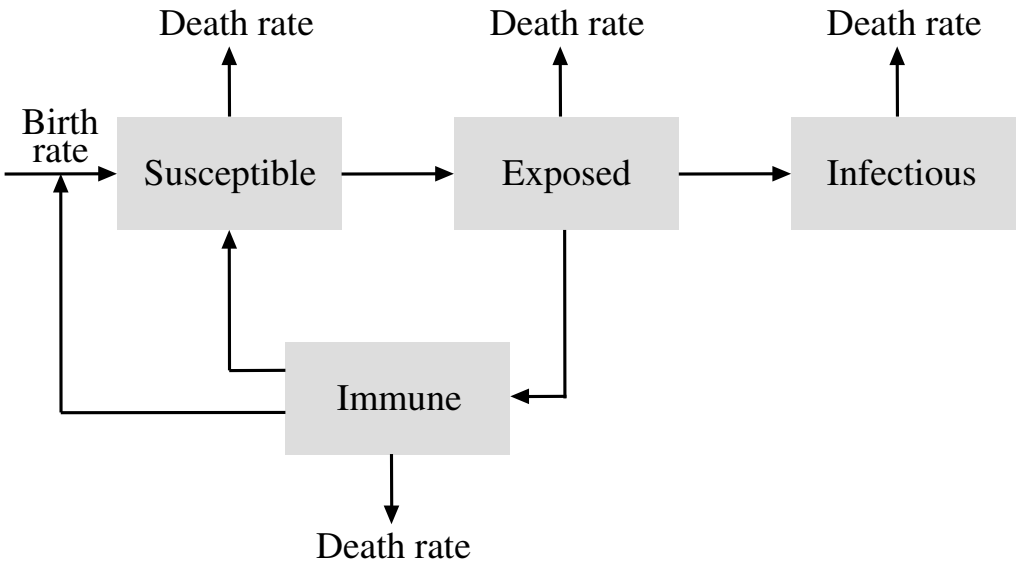


Figure 1 Epidemic model of the interactions between population processes (births and deaths) and disease processes among raccoon groups.

were exposed to an infectious animal. The infectious group is composed of those animals that show clinical symptoms for a short period of 3 to 8 days (13), and can transmit the disease by contact with the group of susceptibles. The infectious period ends in death, and the animal is removed from the population.

Young born to susceptibles and immunes are added to the susceptible group at a rate equal to the birth rate. Birth rate of the exposed group is assumed zero because of a gestation period of 63 days, and a weaning period of 2 to 4 months (4). Non-rabies death occurs in the susceptible, immune, and exposed groups, and is assumed to be the same rate for all groups. All individuals entering the infectious group are removed via the death rate, though death may occur from rabies or a non-rabies cause.

Infectious individuals having adequate contact with susceptible individuals results in a susceptible becoming an exposed individual. The probability of rabies transmission is the same as the contact rate. The amount of virus in the saliva is variable and is determined by a randomly drawn number from a lognormal frequency distribution produced from data on the prevalence of virus in saliva. The amount of virus and the LD50 value determine which category the exposed animals will enter. The dosage determines whether the individual will produce rabies-neutralizing antibodies or incubate the disease.

The incubation period in the exposed group is determined by a random variate drawn from a lognormal frequency distribution, which is produced from data of the known incubation period (15). After the incubation period, the exposed would then enter the infectious category. The length of stay in this category is also determined by a random variate drawn from a lognormal frequency distribution, which is produced from data of the known infectious period (15).

Conclusion

The long incubation period of rabies occurring in the exposed group is a major factor in the transmission cycle. The transmission cycle has two possible pathways, dispersal or home range. Using a geographic information system (GIS), a spatial grid was designed to simulate an extensive geographic area. The GIS allowed the data to be tracked over a period of time. The model uses the spatial procedures in ARC/INFO GIS, including GRID to simulate animal migration. The four status groups were tracked both spatially and temporally to analyze the pattern of the spread of the disease. The epidemic and animal movement models were integrated into the GIS using the ARC/INFO macro language (AML).

The animal movement model is a grid-based spatial model with grid cell size set equal to the average raccoon home range size (Figure 2). Raccoons are assigned to grid cells as individuals or as members of a social group. Individual raccoons then are assigned to one of the four status groups. For each individual in a given cell at time t, the model determines whether that individual moves to an adjacent cell (home range model), disperses to some more distant cell (dispersal model), or remains within the cell (Figure 3). The epidemic model then is used to determine the probability of individuals from the given cell infecting individuals in the cells contacted. Susceptible individuals have a probability of being infected by contact with those individuals that are infectious in contacted cells. After the interactions of individuals in each cell are determined, their group status is updated and the process repeats for each time step in the simulation.

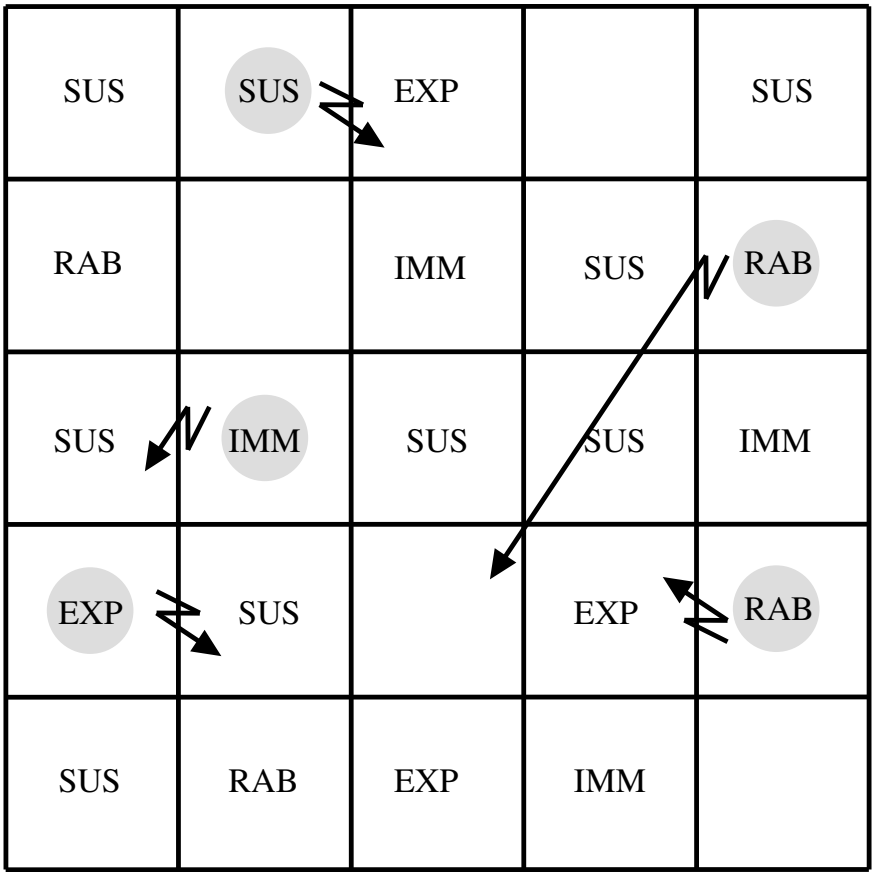


Figure 2 Movement of raccoons through a spatial grid. Each grid cell is equal to the size of the average home range. SUS = susceptible raccoon; EXP = exposed raccoon; IMM = immune raccoon; RAB = rabid infectious raccoon. Arrows represent movement to another grid cell.

Models of animal movement can be based upon any of the following three premises: the matching of spatial patterns of observed behavior (16), the set of rules arising from mechanisms governing the response of an individual to its environment (17), or theoretical constructs such as random walk models (18,19). The complete animal movement model incorporates features of all three methods within the home range. Models of dispersal distance are constrained random walk models. These models use transition probabilities that define the direction and distance moved, based on the animal's position relative to an activity center (20,21).

We anticipate that the integration of computer simulation and a GIS will assist in the development of an epidemic model that simulates the geographic spread of raccoon rabies. The model was designed to investigate disease control scenarios, such as optimizing the placement of rabies oral vaccine to impede the further spread of disease.

Obtaining adequate sample sizes for the estimation of the model's parameters will be essential for future research. Many hundreds of animals would have to be trapped to obtain reliable estimates of contact rates, concentrations of virus in saliva, antibody

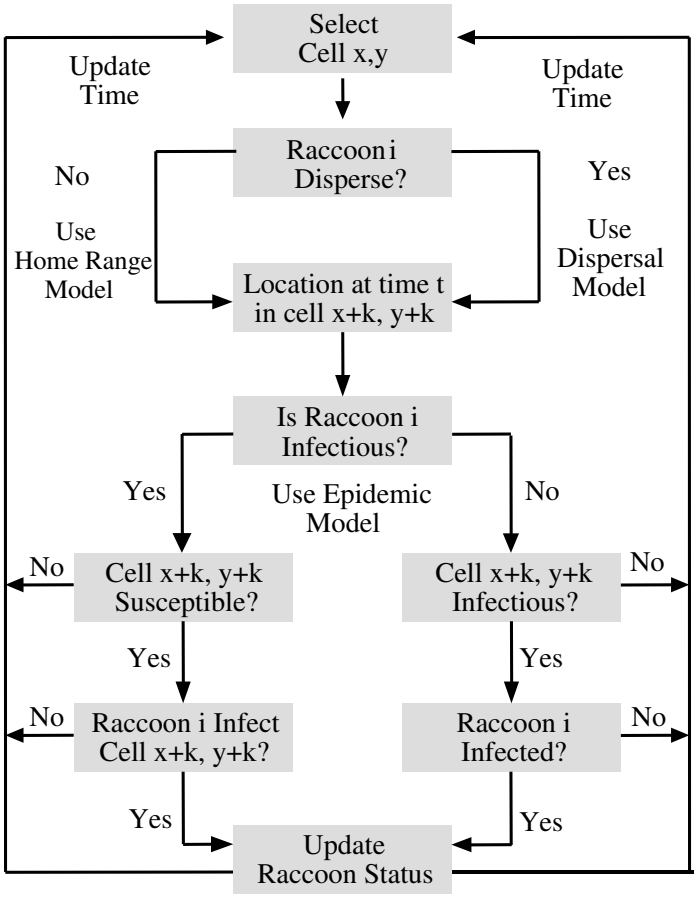


Figure 3 Flow diagram of contact among raccoon groups for the grid-based model.

prevalence, disease prevalence, and other factors. These estimates are necessary to achieve greater insight into the eventual control of wildlife rabies.

Acknowledgments

The authors wish to thank the Department of Biological Sciences and the Institute of Environmental and Human Health at Texas Tech University for financial support and use of facilities. We also thank the GIS in Public Health Conference planning committee for granting a student travel award, allowing attendance at the conference.

References

1. Murray JD, Stanley EA, Brown DL. 1986. On the spatial spread of rabies among foxes. *Proceedings of the Royal Society of London* B229:111–50.
2. Jenkins SR, Winkler WG. 1987. Descriptive epidemiology from an epizootic of raccoon rabies in the middle Atlantic states, 1982–1983. *American Journal of Epidemiology* 126:429–37.

3. Kaufmann JH. 1982. Raccoon and allies. In: *Wild mammals of North America: Biology, management, and economics.* Ed. JA Chapman, GA Feldhamer. Baltimore: The Johns Hopkins University Press. 567–85.
4. Lotze J, Anderson S. 1979. Procyon lotor. *Mammalian Species* 119:1–8.
5. Barash D. 1973. Neighbor recognition in two solitary carnivores: The raccoon and the red fox. *Science* 185:794–96.
6. Schwartz CW, Schwartz ER. 1981. *The wild mammals of Missouri.* Columbia, MO: University of Missouri Press and Missouri Department of Conservation.
7. Ellis RJ. 1964. Tracking raccoons by radio. *Journal of Wildlife Management* 28:363–68.
8. Giles LW. 1943. Evidences of raccoon mobility obtained by tagging. *Journal of Wildlife Management* 7:235.
9. Priewert FW. 1961. Record of an extensive movement by a raccoon. *Journal of Mammalogy* 42:113.
10. Lynch GM. 1967. Long range movement of a raccoon in Manitoba. *Journal of Mammalogy* 48:659–60.
11. Butterfield RT. 1944. Populations, hunting pressure, and movement of Ohio raccoons. *Transactions of the North American Wildlife Conference* 9:337–43.
12. Kaplan C. 1985. Rabies: A worldwide disease. In: *Population dynamics of rabies in wildlife.* Ed. PJ Bacon. London: Academic Press. 1–22.
13. Carey AB. 1985. Multispecies rabies in the eastern Unites States. In: *Population dynamics of rabies in wildlife.* Ed. PJ Bacon. London: Academic Press. 23–42.
14. McLean RG. 1975. Raccoon rabies. In: *The natural history of rabies, vol. 2.* Ed. GM Baer. New York: Academic Press. 53–77.
15. Frerichs RR, Prawda J. 1975. A computer simulation model for the control of rabies in an urban area of Columbia. *Management Science* 22:411–21.
16. Siniff DB, Jensen CR. 1969. A simulation model of animal movement patterns. *Advances in Ecological Research* 6:185–219.
17. Wolff WF. 1994. An individual-oriented model of a wading bird nesting colony. *Ecological Modelling* 72:75–114.
18. Holgate P. 1971. Random walk models for animal behavior. In: *Statistical ecology, vol. 2.* Ed. GP Patil, EC Pielou, WC Waters. University Park, PA: Penn State University Press. 1–12.
19. Tyler JA, Rose KA. 1994. Individual variability and spatial heterogeneity in fish population models. *Reviews in Fish Biology and Fisheries* 4:91–123.
20. Dunn JE. 1978. Optimal sampling in radio telemetry studies of home range. In: *Time series and ecological processes.* Ed. HH Shugart. Philadelphia: SIAM Institute for Mathematics and Society. 53–70.
21. Dunn JE, Gipson PS. 1977. Analysis of radio telemetry data in studies of home range. *Biometrics* 33:85–101.

A GIS Analysis of Motor Vehicle Injuries in Ventura County, California

Paul Van Zuyle*

Department of Geography, University of California, Santa Barbara, CA; Ventura County Public Health Department, CA

Abstract

An analysis of motor vehicle injuries in Ventura County, California, revealed a variety of spatial patterns that depend on the age of the victims. Vehicle injury data were selected from 25,000 transport records collected by Ventura County Emergency Medical Services (EMS) and were geocoded. Of 3,275 motor vehicle injuries in 1996 requiring emergency transport, 554 were to 15- to 19-year-olds. This age group represents only 7% of the population but incurred 17% of vehicular injuries. While no address information was recorded, each record contained a Thomas Brothers map book grid coordinate, which was geocoded using an Avenue script for ArcView 3.0a. The locations of these injuries were aggregated using kernel filtering in ArcView Spatial Analyst to produce a density surface on a 200-meter grid. The collocation of the three greatest peaks in teenage motor vehicle injuries with the three large high schools in the county is notable, and has been used by health educators and EMS to target education and prevention efforts. In addition, an experimental Web site has been set up so that the data can be interactively explored on the Internet.

Keywords: injury, motor vehicle, Internet, Ventura County, public health

Introduction

Motor vehicle accidents were the cause of more than half of all injuries requiring emergency medical transport in Ventura County, California, in 1996. With the cooperation of the Ventura County Emergency Medical Services (EMS) office, the Ventura County Health Department performed a geographic and statistical analysis on selected injury data collected on persons who were transported by ambulance to hospitals. The aim was to quantify the number and location of injuries to children and teenagers due to motor vehicles. An additional goal was to create a prototype Web site, where the map data could be queried online and different aspects of the dataset could be explored.

Data

Out of over 30,000 EMS runs in Ventura County in 1996, 9,542 were to transport persons with injuries. (The data collected were actually from the beginning of December 1995 through the end of November 1996.) Of those 9,542, 4,907 were classified as "MV," or injuries received in motor vehicles. Another 1,563 were attributed to falls. Gunshot wounds, stabbings, and assaults accounted for 1,783 more. Of the rest, 380 were simply

* Paul Van Zuyle, Dept. of Geography, University of California, Santa Barbara, CA USA 93106; (p) 805-893-8652; (f) 805-893-8617; E-mail: vanzuyle@geog.ucsb.edu

classified as "other," while the balance was attributed to pedestrian injuries from cars, burns, near-drownings, and other causes.

Analysis

A spatial analysis was performed on the data using locational information coded for the scene of each injury. As part of the dispatch process, a map grid coordinate for each run was entered from the Thomas Brothers map book (1). While no address information was available, these half-mile grid cells provide sufficient resolution for certain kinds of analysis, including the production of countywide maps. The analysis was performed with ArcView 3.0 and ArcView Spatial Analyst (ESRI, Redlands, CA), along with custom software written to convert the Thomas Bros. grid references to locations on the map.

The center of each grid cell was coded with an attached value for each of the injury types investigated. This resulted in a map with a regular grid of points whose size was displayed proportionate to the number of specified injuries recorded in that area. This is an effective mapping technique for small numbers of occurrences. For large numbers, a different analytic technique, kernel filtering, was utilized. This involved interpolating the value of each grid point to create a smooth surface representing an estimated value for every place on the map. These values were then classed to create a choropleth map.

For presentation on the Internet, the data have been stored on a Web server running ESRI's ArcView Internet Map Server. (The demonstration Web site can be found at http://asosan.geog.ucsb.edu/maps/vtainjury.html.) This application comes with a Java applet that allows users of properly configured Web browsers to view and query the data at a variety of scales. For users of ArcView GIS, the MapCafe Java applet presents map layers in a familiar format. A sample of previously performed analyses can be viewed along with other layers, such as streets, high schools, and political boundaries. These analytic layers (or "themes," in ArcView parlance) can be individually turned on or off to create different map views.

Results

Overall, the geographic distribution of motor vehicle injuries in Ventura County is similar to the spatial distribution of people (see theme "MV accident density per sq. km" at the demonstration Web site). This means that the rate of injury is relatively constant from place to place. Exceptions to this appear in rural areas with busy highways. For instance, Highway 126 east of Fillmore has far more injuries than would be expected based on the size of the local population (see theme "MV injury per 100k pop"). Obviously, however, this highway and other rural roads serve more than the local population, and it is likely that those injured were not local residents. (We cannot verify this, however, because the home addresses of the victims were not recorded in this dataset.)

This general pattern is slightly different when only injuries to children and teenagers are considered (see theme "u20 MV PA injuries per sq. km"). As might be expected, these show more local foci, and a pattern emerges that reflects the locations of schools. This pattern is further reinforced when only motor vehicle injuries to teenagers 15–19 years are mapped (see theme "MV injuries to 15–19 year olds"). Over half of all

motor vehicle injuries to children and teenagers are incurred by this age group. While adolescents of this age represent only about 7% of the population, they incur nearly 17% of the injuries.

Distinctive peaks occur on the map on Moorpark Road near Thousand Oaks High School in Thousand Oaks, and at the intersection of Ventura Road and Gonzales, near Oxnard High in Oxnard. This last place was the site of 17 injuries to teenagers, nearly twice as many as any other single location on the map.

An examination of the numbers of injuries to teenagers compared with juveniles of other ages shows a striking increase once they become old enough to hold a license (Table 1). There is no such corresponding increase in injuries to pedestrian teenagers (Table 2).

Table 1 Motor Vehicle Injuries, Ventura County, CA, 1996

	Age (years)				
	0–4	**5–9**	**10–14**	**15–19**	**Sum**
Girls	41	36	69	259	405
Boys	44	33	39	259	375
Sum	85	69	108	518	

Source: (2)

Table 2 Pedestrian-Auto Injuries, Ventura County, CA, 1996

	Age (years)				
	0–4	**5–9**	**10–14**	**15–19**	**Sum**
Girls	11	5	23	16	55
Boys	14	26	34	25	99
Sum	25	31	57	41	

Source: (2)

Discussion

The overall motor vehicle injury rate for different places in Ventura County is mostly a function of population. There is little difference from city to city in the aggregate. When specific areas are examined, however, there are distinctive local differences. Because of the nature of the data collected for this report, it is not generally possible to locate specific accident sites. In certain cases, however, such as with the data for injuries near Ventura Road and Gonzales in Oxnard, it may be reasonable to assume that most of the accidents in that area occurred at the intersection of those two streets.

When data for specific sub-populations such as teenagers are viewed, the pattern of injuries becomes more distinct from the population distribution. In this case, the spatial behavior of teenagers is revealed in relation to high schools through their motor vehicle injuries. This offers the potential for traffic enforcement, emergency preparedness, and education directed at these specific groups and places.

Presentation of these data in map form on the Internet allows users with many different interests and responsibilities to adjust their view of the county data to fit their specific needs. While there may be a variety of underlying causes for the patterns revealed in these data, the potential for understanding them may be increased by both the presentation and distribution schemes described here.

References

1. Thomas Brothers. 1998. 1*998 Ventura County Thomas guide*. http://store.thomas.com/thomasbros/3053.html.
2. Ventura County Emergency Medical Services. 1996. Vehicle injury data. Ventura County Emergency Medical Services, Ventura County, CA.

Childhood Lead Poisoning: The Potential and Pitfalls of Applying GIS to the Development of Federal Environmental Justice Policy

Max Weintraub, MS*
US Environmental Protection Agency Region 9, San Francisco, CA

Abstract

Childhood lead poisoning is the most common environmental illness facing US children. When the first federal legislation passed in 1971, children in communities of color and low-income communities were disproportionately at risk, a characteristic typical of communities organizing around environmental justice principles. Efforts since then have decreased the extent of childhood lead poisoning, but have increased the disparate impact of the disease. Since the 1994 passage of Executive Order 12898 on environmental justice, the US Environmental Protection Agency (EPA) has begun to develop geographic information system (GIS) programs to assess a broad array of environmental justice issues. This paper will describe the origin of the environmental justice movement and examine childhood lead poisoning as an environmental justice challenge. It will then outline three GIS approaches to childhood lead poisoning and consider what the strengths and weaknesses of those approaches portend not only for EPA's efforts to prevent childhood lead poisoning in the Southwest, but for other GIS uses designed to remedy environmental justice problems.

Keywords: lead poisoning, environmental justice, race, class, public health

Introduction

Potential, and pitfalls, abound in the use of geographic information systems (GIS) to identify and solve public health problems. Childhood lead poisoning is a particularly sensitive example given the role this illness has played in federal recognition of a social movement called environmental justice. The goal of this paper is to relate environmental justice, federal policy, childhood lead poisoning, and GIS in a cautionary tale that recognizes how GIS can promote or stifle positive social change. This paper will describe the origin of the environmental justice movement and examine childhood lead poisoning as an environmental justice challenge. It will then outline three GIS approaches to childhood lead poisoning and consider what the strengths and weaknesses of those approaches portend not only for the US Environmental Protection Agency's (EPA's) efforts to prevent childhood lead poisoning in the Southwest, but for other GIS uses designed to remedy environmental justice problems.

Genesis of the Environmental Justice Movement

Extensive histories have been written about the environmental justice movement (1,2). However, a few key incidents defined the movement for the federal government.

* Max Weintraub, US EPA, 75 Hawthorne St, CMD-4-2, San Francisco, CA 94105-3901 USA; (p) 415-744-1129; (f) 415-744-1073; E-mail: weintraub.max@epamail.epa.gov

Low-income communities of color created the environmental justice movement in recognition of the small share of environmental amenities and large share of environmental burdens they experience. Such burdens range from a disproportionate rate of environmental disease to the disproportionately high concentration of toxic waste facilities sited in their communities. Indeed, it was during the release of the 1987 GIS study Toxic Waste and Race (3), identifying the disproportionate siting of toxic waste facilities in low-income communities of color, that the term "environmental racism" first achieved national prominence.

More than 100 environmental justice groups came together in Washington, DC, in 1991 at the First National People of Color Environmental Leadership Summit to create the "Principles of Environmental Justice." These 17 principles represent the consensus understanding of these groups and guide their approach to the challenge of unequal protection.

Within a year after the summit, EPA issued its first report on the matter (4). That report identified childhood lead poisoning as the only environmental illness disproportionately harming low-income communities and communities of color. As more research was conducted, however, and the scope of the problem beyond childhood lead poisoning became apparent, greater federal involvement became necessary.

In 1994, hundreds of federal officials and activists from environmental justice groups came together at the Symposium for Health Research and Needs to Ensure Environmental Justice to discuss the issue. During the symposium, President Clinton signed an executive order on environmental justice that required federal agencies to develop strategies to address environmental justice concerns and established a federal advisory committee that would provide a regular forum for environmental justice issues to be raised (5). Since then, the National Environmental Justice Advisory Council has met every six months in communities around the country, and dozens of federal agencies have developed environmental justice strategies (6).

Lead Poisoning as an Environmental Justice Challenge

Childhood lead poisoning is the most common environmental disease threatening US children. When the first federal lead poisoning prevention legislation passed in 1971, children in communities of color and low-income communities were recognized as being at high risk for the condition. Efforts since then have dramatically decreased the extent of childhood lead poisoning, but have failed to diminish the disparate impact of the disease. And, in fact, the disparity has grown.

Between 1976 and 1994, the percentage of children with elevated blood lead levels (above 10 micrograms per deciliter [µg/dL]) from the highest income bracket decreased at a rate seven times greater than the percentage of children with elevated blood lead levels from the lowest income bracket. A similar analysis by race indicates that the percentage of white children with elevated blood lead levels declined at a rate four times greater than that of their black counterparts. The benefit of "whiteness" is explicit as Hispanic children are twice as likely, and black children fives times more likely, to have elevated blood lead levels than are white children. As a result, the number of black children with elevated blood lead levels is equal to the number of white and Hispanic children with elevated blood lead levels. Before discussing the implications of these findings, consider how GIS fits in the picture.

Three GIS Approaches to Childhood Lead Poisoning

Childhood lead poisoning occurs as a consequence of lead exposure. In the past, targeting efforts to control childhood lead poisoning followed the "canary in the coal mine" model. That is, after a child was poisoned, health professionals would seek to identify and eliminate the source of lead exposure. GIS offers the opportunity to take a more proactive approach by mapping out risk factors and identifying communities at risk for lead exposure.

The risk factors for assessing children's lead exposure identified by the Centers for Disease Control and Prevention (CDC) include:

- Pre-1950 housing
- Demographic factors
- Industrial sources and parental occupation
- Drinking water
- Hobbies, traditional remedies, ceramicware, and cosmetics (7)

Using GIS to overlay maps of these various risk factors can help EPA childhood lead poisoning prevention personnel determine where greatest needs exist, what type of resources should be allocated, and success over time. The following three different GIS approaches have been developed to support childhood lead poisoning prevention efforts.

The first GIS application that focused on childhood lead poisoning prevention was described in 1991. Researchers at the New Jersey Department of Environmental Protection and Energy and the University of Medicine and Dentistry of New Jersey used GIS to identify areas within Newark, East Orange, and Irvington, New Jersey, where there may be greater environmental exposure to lead. The purpose of the study was to identify areas where further screening and public education may be needed (8).

Table 1 lists the sources of data used in this 1991 study. Data incorporated into the GIS included US Census Bureau demographic and boundary data. Furthermore, data that reflected point industrial and urban corridor sources of lead, as well as blood lead screening records, were included. The strength of the application was reflected in the finding that a strong correlation existed between census tracts with reported high blood lead levels and census tracts predicted by the GIS to support high lead exposure.

In 1992, the Lead Education and Abatement Program (LEAP) in EPA's Region 5 office published its GIS assessment of the spatial and numerical dimensions of young minority children exposed to low-level environmental sources of lead (9). This study of the Great Lakes area did two things. First, it developed a population comparative risk analysis for childhood exposure to lead in census tracts in 83 Midwest cities. Second, it examined whether there was an association in the Minneapolis/St. Paul area between urban transportation corridors and elevated blood-lead levels.

Table 2 lists the data sources used in this 1992 population comparative risk analysis. The study used Census Bureau boundary data. Because the 1990 census data had still not been released, the researchers used the Donnelly Marketing Population database for demographic data extrapolated from 1980 to 1990. Point industrial, urban corridor, and drinking water sources of lead were included.

The study found a statistically weak association between Minneapolis/St.Paul urban corridors and blood lead levels. Furthermore, a weak association was identified

Table 1 Databases Used in 1991 GIS Study of Lead Exposure in Newark, East Orange, and Irvington, NJ

Database	Source
Census tract boundaries, demographics, and housing stock	US Census Bureau
Essex County blood lead screening records	New Jersey Department of Health
Toxic Release Inventory of industrial sites emitting lead	US EPA
Hazardous waste sites contaminated with lead	New Jersey Department of Environmental Protection and Energy
Traffic volume estimates to determine past leaded fuel emissions	New Jersey Department of Transportation

Table 2 Databases Used in 1992 GIS Study of Lead Exposure in the Midwest (with detailed analysis in Minnesota)

Database	Source
Ambient air quality data	Aerometric Information Retrieval System
Ethnicity, sex, age, income, housing age, and location	US Census Bureau; extrapolations available via Donnelly Marketing Population Data
Surface meteorological data	National Climatologic Data Center
Toxic Release Inventory for point sources of lead emissions and facilities that dispose of lead	US EPA
Municipal waste incinerators emitting lead	US EPA
Lead levels in drinking water	US EPA
Abandoned hazardous waste sites where lead is a primary concern	US EPA
Lead concentrations in soil and dust	US Department of Housing and Urban Development; Minnesota Department of Pollution Control
Blood lead screening data	Minnesota Department of Health

between the blood lead levels predicted by the GIS and those measured through screening efforts. The researchers attributed the weak association to the fact that the model is applicable to populations, not individuals, and that an inability to account for ethnicity and socioeconomic status resulted in an underestimate of the at-risk population in lower socioeconomic minority communities. The researchers concluded that the effort should prove useful in identifying hotspots of lead exposure.

The year 1993 witnessed the first results of the EPA Office of Pollution Prevention and Toxics Lead Targeting System, a meeting in Atlanta titled "Mapping Lead Exposure Information" by the EPA Environmental Criteria and Assessment Office, and the release of information about efforts to use GIS to target childhood lead poisoning prevention activities in California and Massachusetts. Unfortunately, this brief flurry of activity did not prove sustainable. The challenge, as summarized by a co-author of the 1991 New Jersey study, was

> . . . to better understand the complexities of lead exposure and vulnerable populations. This effort requires addressing the issues of what information should

be collected, of how to best collect information, how to overcome incompatibility of data, and ways to share information between different software packages and hardware (10).

In 1998, many of those challenges remain. However, the question of what information should be collected has been somewhat simplified. For example, EPA Region 3 is currently using GIS to help identify the human health risk from lead and, in particular, reduce the prevalence of childhood lead poisoning in targeted communities (11). They are doing so by assessing the lead poisoning risk factors mentioned previously and eliminating those factors less often causally related to childhood lead poisoning. They have selected age of housing, poverty, and the presence of children as the focus of their analysis. These researchers are using Census Bureau boundary and demographic data, and US Department of Housing and Urban Development (HUD) housing data in their review. Table 3 identifies additional sources of data used in this study.

Table 3 Databases Used in the 1998 GIS Study of Lead Exposure in Mid-Atlantic States (with detailed analysis in Philadelphia)

Database	Source
Affordable housing Development	US Department of Housing and Urban
Age, income, owner-occupied vs. renter-occupied housing, and housing age	US Census Bureau
Residential lead hazard	US Department of Housing and Urban Development
Blood-lead screening data	Philadelphia Department of Health

The Region 3 study found that targeting major urban areas would, if fully successful, address only 25% of the houses estimated to have lead-based paint, 8% of houses expected to have lead-based paint hazards, and 33% of the children in poverty. In response, Region 3 is beginning to implement a children's initiative that will target lead exposure risks to children in smaller urban areas.

The weakness of this approach is that it does not consider causes of childhood lead poisoning that, while less prevalent, may be significant. That challenge confronts those of us working in Region 9 as we try to develop plans to decrease childhood lead poisoning.

Current Challenges

Region 9's 1993 effort to create a GIS application to measure the potential for elevated blood-lead levels used Census Bureau demographic data to localize childhood lead poisoning in Oakland, California, and the surrounding areas of Alameda County (12). In that one county alone, EPA researchers found thousands of white, black, Hispanic, Asian-American, and other ethnicity children living in poverty. More than 80% of homes were coated with lead-based paint. Potential sources such as water and industrial emissions were considered. And what did they find? A mess that could not be readily sorted out for targeting purposes and that did not closely match the results of screening data that were being collected in the area.

This is not a surprise. Communities in the Southwest are substantially different from those in the Northeast and Midwest. Some of the challenges of preventing childhood lead poisoning in the Southwest (compared with the Midwest or the Northeast) include the following:

- **Space:** Housing is diffuse so hot spots are more difficult to identify and target for action.
- **Time:** Rapid population growth in the Southwest results in data rapidly becoming obsolete.
- **Population:**
  More diverse and more integrated.
  Cultural exposure sources more prevalent.
  Modeling spatial dimension of ethnicity more difficult.
  Class component of childhood lead poisoning weaker.
  Population at greatest risk more likely to speak English as a second language.
- **Medical practice:** Pediatricians are less aware of childhood lead poisoning and less likely to screen.
- **Legislation:** Non-existent or very recently passed.
- **Government agencies:** Relatively recently recognized problem with limited and very localized data to guide action.

Housing stock is more spread out in the Southwest than in the Northeast. Thus, the potential for spatial autocorrelation (which is quite high in densely populated cities of the Northeast) is diminished. Race and class are also much less effective predictors of childhood lead poisoning. The population of California is a bit more than 50% white, 30% Hispanic, 7% black, and 6% Asian-American. Communities in the Southwest are more racially integrated and class does not have as strong an influence upon the prevalence of lead poisoning. Furthermore, cultural characteristics may promote childhood lead poisoning in the Southwest more than in the Northeast, because the Hispanic and Asian populations are more likely to use folk medicines, ceramicware, and cosmetics that contain lead.

For example, a study released by the General Accounting Office in May 1998 noted that the California Department of Health Services has reported that up to 12% of lead-poisoned children in the state may have been poisoned from traditional folk medicines and another 8% of cases may have been linked to lead-glazed pottery, often from Mexico (13). The same report also noted that while Hispanic children in pre-1946 housing had a higher prevalence of elevated blood lead levels than those in newer housing, in either setting the risk of elevated blood lead levels was not appreciably changed by poverty status.

Finally, unlike many cities in the Northeast where lead screening and awareness are relatively high, little screening has taken place in the Southwest and most is quite recent in origin. Indeed, before a 1992 court decision forcing the issue, screening was rarely performed in California, even for children on MediCal. Recent studies indicate pediatricians in California continue to screen children much less often than their counterparts in other parts of the country (14,15). Thus, unlike other areas, it is difficult to analyze GIS applications for lead exposure risks relative to screening data because, until recently, little data existed and, as a result, confidentiality concerns were difficult for researchers to overcome.

As is apparent from these facts, the childhood lead poisoning situation in the Southwest is appreciably different from that in the Northeast where most GIS applications have been tested and where federal policy efforts are focused. Yet, despite these differences, it is important to note one universal truth—that children in old homes are more likely to suffer lead poisoning. California has more homes built before 1950 than any state other than New York or Pennsylvania.

Despite the challenges posed in the Southwest, two national GIS programs are being developed to identify communities at risk for childhood lead poisoning. HUD has recently released "Community 2020" (16). This program allows maps to be created instantly by selecting and displaying census data. The program targets areas by overlaying four characteristics: housing older than 1950; presence of children under six years old; minority status; and, presence of single parent household. The goal is to use this application to target inspections and compliance assistance to better implement the new real estate disclosure law for lead-based paint.

Census data are also the building block for the CDC software that has been developed to provide relevant data on housing and population to help identify high-risk areas for childhood lead poisoning screening (17). This software can be accessed over the Web and, while still in the process of being integrated with mapping, provides county and zip code level data on a broad array of factors including: housing units, pre-1950 housing, children under six years old, race, income, owner or renter status, and percent of children under six years old in poverty.

Future Moves

Both the Community 2020 and new CDC software go a long way toward incorporating the breadth of demographic data absent in some of the earlier applications. However, the need to refine our focus is reinforced by two recent targeting studies contracted by EPA. The first provided support for the notion that real estate disclosure enforcement efforts should target extremely rural areas like Tulare County, California, before targeting Los Angeles (18). The second confirmed once again that race, income, census region, and age of housing are associated with environmental lead exposure and that blood lead level variations by race and class may not be adequately explained by environmental lead measurements (19). Both studies conclude that resources should be focused on Northeast or Midwest communities, yet both fail to acknowledge differences in the epidemiology of childhood lead poisoning in the Southwest that may alter such findings.

Conclusion

Within the last year, the EPA received more than $3 million in applications for lead poisoning prevention activities from non-profit groups in Region 9, though only $200,000 in funding was available. Many of the groups applying are run by, and serve, low-income communities of color most at risk for the disease. These environmental justice groups recognize that despite the decreasing prevalence of childhood lead poisoning, the battle has not been won because their children have been left behind in past federal efforts. GIS can help federal agency personnel decide where to target limited resources for screening, grants, and enforcement in order to eliminate disparities in disease preva-

lence. However, if "garbage in, garbage out" GIS models are produced that fail to recognize the unique situations that different areas face, federal agency personnel not only miss an opportunity to ensure equal protection under the law for all Americans, but also perpetuate the idea that helped generate the environmental justice movement in the first place—that government agencies may not only fail to remedy vestiges of past racism and classism, but through a failure to recognize such challenges may create new barriers to creating a just and fair society.

References

1. Bryant B, Mohai P. 1992. *Race and the incidence of environmental hazard.* Boulder: Westview.
2. Bullard R. 1990. *Dumping in Dixie: Race, class, and environmental quality.* Boulder: Westview.
3. Commission for Racial Justice and Public Data Access, Inc. 1987. *Toxic wastes and race in the United States: A national report on the racial and socioeconomic characteristics of communities with hazardous waste sites.* New York: United Church of Christ Commission for Racial Justice.
4. USEPA. 1992. *Environmental equity: Reducing risk for all communities.* Washington, DC: US Environmental Protection Agency. EPA230-R-92-008.
5. The White House. 1994. *Executive order on federal actions to address environmental justice in minority populations and low-income populations.* Executive Order 12898. Washington, DC: The White House. 11 February.
6. Information on environmental justice strategies for various federal agencies. 1998. http://www.envirosense.com/oeca/ofeds.html.
7. CDC. 1997. *Screening young children for lead poisoning: Guidance for state and local public health officials.* Atlanta: Centers for Disease Control and Prevention. 22.
8. Guthe WG. 1992. Reassessment of lead exposure in New Jersey using GIS technology. *Environmental Research* 39:318–25.
9. USEPA. 1992. *Project LEAP—Phase 1: Spatial and numerical dimensions of young minority children exposed to low-level environmental sources of lead. Report and GIS Appendix.* Washington, DC: US Environmental Protection Agency. EPA905-R-92-002.
10. Luckhardt JC. 1993. Introduction. In: *Proceedings of the Mapping Lead Exposure Information Meeting.* July 24–25. Atlanta: US Environmental Protection Agency.
11. USEPA Region 3. 1998. *Project to characterize the extent of children's health risk from lead Region III.* Philadelphia: US Environmental Protection Agency. http://www.epa.gov/rg3wcmd/leadrept.htm.
12. USEPA Region 9. 1993. *GIS/RCRA—Lead contamination and children as receptors.* San Francisco: US Environmental Protection Agency. (Unpublished.)
13. US GAO. 1998. *Elevated blood lead levels in Medicaid and Hispanic children.* Washington, DC: US General Accounting Office. GAO/HEHS-980169R.
14. Ferguson SC. 1997. *Blood lead testing by pediatricians: Practice, attitudes, and demographics.* American Journal of Public Health 87:1349–51.
15. Campbell JR. 1996. *Blood lead screening practices among US pediatricians.* Pediatrics 98:372–77.
16. HUD. 1998. *Community 2020.* Washington, DC: US Dept. of Housing and Urban Development. http://www.hud.gov/cpd/2020soft.html.
17. CDC. 1998. *CDC childhood lead poisoning prevention program.* Atlanta: Centers for Disease Control and Prevention. http://www.cdc.gov/nceh/programs/lead/lead.htm.

18. Miller D. 1997. *TSCA section 1018 county level targeting (draft).* Science Applications International Corporation.

19. McMillan N. 1998. *Consideration of target population relevance to EPA's lead program.* Battelle Memorial Institute. Contract No. 68-D5-0008.

Disease Surveillance

Cancer Incidence in Southington, Connecticut, 1968–1991, in Relation to Emissions from Solvents Recovery Services of New England

Diane D Aye, MPH, PhD (1),* Gary V Archambault, MS (1), Deborah Dumin (2)
(1) Division of Environmental Epidemiology and Occupational Health, Connecticut Department of Public Health, Hartford, CT; (2) Connecticut Department of Environmental Protection, Hartford, CT

Abstract

Data from Southington, Connecticut, were analyzed using geographic information systems (GIS) to explore a possible association between exposure to contaminants from Solvents Recovery Services of New England (SRSNE) and the incidence of selected types of cancer. Data on the incidence of bladder, kidney, liver, and testicular cancer, leukemia, non-Hodgkin's lymphoma (NHL), and Hodgkin's disease were obtained from the Connecticut Department of Public Health Tumor Registry for the period 1968 to 1991. Improper disposal practices by SRSNE caused the air, public drinking water, and soil in Southington to be contaminated. Possible dose-response relationships between exposure to emissions from SRSNE and cancer risk were explored by calculating age and sex standardized incidence ratios (SIR). No statistically significant increase in the SIR was found for cancer of the bladder, kidney, liver, or testis, leukemia, NHL, or Hodgkin's disease for any of the exposure categories when compared with State of Connecticut incidence rates. The total SIR of all tumor sites combined, however, demonstrated a statistically significant increasing trend in relation to increasing exposure to air emissions. Among individual tumor sites, the risk of NHL among females was elevated in locations where the air exposure levels were estimated to be the greatest, although this elevation did not achieve statistical significance. Non-Hodgkin's lymphoma incidence has been increasing during the past several decades with no clear explanation. This study suggests the need for more evaluation of exposure to environmental contaminants and the development of some types of cancer, with specific attention given to NHL.

Keywords: cancer, non-Hodgkin's lymphoma (NHL), solvent exposure, drinking water, air pollution

Background

Solvents Recovery Services of New England (SRSNE) is a National Priority List (NPL) hazardous waste site located in Southington, Connecticut. SRSNE began its solvent recovery operations in 1955. The facility processed between 3 and 5 million gallons of liquid hazardous waste and 100,000 pounds of solid hazardous waste annually.

The SRSNE facility operations included the distillation of recoverable solvents in batch stills with sludges being placed in unlined on-site lagoons for disposal. Improper disposal practices by the company caused the air, public drinking water, and soil in Southington to be contaminated with waste solvents and metals.

* Diane D Aye, Connecticut Dept. of Public Health, 410 Capitol Ave., PO Box 340308, MS#:11CHA, Hartford, CT 06134 USA; (p) 860-509-7742; (f) 860-509-7785; E-mail: diane.aye@po.state.ct.us

The public drinking water wells #4 and #6 are located approximately 1,200 feet south of SRSNE. Well #4 was installed in 1966 and well #6 in 1976. The wells were identified as being contaminated with volatile organic compounds (VOCs) and possibly heavy metals in 1976 and 1977 (1).

Table 1 presents data on the maximum concentrations of contaminants detected from public water supply wells #4, #5, and #6. Well #5 is located approximately 4 miles south of SRSNE and its contamination is not directly site-related. However, SRSNE-generated wastes were disposed of at the Old Southington Landfill (also an NPL site) near well #5. These wells were taken out of production in 1979.

Table 1 Highest Recorded Contamination of Public Water Supply Wells, Southington, CT, 1978 and 1979

Contaminant	Well #4 (ppb)	Well #5 (ppb)	Well #6 (ppb)
Trichloroethylene	120	45	11
1,1 dichloroethylene	210	No data	No data
1,1 dichloroethane	990	No data	No data
t,1,2 dichloroethane	390	6	No data
1,1,1 trichloroethane	33,500	300	120
Tetrachloroethylene	22	No data	No data
Carbon tetrachloride	35	9	No data
Hexane	91	No data	No data
Methane	400	480	130
Methylene chloride	12	No data	No data
Chlorobutane	930	No data	No data
Methyl ethyl ketone	No data	No data	20

ppb = parts per billion
Source: (20)

An on-site open pit incinerator for the burning of solvent and metal sludges operated with no air pollution controls until 1974, when it was taken out of service. Other sources of air pollution included evaporation from the lagoons and storage tanks at the facility, and 25 recovery wells with uncontrolled air strippers.

Methods

Case Ascertainment

Cases of bladder, kidney, liver, and testicular cancer, leukemia, non-Hodgkin's lymphoma, and Hodgkin's disease diagnosed to Southington residents between 1968 and 1991 were mapped using a geographic information system (GIS) and the census block where the case resided at the time of diagnosis was identified. Data on drinking water and air exposure to trichloroethylene (TCE) emissions from SRSNE were estimated and each census block received a relative exposure score (no actual measurements of contaminant levels at case addresses was available). Census blocks with the same

qualitative exposure scores were grouped for analysis. Age and sex standardized incidence ratios (SIRs) were calculated for each tumor site and exposure category to compare the incidence of each cancer in each exposure category with the incidence for Connecticut as a whole.

The tumor sites included in this study were selected based on epidemiological and toxicological studies and community concerns. Epidemiological studies have been conducted that linked populations exposed to drinking water contaminants with bladder cancer, leukemia, and lymphoma (2–15). Toxicological evidence has linked the contaminants with liver and kidney cancer in animals (5,6,16). Testicular cancer was included in the study because citizens in the community expressed concern about testicular cancer incidence.

Preliminary review of the cancer incidence information on these sites is summarized in Table 2. While Southington did not experience an excess of these tumor types when compared with Connecticut state rates, mapping was done on these cases to determine if the incidence of these tumors was increased in areas of the town that were exposed to air or water contamination from SRSNE. Data on cases of cancer occurring in residents of Southington between 1968 and 1991 were obtained from the Connecticut Department of Health (DPH) Tumor Registry. Individual case information includes the patient's residential address at time of diagnosis, primary site of diagnosis, age, sex, and date of diagnosis.

The residential address at time of diagnosis was assigned digital map coordinates using GIS mapping capabilities. Also, during the geocoding process the census block of residence was identified. Mapping of addresses was conducted by the Connecticut Department of Environmental Protection (DEP) with the assistance of an enhanced TIGER file, Dynamap/2000. Each of the addresses was verified against the Southington Assessors maps or by field investigation to ensure accuracy of the geocoding.

Of the 424 cases identified from the registry, 422 cases were geocoded. The remaining two cases could not be geocoded because the address was not listed in the registry records. A map displaying the location of cancer cases in Southington is not presented here in order to protect the confidentiality of the data.

Water Exposure Modeling

The Agency for Toxic Substances and Disease Registry (ATSDR) and the Georgia Institute of Technology hydrologically analyzed the water supply system to determine the geographic areas with the greatest potential for TCE contamination of the drinking water (17). Data on the contaminants identified by water sampling are presented in Table 1. Each census block was assigned a relative water exposure ranking. For example, those census blocks that did not receive public water and relied on private wells that could not have been contaminated with emissions from SRSNE received the lowest exposure rank. Those areas where hydrologic pressure would have been likely to send most of the contaminated water received the highest score. Those census blocks receiving the same exposure score were grouped for analysis.

The Southington Water Company provided information on the location, diameter, and elevation of pipes, and pipe junctures, which was then assigned digital map coordinates by the University of Connecticut's Department of Geography for use in the GIS. The Southington Water Company also provided information on the elevation and location of reservoirs, location of wells, and the proportion and quantity of the water supply

Table 2 Cancer Incidence by Gender, Southington, CT, 1968–1991

	Cancer Incidence in Total Population				Cancer Incidence in Females				Cancer Incidence in Males			
Cancer Type	**OBS**	**EXP**	**SIR**	**95% CI**	**OBS**	**EXP**	**SIR**	**95% CI**	**OBS**	**EXP**	**SIR**	**95% CI**
Bladder	134	131.60	1.02	0.71, 1.33	24	34.27	0.70	0.20, 1.20	110	99.51	1.11	0.74, 1.47
Hodgkin's disease	29	33.91	0.86	0.30, 1.41	8	14.88	0.54	0.00, 1.20	21	19.08	1.10	0.26, 1.94
Kidney	65	71.67	0.91	0.53, 1.29	23	25.02	0.92	0.25, 1.59	42	47.35	0.89	0.41, 1.36
Leukemia	80	85.92	0.93	0.57, 1.29	22	36.24	0.61	0.16, 1.06	58	50.22	1.15	0.63, 1.68
Liver	21	18.10	1.16	0.28, 2.04	10	5.80	1.72	0.00, 3.63	11	12.52	0.88	0.00, 1.80
Non-Hodgkin's lymphoma	80	92.75	0.86	0.53, 1.20	39	44.37	0.88	0.39, 1.37	41	48.62	0.84	0.38, 1.30
Testis	15	20.22	0.74	0.07, 1.41						20.22	0.74	0.09, 1.39
Southington	424	453.11	0.94	0.78, 1.09	126	160.58	0.78	0.54, 1.03	298	296.39	1.01	0.80, 1.21

OBS = Number of observed cancer cases
EXP = Number of expected cancer cases
SIR = Standard incidence ratios
CI = Confidence interval
Source: (20)

provided by these sources over the study period. DPH and the Southington Water Company provided information on water sampling for contaminants in the water distribution system.

The model relied on a US Environmental Protection Agency (EPA) computer software program, EPANET. EPANET tracks the flow of water within each pipe segment, the pressure at each pipe junction, the height of water in each reservoir or storage tank, and the concentration of a contaminant throughout a distribution system (17).

Estimated daily exposure to TCE in the public drinking water was broken into four water exposure categories and is presented in Table 3 and Figure 1. Two-thirds of the town was not impacted by water contamination from SRSNE or the Old Southington Landfill. Water level 1 is the lowest exposure category and 66.4% of the population lived in those portions of town. The areas northeast of the contaminated wells were estimated to receive the highest exposures. Water level 4 is the highest exposure category and no persons lived in this portion of the town. The shape and distance of the geographic areas impacted by water contamination was influenced by water usage, competing sources of water, and the hydrologic pressures in the water distribution system. The relative ranking of exposure to TCE in the drinking water by census blocks enabled the calculation of SIRs based on potential relative exposure to TCE in drinking water.

Table 3 Population and Contaminant Levels of Water Exposure Categories, Southington, CT

Water Exposure Category	Estimated TCE Level per Category (µg/L)	Population[a]	No. of Census Blocks
Southington		36,723	295
Level 1	No exposure	24,374	236
Level 2	1 to <10	7,186	39
Level 3	10 to <50	5,163	18
Level 4	50 or greater	0	2

[a]1980 census figures

µg/L = micrograms per liter

Source: (20)

Air Exposure Modeling

The development of an air contamination model was completed by Robert Tyler of SciTech Corporation of Wethersfield, Connecticut (18), and funded by ATSDR through this project. Air quality modeling of probable TCE emissions from the site was performed using the EPA Industrial Source Complex Long-Term (ISCLT2) model in conjunction with climatological data from the closest National Weather Service station in Hartford. Emission sources were identified through a review of records on SRSNE at the DEP, EPA, and DPH.

Standard emission factors were used to estimate emissions from the solvent reclamation process, and the receiving/storage and blend tanks. Air stripper emissions were calculated using mass balance equations based on groundwater flow rates and concentrations found in the groundwater and effluent. Engineering calculations were used to

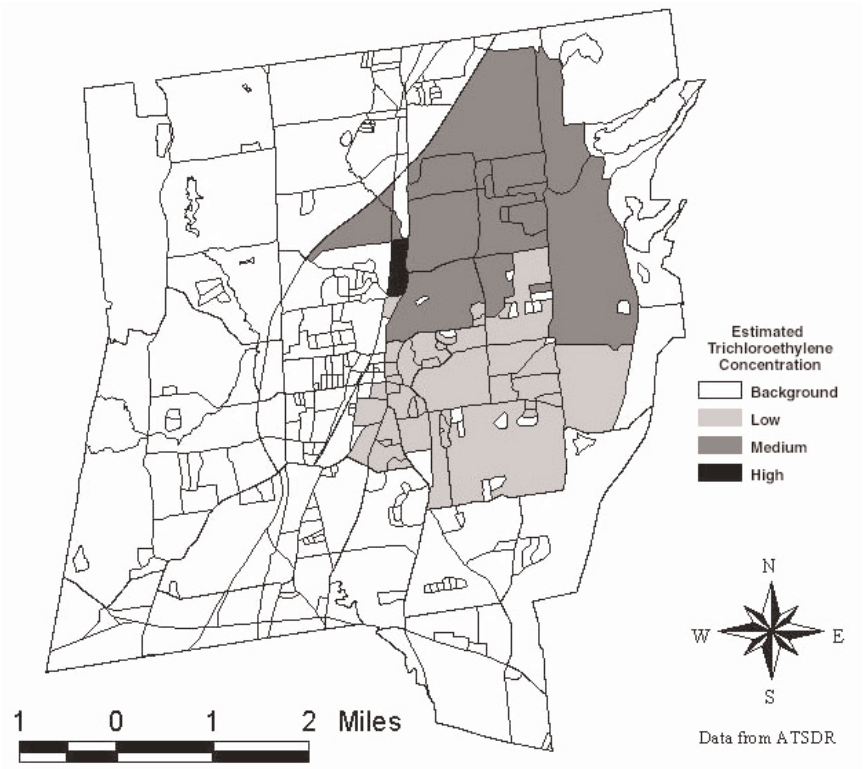


Figure 1 Geographic areas that received contaminated drinking water, Southington, CT, 1980.

estimate emissions from the lagoons and pit incineration. Calculations were based on data and equations provided by the EPA.

The overall air impact of TCE was estimated by summing the individual contributions of each of these sources during the study period and then calculating an average level for the 24-year study period. There were fluctuations in the values, and maximum short-term values were not calculated. TCE was chosen as the indicator pollutant to provide an estimate of the geographic area that was impacted by air emissions from SRSNE. There were, however, many other compounds that were handled by SRSNE that have the potential to be human carcinogens.

The modeling of air releases from SRSNE resulted in Southington being divided into four air exposure categories, which are presented in Table 4 and Figure 2. The majority of the town was not impacted by air emissions from SRSNE. Air level 1 is the lowest exposure category and those areas combined to account for 71% of the Southington population. The areas in closest geographic proximity to the site received the highest exposures. Air level 2 and level 3 combine to account for 29% of the Southington population. Air level 4 is the highest exposure category and no persons lived in this portion of the town. The shape and distance of the geographic areas impacted by emissions was influenced by the topography of the area and the prevailing winds.

Estimates of the Population at Risk

To calculate the SIRs, an estimate of the population at risk was made. The 1980 STF1b

tape from the US Bureau of the Census was used to provide age-specific information for each census block. The 1980 census was chosen for the population estimate because it was the midpoint of the study period. The census blocks with the same exposure scores

Table 4 Population and Contaminant Levels of Air Exposure Categories, Southington, CT

Water Exposure Category	Estimated TCE Level per Category (μg/m³)	Population[a]	No. of Census Blocks
Southington		36,723	295
Level 1	less than 0.01	25,895	200
Level 2	0.01 to <0.015	5,585	46
Level 3	0.015 to <0.10	5,243	46
Level 4	0.10 or greater	0	3

[a]1980 census figures

μg/m³ = micrograms per cubic meter

Source: (20)

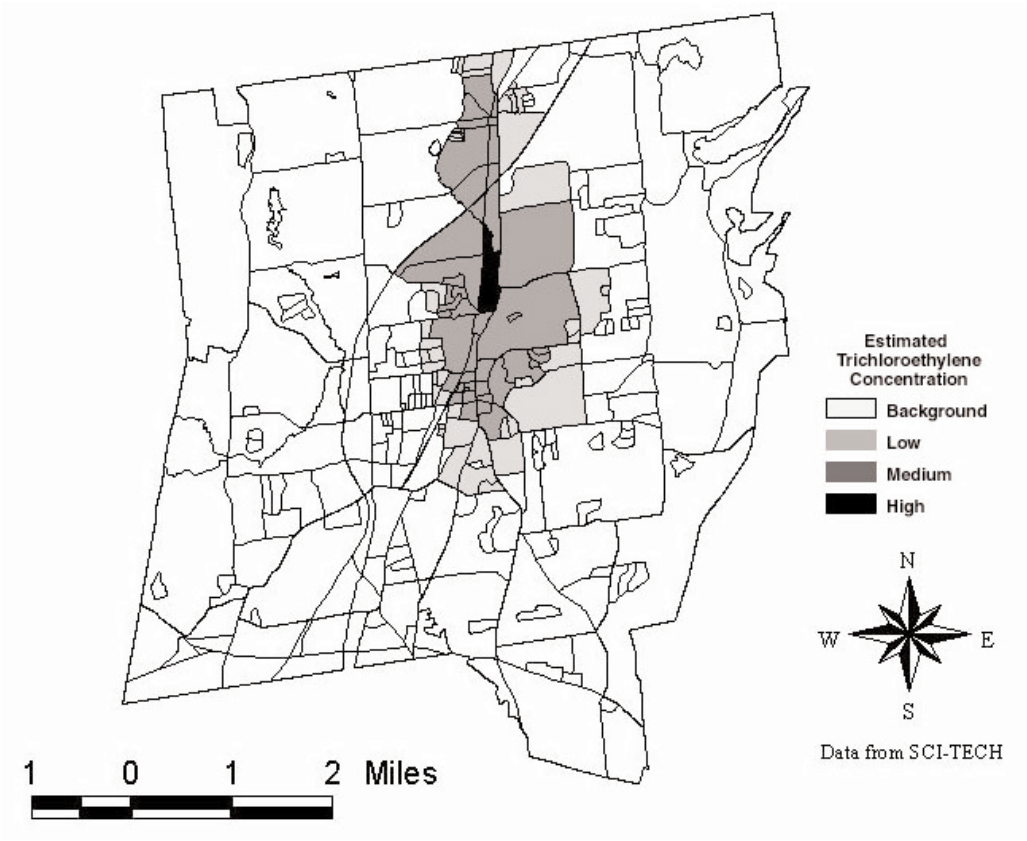


Figure 2 Geographic areas exposed to air emissions from solvents recovery systems of New England, Southington, CT, 1980.

were summed to derive population estimates for the various air and water exposure categories.

Analysis

The indirect method of age standardization and State of Connecticut incidence rates from the DPH Tumor Registry were used to calculate age SIRs. SIRs were calculated for each tumor site and all of the tumor sites combined for each drinking water and air exposure ranking. The indirect rather than direct method of age standardization was chosen for these analyses because the number of cases was small, resulting in incidence rates that would be too unstable for age standardization using the direct method. A Bonferroni correction was made to adjust for multiple comparisons (19).

Results

No statistically significant increase in the SIR was found for cancer of the bladder, kidney, liver, or testis, leukemia, NHL, or Hodgkin's disease for any of the exposure categories when compared with State of Connecticut incidence rates. The total SIR of all tumor sites combined demonstrated a statistically significant increasing trend in relation to increasing exposure to air emissions. For males and females combined, the SIRs were as follows (see Table 5):

- Air level 1: SIR=0.89 (95% confidence interval, 0.70, 1.08)
- Air level 2: SIR=0.99 (95% CI, 0.62, 1.37)
- Air level 3: SIR=1.04 (95% CI, 0.62, 1.47)

Among individual tumor sites, the risk of NHL among females was elevated in locations where the air exposure levels were estimated to be the greatest. Among females, the SIRs were as follows (see Table 6):

- Air level 1: SIR=0.59 (95% CI, 0.09, 1.09)
- Air level 2: SIR=0.59 (95% CI, 0.00, 1.50)
- Air level 3: SIR=2.42 (95% CI, 0.37, 4.46)

The elevation of risk was not consistently shown among males.

The results of the study are presented in more detail in the complete report of the study (20).

Discussion

This study was conducted in response to citizen concerns that persons living in neighborhoods near the SRSNE Superfund site were experiencing higher rates of cancer than was the general population. It was known that these persons had been exposed to emissions from SRSNE, but it was less clear whether the rate of cancer was higher among them than would normally be expected.

Traditionally, cancer rates are calculated for geographic areas with specific political boundaries such as a state or town. Use of a GIS allows disease rates to be calculated at a smaller geographic level. In this study the census block is the unit of analysis.

The GIS was used to improve the exposure assessment and locate each case on a map (geocoding). The mechanization of the geocoding process by the GIS allowed the study to be conducted in a more efficient manner than would have been possible with

Table 5 Cancer Incidence by Gender for Air Exposure Categories, Southington, CT, 1968–1991

	Cancer Incidence in Total Population				Cancer Incidence in Females				Cancer Incidence in Males			
Cancer Type	**OBS**	**EXP**	**SIR**	**95% CI**	**OBS**	**EXP**	**SIR**	**95% CI**	**OBS**	**EXP**	**SIR**	**95% CI**
Southington	424	453.11	0.94	0.78, 1.09	126	160.58	0.78	0.54, 1.03	298	296.39	1.01	0.80, 1.21
Air level 1	265	298.41	0.89	0.70, 1.08	73	105.13	0.69	0.41, 0.98	192	196.21	0.98	0.73, 1.22
Air level 2	84	84.53	0.99	0.62, 1.37	27	30.28	0.89	0.29, 1.49	57	54.90	1.04	0.56, 1.52
Air level 3	73	70.17	1.04	0.62, 1.47	26	25.17	1.03	0.33, 1.74	47	45.27	1.04	0.51, 1.57
Air level 2+3 (any air exposure)	157	154.70	1.01	0.73, 1.30	53	55.45	0.96	0.50, 1.41	104	100.18	1.04	0.68, 1.39

OBS = Number of observed cancer cases
EXP = Number of expected cancer cases
SIR = Standard incidence ratios
CI = Confidence interval
Source: (20)

Table 6 Incidence of Non-Hodgkin's Lymphoma (NHL) by Gender for Air Exposure Categories, Southington, CT, 1968–1991

	NHL Incidence in Total Population				NHL Incidence in Females				NHL Incidence in Males			
	OBS	**EXP**	**SIR**	**95% CI**	**OBS**	**EXP**	**SIR**	**95% CI**	**OBS**	**EXP**	**SIR**	**95% CI**
Southington	80	92.75	0.86	0.53, 1.20	39	44.37	0.88	0.39, 1.37	41	48.62	0.84	0.38, 1.30
Air level 1	44	61.08	0.72	0.34, 1.10	17	28.79	0.59	0.09, 1.09	27	32.45	0.83	0.27, 1.39
Air level 2	13	17.34	0.75	0.02, 1.48	5	8.54	0.59	0.00, 1.50	8	8.85	0.90	0.00, 2.02
Air level 3	22	14.33	1.54	0.39, 2.68	17	7.03	2.42	0.37, 4.46	5	7.32	0.68	0.00, 1.75
Air level 2+3 (any air exposure)	35	31.67	1.11	0.45, 1.76	22	15.57	1.41	0.39, 2.46	13	16.17	0.80	0.03, 1.58

OBS = Number of observed cancer cases
EXP = Number of expected cancer cases
SIR = Standard incidence ratios
CI = Confidence interval
Source: (20)

mapping the location of the cases manually. This enabled the evaluation of whether there was an association between exposure to emissions through the air or public drinking water and the incidence of selected types of cancer.

Census block areas with the same relative exposure rankings were grouped for analysis. Air and water exposures were evaluated separately. Age SIRs were calculated for each tumor site by gender and by relative measure of exposure to TCE emissions in the air and the drinking water. No statistically significant increase in the SIR was found for cancer of the bladder, liver, kidney, or testis, leukemia or Hodgkin's disease for any of the exposure categories when compared with the state of Connecticut. For the total of all tumor sites combined, there was an increase in the SIR for increasing exposure to air contaminants.

The rate of NHL among females was elevated where the air emission rates were estimated to be the highest, although this increase did not achieve statistical significance. Seventeen cases of NHL were diagnosed among women where seven cases would have been expected to occur during this same time period. A similar elevation did not occur among men.

Non-Hodgkin's lymphoma has been increasing dramatically over the past several decades with no clear explanation for the increase. A review of the epidemiology of NHL in Connecticut documented not only the increase in NHL, but also that the histology of the tumors is tending to change. Proportionally, more cases of nodular NHL are occurring. The ratio of diffuse cell type to nodular cell type among women had decreased from 9:1 to about 3:1 from the 1960s to the 1980s (21). Of the 17 women with NHL in the high air exposure area, only 9 had diffuse cell type and 8 had nodular cell type (ratio of 1.1:1). For cases of NHL among women in the unexposed portion of town, 18 had diffuse cell type and only 4 of the 22 cases had nodular cell type (ratio of 4.5:1). Therefore, in the higher air exposure area of this study more of the cases among women are nodular NHL, the same histology of NHL that is increasing in incidence in Connecticut. Previous studies of environmental contamination have shown associations between exposure to solvents and incidence of NHL (14).

These data suggest that women living in the portion of town exposed to emissions from SRSNE through the public water distribution system and the air did experience an increased risk of NHL. Women may have been at their home for a higher portion of the day and therefore experienced a higher exposure than men. However, this study cannot determine whether or not exposures to emissions from SRSNE in the air or water caused any cases of NHL or any other cancer in the town of Southington.

The increasing incidence of NHL over the past few decades supports the need to conduct additional epidemiological research into the possible role of environmental exposure to solvents as a possible risk factor in the development of NHL.

This study of cancer incidence in Southington, Connecticut, has several limitations. While it does use cancer incidence information from the Tumor Registry, these data contain only limited individual data including gender, date of birth, and date of diagnosis. Other relevant risk factors including smoking, family history, and occupation are not routinely available. While Tumor Registry case ascertainment is considered to be very complete, people could have been exposed and then moved from the area prior to diagnosis of their disease.

The population estimates for calculation of the risk measurements were derived from the 1980 census data and do not take growth of the population or migration into

consideration. The population of Southington has grown from 30,746 in 1970, to 36,879 in 1980, to 38,501 in 1990. The 1980 date was selected because it represents the midpoint of the study period.

It must be kept in mind that as an ecological study design this can only be considered hypothesis generating, and that this type of study is not intended to demonstrate a causal relationship (22).

The air exposure estimates are derived from actual chemical use information from the EPA, DEP, and SRSNE company records. Engineering assumptions are used, however, to estimate the amount of TCE released into the air, and average meteorological conditions are used to assist in the estimation of the TCE dispersion.

The water exposure estimates are derived from monitoring data, water usage data, and pipe characteristics. However, water exposure and consumption information for individuals is not known. TCE was modeled as the indicator contaminant because it was found in both the drinking water and air, but it represents only a qualitative indication of the geographic areas in Southington most likely to be impacted by contaminants.

Acknowledgments

This project was funded under a cooperative agreement with ATSDR, U50/ATU199044. We would like to acknowledge the support and involvement of the following people: Marie Tuccitto brought her concerns about the health of the residents of Southington to our attention at the Connecticut Department of Public Health. Morris Maslia of ATSDR worked with Georgia Tech to complete the model of drinking water exposures. Robert Tyler of SciTech developed the air exposure model. Ellen Cromley, an associate professor at the University of Connecticut Department of Geography, worked with graduate student Richard Mrozinski to digitize the public water system so that the water exposure model could be developed. Virginia Lee, a medical officer at ATSDR, provided technical support and guidance during the project. Jennifer Kertanis and Carolyn Jean Dupuy, epidemiologists with DPH, assisted in the review and editing of the document. Michael Knapp, a former epidemiologist with DPH, provided assistance with statistical interpretation of data. Brian Toal, David Brown, Peter Galbraith, and Mary Lou Fleissner of DPH provided supervision and technical advice during the development and completion of the study. Jan AJ Stolwijk, Susan T Mayne, and Mark Wilson, all from Yale University, provided assistance on study design.

References

1. Southington Water Company. 1992. Records on water contaminants, water pumping, and pipe characteristics. Southington, CT.
2. Mallin K. 1990. Investigation of a bladder cancer cluster in northwestern Illinois. *American Journal of Epidemiology* 132:s96–s106.0
3. Silverman DT, Hartge P, Morrison AS, Devesa SS. 1992. Epidemiology of bladder cancer. *Journal of Hematology/Oncology Clinics of North America* 6:1–30.
4. Roush GC, Holford TR, Schymura MJ, White C. 1987. *Cancer risk and incidence trends: The Connecticut perspective*. New York: Hemisphere Publishing Co.

5. Agency for Toxic Substances and Disease Registry (ATSDR). 1994. *Toxicological profile for carbon tetrachloride.* Atlanta: US Dept. of Health and Human Services, Public Health Service.
6. Agency for Toxic Substances and Disease Registry (ATSDR). 1993. *Toxicological profile for trichloroethylene*. Atlanta: US Dept. of Health and Human Services, Public Health Service.
7. Banks P. 1990. The pathology of Hodgkin's disease. *Seminars in Oncology* 17:683–95.
8. Urba WJ, Longo DL. 1992. Hodgkin's disease. *New England Journal of Medicine* 236:678–87.
9. Hartge P, Devesa SS, Fraumeni JF. 1994. Hodgkin's and non-Hodgkin's lymphomas. *Cancer Surveys*. 19/20:423–52.
10. Persson B, Fredriksson M, Olsen K, Boeryd B, Axelson O. 1993. Some occupational exposures as risk factors for malignant lymphomas. *Cancer* 72:1773–78.
11. Newell GR, Mills PK, Johnson DE. 1984. Epidemiologic comparison of cancer of the testis and Hodgkin's disease among young males. *Cancer* 54:1117–23.
12. Lagakos SW, Wessen BJ, Zelen M. 1996. An analysis of contaminated well water and health effects in Woburn, Massachusetts. *Journal of the American Statistical Association* 81:583–96.
13. Fagliano J, Berry M, Bove F, Burke T. 1990. Drinking water contamination and the incidence of leukemia: an ecologic study. *American Journal of Public Health* 80:1209–12.
14. Cohen P, Klotz J, Bove F, Berkowitz M, Fagliano J. 1994. Drinking water contamination and the incidence of leukemia and non-Hodgkin's lymphoma. *Environmental Health Perspectives* 102:556–61.
15. Stuver SO, Trichopoulos D. 1994. Liver cancer. *Cancer Surveys* 19/20:99–124.
16. Agency for Toxic Substances and Disease Registry (ATSDR). 1995. *Draft toxicological profile for benzene*. Atlanta: US Dept. of Health and Human Services, Public Health Service.
17. Agency for Toxic Substances and Disease Registry (ATSDR). 1994. *A public health analysis of exposure to contaminated municipal water supplies at Southington, Hartford County, CT*. Atlanta: US Dept. of Health and Human Services, Public Health Service. 20 December.
18. Tyler R. 1994. *Evaluation of the carcinogenic air impacts from Solvents Recovery Services of New England*. SciTech Corporation. September.
19. McClave J, Dietrich FH. 1988. *Statistics.* San Francisco: Dellen Publishing Co.
20. Aye D, Archambault G. 1998. *Cancer incidence in Southington, CT, 1968–1991, in relation to emissions from Solvents Recovery Services of New England.* US Dept. of Health and Human Services.
21. Zheng T, Mayne ST, Boyle P, Holford TR, Liu WL, Flannery J. 1992. Epidemiology of non-Hodgkin lymphoma in Connecticut, 1935–1988. *Cancer* 70:840–9.
22. Walter SD. 1991. The ecologic method in the study of environmental health; I: Overview of the method. *Environmental Health Perspectives* 94:61–73.

Analyzing Motor Vehicle Injuries with the Connecticut Crash Outcome Data Evaluation System GIS

Ellen K Cromley (1),* Mary Kapp (2), Brian R Pope (1)
(1) Department of Geography, University of Connecticut, Storrs, CT; (2) Connecticut Department of Public Health, Hartford, CT

Abstract

The Connecticut Crash Outcome Data Evaluation System (CODES) geographic information system (GIS) is a statewide GIS application developed by the Injury Prevention Program of the Connecticut Department of Public Health for viewing, analyzing, and reporting information on motor vehicle collisions and the medical care provided to persons injured in them. The Connecticut CODES Project was funded by the National Highway Traffic Safety Administration in 1997. Connecticut is one of a number of states funded to link medical outcome data with motor vehicle collision data. By linking collision, vehicle, and human behavior characteristics to their specific medical and financial outcomes, the project can identify prevention factors. The GIS component of the project uses collision data for 1995 and 1996 from police accident reports coded by the Connecticut Department of Transportation, as well as hospital discharge and emergency department data from the Connecticut Healthcare Research and Education Foundation. The GIS stores these data in a relational database that links to a GIS database of collision locations. The application supports a wide range of GIS functions, including geocoding, querying, and color mapping, through a specially designed user interface. The Injury Prevention Program is using the CODES GIS to identify high-frequency collision locations and to evaluate the effectiveness of safety belts, child passenger safety seats, and motorcycle helmets in preventing (and reducing the severity of) injuries and deaths resulting from motor vehicle collisions. A public-use version of the application and databases enables other stakeholders to retrieve, analyze, and map data of special interest.

Keywords: motor vehicle injury, injury surveillance, injury prevention, data linkage

Introduction

Efforts to reduce deaths and injuries from motor vehicle collisions by improving protection systems like safety belts have been effective, but injuries resulting from crashes continue to be a major public health problem. The focus of government agencies responsible for improving highway safety has broadened from documenting the occurrence of injuries to monitoring the injuries and the subsequent medical care outcomes and health care costs, to establish priorities for prevention. A congressional mandate was included in the Intermodal Surface Transportation Efficiency Act of 1991 calling for a study of the benefits of safety belts and motorcycle helmets. According to the mandate, this study was to go beyond the analysis of fatal injuries to document the severity of non-fatal injuries and the medical care costs associated with treating them.

* Ellen K Cromley, Dept. of Geography, University of Connecticut, U-148, 354 Mansfield Rd., Storrs, CT 06269-2148 USA; (p) 860-486-3656; (f) 860-486-1348; E-mail: ecromley@uconnvm.uconn.edu

Beginning in 1992, the National Highway Traffic Safety Administration (NHTSA) made grants to seven states to implement Crash Outcome Data Evaluation Systems (CODES) projects that would link motor vehicle crash data collected at the state level with medical outcome data (1). In 1997, the Connecticut Department of Public Health received an award to develop a CODES project. This paper describes the Connecticut CODES geographic information system (GIS), a system designed to provide an environment for viewing and analyzing the linked motor vehicle and medical outcome database created for the Connecticut CODES project.

An important NHTSA requirement for CODES projects is creation of a public-use version of the linked database. Although GIS analysis was not specifically required by NHTSA, a number of states have chosen to develop GIS components as part of their CODES projects. This choice reflects a growing interest in the role that GIS might play in motor vehicle injury analysis in the United States and in other countries (2,3). Using GIS in the study of motor vehicle injury supports injury surveillance programs that monitor injury patterns by type and location of occurrence, supports epidemiological analysis through development of more accurate numerators and denominators for particular kinds of crash events, and supports implementation and evaluation of site-specific intervention strategies.

Materials and Methods

The Connecticut CODES project links statewide crash data for 1995 and 1996 to emergency department, hospital inpatient, and trauma records maintained by the Connecticut Health Research and Education Foundation, as well as to mortality records maintained by the Vital Records Section of the Connecticut Department of Public Health. The crash data are compiled by the Connecticut Department of Transportation (DOT) Accident Records Section from police accident reports made by local officers responding to motor vehicle collisions. The data available for 1995 and 1996 include all motor vehicle collisions occurring on state roads, as well as those collisions occurring on local roads for which the responding officer indicated that the collision had resulted in an injury. The medical record linking is an automated process relying on a probabilistic record-linking software package used by all of the funded CODES projects.

The Connecticut CODES GIS database design includes tables of motor vehicle collision attributes, including the linked data, managed in a Microsoft Access relational database, and databases of collision locations, managed as ArcView (ESRI, Redlands, CA) shapefiles. A unique identifying number links the collision attributes to the points representing their locations. The Planning Section of the Connecticut DOT provided latitude/longitude geocodes for all collisions occurring on state roads. CODES GIS project staff geocoded the collisions that occurred on local roads and projected all crash locations based on the Connecticut Coordinate System, a system of state plane coordinates.

A GIS database of collision locations would not be of much use alone. Additional data incorporated into the Connecticut CODES GIS are a 1995 state road network datalayer from the Connecticut DOT Planning Section and town boundary and annotation layers from the Connecticut Department of Environmental Protection. The 1990 TIGER address-ranged street network database from the US Census Bureau can be used to support finding particular addresses and intersections. For viewing crash data, digital

raster graphics from the US Geological Survey of 7.5-minute quadrangle maps for Connecticut can be used as an alternative to the 1995 state road network datalayer.

Because the Connecticut CODES database is a linked database compiled from records maintained by a variety of agencies, CODES database users do not edit data. Instead, the Connecticut CODES GIS is designed to provide users with a tool for easily displaying, querying, and mapping crash data. These functions are especially important because they support distribution of the public-use version of the database.

A GIS application designed specially for the Connecticut CODES project was created to modify the GIS software package purchased from the vendor. This application cannot be modified by the user. It allows the user to search the linked database by asking "what" or "where" questions (for example, "What are the collisions of interest and where did they occur?" or "Where are the places of interest and what kinds of collisions occurred there?"). The application supports a number of important functions required to answer these questions:

- Displays road network, town boundary, and annotation layers.
- Loads one or more years of crash data selected by the user or a user-defined database and joins and links attribute tables.
- Toggles annotation on/off.
- Toggles digital raster graphics on and off.
- Hides/shows table of contents and legends for motor vehicle crash and other databases.
- Identifies collisions where the user clicks.
- Labels collisions by road name (for road where collision occurred or intersecting road).
- Labels a collision site with a count of collisions that occurred at that location in a particular year.
- Finds and pans to a crash location based on a CODES identifier entered by the user.
- Finds and highlights crashes with characteristics identified by the user.
- Provides a range of tools for graphical selection of collisions at a point, within a rectangle, or within a buffered segment of the road network.
- Prints reports of the attributes of selected collisions.
- Prints color maps of crash locations based on a standard map layout, the display created by the user, and title text entered by the user.
- Adds the address-ranged street network database for the town selected by the user and supports finding a particular address or intersection.

Results

An example of the kind of analysis that would be supported by the Connecticut CODES GIS is an investigation of all collisions in 1995 that involved pedestrians and happened in Norwich, Connecticut. This use of the system would begin with a defined set of collisions of interest. The system can be used to develop numerators and denominators for the town as a whole and for the particular collision of interest. For example, in 1995, motor vehicle injuries in Norwich involving pedestrians represented only a small

percentage of all motor vehicle crashes that occurred, but they were more likely to have linked medical records (Table 1).

Table 1 Comparison of Total and Pedestrian Collisions in Norwich, Connecticut, in 1995

	All Types	Involving Pedestrians
Number of collisions	1,163	24
At least one linked medical record	357	17
No linked medical record	806	7

Using the Connecticut CODES GIS, maps and reports of collisions of interest can be prepared. The standard map output (Figure 1) shows that 3 of the 24 pedestrian collisions of interest occurred near the intersection of Connecticut Route 2 and Talman Street, in the context of 7 other collisions that occurred at that location. The collision count and street name labels are shown on the map.

The report on selected crashes provides information based on a set of variables selected by the user. In this example, seven fields were chosen for the report, including the collision identifier number, a variable that indicates whether at least one person involved in the collision had a linked hospital record, and other basic collision attribute information, including the weekday and time of occurrence, the number of vehicles and pedestrians involved, and the collision type. Figure 2 shows a sample report.

In the public-use version of the database, only limited information is provided from the linked medical record data. Basically, users can determine whether linked medical or mortality records were found for individuals involved in a collision and can ascertain the general nature of the injuries and medical care costs associated with treating them. More detailed medical information is available in the full Connecticut CODES database.

Discussion

The Connecticut CODES GIS is advancing the analysis of motor vehicle injury by allowing analysts to select and map collisions based on a complex set of variables that capture the multidimensional characteristics of collision events. The statewide, multiyear, linked database makes it possible to distinguish, for example, the environmental settings in which motorcyclists are involved in collisions with other vehicles, in contrast to those involving only the motorcyclist, as well as the contributing factors associated with those collisions and the different medical care outcomes of the collisions.

In addition to providing a platform for more meaningful epidemiological studies of motor vehicle injury, the Connecticut CODES project, with its GIS component, provides public health professionals, law enforcement and transportation officials, the general public, and other stakeholders a rich database for designing public health interventions to address the problem of motor vehicle injury in the state. Distribution of the data through a public-use version of the database (which protects confidentiality of medical record information) creates an opportunity for interested parties at the local level to develop prevention strategies based on local needs that may not be reflected in statewide priorities for injury prevention. Armed with data on the types of collisions that are

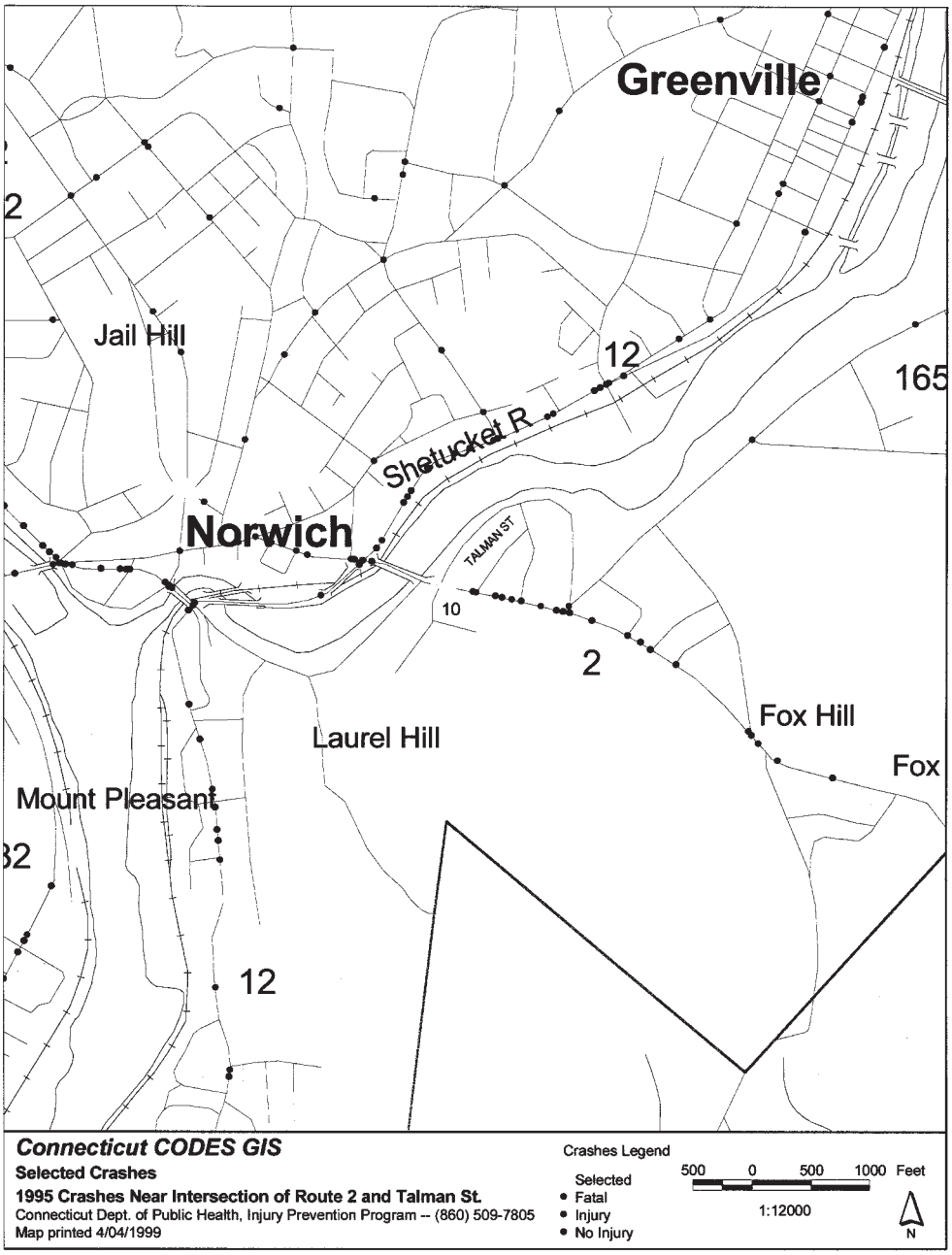


Figure 1 Sample map output from the Connecticut CODES GIS.

Report of Selected Collision Attribute Information

1995 Crashes

CODES_ID	HOSPLINK	WEEKDAY	ACCTIME	NUMVEH	NUMPED	COLLISTYPE
1995101781		Saturday	1615	2	0	Rear-end
1995125135		Thursday	1419	2	0	Turning - Intersecting Paths
1995132971	Yes	Monday	1155	1	1	Pedestrian
1995136910	Yes	Saturday	1200	2	0	Turning - Same Direction
1995139301	Yes	Thursday	2112	1	1	Pedestrian
1995144234	Yes	Wednesday	0705	2	0	Turning - Same Direction
1995145310		Tuesday	1405	1	1	Pedestrian
1995147975	Yes	Monday	1500	2	0	Rear-end
1995152802		Saturday	0900	2	0	Rear-end
1995155309		Thursday	1958	2	0	Rear-end

Figure 2 Sample report output from the Connecticut CODES GIS.

occurring at high-frequency collision locations in each locality, officials can begin to design intervention programs that tailor operator and pedestrian education, traffic enforcement, and environmental modifications to the particular types of collisions occurring at particular sites.

Acknowledgments

This research was funded by the National Highway Traffic Safety Administration, US Department of Transportation.

References

1. US Department of Transportation. 1996. *The crash outcome data evaluation system (CODES).* DOT HS 808 338. Washington, DC: US Department of Transportation.
2. Braddock ME, Lapidus G, Cromley E, Cromley R, Burke G, Banco L. 1994. Using a geographic information system to understand child pedestrian injury. *American Journal of Public Health* 74:1157–61.
3. Raybould S, Walsh S. 1995. Road traffic accidents involving children in north-east England. In: *The added value of geographical information systems in public and environmental health.* Ed. MJC DeLepper, HJ Scholten, RM Stern. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Using a Proximity Filter to Improve Rabies Surveillance Data

Andrew Curtis*
Louisiana State University, Baton Rouge, LA

Abstract

The proximity filter is one of the new exploratory spatial data analysis techniques developed as a result of the visual and interactive capabilities of geographic information systems (GIS). This technique, a variant of the spatial filter, is used to identify significant "holes" in a point surface. This paper presents an example of how the proximity filter can be applied to real-world situations, showing how it can improve rabies surveillance data. The points used in this example are the locations of animals submitted for rabies testing. It is important to identify any holes in a rabies surveillance data surface to see if the low occurrence of points (i.e., of animals submitted) results from natural reasons or from a lack of local education. The proximity filter compares the number of points in an inner ring to those in an outer ring. A Monte Carlo simulation allows for a test of significance to be constructed. A GIS is used to vary the size and shape of the inner and outer rings used in the analysis. Two types of filter shape are discussed; the first is circular, and the second uses the buffer function of the GIS to mirror the shape of an investigated area (such as a county). This paper also provides an example of how data aggregated to a political unit can be turned into a simulated point pattern surface so that the proximity filter can be applied.

Keywords: exploratory spatial data analysis, modifiable areal unit problem, Monte Carlo, point pattern analysis, spatial filter

Background

Recent discussion has focused on the role geographic information systems (GIS) can play in improving spatial analysis (1). One area in which such advancement can be made is exploratory spatial data analysis—the use of a GIS' interactive and visual capabilities to search for problem solutions on the fly (2). Although this practice is still in its infancy, new techniques of analysis that can only be implemented within a GIS environment are now being developed (for a review of these techniques, see reference 3). A good example of these techniques is the spatial filter, a form of analysis that avoids the problems associated with data aggregated to political units (county, zip code, census tract, etc.). Diseases are often continuous across space, and therefore, if the data are available at a disaggregated level, it makes little sense to impose artificial areas (defined by political boundaries) on the analysis. An additional problem in using aggregated data is one of aggregation (scale) and shape (zone)—otherwise known as the modifiable areal unit problem. It has been found that different results can occur depending on which scale and zone scheme is chosen (4).

* Andrew Curtis, Louisiana State University, 110 Howe/Russell Geoscience Complex, Baton Rouge, LA 70803 USA; (p) 225-388-6198; (f) 225-388-4420; E-mail: acurti1@lsu.edu

The spatial filter avoids the modifiable areal unit problem by using original point data (5,6). A fine grid is layered onto the research surface, with the intersection points of the grid acting as foci for a series of overlapping "filters" that are used to calculate "rates" from the point data. These rates, when assigned to the intersection point, can then be mapped, usually as a contour surface. This creates a visual impression of clusters from a continuous surface. Different-sized filters will "smooth" the data to different degrees, but the most dramatic cluster groupings should emerge irrespective of filter size. This form of analysis is truly "exploratory" because the visualization of the data drives the analysis. A test of significance can then be conducted using a Monte Carlo simulation. In such a test, the original points are assigned probabilities of occurrence (such as the chance that a particular birth will become a death). The same filter analysis is then performed on the simulation surface. Through multiple repetitions of this simulation, a distribution can be created against which the original clusters can be compared to see their probability of occurrence.

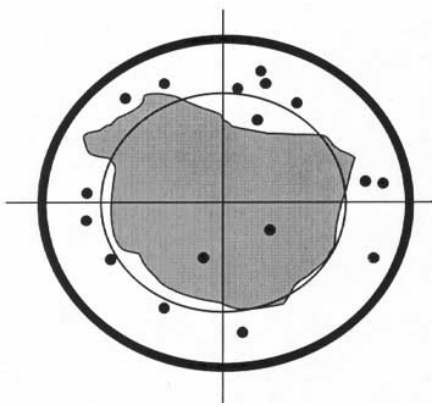
Although the spatial filter is useful for identifying significant clusters on a point surface, the research problem for this paper required a technique that could find significant "holes" in a point surface. A variation of the spatial filter was needed—the proximity filter, which can use a GIS to compare two distributions of points and identify significant differences between them.

Rabies Surveillance Data

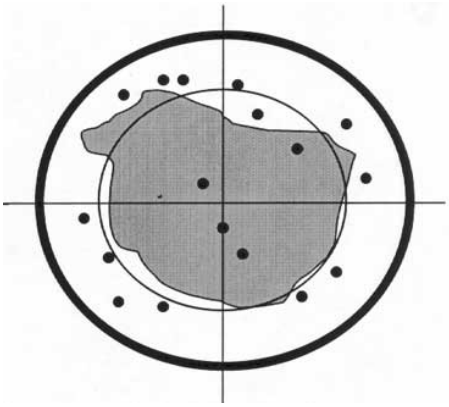
A raccoon rabies epizootic has been spreading through the eastern seaboard states since 1977. The epizootic started when rabid raccoons were translocated to West Virginia/Virginia by hunters (7). Since then, most of the eastern seaboard states have been affected by the spread (8). In a state impacted by rabies, surveillance data provide the main source of information about the disease. Both the public and local officials are supposed to submit animals for rabies testing if they interact with people (i.e., bite or scratch them) and cannot be quarantined, or if they display potential rabies symptoms. Data from this testing are used to determine how far the disease has progressed in the state and where to place countermeasures such as oral vaccine barriers. These data are also used to determine the success of these countermeasures. Unfortunately, evidence suggests that data quality varies considerably between areas (9). There is, therefore, a need for a means of analysis that can identify areas where fewer animals are submitted for rabies testing so that resources can be targeted to "educate" both the public and local officials.

The Proximity Filter

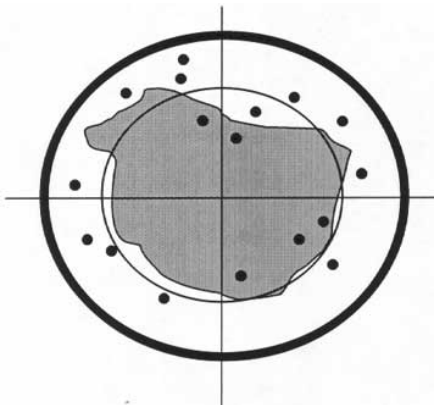
The general principle of the proximity filter is the same as that of the spatial filter. It uses a floating kernel to analyze a point data surface. The main difference between the proximity and spatial filters is that the former compares an inner and outer spatial point distribution. Figure 1a shows a series of points (the locations of animals submitted for rabies testing), the boundary of the county under investigation, and an inner and outer ring of the spatial filter (centered on the cross hair). In this example, there are 17 points in the inner and outer rings combined. The number of points in the inner ring (in this case, three) can be compared with the number in the outer ring (in this case, 14) to



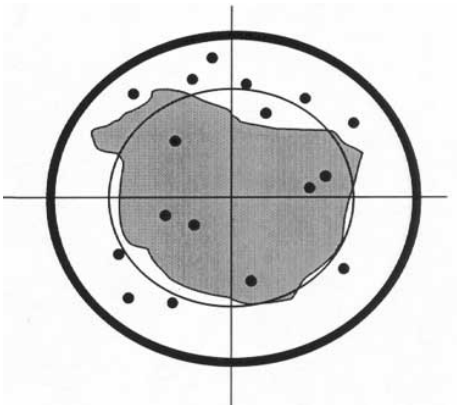
1a Actual point data.



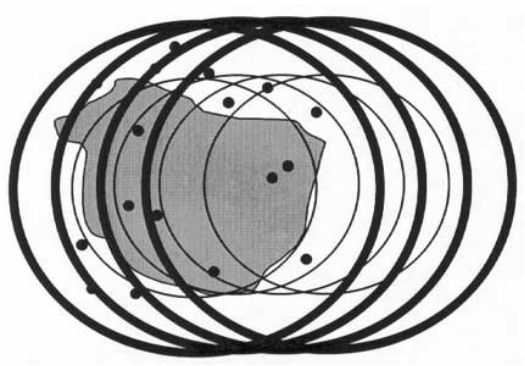
1b Simulation run #1.



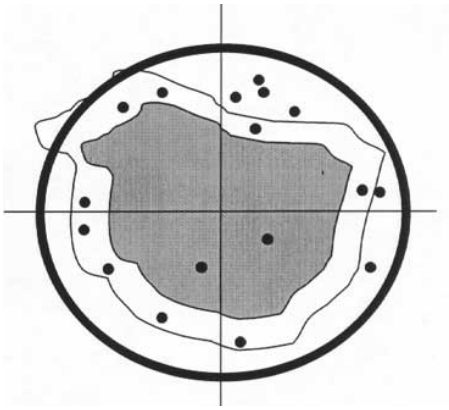
1c Simulation run #2.



1d Simulation run #3.



1e Overlapping proximity filter analysis.



1f Using the GIS buffer function to create the outer ring.

Figure 1 Applying the proximity filter to point data.

determine if, given the latter number, there are significantly fewer points in the inner ring than expected. GIS allows us to answer this question with a Monte Carlo simulation. A fine grid of coordinates is layered over both the inner and outer rings. All the points are then randomly redistributed across the combined area, with each coordinate having an equal chance of accepting one point. This simulation is repeated 100 times. If three or fewer points occur in the inner ring on no more than five of the simulation runs, then there is (at least) a 95% chance that the low number of points in the inner ring did not occur by chance. Figures 1b, 1c, and 1d show three typical simulation runs, with five, six, and seven points being placed in the inner circle.

The size of the outer ring obviously affects the number of points being compared with the inner ring—the larger the outer ring, the more points will be included in the analysis. One possible general rule is to use a comparable land area in the analysis; that is, the land area of the outer and inner rings should be the same. In order to avoid any bias in selecting the centroids of the filters, overlapping proximity filters should cover the entire area (Figure 1e) as in traditional spatial filter analysis. The rings should be centered on a fine grid of coordinates. Repeating the same simulation procedure for each inner and outer ring makes it possible to identify significant data holes across the entire surface.

A further modification of the proximity filter could involve using the shape of the county as the inner ring. If it is believed that this particular political boundary does exert some influence on the analysis—one county, for example, may have officials who are less compliant with submission laws—then it may be useful to include the shape of the county in the analysis. One way to do this is to leave the outer ring the same, but use the shape of the county as the inner ring. This produces changes like those shown in Figure 1f (using the same original point distribution as in Figure 1a), in which there are 2 points in the inner ring and 15 in the outer ring. (Note: Larger county-shaped outline not applicable here.) The simulation procedure described above would be used to create a test of significance.

If the shapes of the counties vary considerably (e.g., if some are compact and others elongated), then the buffer function of a GIS can be used to mirror the shape of the investigated county, making the outer ring a projected extension of the county shape. Figure 1f (mentioned above as showing a county-shaped inner ring with a circular outer ring) is also an example of this method. In it, there are nine points in the county-shaped outer ring. Again, the size of the buffer affects the number of points in the outer distribution; a possible general rule is to use the same land area for both buffer and investigated county.

The size and shape of the proximity filter depends on the background of the analysis. With no prior reason to suspect a particular county of low compliance, a complete coverage of overlapping filters is best. In the case of animals submitted for rabies testing, if there is reason to suspect a political unit such as the county, then the shape of the political unit could be used as the inner ring of the proximity filter, with either a circular or a county-shaped buffer as the outer ring.

It must be stressed that the proximity filter only works if combined with local expertise. Once a significant point hole has been identified, the next stage of the analysis is to investigate causative factors. These can include a difference in terrain between the inner and outer rings (the most extreme example being that no animal submissions are coming from the inner ring because it is a lake). A difference in human populations (the

fewer the people, the less likely the human/animal interaction) or a different environment leading to human/animal interaction (e.g., the outer proximity filter containing a state park) could also explain away significantly low point counts for the inner ring. All of these factors can be investigated after the initial analysis, and some of the holes discounted accordingly.

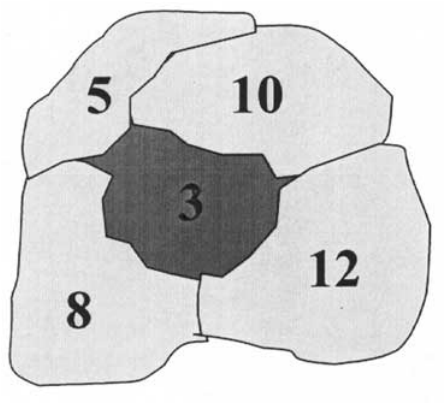
Using a Proximity Filter on Aggregated Data

Unfortunately, not all states have precise spatial locations allocated to animals submitted for testing. This is the case for Kentucky, where these data are aggregated to the county from which the submission came. Obviously, this makes it difficult to apply the filter. The only solution for cases such as this is to distribute the number of animal submissions across the host county at random.

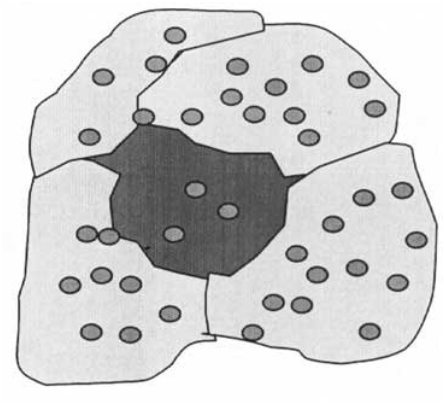
This is easy to do using a GIS. A fine grid of coordinates is layered over the county. All the points (animals submitted for testing) are then randomly distributed across this surface, with each coordinate having an equal chance of accepting a point. Once the entire surface has been covered in this way, the proximity filter can be layered on top of "suspect" counties. There is little point in applying an overlapping filter analysis because the lack of precise locations already produces error.

Figure 2 shows the typical steps of this type of analysis. The aggregated total number of animals submitted for rabies testing is randomly distributed across the county space as a set of points, using a fine lattice of coordinates in the GIS (Figure 2b). The county under investigation (which has three points in this case) is then investigated to see if the number of submissions is significantly low given submission numbers from the surrounding counties. Either a circular or a county-shaped proximity filter can be used in the analysis of the county under investigation. In this case, the circular proximity filter contains 22 points (Figure 2c), and the county-shaped buffered proximity filter contains 12 points (Figure 2d). When dealing with aggregated data and a simulated surface, it is preferable to include the shape of the county as both the inner and outer rings of the proximity filter. This ensures that the same distance extends from the boundary of the county in all directions, reducing variations in the analysis that the county shape might cause.

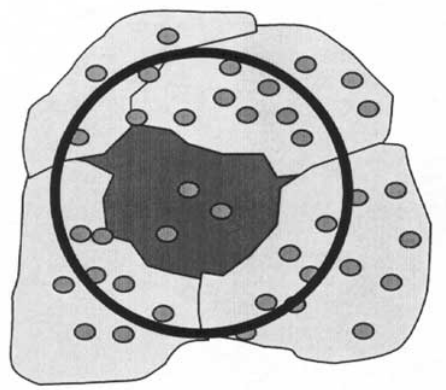
The proximity filter is layered on the randomly distributed points and the points falling inside the inner and outer rings are counted. (The initial simulation generating the points from the aggregated data should be repeated several times in order to obtain an average number of points in the buffer area.) If the inner ring of the proximity filter is the county shape, then the number of points in it will always equal the total points submitted from that county. All the points falling within the inner and outer ring are then randomly redistributed across the same area, using the same fine, layered grid of coordinates within the GIS. Two examples of this simulated run can be found in Figure 2e (for the circular outer ring) and Figure 2f (for the buffer-shaped outer ring). This simulation is repeated 100 times. The distribution of points falling within the county under investigation is then compared with the simulation runs. If there is a significantly low number of points in the inner ring (investigated county) on no more than five of the simulation runs, then there is (at least) a 95% chance that the low number of points in the inner ring did not occur by chance.



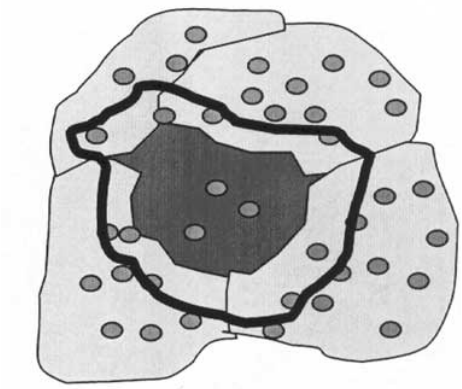
2a Counties with a total number of points.



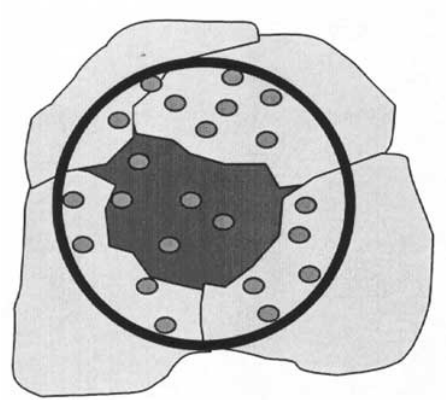
2b Randomly distributing the points.



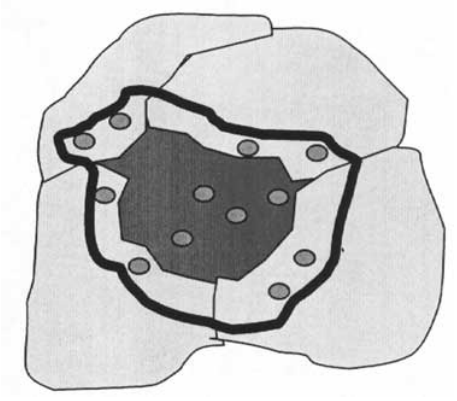
2c Using a circular outer ring.



2d Using a projected-county-shape outer ring.



2e Simulation run using a circular outer ring.



2f Simulation run using a projected county shape.

Figure 2 Creating a point surface from aggregated data.

Improvements

Although it has been stated that the presence of data holes should stimulate a search for factors leading to their explanation, it is possible that some of these factors could be considered in the initial analysis. During the simulation, certain terrain may be more likely to accept a point than other terrain (e.g., suburban fringe compared to marshland). Using GIS makes it possible to identify land cover and create a probability surface of point acceptance. However, the question remains whether adding this complexity to the initial analysis is more efficient than investigating the holes in a homogeneous surface.

It is also useful to repeat the analysis temporally to see if there are any changes in the pattern. If the analysis identifies a county as a significant hole every year, then there is a definite reason for that hole (a difference in terrain, a problem with local officials, etc.). If a county is identified as a significant hole for only certain years, then terrain is unlikely to be the cause. An even more disturbing result would be if there were a clear temporal break point, with a series of normal years being replaced by a series of significant holes. This would suggest that something had changed in the reporting environment of the county.

The proximity filter is a relatively easy technique that can be performed using a PC-based GIS. It enables a health official to investigate any point-data source for which the presence of a hole is as important as that of a cluster. Although the example given here is for animals submitted for rabies testing, any reportable disease with an environmental association (Lyme disease, histoplasmosis, etc.) could be investigated in the same way.

Acknowledgments

Andrew Curtis would like to thank MB Auslander, DVM, MSPH, for his help at the Kentucky Department for Health Services; Ron Mitchelson for discussions during manuscript preparation; and Morehead State University (Kentucky) for its internal grant supporting this project.

References

1. Anselin L. 1999. Interactive techniques and exploratory spatial data analysis. In: *Geographical information systems: Principles, techniques, applications and management.* Vol. 2. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: John Wiley & Sons. 253–66.
2. Openshaw S, Alvanides S. 1999. Applying geocomputation to the analysis of spatial distributions. In: *Geographical information systems: Principles, techniques, applications and management.* Vol. 2. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: John Wiley & Sons. 267–82.
3. Gatrell A, Senior M. 1999. Health and health care applications. In: *Geographical information systems: Principles, techniques, applications and management.* Vol. 2. Ed. PA Longley, MF Goodchild, DJ Maguire, DW Rhind. New York: John Wiley & Sons. 925–38.
4. Curtis A, MacPherson AD. 1996. The zone definition problem in survey research: An empirical example from New York state. *The Professional Geographer* 48:310–20.

5. Openshaw S, Charlton M, Craft AW, Birch JM. 1998. Investigations of leukemia clusters by the use of a geographical analysis machine. *The Lancet* I:272–3.
6. Rushton G, Lolonis P. 1996. Exploratory spatial analysis of birth defect rates in an urban population. *Statistics in Medicine* 15:717–26.
7. Jenkins SR, Winkler WG. 1987. Descriptive epidemiology from an epizootic of raccoon rabies in the middle Atlantic states, 1982-1983. *American Journal of Epidemiology* 126:429–37.
8. Krebs JW, Strine TW, Smith JS, Rupprecht CE, Childs JE. 1995. Rabies surveillance in the United States during 1994. *Journal of the American Veterinary Medical Association* 207:1562–75.
9. Heidt G, Ferguson D, Lammers J. 1982. A profile of reported skunk rabies in Arkansas: 1977–1979. *Journal of Wildlife Diseases* 18:269–77.

A GIS-Based, Case-Control Analysis of Cancer Incidence and Land Use Patterns

Steve Dearwent*
Department of Environmental Health Sciences, School of Public Health, University of Alabama at Birmingham, Birmingham, AL

Abstract

Geographic information systems (GIS) have been used in environmental epidemiologic studies primarily for ecologic analyses. However, many public health researchers are aware of the limitations of the ecologic study when compared with cohort and case-control designs. This paper outlines an approach to be used in a GIS-dependent, case-control analysis of cancer incidence and land use patterns. The study base consists of residents in Jefferson County, Alabama, a large metropolitan area with a population of approximately 650,000. Incident cases of three primary cancers (brain/central nervous system, non-Hodgkin's lymphoma, and pancreas) are identified through the Alabama Statewide Cancer Registry. A static residential requirement of five years is imposed on study subjects to estimate a minimal latency period for neoplastic development and control for population mobility. Georeferencing of cases and controls is anticipated to be highly accurate due to linkage with tabular data and related digitized parcel coverages maintained by the county. As with many GIS-based health studies, distance is a surrogate for exposure and is assessed using buffers generated around residential parcels. Land use characteristics are defined for every parcel in the county (approximately 290,000) and are divided into 16 classes ranging from agriculture and low-density residential to heavy industrial and resource extraction (mining). This study should describe the spatial distribution of these particular cancers in a major metropolitan area as well as address the potential relationships between environmental determinants and disease incidence.

Keywords: cancer, incidence, land use

Introduction

Cancer is a multifactorial disease frequently having etiologies of both environmental and genetic influence. Because of the diverse nature of cancer and the variability of anatomical characteristics exhibited by this disease, health researchers use many scales of analysis. These differing scales range from the study of the disease at molecular and cellular levels, to individual cases and large, population-based analyses. This variety of approaches has been beneficial in understanding biological processes, risk factors, and prevention efficacy as it relates to cancer morbidity and mortality.

Cancer has been recognized as a valuable indicator for environmentally related health effects because there is a definable endpoint (1). Cancers affecting many anatomical sites including bladder, blood, brain, kidney, liver, lung, prostate, and skin have been associated with exposure to synthetic chemicals in the occupational setting (2). It

* Steve M Dearwent, UAB School of Public Health, Dept. of Environmental Health Sciences, Rm. 317, Birmingham, AL 35294-0022 USA; (p) 205-934-6080; (f) 205-975-6341; E-mail: cryptcl@wwisp.com

is probable that a portion of the cancer burden also results from environmental (nonoccupational) exposures. These may include effects from natural substances (sunlight, radon), man-made influences (organic products of incomplete combustion), or a combination of both (asbestos, metals, nitrates, fluorides, exogenous hormones).

Environmental pollutants and resulting adverse health effects have an inherent spatial relationship. The distance from a contaminant source to a given population can influence the magnitude of exposure. Therefore, one may infer that proximity to a source may be a good predictor of the extent of adverse health effects attributable to that source. A geographic information system (GIS) can be used to organize and analyze data in studies designed to consider distance and locational attributes.

Historically, studies using GIS have been descriptive in nature. They have also tended to aggregate exposure/outcome into areas or groups (the ecologic analysis). From an epidemiological perspective, case-control and cohort designs hold more promise in quantifying associations between exposure and disease. Therefore, researchers using GIS should strive to incorporate location-specific measures for both exposure and disease, avoiding data aggregation techniques. Using location-specific measures increases study precision and validity. Accurate georeferencing of study subjects increases precision by reducing random error. Validity is improved with accurate exposure estimation because it increases the chance for correctly assessing cases or controls (minimizing nondifferential misclassification).

The purpose of this study is to describe the spatial variation of cancer incidence in Jefferson County, Alabama, particularly as it relates to land use. It should be emphasized that the methods for defining disease incidence and environmental determinants in this study are not based on an aggregate model. This manuscript describes the data sets, study design, and rationale for research. Analysis is not complete so results are not presented.

Data Sources/Descriptions

The three primary data sets being used in this study detail cancer incidence, residential parcel history, and land use in Jefferson County. The sources for this information are described below. The databases for parcel history and land use are already spatially referenced and accessible via the Jefferson County Information Services network. The data set for cancer incidence is spatially referenced through matching to the county's master address database and subsequent linkage with digitized parcel maps.

Cancer incidence data for Jefferson County are available from the Alabama Statewide Cancer Registry (ASCR). The ASCR began data collection on January 1, 1996. This data set is anticipated to be particularly complete for the study area because all of the hospital-based registries in Jefferson County providing data to the ASCR existed prior to the beginning of statewide data collection. There are approximately 500 combined cases for the cancers (brain/central nervous system, non-Hodgkin's lymphoma, pancreas) and time period (1996–1997) under analysis. Case totals for each category of cancer examined in this study are documented in Table 1.

The Office of Stormwater Management (OSWM) maintains the land use database. The OSWM is a nonprofit public entity that deals with environmental compliance issues pertaining to stormwater discharge in Jefferson County. It was created in response to the National Pollution Discharge Elimination System (NPDES). The stormwater

Table 1 Cancer Cases by Anatomical Site, Jefferson County, AL, 1996–1997

	Number of Cancer Cases	
Anatomical Site	**1996**	**1997**
Brain/central nervous system	46	34
Non-Hodgkin's lymphoma	129	118
Pancreas	80	84

coverage used in this study was developed over a span of four years (1991–1994) by Walter Schoel Engineering. Aerial photography and field-verified land use maps were the primary sources for creating the coverage (3). The series of aerial photographs used in this project were taken in 1990 for over 1,300 1-mile sections in the county. For regions of the county experiencing heavy growth rates, personnel were sent into the field to visually verify documented patterns. The coverage classifies every parcel in the county into one of 16 categories of land use. The magnitude of this database is immense, considering that Jefferson County is an area of over 1,120 square miles and there are approximately 290,000 individual parcels of land in the county ranging in size from small residential plots (fractions of an acre) to large commercial and government-owned properties (multiple acres/square miles). The NPDES data classify every parcel according to the scheme outlined in Table 2.

The Jefferson County Tax Assessor database is used in the analysis for many functions. This source details parcel information for the entire county and assists in enumerating the study base, georeferencing all study participants, and obtaining residential parcel characteristics (zoning, length of ownership, property value). Specific study restrictions have been applied during the query of the tax assessor database to identify all "eligible" parcels within the county. These parameters and the rationale for imposing them are discussed subsequently.

Methods

The data described above are being used in a cumulative incidence, case-control analysis of cancer and land use patterns in Jefferson County. Cases are defined as primary cancers occurring in Jefferson County that are identified through the ASCR for the period of 1996 through 1997.

Study restrictions insure that cases are derived from the same cohort (the study base) out of which controls are selected (4). These parameters are applied to the parcel of land where subjects reside. Eligible parcels must meet the following restrictions:

- Should lie completely within the boundaries of Jefferson County
- Should be zoned for residential use
- Cannot have a deed date (transaction) during the period 1992–1997
- Must have homestead status

For obvious reasons, a study subject's residential parcel must fall within the county boundaries. If this condition is not met, then that individual is not considered to be a

Table 2 Land Use Categories, Jefferson County, AL, 1991

Database Coding	Field Description	Number of Polygons[a]	Percentage of Area within Jefferson County
AG	Agriculture	960	6.3%
CH	Heavy commercial	117	0.6%
CL	Light commercial	661	0.3%
CU	Urbanized commercial	252	0.7%
HW	Highway	9	1.3%
IH	Heavy industrial	342	1.7%
IL	Light industrial	617	1.2%
INST	Institutional (schools, churches)	1,247	1.3%
OS	Open space (parks, recreational areas, greenways)	142	1.0%
RE	Resource extraction (mining)	457	1.9%
RH	High-density residential	864	3.4%
RL	Low-density residential	1,137	4.9%
RM	Moderate-density residential	1,091	9.2%
RT	Mobile home parks	128	0.1%
U	Undeveloped	1,698	64.6%
W	Water	398	1.4%

[a] The number of polygons per land use category does not correspond with the number of parcels for each respective group. Adjoining parcels with the same land use coding have been concatenated into one polygon using the dissolve function in ARC/INFO.

resident of Jefferson County. Study subjects include only Jefferson County residents because the tax assessor database, digitized parcel maps, and land use coverage are all limited to this region.

All participants must live on a parcel zoned for residential use. This restriction is imposed to enumerate a control population more precisely. By eliminating all parcels of land used for commercial, industrial, government, and other nonresidential purposes, the remaining set should constitute a viable group of parcels where people actually live.

Eligible parcels cannot have a deed date between 1992 and 1997. Deed dates within the tax assessor database indicate a parcel transaction. By imposing this restriction, all members of the study base should have lived at their current residence for a minimum of five years. The five-year period is chosen arbitrarily but provides an estimate for static residential populations. This residential exclusion period serves many purposes. It provides more plausibility to the study design by establishing a minimal latency period for initiation and progression of neoplastic growth to diagnostic levels. Five years is an extremely short latency period, but extending this to 10, 15, or even 20 years would severely compromise the size of the study population. The five-year exclusion also assists in controlling for population mobility, an important consideration when studying locational attributes of disease status within urban areas. This is particularly true for urban residents in the US because they exhibit some of the highest levels of mobility for

any industrialized nation. The main limitation in imposing the five-year residential requirement is that it will decrease study precision (power) by eliminating cases that do not meet this parameter.

The exclusion of parcels without homestead status will eliminate potential cohort members who rent their residence. The exclusion of renters is necessary because it is virtually impossible to follow their residential history with the county records used in this study. Renters are a more mobile population and many would probably not meet the five-year static residential requirement. Imposing this parameter increases study validity because it strengthens the definition of the study base by minimizing selection bias. However, it also decreases precision because some cases will be excluded from the analysis.

The time frame under analysis in this study is used for many reasons. Case information from the ASCR is available only for this span. Also, this period corresponds well with potential exposure to the 1991 land use coverage combined with the five-year static residential requirement. In other words, all study subjects identified during the 1996–1997 time frame must have lived at their current residence since 1991/1992, approximately the same period during which the land use audit (exposure assessment) was conducted.

The processes for georeferencing cases and controls differ slightly. They are both linked to their residential parcel via digitized parcel maps. However, controls are initially defined by querying the tax assessor database for "eligible" parcels, while cases are provided by ASCR. Because all eligible controls are identified and accurately georeferenced by querying the tax assessor database, there is no need for matching address fields. If a parcel meets all the study restrictions, then it is selected along with the spatial references documented in the digitized parcel coverage. Cases, however, are georeferenced by matching street addresses documented in the ASCR database to the county's master address database, with subsequent linkage to digitized parcel maps. The matching of text-based address fields between databases is more cumbersome and problematic.

Once all study subjects are georeferenced, exposure is assessed by generating concentric buffers of predefined size around each parcel polygon label point and aggregating land use characteristics found therein. Because distance is a surrogate for exposure, varying buffer sizes will assist in dose-response and trend analyses. Buffering around points instead of parcel boundaries insures that the area encompassed by buffers using the same predefined diameter will not vary. If boundaries (polygons) of residential parcels were used to determine buffering dimensions, the buffered regions would vary in size corresponding with parcel size. They would also take on heterogeneous shapes. These factors may produce "spatial" confounding.

Discussion

Health outcome data are often georeferenced to areal units such as state, county, municipality, zip code, or census tract. They are infrequently assigned a point value even though this provides a much more accurate, non-aggregate, locational description. This practice stems from the fact that most health outcome datasets include either information on an areal unit or have a data element that can be easily related to a region.

The method of georeferencing used in this analysis is anticipated to be highly

accurate compared with typical procedures involving linkage with street address ranges. The US Census Bureau's TIGER/Line files provide a common means for matching a list of addresses to street segments and their respective address ranges. However, the use of address ranges can be problematic. Most address matching programs allocate addresses at evenly spaced intervals along a street line segment, recessed off the street by a predefined value. An apartment complex located at one end of a street may account for the majority of addresses on that segment, yet addresses will be allocated at equal intervals along the entire path. With the method for georeferencing used in this analysis, study subjects will be linked directly to the parcel of land where their residence is located. This will eliminate the erroneous assumption of evenly spaced, single-dwelling residences.

Conclusion

Geographic information systems are being integrated into many of today's information management sectors. GIS is already an important component of earth sciences. This growth will eventually have a substantial impact on the collection, management, and analysis of health outcome data. GIS provides environmental health researchers with the ability to combine data from population-based cancer registries and environmental hazard assessments. For cancers in which environmental exposures are potential risk factors, it will mature as a more useful analytical tool. Public health is beginning to witness the use of GIS in many facets, although this is not always published in the standard epidemiologic and environmental health literature (5,6,7). This growth should continue, as information technologies become an increasingly important part of our society.

Acknowledgments

This research is supported in part by grant CA47888 from the National Cancer Institute. The author is also indebted to Jefferson County Information Services for providing GIS facilities and technical guidance.

References

1. Sherman JD. 1994. *Chemical exposure and disease.* New Jersey: Princeton Scientific Publishing Co.
2. Doll R, Peto R. 1981. The causes of cancer: Quantitative estimates of avoidable risks of cancer in the United States. *Journal of the National Cancer Institute* 66:1191–1308.
3. Haynes D. 1998. Personal communication. Walter Shoel Engineering.
4. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. 1992. Selection of controls in case-control studies. *American Journal of Epidemiology* 135:1019–28.
5. McGarigle B. 1998. GIS takes on TB. *Government Technology* 6:19–20.
6. Nygeres T, et al. 1997. Geographic information systems for risk evaluation: Perspectives on applications to environmental health. *Cartography and Geographic Information Systems* 3:123–44.
7. Zhou Y, et al. 1996. GIS-based network models of Schistosomiasis infection. *Geographic Information Sciences* 2:51–7.

Power Lines, Line Transects, and GIS

J Wanzer Drane, PE, PhD (1),* Heidi L Weiss, PhD (2), Tim E Aldrich, MPH, PhD (3), Dana L Creanga, PhD (4), Gerald F Pyle, PhD (5)

(1) School of Public Health, University of South Carolina, Columbia, SC; (2) Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL; (3) Department of Health and Environmental Control, Centers for Disease Control and Prevention, US Department of Health and Human Services, Columbia, SC; (4) Computer Horizons, Indianapolis, IN; (5) Professor of Geography, University of North Carolina at Charlotte, Charlotte, NC

Abstract

With the advent of address matching, line and band transects can be used to take useful random samples of homes, workplaces, or other such objects in an area. In this study of power lines and cancers in a part of North Carolina, we created band transects by adding buffers to each side of a line transect. The band transects were used to sample homes of cancer victims in order to estimate the density of new cases within a fixed geographic region. The entire set of new cases over one year defines the sampling frame. Trunk power lines were the assumed hazard. Therefore, the band transects were modified by removing a center strip representing the rights of way of the power lines. The intersections of addresses of cancer victims with buffer zones outside the rights of way created a null distribution of cancer densities within the study area. The power line rights of way, with buffers added to each side, were then intersected with the frame of addresses to provide a sample of affected persons. This paper presents four major conclusions of the study. First, the distributions of the numbers of cancer cases were very well modeled by Poisson distributions. Second, 40-meter-wide buffers were more efficient at capturing cases than were those 20 and 70 meters wide. Third, the density of cancer cases within the buffers of power lines was approximately half that within the randomly created band transects. Finally, a case is made for the continued development of this methodology. From this experience was born the concept of "the shadow of the hazard," an appropriately shaped quadrat that represents a hazard's area of influence and can be used to sample the affected and unaffected portions of a population.

Keywords: linear hazards, cancer, spatial statistics, sampling, prevalence

The Line Transect Method

Background

The line transect and its extension, the band transect, are normally used to estimate D, the population density of wildlife in an area. (D is the number of subjects per unit area.) The technique's application consists of walking a straight path of random or fixed length, L, through the region of interest. The observer records the number of subjects identified and their estimated right angle distances, y, from the line transect when first observed. This method of sampling is well documented (1–8). Let the function

* J Wanzer Drane, Professor of Biostatistics, University of South Carolina, School of Public Health, Green and Sumter Streets, Columbia, SC 29201 USA; (p) 803-777-5053; (f) 803-777-2524; E-mail: Wdrane@sph.sc.edu

$g(y)=P$(observing an object | y) (1). It is assumed that $g(0)=1$, and $g(\infty)=0$. That is, a subject exactly on the line is seen with probability equal to 1, and at great distances it will not be seen at all. During the traverse, n subjects will be seen at distances Y_i, $i=1, \ldots, n$. Let $f(y)$ denote the probability density function of Y.

Estimation of the Density of Cancer Cases Using the Line Transect Method

Let the object being observed be the residence of a cancer victim. Let $g(y)$ and $f(y)$ be defined as above. D, instead of the number of wildlife subjects, is the number of cancer cases per unit area. Consider a random line of length L with a fixed observation width W within a study area. The band has width $2W$, length L, area $2LW$, and a number n of cases within it. The estimate of D is given by $\hat{D} = \frac{n}{2L\mu}$, where $\mu = \int_0^W g(y)dy$ In this application, $g(y)=1$ for all y, and $0 \le y \le W$, because every cancer case within a band will be detected and counted with certainty. Using a geographic information system (GIS), precise coordinates of the home of each cancer case are encoded. Thus, the estimate of cancer case density is $\hat{D} = \frac{n}{2LW}$, which is simply the count divided by the area of the band transect.

Statistical Properties of the Estimate of D

In typical transect studies, replicate random lines are created over the entire study area. Let there be R replicate random lines with lengths h_i, with respective counts n_i, where $i=1, 2, \ldots, R$. The estimate of D can be calculated as above for each individual transect or as $\hat{D} = \frac{\sum n_i}{2W\sum h_i}$, where the summations are over the R replicates. This estimate is the maximum likelihood estimate of $\hat{D}$ (3). We will estimate the variance of D directly using replicate transects and also using the ever-popular jackknife.

Spatial Analysis of Disease and Environmental Hazards

Here we are interested in describing and analyzing disease patterns in physical space, defined, for example, by longitude and latitude or their transformations of location and scale. As an example, many epidemiological studies have focused on residential exposures to electromagnetic fields created by the passage of electrical currents through power lines (9–16). Residential proximity to environmental factors, presumed to be hazards, has been central to some of those studies. In the same studies, moreover, maps showing power lines near homes—or, for other hazards, similarly indicative maps showing residential proximities to the presumed hazards—were sufficient to indicate exposure to environmental hazards.

It is therefore beneficial for the inquiry into the effects of distance between potential environmental hazards and cancer occurrence to capitalize on the advances in GIS technology. This study demonstrates the application of the line transect method, by way of GIS, to this inquiry, the purpose of which is to estimate the densities of cancer

cases within proximities of linear objects. The linear objects in our study were the electrical trunk lines serving a large city, Charlotte, North Carolina.

Background on Geographic Information Systems

A GIS is any manual or computer-based set of procedures used to store and manipulate geographically referenced data (17). A GIS should be capable of data input, data management, and manipulation and analysis of geographically referenced data. A distinction should be made between cartographic systems and GIS. The main function of a cartographic system is to store maps in automated form and generate computer-based maps. A GIS should have additional capabilities—it should be able to integrate layers of geographically referenced data, perform analysis on these layers of data, and predict or evaluate spatial relationships and outcome phenomena. A GIS layer is a set of data with geographic references compatible in scale and location with cartographic features stored and manipulated by the GIS. Layers include, but are not limited to, spatial attributes such as longitude and latitude, area, length, use, population within polygons (e.g., city blocks), or boundary designations. Layers can also include population densities, socioeconomic status variables, census information, or personal information together with the coordinates of the place where a person lives or works. The ability to explore the interactions or interrelationships between geographical variables and geographically referenced other variables makes GIS potentially a very powerful tool for geographic analysis.

The area of geographic health information systems is an emerging area of GIS applications. GIS has been used for the planning and delivery of health services (18). Gisler (19) cited the use of GIS for assessing environmental risks as one of the future research directions of spatial analysis of diseases. GIS facilitates the application of spatial techniques in generating hypotheses about environmental hazards and cancer risks. Furthermore, GIS provides an initial impression of the relationship between cancer cases and potential environmental hazards—a starting point from which to analyze data from typical state cancer registries and surveillance programs.

Materials and Methods

Sample

The line transect method of sampling and ARC/INFO (20), a GIS, were applied to data from Mecklenburg County, North Carolina. This study was limited to the area covered by the Dual Independent Map Encoding (DIME) files. These datasets, called coverages, included digitized spatially related data on homes of cancer victims, roads, streams, power lines, census tracts, and street addresses. Each coverage has a corresponding attribute table that describes the geographic features within the area. The US Census Bureau's TIGER/Line files (21) were used to create the coverages. TIGER files provide digital data on all census geographic boundaries, codes, longitude and latitude coordinates, feature names, addresses, and zip codes. Mecklenburg County was divided into sub-areas, and only the sub-areas defined by the DIME files were investigated. This ensured address matching of all cancer cases. The North Carolina Central Cancer

Registry's population-based cancer incidence database for 1990 supplied the incidence data on cancer for this study (22,23).

ARC/INFO was used to perform a cross-sectional geographic analysis by integrating these coverages. Addresses of newly diagnosed cancer cases for 1990 were represented as points, roads and power lines as arcs, and census tracts as polygons. To perform spatial analysis, ARC/INFO was used to create random transects across the county, create buffers of specified widths around the transects, overlay the resulting buffered transects with other coverages such as cancer case coverage or census tract coverage, and calculate such measures as the number of cancer cases within the transect and the total area of the band transects created from the line transects and their respective buffers.

Spatial Sampling Procedures

The geographic area considered in this study is the DIME area of Mecklenburg County, North Carolina. The population under investigation is the newly diagnosed cancer cases for 1990 who live within that DIME area. The cancer cases are represented as points whose longitude and latitude coordinates locate them within the area. Those cancer case points constitute the cancer case coverage or layer, along with an attribute table associated with them. That attribute table contained ID, RACE, AGE, GENDER, ZIPCODE, ADDRESS, CENSUS TRACT, and ICDOCODE (type of cancer). An address coverage was used to create the cancer case coverage. Within it were stored addresses and corresponding IDs. Address matching was used to obtain the coordinates for each cancer case address. Another coverage was composed of the census tracts and their associated polygon attribute table. This attribute table included data on census tract number, census tract ID, and population size.

Line transects with their respective buffers were the sampling units. A simple random sample of lines throughout the entire county was drawn. This was done by creating a random number of random point pairs $[(x_1, y_1), (x_2, y_2)]$ using a random number generator. The resulting data were read into ARC/INFO. Using the program's GENERATE command, line segments were drawn between the pairs of points. Those segments were considered to be one replicate. This process was repeated 10 times, so 10 replicates were drawn. Each replicate consisted of, at most, 20 line transects.

The lengths of the transects were calculated and a 30-meter corridor was drawn around each. These corridors represented the rights of way beneath the power lines, where there could be no houses. Beyond the corridors, buffers were created of widths 20, 40, and 70 meters. These buffers of varying width acted as surrogates for exposure to electromagnetic fields. The resulting coverage of transects and buffers was overlaid on the cancer case coverage. The cancer cases denoted as points within the buffered transects were elements within the sampling units. Note that the lengths of the transects differed, and many transects intersected one another.

Transects from a single replicate were pooled to provide an estimate of the density of cancer cases. Ten estimates of the density of cancer cases were calculated, one for each of the ten replicates. From these an empirical distribution of the density could be calculated and then used to calculate an overall estimate, $\boldsymbol{D}_0$. The next step was to create line transects using a geographic feature in the environment. The high-power electrical trunk lines were the linear feature of choice. The entire segment of the trunk lines feeding the greater Charlotte area within the Mecklenburg County DIME area was

considered a single transect. Using the same procedure that was used for the random line transects, a single estimate of the density of cancer cases was calculated. A test of equality between the density of cancer cases along the random transects and the density along the power line transect was carried out. In both cases, the number of cases found by the buffers followed a Poisson probability mass function with a mean of $2LWD$, and under H_0, $\boldsymbol{D}_0$ is substituted for D (1). Denoting $\boldsymbol{D}_p$ as the density near the power lines, n_p was the Poisson test statistic with mean (parameter) $\Theta_0=2WLp\boldsymbol{D}_0$. That is,

$$f(n) = P(n \mid \Theta_0) = \frac{e^{-\Theta_0}\Theta_0^n}{n!}, \text{ n} = 0,1,2, \ldots, \infty.$$

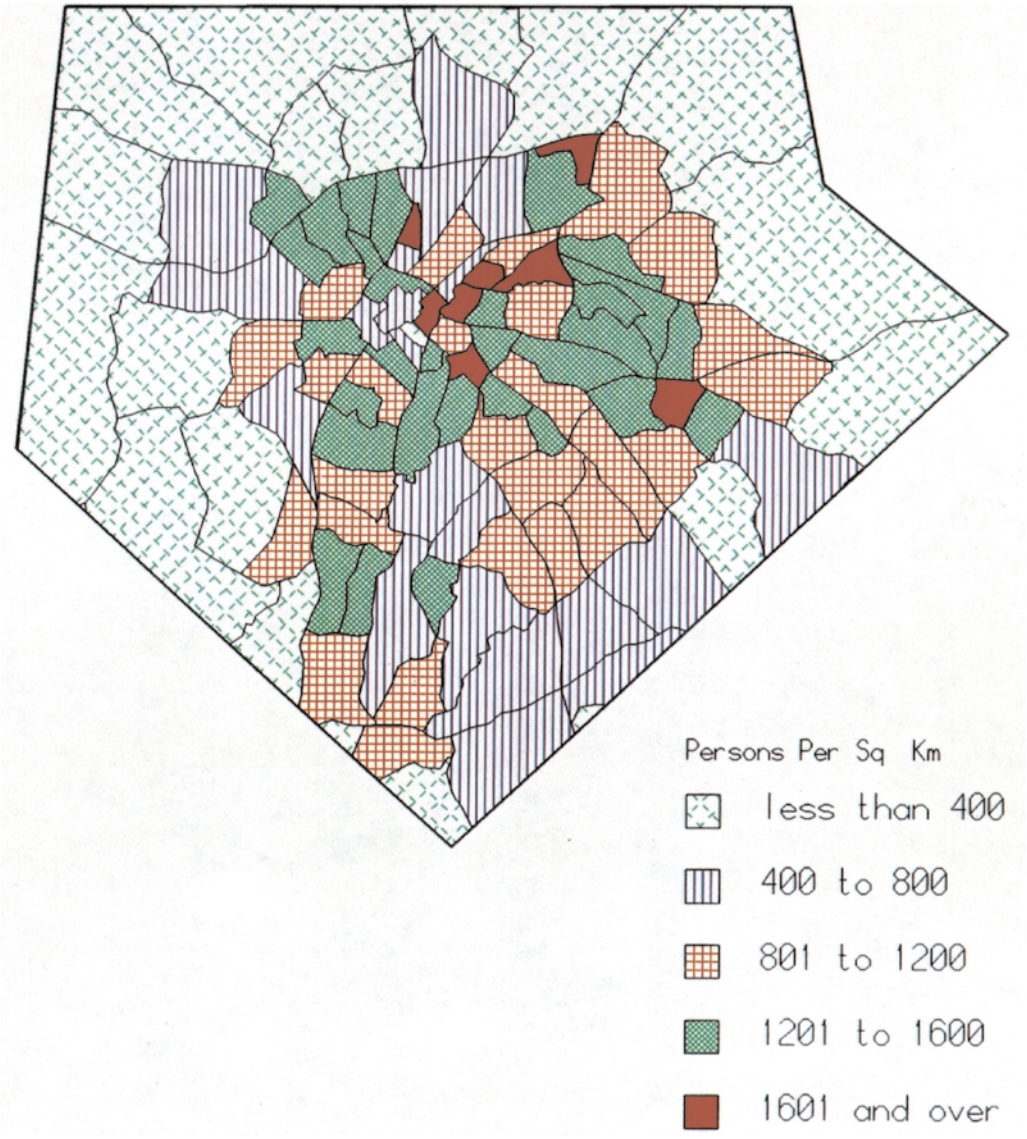


Figure 1 Population density by census tract, DIME file area of Mecklenburg County, North Carolina.

The p-values were calculated using the minimum likelihood method (24). The above procedure was done for *W*=20, 40, and 70 meters.

Results

Estimation of Cancer Density Using Random Line Transects

The final coverages used in the analysis include the census tracts (Figure 1), the high-power transmission lines (Figure 2), and the roads (Figure 3), all within the DIME area. The result of address matching is shown in Figure 4.

Each replicate, consisting of many line transects, formed a new coverage. For each replicate, a new analysis was done as follows. The coverage was clipped to fit within the boundary of the DIME area. Right-of-way corridors of 30 meters were drawn around the line transects (Figure 5), and buffers of 20, 40, or 70 meters were drawn outside the corridors (Figure 6). The corridors were then erased and the new buffered coverage was overlaid on the cancer case coverage (Figure 7).

The areas of the buffers and the numbers of cancer cases within them were

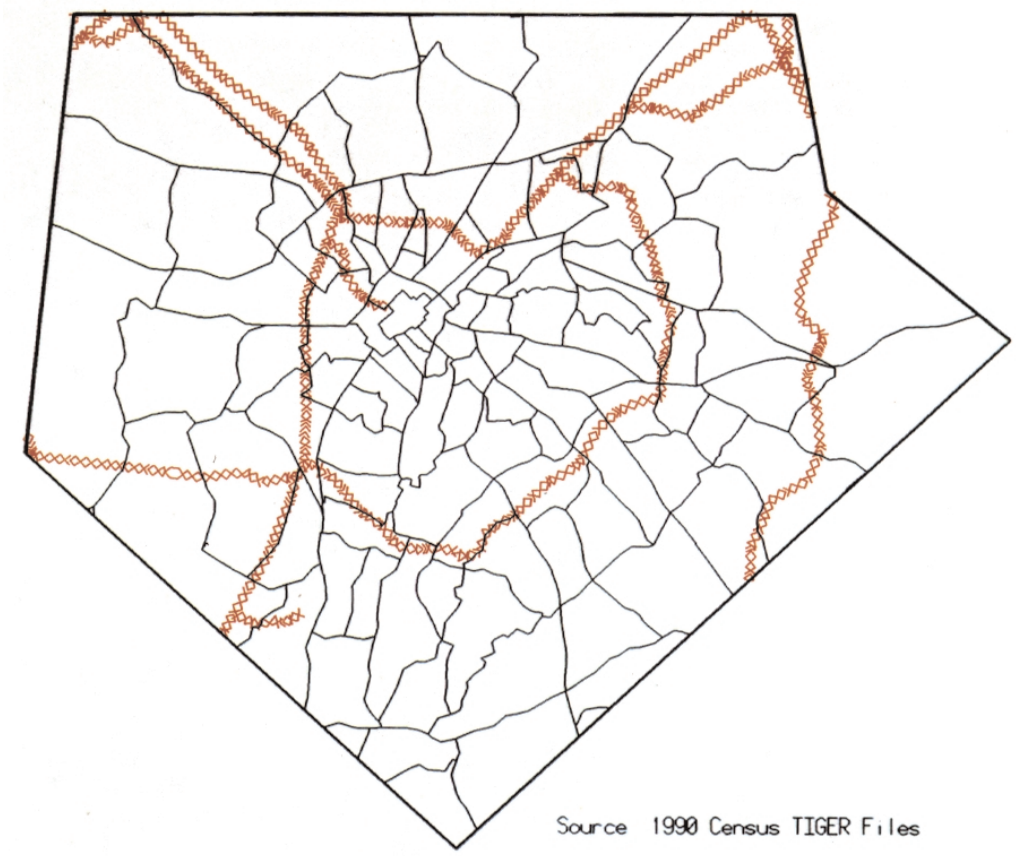


Figure 2 High-power transmission lines, DIME area of Mecklenburg County, North Carolina. Source: (21)

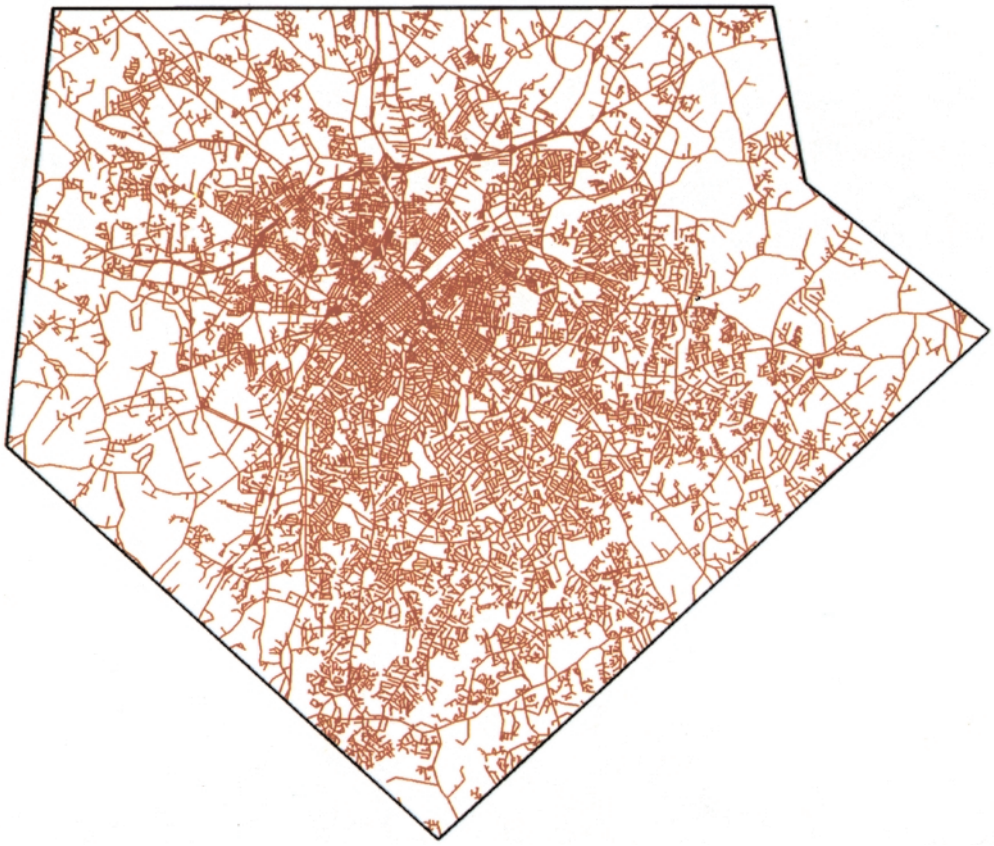


Figure 3 Roads coverage, DIME area of Mecklenburg County, North Carolina. Source: (21)

calculated using ARC/INFO. Note that the buffers are not entirely rectangular, and their areas are therefore not exactly $2LW$. ARC/INFO calculates the areas accurately, taking into account the rounded or otherwise altered ends of the buffers.

The first analysis was done for a buffer 20 meters wide, and the results of the density estimates for that and the two other buffer widths were stored in a table (Table 1). These 10 replicates provide an empirical distribution of cancer cases within the buffered transects. The combined estimates of D_0, shown in Table 2, are 1.72, 1.77, and 1.63 cases per square kilometer, respectively, for buffers 20, 40, and 70 meters wide. Table 2 also shows estimates of the variances of the estimates of cancer densities.

Estimates of Cancer Density Using Power Transmission Lines

The null hypothesis of $D_0=D_p$ versus not equal was tested for each of the buffer widths; the corresponding p-values are given in Table 3. The several D_p shown in Table 4 were compared with their respective D_0 counterparts of Table 2. The p-values in each case gave no evidence to support a conclusion that D_p and D_0 were different.

At least for this study, which is *not* an epidemiological investigation, there is no

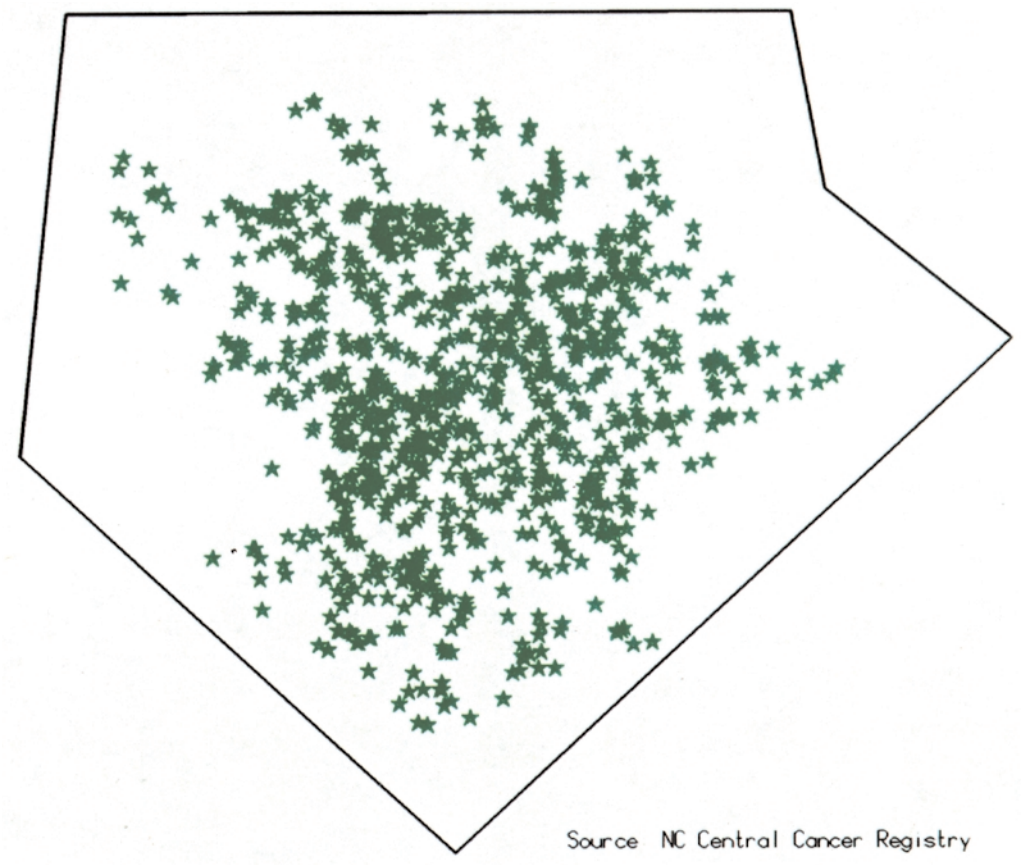


Figure 4 1990 cancer cases, DIME area of Mecklenburg County, North Carolina. Source: (22,23)

evidence to implicate trunk power lines as a source of cancer formation. This will be addressed again in the discussion.

Discussion

The line transect method has not been previously applied to spatial analysis of health data, at least to the knowledge of these writers. This study's goal was to demonstrate the plausibility of the method's use in estimating cancer case density. This we have done. The densities of cancer cases near supposed hazards provide evidence to suggest further studies of situations in which the densities are higher than those in the non-exposed portions of the population. The methodology used in this study, however, does need refining.

GIS goes hand-in-glove with analyses of this type. Aside from making it possible to overlay various layers on a base map, GIS enables users to manipulate and analyze spatially indexed data. In this study, for example, GIS made it possible to generate random

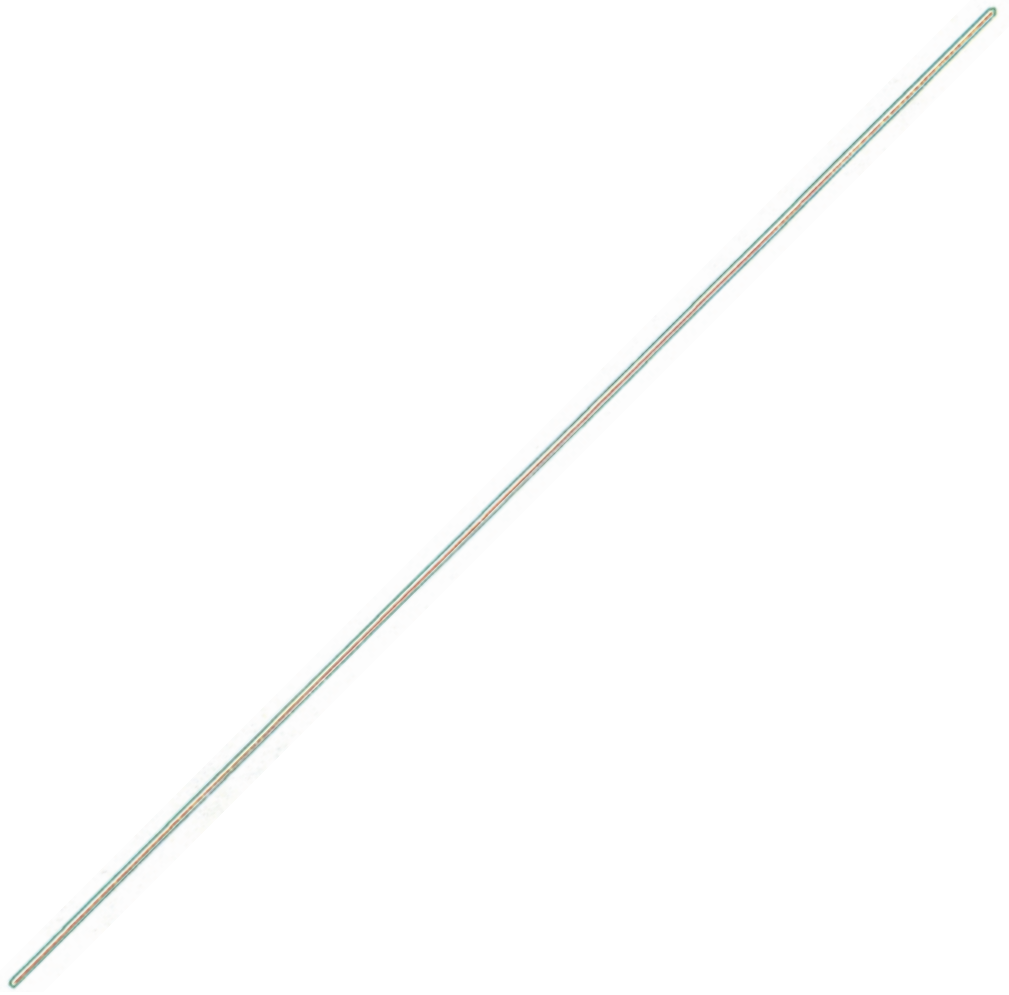


Figure 5 Random line transect with 30-meter corridor.

lines within a well-defined geographic area, remove corridors, add buffers outside the corridors, and overlay and intersect the point pattern of cancer cases with the transects, thereby providing estimates of cancer case densities.

Many insights can be garnered by applying GIS to such data. First, many existing surveillance programs could expand their reporting procedures to include such characteristics as occupations, industries, and time spent at home, at work, and in transit. The ability of GIS to integrate non-spatial data increases our ability to perform ecological analyses on human health data.

It should be pointed out again that this study was limited to hazards of a linear form such as power lines. The buffers outside linear hazards we call "shadows of the hazard." This idea and name can be generalized to hazards of varying geometries and

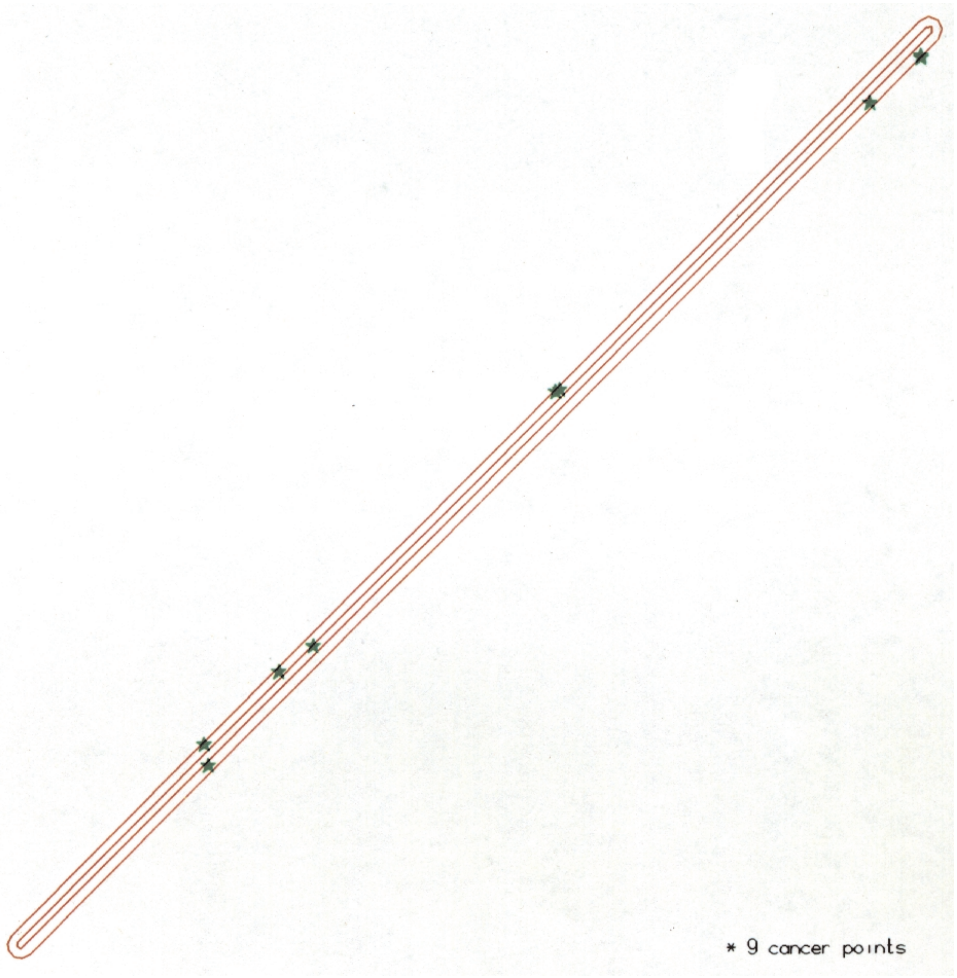


Figure 6 Random line transect with 70-meter buffer around 30-meter corridor.

sizes. Wind roses, for example, are shadows of the hazard of exhaust gases emitted to the atmosphere. A "cookie cutter" could be created from a wind rose with its core—which represents the hazard area itself—omitted. Randomly placed quadrats (the cookie cutter) would be used to sample areas and obtain densities, as we have done here with linear band transects with their cores (power line rights of way, in our case) removed.

Finally, it should be noted that we did not discriminate by cancer type, and population densities in the various census tracts were not taken into consideration. Now that our method has been demonstrated, you can improve on it at least in these two ways.

Acknowledgments

Thanks to Doctors C Macera, J Hussey, and S Weinrich for their contributions to this project; L Shirley, J White, and T White, who were resource persons for ARC/INFO; and T Weiss for his helpful editorial suggestions.

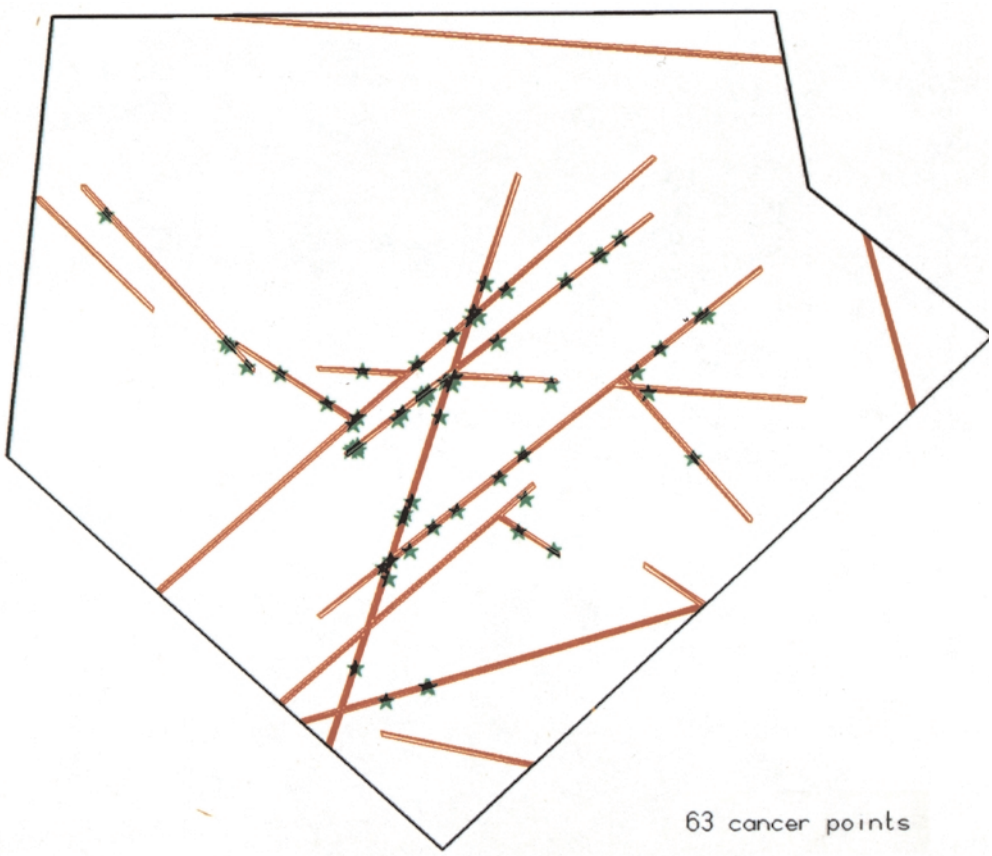


Figure 7 Random line transects and cancer cases within 70-meter buffers.

Table 1 Estimates of Cancer Case Density for Random Line Transects Using 20-, 40-, and 70-Meter Buffers

	Density Estimates per Square Kilometer		
Replicate	**20-Meter Buffer**	**40-Meter Buffer**	**70-Meter Buffer**
1	2.00	2.42	1.89
2	1.75	1.57	1.60
3	2.37	2.78	1.68
4	1.33	1.73	1.83
5	0.98	1.16	1.13
6	0.93	1.60	1.42
7	2.86	1.86	1.71
8	1.55	1.42	1.40
9	1.45	1.54	1.97
10	1.77	1.85	1.91

Table 2 Weighted Means of Estimates of D_0 and Variances of D_0

		Variance of D_0			
Buffer Width (meters)	D_0	**Direct**	**Indirect**	**Jackknife**	D_0/**Area**
20	1.72	0.043	0.037	0.039	0.023
40	1.77	0.028	0.019	0.015	0.012
70	1.63	0.008	0.007	0.007	0.006

Table 3 Tests of the Hypothesis that $D_p=D_0$

Buffer Width (meters)	D_0	D_p	**p-value**
20	1.72	0.53	0.427
40	1.77	0.53	0.432
70	1.63	1.15	0.999

Table 4 Estimates of Cancer Case Densities Near Power Lines Using 20-, 40-, and 70-Meter Buffers

Buffer Width (meters)	**Number of Cases**	**Area (square kilometers)**	D_p **(cases per square kilometer)**
20	3	5.64	0.53
40	6	11.34	0.53
70	22	19.99	1.15

References

1. Buckland ST. 1985. Perpendicular distance models for line transect sampling. *Biometrics* 41:177–95.
2. Upton GJG, Fingleton B. 1985. *Spatial data analysis by example; Point pattern and quantitative data, volume 1.* New York: John Wiley & Sons.
3. Burnham KP, Anderson DR, Laake JL. 1980. Estimation of density from line transect sampling of biological populations. *Wildlife Monograph No. 72, Supplement to Journal of Wildlife Management* 44.
4. Burnham KP, Anderson DR. 1976. Mathematical models for nonparametric inferences from line transect data. *Biometrics* 32:325–36.
5. Pollock KH. 1978. A family of density estimators for line transect sampling. *Biometrics* 34:475–8.
6. Eberhardt LL. 1968. A preliminary appraisal of line transects. *Journal of Wildlife Management* 32:82–8.
7. Gates CE, Marshall WH, Olson DP. 1968. Line transect method for estimating grouse population densities. *Biometrics* 24:135–45.
8. Seber GAF. 1973. *The estimation of animal abundance and related parameters.* London: Charles Griffin.

9. Wertheimer N, Leeper E. 1979. Electrical wiring configurations and childhood cancer. *American Journal of Epidemiology* 109:273–84.
10. Wertheimer N, Leeper E. 1982. Adult cancer related to electrical wires near the home. *International Journal of Epidemiology* 109:345–55.
11. Fulton JP, Cobb S, Preble L, et al. 1980. Electrical wiring configuration and childhood leukemia in Rhode Island. *American Journal of Epidemiology* 111:292–6.
12. McDowall ME. 1986. Mortality of persons resident in vicinity of electricity transmission facilities. *British Journal of Cancer* 53:271–9.
13. Tomenius L. 1986. 50-Hz electromagnetic environment and the incidence of childhood tumors in Stockholm County. *Bioelectromagnetics* 7:191–207.
14. Coleman MP, Bell CMJ, Taylor HL, Primic-Zakelj M. 1989. Leukaemia and residence near electricity transmission equipment: A case-control study. *British Journal of Cancer* 60:793–8.
15. Savitz DA, Wachtel H, Barnes FA, John EM, Tvrdik JG. 1988. Case-control study of childhood cancer and exposure to 60-Hz magnetic fields. *American Journal of Epidemiology* 123:21–38.
16. Stevens RK, Stevens RG, Kaune, et al. 1988. Acute nonlymphocytic leukemia and residential exposure to power frequency magnetic fields. *American Journal of Epidemiology* 123:10–20.
17. Starr J, Estes J. 1990. *Geographic information systems: An introduction.* Princeton: Prentice Hall.
18. Twigg L. 1986. Health-based GIS: Their potential examined in the light of existing data sources. *Social Science in Medicine* 30:963–73.
19. Gisler W. 1986. The uses of spatial analysis in medical geography: A review. *Social Science in Medicine* 23:963–73.
20. Environmental Systems Research Institute, Inc. (ESRI). 1991. *ARC/INFO version 5.* Redlands, CA: ESRI.
21. US Census Bureau. 1990. TIGER: *The coast-to-coast digital map data base.* Washington, DC: US Department of Commerce, Bureau of the Census. November.
22. North Carolina Central Registry. 1992. Cancer incidence in North Carolina: 1990 county specific numbers. Raleigh, NC: State Center for Health and Environmental Statistics, North Carolina Department of Environment, Health, and Natural Resources. November.
23. North Carolina Central Registry. 1992. *North Carolina health statistics pocket guide.* Raleigh, NC: State Center for Health and Environmental Statistics, North Carolina Department of Environment, Health, and Natural Resources. December.
24. Gibbons JD, Pratt PD. 1975. P-values: Interpretation and methodology. *American Statistician* 29:20–5.

Data Issues and Cartographic Techniques as Applied to the Use of GIS in Epidemiology: The Alberta Health Model

Erik A Ellehoj (1),* Dr Fu-Lin Wang (2), Dr Stephan Gabos (2)
(1) Ellehoj Redmond Consulting, Edmonton, Alberta, Canada; (2) Surveillance Branch, Alberta Health, Alberta, Canada

Abstract

Data used for spatial analysis of health research are available from several sources and in a variety of formats. Geographic boundaries used to define these datasets are not consistent, which results in problems with data presentation. A consistent and appropriate geographic segmentation of a region is necessary to reveal geographic patterns and ensure that their implied relationships are real and not a result of boundary placement. Cartographic literature provides limited assistance to the researcher attempting to manage these difficulties effectively. Inconsistencies in the literature and discussions of mapping technique limitations force the researcher to deal with these issues on a case-by-case basis. This paper outlines a variety of cartographic techniques and discusses methods of presenting spatial data used by Alberta Health, Surveillance Branch.

Keywords: mapping techniques, geographic boundaries, Alberta Health, Canada

Introduction

The mission of the Government of Alberta's Department of Health (Canada) is to protect, maintain, restore, and enhance the health of Albertans. To achieve this, the Surveillance Branch of Alberta Health is implementing geographic information systems (GIS). GIS enable researchers to examine health data from a spatial perspective. The use of GIS in epidemiology research has recently received much attention in research literature and from agencies responsible for the epidemiological research, and has been well documented by Clarke et al. (1). GIS can help researchers describe the spatial perspective of disease, but this must be done using a consistent spatial foundation.

Researchers at Alberta Health, Surveillance Branch, have examined the temporal-spatial patterns of health events and the possible determinants for these patterns. As a result of this work, researchers have become aware of the problems associated with data collection and issues of replication of results. The following sections outline the concerns related to geographic boundaries used in spatial analysis, and a method is presented to overcome data collection and replication concerns.

Health Data and Boundaries

The two main issues confronting GIS research as applied to epidemiology are inconsistent geographic boundaries and the size of mapping units. Locating and assigning

* Erik A Ellehoj, Ellehoj Redmond Consulting, 11456 43 Ave., Edmonton, AB T6J-0Y4 Canada; (p) 780-434-1943; E-mail: ellehoj@supernet.ab.ca

individuals into geographic units is essential in the use of GIS in epidemiological research. Ideally, for spatial analysis, all individuals are linked to a fixed geographic location. The movement of individuals from this fixed location is tracked. (To ensure client confidentiality, the health data collected by an agency are only made available in an amalgamated format.) Aggregated data vary depending on how and why the data are collected. Boundaries assigned to geographic areas by research agencies are not consistent; therefore, the grouping of individuals into geographic units varies among agencies. For example, Statistics Canada uses census boundaries, while Alberta Health uses Regional Health Authority (RHA) boundaries.

Local spatial variations within a larger mapping unit are often masked due to data collection methods. For example, rural census subdivisions in the northern portion of the province are very large. Regional variations are expected within these areas, but the variations are not visible because information for the areas is amalgamated. The use of smaller data boundaries reduces this effect.

To examine the challenges associated with spatial analysis of health data in the Alberta context, it is necessary to consider the different kinds of boundaries Alberta Health could use in GIS research.

County

Counties are not as important in Canada as they are in the United States. The county is a logical spatial unit in the United States, but in Canada, health provision is organized according to other boundaries described below. It is quite difficult to obtain health data at the county level, so the county is not a recommended land unit for GIS use. Additionally, county boundaries in many provinces, including Alberta, are subject to change, as are their equivalents—municipal districts, improvement districts, special areas, etc.

Regional Health Authorities

The province of Alberta is divided into 17 RHAs. Spatial analysis of this geographic unit is relatively simple because most medical data are organized by RHA. RHA regions often encompass large areas, though; because regional differences within an RHA are not visible at this level of aggregation, detail is lost.

Census-Related Boundaries

Figure 1 shows four census-related boundaries—census divisions (CDs), consolidated census subdivisions (CCSDs), census subdivisions (CSDs), and enumeration areas (EAs)—and illustrates their hierarchy.

Census Divisions

CDs, as outlined by Statistics Canada, provide a level of aggregation similar to that of RHAs. As with RHAs, most needed medical data can be obtained at this level, but working at this level creates similar challenges to those encountered when using RHAs. CDs are suitable for census data (because they represent aggregations of the smaller census-related areas), but not for tracking health events.

Census Subdivisions

A division of the CDs into smaller units has resulted in the CSDs. The more than 400

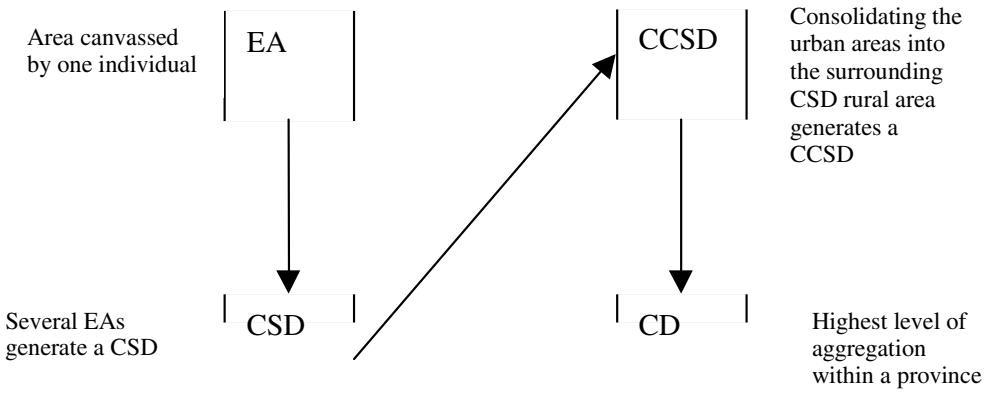


Figure 1 Summary of census boundaries.

CSDs in the province correspond to the municipal political boundaries, including counties, improvement districts, municipal districts, special areas, cities, towns, national parks, and reserves. Regional differences are more apparent when this level of aggregation is used. Unfortunately, many of the CSDs are very small and prove a challenge to cartographers who wish to use visual patterns to differentiate CSDs according to a health variable. If a pattern (e.g., crosshatching, shading) is applied to each CSD according to a health variable, many of the smaller units (such as towns or small cities) are not visible on a map of reasonable dimension. Many of these spatial units are not visible even when the page size for the map of the province is set to "E" size (34" by 44", the largest paper size most plotters can use).

Because CSDs represent legal entities, populations range from over 600,000 for the larger urban centers to under 100 for some native reserves and villages. It is difficult to create a comparable set of statistics for the entire province based on such variation. The larger urban centers cannot be subdivided to examine patterns in regions of the city; meanwhile, a single case can change the mortality or morbidity rate in a smaller community from a "No Cases" category to the highest category.

Consolidated Census Subdivisions

CCSDs, created by Statistics Canada, are amalgamations of CSDs. There are over 70 CCSDs in Alberta. Each consists of a rural envelope—counties, improvement districts, municipal districts, and special areas—and the amalgamation of all cities, towns, villages, and reserves within it. (All populations within a CCSD are added together to create larger populations and mapping regions.) Health data can be obtained at this level by using a CSD-to-CCSD lookup file and adding together all the necessary components. This level of detail is a good compromise between the RHA and CSD levels, especially when mapping incidence of rare diseases. CCSD boundaries make some regional variations evident, but some of the differences between urban and surrounding rural areas are lost.

Enumeration Areas

Each EA is an agglomeration of approximately 800 individuals, and represents the

region canvassed by a single census taker. Rural areas remain quite large when divided according to this scheme, especially in northern Alberta, where population densities are lower. Urban areas are subdivided into many smaller areas; the small size of these units makes this set of boundaries unsuited to traditional mapping techniques, although EAs are used to show detail in one small region. EA-level mapping is most often used for marketing, in which a city is divided into socioeconomic areas and categorized for potential sales. Many health data are not always available at the EA level, and therefore EA mapping is not recommended on a province-wide basis.

Postal Codes

Postal codes are also used to determine the location of a group of individuals. Locations of postal codes can be estimated as points. Medical data can be obtained at this level of aggregation because the mailing address of each Alberta resident is stored as part of Alberta Health's registration data. Postal code populations can be obtained from the same source. Some postal codes share the same geographic location because they are based at the same post office.

There are more postal codes in urban areas than in rural areas. This makes it possible to analyze regions of an urban center for differences. For example, an outbreak of giardiasis can be examined as it relates to the location of a specific water-processing plant. Such specificity is useful to an RHA or for a particular study that examines a disease or a geographic region in detail. The postal code technique is also useful in mapping rural areas. While rural areas defined as postal code regions may be large, they are smaller than rural areas defined by the previously described geographic boundaries.

Postal code areas are inferred using the locations of the post offices, but there are no digital files available that store the geographic boundaries for each postal code. When possible, overlapping postal codes are moved to create unique boundaries for each. The postal codes and populations are pooled together when moving these points is impossible. The urban areas—considered to be those regions served by a postal code *not* starting with "T0"—are examined at the 3-digit postal code (also known as forward sortation area) level. All data for these non-T0 postal codes are amalgamated, and the average geographic location is calculated. (There are some urban areas that are served by T0 postal codes, notably Drumheller and Rocky Mountain House.) The amalgamation of urban postal codes to 3 digits, in conjunction with all 6-digit rural postal codes, results in over 800 regions. Some rural postal codes may themselves be amalgamated to create geographic units with larger minimum populations. Amalgamations of rural and urban postal codes, as described, have been given the name "consolidated postal codes."

Findings

Based on these findings, the use of CSDs for health surveillance purposes is not recommended in the province of Alberta. Using RHA, CCSD, or CD boundaries results in large regions, in which local detail may not be visible. More importantly, the boundaries of all these regions change on a regular basis. Surveillance requires a consistent spatial base, so none of these boundaries are suitable.

Postal code data—and especially consolidated postal code data—are far better suited to this task in the province of Alberta because the numerator and denominator

are obtained from the stakeholders registration database.[1] The Personal Health Number (the number identifying each person in the database) is then used to link cases to this file. A constant geographic structure facilitates long-term surveillance research, but postal codes are always changing as old postal codes are retired and new ones are added. Disease surfaces created from these data may reflect changes that are the consequence of changes in postal codes, not significant changes in the data.

Clearly, an alternative method must be devised to enable researchers to effectively manage the problems described in this section (Table 1). Using latitude/longitude (lat/lon) blocks provides an innovative solution to these spatial issues.

Latitude/Longitude Blocks

Lat/lon blocks were devised in order to overcome the challenges associated with the

Table 1 Advantages and Disadvantages of Different Regional Boundary Types Used in GIS Analysis

Technique	Advantages	Disadvantages	Result
County	Similar unit used in the United States.	Data are not available at this level.	Not usable
Regional Health Authority (RHA)	Data are available at this level. Rates are somewhat stable. Decisions are made at the RHA level.	Boundaries change frequently. Regional differences within RHAs are not visible. Urban and rural data are mixed.	Used to report province-level data for public reports
Census division	Stable rates.	Boundaries change. Local variation not visible.	Not usable
Census subdivision	Population data are readily available. Regional differences are apparent. Data collection geography does not match these boundaries.	Stability of rates is questionable because large populations may be compared with small populations.	Not usable
Consolidated census subdivision	Data are readily available.	Difference between urban and rural communities is lost.	Limited use
Enumeration area	Smaller, more comparable populations.	Changes frequently. Size of EAs varies.	Not usable
Postal code	Data collection is based on these boundaries.	Changes frequently.	Limited use

[1] In Alberta, every person who wishes to receive health services must make monthly payments into the health plan. The stakeholders registration database exists to keep track of every person who makes these payments. The database also contains records of every health service; the action taken, diagnosis, etc., are assigned unique codes. For invoicing purposes, the database also contains a mailing address for each person who receives health services. The postal code information for each health event can be retrieved from this database.

techniques listed above. The northern, southern, and eastern boundaries of Alberta, as well as half of the western boundary, are formed by latitude and longitude lines (60° N, 49° N, 110° W, and 120° W, respectively). All political boundary lines are liable to change over time, but the framework that is used to describe their location (i.e., the geographic coordinate system) is not. This means that long-term surveillance can continue regardless of changes in boundaries.

Blocks based on the Universal Transverse Mercator (UTM) grid have been used to define consistent geographic regions to collect data. Many bird atlases use these blocks as a foundation, guaranteeing that any changes in the avian populations are real, not based on changes in boundaries. *The Atlas of Breeding Birds of Alberta* (2) and others have used this method. The challenge that arises with the use of UTM blocks is that blocks in which two UTM zones join become triangular, and therefore are not of the same size, meaning that populations may become too small. This may also present a cartographic problem: ensuring that the smaller blocks are visible.

Blocks generated using latitude and longitude lines may be a better choice for public health purposes. Although the blocks become smaller as latitude increases, these changes are predictable and measurable and are not as dramatic as those experienced with the UTM blocks. These lat/lon blocks can be set at any size, based on the population distribution characteristics of the region and also based on the number of cases associated with the health event under consideration.

The case and population structure information from each block's corresponding postal codes are used to amalgamate and display needed information. This ensures the protection of the confidential information. Because it creates blocks with larger populations, amalgamating information also makes it possible to calculate rates with more reliability and statistical confidence.

A number of diseases have been mapped using 1×2 lat/lon degree blocks—blocks of 1° latitude by 2° longitude. This technique is very effective because 1×2 lat/lon blocks are small enough to allow visual identification of regional patterns, but large enough to ensure large enough populations to allow accurate calculation of rates (Figure 2). Two-by-two (2×2) lat/lon degree blocks have been used to map diseases that affect smaller populations (Figure 3). Smaller grids (Figures 4 and 5) and variable-size grids (grids that are smaller in densely populated areas and larger in other areas) have been tested on some diseases, but the 1×2 and 2×2 grids are well suited for the province. It would be easy to use smaller grid units if necessary. It would also be easy to extend the same system to other regions. This method will be used in the pilot of a national surveillance system for Canada.

The method has also proven effective in mapping potential determinant data. Socioeconomic information is available from Statistics Canada, presented using any of the boundaries listed above. It is often difficult to examine data presented in a spatial format in context of health data that use other boundaries, but the EAs are small enough that the data can be aggregated to the same lat/lon blocks as the health data. The resulting maps can be readily compared at a glance (Figure 6). Environmental data are often more difficult to examine because the density of data collection can vary vastly from region to region. Data on each water well data in the province, for example, can be obtained from the provincial government, but in examining this information it is difficult to discern any patterns that may have an influence on human health. One solution to this problem is showing the total number of wells as a graduated circle, while

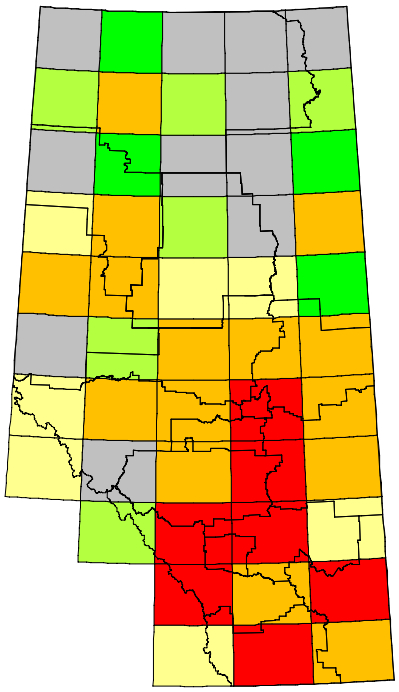


Figure 2 1×2 grid.

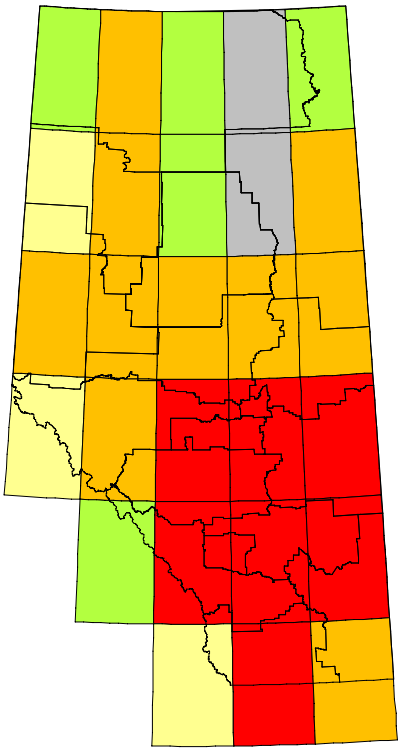


Figure 3 2×2 grid.

showing the proportion of wells that exceed national guidelines for a particular element as a pie sector (Figure 7). It is possible, then, to identify at a glance those areas that have large numbers of wells that exceed guidelines, and the reporting boundaries are identical to those used to map a potentially related health event. This method also makes it possible to ensure that confidentiality is maintained for health data as well as potential determinants.

The lat/lon blocks are an excellent vehicle for surveillance and for identifying regions in which a series of health events occur more often than is normal. The technique is not as well suited to analyzing smaller regional patterns. The size of the blocks can be decreased, but this can generate errors, because the allocation of a postal code to a block has greater repercussions when the population of each of these blocks is smaller. The small blocks will be very useful when health data can be obtained that are more precise than postal code-level data.

Conclusion

GIS is a useful tool within the field of epidemiology, but software and training costs often prevent successful implementation, as does access to base data. The researcher must be aware of the challenges associated with using various spatial units. The discussion above outlines a method developed at Alberta Health Surveillance for presenting and analyzing health and determinant data from a spatial perspective. This method

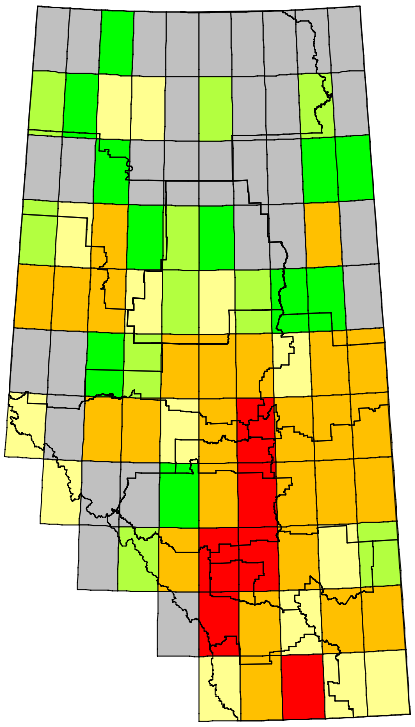


Figure 4 1×1 grid.

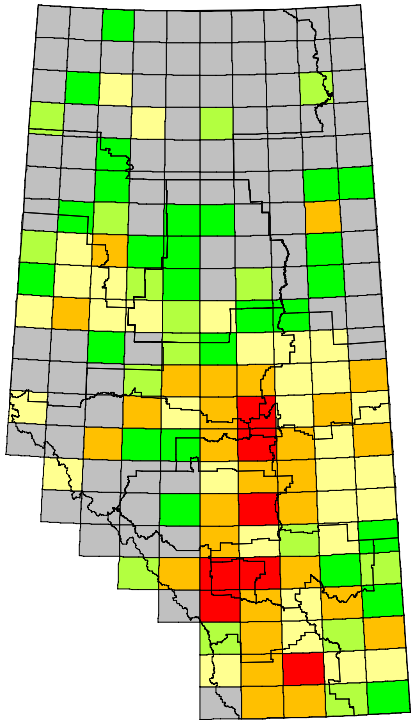


Figure 5 .5×1 grid.

is well suited to examining provincial data from a long-term surveillance effort. (The examination above of other possibly suitable boundaries provides a context for the creation of this method.)

The use of lat/lon blocks is recommended in jurisdictions where data to be compared are collected using different boundaries. The method is also well suited for jurisdictions in which limited budgets have not allowed the development of spatial health analysis, because much of the work can be performed without access to GIS tools. Block membership can be obtained from the lat/lon coordinates of any point; maps can be generated by filling in the colors in each block using the graphics editing capabilities of any current word processor; and the bubble chart option in many spreadsheet programs is able to generate simple pie chart maps.

The method demands few resources, provides a consistent geographic structure, simplifies comparisons among datasets collected by different agencies, and provides a means of examining issues that cross political boundaries. For these reasons, the use of the lat/lon block method is recommended.

References

1. Clarke KC, McLafferty SL, Tempalski BJ. 1996. On epidemiology and geographic information systems: A review and discussion of future directions. *Emerging Infectious Diseases* 2(2).
2. Semenchuk Glen P, Ed. 1992. *The atlas of breeding birds of Alberta.* Edmonton, Alberta, Canada: Federation of Alberta Naturalists.

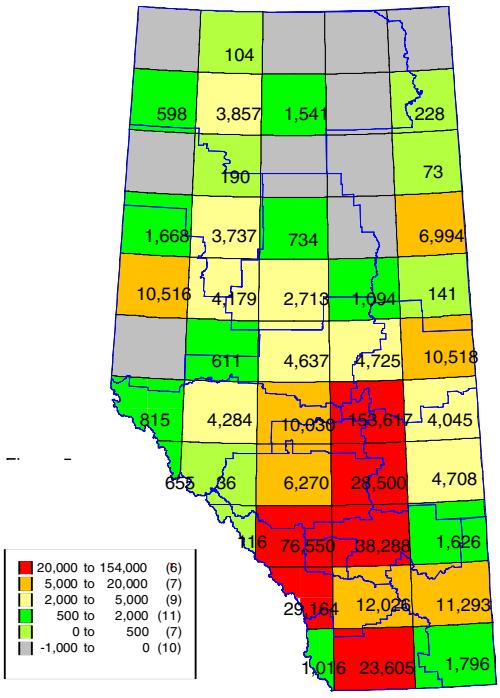


Figure 6 Socioeconomic data summarized to lat/lon block.

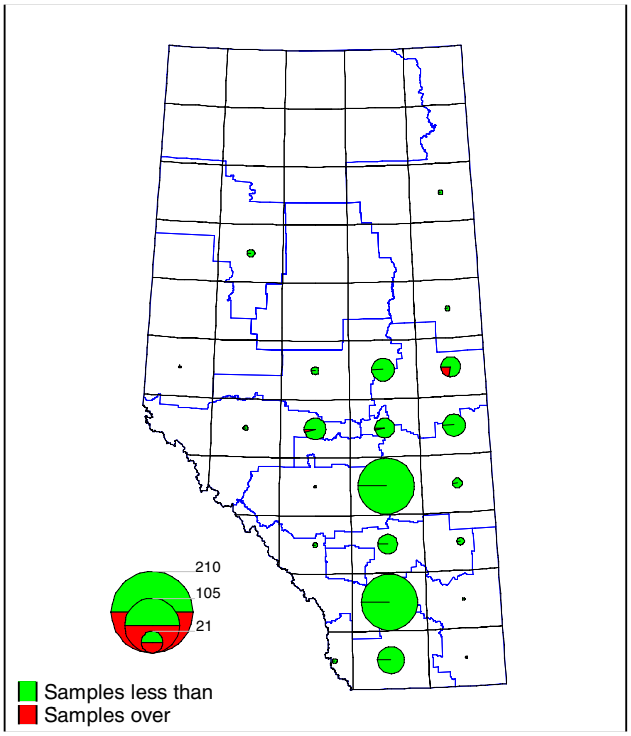


Figure 7 Water contamination summarized to lat/lon.

Spatial and Environmental Risk Factors for Diarrheal Disease in Matlab, Bangladesh

Michael Emch*
Department of Geography, University of Northern Iowa, Cedar Falls, IA

Abstract

The objective of this research project is to assess risk for diarrheal disease in rural Bangladesh by analyzing the complex and dynamic interaction of biological, socioeconomic, cultural/behavioral, and environmental factors over time and space. Risk factors of cholera and non-cholera watery diarrheal disease are calculated to compare the relative importance of risk for several independent variables. Diarrheal disease data were collected for people who were hospitalized at the International Centre for Diarrhoeal Disease Research (ICDDR) hospital (Matlab, Bangladesh) from January 1, 1992, to December 31, 1994. Using laboratory and hospital records, cases were assigned to one of two diarrhea disease categories (cholera or non-cholera watery diarrhea) that were used as dependent variables in the analysis stage of the research. Age-matched individuals were randomly chosen from the community to be controls. Information was collected for independent variables that were hypothesized to be related to watery diarrhea. This information was collected by administering questionnaires, obtaining secondary data from the ICDDR's demographic surveillance system records and community health worker record books, and calculating variables using a geographic information system database. Sanitation and water availability and use are extremely important in the effort to reduce secondary transmission of cholera and non-cholera watery diarrhea. Water use and availability variables were more important for non-cholera watery diarrheal risk than for cholera, but they were important for both. Socioeconomic status is an important indirect cause of both of these diseases because poverty is the root cause of many of the other variables, such as lack of sanitation and clean water. Flood control was related to both types of diarrhea, but it is not understood why. Because the Bangladesh Flood Action Plan maintains and will continue to build flood-control embankments, it is important to investigate whether there is a pattern to this relationship throughout the country and to investigate why the relationship exists.

Keywords: diarrheal disease, cholera, Bangladesh, medical geography

Introduction

Diarrheal diseases cause one-third of the 15 million annual deaths in children under five years old in the developing world (1) and they are the largest cause of death among children under five in Bangladesh (2,3). The people of Bangladesh suffer not only directly, when they contract the disease, but also indirectly, from economic hardship due to lost productivity and medical expenses. Because of resource constraints in developing countries like Bangladesh, it is necessary to identify risk factors so preventative

* Michael Emch, Department of Geography, University of Northern Iowa, Cedar Falls, IA 50614 USA; (p) 319-273-7768; (f) 319-273-7103; E-mail: Mike.Emch@UNI.EDU

health programs can focus on particular interventions. Assessing risk for diarrheal disease requires knowledge of the complex and dynamic interaction of biological, socioeconomic, behavioral, cultural, and environmental factors over time and space. The objective of this study is to advance such knowledge in the context of rural Bangladesh.

Humans were the only known reservoir of *Vibrio cholerae* until the mid-1980s, when theories of the ecology of cholera were substantially revised. During this time, Colwell et al. (4) published the results of a study claiming that vibrios can live freely in an aquatic environment, even under conditions of nutrient deprivation, if the environment is not sodium-free. Prior to this study, it was maintained that cholera was only transmitted by ingestion of feces-contaminated food or water. However, Colwell's research suggests that transmission can occur through water without fecal contamination. If transmission can occur without fecal contamination, then these findings dramatically change long-standing conceptions of the ecology of cholera.

This study differentiates between two types of diarrhea: cholera and non-cholera. Cholera watery diarrhea is defined as watery diarrhea caused by the bacterium *Vibrio cholerae.* Non-cholera watery diarrhea is defined as watery diarrhea caused by microorganisms other than *Vibrio cholerae.* Ideally, this study would have distinguished between all of the non-cholera diarrheal agents; however, the microbiological tests associated with obtaining this information would have been exorbitantly expensive. This study differentiates between risk factors for cholera and non-cholera diarrhea.

The research was conducted at the International Centre for Diarrhoeal Disease Research (ICDDR). The ICDDR has a field station called Matlab, where the Centre's diarrhea treatment hospital is located. It is in south central Bangladesh, approximately 50 kilometers southeast of Dhaka, adjacent to where the Ganges River meets the Meghna River forming the Lower Meghna River. Figure 1 shows the study location within Bangladesh relative to Dhaka City, three major South Asian rivers, and the Bay of Bengal.

Conceptual Framework

Analyzing risk of contracting watery diarrheal disease in Bangladesh requires a conceptual framework that addresses the complexities of biological, socioeconomic, cultural/behavioral, and environmental factors over time and space. A medical geographic theoretical approach that addresses these issues is disease ecology, which maintains that disease results from a dynamic complex of variables that coincide in time and space (5–15). Hunter (16) argues that researchers must not take a pathogen-centric view of disease, that is, one that focuses only on the disease agent. He suggests that studies of disease "must co-jointly involve pathogen, host, and environment" (16). He views "environment" broadly, as consisting of "diverse physical, biological, social, cultural, and economic components" (16). Hunter defines geography as a discipline that bridges the social and environmental sciences and writes that "its integration and coherence derive from systems-related analysis of man-environmental interactions through time and over space" (16).

This paper is intended to demonstrate the value of a medical geographic approach that is holistic and that integrates many different types of variables responsible for disease. The types of variables to be investigated have been classified in many different ways, but Mayer's classification system (17) is most useful. Mayer differentiates

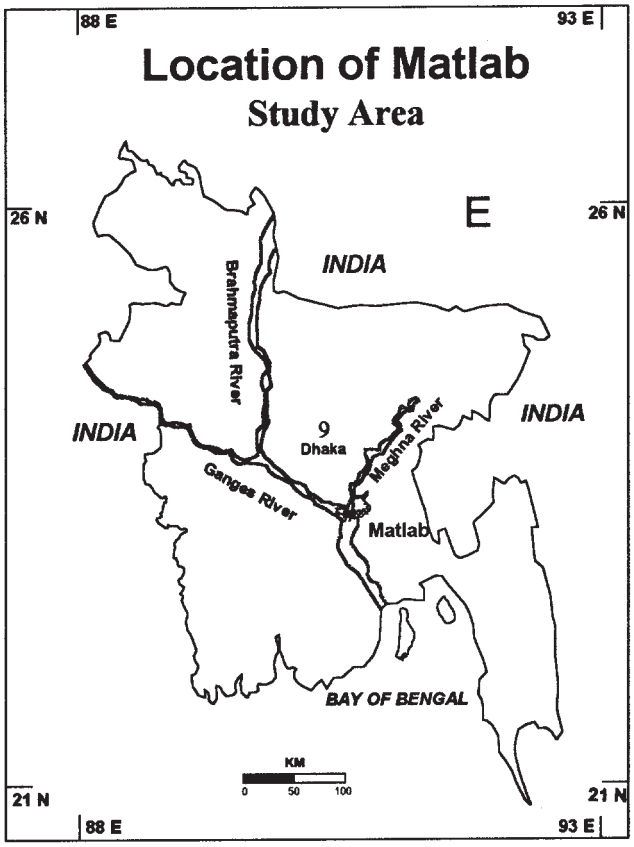


Figure 1 Location of diarrhea treatment center, Matlab, Bangladesh.

between biological, socioeconomic, behavioral, and environmental variables. Biological variables are those that describe biological characteristics of the host, such as blood type. Behavioral variables are those that describe individual or group behaviors, and may be related to culture or individual decision-making (for example, what types of food people eat). Environmental variables are those of the biophysical environment, such as climatic variables. Socioeconomic variables are variables that affect the coincidence of agent and host, such as wealth or class. Different patterns of socioeconomic, behavioral, and environmental variables result in different spatial and temporal patterns of disease. Virtually every disease exhibits spatial and temporal variation, and medical geographers attempt to explain this variation.

Research Design

The author created a vector geographic information system (GIS) database of the Matlab field research area. Features in digital format include baris, rivers, roads, and a flood-regulating embankment. Baris are patrilineally related clusters of households. Figure 2 shows three features in the GIS database: the flood-regulating embankment,

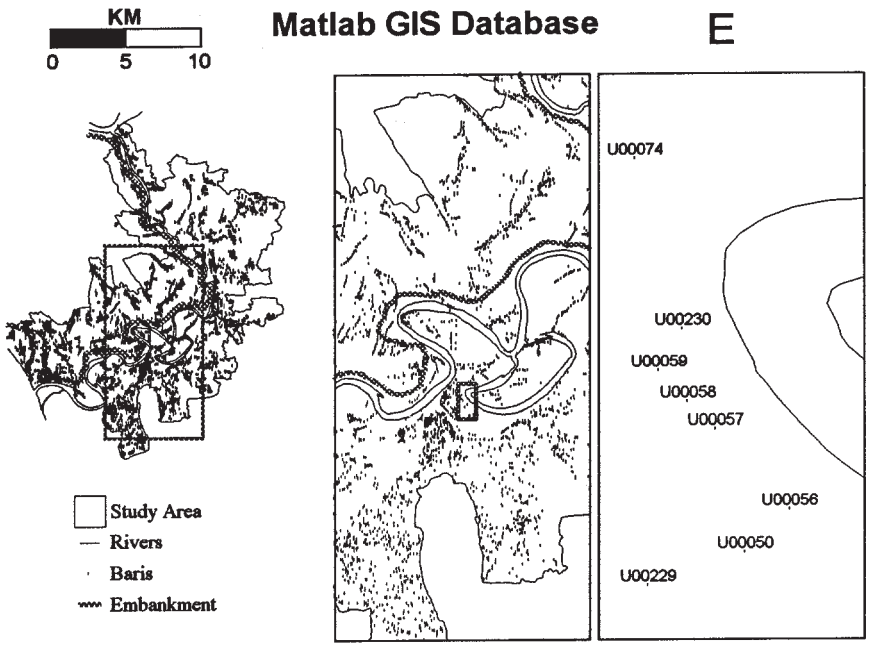


Figure 2 Study area of GIS database.

the Dhonagoda River, and baris. The three map views in Figure 2 are displayed at different scales. The map view on the far right has the individual bari identification numbers visible. The baris are all identified by an ICDDR demographic surveillance system (DSS) census number within the structure of the GIS database. This allows attribute data to be linked to the spatial database. Thus, disease incidence data can be linked to specific bari locations.

The Matlab field research center is a diarrhea treatment center (DTC) that has in- and outpatient services, a laboratory for the identification of pathogens, and research facilities. The DTC laboratory consists of microbiology, clinical pathology, and biochemistry units, which provide diagnostic services to the hospital and for field research activities. There are motorized boats that function as a free ambulance service for diarrhea patients, so access to the hospital is remarkably good. All DTC services are free as well. The research center maintains a community-based data collection system. One hundred twenty community health workers (CHWs) visit each household every two weeks to collect demographic, morbidity, and other data. The DSS conducts periodic censuses (most recently in 1993) and uses CHWs to update demographic data (births, deaths, and migrations).

Diarrheal disease data were collected for people from the Matlab treatment area who were hospitalized at the diarrhea treatment center with watery diarrhea between January 1, 1992, and December 31, 1994. The cases were assigned to one of two diarrhea disease categories (cholera or non-cholera watery diarrhea) that were used as dependent variables in the analysis stage of the research. For each patient admitted to the

Matlab DTC, a stool sample was regularly collected and routinely tested for *Vibrio cholerae* and *Shigella*, a dysenteric agent. In this study, laboratory records of the patients were used to assign one of the two above agent categories. Hospital records specify whether there was blood in each patient's stool. Patients who tested positive for *Shigella* or who had blood in their stool were excluded because this study is not concerned with dysentery. Patients who did not have dysentery or cholera were assigned to the non-cholera watery diarrhea category. For each patient with cholera or non-cholera watery diarrhea, the bari identification number was collected for mapping.

Individuals were randomly chosen from the community to be controls. After the cases were identified, a list of potential controls was compiled from DSS records. A person was eligible to be a control if she/he lived in the Matlab surveillance area, was not admitted to the DTC during the study period, and did not die of a diarrheal disease during the study period. The controls were age-matched. For cases of diarrhea in persons older than five years of age, controls were chosen who were born in the same year. For those less than five years old, controls were chosen who were born in the same month. Children under five had a stricter age-matching interval because there were more potential controls who were in this age group. In addition, calculating certain biological independent variables for children required a smaller age-matching interval because the status of these variables was collected on a monthly basis.

Information was collected for independent variables that were hypothesized to be related to watery diarrhea. This information was collected by administering questionnaires, obtaining secondary data from DSS records and community health worker record books, and calculating variables using the GIS database. These data were collected for both cases and controls. Tables 1, 2, and 3 summarize the different variables that were collected.

Results

Cholera

Two of the environmental independent variables—living in a household that shared its latrine with other households and living in a flood-control area—were strongly associated with cholera hospitalization. Participants whose households shared latrines with other households had a 2.8 times greater chance of being hospitalized with cholera. Sharing latrines represents increased exposure to the fecal material of others, which can lead to secondary transmission. Individuals living in flood-controlled areas were 2.47 times more likely to be hospitalized with cholera. It is not entirely clear why this is true. One theory is that flood control exacerbates cholera bloom by some unknown mechanism (18). Flood control may change salinity levels or may impede the natural flushing out of cholera-laden water. The association between cholera and flood control may, however, be entirely unrelated to flood control. There may be another variable that is associated with flood control, creating a spurious association between flood control and cholera hospitalization. In the future, cholera incidence rates in other flood-controlled areas of Bangladesh should be compared with rates in their surrounding areas. Similarities of the environments of these flood-controlled areas should be identified so that, if cholera incidence is higher in these areas, a causal pathway can be determined. The multi-billion dollar Bangladesh Flood Action Plan may or may not be responsible

Table 1 Summary of Categorical Independent Variables with Two Classes

Variable	Variable Type	Description
Gender	Cultural/behavioral and biological	Male or female
Source of drinking water	Cultural/behavioral	Tubewell or other
Source of cooking water	Cultural/behavioral	Tubewell or other
Source of bathing water	Cultural/behavioral	Tubewell or other
Source of washing water	Cultural/behavioral	Tubewell or other
Working tubewell in bari	Environmental	Yes or no
Adult male defecation	Cultural/behavioral	Latrine or other
Adult female defecation	Cultural/behavioral	Latrine or other
Male child defecation	Cultural/behavioral	Latrine or other
Female child defecation	Cultural/behavioral	Latrine or other
Presence of latrine in household	Environmental	Yes or no
Type of latrine drainage	Environmental	Septic or not
Number of households using a latrine	Environmental	Single or multiple
Consumption of shellfish	Cultural/behavioral	Yes or no
Flood-controlled area	Environmental	Yes or no
Breastfeeding status of children under 5	Biological	Yes or no
Nutritional status of children under 5	Biological	Malnourished or not

Table 2 Summary of Categorical Independent Variables with More Than Two Classes

Variable	Variable Type	Description
Years of education: adult (over 15) participant	Socioeconomic	More than six; one to six; none
Years of education: mother	Socioeconomic	More than six; one to six; none
Years of education: father	Socioeconomic	More than six; one to six; none
Knowledge of prevention of diarrhea	Cultural/behavioral	Full; good; partial; none
Knowledge of source of diarrhea	Cultural/behavioral	Good; partial; none
Household construction material	Socioeconomic	Brick,/tin; bamboo/tin; jute/tin; straw/stick/bamboo

for increased cholera rates. Thus, it is important to investigate whether or not flood control is contributing to transmission of this disease.

Several of the cultural/behavioral variables that describe the water and sanitation situation of study participants did not reveal associations. Use of tubewell[1] water for drinking, cooking, bathing, or washing was not related to cholera hospitalization. This certainly does not mean that people do not need to use tubewell water to avoid contracting cholera. Almost all of the questionnaire respondents (95%) said that they regularly use tubewell water for drinking, so there is not a major problem with drinking water use. Defecation in places other than a latrine, households without latrines, and households with open latrines were not associated with cholera transmission. It is unclear why these variables (which represent an unsanitary environment) were not

[1] A tubewell is a drinking water well with a pump for extracting water from the shallow aquifer.

Table 3 Summary of Continuous Independent Variables

Variable	Variable Type	Description
Number of open latrines	Environmental	Count
Number of non-septic latrines	Environmental	Count
Number of ring septic latrines	Environmental	Count
Number of concrete septic latrines	Environmental	Count
Number of other households using latrines	Cultural/behavioral, environmental	Count
Latrines per person (excluding open)	Environmental	Latrines per 100 people
Number of tubewells in bari	Environmental	Count
Number of households sharing a common tubewell in bari	Cultural/behavioral, environmental	Count
Tubewells per person	Environmental	Tubewells per 100 people
Household area	Socioeconomic, environmental	Square feet
Bari population	Cultural/behavioral, environmental, socioeconomic	Count
Population density around baris	Cultural/behavioral, environmental, socioeconomic	Persons within half-kilometer radius
Total household assets	Socioeconomic	Taka
Annual income	Socioeconomic	Taka
Mid-arm circumference (children under 5 ears old)	Biological	Millimeters
Distance from main river	Environmental	Meters

associated with cholera, because they are no doubt responsible for secondary cholera transmission. There were also continuous variables associated with sanitation and water availability; these variables will be discussed below.

Another cultural/behavioral variable, shellfish consumption, was not associated with cholera transmission either. This is contrary to one of Colwell's (4) theories about an environmental reservoir for cholera. She believes that shellfish are one of the attachment sites for the bacteria. The lack of an association might be due to the fact that 92% of the people in the study population consume shellfish. The only people who might not consume any type of shellfish are extremely poor, and are thus more prone to contracting cholera because of other variable types (such as socioeconomic variables and variables involved with their access to and use of clean water and proper sanitation).

Two biological variables, breastfeeding and malnutrition, were not associated with cholera transmission. This may be attributed to the low number of child participants who were not breastfeeding during the month before hospitalization (23%) or who had a mid-arm circumference below 120 millimeters (12%).

The independent variables that had more than two ordinal classes included level of education for different household members, household construction material, and knowledge of diarrhea prevention and source. Education level and household construction material are socioeconomic variables that were hypothesized to show a negative association with cholera incidence; surprisingly, there were no associations. Knowledge about the source and prevention of diarrhea were hypothesized to be

inversely associated with cholera hospitalization, but there were no associations in that case either.

Modeling a complex problem such as what makes someone susceptible to contracting cholera requires that a variety of methods be used. Non-parametric statistics were used to measure associations between cholera and potential risk factors, and simple regression analysis was used for the continuous variables. The larger the number of open latrines in a bari, the more likely a resident was to contract cholera. Open latrines are basically fixed sites where people regularly defecate. These fixed sites are an indicator of an unsanitary environment. The number of households using tubewells was positively related to cholera hospitalization. It is unclear why this association exists but a speculation is offered. If many households share a tubewell, it may decrease access to that tubewell; thus, this relationship might indicate that access to tubewell water is important to preventing cholera.

Bari population and population density were positively related to cholera incidence. While it is not completely clear why bari population size is related to cholera hospitalization, one conjecture is that the larger the bari population, the larger the number of human contacts people have. The last variable that was related to cholera hospitalization was household area, a socioeconomic and environmental variable. (Household area is a socioeconomic indicator because people with smaller households are usually poorer, and it is environmental because smaller households represent a condition of crowding.) Household area was inversely related to hospitalization for cholera. There were two other socioeconomic indicators, assets and income, that were built into simple logistic regression models. However, neither was found to be related to hospitalization for cholera. Conroy (19) suggests that socioeconomic status in the developing world is a complex issue and that assets and income measure different parts of socioeconomic status. He states that income is an indicator of purchasing power and consumption, while assets are an indicator of a person's ability to develop options for improving their quality of life (e.g., participating in poverty alleviation programs). A house is part of a family's assets, although houses were not included in this study's original measurement of assets. Household area indicates how much a person is able to invest in their home, which is why the ICDDR collects this information regularly. While the variation of assets and income is quite small, there is a much larger variation in household area. The inverse relationship between household area and cholera shows that it is an important factor. The author believes that the environmental part of household area, which is a measure of crowding, and the socioeconomic part of this variable, which describes the socioeconomic status of a family, are inseparable yet both important. Crowding, however, is more likely to occur in poorer households. Poor people are at a major disadvantage in many other parts of their lives in rural Bangladesh. They are forced to eat cheaper food, which may be unsanitary. They may not be able to invest in proper water and sanitation facilities. Even if an outside organization is paying for the water and sanitation facilities, poorer people are less likely to have these facilities in their baris because they have less social power to influence how these resources are distributed. It is the belief of the author that poorer people are exposed to diseases at higher rates.

Because several different variables may interact in affecting risk of cholera hospitalization, a multiple logistic regression model was built using many independent variables. Because models were devised only for observations for which there were

data for all of the variables, the relationships do not refer to the same sample to which the simple regression models refer. Four of the variables that were significant in simple logistic regression models were also significant in the multiple logistic regression model. These were the number of open latrines in a bari, the household area, the bari population, and flood control.

Non-cholera

The risk factors for non-cholera watery diarrhea were somewhat different from the risk factors for cholera. Although some of the significant variables were the same as for cholera, the strengths of the associations were different. Four of the binary dependent variables were significantly associated with hospitalization for non-cholera watery diarrhea. Female participants were only 0.81 times as likely to be hospitalized with non-cholera watery diarrhea as males. In rural Bangladesh, males have more freedom of movement than females, so they are more likely to come into contact with a larger number of people. Contact with more people can lead to increased exposure to people infected with non-cholera watery diarrhea.

Participants who did not use tubewell water for drinking were 8.49 times more likely to be hospitalized with non-cholera watery diarrhea than were those who did use tubewell water. This extremely high association highlights the importance of clean drinking water for avoiding non-cholera watery diarrhea. There was also a relatively high association between not using tubewell water for bathing and hospitalization for non-cholera watery diarrhea. (Very few people actually bathe with tubewell water, and it is not a feasible public health option to change this. It would require a large educational effort to change the custom of bathing in rivers or ponds.) Individuals living in flood-controlled areas were 1.42 times more likely to be hospitalized with non-cholera watery diarrhea than individuals not living in flood-controlled areas. This was not as strong an association as with cholera, and, again, it is not entirely clear why there is an association. Future work must be conducted to ascertain the reason for this association and to investigate whether it exists in other flood-controlled areas of Bangladesh.

Several variables involving water availability and sanitation were not associated with non-cholera hospitalization. The absence of a working tubewell in a participant's bari did not lead to a greater hospitalization rate for non-cholera watery diarrhea. Defecation in places other than latrines was not associated with hospitalization, nor were participants who lived in households without latrines or who had latrines with open drainage systems more likely to be hospitalized with non-cholera watery diarrhea. Participants who shared latrines with other households did not have a greater chance of being hospitalized. The finding that these water and sanitation variables were not associated with watery diarrhea incidence does not mean that they are not important. The reason why so many water and sanitation variables were collected is that previous research has identified them as important. The strongest association of any water and sanitation variable for non-cholera watery diarrhea was the negative association found for using tubewell water as a drinking source. Thus, this portion of the issue of overall water and sanitation is most important.

Shellfish consumption, breastfeeding, and malnutrition were not associated with non-cholera watery diarrhea incidence, possibly because there was little variation in these variables. None of the ordinal-level variables (education for different household

members, household construction material, or knowledge of diarrhea prevention and source) was related to non-cholera watery diarrhea.

Simple regression analysis for continuous variables was also used to calculate risk of non-cholera watery diarrhea. Household area was inversely related to hospitalization for non-cholera watery diarrhea, as with hospitalization for cholera. Also as with cholera, assets and income—the two other socioeconomic indicators—were not related to non-cholera hospitalization. In accordance with the non-parametric test, people living in a flood-controlled area were more likely to be hospitalized with non-cholera watery diarrhea than were those not living in a flood-controlled area. There was one biological variable associated with non-cholera watery diarrhea that was not associated with cholera. Mid-arm circumference was related to non-cholera watery diarrhea hospitalization at the 95% confidence level. The ICDDR considers a mid-arm circumference of less than 120 millimeters to indicate malnutrition in children under five years old, in this study population. Stratifying this variable as above and below 120 revealed no association, but the raw data values show a relationship. There were only 37 observations for this variable, which may explain the absence of an association using the non-parametric test.

Three variables were significant in a multiple logistic regression model for non-cholera watery diarrhea using variables that were at least moderately significant in the simple regression models. The variables were household area, flood control, and tubewell density. Household area was negatively related to non-cholera hospitalization—the smaller the household, the more likely it was that an individual would be hospitalized. As with cholera, people living in flood-controlled areas were more likely to be hospitalized with non-cholera watery diarrhea. Tubewell density was highly significant when built into a multiple logistic regression model, but was only moderately significant in a simple regression model. This indicates that there was interaction with some other variable. As tubewell density increased, non-cholera watery diarrhea hospitalization decreased. This relationship highlights the importance of clean water availability as a protective barrier to non-cholera hospitalization.

Discussion

It becomes apparent on arrival in rural Bangladesh that people there are always in close contact with the aquatic environment. The aquatic environment is an important source of income for fishers and farmers and it provides the most important system of transportation for people living in the study area. Another readily apparent characteristic of the study area is that there are many people living in a small area and that all land seems to be used for some economic activity. By developing-world standards, it is also clear that almost everyone living in the study area is extremely poor. Figure 3 displays the factors that were found to be statistically significant in cholera transmission.

Only two variables that describe characteristics of water and sanitation infrastructure or use were related to cholera transmission. Secondary cholera transmission is by the fecal-oral route, and is thus due to the lack of clean water and good sanitation. (In fecal-oral transmission, people are infected when they ingest something that has been contaminated with fecal material.) Thus, it is no surprise that two water- and sanitation-related variables are associated with cholera.

Other variables related to cholera transmission describe the number of people

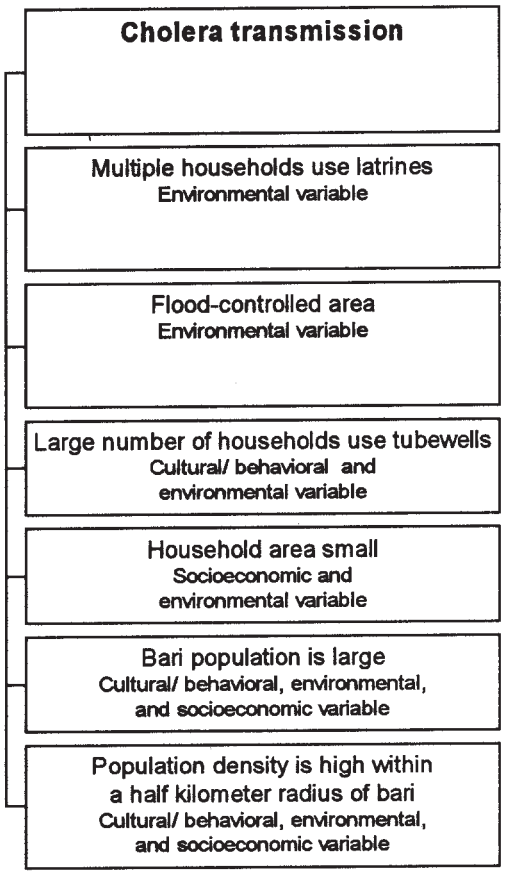


Figure 3 Variables involved in cholera transmission.

living in a bari, the population density near a bari, and the size of a housing structure. All of these have to do with the environmental circumstances in which people are living. Several of these variables show that people living in crowded areas get cholera more often.

The last variable related to cholera transmission is flood control, another environmental variable. People living inside a flood-controlled area are living in an environment that has been significantly altered by humans. This alteration certainly changes the way people interact with their environment in these areas. For example, the agricultural system in the flood-controlled area is more reliant on irrigation. It is unclear why there is an association between flood control and cholera, but it may have something to do with how people are interacting with the aquatic environment in this area.

Non-cholera transmission is exclusively secondary, via the fecal-oral route. The study area is littered with latrines that hang over water bodies that are used for bathing, washing clothes, cooking water, and occasionally for drinking water. In such a densely populated area, it is safe to say, the surface water is not fit for drinking. Figure 4 displays the factors important to non-cholera watery diarrhea transmission.

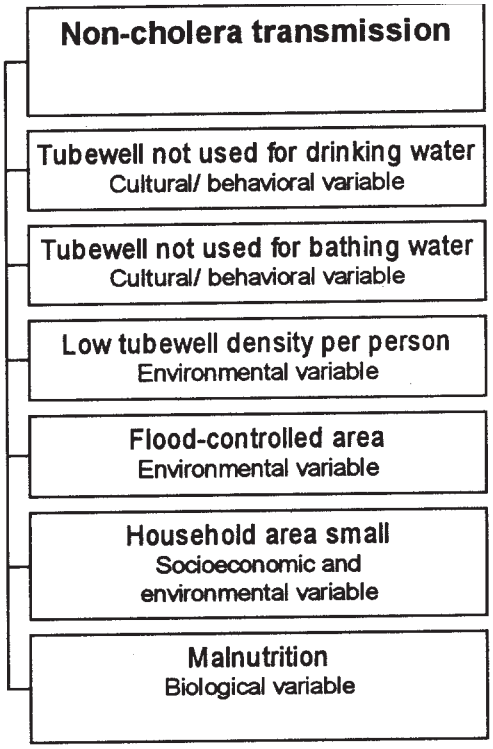


Figure 4 Variables involved in non-cholera diarrhea transmission.

Three of the variables shown in Figure 4 involve tubewell water use. Because there is no water treatment facility in the area, tubewells are the only clean source of water. Other variables associated with transmission of non-cholera watery diarrhea transmission include household area, malnutrition, and flood control. There are many types of variables that predispose people to secondary transmission of diarrhea. If clean water and sanitary latrines were used, however, then secondary transmission would be much less of a problem.

Conclusion

It is clear that sanitation and water availability and use are extremely important in the effort to reduce transmission of secondary cholera and non-cholera watery diarrhea transmission. While this may seem obvious to many outsiders, health policy makers in Bangladesh and international aid organizations continue to debate whether the appropriate tubewell coverage threshold has been achieved in rural Bangladesh. The water use and availability variables were more important for non-cholera watery diarrhea than for cholera, but they were important for both. With the exception of UNICEF, there has been very little effort to provide septic latrines to the people of rural Bangladesh even though only 10% of the study area population had concrete septic latrines. Another debate among health policy makers concerns how to increase latrine coverage.

The status quo has been that latrines are usually provided at the expense of the family or community. Because diarrhea is a poor person's disease, however, the people who need proper sanitation most are those who are least likely to be able to afford it. This research project found significant relationships between sanitation-related variables and cholera, but not between the same variables and non-cholera diarrhea. It is important to note, however, that the sanitation situation in the entire study area was very poor, so comparisons with ideal sanitation conditions could not be made.

One of the socioeconomic status indicators was related to both cholera and non-cholera watery diarrhea. The author suspects that if the study population were compared with a more affluent group, more relationships would become apparent between socioeconomic variables and the diseases. Socioeconomic status is probably the single most important indirect cause of both of diseases of concern because poverty is the root cause of many of the other variables, such as lack of sanitation and clean water. The educational level, income, assets, and living environment of the study population are abysmal. The poverty, however, will no doubt continue and these diseases will most likely continue as well. A stronger national and international policy directed at poverty alleviation in rural Bangladesh is necessary to tackle such a difficult problem. A relationship was found between malnutrition and non-cholera watery diarrhea but not for cholera. There is contradictory information in the health literature concerning the effect of malnutrition on diarrhea. It is obvious, however, that malnutrition is already a health policy concern and thus is already on the health care agenda.

Flood control was related to both types of diarrhea, but it is not understood why. Because the Bangladesh Flood Action Plan will continue to build and maintain embankments into the distant future, it is very important to investigate whether there is a pattern to this relationship throughout the country and to investigate why the relationship exists. This will no doubt require a multi-disciplinary effort involving ecologists, hydrologists, engineers, epidemiologists, and medical geographers.

References

1. Snyder JD, Merson MH. 1982. The magnitude of the global problem of acute diarrhoeal disease: A review of active surveillance data. *Bulletin of the World Health Organization* 60:605–13.
2. D'Souza S. 1985. *Mortality case study Matlab, Bangladesh.* Special Publication No. 24. The International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh.
3. Hoque BA, Hoque MM. 1994. Environment and health. In: *Environment and development in Bangladesh.* Ed. AA Rahman, S Huq, R Haider, EG Jansen. Dhaka, Bangladesh: The University Press Limited. 359–73.
4. Colwell RR, Brayton PR, Grimes DJ, Roszak DR, Huq SA, Palmer LM. 1985. Viable, but non-culturable *Vibrio cholerae* and related pathogens in the environment: Implications for release of genetically engineered micro-organisms. *Bio/Technology* 3:817–20.
5. May JM. 1958. *The ecology of human disease: Studies in medical geography.* New York: MD Publications.
6. May JM. 1977. Medical geography: Its methods and objectives. *Social Science and Medicine* 11:715–30.
7. Mayer JD. 1982. Relations between two traditions of medical geography: Health systems planning and geographical epidemiology. *Progress in Human Geography* 6:216–30.

8. Mayer JD. 1984. Medical geography: An emerging discipline. *Journal of the American Medical Association* 251:2680–3.
9. Mayer JD, Meade MS. 1994. A reformed medical geography reconsidered. *The Professional Geographer* 46:103–6.
10. Meade MS. 1977. Medical geography as human ecology: The dimension of population movement. *The Geographical Review* 67(4):379–93.
11. Meade MS, Florin JW, Gesler WM. 1988. *Medical geography.* New York: The Guilford Press.
12. Learmonth A. 1988. Disease ecology. New York: Basil Blackwell.
13. Paul BK. 1985. Approaches to medical geography: An historical approach. *Social Science and Medicine* 20:399–409.
14. Pyle GF. 1977. International communication and medical geography. *Social Science and Medicine* 11:679–82.
15. Pyle GF. 1979. Elements of disease ecology. In: *Applied medical geography.* Ed. GF Pyle. New York: John Wiley and Sons.
16. Hunter JM. 1974. The geography of health and disease. In: *The challenge of medical geography: Studies in geography no. 6.* Ed. JM Hunter. Chapel Hill, NC: Department of Geography, University of North Carolina, Chapel Hill. 3–6.
17. Mayer JD. 1986. Ecological associative analysis. In: *Medical geography: Progress and prospect.* Ed. M Pacione. London: Croom Helm. 64–81.
18. Colwell RR, Spira WM. 1992. The ecology of Vibrio cholerae. In: *Cholera.* Ed. D Barua, WB Greenough. New York: Plenum Medical Book Company. 107–28.
19. Conroy ME. 1997. *The challenges of economic geography in defining, creating, and defending sustainable communities.* Presented at the National Science Foundation Workshop on the Future of Economic Geography. Washington, DC. September 27, 1997.

Design and Implementation of a Geographic Information System for the General Practice Sector in Victoria, Australia

Julie B Green (1),* Francisco J Escobar (2), Elizabeth Waters (1), Ian P Williamson (2)

(1) Centre for Community Child Health, University of Melbourne, Royal Children's Hospital, Melbourne, Victoria, Australia; (2) Department of Geomatics, University of Melbourne, Melbourne, Victoria, Australia

Abstract

This paper details a collaborative research project to develop a geographic information system (GIS) for two diverse administrative areas of general medical practitioners in Victoria, Australia. The study is one of a small number of initiatives in the use of geospatial information and application of GIS technology to the health sector in Australia. Australia's use of divisions of general practice is described, depicting the role of divisions in improving the health of the Australian population. An outline is given of the role of data and information technology in improving effectiveness and efficiency in the operation of these divisions. The paper describes the methodology of the pilot project, which was aligned to the divisions' needs and future directions. Data were drawn from routinely collected demographic, health, and road network datasets; the datasets themselves came from local, state, and federal sources. Additional data were collected using a questionnaire that profiled general medical practices. The rationale for using the Internet to present the GIS prototype is given. The paper also presents a range of data analysis that depicts the role of this integrated information in identifying strategic decision-making and further research possibilities. This project demonstrates the potential of a GIS, with its ability to answer spatial questions and illustrate spatial relationships, to assist in decision-making in local health areas. Routine collection of morbidity and treatment information at the general practice level would enhance data quality at that level. The methodology and outcomes of this project are serving as a springboard to broader interest in the uptake of GIS in the health sector, given the diversity and widespread location of the population.

Keywords: general practice, service planning, Australia

Introduction

This paper details a collaborative research project to develop a geographic information system (GIS) for two diverse administrative areas of general medical practitioners in Victoria, Australia. The study is one of a small number of initiatives in the use of geospatial information and application of GIS technology to the health sector in Australia. The paper first briefly describes Australian initiatives in GIS and health, and sets the scene of this particular study. The paper then describes the methodology of the

* Julie B Green, Centre for Community Child Health and Ambulatory Paediatrics (University of Melbourne), Royal Children's Hospital, Flemington Road, Parkville, VIC 3053 AUS; (p) 61-3-9345-5356; (f) 61-3-9345-5900; E-mail: greenju@cryptic.rch.unimelb.edu.au

pilot project, which was aligned to the end users' needs and future directions, and details data sources and a range of data analysis. Finally, the outcomes of the project are highlighted and ongoing aspects of the research identified.

Overview of Australian Initiatives in GIS and Health

Over the last decade, interest has increased in initiatives to make the best possible information available to health and community service providers at the national, state, and local levels. To date in Australia, spatial health research has concentrated more on analyzing health care service needs for purchasing and planning, and less on patterns of disease distributions or what is more commonly known as geographic or environmental epidemiology. GIS is most commonly used across the Australian health sector within a social health framework. Because of this, using GIS has involved integrating data collections such as socioeconomic and specific health datasets of hospital admission rates, mortality, and birth events. Examples in Australia include the study of possible relationships between locational disadvantage and uptake of health services (1); emergency services dispatch developed by the Intergraph Corporation in the state of Victoria; drug research and harm reduction strategies (2); the South Australian Health Commission Social Health Atlas (3); and the National Social Health Database, known as HealthWIZ (4), which contains local-area health data on deaths, population characteristics, cancer registry details, social security, and Medicare, the Australian universal public health financing system.

Current reforms in Australia have redefined funding formulae for the health system. This has, in turn, caused a growing recognition of the importance of decision-making tools like GIS.

GIS for General Practice Project

This particular project is the result of a willingness to improve communication, information technology, and information management between the state-funded health services and the federally funded primary care infrastructure of family doctors (general practitioners, commonly known as "GPs"). Groups of GPs were brought together in recent years to form "divisions of general practice," a relatively new organizational structure designed to enable GPs to work together and to work within the wider health care system, to improve the quality of care, to meet local health needs, to promote preventive care, and to respond rapidly to community health needs. There are 118 divisions in Australia, with a median population of 152,920 per division (5). The Victorian state health department (known as the Department of Human Services) committed funds to develop and implement a prototype GIS as a tool for planning, education, and research in relation to the health needs and health status of the population groups within each division.

Project Team

The research team involved in this project includes the Centre for Community Child Health (University of Melbourne, Royal Children's Hospital, Melbourne), the Department of Geomatics, University of Melbourne, and the National Key Centre for

Social Applications of GIS, University of Adelaide. The project also received support from Land Victoria, Department of Natural Resources and Environment.

The Centre for Community Child Health plays a national role in child health education and research across the range of health professions, including general practice. The University of Melbourne's Department of Geomatics conducts education and research on a wide variety of GIS topics and plays an important role in the diffusion of GIS technology, assisted by Land Victoria, which deals with geospatial policy and geospatial datasets. The National Key Centre for Social Applications of GIS, as its name suggests, has expertise in the application of GIS technologies to social and community planning programs.

Instrumental partners in the project have been the two divisions of general practice for which the project was developed. Both of these divisions are in Victoria, a state in southeast Australia. One division is in northwest Melbourne, an inner metropolitan area of Victoria's capital city (Figure 1). The northwest Melbourne division has 234 of the 436 GPs known to be practicing in that geographic area (membership in a division is voluntary). The total population is 281,856 persons (6), giving a GP-to-person ratio of 1:646.

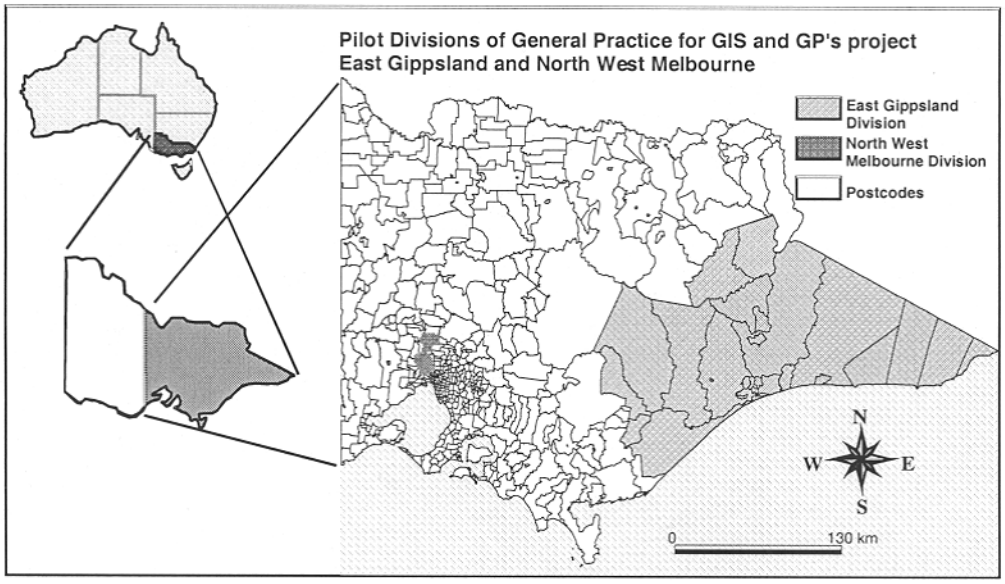


Figure 1 Pilot divisions of general practice, Victoria.

The second division is located in East Gippsland, a rural, coastal area in the southeast of Victoria that covers 12.5% of the state. The East Gippsland division of general practice is situated approximately 200 to 500 kilometers southeast of Melbourne. Most of East Gippsland's population lives in two major town centers (7). There is an average GP-to-person ratio of 1:1679 in this division; however, due to the seasonal nature of the population (East Gippsland is a popular seaside vacation area), the ratio can vary from 1:954 to 1:4753 (7).

General Practice Data and Outcome-Based Funding

Eighty-six percent of Australians visit their doctor at least once a year, giving GPs a principal role in the management of health concerns and, consequently, overall resource spending, including patterns of prescribing, uptake of preventive activities, uptake of other health services, the use of diagnostic imaging services, and referrals to specialists (5). General practice divisions have been identified as an organizational structure that will likely effect improvement of health outcomes. Divisions are required to identify key areas in which outcomes can be measured over time; with the growing recognition of the importance of decision support systems in measuring these outcomes, the setting of general practice divisions is an important one.

Funding is made available for GP members of divisions to become involved in cooperative activities. A proportion of a division's income, however, is tagged to its ability to demonstrate improvement in previously agreed-upon health outcomes for its population. These outcomes—and, therefore, the income—are information-dependent.

In terms of the role of routine data collection in the general practice sector, the collection of morbidity data or practice patterns is currently not at all systematic. In Australia, in contrast to the United Kingdom and the United States, there is an unfortunate lack of reliable morbidity data collected at a population level and inclusive of any useful geospatial variables such as address, postcode, or statistical local areas. Each practice chooses how or when to computerize its business, what data it collects, and how the information is used. To date, there has also been a lack of information on the outcomes of GPs' activities, which can partially be attributed to a lack of data collection systems and technological approaches to advancing information for a health outcome decision-making system. The development of this GIS sought to redress some of these past limitations.

Project Methodology

Briefly, the phases of the project consisted of determining information needs, collecting data, implementing the system, delivering it, and evaluating it. The methods and data sources for this project were closely aligned to the needs and future directions of the divisions in their provision of clinical and preventive general practice services to their communities.

Data Collection

The information needs of the two divisions were determined early, relative to the scope of the project and dependant on whether the data collections had a geospatial variable included within their data structure. Divisions identified the important areas of decision-making and these were linked to potential sources of available data. Some of these data are routinely collected by leading health agencies at national and state levels, but more local data needed to be gathered to provide a more complete picture.

The digital map base of Victoria was provided by Land Victoria. Demographic data (country of origin, age, sex, and income) came from the 1996 Population and Housing Census (6). Hospital admissions data were obtained from the state government's hospital inpatient database, known as the Victorian Inpatient Minimum Dataset (8). One of the most important priorities that divisions of general practice identified is immunization coverage of young children, so data on coverage rates came from a national population-based immunization register (9).

Redressing the Gaps in Data Availability

Data were attainable on the population within the divisions, but very little information was available on the general practices themselves. This gap in data availability was partially redressed through questionnaires administered to each practice location. These questionnaires collected information relating to types of data held by general practices (electronic or paper-based patient health records, availability of data summaries, knowledge of peak service times) and questions that helped build a picture of the size of the practice by number of staff and patients seen, other co-located services, and other relevant data such as the distances patients traveled to see their doctors.

Developing the System for Divisional Implementation

The increasing emergence and widespread uptake of communication technologies in Australia was considered in the preparation for presentation of the GIS prototype to the divisions. The team chose the Internet as the optimum medium for the delivery and placement of the product. For the GIS software itself, ArcExplorer (ESRI, Redlands, CA; http://www.esri.com) was chosen, because it can perform elementary queries and provide good quality display, desirable by the divisions. While this package does not have the full analytical capabilities of other GIS packages, the selection of user-friendly software was a high priority. Because the pilot divisions already have access to the Internet, software costs and the acquisition of additional hardware were eliminated. This project did involve posting confidential information. To address this concern (one not unique to working with the Internet), a password-protected Web site (http://www.sli.unimelb.edu.au/gdv/gdv_health.html) was used. The password system allows only the pilot divisions to access the confidential information.

Figure 2 shows the model of integration of databases in the GIS for GPs. All the databases have been integrated into the system through common GIS operations such as tabular linking and address geocoding. Common identifiers like postcode boundaries and divisional boundaries permit the integration of all data into the system.

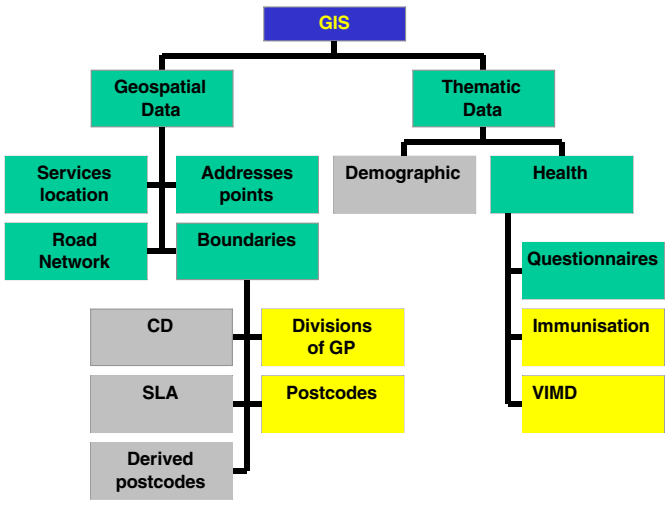


Figure 2 Model of integration of databases in the GIS for General Practice project.

Project Outcomes

Queries

The possibilities for making queries and analysis are many. An example of a two-stage query is as follows:

1. "Show me the postcode areas in the northwest Melbourne division where fewer than 80% of children between 15 and 20 months of age are fully immunized" (corresponding to the Australian Childhood Immunisation Schedule). The postcodes are both tabled and highlighted (Figure 3).
2. "What immunization providers are located in these postcodes?" This reveals all family doctors, community health centers, and maternal and child health nurses in the area who may be targeted for inclusion in immunization initiatives (Figure 4).

Ideally, all the datasets would have boundaries that articulated, making it simpler to integrate the data. Because this is not the case, a new query must be formulated for each of the themes. Spatial queries can be performed using the information tool (***i***) in each of the layers.

Training and Evaluation

The end stages of the project involved providing training and education to the two di-

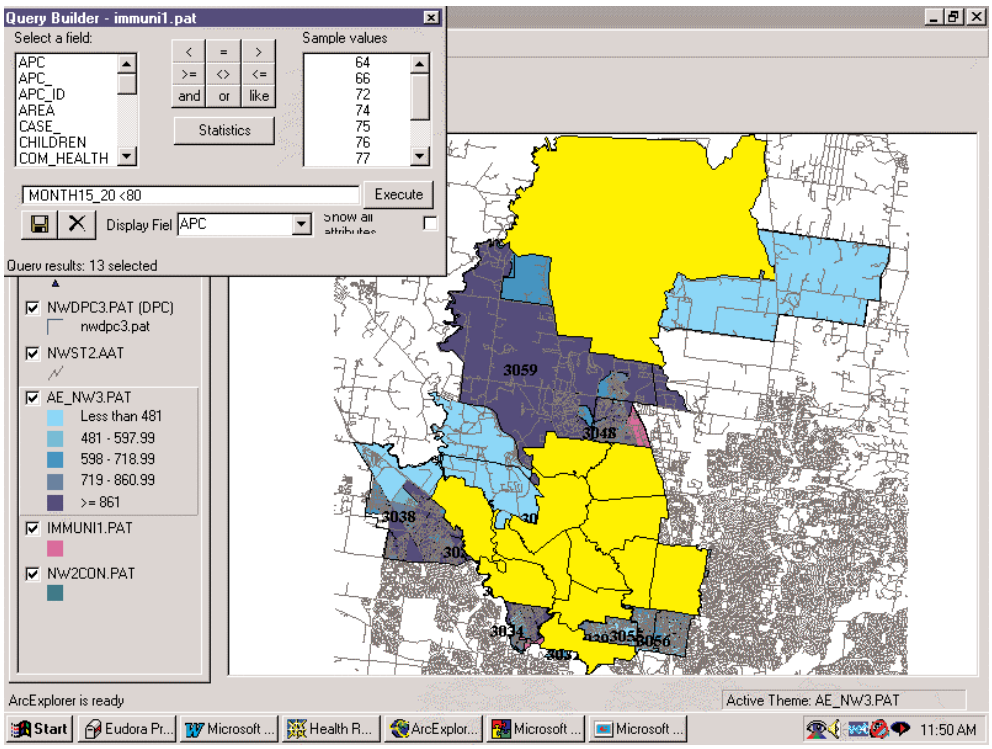


Figure 3 Postcode areas in northwest Melbourne division where fewer than 80% of children between 15 and 20 months of age are fully immunized.

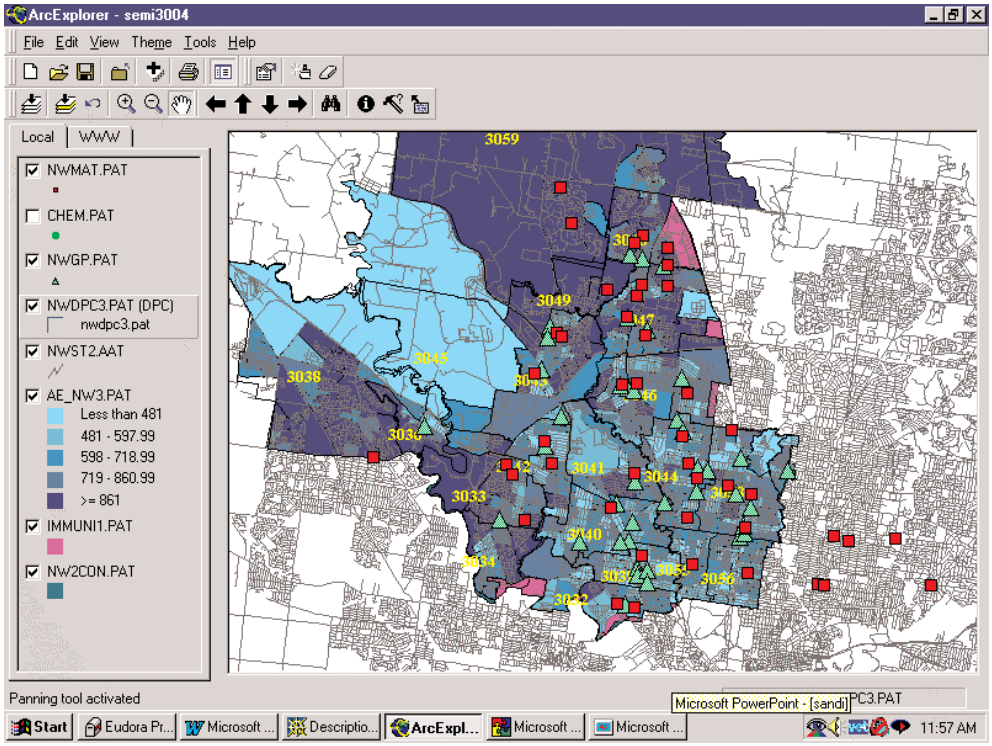


Figure 4 Immunization providers located in the postcodes with low immunization rates.

visions because divisional staff have no prior experience with GIS or geospatial data. The training provided some basic theory, and there were practical sessions on the use of spatial information systems, geospatial data, introduction to ArcExplorer, basic querying skills, and spatial data visualization. While it is important to consider that the technology is being adopted by the organization rather than individual users, consideration needs to be given to the variation in the ability and familiarity of divisional staff with information technology and whether this affects how they use the database. Divisional staff will use the GIS according to their job role within the organization. Program planners, administration support, and executive staff, for example, may make varying uses of the database.

For the 12 months after the GIS is implemented, an evaluation is planned of the ways in which the technology is adapted and reinvented to meet divisional needs. The timeline of the evaluation takes critical planning processes of the divisions into consideration and ensures that the evaluation occurs within the lifespan of the data in the GIS. Qualitative and quantitative data collection methods will be used, such as in-depth interviews with staff and a systematic review of documents and administrative records that incorporate data or use the GIS' capabilities. Of particular interest will be determining what new things the GIS enables the divisions to do, as well as how it helps them perform activities in which they were already involved.

Because the adoption of GIS in the primary health sector is a relatively recent phenomenon, there are relatively few examples of adoption of the technology and, conse-

quently, a limited amount of experience to support claims that GIS implementation means improved information processing and more informed decision-making. Results of the evaluation will be used to inform the development of future information-based decision support technology within the divisions-of-general-practice environment.

Conclusions

The GIS for General Practice study was undertaken to develop a methodology for the provision of a GIS to a particular group of providers of primary health care, and has achieved its aims. The research has highlighted a number of constraints in the development of a GIS for the health sector, the major challenge being the variety of geographic classifications that have been used for health data over the last decade (including numerous versions of regional and subregional classifications used by national and state authorities). There are also widespread differences in data collection methods, data quality, and data access.

A relatively underdeveloped technological infrastructure within general practice in Victoria minimizes access to the Internet and information systems in general, though current trends in the acquisition of computers will go some way to overcoming this constraint.

The Victorian government's policy to make available geospatial data to all Victorians (10) details its intention to face the information age in the 21st century. Other initiatives in the state of Victoria this year include the state Department of Human Services' commitment to drawing up a GIS for Health Strategy to support spatial information systems and improvement of decision-making by health planners (11), and the development of an Australian Research Council-funded project to develop a GIS for Health Research Strategy as well.

Until recently, GIS in health has depended on quantification methods of monitoring and measuring the population. Statistical surveys, epidemiological assessments, evaluations, and health outcomes are currently a central influence on policy, planning, and resource allocation. If there is a desire to study the geography of health rather than the geography of disease (12), consideration needs also to be given to ways in which qualitative health data—which include lay perceptions of health and illness and the "lived," or socially experienced, dimension of health (13)—can be incorporated into a GIS framework.

The research team that developed the GIS for General Practice prototype is not the end user. The end users, the divisional teams, have not previously had experience using information systems to help them make decisions. They are also under a great deal of pressure to change how they make decisions, and change what techniques they use in their decision-making. The final evaluation of this project will be testimony to the ultimate success of the GIS for General Practice product, but the outcomes of the actual creation process are already tangible. It is hoped that these initiatives will further the use of GIS technology in the health sector in Australian states and territories.

Acknowledgments

The authors would like to thank the Department of Human Services, Victoria, the National Key Centre for Social Applications of GIS (University of Adelaide), and Land Victoria for their collaboration with this project.

References

1. Hugo G. 1995. *Locational disadvantage: Development of a data model to support government decision making.* Prepared for the Social Justice Research Program into Locational Disadvantage and the Office of Geographic Data Coordination, Department of Premier and Cabinet, Victoria. March.
2. Rumbold G. 1998. *Two examples of the application of geographic information systems in drug research.* Presented at the Second Symposium on GIS and Health, Developments in the Application of Geographic Information Systems within the Health Sector. Sponsored by the University of Melbourne and Land Victoria. June 10. http://www.sli.unimelb.edu.au/HealthGIS98/presentations.
3. Glover J. 1998. *GIS for health in Australia without the building blocks; Are we missing the point?* Presented at the Second Symposium on GIS and Health, Developments in the Application of Geographic Information Systems within the Health Sector. Sponsored by the University of Melbourne and Land Victoria. June 10. http://www.sli.unimelb.edu.au/HealthGIS98/.
4. Prometheus Information Proprietary Limited. 1998. *HealthWIZ: Divisions of General Practice.* Prometheus Information Pty., Ltd., Dickson, Australian Capital Territory. http://www.prometheus.com.au.
5. Commonwealth Department of Health and Family Services. 1996. *General practice in Australia: 1996.* Australia: General Practice Branch.
6. Australian Bureau of Statistics. 1996. *1996 census of population and housing.* Belconnen, Australian Capital Territory: Australian Bureau of Statistics. http://www.abs.gov.au.
7. Hind J, Hind J. 1998. The health status of people in the East Gippsland region: A report to the East Gippsland Division of General Practice. Bairnsdale, Victoria: East Gippsland Division of General Practice. May.
8. Victorian Department of Human Services. 1997. Victorian inpatient minimum dataset. Victoria: Victorian Department of Human Services.
9. Health Insurance Commission. 1997. *Australian childhood immunisation register.* Australian Capital Territory: Health Insurance Commission. http://www1.hic.gov.au/general/acircirghome.
10. Land Victoria. 1997. *Victoria's geospatial information strategic plan: Building the foundations, 1997–2000.* Land Victoria, Department of Natural Resources and Environment, State Government of Victoria. October. http://www.giconnections.vic.gov.au/content/stratgie/stratgie.htm.
11. Catford J. 1998. *Geographic information systems and health.* Presented at the Second Symposium on GIS and Health, Developments in the Application of Geographic Information Systems within the Health Sector. Sponsored by the University of Melbourne and Land Victoria. June 10. http://www.sli.unimelb.edu.au/HealthGIS98/presentations.
12. Gattrell AC, Loytonnen M. 1996. *GIS and health research in Europe: A position paper.* Prepared for GISDATA Specialist Meeting, Helsinski workshop. January.
13. Janes C, Stall R, Gifford S. 1986. *Anthropology and epidemiology: Interdisciplinary approaches to the study of health and disease (culture, illness, and healing).* Dordrecht. The Netherlands: D. Reidel Publishing Company.

The Knox Method and Other Tests for Space-Time Interaction

Martin Kulldorff (1),* Ulf Hjalmars (2)
(1) Division of Biostatistics, Department of Community Medicine and Health Care, University of Connecticut School of Medicine, Farmington, CT; (2) Department of Pediatrics, Östersund Hospital, Östersund, Sweden

Abstract

The Knox method, like other tests for space-time interaction, is biased in situations in which there are geographical population shifts; that is, when there are different percentages of population growth in different regions. In this paper, the size of the population shift bias is investigated for the Knox test, and it is shown that it can be a considerable problem. This paper then presents a Monte Carlo method for constructing unbiased space-time interaction tests, illustrating the method for the Knox test and for a combined Knox test. Practical implications are discussed in terms of the interpretation of past results and the design of future studies.

Keywords: bias, Jacquez' test, Knox test, Mantel's test, population shifts

Introduction

Space-time interaction tests are used to evaluate whether there is space-time clustering of events after purely spatial and purely temporal clustering are adjusted for. These tests are frequently applied in epidemiological studies, in which it is of interest to know whether cases of some disease are more clustered than would be expected based on the underlying geographical population distribution and on any purely temporal trend. Two excellent surveys on space-time interaction tests have been written by Mantel (1) and Williams (2). Comparative evaluations and power studies have been done by Chen, Mantel, and Klingberg (3) and by Jacquez (4).

The most widely used statistical technique for testing space-time interaction was developed by Knox (5). In the Knox test, the time and geographical location of each case are noted, and the distance between each possible pair of cases is calculated in terms of both time and space. If many of the cases that are "close" in time are also "close" in space ("close" is defined by the user), or vice versa, then there is space-time interaction. This could be an indication that a disease is infectious or that it is caused by some other type of agent that appears locally at specific times, such as food poisoning.

In a survey of epidemiological articles published between 1960 and 1990, Daniel Wartenberg and Michael Greenberg (6) found 59 different studies that used the Knox method. Many of these were concerned with leukemia, and the results from such studies have been used as evidence supporting a viral etiology of the disease (7,8).

The Knox test is an elegant and, in many ways, attractive method. For example, it

* Martin Kulldorff, University of Connecticut School of Medicine, 263 Farmington Avenue, Farmington, CT 06030-6325 USA; (p) 860-679-5473; (f) 860-679-5464; E-mail: martink@cortex.uchc.edu. Note: this work was conducted while this author was at the Biometry Branch, Division of Cancer Protection, of the National Cancer Institute, Bethesda, MD.

is simple and straightforward to calculate the test statistic, and using the test requires knowledge only of cases, not controls. There is, however, a well-known problem with the method.

Mantel (1) pointed out that the Knox test is biased if the rate of population growth is not constant for all geographic sub-areas. We call this the *population shift bias.* Shifts in the population distribution create space-time interaction among any random sample of individuals, including sets of cases generated under the null hypothesis of equal disease risk. The Knox statistic is constructed so as to pick up any type of space-time interaction; it does not distinguish whether that interaction is due to shifting population distributions or to some disease-related phenomenon. This is not a flaw of the test per se, and is not a problem if one is looking for any type of space-time interaction. However, interest is typically focused—as in epidemiology—on disease-related phenomena, not shifts in population distribution, so the latter should be adjusted for.

While the existence of the population shift bias has long been known, the magnitude of the bias has not been studied for any real datasets, and the bias has typically been ignored in practical applications. In the "Estimation of the Population Shift Bias" section, the bias of the ordinary Knox test is estimated for two different datasets: the child population in Sweden from 1976 to 1994 (a fairly stable population) and the total population in New Mexico (where there have been large population shifts) from 1973 to 1991. The estimations show that the bias is considerable for some cases.

Klauber and Mustacchi (9) suggested that the population shift bias could be reduced by dividing the data into several parts corresponding to different time periods. Within these parts, the population would be more stable. A test statistic would then be calculated separately for each part, and the statistics would be summed to get an overall test. This method reduces the bias but does not eliminate it. Unfortunately, it also decreases the power of the test; pairs of cases falling in different data parts would not be used, leading to loss of information.

A simple unbiased version of the Knox test is presented in the section entitled "An Unbiased Knox Test." This test adjusts not only for purely spatial and purely temporal variations, but also for the space-time interaction inherent in the background population. It does so without the loss of power associated with the Klauber-Mustacchi approach. Its one drawback is that it requires knowledge of the underlying population distribution.

While this paper is focused on the Knox method, which is the most commonly used space-time interaction test, other space-time interaction tests suffer from the same population shift bias. This includes the methods proposed by David and Barton (10), Mantel (1), Pike and Smith (11), Diggle et al. (12), Jacquez (4), and Baker (13). This paper's approach for constructing an unbiased Knox test can also be used to construct unbiased versions of these other methods.

A second issue with the Knox method relates to the choice of critical distances to define which pairs of cases are close in space and time respectively. Unless the investigator has a fairly clear idea of the scale at which clustering may occur, this is a problem. Separate tests are often performed for a number of different critical distances (e.g., Gilman and Knox, 1995 [14]). It is possible to do a Bonferroni-type adjustment for the multiple testing inherent in such a procedure, but because the test statistics calculated for adjacent critical distances are highly correlated, there is loss of power when using such a method. In practice it is seldom used. Baker (13) has presented a combined Knox

test, providing a single hypothesis test with multiple critical distances. The approach presented in the section entitled "An Unbiased Combined Knox Test" uses the same basic idea to deal with multiple testing.

If the simple modification to the Knox test described here were implemented in actual studies, the value of those studies would greatly increase. There would no longer be any uncertainty about whether a significant result is due simply to shifts in the geographical population distribution, and there would be no issue of multiple testing. The Knox test is an intuitive, elegant method. With its major weaknesses resolved, we hope, it will continue to be used for years to come.

The Knox Test

Let n be the total number of cases, so that there are $N=n(n-1)/2$ distinct pairs of cases. Let N_t be the number of case pairs that are closer to each other in time, compared to some specified temporal distance. Likewise, let N_s be the number of pairs close in space as defined by some geographic distance. Finally, let X be the number of case pairs that are close both in time and space.

The observed value of X is the test statistic of the Knox method (5). To adjust for purely spatial and purely temporal inhomogeneities in the data, the test statistic is evaluated conditionally on N_t and N_s. Under the null hypothesis of no space-time interaction, the expected value of X is $E[X \mid N_t,N_s]=N_tN_s/N$ (15).

Knox (5) conjectured that X is approximately Poisson-distributed. Barton and David (15) showed this to be true when N_t and N_s are small compared to N, in the sense that the variance of X is then approximately equal to its expected value. More importantly, by application of graph theory and by also conditioning on the second-order terms, they obtained an exact formula for the variance:

$$V[X|N_t, N_s, N_{2s}, N_{2t}] =$$

$$\frac{N_sN_t}{N} + \frac{4N_{2s}N_{2t}}{n(n-1)(n-2)} + \frac{4[N_s(N_s-1)-N_{2s}][N_t(N_t-1)-N_{2t}]}{n(n-1)(n-2)(n-3)} - \left(\frac{N_sN_t}{N}\right)^2$$

where N_{2s} is the number of pairs of case pairs close in space that have one case in common, and where N_{2t} is defined equivalently for time.

In practical applications, different approximations of the test statistic's distribution have been used. The Cluster software package, written by Aldrich and Drane (16), uses the Poisson approximation, as originally proposed by Knox (5). Gilman and Knox (14) and many others have done likewise, except that they have used the normal approximation for the Poisson distribution, keeping the variance equal to the mean. We will call this approach the *Poisson-based approximation*. An alternative approach is to use a normal approximation with the mean and variance given by Barton and David (15). We will call this the *Barton-David-based approximation*. Yet another option, originally proposed by Mantel (1), is to use Monte Carlo hypothesis testing (17) by permuting the times among the fixed spatial locations. This is implemented as part of the Stat! software package (18); Petridou et al. (8) provide one example of its use.

Before estimating the population shift bias, as in the next section, it is important to look at any potential bias due to the distributional assumptions of the Knox test

statistic. Table 1 contains bias estimates for the Poisson- and Barton-David-based approximations when the Knox test is applied to a hypothetical child population in Sweden. For all years from 1976 to 1994, the population is artificially fixed at the 1982 level so that there are no population shifts. The data are aggregated into 2,507 parishes. The parish and month were randomly selected for each of 1,000 and 10,000 cases in proportion to the 1982 population for each parish, and in proportion to the length of each month.

When N_t and N_s are small compared to N, the Poisson-based approximation works well. When N_t and N_s are larger, though, there is some bias. This is as expected based on the theoretical results of Barton and David (15). The Barton-David-based approximation, on the other hand, works well across the board for the Swedish data. This is important to remember when estimating the population shift bias, as in the next section. By definition, the Monte Carlo procedure provides an unbiased test when there are no shifts in the population distribution.

Estimation of the Population Shift Bias

Differential population growth can be caused by internal migration between different regions, by geographically differential emigration or immigration rates, or by geographically differential birth or death rates. If the disease risk is related to age, the bias can also be caused by different age structures in different regions, whether that structure changes over time or not; as the population ages, the age-specific population counts change over time to different degrees in different regions.

The magnitude of the population shift bias of any test for space-time interaction depends on the specific geographic area and time period under study. In general, shorter overall time periods result in less bias because there is less time for population shifts to occur, as pointed out by Klauber and Mustacchi (9). Nothing general can be said about specific geographic areas. To give some idea of the extent of the bias, we have calculated the population shift bias of the ordinary Knox test for two different datasets.

The first dataset is the child population in Sweden from 1976 to 1994, aggregated to the 2,507 parishes. The second dataset is the total New Mexico population from 1973 to 1991, aggregated to 32 counties. (The second dataset is available at http://dcp.nci.nih.gov/BB/datasets.html.) For the New Mexico dataset, Cibola and Valencia are counted as one county for the whole time period even though they became two different counties in 1981. The geographic distance between cases is the distance between the parish/county centroids to which they belong. When the critical geographic distance is 0, only those cases located in the same parish are considered spatial neighbors. For both datasets, the case times are noted in months. When the critical temporal distance is zero months, neighboring cases are only those occurring in the same calendar month; when it is three months, neighboring cases are those occurring in months at most three calendar months apart (e.g., January and April, but not January and May); and so on.

To put these datasets in a proper context, the population growth for various subregions is provided in Table 2. For the child population in Sweden, Table 2 shows the population growth in each of the country's 24 counties, or läns. Table 2 shows only part of the picture, though; the data were analyzed at the much finer level of 2,507 parishes. The percentage of change, naturally, varies more for the smaller parishes. The 470

Table 1 Estimated True Significance Levels for the Ordinary Knox Test When Applied to the Childhood Population in Sweden

		Poisson Approximation				Barton-David Approximation			
		$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.01$	
cut-off points		# cases		# cases		# cases		# cases	
km	months	1000	10000	1000	10000	1000	10000	1000	10000
0	0	.064	.072	.021	.018	.064	.073	.021	.019
0	3	.052	.052	.013	.010	.054	.056	.014	.011
0	6	.051	.050	.010	.006	.055	.056	.012	.009
0	12	.044	.046	.009	.008	.052	.053	.012	.010
0	24	.034	.044	.006	.007	.050	.051	.011	.009
2	0	.064	.075	.020	.018	.064	.076	.020	.018
2	3	.053	.053	.013	.010	.055	.054	.014	.011
2	6	.048	.050	.011	.007	.053	.055	.013	.008
2	12	.043	.044	.009	.008	.052	.050	.012	.010
2	24	.035	.041	.006	.007	.050	.050	.012	.010
5	0	.062	.059	.020	.012	.063	.060	.020	.012
5	3	.052	.056	.012	.009	.054	.058	.014	.010
5	6	.047	.051	.011	.010	.052	.055	.013	.012
5	12	.047	.052	.009	.011	.055	.053	.013	.011
5	24	.040	.062	.007	.014	.054	.050	.013	.011
10	0	.055	.054	.015	.013	.056	.056	.015	.013
10	3	.049	.058	.012	.012	.052	.058	.012	.012
10	6	.047	.056	.010	.010	.053	.053	.011	.010
10	12	.049	.067	.010	.018	.056	.051	.013	.011
10	24	.046	.118	.011	.050	.053	.050	.014	.012
20	0	.052	.054	.013	.010	.055	.056	.014	.011
20	3	.049	.057	.011	.011	.053	.056	.013	.010
20	6	.049	.069	.010	.017	.055	.056	.012	.010
20	12	.054	.109	.015	.039	.059	.056	.016	.015
20	24	.064	.196	.019	.120	.058	.055	.017	.012
50	0	.049	.051	.010	.010	.053	.056	.012	.012
50	3	.047	.056	.011	.011	.054	.057	.014	.012
50	6	.046	.070	.011	.019	.052	.055	.014	.013
50	12	.049	.116	.013	.049	.053	.054	.015	.012
50	24	.067	.221	.022	.139	.058	.051	.017	.011

Note: Cases were randomly generated according to the 1982 population, so there is no population shift bias. The bias due to the Poisson and Barton-David approximations for the distribution of the test statistics is the difference between the numbers reported and the nominal significance level.

Table 2 Population Changes in the 24 Läns of Sweden and in the 32 Counties of New Mexico

Sweden				New Mexico			
	child population				total population		
län	1976	1994	change	county	1973	1991	change
Stockholm	319901	335044	+5%	Bernalillo	353813	490248	+39%
Uppsala	54445	60829	+12%	Catron	2372	2507	+6%
Södermanland	57765	52666	-9%	Chaves	45204	58699	+30%
Östergötland	86058	83198	-3%	Colfax	12577	12743	+1%
Jönköping	68603	64988	-5%	Curry	42709	44613	+4%
Kronoberg	38228	36530	-4%	DeBaca	2509	2310	-8%
Kalmar	51718	48694	-6%	Doña Ana	76915	140696	+83%
Gotland	12389	12426	0%	Eddy	40940	49998	+22%
Blekinge	34501	29280	-15%	Grant	23549	27986	+19%
Kristianstad	60617	59318	-2%	Guadalupe	4889	4102	-16%
Malmöhus	157924	155186	-2%	Harding	1234	987	-20%
Halland	52084	56385	+8%	Hidalgo	5108	5937	+16%
Göteborg-Bohus	149254	146911	-2%	Lea	48907	55584	+14%
Älvsborg	96494	94708	-2%	Lincoln	8395	12824	+53%
Skaraborg	60117	59042	-2%	Los Alamos	15315	17908	+17%
Värmland	58583	55198	-6%	Luna	13493	18984	+41%
Örebro	58775	54432	-7%	McKinley	46826	62746	+34%
Västmanland	61058	52042	-15%	Mora	4712	4208	-11%
Kopparberg	59303	58938	-1%	Otero	42303	52256	+24%
Gävleborg	61696	55583	-10%	Quay	10980	10564	-4%
Västernorrland	56258	49316	-12%	Rio Arriba	27339	34330	+26%
Jämtland	27216	26881	-1%	Roosevelt	16477	17258	+5%
Västerbotten	52376	54642	+4%	Sandoval	23858	65975	+177%
Norrbotten	63038	53520	-15%	San Juan	58718	94028	+60%
				San Miguel	23452	26074	+11%
				Santa Fe	61250	101675	+66%
				Sierra	7976	10098	+27%
				Socorro	10492	14696	+40%
				Taos	18053	23679	+31%
				Torrance	5730	10658	+86%
				Union	5060	4136	-18%
				Valencia	43192	70135	+62%
Total	1798401	1755757	-2%	Total	1104347	1548642	+40%

Note: For the Swedish data, the actual analysis was done using much less aggregated data.

parishes with more than 1,000 children in 1976 had an average population decrease of 2.5% from 1976 to 1994, with a standard deviation of 30.4 percentage points. The equivalent standard deviations for other subgroups were 33.3 for 275 parishes with 1976 populations in the 500–1000 range, 31.4 for 486 parishes in the 200–500 range, 72.0 for 467 parishes in the 100–200 range, and 122.8 for 809 parishes with 1976 populations of less than 100. The population growth in New Mexico is also presented in Table 2. Between

1973 and 1991, one county's population doubled while many other counties had a fairly constant population.

To estimate the population shift bias, cases were randomly assigned to a parish (or county, for New Mexico) and to a particular month with probability proportional to the actual population in that parish during that month. In this way, the cases were randomized with population shifts taken into account. The population for a particular month was obtained through linear interpolation, using yearly population data for New Mexico and the years 1976, 1982, 1988, and 1994 for Sweden. Separate calculations were done for 1,000, 4,000, and 10,000 randomized cases. For each random Monte Carlo replication of the fixed number of cases, the test statistic was calculated and compared with its nominal critical region using the Barton-David distributional approximation. Without bias, 5% of the test statistics from the Monte Carlo replications should fall within the critical region. The actual numbers are given in Tables 3 and 4.

The population shift bias for the Swedish data is the difference between the numbers reported in Table 3 and those reported in Table 1 for the Barton-David approximation. The total bias is the difference between Table 3 and the nominal significance levels. For the Swedish data, there is very little bias using the original Knox test when the total number of cases observed is 1,000. With more cases, the bias increases. It is a considerable problem with 10,000 cases observed.

For New Mexico, the bias is considerable for 1,000, 4,000, or 10,000 cases, as can be seen from Table 4. Note that it is not the total population increase of 40% that causes the bias. If the increase were the same in all counties, the population shift bias would be zero.

The bias estimates in Tables 1, 3, and 4 were calculated using 20,000 random replications of the fixed number of cases. The 95% confidence intervals are ±0.007 when the estimate is around 0.50, and ±0.003 when the estimate is around 0.05. If the Poisson approximation is used instead of the Barton-David approximation, the total bias is about the same or higher (not shown), as would be expected considering Table 1.

As Tables 3 and 4 show, the population shift bias increases with an increased number of total cases observed. Why? By definition, the population shift bias is the probability that a method will detect space-time interaction due to the population shift when there is no space-time interaction of any other kind. That is, a method's population shift bias is identical to its power to detect a population shift using a random sample from the population. The larger the random sample, the greater the power; by consequence, the more cases, the bigger the population shift bias. In a sense, this is a Catch-22 situation. We could reduce the population shift bias by analyzing a smaller number of cases, but that would also reduce the power to detect space-time interaction due to any biological phenomena of interest.

The population shift bias also varies with the choice of critical geographical distance. Such differences are data-dependent. Consider a situation in which the child population over time is identical in several cities, but in which, within those cities, there is a continuous child population shift. New suburbs have many small children who grow older together with the suburbs until they move out and leave a predominantly adult population behind. This will lead to a population shift bias for small values of the critical geographic distance, but not necessarily for large ones. On the other hand, increased critical distances will result in more space-time case pairs, increasing the power

Table 3 Estimated True Significance Levels for the Ordinary Knox Test Using the Barton-David Approximation, When the Nominal Levels Are α=0.05 and α=0.01, for the Childhood Population in Sweden, 1976–1994

		$\alpha = 0.05$			$\alpha = 0.01$		
cut-off points		number of cases			number of cases		
km	months	1000	4000	10000	1000	4000	10000
0	0	.069	.074	.105	.021	.019	.032
0	3	.062	.079	.141	.016	.020	.040
0	6	.063	.087	.181	.018	.021	.064
0	12	.062	.102	.246	.018	.026	.101
0	24	.066	.126	.299	.018	.033	.132
2	0	.066	.072	.105	.021	.018	.034
2	3	.063	.086	.155	.017	.022	.041
2	6	.063	.091	.198	.017	.026	.071
2	12	.064	.105	.267	.017	.029	.105
2	24	.068	.138	.331	.018	.039	.146
5	0	.065	.070	.102	.018	.018	.029
5	3	.059	.087	.164	.016	.022	.050
5	6	.059	.100	.214	.016	.030	.077
5	12	.063	.114	.240	.016	.036	.097
5	24	.065	.116	.231	.017	.036	.090
10	0	.056	.061	.088	.015	.015	.020
10	3	.057	.081	.134	.014	.021	.043
10	6	.058	.092	.166	.015	.025	.058
10	12	.060	.098	.158	.017	.031	.059
10	24	.061	.092	.108	.017	.032	.037
20	0	.054	.061	.081	.013	.016	.019
20	3	.058	.071	.100	.015	.020	.030
20	6	.058	.082	.108	.016	.024	.032
20	12	.060	.075	.087	.018	.023	.024
20	24	.059	.066	.054	.016	.020	.014
50	0	.051	.055	.059	.013	.013	.014
50	3	.058	.062	.080	.014	.016	.022
50	6	.056	.068	.082	.015	.023	.024
50	12	.061	.064	.063	.016	.019	.019
50	24	.057	.058	.050	.017	.013	.015

Note: The difference between these numbers and those in Table 1 is the population shift bias.

Table 4 Estimated True Significance Levels for the Ordinary Knox Test Using the Barton-David Approximation, When the Nominal Levels Are α=0.05 and α=0.01, for the Population in New Mexico, 1973–1991

		$\alpha = 0.05$			$\alpha = 0.01$		
cut-off points		number of cases			number of cases		
km	months	1000	4000	10000	1000	4000	10000
0	0	.064	.085	.101	.016	.023	.027
0	3	.071	.098	.148	.020	.028	.048
0	6	.076	.105	.155	.022	.034	.049
0	12	.077	.110	.151	.023	.036	.051
0	24	.078	.104	.134	.025	.033	.041
50	0	.069	.127	.240	.019	.039	.088
50	3	.091	.212	.450	.030	.081	.232
50	6	.109	.239	.469	.036	.100	.250
50	12	.114	.241	.444	.042	.103	.227
50	24	.115	.208	.353	.041	.084	.159
100	0	.072	.157	.370	.019	.050	.165
100	3	.108	.313	.659	.034	.142	.427
100	6	.130	.351	.678	.044	.166	.445
100	12	.146	.351	.638	.055	.168	.399
100	24	.141	.295	.508	.054	.128	.278
200	0	.067	.116	.226	.017	.033	.081
200	3	.088	.200	.412	.026	.077	.207
200	6	.101	.222	.421	.034	.091	.212
200	12	.111	.217	.380	.041	.090	.183
200	24	.109	.177	.276	.036	.069	.115

to detect a population shift and thus increasing the population shift bias. Hence, different phenomena may work in opposite directions.

The population shift bias also depends on the level of aggregation. If there are very local population shifts, then it is possible to reduce the bias of the ordinary Knox test by combining areas in which the shifts are in opposite directions from the overall average population growth. Taking this one step further, it is worth pointing out that one way to construct an unbiased Knox test is to aggregate data in such a way that each aggregated area has the same population growth curve. In practice, though, this is hard to accomplish because populations aggregated into the same area must be very close to each other for the test to be meaningful. A better way to obtain an unbiased test is proposed in the following section.

An Unbiased Knox Test

To obtain an unbiased version of the Knox test, it is necessary to know the background population and its temporal trends. Using such data, one can obtain random replications of cases generated under the null hypothesis. These replications can then be used for hypothesis testing using the Monte Carlo procedure. Randomizing in proportion to the population size at each time and place adjusts for the population shifts.

In creating an unbiased Knox test, one must be careful as to how to implement the Monte Carlo method. For example, the Monte Carlo approach suggested by Mantel (1) does not work for this purpose. This is because Mantel proposes to randomize cases using random permutations of spatial and temporal observations conditioned on the set of spatial and set of temporal values, rather than randomizing completely new cases from the background population. The former is the preferred way to do the test when there are no population shifts—when it is not necessary to make distributional approximations—but it does not eliminate the population shift bias.

Neither does it work to simply calculate the Knox test statistic X and, in the normal Monte Carlo fashion, compare its values in the real and randomized datasets. Doing so would give a valid unbiased test, but the value of the test statistic would be high due to purely spatial clustering, purely temporal clustering, or temporal trends. Hence, it would no longer be a test for space-time interaction, but instead a test for global space-time clustering, as discussed by Kulldorff (19).

A way to eliminate the population shift bias and at the same time retain the space-time interaction test is as follows:

1. Generate random datasets for which each random replication has the same number of cases as the real data. The location and time of each case should be random, with probability proportional to the population size for that location and time or to the expected number of cases under the null hypothesis, adjusted for potential confounders such as age.
2. Calculate the test statistic X for the real and random datasets.
3. For each dataset, normalize X using the Barton-David-based approximation:

$$N(X) = \frac{X - E[X|N_t, N_s]}{V[X|N_t, N_s, N_{2s}, N_{2t}]}$$

   This is necessary because N_t, N_s, N_{2s}, and N_{2t} change in each simulated dataset.
4. Rank $N(X)$ for the real and random datasets. If the former is among the 5% highest, reject the null hypothesis of no space-time interaction at the 5% significance level. The corresponding simulated p-value is $R/(REP+1)$, where R is the rank of $N(X)$ from the real dataset and REP is the number of Monte Carlo replications.

For the third step we chose to use the Barton-David-based approximation. Using the Poisson-based approximation will also give an unbiased test. Because only the relative rank is of interest, the accuracy of the approximation is unimportant as long as the ranking it creates is unchanged. The Monte Carlo option for approximating the distribution of X is less practical, as choosing it would mean running one Monte Carlo simulation embedded within another, quite a time-consuming task even for a computer.

Table 5 shows the application of the unbiased Knox test to lung cancer in New Mexico from 1973 to 1991. These data were collected by the New Mexico Tumor Registry for the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program. Table 6 presents the unbiased Knox test as applied to all types of childhood leukemia in Sweden from 1973 to 1994. In both cases, 19,999 Monte Carlo replications were performed. The resulting p-values are given for a range of spatial and temporal critical distances. For comparison, the nominal but biased p-values using the Poisson and Barton-David approximations are given in parentheses.

For the lung cancer data, the unbiased Knox test gives no evidence of any space-time interaction. In contrast, when the population shift bias is not adjusted for, some of the p-values are very small, giving a false impression of space-time interaction. For the leukemia data, out of 30 tests for different critical distances, 8 are significant at the 0.05 level when the unbiased Knox test is used. This may indicate some level of space-time interaction, but it is hard to judge because there is considerable multiple testing involved. This is discussed in the next section.

Table 5 Unbiased p-values When the Knox Test Is Applied to 9,254 Cases of Lung Cancer in New Mexico, 1973–1991, Using Different Critical Distances

	months				
	0	3	6	12	24
0 km	**.361**	**.796**	**.910**	**.888**	**.895**
	(.092/.125)	(.203/.345)	(.540/.512)	(.452/.488)	(.749/.552)
50 km	**.187**	**.396**	**.551**	**.555**	**.696**
	(.005/.012)	(<.0001/.010)	(<.0001/.024)	(<.0001/.030)	(<.0001/.117)
100 km	**.482**	**.423**	**.366**	**.418**	**.544**
	(.016/.027)	(<.0001/.001)	(<.0001/.0004)	(<.0001/.001)	(<.0001/.014)
200 km	**.228**	**.174**	**.138**	**.196**	**.231**
	(.032/.026)	(<.0001/.002)	(<.0001/.0007)	(<.0001/.003)	(<.0001/.012)

Note: In parentheses are the biased p-values from the ordinary Knox test using the Poisson and Barton-David approximations (Poisson/Barton-David). Adjusting for the multiple testing, the unbiased combined p-value is .472.

An Unbiased Combined Knox Test

When the ordinary Knox test is applied, a key feature is the choice of critical distances. Because the scale at which clustering may exist is often unknown, the test has often been applied for a whole range of possible values (e.g., Gilman and Knox, 1995 [14]). This is valuable for estimating the scale of clustering, but it also introduces multiple testing, and if the test is significant for some critical distances but not for others, as in Table 6, then the result is hard to interpret. One solution is to apply some Bonferroni-type adjustment, but because the different tests for different critical distances are statistically dependent, such a procedure is overly conservative and is not commonly used. Using the same basic idea as Baker (13), one can obtain an unbiased combined Knox test as follows.

Table 6 Unbiased p-values When the Knox Test Is Applied to 1,592 Cases of Childhood Leukemia in Sweden, 1973–1994, Using Different Critical Distances

	months				
	0	3	6	12	24
0 km	**.077**	**.135**	**.295**	**.282**	**.074**
	(.061/.061)	(.119/.116)	(.255/.249)	(.246/.236)	(.055/.040)
2 km	**.113**	**.125**	**.238**	**.188**	**.040**
	(.097/.097)	(.109/.106)	(.200/.194)	(.157/.146)	(.028/.018)
5 km	**.224**	**.028**	**.077**	**.423**	**.369**
	(.210/.210)	(.020/.019)	(.058/.054)	(.359/.353)	(.292/.277)
10 km	**.611**	**.049**	**.134**	**.293**	**.398**
	(.609/.609)	(.038/.036)	(.108/.105)	(.250/.247)	(.341/.340)
20 km	**.851**	**.029**	**.028**	**.020**	**.143**
	(.840/.842)	(.020/.020)	(.018/.018)	(.009/.011)	(.100/.121)
50 km	**.362**	**.091**	**.054**	**.033**	**.023**
	(.360/.357)	(.084/.080)	(.042/.041)	(.019/.022)	(.007/.015)

Note: In parentheses are the biased p-values from the ordinary Knox test using the Poisson and Barton-David approximations (Poisson/Barton-David). Adjusting for the multiple testing, the unbiased combined p-value is .237.

1. For the real and random datasets, calculate the test statistic X_d for each of several combinations of critical distances.
2. For each choice of critical distances, calculate the normalized test statistic $N(X_d)$ as described in "An Unbiased Knox Test."
3. For each dataset, select the maximum value of $N(X_d)$ taken over all sets of critical distances, $M=\max_d N(X_d)$.
4. Rank the maximum values M coming from the real and random datasets. If the former is among the 5% highest, reject the null hypothesis of no space-time interaction at the 5% significance level. The corresponding simulated p-value is as before—$R/(REP+1)$, where R is the rank of M from the real dataset and REP is the number of Monte Carlo replications.

For the Swedish childhood leukemia data presented in Table 6, the p-value for the unbiased combined Knox test is 0.237. This indicates that there was no significant space-time interaction of childhood leukemia in Sweden during the period 1973–1994. From an epidemiological viewpoint, though, it is not necessarily the union of all types of leukemia that is of primary interest in a space-time analysis. More detailed analyses by subgroup will be presented in a medicine-oriented paper.

A combined Knox test can be seen not only as a way to account for the multiple testing of several Knox tests, but also as a test in itself to be compared with other space-time interaction tests. Some of these, including Mantel (1), were proposed precisely to avoid the arbitrariness in the choice of critical distances. They are not the same as the combined Knox test, though.

Mantel (1) and Diggle et al. (12) sum up the value of several Knox tests and use the combined sum as an omnibus test statistic. Diggle et al. do the summation for a finite set of critical distances, while Mantel uses a general function leading to continuous summation (integration) if the function is continuous, and to the ordinary Knox test if a dichotomous indicator function is used. The combined Knox test, on the other hand, picks the maximum rather than the sum over a finite set of critical distances.

The choice of method depends on the set of alternative hypotheses for which the user wants to maximize the statistical power. An advantage of the approaches taken by Mantel and Diggle et al. is that they model a gradual decrease in the strength of space-time clustering with increasing distance. A drawback is that the relative strengths at different distances have to be specified a priori. The combined Knox test, on the other hand, models an abrupt cutoff point just like the ordinary Knox test, in which the strength of space-time clustering is constant within the critical distance and zero outside. Unlike the Knox test, though, the critical distances do not need to be specified a priori, and unlike the Mantel and Diggle et al. tests, the relative strengths of clustering at different distances need not be specified. This has two advantages. It is not necessary to limit the scale of space-time interaction to be tested for, and the result provides not only an overall p-value but also, if the result is significant, an indication of the scale at which the space-time interaction operates.

Discussion

In looking at the population shift bias of space-time interaction tests, we have focused on the Knox method because it is the method most widely used for epidemiological data. Such bias is also present in other space-time interaction tests, proposed by David and Barton (10), Mantel (1), Pike and Smith (11), Diggle et al. (12), Jacquez (4), and Baker (13). The Mantel test, and even more so the Jacquez test, have been shown to have higher power than the Knox test for certain alternative hypotheses (4). Ironically, this also means that the population shift bias is higher, because a test's population shift bias is simply its power to detect the space-time interaction inherent in the population distribution. Fortunately, the procedure for constructing the unbiased Knox test can also be used for the Mantel and Jacquez tests, in the same simple fashion.

An unbiased combined Jacquez test would be especially attractive. Rather than using fixed geographic distances as Knox (5), Mantel (1), and Diggle et al. (12) have done, Jacquez (4) defines distances in terms of nearest neighbors, so that cases 1 kilometer apart are considered to be close to each other in a rural area but not necessarily so in a densely populated city. This increases the power when there is space-time interaction in less-populated areas.

No matter which space-time interaction test is used, it would have been ideal to show that, for practical purposes, the population shift bias is more or less irrelevant. Unfortunately, that is not the case (see "Estimation of the Population Shift Bias"). This leads to two questions: How do we do this type of analysis in the future? How do we interpret past results in light of the bias that may be associated with them?

To perform an unbiased test for space-time interaction in an area, we need underlying population data for that area. These data are sometimes harder to get than the case data. If a proper test is to be performed, there is no way around this, but in some cases there is a shortcut. The ordinary space-time interaction tests are all liberal. Therefore,

we know that if there is no significant space-time interaction using the ordinary test, then space-time interaction will not be significant according to the unbiased version. This suggests a two-stage procedure. First, collect only the case data and use one of the ordinary space-time interaction tests. If the result is non-significant, then there is no need to obtain the population data and the negative results can be published as such. If the result is significant, though, the population shift bias may be affecting it. It is then important to obtain population data and apply the unbiased version before making any conclusions.

Caution should be used in interpreting results that have already been published. If a result is non-significant, then it is fine. If the study period was only one or two years, the population shift bias is probably not a major problem because differential changes in population sizes did not have much chance to accumulate. For datasets spanning 10 or 20 years, though, there is really no way of knowing how reliable the results are without reanalyzing the data using an unbiased approach. For any past results that are considered important from an etiological or public health standpoint, we recommend that the data be reanalyzed using the unbiased version of any of the space-time interaction tests.

Acknowledgments

Valuable suggestions from Geoffrey Jacquez are gratefully acknowledged.

References

1. Mantel N. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209–20.
2. Williams GW. 1984. Time-space clustering of disease. In: *Statistical methods for cancer studies.* Ed. RG Cornell. New York: Marcel Dekker.
3. Chen R, Mantel N, Klingberg MA. 1984. A study of three techniques for time-space clustering in Hodgkin's disease. *Statistics in Medicine* 3:173–84.
4. Jacquez GM. 1996. A *k* nearest neighbor test for space-time interaction. *Statistics in Medicine* 15:1935–49.
5. Knox G. 1964. The detection of space-time interactions. *Applied Statistics* 13:25–9.
6. Wartenberg D, Greenberg M. 1994. Personal communication.
7. Alexander FE. 1992. Space-time clustering of childhood acute lymphoblastic leukemia: Indirect evidence for a transmissible agent. *British Journal of Cancer* 65:589–92.
8. Petridou E, Revinthi K, Alexander FE, Haidas S, Koliouskas D, Kosmidis H, Piperopoulou F, Tzortzatou F, Trichopoulos D. 1996. Space-time clustering of childhood leukemia in Greece: Evidence supporting a viral etiology. *British Journal of Cancer* 73:1278–83.
9. Klauber MR, Mustacchi P. 1970. Space-time clustering of childhood leukemia in San Francisco. *Cancer Research* 30:1969–73.
10. David FN, Barton DE. 1966. Two space-time interaction tests for epidemicity. *British Journal of Preventive Social Medicine* 20:44–8.
11. Pike MC, Smith PG. 1968. Disease clustering: A generalization of Knox's approach to the detection of space-time interactions. *Biometrics* 24:541–56.

12. Diggle P, Chetwynd AG, Häggkvist R, Morris SE. 1995. Second-order analysis of space-time clustering. *Statistical Methods in Medical Research* 4:124–36.
13. Baker RD. 1996. Testing for space-time clusters of unknown size. *Journal of Applied Statistics* 23:543–54.
14. Gilman EA, Knox EG. 1995. Childhood cancers: Space-time distribution in Britain. *Journal of Epidemiology and Community Health* 49:158–63.
15. Barton DE, David FN. 1966. The random intersection of two graphs. In: *Research papers in statistics: Festschrift for Jerzy Neyman.* Ed. FN David. London: John Wiley & Sons. 445–59.
16. Aldrich TE, Drane JW. 1993. *Cluster, v. 3.0: A program for identifying and analyzing the spatial and temporal structure of chronic disease patterns.* Atlanta, GA: Agency for Toxic Substances and Disease Registry.
17. Dwass M. 1957. Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28:181–7.
18. Jacquez GM. 1994. *Stat!: Statistical software for the clustering of health events.* Ann Arbor, MI: BioMedware.
19. Kulldorff M. 1998. Statistical methods for spatial epidemiology: Tests for randomness. In: *GIS and health in Europe.* Ed. A Gatrell, M Löytönen. London: Taylor & Francis.

Exploratory Data Analysis in a Study of Breast Cancer and the Environment

Steven J. Melly (1),[1] Nancy I. Maxwell (1), Yvette T. Joyce (2), Julia G. Brody (1)
(1) Silent Spring Institute, Newton, MA; (2) Applied Geographics, Inc., Boston, MA

Abstract

In the first phase of the Cape Cod Breast Cancer and Environment Study we used a geographic information system (GIS) as the central management and analysis tool in a detailed cancer surveillance effort and community-level study. We mapped patterns of breast cancer incidence in relation to environmental exposures of concern including infiltration of waste water into drinking water and large-scale historical pesticide use. We developed methods to compensate for some limitations in the data including differences in source scales. Part of our work included using the GIS together with a statistical program for exploratory data visualization. We used the statistical program to explore the effects of using different cutpoints to define categories of both exposure and disease. This exploratory analysis brought together data on US Census units with geographic information on point and non-point sources of environmental pollution from a range of data sources. Results from this first phase of research were used to plan a case-control study that began in the fall of 1998.

Keywords: breast cancer, endocrine disrupters, drinking water, waste water

Introduction

With increasing access to health surveillance data from state cancer registries and other sources, communities across the country are learning how disease rates in their area compare with those in other areas. As new statistics are published, high incidence communities want to know why their rates are high and how to bring them down. The Cape Cod Breast Cancer and Environment Study illustrates how geographic information system (GIS) technology can be used both to develop more accurate and detailed descriptions of disease incidence and to explore the causes. The study is being conducted by the Silent Spring Institute, a nonprofit research organization dedicated to studying the links between women's health and the environment. The Institute is funded by the Massachusetts Department of Public Health.

When the Cape Cod Study began in 1994, Massachusetts Cancer Registry data indicated that age-adjusted breast cancer incidence was significantly higher in a majority of Cape Cod towns than in the state as a whole. Alarmed by these statistics, citizen activists, public health officials, and researchers began sifting through possible explanations. Were high breast cancer rates due to characteristics of women who live on the Cape, better mammography screening, or something about the environment?

Information about the population of Cape Cod suggested that it was substantially similar to the rest of the state. In contrast, the environment of the Cape is obviously

[1] Steven J. Melly, Silent Spring Institute, 29 Crafts Street, Newton, MA 02158 USA; (p) 617-332-4288; (f) 617-332-4284; E-mail: melly@silent.shore.net

different. Nearly all of the population relies for its drinking water on groundwater from a sand and gravel aquifer overlain by sandy soils, so drinking water wells are vulnerable to contamination from septic tanks and other land uses. Historically, pesticide use on the Cape may have been greater than elsewhere because of the large number of cranberry bogs, golf courses, and trees susceptible to gypsy moth and other pests. In addition, the Cape hosts two military facilities. All of these factors pointed to the environment as a possible key to understanding breast cancer on the Cape. Because no *single* factor stood out as a likely cause of the elevated breast cancer incidence, however, it was important to design a study that would allow the exploration of many factors. Considering multiple factors also was important given the complexity of the disease and the possible effects of carcinogens, tumor promoters, and genetics.

Methodology

Faced with the challenge of investigating the relationship between a complex disease and multiple environmental factors within a 440-square-mile region, the Silent Spring Institute proposed developing a GIS to be used as the central data management and analysis tool for the study. We used the GIS to further define the problem of breast cancer on Cape Cod by conducting a detailed cancer surveillance effort. We also began to characterize the environment of the Cape, identify differences within the Cape, and explore relationships between the breast cancer incidence and environmental features.

In our cancer surveillance effort we used the GIS to geocode the addresses of 2,173 women with breast cancer for the period 1982 to 1994 (the full set of addresses currently available from the Cancer Registry). We also used residential land use data to refine estimates of populations for intercensal years. We used these refined population estimates to calculate standardized incidence ratios (SIRs) for towns, census tracts, and census block groups. The results of this cancer surveillance effort indicate that breast cancer incidence is about 20% higher on the Cape compared with the rest of Massachusetts.

When we looked at breast cancer incidence in geographic units smaller than the town, we identified six geographic areas scattered across the Cape where the excess cancer incidence is focused (Figure 1). We were particularly interested to note elevated incidence in the area of the Massachusetts Military Reservation (MMR), a Superfund site. A case-control study previously conducted in this part of the Cape by members of our research team also identified a statistically unstable association between breast cancer and the gun and mortar positions at the MMR (1). These sites were not only used for artillery practice, but also for burning propellant bags. Dinitrotoluene, a known mammary carcinogen in animals, is one of the chemicals found in the propellant. Other areas of elevated incidence are southern Falmouth; south Barnstable; a mid-Cape area including parts of Yarmouth, Dennis, and Harwich; a section of Orleans and Chatham; and south Truro.

We began using the GIS to explore how the environmental features in the areas of elevated incidence differ from the features in the rest of the Cape and in the rest of the state. We especially focused on environmental features that might be related to exposure to endocrine disrupting compounds (EDCs). EDCs include several chemicals that act like estrogens and are found in common commercial products and in the environment. Because breast cancer has been shown to be associated with lifetime exposure to natural estrogen, it is plausible that there might be a link between synthetic chemicals

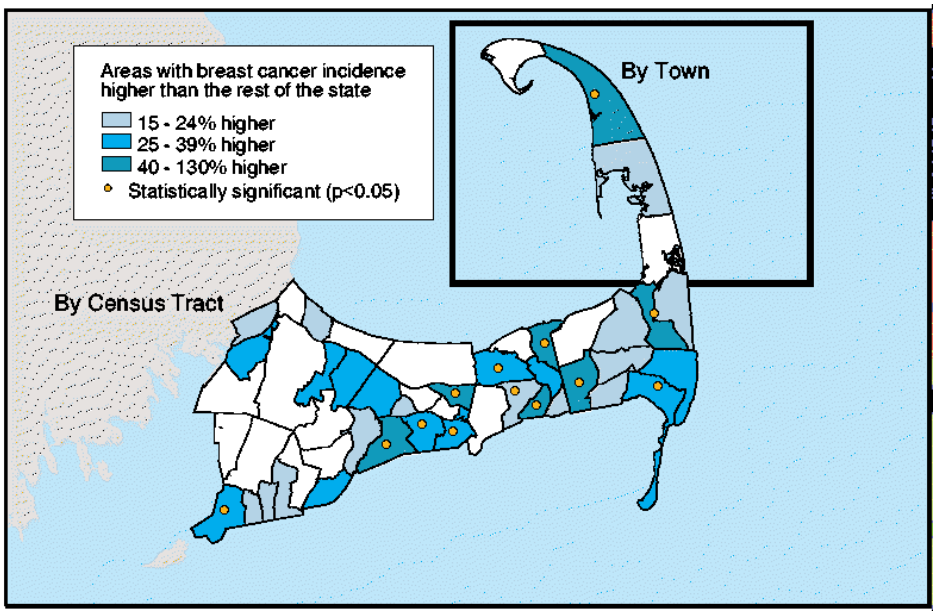


Figure 1 Breast cancer incidence by census tract, 1982-1994. Cape-wide breast cancer incidence was 20% higher than the rest of Massachusetts for this period. The circled areas are subregions of Cape Cod where the excess breast cancer incidence is concentrated.

that act like estrogen and breast cancer. We gathered data about two potential routes of exposure to EDCs on Cape Cod: exposure to pesticides through inhalation, dermal contact and ingestion, and exposure to drinking water impacted by waste water.

We used the GIS along with historical records and land use data to map areas of pesticide use. We focused on pesticides used on cranberry bogs and golf courses and those used to control tree pests. Our work using the GIS to study pesticides was described in a demonstration project for this conference.

Waste water has been shown to contain endocrine disrupting compounds. For several years, local, state, and federal agencies have been concerned about the impact of development on the Cape and of waste water disposal practices on the aquifer. The US Geological Survey collected data on analyses of private wells conducted by Barnstable County Health and Environment Department. Researchers have focused on nitrate-nitrogen as an indicator of water quality. Natural nitrate concentrations on the Cape are low, less than 2 mg/L. Waste water and fertilizer can cause nitrate concentrations to be elevated.

In an ecological epidemiology analysis we conducted using data from the cancer registry, we did not see any association between breast cancer incidence and elevated nitrate levels in drinking water. In the interest of getting the most out of readily available data, we explored the data further with statistical and data visualization software in order to generate hypotheses for further study.

We used the GIS to investigate how much the population and environmental features vary within the Cape and found substantial variation. One difference within the Cape that stands out is the distinction between the Lower Cape—the easternmost

part—and the rest of the Cape. The population density in this region, with the exception of Provincetown at the tip of Cape Cod, is lower than most of the rest of the Cape (Figure 2). The Lower Cape towns of Truro, Eastham, and Wellfleet are almost entirely dependent on private wells for their drinking water supply.

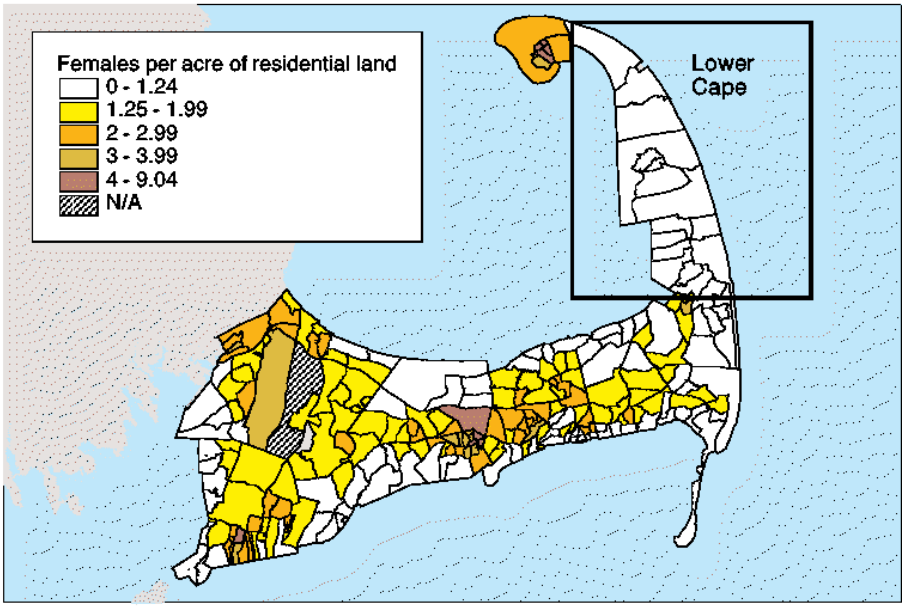


Figure 2 Population density by census block group, 1990. There are differences in both the environment and the population within the Cape. The towns of the Lower Cape, inside the rectangle, have lower population density.

Two factors that we speculated might be associated with water quality in areas served by private wells are housing density and housing age. Obviously, areas of dense housing would contribute more waste water to the aquifer. Older housing may include homes with cesspools and other waste water disposal systems that do not meet today's standards. In addition, the longer a septic tank is in operation, the greater the impact it will have on the aquifer. Housing density information was available from land use data and census data. Age of housing information was available from the census at the block group level.

Within the Lower Cape we found some variation in housing density (Figure 3) and age of housing (Figure 4). Denser and older residential areas were concentrated in the town centers. Wellfleet Center stands out as a location with particularly high density and old homes.

We created a series of scatter plots of SIRs by census block group versus percent housing greater than a specified number of years for those block groups of the Cape that are primarily dependent on private wells. We found that there did not appear to be an association when the x-axis was percent housing greater than 20 years. When we looked at percent housing greater than 50 years old there appeared to be a weak positive correlation (Figure 5). The correlation was strongest if only the census block groups

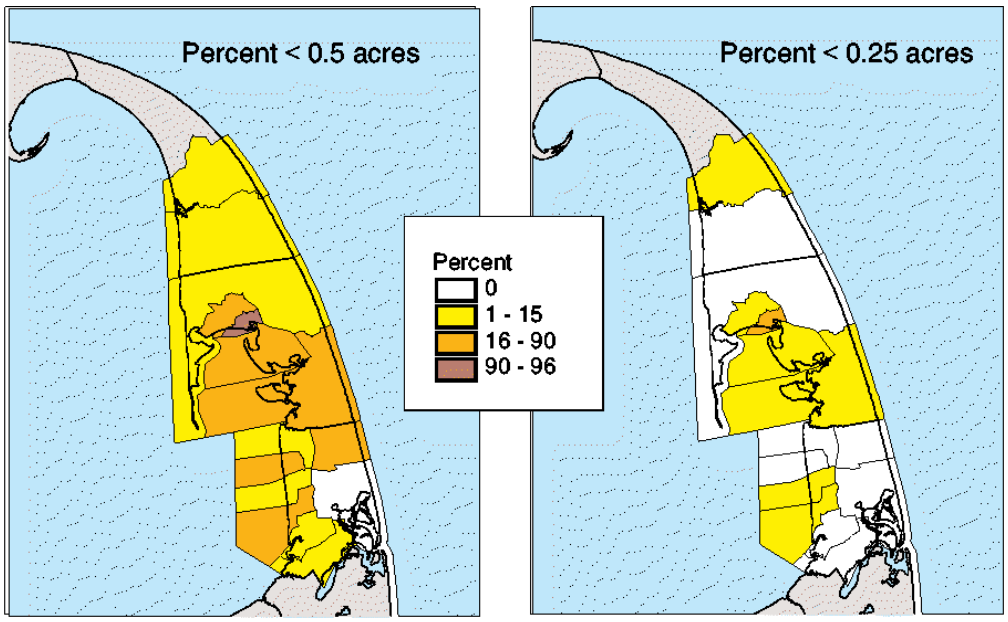


Figure 3 Percent high-density housing, 1971. Within the Lower Cape there are differences in housing density. Areas of higher density may have more wells impacted by waste water disposed of in septic tanks.

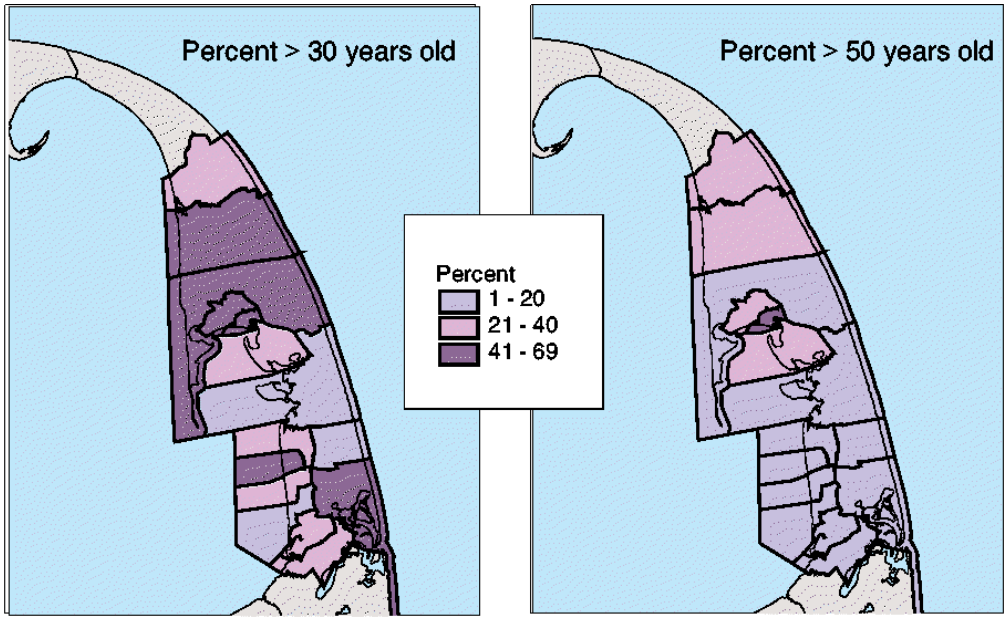


Figure 4 Percent old housing. The age of housing also varies within the Lower Cape. Wellfleet Center, in the middle of the figure, stands out as an area with old, dense housing.

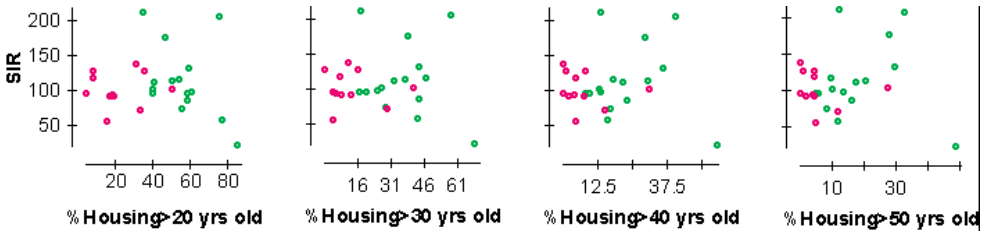


Figure 5 Breast cancer incidence and age of housing. Standardized incidence ratios (SIRs) by census block group were plotted against percent of old housing for the block groups where private wells are the primary source of drinking water. The correlation between breast cancer incidence and age of housing is highest when housing over 50 years old is used for the x-axis. The lighter dots are the block groups of the Lower Cape and the darker dots are the block groups for the rest of the Cape. Wellfleet Center stands out as a block group with low breast cancer incidence and old housing in the lower right of the plots.

of the Lower Cape were considered. Wellfleet Center stands out as an outlier with old homes and dense housing but low breast cancer incidence.

We also developed a simple scheme to take into account age of housing and housing density together. We calculated a housing risk index (HRI) according to the following formula:

$$HRI=a+d$$

where:
a=percent of old housing
d=percent of residential land in smaller than ½-acre lots up to a maximum value of 'a'

We assumed that if there was more dense housing than old housing then the excess dense housing must be new. We created a series of plots of SIRs versus the HRI (Figure 6). We found that the association between incidence and age of housing for the Lower Cape became even stronger when housing density was taken into account using the HRI.

Conclusion

We do not intend this sort of exploratory analysis to be used as evidence that certain environmental factors are responsible for the elevated breast cancer. A major limitation of this analysis is the lack of information available about confounding factors. For example, risk of breast cancer incidence has been shown to be correlated with elevated socioeconomic status (SES). It is possible that in the Lower Cape, the areas of older homes may be areas of higher SES because the older homes might be more desirable in this area. Ecological epidemiology analyses are intended to be hypothesis generating. In our case, the analyses suggest that the hypothesis that breast cancer incidence is associated with exposure to drinking water from shallow wells in areas of old and dense housing should be refined to focus on exposure in areas with high percentages of housing greater than 50 years old.

We are gathering data about confounding factors in a case-control study begun in the fall of 1998. GIS has been incorporated into the design of this study from the

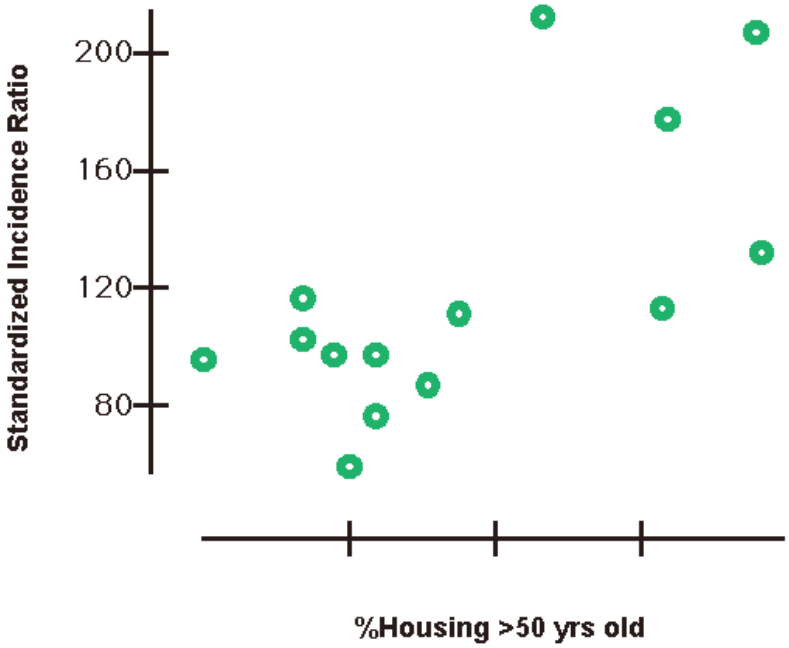


Figure 6 Breast cancer incidence, age of housing and housing density in the Lower Cape. The correlation between breast cancer incidence and age of housing is stronger when the age of housing is weighted to account for housing density. Wellfleet Center was excluded from this plot.

beginning. GIS data compiled in the first phase of our study are being used to develop exposure variables in the current case-control phase. The exploratory analyses we conducted in the first phase of our study illustrate some of the kinds of analyses that can be done when GIS data are combined with disease incidence data.

Acknowledgements

This work was supported by funds appropriated by the Massachusetts Legislature and administered by the Massachusetts Department of Public Health under Contract DPH79007214H11, and by the generous contributions of the Susan G. Komen Breast Cancer Foundation/Boston Race for the Cure.

Reference

1. Aschengrau A, Ozonoff DM. 1991. *The Upper Cape cancer incidence study: Final report*. Boston, MA: Boston University School of Public Health. September.

Temporal and Spatial Distributions of Cases of Verocytotoxigenic *Escherichia Coli* Infection in Southern Ontario

Pascal Michel (1),* Jeff Wilson (1,2), Wayne Martin (1), Scott McEwen (1), Robert Clarke (3), Carlton Gyles (4)

(1) Department of Population Medicine, OVC, University of Guelph, Guelph, Ontario, Canada; (2) Laboratory Centre for Disease Control, Health Canada, Canada; (3) Guelph Laboratory, Health Canada, Guelph, Ontario, Canada; (4) Department of Pathobiology, OVC, University of Guelph, Guelph, Ontario, Canada

Abstract

The distribution of 3,001 cases of verocytotoxigenic *Escherichia coli* (VTEC) reported in the province of Ontario, Canada, was examined to describe the magnitude of this condition geographically and to evaluate the spatial relationship between livestock density and human VTEC incidence using a geographical information system (GIS). Incidence of VTEC cases had a marked seasonal pattern with peaks in July. Areas with a relatively high incidence of VTEC cases were situated predominantly in areas of mixed agriculture. Spatial analyses were done for the southern regions of the province. Spatial models indicated that cattle density had a positive and significant association with VTEC incidence of reported cases (p=0.000). An elevated risk of VTEC infection in rural population could be associated with living in areas with high cattle density. Results of this study suggested that the importance of contact with cattle and the consumption of contaminated well water or locally produced food products may have been previously underestimated as risk factors for this condition.

Keywords: VTEC, mapping, surveillance, spatial analysis, cattle density

Introduction

Data on 3,001 verocytotoxigenic *Escherichia coli* (VTEC) cases reported in Ontario, Canada, from January 1990 to December 1995 were extracted from the Ontario Ministry of Health's Reportable Disease Information System database. Cases of VTEC infection are defined as persons with compatible clinical signs for which verocytotoxin was detected from stool specimens; persons with compatible clinical signs and for which one or more strains of VTEC was isolated from stool or blood; or, any symptomatic person with an epidemiologic link to two or more laboratory-confirmed VTEC cases. Farm animal distributions and land use data were obtained from the 1991 Census of Agriculture for Ontario (1).

Spatial Regression and Mapping

All cartographic outputs were produced by ArcView 3.1 (ESRI, Redlands, CA). In addition to providing a relational linkage between databases and the production of maps,

* Dr Pascal Michel, Health Canada, Quebec, Canada; (p) 450-773-8521 x8475; (f) 450-778-8120; E-mail: pascal_michel@hc-sc.gc.ca

the software was used to perform several spatial manipulations including the calculation of the county's total area and centroid coordinates (latitude and longitude), the Euclidian and geographical distances between each pair of counties, and the production of contiguity and inverse distance matrices used for spatial autocorrelation and regression. The Moran's I (2) and G statistics were calculated to explore the spatial distribution of VTEC cases across the 49 counties of Ontario (Figure 1).

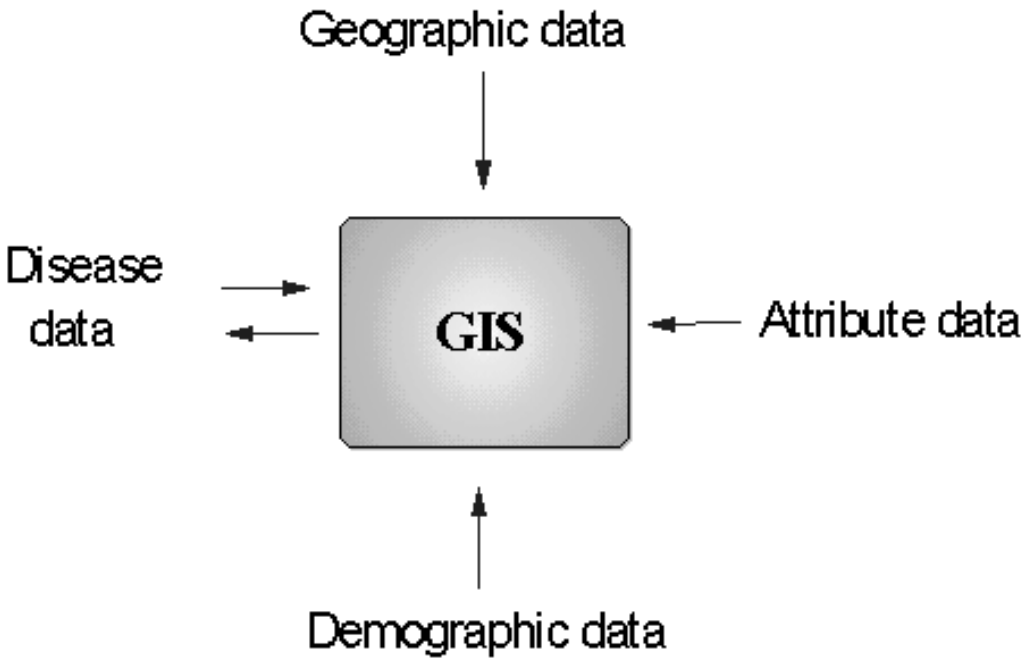


Figure 1 Sources of data.

Variables used in the modeling process included proportion of the total land that is cultivated (PCULTI); cattle density (TCDENS); dairy cattle density (TDDENS); density of livestock other than cattle (NOCATDENS); and livestock density (AUDENS) (3). An additive seasonal variation model was used to describe the temporal variation of VTEC cases over the six-year study period. This model includes a trend (Tt), a seasonal effect (St), and an error component (It) (Figure 2).

Geographical Distribution

The Moran's I index indicated an overall significant spatial autocorrelation of VTEC incidence in Ontario regardless of the underlying null distribution or the weight matrix chosen for the calculation ($p < 0.005$). For most regions, the geographic distribution of cattle density by county presented a geographic pattern similar to the one described for human VTEC cases (Figure 3). Counties with higher cattle density were located in three different areas of Ontario (Figure 4).

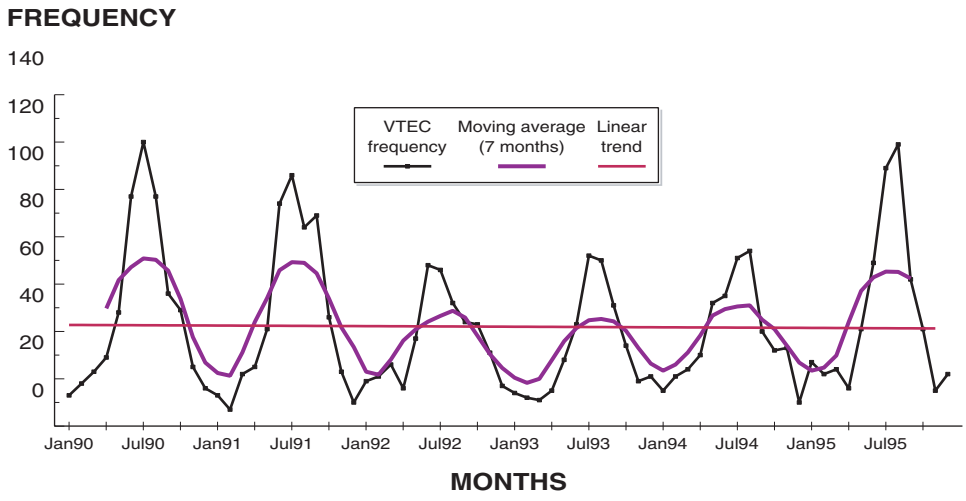


Figure 2 Linear trend and moving average for VTEC time-series, Ontario, Canada, 1990–1995.

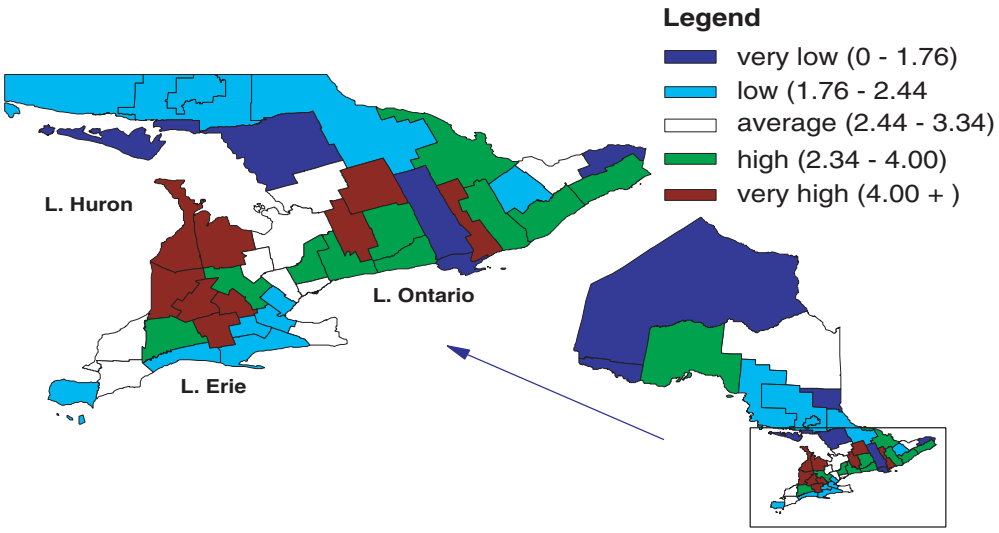


Figure 3 VTEC incidence in Ontario, Canada, 1990–1995. Direct standardized VTEC rates per county (per 10,000 population).

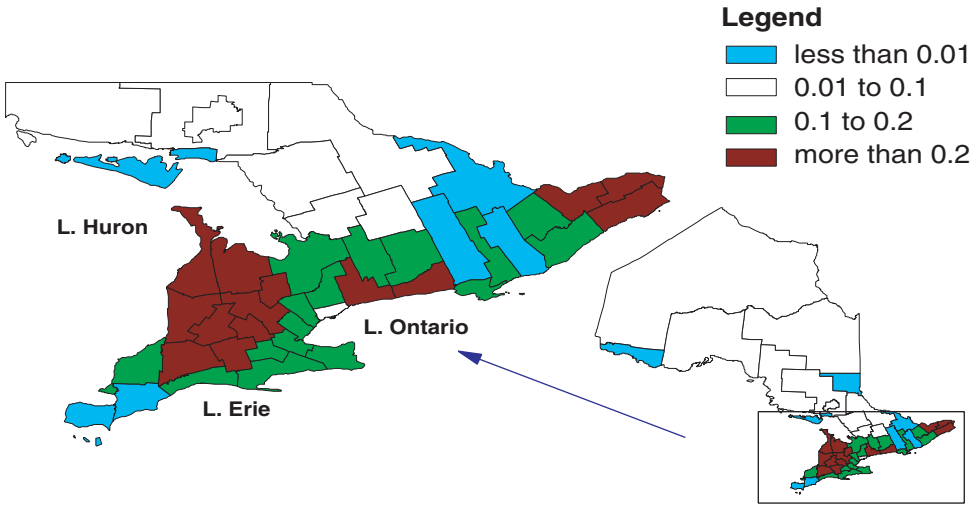


Figure 4 Cattle density in the Province of Ontario. Total cattle per hectare of total land.

Spatial Regression

In the first spatial regression approach, southern regions (south, west, central, and east) were analyzed separately from the northern region. The northern section of Ontario was omitted from the analysis. Besides the spatial error coefficient—latitude (YCOORD) and longitude (XCOORD)—the best predictors for the southern models included cattle density (TCDENS) (p=0.012) in a first model (A) and livestock density (AUDENS) (p=0.005) in a second model (B). In these models, the strong effect of latitude and longitude variables suggested some meaningful and undescribed spatial process influencing the outcome.

In the second approach, a spatial regimes model (4,5) was implemented to take into consideration different intercepts and/or slopes in the regression equation for the five agricultural regions of Ontario. The variable selection process resulted in a third model that included only the explanatory variable "total cattle density." Positive and significant coefficients representing the regional effects of cattle density on the incidence of human VTEC cases were calculated for the southern and western regions and negative and significant coefficients were associated with the eastern and northern regions. A positive but not statistically significant coefficient was estimated for the cattle-VTEC relationship in the central region.

Time Distribution

Observed rates of VTEC cases and predicted values from the additive seasonal model are presented in Figure 2. Visual assessment shows that the model fit the observed values closely, with some underestimation for the summers of 1990, 1991, and 1995 (coefficient of determination for the model: R=71.4%).

Conclusion

Results of the present study suggest that farm animal density, and particularly cattle density, is a significant predictor of human VTEC incidence in many regions of Ontario. This finding supports the possibility that direct and indirect human contact with reservoir animals is an important mode of transmission of VTEC organisms. In areas with higher cattle density, factors that could be responsible for VTEC transmission include the contamination of surface water and shallow wells by cattle manure; working with, or being in close contact with cattle; and, consumption of food produced and processed locally. It is understood that, under the limitations of the present study design, the observed association between human VTEC incidence and cattle density may not be causal. The importance, however, of such information for the public health and agriculture sectors underscores the need to promote further studies, including specific evaluations of the comparative risk of disease acquisition between rural and urban human populations, as well as investigations of environmental risk factors associated with human exposure to the cattle.

The temporal distribution of human VTEC cases reported in Ontario is regular with one seasonal peak in mid-summer. The regularity in the provincial cyclical pattern generated a very good fit between the observed monthly VTEC incidence and the expected level based on an additive seasonal variation model. A secondary objective in estimating the seasonal model was to provide reference values for a comparison of observed regional VTEC incidence and a 95% prediction interval based on the Ontario model. For most areas, the model could therefore be used to monitor observed surveillance data and point out unexpected temporal clusters of VTEC cases in a given region.

Future Directions

The national technical steering committee on foodborne, waterborne, and enteric disease surveillance of the Laboratory Centre for Disease Control (LCDC), Health Canada, has recently suggested that the project on the geographic surveillance of VTEC data in Ontario be expanded nationally to include other reportable enteric conditions such as campylobacteriosis and salmonellosis. This impulse has lead to the development of the National GIS Enteric Surveillance Initiative. The main objective of this initiative is to develop the capabilities and expertise to analyze and interpret geographically referenced surveillance data on priority foodborne pathogens with corresponding demographic and environmental information, and to make use of these resources in various surveillance activities and targeted studies in collaboration with public health partners. The initiative is also closely linked with various components of the National Health Surveillance Infostructure, which is supporting the development of a nationally integrated electronic health information network and includes the Canadian Integrated Public Health System (CIPHS), the Spatial Public Health Information eXchange (SPHINX), and the Geomatic Information System Infrastructure. The current research focus of the GIS surveillance initiative includes the epidemiology of high priority foodborne and waterborne enteric pathogens, the antimicrobial resistant enteric organisms transmitted from food and animals to humans, and the development of indices describing the environmental hygienic pressure linked to intense agriculture and livestock density.

Acknowledgments

The authors gratefully acknowledge Dr C Leber and Dr J Carlson of the Ontario Ministry of Health, the Medical Officers of Health of the regional health units of Ontario, Dr B MacDonald of Agriculture and Agri-Food Canada, Guelph, the Laboratory Centre for Disease Control (LCDC), Health Canada, and the agroecosystem research group of the University of Guelph for their participation in this study.

References

1. Statistics Canada. *1991 census of agriculture for Ontario*. Statistics Canada.
2. Moran P. 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society* B10:243–51.
3. Ministry of Agriculture and Food, Ministry of the Environment, Ministry of Housing of Ontario. 1996. *Agricultural Code of Practice.* 14–15.
4. Quandt R. 1958. The estimation of parameters of a linear regression system obeying separate regimes. *Journal of the American Statistical Association* 53:873–80.
5. Anselin L, Hudak S. 1992. Spatial econometrics in practice: A review of software options. *Regional Science and Urban Economics* 22:509–36.

Spatial Patterns of Malaria Case Distribution in Padre Cocha, Peru

Martha H Roper (1),* O Jaime Chang (2), Adeline Chan (1), Claudio G Cava (3), Javier S Aramburu (3), Carlos Calampa (3), Carlos Carrillo (2), Alan J Magill (1), Allen W Hightower (4)

(1) US Naval Medical Research Institute Detachment, Lima, Peru; (2) Instituto Nacional de Salud, Lima, Peru; (3) Direccion Regional de Salud de Loreto, Iquitos, Peru; (4) Centers for Disease Control and Prevention, Atlanta, GA

Abstract

Padre Cocha is a village of 1,400 inhabitants, situated in an area of epidemic vivax and falciparum malaria in the Peruvian Amazon. During the 1997–1998 transmission year, there were 1,157 *Plasmodium vivax* infections and 232 *Plasmodium falciparum* infections diagnosed at the village health post. As part of an ongoing study of malaria transmission in Padre Cocha, the village was mapped using global positioning system (GPS) hardware over the course of one week. Differential GPS correction of locations of all features mapped yielded a positional standard deviation of ±0.2 meters. Mapping of household malaria incidence data revealed areas of consistently high malaria infection density and a central area of low malaria incidence. This pattern suggests that transmission dynamics are heterogeneous within this village of approximately 1 square kilometer. The use of geographic information system (GIS) techniques to explore spatial relationships contributed to generating hypotheses when approaching this previously unstudied site, to exploring patterns of malaria case distribution, and to directing further entomological and epidemiological field work and malaria control measures. Proficiency with the required GPS equipment and GPS/GIS software was achieved by previously inexperienced users during a one-week training session, after which the on-site team was able to continue to use the system to successfully complete the project.

Keywords: malaria, Amazon, *Plasmodium falciparum*, *Plasmodium vivax, Anopheles darlingi*

Introduction

Malaria in Peru has undergone explosive growth during the 1990's, particularly in the Amazonian region of Loreto where more than 60% of the cases have occurred (Figure 1). Causes for the epidemic rise in *Plasmodium vivax* (*P. vivax*) and *Plasmodium falciparum* (*P. falciparum*) infections are thought to include the arrival of the highly efficient vector, *Anopheles darlingi* (*An. darlingi*); the dismantling of household residual insecticide spraying programs that took place prior to the 1990's; the changing patterns of river use; and the ecological disruption related to increasing jungle settlement and natural resource exploitation (1,2,3,4).

There has been relatively little study of specific factors related to malaria acquisition and transmission in the Amazon basin, despite the resources that have been

* Martha H Roper, 200 Morningside St., Middlebury, VT 05753 USA; (p) 802-388-4507; (f) 802-388-4507; E-mail: mroper@hsph.harvard.edu

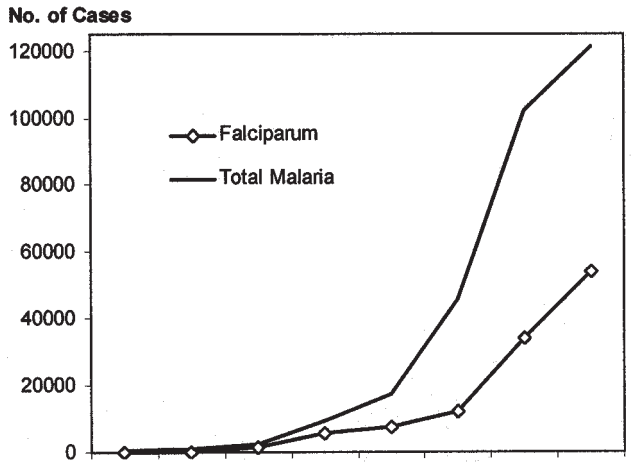


Figure 1 Malaria in Loreto, Peru, 1990–1997.

expended on treatment and control. Studies in Rondônia, Brasil, have shown that malaria is principally related to forest-based occupations such as gold mining and logging (5,6). In the Department of Loreto, Peru, malaria is more common in adults, particularly males, suggesting occupational risk as well (2,3). In other parts of the world, studies have highlighted the heterogeneous nature of malaria incidence and vector distribution within small areas (7,8), as well as the relationships of household malaria risk to vector abundance and distance from vector breeding sites (9,10,11). There have been no studies investigating specific malaria risk factors or spatial relationships in malaria transmission in Loreto to date. As part of an ongoing study of the epidemiology and transmission of malaria in Padre Cocha, Peru, the village was mapped using differential global positioning system (GPS) technology and spatial patterns of the distribution of malaria during the 1997–1998 transmission season were explored.

Materials and Methods

Padre Cocha is a village of 1,400 inhabitants, situated 5 kilometers (km) from Iquitos, the capital of Loreto (Figure 2). The village lies at the side of the Nanay River in a high malaria transmission zone (latitude 3°41'55" S; longitude 73°16'39" W; altitude 122 meters [m]). Between the river and village lies a cocha, a lake fed by the river, which expands and contracts with the river level. The central portion of the village has been cleared; the periphery is ringed by scrub and secondary forest, some of which is inundated during high-water months. Mean annual rainfall is 4.3 m, and the river fluctuates approximately 10 m throughout the year, peaking in April and May. There are 235 households with a mean of 6 occupants (sd=±3). House construction consists of two basic types—traditional boards or logs and thatched roofs, and newer brick and concrete with sheet metal roofs. Malaria is perennial with peak transmission during the wetter months of January through June.

In November 1997, all houses, streets, public buildings, and other features of interest in Padre Cocha were mapped in a five-day period using Trimble ProXR GPS

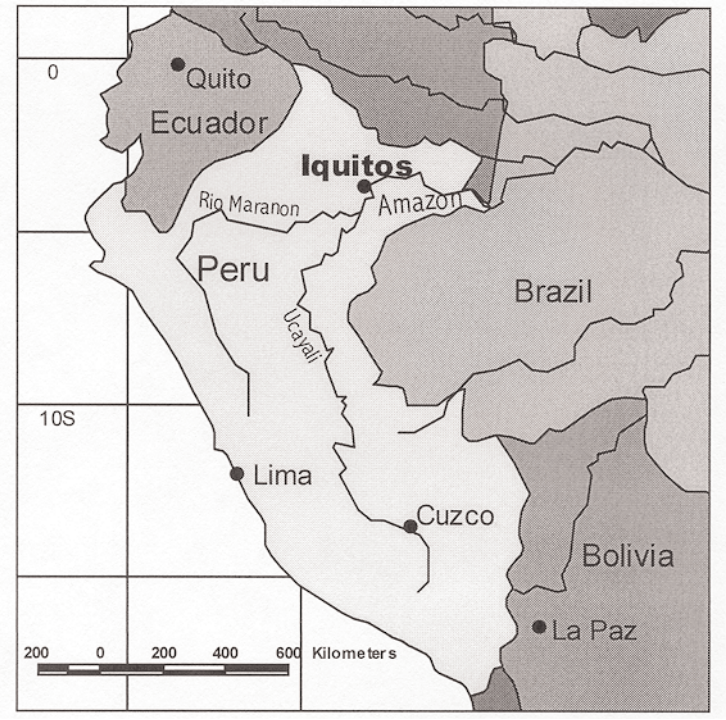


Figure 2 Map of Peru. The study site of Padre Cocha is located 5 km northwest of Iquitos.

receivers. GPS receivers capture radio frequency signals continuously emitted by 21 navigational satellites orbiting the earth at an altitude of approximately 20,200 km (12). During the Padre Cocha mapping, data were received from at least four satellites at any given time, permitting the calculations of latitude, longitude, and altitude. The exact methodology for how position fixes are computed is described in detail elsewhere (12,13).

Error in GPS position determinations arises from several uncontrollable sources, including atmospheric and topographic conditions, orbital and clock error, receiver noise, and most importantly, selective availability, the intentional error component built into the signal of each satellite. Selective availability varies with time and from one satellite to the next. Thus, errors are highly correlated with respect to time if readings are all from the same satellite group, while changes in satellite group members being used to measure positions yield abrupt changes in computed locations. The resultant variation in measurement error can lead to significant compromise of the accuracy of calculated positions. In the absence of adjustment, error in positional accuracy can be as high as 100 m (10,12,13).

To circumvent these sources of errors, post-processing differential correction of locational data, or differential GPS, was employed. For this technique, two receivers storing simultaneous signal information and subject to the same sources of error are required. One receiver is placed at a fixed known location, the base station, while the other is used to record information at remote sites of interest. The data from each receiver are downloaded to a computer program with differential correction capability,

are synchronized, and the base station information is used to calibrate the positions recorded by the mobile unit.

For the Padre Cocha mapping, one receiver was placed on the middle of a marker with coordinates previously identified by a US Geologic Service survey at the Peruvian Naval Base in Iquitos (latitude 3°44'05" S; longitude 73°14'25" W; altitude 95.5 m). This receiver served as the base station and collected data continuously, while the second receiver simultaneously recorded village feature positions. The locations of all village point features (houses, shops, public buildings, wells) were measured for two minutes with positional fixes taken at one-second intervals. For line and area features, positions were recorded at three-second intervals as lengths or borders were walked. The perimeter of the cocha was recorded from a canoe paddled around its circumference, using a constant offset from the shore. Each house was assigned a unique household identification number at the time that it was mapped, and information about construction type and the presence of home businesses was also recorded. Pathfinder Office software, version 1.10 (Trimble Navigation, Sunnyvale, CA) was used to perform differential correction of all feature locations and to create a locational database for use in geographic information system (GIS) analyses. In May 1998, new houses erected during the study year were mapped and added to the database.

At the time of the initial Padre Cocha mapping, a community-wide census was performed to obtain basic demographic information. Each identified resident was assigned a unique personal identifier and household address. During the course of the study year, the census database was updated as new residents were identified and former residents moved away. In May 1998, a second village-wide census was performed to verify the completeness and accuracy of all demographic information.

Malaria case data were collected continuously at the Padre Cocha Health Post from August 1, 1997, through July 31, 1998. Individuals with symptoms suggestive of malaria received Giemsa-stained blood smear examinations by experienced Ministry of Health (MOH) microscopists. Those whose blood smears demonstrated malaria parasites were considered cases and were treated in accordance with Peruvian MOH protocols. Information about smear positivity and treatment response was entered into a registry kept at the Health Post, and was subsequently abstracted and entered into an EpiInfo, version 6.04 (CDC, Atlanta, GA), computer database. Individual malaria case data were grouped by the household in which the infection occurred, and incidence for each household was calculated as the number of malaria episodes occurring during a given time period in each home, divided by the number of people residing in that home during that time period. Because there was movement of residents in and out of households during the year, the incidence for the study year was determined by calculating household incidence for each half of the year, and then summing the two six-month incidences. Household incidence for the low-transmission dry season (August–November), high-transmission wet season (December–March) and the transitional season of declining transmission (April–July) were calculated in the same manner.

Entomologic investigations began in Padre Cocha in March 1998. The results of a pilot study of vector abundance performed in late March to July 1998 are the source of data used in this analysis. Adult female mosquitoes were captured at four study stations using indoor and outdoor human landing catches. Mosquitoes were collected by a team of technicians working four consecutive nights from 6:00 PM to 6:00 AM, in twice-monthly cycles. The numbers of female anopheline mosquitoes detected were

entered into a computer database by species, parity, and location of capture. The total number of *Anopheles* mosquitoes captured both indoors and outdoors at each station was calculated and used for comparison with the human malaria case distribution for April through August 1998.

ArcView software, version 3.0 (Environmental Systems Research Institute, Inc., Redlands, CA), was used for GIS analysis. The Pathfinder locational database was exported as ArcView shape files, entered into ArcView and then merged with household malaria data exported from EpiInfo in Dbase format. Surface interpolations were performed with ArcView Spatial Analyst in order to make patterns of household-specific malaria transmission easier to interpret. This methodology summarizes information collected for all data points within a fixed distance from a given location; or, for a predetermined number of nearest neighbors, computes the mean or median and applies this smoothed estimate to the entire area under consideration. Observations are weighted proportionally to their distance from the center of the defined area. We used linear weighting and restricted consideration to only those households lying within 50 m of each specified point.

This work was conducted under research protocols approved by the scientific and human use committees of the Walter Reed Army Institute of Research (WRAIR Protocol No. 727) and the United States Army Medical Research Institute of Infectious Diseases (USAMRIID, HSRRB Protocol Log No. A-7421, DoD Protocol No. 30558), and the corresponding ethical review committee of the Direccion Regional de Salud de Loreto. In addition, the studies were conducted under Technical and Scientific Letters of Intention between the US Naval Medical Research Center Detachment (NAMRCD), Lima, Peru, and the Direccion Regional de Salud de Loreto and the Vice-Minister of Health (RM No. 237-97-SA/DM of 5 May 97) for the government of Peru.

Results

Figure 3A shows the map of Padre Cocha using the corrected locations of the November 1997 and May 1998 mappings. The village lies along two main axes with fairly straight streets and house alignments. Figure 3B shows the results of the initial village mapping prior to differential correction. Of note, the cocha appears wrapped over on itself; streets and paths appear crooked and haphazard, and houses have lost relation to each other and the streets. Figure 3C shows readings that were taken at the fixed base station during a five-hour mapping session in November 1997. Rather than showing a single point, the recorded locations follow a randomly meandering line with occasional abrupt changes in direction or position that reflect moments when satellites enter or leave receiving range. The linear pattern signifies that the errors are highly correlated from one reading to the next, and not statistically independent. Thus, readings taken over short periods of time contain approximately the same error term, and averaging them will not eliminate the error. This persistent error is the underlying cause of the inaccuracies that produced the chaotic uncorrected village map. Once differential correction was performed, the mean standard deviation of point feature positions measured in Padre Cocha was 0.2 m (sd=±0.09).

During the study year, there were 232 episodes of *P. falciparum* infections (incidence of 16.6%), 1,157 episodes of *P. vivax* (incidence of 82.6%), and a total of 1,300 independent episodes of malaria of either or both species (incidence of 92.9%). Mean household

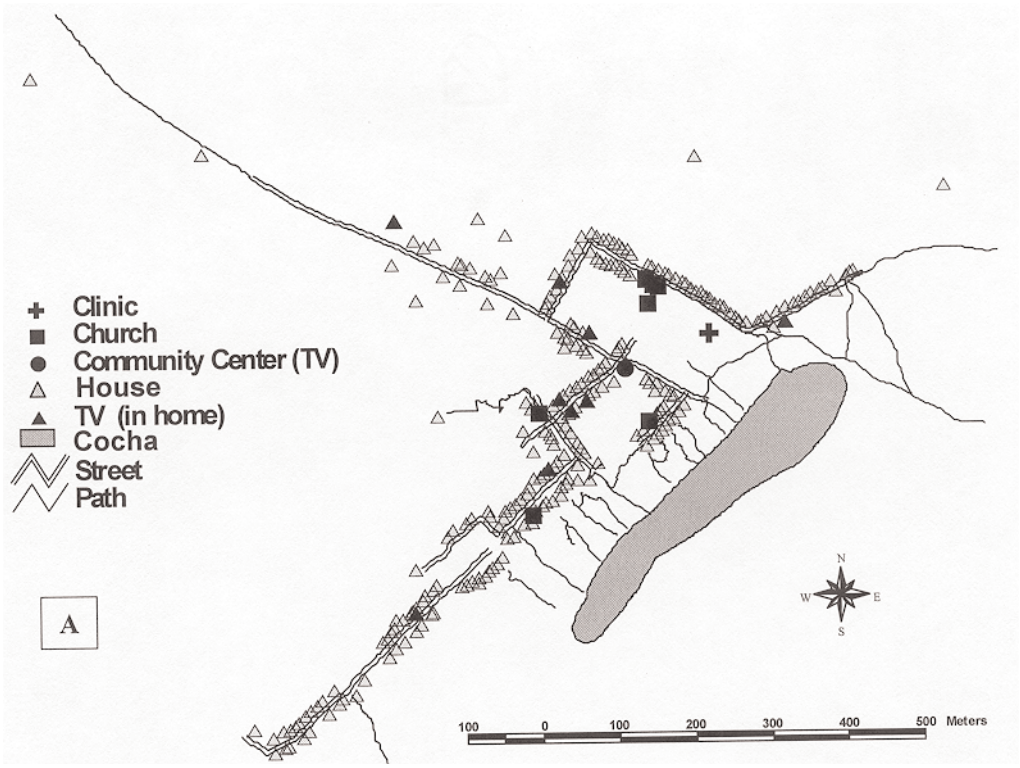


Figure 3A Map of Padre Cocha after differential correction of GPS readings.

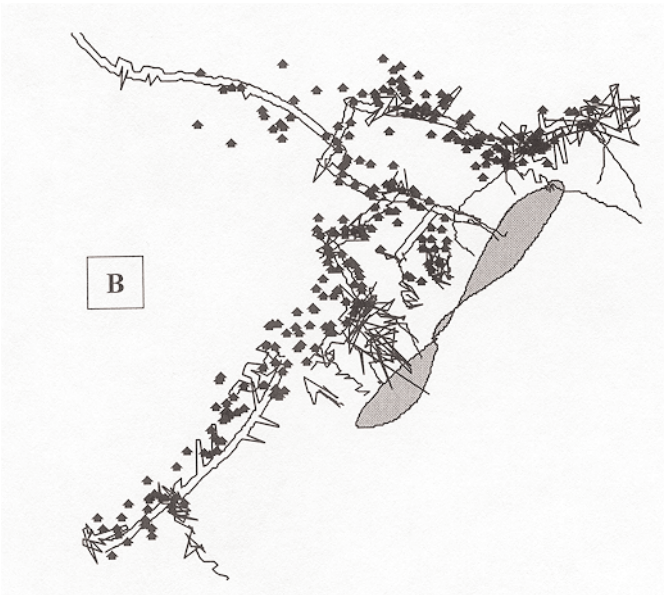


Figure 3B Map of Padre Cocha before differential correction.

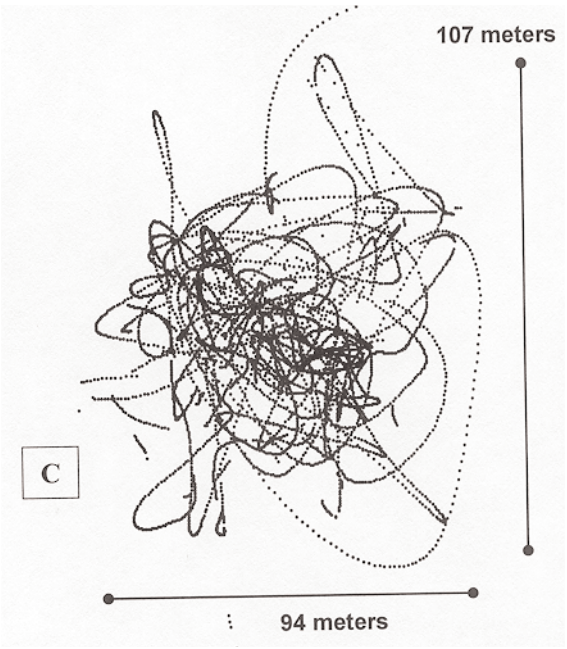


Figure 3C Plot of 5 hours of uncorrected GPS readings at the fixed base station.

incidences for the year were 15.3% for falciparum malaria, 82.1% for vivax, and 96.7% for either or both. Figure 4 shows the monthly distribution of malaria attack rates in Padre Cocha during the study year.

In Figure 5A, the cumulative number of malaria episodes occurring in each household during the study year was mapped using GIS. Several areas appear to have greater

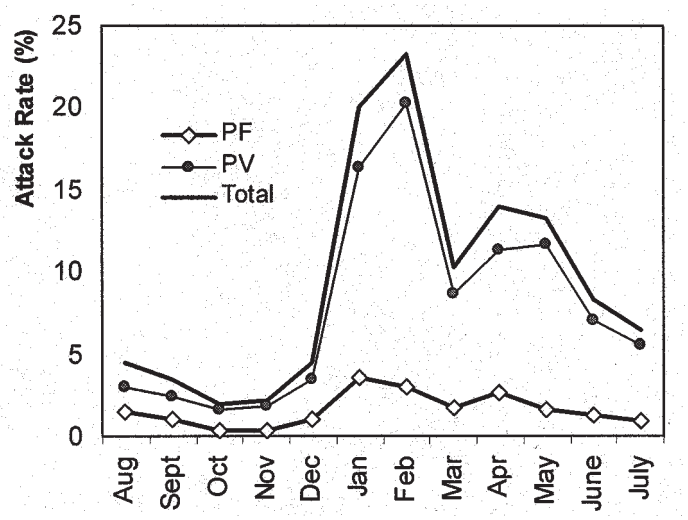


Figure 4 Monthly malaria attack rates in Padre Cocha, 1997–1998.

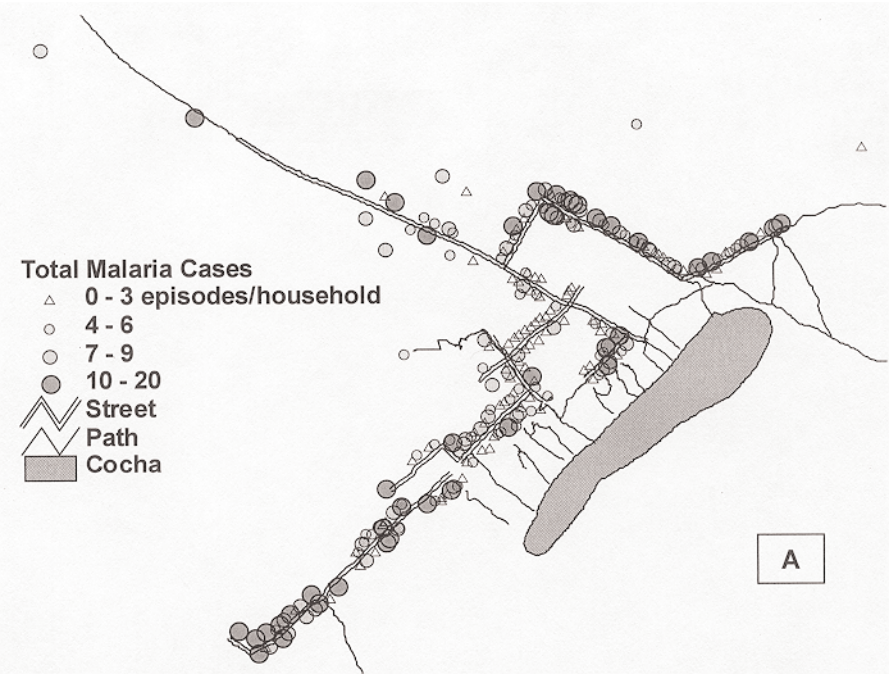


Figure 5A Household distribution of the cumulative number of malaria cases in Padre Cocha, 1997–1998.

concentrations of infections, while one central area appears to have fewer. To control for household size, household incidence was calculated and plotted, and a surface interpolation was performed to determine area incidence distribution. Clusters of high malaria density and the central area of low malaria intensity were confirmed (Figure 5B).

To exclude neighborhood population density as a factor in malaria occurrence, plots and interpolations of household size were performed. These analyses demonstrated a homogenous population distribution in the inhabited areas of the village throughout the year, supporting the conclusion that population density and household size are not significant determinants of malaria distribution. Similarly, altitude variation within the village perimeter was small and did not match the pattern of malaria distribution. Analysis of the spatial distribution of houses of the two construction types also showed a homogenous pattern throughout the village, and analysis of the relationship of household malaria cumulative incidence to house construction type showed no association (one-way analysis of variance: $F=0.5$; $p=0.61$).

Figure 6 demonstrates the results of the examination of spatial patterns of *P. vivax* (left) and *P. falciparum* (right) during the three phases of transmission during the year. At the top (sections A and B), the distribution of infections during the dry season from August through November 1997 is shown. The 1997 dry season was unusually prolonged, and little malaria occurred during that time. The middle sections (C and D) depict the dramatic rise in both vivax and falciparum malaria during the wetter months of December 1997 through March 1998. In the lower sections (E and F), the declining in-

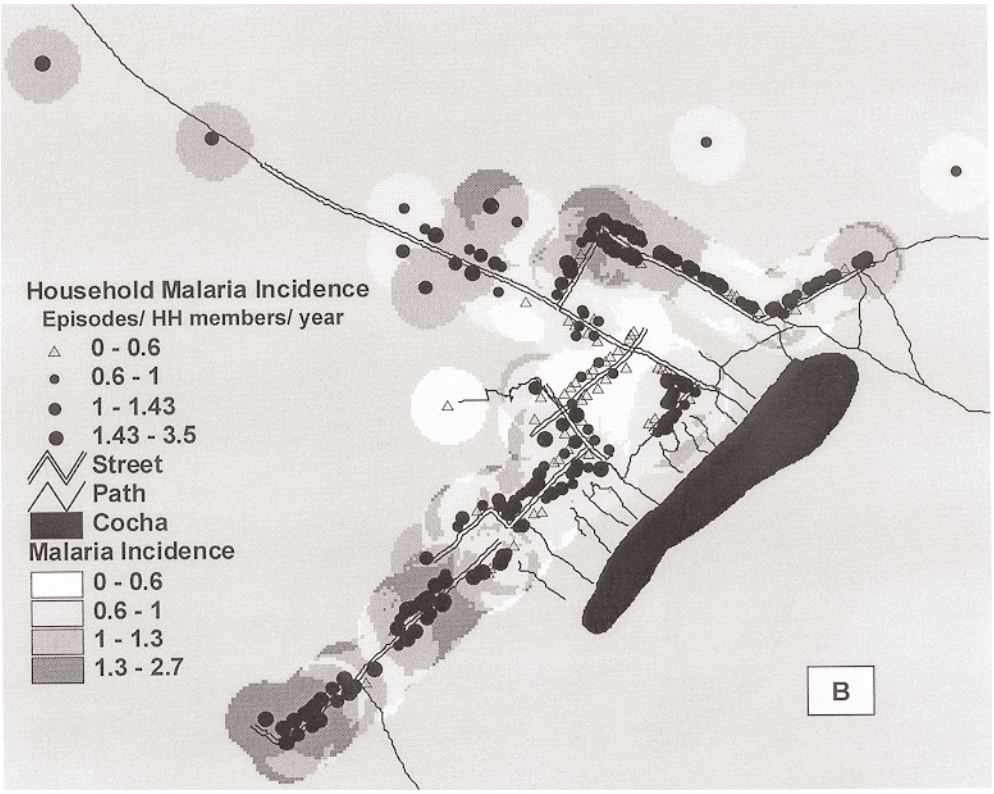


Figure 5B Household malaria incidence and surface interpolation of household incidence demonstrating the spatial distribution of malaria for the study year.

tensity of infections that occurred during the transitional period of April through July 1998 is shown. The general pattern of high and low density clustering noted in Figure 5 was again apparent for each species, and for all three parts of the year.

The results of adult female mosquito collections are displayed in Figure 7. The total number of mosquitoes collected indoors and outdoors at each of the four stations is listed in the legend. Virtually all mosquitoes collected (>97%) were identified as *An. darlingi*. The underlying surface interpolation represents the malaria incidence for the months of April through August 1998, months during which new human malaria infections would likely have been caused by mosquitoes of the same generations as those being captured. Of note, the capture stations with lower total numbers of indoor and outdoor mosquitoes catches were situated in areas of low malaria density, while those with higher numbers of mosquitoes captured were located in high malaria density areas.

Discussion

The use of GPS to map study sites is essential for GIS investigations of spatial relationships between exposure factors and disease occurrence in areas for which accurate maps are not available. Further, when the distances between features are small, differ-

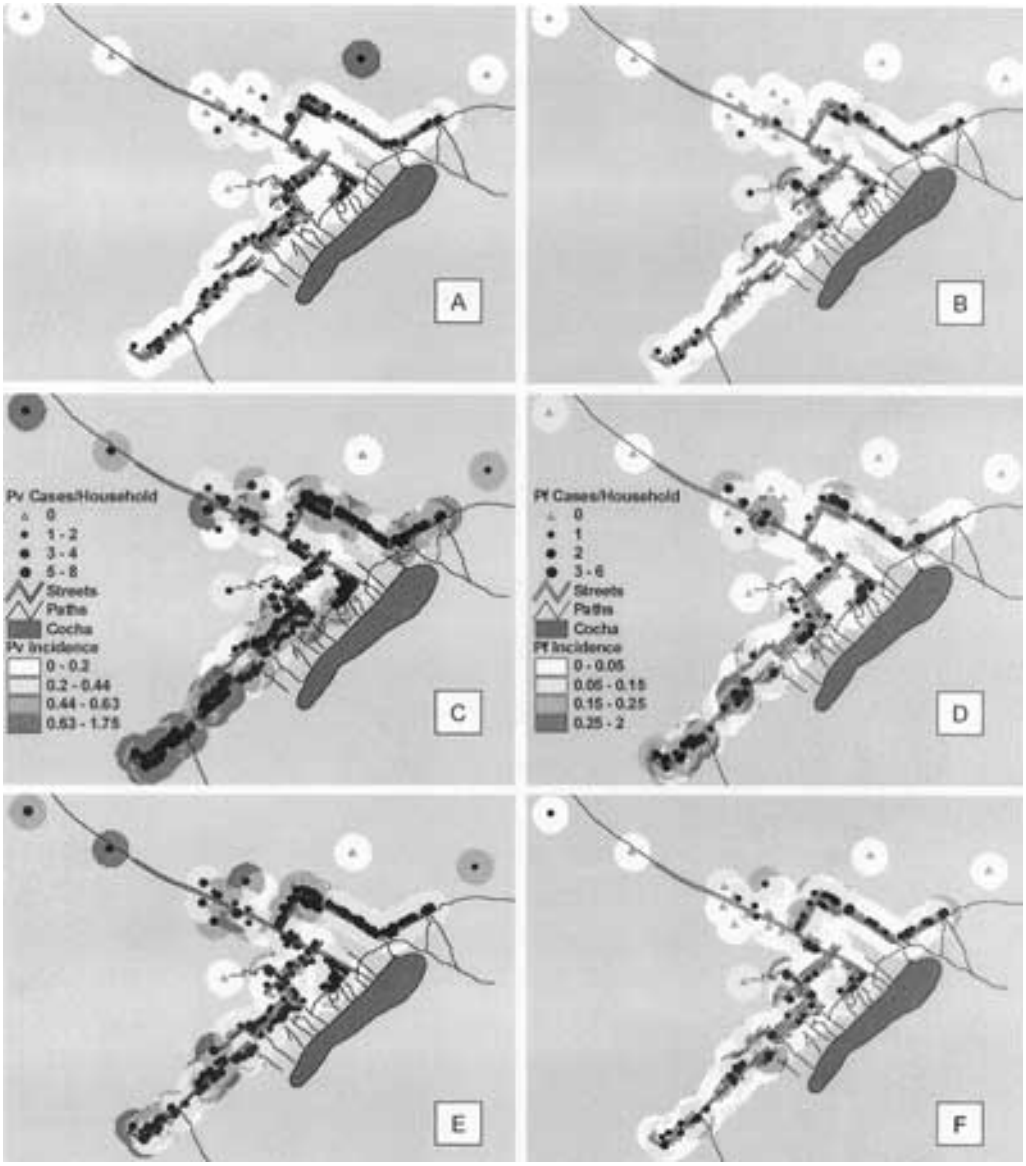


Figure 6 The distribution of cases and incidence of *P. vivax* infections (on left: A, C, E) and *P. falciparum* infections (on right: B, D, F) during the three transmission phases of the study year. A & B depict the dry season of low transmission, August–November. C & D represent the peak transmission period, December–March. E & F show the transitional period of declining transmission, April–July. The scales for cases per household and incidence were held constant throughout the seasons for each species.

ential GPS is necessary for adequate discrimination among sampling units. Our study of malaria distribution in Padre Cocha, Peru, is a good example of the importance and power of this technology. In an area of approximately 1 square kilometer, there was

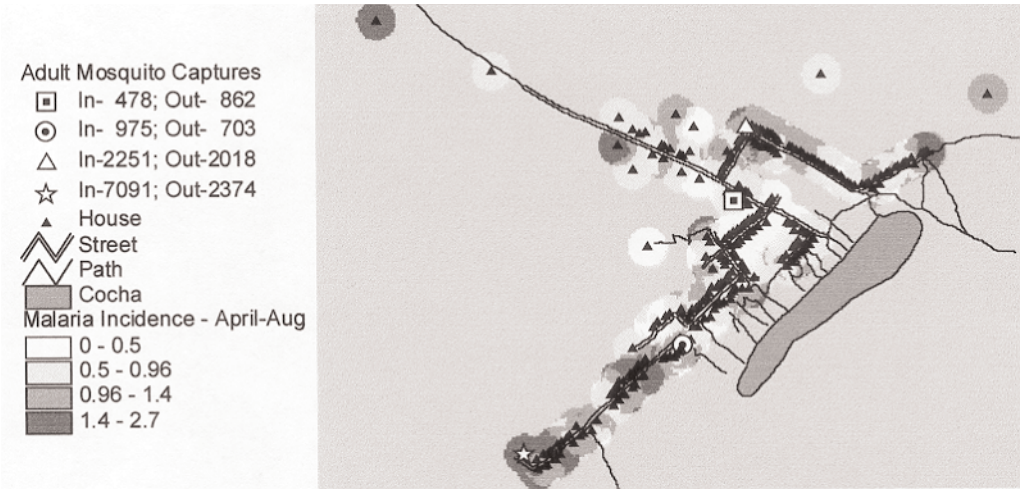


Figure 7 Location of mosquito capture stations and results of adult anopheline mosquito collections, April–August 1998. The underlying distribution of malaria incidence for the corresponding months is also shown.

clear and consistent spatial clustering of high and low malaria infection density. This spatial heterogeneity was true for both malaria species transmitted in the area, and was evident during periods of both high and low transmission.

A number of potential determinants of the distribution of malaria in Padre Cocha were investigated. Population density distribution, neighborhood altitude, and the spatial distribution of house construction type were unassociated with the pattern of malaria occurrence.

An epidemiological study performed during the 1997–1998 transmission year in Padre Cocha examined data about a variety of potential individual risk factors for malaria acquisition (14). Factors significantly associated with malaria incidence included age, working in Padre Cocha or its vicinity rather than elsewhere, time of arising (adults), evening strolling around the village (adults), and evening church attendance (children). The magnitude of associations was modest (range of RRs: 1.22–1.5), and none of these factors had any apparent geographic association with malaria distribution. Television viewing in the evenings was negatively associated with malaria incidence (RR 0.84; 95% CI, 0.73–0.96; p=0.01). Of interest, most of the locations where inhabitants congregate to watch television lie in the central zone of low malaria occurrence and low anopheles captures. Factors unassociated with malaria risk included bedtime, hour of bathing, specific occupation (farmer, fisherman, artisan, etc.), and bednet use. The overall lack of strong risk factor associations to explain the distribution of malaria at the individual or household level argues for the need to identify other factors related to risk, particularly factors with a significant spatial component.

Our analysis of the relationship between entomologic factors and human malaria infections is limited by the preliminary nature of the entomologic data. The finding that capture stations with low total *An. darlingi* catches were located in areas of low malaria infection densities and that, conversely, stations with high total counts were located in areas of high malaria intensity, is suggestive that patterns of vector abundance in and

around Padre Cocha are important determinants of the spatial distribution of infections in the village. The entomologic data being collected during the 1998–1999 transmission season, including information about adult mosquito abundance and behavior, and anopheline larval breeding site distribution, will permit a more complete analysis of this hypothesis.

The addition of GPS and GIS technologies to malaria and other vector-borne disease studies affords the possibility of exploring spatial dimensions of disease transmission not easily examined in the absence of these capabilities. Further, the incorporation of these techniques can be performed with a limited addition of time and resources. Our project team consisted of previously inexperienced GPS and GIS users and a consultant who provided one week of training and supervision. All subsequent work was successfully performed by the on-site team.

Village mapping required six separate four- to six-hour sessions, with approximately one hour of subsequent computer time for data downloading and processing. Learning how to set up and operate GPS equipment and software required one didactic and practice session, and then several sessions in the field to attain full confidence in use of the system. Sufficient mastery of GIS techniques to allow the production of basic data maps was achieved within days of introduction to the software. Training by an experienced user on site, while not strictly necessary, greatly simplified the learning process and assured that decisions about base station siting, GPS unit parameterization, mapping plan, and database design were appropriate and efficient.

The single greatest cost in performing the GPS/GIS work at Padre Cocha was that of the GPS equipment and software, which totaled approximately $20,000. Our system was extremely easy to learn and use, and provided accuracies that met our need for analysis within a small area. The very simplicity of the system eliminated any ongoing need for personnel with prior technical training, reducing personnel costs significantly. However, sophisticated GPS units that offer such ease of use are costly, and if site mapping is likely to be accomplished in a few sessions, leasing units may be preferable to purchase. Further, it is now possible to subscribe to a satellite service that can provide real-time differential correction, abrogating the need for a base station unit. Taking advantage of these cost-reducing measures could make the application of GPS/GIS techniques more feasible for projects with limited budgets.

In summary, differential GPS and GIS systems can add critical information to the understanding of malaria transmission and epidemiology. Technologies that are currently available can be used by researchers without specialized backgrounds, and for moderate cost. In our work at Padre Cocha, spatial analysis contributed to generating study hypotheses, to understanding malaria distribution in the community, and to focusing entomologic research and MOH vector control efforts.

Acknowledgments

We would like to acknowledge the invaluable assistance of Padre Cocha Health Post workers Maria Ricopa Huanaquiri, Leny Curico Manihuari, Miriam Ojaicuro Pahanaste, and Juan Cumapa Whuayambahua, and the gracious cooperation of the residents of Padre Cocha, Peru. This work was supported by funds from the United States Army Medical Materiel Development Activity (USAMMDA) and the Military Infectious Disease Research Program (MIDRP).

References

1. Ministerio de Salud del Peru, Oficina General de Epidemiologia. 1997. *Situation epidemiologica Daños Trazadores.* Lima.
2. Ministerio de Salud, Direccion Regional de Salud de Loreto. 1998. *Boletin de malaria 1997*. Iquitos.
3. Aramburu F, Ramal A, Witzig R. 1999. Malaria re-emergence in the Peruvian Amazon region. *Emerging Infectious Diseases* 5(2):209–15.
4. Roberts DR, Laughlin LL, Hsheih P, Legters LJ. 1997. DDT, global strategies, and a malaria control crisis in South America. *Emerging Infectious Diseases* 3(3):295–302.
5. Camargo LMA, Ferreira MU, Krieger H, De Camargo EP, Da Silva LP. 1994. Unstable hypoendemic malaria in Rondonia (Western Amazon region, Brasil): Epidemic outbreaks and work-associated incidence in an agro-industrial rural settlement. *American Journal of Tropical Medicine and Hygiene* 51(1):16–25.
6. Camargo LMA, Colletto GMD, Ferreira MU, Gurgel SM, Escobar AL, Marques A, Krieger H, Camargo EP, Da Silva LHP. 1996. Hypoendemic malaria in Rondonia (Brasil, Western Amazon region): Seasonal variation and risk groups in an urban locality. *American Journal of Tropical Medicine and Hygiene* 55(1):32–8.
7. Snow RW, Armstrong Schellenberg JRM, Peshu N, Forster D, Newton JC, Winstanley PA, Mwangi I, Waruiru C, Warn PA, Newbold C, Marsh K. 1993. Periodicity and space-time clustering of severe childhood malaria on the coast of Kenya. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 87:386–90.
8. Ribeiro JMC, Seulu F, Abose T, Kidane G, and Teklehaimanot A. 1996. Temporal and spatial distribution of anopheline mosquitos in an Ethiopian village: Implications for malaria control strategies. *Bulletin of the World Health Organization* 74(3):299–305.
9. Gunawardena DM, Wickremasinghe AR, Muthuwatta L, Weerasingha S, Rajakaruna J, Senanayaka T, Kotta PK, Attanayake N, Carter R, Mendis KN. 1998. Malaria risk factors in an endemic region of Sri Lanka, and the impact and cost implications of risk factor-based interventions. *American Journal of Tropical Medicine and Hygiene* 58(5):533–42.
10. Hightower AW, Ombok M, Otieno R, Odhiambo R, Oloo AJ, Lal AA, Nahlen BL, Hawley WA. 1998. A geographic information system applied to a malaria field study in western Kenya. *American Journal of Tropical Medicine and Hygiene* 58(3):266–72.
11. Thomson R, Begrup K, Cuamba N, Dgedge M, Mendis C, Gamage-Mendis A, Enosse SM, Barreto J, Sinden RE, Hogh B. 1997. The Matola malaria project: A temporal and spatial study of malaria transmission and disease in a suburban area of Maputo, Mozambique. *American Journal of Tropical Medicine and Hygiene* 57(5):550–59.
12. French GT. 1996. *Understanding the GPS: An introduction to the global positioning system.* Bethesda, MD: GeoResearch, Inc.
13. Herring TA. 1996. The global positioning system. *Scientific American.* February. 44–50.
14. Roper MH, Carrion RS, Cava CG, Andersen EM, Aramburú JS, Calampa C, Hightower AW, Magill AJ. 2000. The epidemiology of malaria in an epidemic area of the Peruvian Amazon. *American Journal of Tropical Medicine and Hygiene* (in press).

Evidence for Geographic Clustering of Reported Gonorrhea Cases: A Neighborhood-Level Analysis of Environmental Risk

Richard A Scribner, MD, MPH (1),* Deborah A Cohen, MD, MPH (1), Thomas A Farley, MD, MPH (2)

(1) Department of Public Health and Preventive Medicine, School of Medicine, Louisiana State University, New Orleans, LA; (2) Department of Epidemiology, Louisiana Office of Public Health, New Orleans, LA

Abstract

The availability of alcohol, as measured by alcohol outlet density, is associated with numerous alcohol-related outcomes in a small area analysis. A number of studies suggest that high-risk sexual behavior should also be considered an alcohol-related outcome. This study assessed the geographic relationship between alcohol availability and high-risk sexual behavior at the neighborhood level. Ecological analysis tested the geographic relationship of off-sale, on-sale, and total alcohol outlet density with reported gonorrhea rates among 155 urban residential census tracts in New Orleans, Louisiana, during 1995. All alcohol outlet density variables were positively related to gonorrhea rates. Off-sale outlets per square mile was most strongly related to gonorrhea rates (β=.582±.073 [standard error]), accounting for 29% of the variance in gonorrhea rates. Interpreted as an elasticity, a 10% increase in off-sale alcohol outlet density accounts for a 5.8% increase in gonorrhea rates. Including the covariates of percent African American and percent unemployed in the model reduced, but did not remove, the effect of off-sale outlet density (β=.192±.047). These results indicate that there is a geographic relationship between alcohol outlet density and gonorrhea rates at the census tract level. Although these results cannot be interpreted causally, they do justify public health intervention as a next step in defining the relationship between alcohol availability and high-risk sexual behavior.

Keywords: STD, gonorrhea, alcohol

Introduction

A small literature exists that documents an association between alcohol consumption and high-risk sexual behavior. Among both heterosexuals and homosexuals, alcohol use is associated with a greater likelihood of unprotected sex, multiple sexual partners, and anal intercourse (1–5). Although the relationship between alcohol use and high-risk sexual behavior is complex, the pervasiveness of alcohol consumption in the United States would make even a small effect relevant from a public health perspective.

The availability of alcohol, as measured by alcohol outlet density, has been demonstrated to be geographically linked to numerous alcohol-related outcomes, including

* Richard A Scribner, MD, Department of Public Health and Preventive Medicine, School of Medicine, Louisiana State University, 1600 Canal St., 8th floor, New Orleans, LA 70112 USA; (p) 504-568-6951; (f) 504-568-6905; E-mail: rscrib@lsumc.edu

alcohol consumption (6), drunk driving arrests (7,8), fatal and injury traffic accidents (9), alcoholism rates (10–12), cirrhosis mortality (13), and assaultive violence (14,15). If high-risk sexual behavior is considered an alcohol-related outcome, the distribution of high-risk sexual behavior should also be geographically linked to alcohol outlet density.

High endemic rates of a sexually transmitted disease within a population are explained by a high reproductive rate of infection. The existence of a high reproductive rate of infection within a high-risk core population is determined by three factors, including a high rate of partner change (16). Rothenberg (17) and Potterat et al. (18) have demonstrated that gonorrhea is geographically concentrated in certain neighborhoods where a core group of high-risk individuals is found. Consequently, the existence of high endemic rates of a sexually transmitted disease within a particular geographic area may be explained in part by a higher rate of alcohol consumption among residents of the area.

The present study analyzed the geographic relationship between the density of alcohol outlets and a proxy for high-risk sexual behavior—reported gonorrhea rates—among 155 census tracts in New Orleans, Louisiana. The analysis tested the hypothesis that high-risk sexual behavior in New Orleans is geographically clustered in neighborhoods and that the clustering of high-risk behavior is predicted by the density of alcohol outlets.

Methods

In 1995 there were 1,834 licenses for alcohol outlets in New Orleans. These licenses are classified as either on-sale—alcohol is purchased for consumption on the premises—or off-sale—alcohol is purchased for consumption off the premises. On-sale outlets include bars and restaurants, while off-sale outlets include liquor stores and grocery or convenience stores. New Orleans, a city of approximately 420,521 residents, has one alcohol licensee for every 230 residents.

Rates of reported gonorrhea cases were used as a proxy for high-risk sexual behavior. The high prevalence of gonorrhea, as compared with syphilis or HIV, makes it a sensitive indicator of high-risk sexual behavior at the census tract level. While chlamydia also has a high prevalence, gonorrhea records are more complete because there is a longer history of reporting this disease. One problem associated with using gonorrhea case reports as a marker of high-risk sexual behavior is the possibility of differential reporting rates by public- and private-sector physicians. This limitation can be addressed by using the variable of mean employment status to estimate the proportion of the population served by public- or private-sector physicians within a census tract.

Alcohol outlet license data were obtained from the Louisiana Office of Alcoholic Beverage Control (ABC) for March of 1995. Only active licensees were included in the analysis. Data from the ABC included trade addresses for active alcohol licenses were georeferenced utilizing MapMarker (MapInfo Corporation, Troy, NY), which used 1995-updated TIGER files. A trade address was available for 99% of all listings (1,868 of 1,893). Georeferencing to street address (1,635) or to zip+4 centroids (186) achieved a georeferencing rate of 97%. Reported gonorrhea cases for 1995 were obtained from the Louisiana Office of Public Health (OPH), Office of Epidemiology. OPH reported a georeferencing rate of 95%.

Georeferenced alcohol outlets and reported gonorrhea cases were then linked to

their census tract by overlaying census tract boundary files. Density of outlets and rates of gonorrhea were obtained by dividing the total number of alcohol outlets and reported cases of gonorrhea within a census tract by census tract population estimates. Population estimates for 1994 were obtained from projections of US Census data made available by the Claritas Corporation (19). An additional density statistic, outlets per square mile, was calculated for alcohol outlets by dividing the number of outlets by the size of the census tract in square miles.

Sociodemographic data aggregated by census tract were also obtained from the Claritas Corporation (19). Sociodemographic data incorporated in the analysis included percent of population that was African American and percent of population over 16 years of age that was unemployed or not in the labor force.

To assure that all census tracts represented urban residential neighborhoods, certain census tracts were omitted from the 184 found in New Orleans. Rural census tracts (5) were removed by excluding those tracts with a population of less than 2,000 persons per square mile. Commercial or tourist census tracts (7) were removed by excluding those tracts with on-sale outlet densities of greater than 200 outlets per 1,000 persons. Industrial or nonresidential census tracts (17) were removed by excluding those tracts with a total population of less than 500. In all, 155 urban residential census tracts were included in the analysis.

Data Analysis

Least squares regression analysis was used to examine the relationship between gonorrhea rates and the covariates. Percent African American and percent adults unemployed were selected to control for the higher rates of reported gonorrhea among African Americans and the potential underreporting of gonorrhea by private-sector physicians, respectively. The complete model was composed of all these covariates.

All variables included in the analysis were transformed to their base-10 logarithm to adjust for skew and to permit analysis of the results as an elasticity. Observations with zero values were assigned a value equal to one-half of the value of the lowest observation before transformation. After transformation, the regression slope estimates the percent change in the dependent variable associated with a 1% increase in the independent variable.

Separate analyses were conducted for the two primary independent variables: outlet density, measured as outlets per person; and outlet density, measured as outlets per square mile. In each analysis, the three outlet density categories—off-sale outlet density, on-sale outlet density, and total outlet density—were added to the basic model.

Results

New Orleans census tracts in the study had a mean of 3,013 residents, 9.3 licensed alcohol outlets, and 17 reported cases of gonorrhea for the year. In addition, the census tracts had a mean of 61.4% African American residents and an unemployment rate of 14.9% (Table 1).

In the initial analysis, each outlet density variable was regressed with census tract gonorrhea rates. From this analysis we observed a strong relationship between reported gonorrhea rates and off-sale outlet density, measured either as outlets per square mile

Table 1 Means and Standard Deviations for Study Variables (n=155), New Orleans, LA, 1995

	Mean	SD
Sociodemographic Variables		
Percent African American	61.4%	32.6
Percent adults unemployed	14.9%	10.03
Population 1994	3,013	2,077
Outlet Density Variables		
Total number of outlets	9.28	7.12
Off-sale outlets per 1,000	1.75	1.55
On-sale outlets per 1,000	2.31	3.13
Total outlets per 1,000	4.05	4.07
Off-sale outlets per sq. mile	18.02	18.24
On-sale outlets per sq. mile	22.31	3119
Total outlets per sq. mile	40.17	42.58
Gonorrhea Variables		
1995 gonorrhea cases	17.1	15.3
1995 gonorrhea rate per 1,000	6.15	4.59

(β=.582±.073) or outlets per person (β=.374±.061). The density of off-sale outlets per square mile accounted for 29% of the variance in gonorrhea rates, while the density of off-sale outlets per person accounted for 20% of the variance. On-sale outlet densities (i.e., outlets per person and outlets per square mile) demonstrated smaller but significant relationships with gonorrhea rates (Table 2). As with off-sale outlet density, the relationship was greater for on-sale outlet density measured as outlets per square mile (β=.300±.069) than as outlets per person (β=.201±.061).

The relationship between alcohol outlet density and reported gonorrhea rates can be interpreted as an elasticity because all covariates had been transformed to their base-10 logarithm. A 1% higher off-sale outlet density was associated with a 0.582% higher gonorrhea rate. Therefore, a 25% higher off-sale outlet density (one more off-sale outlet

Table 2 Coefficients with Standard Errors and Proportion of Variance of Explained (r^2) for Census Tract Gonorrhea Rates (n=155) Regressed on Different Independent Variables

	Coefficient (Standard Error)	r^2
Outlets per Square Mile		
Off-sale outlets	.582(.073)	.29
On-sale outlets	.300(.069)	.11
Total outlets	.476(.073)	.21
Outlets per Person		
Off-sale outlets	.374(.061)	.20
On-sale outlets	.201(.061)	.07
Total outlets	.488(.091)	.16

in a census tract, with a mean of four off-sale outlets) translates into a 14.5% higher gonorrhea rate, or 2.5 additional cases of gonorrhea.

The geographic association between gonorrhea cases and off-sale outlet density is illustrated in Figure 1. Areas where the outlet density is highest tend to be areas where the number of gonorrhea cases is the greatest.

To address the possibility that higher rates of gonorrhea among African Americans or underreporting of gonorrhea cases by private physicians could account for the relationships between outlet densities and gonorrhea rates, the covariates percent African

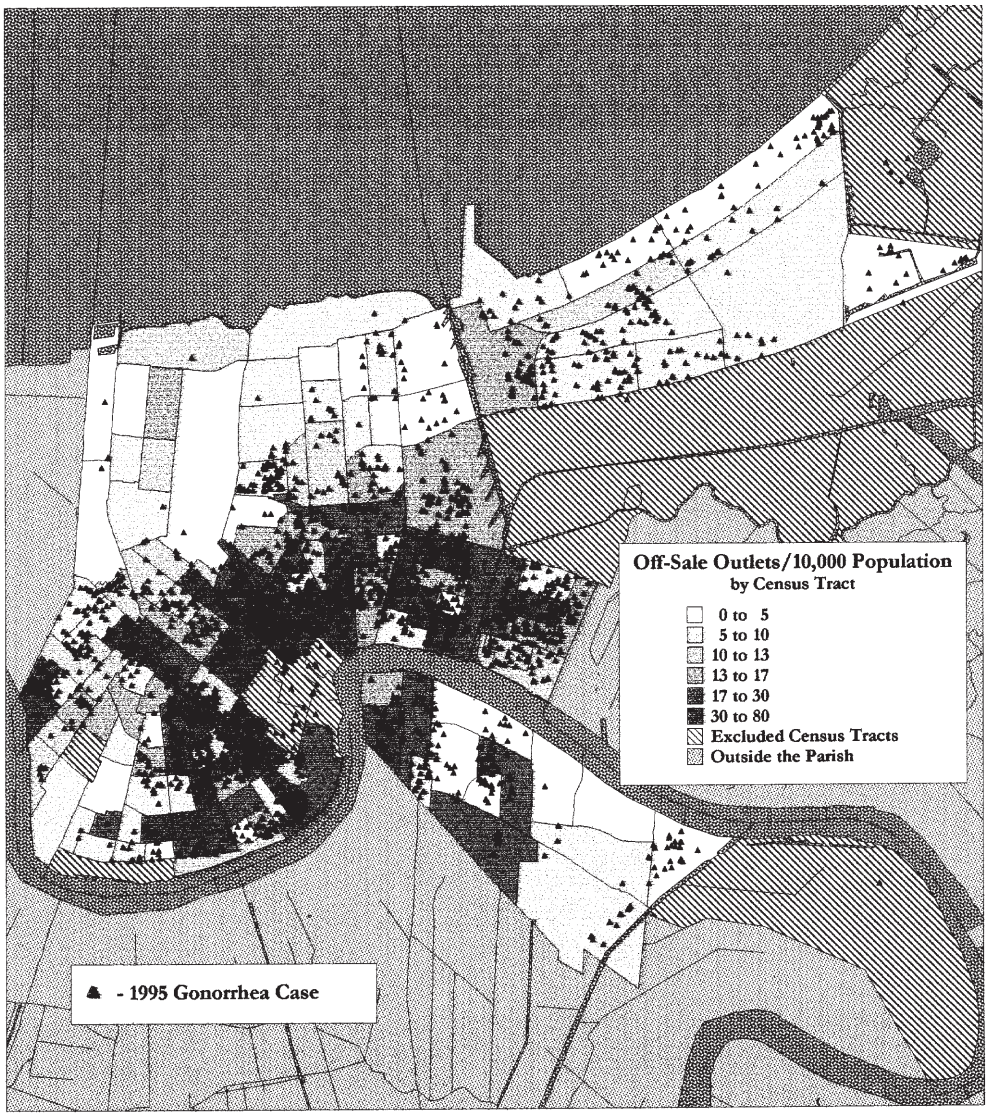


Figure 1 The 1995 census tract density of licensed off-sale alcohol outlets overlaid with reported cases of gonorrhea in New Orleans, LA, 1995.

American and percent unemployed were added to the model. Both percent African American (β=.445±.051) and percent unemployed (β=.789±.106) demonstrated strongly positive relationships with gonorrhea rates (Tables 3 and 4, Model 1). Adding these two covariates to the model increased the amount of variance explained to over 78%. The magnitude of the effect for each of the outlet density variables was reduced by the inclusion of these variables into the respective models. In every case, however, the effect of alcohol outlet density remained statistically significant. Outlet density measured as off-sale outlets per square mile demonstrated the strongest relationship, increasing the amount of additional variance explained by 3% (β=.192±.047) (Table 3, Model 2). The on-sale outlet density variables, on-sale outlet per square mile (β=.108±.036) and on-sale outlets per person (β=.102±.030), had the weakest relationships (Tables 3 and 4, Model 3). Interpreted as an elasticity, the relationship between off-sale outlet density and gonorrhea rates, controlling for covariates, indicates one additional off-sale outlet is associated with nearly one (.81) additional case of gonorrhea in the average New Orleans census tract during 1995.

Table 3 Coefficients (Standard Errors) for Regression Models in which the Dependent Variable is Census Tract Gonorrhea Rates and the Outlet Density Independent Variables Were Calculated as Outlets per Square Mile (n=155)

	Model 1	Model 2	Model 3	Model 4
Covariates				
Percent African American	.445(.051)*	.442(.048)*	.453(.049)*	.448(.048)*
Percent adults unemployed	.789(.106)*	.632(.107)*	.714(.106)*	.660(.105)*
Outlets per Square Mile				
Off-sale outlets		.192(.047)*		
On-sale outlets			.108(.036)*	
Total outlets				.171(.042)*
R^2	76	.79	.77	.78

* p<0.001

Table 4 Coefficients (Standard Errors) for Regression Models in which the Dependent Variable is Census Tract Gonorrhea and the Outlet Density Independent Variables were Calculated as Outlets per Person (n=155)

	Model 1	Model 2	Model 3	Model 4
Covariates				
Percent African American	.445(.051)*	.441(.049)*	.451(.054)*	.452(.048)*
Percent unemployed	.789(.106)*	.691(.105)*	.738(.160)*	.680(.102)*
Outlets per 1,000 Residents				
Off-sale outlets		.123(.034)*		
On-sale outlets			.102 (.030)	
Total outlets				.221 (.047)
R^2	.76	.78	.78	.79

* p<0.001

Discussion

The analysis demonstrates that a strong geographic relationship exists between alcohol outlet density and reported gonorrhea rates at the census tract level, and that the relationship was partially independent of the effects of racial composition and level of unemployment across census tracts. These findings confirm the findings of Rothenberg, who showed that cases of reported gonorrhea cluster geographically when analyzed across urban residential neighborhoods (17). They also indicate that the grouping of cases is, in part, predicted by alcohol outlet density. The implications of these findings are significant, both in terms of using alcohol outlet density as a possible geographic indicator of an STD core group (20), and in terms of supporting theories that postulate a role for alcohol in promoting high-risk sexual behavior (21–23).

The fact that both outlets per person or outlets per square mile were associated with gonorrhea rates indicates that both these variables are reliable measures of the effect of alcohol outlet density. It should be noted that in studies of the effect of alcohol outlet density on other alcohol-related outcomes, limiting the analysis to a particular type of neighborhood (i.e., urban residential census tracts) is responsible for this consistency (15).

The covariates of percent African American and percent unemployed were also strongly associated with gonorrhea rates, independent of each other. The effect of the percent African American variable on gonorrhea rates is consistent with a higher risk of gonorrhea infection evidenced in African-American populations across the country. It may be that the same effects of concentrated disadvantage that are believed to be responsible for the higher rates of crime and mortality in African-American communities are also operating with regard to high-risk sexual behavior (24). On the other hand, the percent unemployed variable was introduced to control for the fact that gonorrhea cases are more likely to be reported by public-sector physicians as compared with private-sector physicians. The role employment plays in obtaining health insurance is undoubtedly a factor in determining whether an individual seeks treatment in the public or private sector. Unemployment, however, is also a marker for lack of access to treatment services. Lack of access to treatment services results in longer periods of infection, which increases the risk of transmission. It is impossible to differentiate between these two possible explanations for the effect of unemployment.

It should be noted that the data are cross-sectional and therefore do not permit a determination of directionality for the effects described. The analysis does not differentiate between competing explanations of the relationship between alcohol availability and high-risk sexual behavior in terms of cause and effect. In addition, the study represents a group-level analysis. Interpreting these findings in terms of an individual-level explanation could result in an ecologically fallacious inference.

With these limitations in mind, there are a number of hypotheses regarding the relationship between alcohol use and high-risk sexual behavior that are consistent with these findings (22,23). The census tract-level design permits multi-level explanations for the relationship involving both individual- and group-level mechanisms. Individual explanatory models view alcohol use both as a causal factor and as a confounder. Causal explanatory models at the individual level view alcohol outlet density as a direct indicator of increased individual access to alcohol that results in greater alcohol consumption. It has been shown that alcohol affects judgement and has a disinhibiting effect on

socially learned restraints (25,26). Alternatively, alcohol use may merely serve as a marker for a risk-taking personality (21,22). Risk takers may engage in a host of high-risk behaviors, including both alcohol use and high-risk sexual behavior. This analysis does not differentiate between these two possibilities because both risk takers and drinkers could be clustered in the high outlet density neighborhoods.

Group-level explanatory models view the "wetness" of the neighborhood as a social context that affects individual behavior (21). A wet neighborhood can be viewed as a geographic area where the norms of residents regarding alcohol consumption and alcohol-related behaviors are more likely to be conducive to high-risk sexual behavior. The high-risk norms in these areas evolve over time as residents come to expect the high-risk behaviors associated with drinking that they are more likely to observe in their daily social interactions (14). Such an effect for alcohol outlet density on neighborhood norms would affect all residents of the neighborhood to varying degrees, as opposed to an individual-level effect that only affects the drinker.

In any case, the geographic relationship between alcohol outlet density and gonorrhea cases has implications for future research. Neighborhood-level factors, such as alcohol outlet density, need to be considered as risk factors in the design and evaluation of preventive interventions. Population surveys should be designed to stratify by neighborhood to account for local risk factors like alcohol outlet density. In addition, multi-level studies of individuals within neighborhoods need to be conducted to distinguish between individual- and neighborhood-level risk factors. An individual-level effect for alcohol outlet density would reveal that individual access to alcohol is the primary factor accounting for the relationship; a neighborhood-level effect would reveal an effect for alcohol outlet density independent of individual access to alcohol outlets. Again, a neighborhood-level effect could mean that shared norms of sexual behavior for individuals living in high outlet density neighborhoods have been liberalized as a result of living in the neighborhood for a period of time.

Finally, these findings indicate the need for intervention studies. A preventive intervention designed to reduce alcohol outlet density is the logical next step. An intervention study would account for potential unmeasured confounders and help to determine the directionality of a potential causal association. Moreover, in a number of cities across the country, community groups are working toward the goal of reducing alcohol outlet density in problem neighborhoods. For example, some Chicago residents are organizing to take advantage of a city ordinance that permits precincts to vote out alcohol sales in the name of cleaning up the neighborhood (27). Preventive interventions could be organized around these efforts and evaluated in terms of their effect on STD rates.

References

1. Caetano R, Hines AM. 1995. Alcohol, sexual practices, and risk of AIDS among blacks, Hispanics, and whites. *Journal of Acquired Immune Deficiency Syndromes* 10:554–61.
2. Leigh BC, Temple MT, Trocki KF. 1994. The relationship of alcohol use to sexual activity in a US national sample. *Social Science and Medicine* 39:1527–35.
3. Robinson JA, Plant MA. 1988. Alcohol, sex and risks of HIV infection. *Drug and Alcohol Dependence* 22:75–8.

4. Hingson RW, Strunin L, Berlin BM, Heeren T. 1990. Beliefs about AIDS, use of alcohol and drugs, and unprotected sex among Massachusetts adolescents. *American Journal of Public Health* 80:295–9.
5. Trocki KF, Leigh BC. 1991. Alcohol consumption and unsafe sex: A comparison of heterosexuals and homosexual men. *Journal of Acquired Immune Deficiency Syndromes* 4:981–6.
6. Gruenwald PJ, Ponicki WR, Holder HD. 1993. The relationship of outlet densities to alcohol consumption: A time series cross-sectional analysis. *Alcoholism, Clinical and Experimental Research* 17:38–46.
7. MacKinnon DP, Scribner RA, Taft KA. 1995. Development and application of a city-level alcohol availability and alcohol problems database. *Statistics in Medicine* 14:591–604.
8. Rabow J, Watts RK. 1982. Alcohol availability, alcoholic beverage sales and alcohol-related problems. *Journal of Studies on Alcohol* 43:767–801.
9. Scribner RA, MacKinnon DP, Dwyer JH. 1994. Alcohol outlet density and motor vehicle crashes in Los Angeles County cities. *Journal of Studies on Alcohol* 55:447–53.
10. Smart RG. 1977. The relationship of availability of alcohol beverages to per capita consumption and alcoholism rates. *Journal of Studies on Alcohol* 38:891–6.
11. Harford TC, Parker D, Paulter C, Wolz M. 1979. Relationship between number of on-premise outlets and alcoholism. *Journal of Studies on Alcohol* 40:1053–7.
12. Parker DA, Wolz M, Harford T. 1978. The prevention of alcoholism: An empirical report on the effects of outlet availability. *Alcoholism, Clinical and Experimental Research* 2:339–43.
13. Colon I. 1981. Alcohol availability and cirrhosis mortality rates by gender and race. *American Journal of Public Health* 71(12):1325–8.
14. Scribner RA, MacKinnon DP, Dwyer JH. 1995. The risk of assaultive violence and alcohol availability in Los Angeles County. *American Journal of Public Health* 85:335–40.
15. Scribner RA, Cohen DA, Kaplan S, Allen SH. [in press]. Alcohol availability and homicide in New Orleans: Conceptual considerations for small area analysis of the effect of alcohol outlet density. *Journal of Studies on Alcohol.*
16. May RM, Anderson RM. 1987. Transmission dynamics of HIV infection. *Nature* 326:137–42.
17. Rothenberg RB. 1983. The geography of gonorrhea. *American Journal of Epidemiology* 117:688–94.
18. Potterat JJ, Rothenberg RB, Woodhouse DE, Muth JB, Pratts CI, Fogle JS. 1985. Gonorrhea as a social disease. *Sexually Transmitted Diseases* 12:25–32.
19. Claritas Corporation. 1995. *Trendmap 95.* Ithaca, NY: Claritas Corporation.
20. Aral SO, Holmes KK, Padian NS, Cates W. 1996. Overview: Individual and population approaches to the epidemiology and prevention of sexually transmitted diseases and human immunodeficiency virus infection. *Journal of Infectious Diseases* 174(suppl):S127–33.
21. Biglan A, Metzler CW, Wirt R, Ary D, Noell J, Ochs L, French C, Hood D. 1990. Social and behavioral factors associated with high-risk sexual behavior among adolescents. *Journal of Behavioral Medicine* 13:245–61.
22. Stall R, McKusick L, Wiley J, Coates TJ, Ostrow DG. 1986. Alcohol and drug use during sexual activity and compliance with safe sex guidelines of AIDS: The AIDS behavioral research project. *Health Education Quarterly* 13(4):359–71.
23. Leigh BC. 1990. Alcohol and unsafe sex: An overview of research and theory. Progress in clinical and biological research. *Alcohol, Immunomodulation and AIDS,* Vol. 325. New York: Alan Liss, Inc. 35–46.

24. McCord C, Freeman HP. 1990. Excess mortality in Harlem. *New England Journal of Medicine* 322:173–7.

25. Crowe LC, George WH. 1989. Alcohol and human sexuality: Review and integration. *Psychological Bulletin* 105:374–86.

26. Steele CM, Josephs RA. 1990. Alcohol myopia: Its prized and dangerous effects. *American Psychologist* 45:921–3.

27. Novak T. 1998. City's liquor crackdown raises bar for taverns. *Chicago Sun-Times* 19 July. 4A.

Social and Demographic Analysis

The New Mexico Mammography Project: Using GIS to Determine Geographic Variation in Mammography Utilization

Andrew M Amir-Fazli,* Patricia M Stauber, Meg Adams-Cameron, Charles R Key
New Mexico Tumor Registry, Cancer Research and Treatment Center, University of New Mexico, Albuquerque, NM

Abstract

The New Mexico Mammography Project (NMMP) is establishing a population-based mammography registry for the state of New Mexico. One of the aims of this project is to examine the effectiveness of mammography screening in a community setting. In order for mammography screening to achieve the degree of reduction in mortality from breast cancer that has been demonstrated in randomized control trials, women at risk must undergo screening on a regular schedule. According to a previous analysis of over 135,000 women's records in the NMMP database, more than 60% of all women had at least one mammogram between 1992 and 1995, but only 22% had three exams in those four years. There are many reasons for lack of compliance. One that few studies have examined is the effect of travel distance to a mammography facility. Information available in the NMMP database enabled us to determine, for each examination, where the woman being examined lived and where the examining facility was. Over 80% of recent addresses were geocoded to the census tract level and 99% to the zip code level. Geographic information system (GIS) technology, specifically ArcView 3 software with the Network Analyst extension, was used to calculate the distances women drove to mammography facilities. Mammography screening rates for various areas of the state were then calculated and analyzed (using US census data) for differences by driving distance versus age, rural or urban status, education, and household income. The results of this analysis are not yet available; this paper presents a discussion of the issues involved and the problems encountered in using GIS methodology with registry data.

Keywords: health care access, mammography, driving distance, geocoding

Introduction

The New Mexico Mammography Project (NMMP) is establishing a population-based mammography registry for the state of New Mexico (1). One of the aims of this project is to examine the effectiveness of mammography screening in a community setting. In order for mammography screening to achieve the degree of reduction in mortality from breast cancer that has been demonstrated in randomized control trials, women at risk must undergo screening on a regular schedule. According to a previous analysis of over 135,000 women's records in the NMMP database, more than 60% of all women

* Andrew M Amir-Fazli, New Mexico Tumor Registry, Cancer Research and Treatment Center, University of New Mexico, 2325 Camino de Salud NE, Albuquerque, NM 87131 USA; (p) 505-272-8575; (f) 505-272-8572; E-mail: aamirf@nmtr.unm.edu

had at least one mammogram between 1992 and 1995, but only 22% had three exams in those four years. There are many reasons for lack of compliance. One that few studies have examined is the effect of travel distance to a mammography facility. In 1997, the New Mexico Tumor Registry (NMTR) at the University of New Mexico Cancer Research and Treatment Center began a SEER (Surveillance, Epidemiology, and End Results) Special Study for the National Cancer Institute, "Geographic Variation in Breast Cancer Treatment and Mammography Use." An objective of this study was to use geographic information system (GIS) technology to determine the distances from patients' residences to treatment or diagnostic facilities, and analyze how these distances affect treatment choice or service utilization.

Overview

To compute mammography utilization rates, NMTR selected records from the NMMP database for all women age 40 and older who had a mammogram in 1994 or 1995 at any of the screening facilities in the five-county area surrounding Albuquerque, the largest city in New Mexico. At the time of the study, this was the only area in New Mexico for which data on all mammograms (more than 95%) had been collected. Mammograms were grouped into series called mammographic events; an initial mammogram and immediate follow-up examinations (all examinations within 90 days) were considered one mammographic event. Information in the NMMP database was used to determine where the women lived at the time of their examinations and locate the mammography facilities where the examinations were performed. For most areas, over 80% of recent addresses were geocoded to the street address (census tract) level and 99% to the zip code level. Age-adjusted rates of mammography utilization (the number of mammograms per 100 women per year) were calculated. The population numbers for the denominators were taken from the 1990 US Census Bureau zip code files.

An initial statewide study, "Geographic Variation in Breast Cancer Treatment," used a smaller number of cases and facilities from the years 1994 and 1995 to develop and test GIS methods. GIS technology was used to calculate the distances women drove to radiation treatment centers and mammography facilities. Once methods were established, they were applied to the larger mammography database so that mammography screening rates could be calculated for various areas of the state. US census data were used to analyze the rates in these areas for differences by driving distance versus age, rural or urban status, education, and household income. The hypotheses for the study were that rates are lower in areas with the greatest distance to travel to receive a mammogram, rates are lower in rural areas than in urban areas, and rates are lower in areas with lower educational levels or lower household income. As of this writing, the final analysis phase of the study is not complete; this paper presents results of and problems arising from using GIS.

Using GIS

Methods

In order to allow analysis of geographic variation in the NMMP data collected, it was necessary to develop GIS capability at NMTR. The GIS was created specifically for two

main operations, geocoding and routing. Geocoding is the process of assigning an absolute location (in this case, latitude and longitude coordinates) to a geographic feature referenced by a relative location (such as a street address or zip code). Once facility and patient locations were determined by geocoding, a Euclidean (or straight line) distance could be calculated between patients and facilities and a measure of geographic variation, such as a distance to the closest facility, could be determined. A GIS, however, is capable of more sophisticated routing analysis, determining a shortest distance along available road networks. Because it is not possible in most cases to travel in a straight line in New Mexico, a GIS network analysis was used to compute driving distance (and to consider driving time) as a more realistic measure of geographic access.

The GIS installed at NMTR for this study was ArcView 3.0a (ESRI, Redlands, CA), with the Network Analyst extension. This software was available at low cost through the University of New Mexico site license and included the ESRI StreetMap product as a data resource. The system was installed on a generally available personal computer, a PC running Microsoft Windows 95. The PC was equipped with a Pentium Pro 233 MHz processor with 64 Mb RAM, an 8 Gb hard drive, and a 21-inch display. This configuration was adequate to the task, although some operations took several hours.

Geocoding

Geocoding involved many steps, including obtaining addresses of facilities and patients, standardizing addresses, obtaining a street reference file, configuring and running the geocoding process, and mapping the resulting locations. Where a patient's street address could not be successfully geocoded, the patient's zip code was used to determine a location of residence. To use zip codes required an additional step, assigning a point location for each zip code area. Geocoding by zip code allowed all patients to be assigned a location.

A file of 12 functioning radiation treatment facilities and their street addresses was created from NMTR's facility records and from consultation with radiation treatment personnel. All radiation treatment facilities were successfully geocoded to a street address. Addresses of mammography screening facilities were obtained from the Food and Drug Administration (FDA) Web site listing facilities certified by the Mammography Quality Standards Act. To ensure completeness, the FDA list was compared with the facility list from which patient records were obtained. A missing facility address for the Veterans' Administration Hospital/Kirtland Air Force Base was then added, making a total of 57 geocodable facilities. After mapping these locations, creating a five-county service area, and eliminating facilities located outside the service area, the number of facilities in the study was reduced to 21.

The addresses of radiation treatment cases were obtained from NMTR files, in which an address at time of diagnosis is routinely recorded in the course of entering cancer data. The addresses of mammography cases were obtained via records collected by the NMMP from screening facilities. In these mammography records, a site and subsite code were intended to identify the facility where the screening was performed. In the 1994–1995 patient records, the subsite code is currently missing for almost all cases (although it is believed this information can still be obtained). Without a subsite code, it cannot be known whether a patient record shows that the screening was performed in an outlying clinic and later read at a main facility, or whether the record shows that the screening was actually performed at a main facility. A distribution of the geocoded

1994–1995 mammography patient addresses shows that at least 16% of patient residences were outside the five-county service area defined for our mammography facilities. These records were excluded because it has yet to be determined whether these patients actually traveled great distances—away from closer mammography facilities—to use the facilities in the center of the state.

The geocoding component of a GIS is most successful when street addresses follow a standardized format. NMTR uses address-editing routines that conform to ArcView's expectations. A separate step was not required to reformat addresses recorded by NMTR. Records from mammography screening facilities varied in format quality. In the future, additional address-formatting software could be used to process these records. This may improve the geocoding match rates.

In order to geocode and perform routing analysis, a reference dataset of street locations and address ranges is required. This dataset is commonly called a street network file. The street network file used for this study was ESRI's StreetMap product, a dataset based on Geographic Data Technology's Dynamap/1000 dataset (GDT, Lebanon, NH), which in turn was constructed from the original US Census Bureau TIGER/Line files with enhancements and corrections. StreetMap was used to geocode facility and patient addresses directly. To be used as a street network file, portions of StreetMap were converted to the "shapefile" format used by ArcView. The reduced size of these StreetMap "regions" limited the area to be searched during network analysis and sped up processing on the PC. The large distances between radiation facilities made it possible to break up the state of New Mexico into separate StreetMap regions. These regions were overlaid on a map that contained zip code boundaries and major roads, and an analysis was performed on the zip codes that fell along the boundary of a region. These zip codes were then assigned to the region that had the nearest facility. For the five-county mammogram study, a region was constructed to cover the five counties and any outlying facilities that were near enough to the five-county boundary that they might be candidates for the nearest-facility calculation.

Late in the project time frame, TIGER/Line97 data became available for New Mexico. An evaluation was made to see if TIGER/Line97 would prove a better street network file resource than StreetMap. A software utility, TGR2SHP (GIS Tools, Knoxville, TN), was obtained and used to convert TIGER/Line97 files to a format compatible with the ArcView GIS software. The converted TIGER/Line97 files were then used to rerun both geocoding and street network conversion. A spot check of certain regions of the state indicated that the road classification problems (described below) in the StreetMap product were largely corrected. This indicated that the use of TIGER/Line97 could produce more realistic driving distances and could enable the use of driving time analysis. With properly classified roads, each type of road could be assigned an average speed limit and a time expended to traverse each road segment could be calculated and summed.

Geocoding with Tiger/Line97, however, produced significantly fewer matches and did not improve earlier methods. It appears that although TIGER/Line97 contains more recent information than the StreetMap product, which is based on earlier TIGER files, there are fewer total address ranges in TIGER/Line97's underlying database, resulting in a lower address match rate.

In New Mexico, many people, especially those living outside the main urban centers, use post office boxes, local road names, and rural route addresses. It was

recognized that not all patient records would have street addresses that would geocode. A secondary marker for location was needed, so all records had zip codes assigned. In order to geocode to a zip code, a reference dataset of zip codes was obtained. Two types of zip code datasets were originally available: a point coverage and a polygon coverage. These datasets were bundled along with StreetMap and were derived from information matching the census in the early 1990s. In the point coverage file were the zip code number, the associated postal name, a code indicating a type of zip code (whether an area or single site such as a PO box, office building, or entity such as a university), and an area in square miles. When the zip code points and polygons were overlaid on a map of New Mexico, the zip code points were found in most cases to be located at the geographic centers of each corresponding zip code polygon (area). A concern of the study was that certain zip codes in New Mexico cover very large geographical areas; for example, Roswell's zip area is over 5,000 square miles. In these cases, zip code would be a poor proxy for the location of a patient's actual residence.

When it was observed that many zip code points located at the centroid of a zip code area were positioned in roadless and uninhabited areas and far from the true population centers of the area, it was decided to use another zip code point file. The new file was obtained from the US Census Bureau LandView III system and contained zip code points established by the US Postal Service as of January 1, 1997. It was noted from a comparison of this file with the previous zip code file that the assignment of zip code numbers had not changed. The new zip code points represented, for the most part, actual post office locations (which as a general rule are close to the population centers of zip code areas). A plot of these new points overlaid on the older zip code points and New Mexico population centers showed a closer match between the new zip code points and New Mexico population centers. The new zip code file, however, assigned a number of less-used zip code points to the centroids of the counties. In the statewide radiation treatment study, any zip code point found in the patient data that was outside the mapped polygon boundary of the zip code was moved from the centroid of its county to the nearest major road segment within its proper zip code boundary. This was not done for the five-county screening study, although the newer zip code point file was used. A review of the five-county region map showed that very few of these county centroid points would have been moved any significant distance.

ArcView's geocoding function tags each address record processed as "matched" if ArcView can locate the address along a road segment in the street network file. All radiation treatment and mammography screening facilities were successfully matched to a street address, in some cases after the street address was corrected. Seventy-one percent (71%) of patient addresses for the radiation treatment study were geocoded to the street address. The original geocoding run had produced a match rate of only 66%. Unmatched records, however, were reprocessed interactively with the opportunity to make corrections and rerun the geocoding process. Corrections were made of obvious spelling errors, street numbers that extended existing ranges but did not cross zip code boundaries, non-standard address formats, and street suffixes that were different but still in the same zip code boundary. This type of "reject" processing was not done for the much larger file of mammography patient addresses.

Seventy-three percent (73%) of patient addresses for the mammography study were geocoded to the street address. Due to the number of missing address entries and the uneven quality of address information reported by mammography facilities, additional

address fields matched from the State of New Mexico Motor Vehicle Department (MVD) were appended to the patient records. Use of the MVD address fields consistently showed a higher geocoding match rate and was recommended over use of the original address fields.

Match rates for the mammography patient records were also increased by relaxing the matching sensitivity parameters of the ArcView geocoding process. This caused many "partial" matches to be assigned a map location and was similar in effect to manual corrections performed by interactive reprocessing of unmatched records. The lower-sensitivity settings accommodated errors such as spelling variations and address prefix or suffix differences (e.g., "Road" instead of "Drive"). Because zip code was used as a restriction, it was feasible to match additional records with the assurance that the resulting location would remain in the same zip code.

A concern of the study was the large discrepancy in geocoding rates between "urban" and "rural" counties. The two urbanized counties in the five-county study area had geocoding rates of 86% and 74%, while the three rural counties had rates of 56%, 49%, and 3%.

Whether or not geocoding to the street address was successful, all patient records were matched to mapped zip code points, so that these records could also be assigned driving distance values calculated from the zip code points.

Routing and Distance Calculation

Once facilities and patients and/or their associated zip code points were mapped, the shortest routes between facilities and patients were determined. For this analysis, Arcview's Network Analyst extension was used, specifically the FindClosestFac(ility) function. Because many possible routes can be mapped between patient and facility, the Network Analyst used a generally accepted heuristic algorithm to determine the shortest route. Driving distance was then calculated by summing the lengths of all road segments along the shortest path from the patient location to the facility location.

By default, ArcView reports the lengths of line segments in units of decimal degrees, which are not really units of linear measure. It was observed that summing these line lengths and using the UNITS.CONVERT function to change the total length into units of miles produced incorrect results. Therefore, a MILES field was added to any street network file used for reference and an Avenue (ArcView's programming language) script was modified to compute the proper length in miles for each road segment. Within this script, the function UNITS.CONVERTDECIMALDEGREES was called to correctly calculate a "great circle arc length" between the starting and ending point of each line segment. This is the proper calculation for line length on a latitude/longitude grid. This MILES field then was identified as the "cost" field to be used by Network Analyst in reporting the results of a shortest path analysis. The cost field is an additional value that can be summed and reported by Network Analyst.

An alternative considered was to calculate driving time. If the cost field could be recorded in terms of hours or minutes, then a driving time would be reported instead of a driving distance. This would have been useful, because the shortest distance computed is not always the shortest time. Many roads in New Mexico are barely passable at low speeds; while these roads may be direct paths, they are rarely driven, because a circuitous route via state or county roads is much faster. The Network Analyst function does not directly distinguish between fast and slow routes. When using an associated

cost field, however, the same effect can be achieved if there is a way of classifying roads in the street network file.

The line segment record in the data table associated with StreetMap retains the TIGER/Line census feature classification code, which serves as a road type attribute. The code scheme is as follows:

- A00 (found in the data but not a specified code)
- A11 through A18: primary roads and major highways
- A20 through A28: secondary roads and minor highways
- A30 through A38: connecting and county roads
- A40 through A48: city and neighborhood streets
- A50 through A73: service roads, 4WD trails, etc.

Computing different "costs" by road type would have allowed a driving time analysis. When the road network was mapped, however, it became evident that many thousands of unusable road segments were improperly classified as significant roads. A query and extract process selecting for the roads of types A11 through A38 was performed on the street network file created for each region. This process did filter out most non-drivable routes, and did leave intact the major routes.

Using the "filtered" street network file, a first pass was made to calculate paths from each zip code point in a region to the nearest facility. (Some zip code points were not located near enough to a line segment in the street network. These zip code points were moved small distances to the nearest line segment.) The resulting routes were checked individually by running Network Analyst in an interactive mode. This process created a driving distance value for each patient in each zip code area and allowed a visual check of the Network Analyst choice of paths. Several "broken" major routes were identified in this manner and repaired.

A limitation of Network Analyst as delivered is that it only computes one solution at a time from a user-selected menu option. This is impractical for studying thousands of cases. To remedy this, an Avenue script was written to "batch process" the cases in each region. This script identified, via a user input dialog, the street network, a file of "events" (in this case, either zip code points or street address points) and a key field identifier for each event, a file of facilities and associated key fields, and an output file in which to store results. The script then cycled through each event, performing the FindClosestFac function. At each cycle, an output record was stored indicating the event key, facility key, and minimum distance in cost units. John Fortney, Kathryn Rost, and James Warren of the University of Arkansas for Medical Sciences pioneered this approach and provided suggestions for our study (2).

The batch Network Analyst script was then used to process treatment patient records region by region. The results were appended to the treatment database for statistical analysis. Because both a zip code-derived distance and a street address-derived distance were stored with each patient record, a comparison of the two distances was made. A strong correlation (R=.97132) suggested that zip code distances might reasonably substitute for street address distances when processing the larger mammography records file. One hundred seven thousand eight hundred thirty-four (107,834) mammographic events were processed for the five-county study area for the two years studied. Zip code-derived distances were then also computed and assigned to each event record as a measure of quality control.

Further Study

This study will be expanded in future years, as more complete mammography records become available for more of New Mexico. It is believed that the effect of distance to treatment and screening facility on mammography utilization will be more pronounced and better understood as larger areas are studied. An alternate approach to individual route calculations is being developed to process the large numbers of mammography records more efficiently. Service areas will be computed from treatment and screening centers, and mammography events will be assigned driving distances and times from GIS overlay functions. This approach should be possible after analysis of the current routing results establishes distance classifications that can be used to size service areas.

Additional improvements could be made to the GIS methods in the study in both geocoding and routing. Further address cleanup and comparison with MVD records would increase the geocoding match rate. Better protocols are needed to handle non-standard addresses such as rural routes and post office boxes. Assigning average travel speeds to a better road network would allow more realistic route choices and the ability to determine driving times instead of just driving distances. Additional GIS studies comparing distance to actual facility used with distance to closest facility would refine our measures of geographic variation in mammography utilization.

Acknowledgments

This research was supported by the National Cancer Institute SEER Special Study N01-PC-67007 and Breast Cancer Surveillance Consortium Project U01-CA-69976.

References

1. Rosenberg RD, Lando JF, Hunt WC, Darling RR, Williamson MR, Linver MN, Gilliland FD, Key CR. 1996. The New Mexico Mammography Project: Screening mammography performance in Albuquerque, New Mexico, 1991 to 1993. *Cancer* 78(8):1731–9.
2. Fortney J, Rost K, Warren J. *Comparing alternative methods of measuring geographic access to health services.* Working paper.

Population-Based Prevalence of Cocaine in Newborn Infants—Georgia, 1994

Mary D Brantley (1),* Roger W Rochat (2), Cynthia D Ferre (1), M Louise Martin (1), L Omar Henderson (1), W Harry Hannon (1), Brian J Ziegler (3), Paul M Fernhoff (3), Lori M Mayer (3), Elizabeth A Franko (2), Virginia D Floyd (2), Eric J Sampson (1), David J Erickson (1)

(1) Centers for Disease Control and Prevention, Atlanta, GA; (2) Georgia Division of Public Health, Atlanta, GA; (3) Georgia Chapter of the March of Dimes Birth Defects Foundation, Atlanta, GA

Note: This paper is an outline of a poster presentation made at the 1998 GIS in Public Health Conference, Phoenix, Arizona. For other reports about the study discussed in this paper, see the Centers for Disease Control and Prevention's Morbidity and Mortality Weekly Report (MMWR) of October 18, 1996 (1) at http://www.cdc.gov/epo/mmwr/preview/mmwrhtml/00044121.htm, or the Georgia Epidemiology Report of February 1997 (2) at http://www.ph.dhr.state.ga.us/epi/manuals/ger/feb97ger.pdf (adapted from the MMWR article).

Abstract

In 1994, the Georgia March of Dimes Birth Defects Foundation, the Georgia Department of Human Resources, and the federal Centers for Disease Control and Prevention collaborated to assess the feasibility of using dried blood spots (DBSs), routinely collected from newborn infants for metabolic disease screening, to conduct low-cost population-based screening to ascertain the prevalence of cocaine exposure. This is the first known application of population-based screening of newborns to detect exposure to cocaine. Georgia birth certificate records were electronically linked to metabolic records using probabilistic linkage. Maternal and infant characteristics associated with increased infant mortality were kept for analysis, and personal identifiers and linkage information were removed. The cocaine testing was performed on blinded DBSs using a modified radioimmunoassay to screen for benzoylecognine (BE). Positives were confirmed by mass spectrometry. The analysis file includes 17,230 infants born during a two-month period; of these infants, 91% had a DBS. Infants who were older than seven days, who had had less than 31 weeks of gestation, or whose birth weight was less than 1,500 grams were excluded. Specimens for 73 infants (4.7 per 1,000 statewide) tested positive for BE. Maternal characteristics associated with increased rates of BE in infants include greater age, three or more previous live births, cigarette smoking or alcohol drinking by the mother during pregnancy, inadequate weight gain during pregnancy, no father's name on birth certificate, black race, a short interpregnancy interval, education less than three years of college, late or no prenatal care with fewer visits, and delivery in perinatal centers or hospitals with no obstetric services. Six infants in the study were not delivered at a hospital. Mothers of BE-positive infants resided in 16 of the 19 health districts in Georgia, but 45% resided in a particular district. Antepartum cocaine

* Mary D Brantley, Centers for Disease Control and Prevention, 4770 Buford Hwy. NE, Mail Stop K-22, Atlanta, GA 30341 USA; (p) 770-488-5227; (f) 770-488-5240; E-mail: mdb4@cdc.gov

exposure is found in newborn infants throughout Georgia and in diverse population groups. This methodology is suitable for screening large populations and can be adapted for use with other substances of abuse.

Keywords: cocaine, newborns, zip code, surveillance, pregnancy

Background

A political history of the study of cocaine in newborns, and this study's time line:

- Late 1980s: Perceived cocaine epidemic
- 1990: Infant Health in Georgia report
- 1990: Georgia General Assembly Conference on Children of Cocaine
- 1990–1993: Protocol and prototype study
- 1994: Specimen collection and test for cocaine
- 1995: Data analysis and report writing
- 1996: Publication—*Morbidity and Mortality Weekly Report,* October 18, 1996 (1)
- 1996: Ground breaking for the March of Dimes Home of New Beginnings

Public policy recommendations from the Georgia General Assembly Conference on Children of Cocaine:

- Base public health policy on valid research
- Legislate comprehensive, holistic approach to control of substance abuse crisis
- Declare moratorium on legislation that could prosecute drug-dependent pregnant women
- General Assembly should develop and fund appropriate substance abuse treatment facilities for pregnant women

Methods

Research goals:

- Determine prevalence of cocaine-positive infants
- Determine geographic distribution
- Determine if cocaine-positive mothers get health care
- Identify possible interventions for mothers
- Define the methodology: strengths, weaknesses, biases

The analysis files, designed to assure anonymity of mothers:

- "Selection file" to study biases—no lab test results
- 21 analysis files with lab test results
- No file with more than four maternal characteristics
- No category with fewer than 50 infants

Strengths of the method:

- Population-based
- Large sample
- Low cost per individual tested
- Linked data to define birth cohort and to improve quality of data on maternal characteristics

Weaknesses of the method:

- Requires high-quality laboratory
- Bias against including premature infants
- Requires linkage to birth file for good epidemiology
- Fear of prosecution mandates extreme concern for anonymity
- Anonymity requirements limit datafiles for careful definition of maternal characteristics

Results

See Tables 1 through 5 and Figures 1 through 3.

Table 1 Rate of Maternal Cocaine Use by Selected Socio-Demographic Characteristics

Maternal and Child Characteristics	Sample Size	Number Positive	Rate per 1,000	Poisson 95% CI	
				Low	High
Age group (years)					
<20	2,975	2	0.7	0.1	2.4
20–24	4,168	15	3.6	2.0	5.9
25–29	3,921	34	8.7	6.0	12.1
30 or over	3,903	22	5.6	3.5	8.5
Missing	1	0	*	*	*
Education (years)					
<12	3,449	25	7.2	4.7	10.7
12	5,406	34	6.3	4.4	8.8
13–14	2,482	12	4.8	2.5	8.4
15 or more	3,511	2	0.6	0.1	2.1
Missing	120	0	0.0	0.0	30.7
Race/Ethnicity					
Other	287	0	0.0	0.0	12.9
White Hispanic	491	0	0.0	0.0	7.5
White non-Hispanic	9,139	12	1.3	0.7	2.3
Black	5,049	61	12.1	9.2	15.5
Missing	2	0	*	*	*
Urban Residence					
Large MSA/city	2,766	39	14.1	10.0	19.3
Small MSA/city	1,643	9	5.5	2.5	10.4
Non-MSA/city	1,051	6	5.7	2.1	12.4
Large MSA/non-city	4,705	9	1.9	0.9	3.6
Small MSA/non-city	1,360	3	2.2	0.5	6.4
Non-MSA/non-city	3,442	7	2.0	0.8	4.2
Missing	1	0	*	*	*
Marital Status					
Married	9,851	19	1.9	1.2	3.0
Not married	5,116	54	10.6	7.9	13.8
Missing	1	0	*	*	*
Total	14,968	73	4.9	—	—

* = sample size less than 5

— = not applicable

CI = confidence interval

MSA = metropolitan statistical area

Table 2 Rate of Maternal Cocaine Use by Selected Health Care Information

Health Care Characteristics	Sample Size	Number Positive	Rate per 1,000	Poisson 95% CI	
				Low	High
Adequacy of prenatal care					
None	167	15	89.8	50.3	148.1
Inadequate	1,702	27	15.9	10.5	23.1
Intermediate	2,236	3	1.3	0.3	3.9
Adequate	6,780	12	1.8	0.9	3.1
Adequate plus	3,920	12	3.1	1.6	5.3
Missing	163	4	24.5	6.7	62.8
Perinatal hospital service level					
None	149	6	40.3	14.8	87.6
Minimal (Level I)	2,817	10	3.5	1.7	6.5
Intermediate (Level II)	4,844	12	2.5	1.3	4.3
Specialized (Level III)	4,649	16	3.4	2.0	5.6
Regional perinatal center	2,503	29	11.6	7.8	16.6
Missing	6	0	—	—	—
Teaching hospital					
Yes	3,719	36	9.7	6.8	13.4
No	11,243	37	3.3	2.3	4.5
Missing	6	0	—	—	—
Trimester prenatal care began					
None	167	15	89.8	50.3	148.1
First (1–3 months)	12,080	25	2.1	1.3	3.1
Second (4–6 months)	2,139	21	9.8	6.1	15.0
Third (after 6 months)	447	8	17.9	7.7	35.3
Missing	135	4	29.6	8.1	75.9
Total	14,968	73	4.9	—	—

— = not applicable

CI = confidence interval

Table 3 Rate of Maternal Cocaine Use by Selected Maternal Risk Factors

Maternal Risk Factor	Sample Size	Number Positive	Rate per 1,000	Poisson 95% CI Low	Poisson 95% CI High
Smoking tobacco and/or drinking alcohol during pregnancy					
Both	106	13	122.6	65.3	209.7
Tobacco only	1,584	28	17.7	11.7	25.5
Alcohol only	111	3	27.0	5.6	79.0
Neither	13,117	29	2.2	1.5	3.2
Missing	50	0	—	—	—
Weight gain during pregnancy (pounds)					
Less than 15	996	13	13.1	6.9	22.3
15–24	3,001	18	6.0	3.6	9.5
25 or more	9,995	35	3.5	2.4	4.9
Missing	1,016	7	6.9	2.8	14.2
Previous births					
None	6,520	6	0.9	0.3	2.0
1	5,015	14	2.8	1.5	4.7
2	2,262	16	7.1	4.0	11.5
3 or more	1,171	37	31.6	22.2	43.6
Interpregnancy interval (months)					
No previous birth	6,520	6	0.9	0.3	2.0
0–6	675	15	22.2	12.4	36.7
7 or more	7,542	44	5.8	4.2	7.8
Unknown	231	8	34.6	15.0	68.2
Father's name present on birth certificate					
Yes	12,360	32	2.6	1.8	3.7
No	2,608	41	15.7	11.3	21.3
Total	14,968	73	4.9	—	—

— = not applicable

CI = confidence interval

Table 4 Relative Risk of Maternal Cocaine Use by Selected Maternal Risk Factors

Maternal Risk Factor	Sample Size	Number Positive	Rate per 1,000	Poisson 95% CI Low	Poisson 95% CI High
Smoking tobacco and/or drinking alcohol during pregnancy					
Both	106	13	55.5	38.8	79.4
Tobacco only	1,584	28	8.0	5.2	12.4
Alcohol only	11	3	12.2	4.8	30.9
Neither	13,117	29	1.0	+	+
Missing	50	0	—	—	—
Father's name present and marital status					
Neither	2,437	34	11.3	6.7	18.9
Married only	170	7	33.2	18.5	59.5
Father's name only	2,679	20	6.0	3.2	11.3
Both	9,681	12	1.0	+	+
Missing	1	0	—	—	—
Age at conception and parity					
25 or older and 2 or more children	2,348	43	15.8	8.8	28.6
Under 25 and 2 or more children	1,085	10	8.0	3.5	17.9
25 or older and 1 or no children	5,476	13	2.1	0.8	5.0
Under 25 and 1 or no children	6,058	7	1.0	+	+
Adequacy of prenatal care and parity					
Inadequate and 2 or more children	671	36	54.6	37.1	80.3
Inadequate and 1 or no children	1,361	10	7.5	3.6	15.7
Adequate and 2 or more children	2,762	17	6.3	3.2	12.4
Adequate and 1 or no children	10,174	10	1.0	+	+
Gestational age and weight gain during pregnancy					
Preterm infant and <25 pounds gained	877	23	10.0	6.3	16.0
Preterm infant and 25+ pounds gained	1,145	12	4.0	2.1	7.6
Term infant and <25 pounds gained	4,136	15	1.4	0.7	2.7
Term infant and 25+ pounds gained	8,810	23	1.0	+	+
Total	14,968	73	4.9	—	—

— = not applicable

+ = reference value

CI = confidence interval

Table 5 Rate of Maternal Cocaine Use by Selected Pregnancy Outcomes

Maternal and Child Characteristics	Sample Size	Number Positive	Rate per 1,000	Poisson 95% CI	
				Low	High
Birth weight category (grams)					
Normal (2,500 and over)	14,256	57	4.0	3.0	5.2
Low (1,500–2,499)	703	16	22.8	13.0	37.0
Gestational age (weeks)					
38 or more	12,9426	38	2.9	2.1	4.0
32–37	2,003	29	14.5	9.7	20.8
Missing	19	6	315.8	115.9	687.3
Total	14,968	73	4.9	—	—

— = not applicable

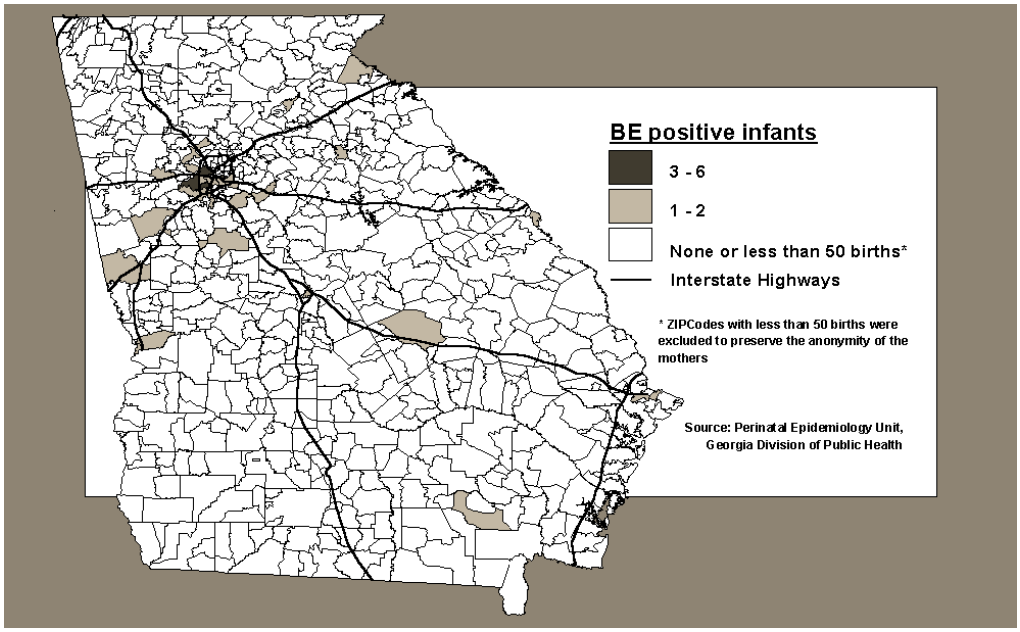


Figure 1 Geographic distribution of cases of benzoylecognine (BE)—a cocaine metabolite—in newborn infants, by zip code of mother's residence; Georgia, February 22 through April 23, 1994.

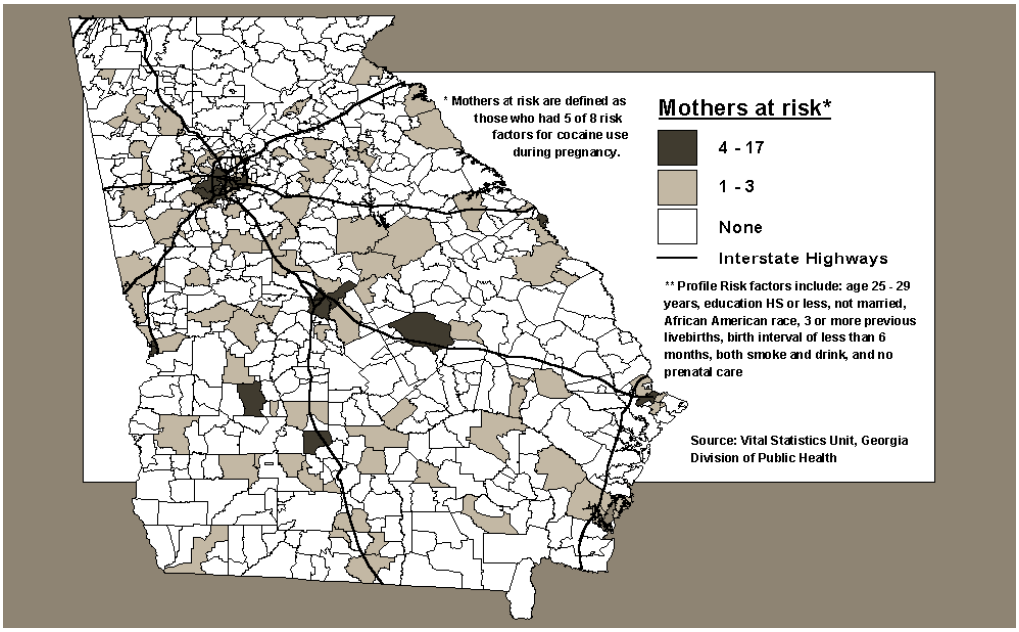


Figure 2 Areas with mothers at high risk for using cocaine during pregnancy (as defined by a risk profile** developed from the study), by zip code of mother's residence; Georgia, February 22 through April 23, 1994.

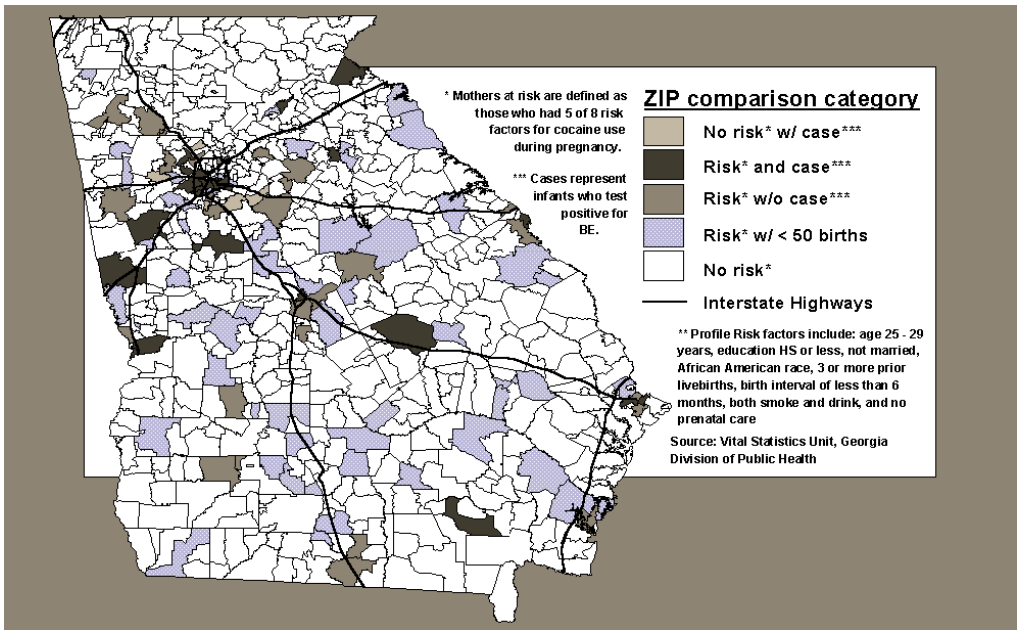


Figure 3 Comparison of high-risk zip codes identified by a risk profile developed from the study and zip codes with confirmed cases of infants testing positive for BE, by zip code of mother's residence; Georgia, February 22 through April 23, 1994.

Conclusions

Highlights of the study:

- Higher rates observed among users of other substances (tobacco and alcohol)
- Higher rates observed in women 25 and older, and in women who have had children before
- Higher rates observed in the city; however, events were observed throughout Georgia
- Higher rates received late or no prenatal care; however, three-quarters of mothers of BE-positive infants receive some prenatal care, and one-third receive care in the first trimester
- Some women deliver outside a hospital environment

A Georgia perspective on infant health and maternal behaviors during pregnancy that can adversely affect the fetus:

- Infant health
  - — 1 in 50 have very low birth weight (less than 3 pounds, 5 ounces)
  - — 1 in 50 have a major birth defect
  - — 1 in 200 have perinatal exposure to cocaine
  - — 1 in 500 test positive for HIV (maternal antibody); an estimated 15% to 20% of these will develop AIDS
  - — 1 in 4,000 are born with fetal alcohol syndrome

See Figures 4 through 7.

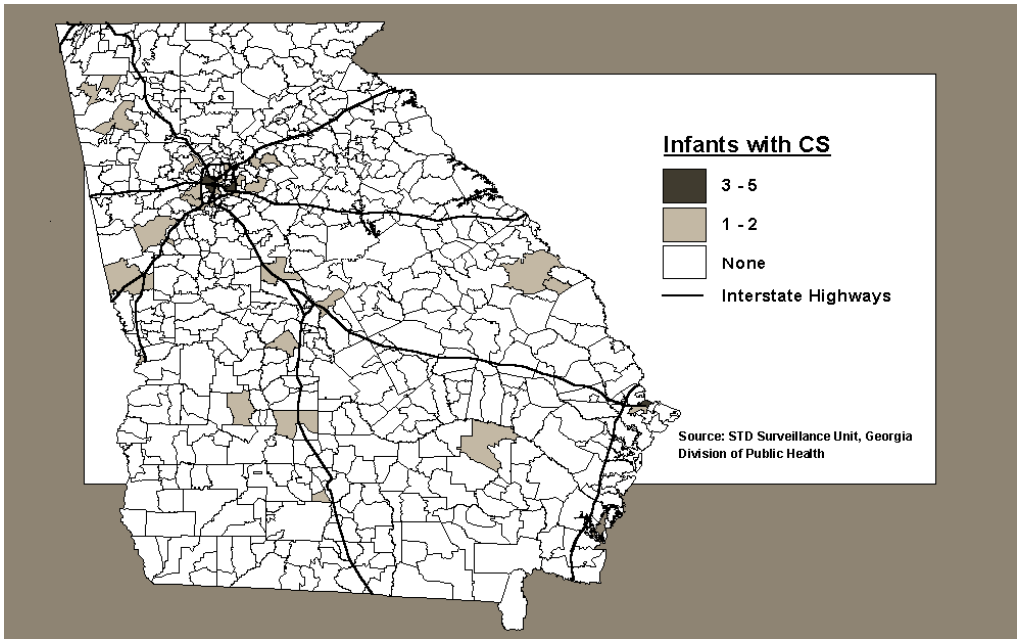


Figure 4 Geographic distribution of cases of congenital syphilis (CS) in newborn infants, by zip code of mother's residence; Georgia, 1994.

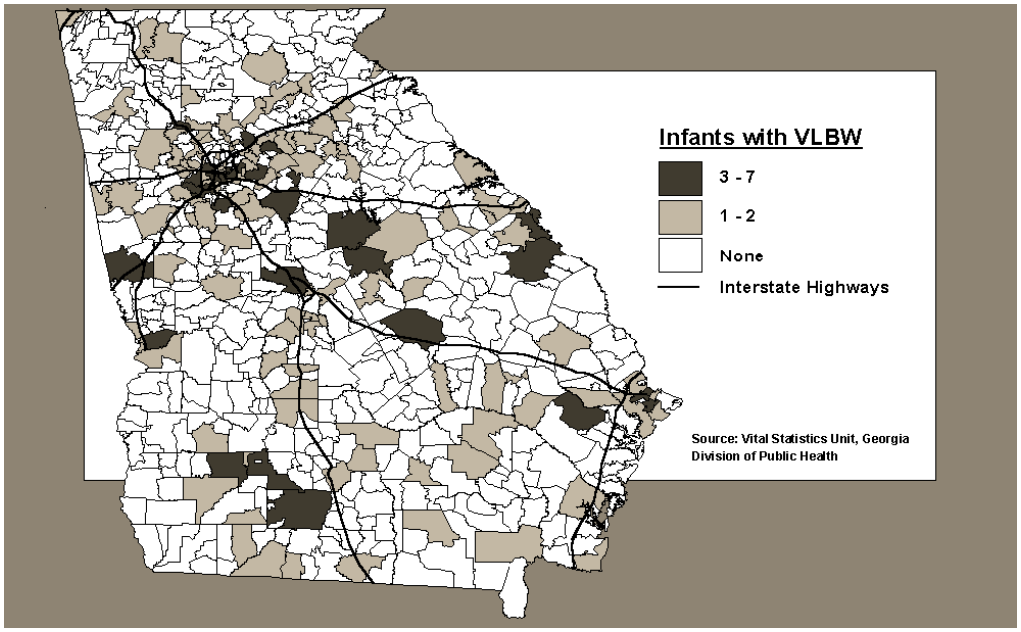


Figure 5 Geographic distribution of newborn infants with very low birth weight (VLBW), by zip code of mother's residence; Georgia, February 22 through April 23, 1994.

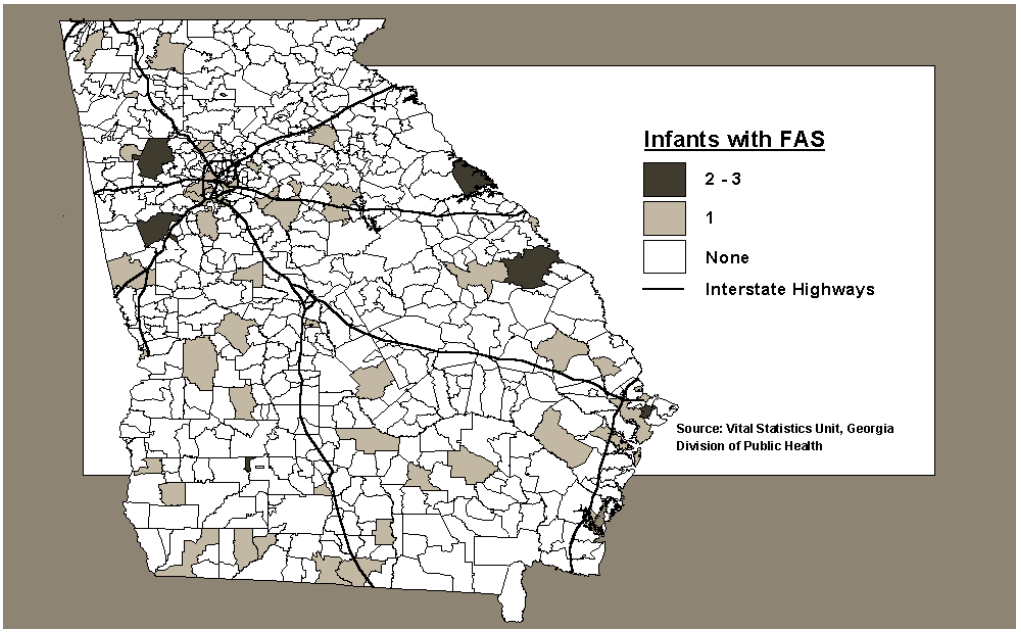


Figure 6 Geographic distribution of fetal alcohol syndrome (FAS) in newborn infants, by zip code of mother's residence; Georgia, 1990 through 1994.

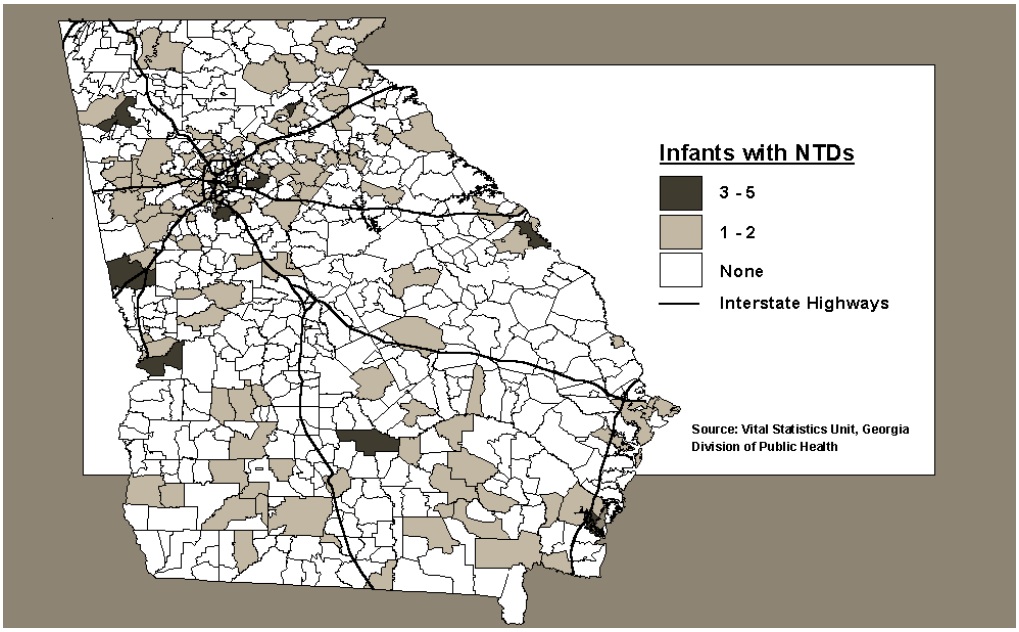


Figure 7 Geographic distribution of neural tube defects (NTDs) in newborn infants, by zip code of mother's residence; Georgia, 1990 through 1994.

- Maternal behaviors
  — 1 in 2 are not taking multivitamins with folate during pregnancy
  — 1 in 6 smoke tobacco during pregnancy
  — 1 in 10 consume alcohol during pregnancy
  — 1 in 100 receive no prenatal care
  — 1 in 200 use cocaine during pregnancy

See Figures 8 through 10.

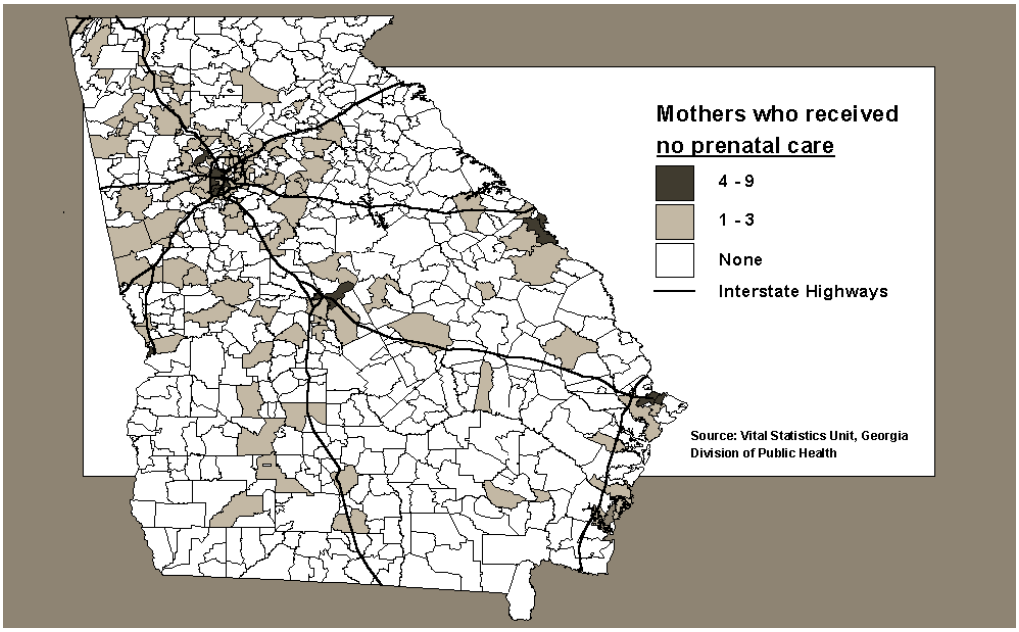


Figure 8 Geographic distribution of mothers who received no prenatal care during pregnancy, by zip code of mother's residence; Georgia, February 22 through April 23, 1994.

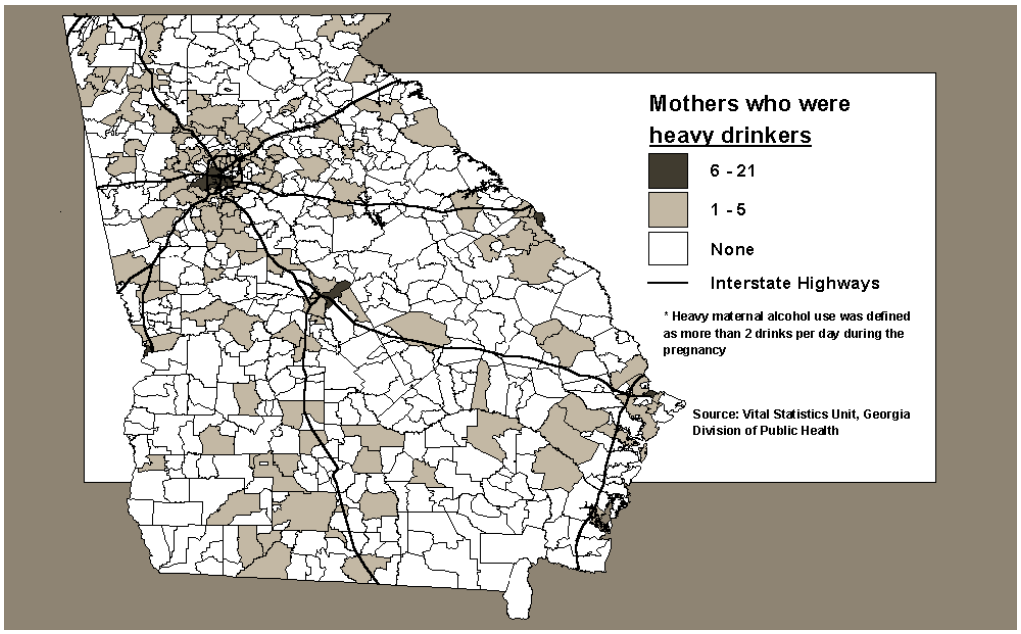


Figure 9 Geographic distribution of heavy maternal alcohol use* during pregnancy, by zip code of mother's residence; Georgia, 1990 through 1994.

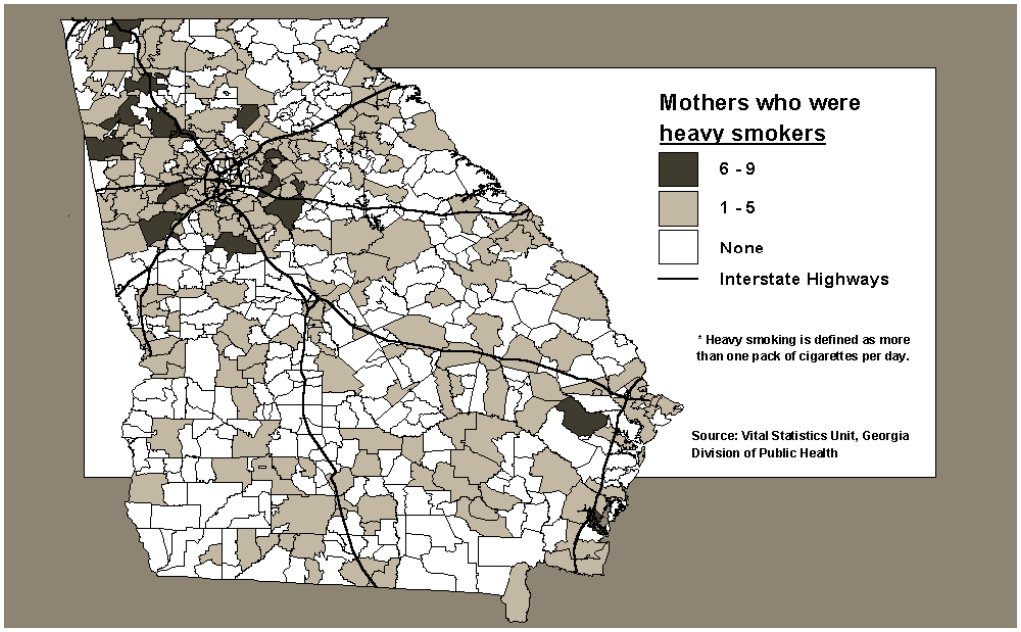


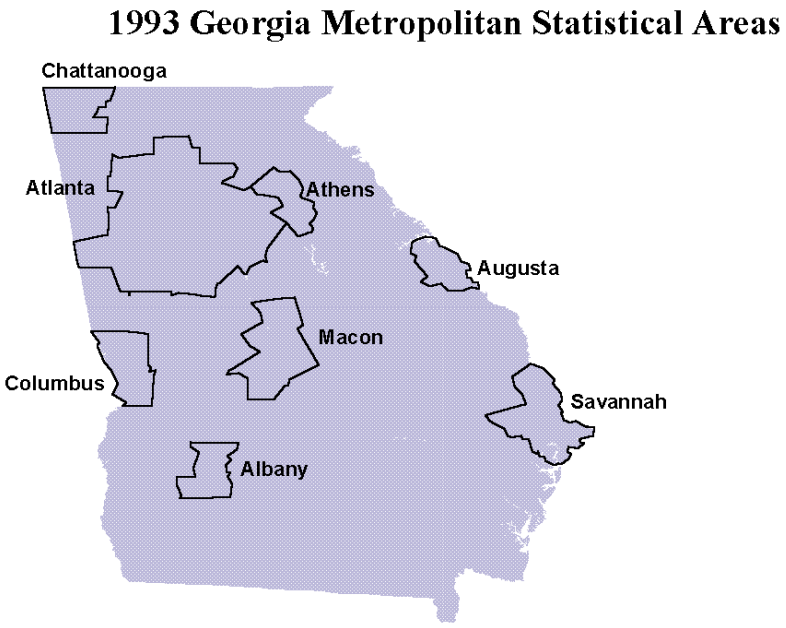
Figure 10 Geographic distribution of heavy maternal smoking* during pregnancy, by zip code of mother's residence; Georgia, February 22 through April 23, 1994.

Acknowledgments

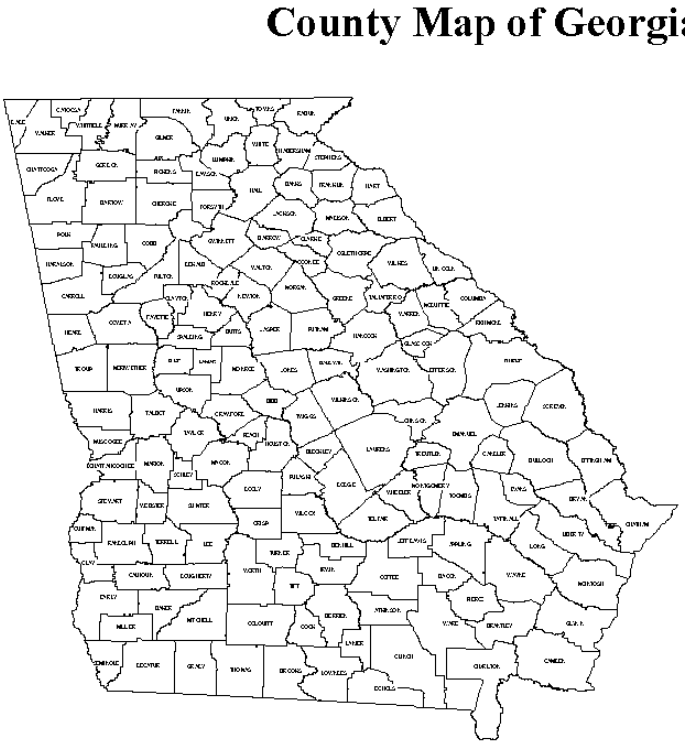
Funding for this project was provided by the Georgia Chapter of the March of Dimes Birth Defects Foundation.

References

1. Centers for Disease Control and Prevention. 1996. Population-based prevalence of perinatal exposure to cocaine—Georgia, 1994. *Morbidity and Mortality Weekly Report* 45(41):887–91. October 18.
2. Rochat R, Brantley M, Floyd V, Norris D, Franko E, Blake P, Toomey, Fernhoff P, Ziegler B, Mayer L, Henderson O, Hannon H, Martin L, Ferre C. 1997. Population-based prevalence of perinatal exposure to cocaine in Georgia, 1994. *Georgia Epidemiology Report* 13(2):1–3. February.
3. Henderson LO, Powell MK, Hannon WH. 1997. An evaluation of the use of dried blood spots from newborn screening for monitoring the prevalence of cocaine use among childbearing women. *Biochemical and Molecular Medicine* 61(2):143–51.


1993 Georgia Metropolitan Statistical Areas
Chattanooga
Atlanta
Athens
Augusta
Macon
Columbus
Savannah
Albany

Map 1


County Map of Georgia

Map 2

Issues in Environmental Justice Research

Susan L Cutter*
Director, Hazards Research Laboratory, Department of Geography, University of South Carolina, Columbia, SC

Abstract

This paper examines some of the problems and constraints in conducting environmental justice research. The disproportionate impact of environmental threats on public health, especially on minority and low-income populations, is the key concept in environmental justice. After a brief review of environmental equity, public policies, and the environmental justice movement, the paper highlights some of the empirical and GIS-related research supporting and/or refuting environmental justice claims. The constraints include the choice of the specific environmental threat or threats and their comparability; the geographic scale of analysis; the particular subpopulation selected; and the time frame for the analysis. The paper concludes with a suggestion to move from static interpretations of environmental injustices to more dynamic approaches that rank the relative hazardousness of spatial units (census tracts, counties, etc.) based on the magnitude and toxicity of releases within them, rather than the mere presence or absence of industrial facilities. Measures of relative risk provide more meaningful indicators of the potential sources of environmental threats within communities and can make it easier to understand local sensitivity to claims of environmental injustice.

Keywords: environmental justice, toxic releases, geographic scale

Introduction

Environmental equity—preventing disproportionate effects of environmental degradation on people and places—has been a federal concern for at least three decades (1–3). In the early 1990s, coalitions of civil rights and environmental activists transformed environmental equity concerns into the environmental justice movement, ostensibly because of concerns about the placement of toxic waste facilities in low-income and minority communities (4–8). The First National People of Color Environmental Leadership Summit was held in 1991 and immediately was followed by the establishment of the US Environmental Protection Agency's (EPA's) Office of Environmental Equity. In 1994, environmental justice was institutionalized within the federal government through Executive Order 12898, which focused federal attention on human-health and environmental conditions in minority and low-income communities. It also provided for greater public participation and access to environmental information in these impacted communities.

While the environmental justice movement has been successful in bringing the issue to the attention of policy makers and the public, there is some skepticism as to whether or not injustices do, in fact, exist. In other words, there is a federal policy in place despite little empirical verification of the true extent of the problem. While we may see a correlation between the presence of facilities and the demographic composition

* Susan L Cutter, Dept. of Geography, University of South Carolina, Columbia, SC 29208 USA; (p) 803-777-5236; (f) 803-777-4972; E-mail: scutter@sc.edu

of places in the late 1990s, we have little knowledge of how this situation arose. The outcome is important, but the process behind it is equally relevant in environmental justice considerations (9–10). Were these sources of environmental threats intentionally located in communities that were poor, minority, and/or politically weak? Or is there an alternative explanation, one suggesting that facilities were located without any reference to the race and economic status of communities, and that the demographics of communities with facilities simply changed over time, producing the inequity that we see today?

This paper examines some of the geographic factors involved in proving or disproving environmental justice claims. Six issues are highlighted as the most salient.

Precise Locations of Threats

The vast majority of social science studies of environmental justice use a standard correlation methodology to examine the relationship between toxic facilities and the demographics of their locations. There is an implicit assumption that the reported locations are correct. While many of the studies comment on some of the reliability issues of the EPA's Toxics Release Inventory (TRI)—the database most often used—there is rarely comment on the locational accuracy of the sites (11). In a statewide study of South Carolina using EPA databases (TRI, the National Priorities List [NPL], and the Biennial Reporting System [BRS]) for the 1987–1992 time period, we found that nearly 60% of the facilities listed were located in the wrong census block group (12). Comparing inaccurately recorded locations with accurately recorded locations can lead to widely different conclusions about the racial makeup of the surrounding community. For example, Westinghouse Nuclear Fuels Division (in Columbia, South Carolina) was improperly located on an EPA Web site that illustrated the use of geographic information systems (GIS) in environmental justice. When the site was located correctly and the concentric zones redrawn, the profile of the community changed from completely white (according to EPA) to 91% nonwhite. As Table 1 shows, there was very little percentage change in the 0- to 5-mile range, although the absolute number of potentially affected residents was lessened significantly.

Table 1 Differences in Socioeconomic Characteristics Before and After Correcting for the Location of the Westinghouse Nuclear Fuels Facility in Columbia, SC

	Block Group		0 to 1 mile		0 to 3 miles		0 to 5 miles	
	Old[a]	New[b]	Old[a]	New[b]	Old[a]	New[b]	Old[a]	New[b]
% nonwhite	0	91	25	61	36	56	38	41
% poverty	11	38	19	21	19	15	16	13
Total population	536	2,051	15,804	190	74,076	5,099	148,660	26,964
Total minority population	0	1,873	3,300	143	27,026	4,308	61,206	11,377
Density (per square mile)	1,703	14	2,043	21	1,562	185	1,366	564

[a] Demographic characteristics according to an EPA Region 4 Web site illustrating the use of GIS in environmental justice

[b] Revised demographic characteristics based on the correct location of the facility. Conducted by the Hazards Research Lab, University of South Carolina.

Choice of the Environmental Threat

The potential for scientific replication and generalization of findings is often thwarted by the lack of comparability between empirical studies. Depending on the type of threat examined (e.g., a landfill, a TRI facility, a Superfund site, actual emissions), very different patterns can be observed, leading to conflicting results in the literature. Hird (13) found that affluent counties were more likely to host NPL sites than non-affluent counties. On the other hand, no relationship between poverty and host/non-host counties was found when hazardous waste treatment, storage, and disposal facilities were examined (14).

Very few studies have been conducted comparing two different sources of threats and their spatial manifestations for particular places. In a study of the Southeast comparing acute releases (reported by the federal Emergency Response and Notification System) with more chronic releases (reported under TRI), little association with race was found (15). Wealth indicators, on the other hand, were positively correlated with releases, with TRI releases being dominant in urban areas. In an earlier study of the same region, considering only TRI emissions, Stockwell et al. (16) developed a GIS-based profiling method to delineate high-risk from low-risk counties and found that high-risk counties were correlated with population density and more urbanized places. Depending on the nature of the environmental threat, we can see radically different conclusions from the empirical literature.

Geographic Scale of Analysis

This issue of geographic or spatial scale is perhaps one of the most important issues in environmental justice research. While the scale of research studies varies widely (census block, tract, metropolitan area, county), there is no assessment or uniform opinion as to which scale is the most appropriate for proving or disproving environmental justice claims (17). Because of the aggregation bias and the modified areal unit problem (well-known spatial considerations in geography), the selection of the enumeration unit is critical in the proof. A number of studies demonstrate that the statistical results change as the geographical unit of analysis is varied (18,19). For example, there was no association between any of the three indicators examined (TRI, BRS, NPL) and the racial or economic composition of host census tracts or blocks for a study of South Carolina. However, when data were aggregated to the county level, larger numbers of facilities were associated with higher-income white counties—just the opposite of what most people expected (20).

A secondary scale issue involves the methods for stratifying the population around the facilities and the resulting classification problems. The host/non-host methodology (i.e., using statistical analyses—such as difference of means and difference of proportions tests—to compare host and non-host communities, thus ascertaining the statistical significance of facility distribution) is the most common way to differentiate the local geography. Recent research is moving toward using buffers (at varying distances from the source) as the classification tool (21–24). However, major questions arise concerning the actual interval distance to use (0.5 miles, 1 mile, etc.) and the basis for that selection (worst-case events, modeled effects, convenience). Obviously, the choice of buffer distance is an important variable in determining whether inequities exist or not.

Subpopulation Selected

Thus far, most environmental justice research has targeted two subpopulations, delineated by race/ethnicity and income levels. There are few studies that examine the disproportionate effects on subpopulations delineated by gender or age, despite very real differences in susceptibility, vulnerability, and ability to cope with and recover from environmental threats (25,26) including disasters (27).

The following example illustrates why other subpopulations should be considered in addition to those based on race/ethnic and wealth indicators. Again drawing on the experience with South Carolina, the state's block groups were categorized based on the percentage of children under 18 in each block group, compared with the statewide average for the same percentage. The comparisons were expressed as ratios. Once they were categorized as either high (>1.1), medium (0.9:1–1.1), or low (<0.9) regions, the block groups were mapped and statistically compared with the block-group locations of all TRI facilities in South Carolina, and also those TRI facilities in the state that emitted heavy metals, using a host/non-host methodology. The differences between each group were statistically significant (based on chi-square tests), as shown in Table 2. More importantly, when the sites were desegregated, we found that nine out of the top ten heavy emitters were located in block groups with greater than the statewide average of children under 18. How this relates to potential health outcomes is unclear at this time, but it does provide another perspective on who bears the burdens of toxic releases.

Table 2 Children and the Location of Toxic Facilities in South Carolina, 1990

Block Group Characterization for Children <18 Years[a]	Percentage of Top 100 TRI Releasers in Block Group Category	Percentage of TRI Heavy Metal Emitters in Block Group Category	Number of TR Heavy Metal Emitters in Block Group Category
High	44%	45%	61
Medium	31%	28%	38
Low	25%	27%	36
Chi-square (significance)	6.08 (.0478)	10.09 (.0064)	

[a] Block group characterization (high, medium, low) is calculated as the percent younger than 18 in a block group, divided by the percent younger than 18 for the state. "High" indicates block groups with a ratio of children <18 years greater than 1.1:1, "medium" indicates a ratio between 0.9:1 and 1.1:1, and "low" indicates a ratio lower than 0.9:1.

Time Frame

As noted earlier in this paper, environmental justice research needs to address the fundamental issue of "which came first," so that we more accurately understand the processes that gave rise to the patterns that are observed today. As Greenberg (28) suggests, it is important that environmental justice research focus on both outcome (current patterns) and process. However, the process-oriented studies require detailed time-series analyses of demographic change in communities. These are difficult to perform because of limitations in historical databases, knowledge of facility start dates, and most importantly, matching historical census boundaries. However, a number of

studies (10,29,30) have conducted these time-series analyses with inconclusive results. While inequities may currently exist, they came about by a process more likely explained by regional and state migration patterns, market dynamics, and unique and localized sociospatial contexts. For example, in an attempt to solve income and employment disparities within a region, states may embark on economic development plans that promote economic growth but may also promote environmental inequities. This may help to explain some of the disparities found in rural, southern states, for example, though this explanation may not address the processes giving rise to inequities in northern urban-industrial areas.

Relative Hazardousness of Spatial Units

The late 1990s have seen an increasing sophistication of environmental justice research. There is movement away from the static indicators of injustices—mere presence or absence in a community—to more consideration of the underlying processes and potential impacts of facilities and their emissions. Of critical concern are the quantity and toxicity of emissions from facilities. Secondary concerns are the spatial variability of emissions and their potential impact on local populations. While many of the existing databases (such as TRI) provide estimates of releases or emissions, there is no consistent source of data regarding the toxicity or potential health impact of these emissions. One of the biggest stumbling blocks to this line of inquiry is the lack of a consistent measure of toxicity that the social science community can use in comparing risk.

Only a handful of studies have tried to incorporate toxicity measures into environmental justice research (11,16,24,31,32). While each study used TRI data, they employed different toxicity measures, so comparability between them and generalizations from them are difficult. Much more work needs to be done in this area to develop a representation of relative risk and thus a prioritization or action.

Conclusions

This paper has described the evolution of environmental justice policy. Based on the research and empirical work to date, a number of issues or lessons have been highlighted. First, there is a critical need for spatial accuracy in the location of toxic facilities. Second, geographical scale is important because injustices may statistically exist at one scale, but disappear when using another. The optimal scale depends on the initial questions asked of the research and/or policy. Third, the choice of environmental threat, subpopulation, and time frame affect the comparability of findings and their replication. A number of important subgroups are missing from much of the research (e.g., the elderly, children). Thus far, there is little replication of results, largely due to the differences mentioned. Fourth, the current physical distribution of environmental threats may not lead to differences in potential exposures. Just because a facility is located in a minority tract, for example, does not mean that there is more potential exposure to that tract's residents. Fifth, the spatial delineation of toxicity indicators can help to define the relative hazardousness of places. Finally, and perhaps most important, the empirical "proof" of an injustice may be less important than the local perception of and sensitivity to the issue. Demanding complete certainty in the existence of environmental injustice before policy initiatives are undertaken may pose greater risks to the well-being

and functioning of communities than simply responding to the perception of the threats.

References

1. Berry BJL. 1977. *The social burdens of environmental pollution.* Cambridge: Ballinger.
2. Cutter SL. 1995. Race, class, and environmental justice. *Progress in Human Geography* 19:107–18.
3. US Government Accounting Office. 1995. *Hazardous and nonhazardous waste: Demographics of people living near waste facilities.* Washington, DC: GAO/RCED-95-84.
4. United Church of Christ Commission for Racial Justice. 1987. *Toxic wastes and race in the United States.* New York: United Church of Christ.
5. Bullard RD. 1990. *Dumping in Dixie: Race, class, and environmental quality.* Boulder, CO: Westview Press.
6. Bullard RD. 1994. *Unequal protection: Environmental justice and communities of color.* San Francisco: Sierra Club Books.
7. Bullard RD. 1996. Environmental justice: It's more than waste facility siting. *Social Science Quarterly* 77 (3):493–9.
8. Bryant B, Ed. 1995. *Environmental justice: Issues, policies, and solutions.* Washington, DC: Island Press.
9. Been V, Gupta F. 1997. Coming to the nuisance or going to the barrios? A longitudinal analysis of environmental justice claims. *Ecology Law Quarterly* 24:1–56.
10. Yandle T, Burton D. 1996. Reexamining environmental justice: A statistical analysis of historical hazardous waste landfill siting patterns in metropolitan Texas. *Social Science Quarterly* 77:477–92.
11. McMaster RB, Leitner H, Sheppard E. 1997. GIS-based environmental equity and risk assessment: Methodological problems and prospects. *Cartography and Geographic Information Systems* 24(3):172–89.
12. Scott MS, Cutter SL, Menzel C, Ji M, Wagner DF. 1997. Spatial accuracy of the EPA's environmental hazards databases and their use in environmental equity analyses. *Applied Geographic Studies* 1(1):45–61.
13. Hird JA. 1994. *Superfund: The political economy of environmental risk.* Baltimore: Johns Hopkins University Press.
14. CleanSites, Inc. 1990. *Hazardous waste sites and the rural poor: A preliminary assessment.* Alexandria, VA: Clean Sites, Inc.
15. Cutter SL, Solecki WD. 1996. Setting environmental justice in space and place: Acute and chronic airborne toxic releases in the southeastern United States. *Urban Geography* 17(5):380–99.
16. Stockwell JR, Sorenson JW, Eckert JW Jr, Carreras EM. 1993. The US EPA geographic information system for mapping environmental releases of Toxic Chemical Release Inventory (TRI) chemicals. *Risk Analysis* 13:155–64.
17. Zimmerman R. 1993. Social equity and environmental risk. *Risk Analysis* 13:649–66.
18. Anderton DL. 1996. Methodological issues in the spatiotemporal analysis of environmental equity. *Social Science Quarterly* 77(3):508–15.
19. Zimmerman R. 1994. Issues of classification in environmental equity: How we manage is how we measure. *Fordham Urban Law Journal* 21:633–69.

20. Cutter SL, Holm D, Clark L. 1996. The role of geographic scale in monitoring environmental justice. *Risk Analysis* 16(4):517–26.
21. Opaluch JJ, Swallow SK, Weaver T, Wessells CW, Wichelns D. 1993. Evaluating impacts from noxious facilities: Including public preferences in current siting mechanisms. *Journal of Environmental Economics and Management* 24:41–59.
22. Glickman TS, Hersh R. 1995. Evaluating environmental equity: The impact of industrial hazards on selected social groups in Allegheny County, Pennsylvania. Discussion Paper 5-13. Washington, DC: Resources for the Future.
23. Chakraborty J, Armstrong MP. 1997. Exploring the use of buffer analysis for the identification of impacted areas in environmental equity assessment. *Cartography and Geographic Information Systems* 24(3):145–57.
24. Neumann CM, Forman DL, Rothlein JE. 1998. Hazard screening of chemical releases and environmental equity analysis of populations proximate to Toxic Release facilities in Oregon. *Environmental Health Perspectives* 106(4):217–26.
25. Cutter SL. 1995. The forgotten casualties: Women, children, and environmental change. *Global Environmental Change* 5(3):181–94.
26. Swanston SF. 1994. Race, gender, age, and disproportionate impact: What can we do about the failure to protect the most vulnerable? *Fordham Urban Law Journal* 21:577–604.
27. Fothergill A. 1996. Gender, risk and disaster. *International Journal of Mass Emergencies and Disasters* 14(1):33–56.
28. Greenberg MR. 1993. Proving environmental equity in siting locally unwanted land uses. *Risk—Issues in Health and Safety* 4:235–52.
29. Anderson AB, Anderton DL, Oaks JM. 1994. Environmental equity: Evaluating TSDF siting over the past two decades. *Waste Age* 25(7):83–100.
30. Mitchell JT, Thomas DSK, Cutter SL. 1999. Dumping in Dixie revisited: The evolution of environmental injustices in South Carolina. *Social Science Quarterly* 80(2):229–43.
31. Bowen WM, Salling MJ, Haynes KE, Cryan EJ. 1995. Toward environmental justice: Spatial equity in Ohio and Cleveland. *Annals of the Association of American Geographers* 85:641–63.
32. Perlin SA, Setzer RW, Creason J, Sexton K. 1995. Distribution of industrial air emissions by income and race in the United States: An approach using the Toxic Release Inventory. *Environmental Science and Technology* 29:69–80.

Using GIS to Study the Health Impact of Air Emissions

Andrew L Dent (1),* David A Fowler (2), Brian M Kaplan (3), Gregory M Zarus (4)
(1) GIS Analyst, Electronic Data Systems, Plano, TX; (2) Toxicologist, Agency for Toxic Substances and Disease Registry, Exposure Investigations and Consultation Branch, Atlanta, GA; (3) Environmental Health Scientist, Agency for Toxic Substances and Disease Registry, Federal Facilities Assessment Branch, Atlanta, GA; (4) Atmospheric Scientist, Agency for Toxic Substances and Disease Registry, Exposure Investigations and Consultation Branch, Atlanta, GA

Disclaimer

The use of company or product names is for identification only and does not constitute endorsement by the Agency for Toxic Substances and Disease Registry or the US Department of Health and Human Services.

Abstract

Geographic information system (GIS) technology is fast-developing with an ever-increasing number of applications. Air dispersion modeling is a well-established discipline that can produce results in a spatial context. The marriage of these two applications is optimal because it leverages the predictive capacity of modeling with the data management, analysis, and display capabilities of GIS. In the public health arena, exposure estimation techniques are invaluable. The utilization of air emission data, such as the US Environmental Protection Agency's Toxics Release Inventory data, and air dispersion modeling with GIS enable public health professionals to identify and define a potentially exposed population, estimate the health risk burden of that population, and determine correlations between point-based health outcome results and estimated health risk.

Keywords: air pollution, emissions, toxics release inventory, public health, Air Force

Introduction

The federal Agency for Toxic Substances and Disease Registry (ATSDR) is often charged with investigating past exposures to determine their associations with health outcomes observed within specific communities. The task often requires the use of information specific to a given locality, namely: toxic substance release data, topographical data, land use data, meteorological data, mathematical modeling data, population density data, demographic data, and health outcome data. Proper evaluation and correlation of these data are essential to reveal possible associations between observed health effects and chemical exposures. This paper details a technique by which air dispersion model results are integrated with other spatial datasets on a geographic information system (GIS) platform. Once the component datasets are gathered, they can be

* Andrew L Dent, Electronic Data Systems, 5400 Legacy Dr., Plano, TX 75024 USA; (p) 404-639-6099; E-mail: aed5@cdc.gov

analyzed using standard GIS functions to provide support for interpretation of site data and evaluation of exposure hypotheses.

Public Health Context

The analytical technique explained in this paper will be showcased in the context of a public health concern brought on by environmental contamination. The public health concern existed at a US Air Force base charged with the management and maintenance of aircraft engines, weapons systems, support equipment, and aerospace fuels. The base hosted, maintained, and repaired various jet aircraft. Specific activities that were potential sources of off-site air contamination included painting, chrome plating, fueling, and fuel storage. Members of the community neighboring the base expressed concern about fuel vapor and other odors, and questioned the relationship between the odors and the occurrence of health effects such as nausea, headaches, difficulty breathing, and cancer.

Integration Technique

The technique integrates conventional air dispersion modeling with GIS technology. The marriage of these two technologies is appropriate because it takes advantage of each component's strengths—the high-powered, predictive capacity of computer models and the editing, data handling, and interpolation capabilities of GIS (1). This paper first explains the basics of air dispersion modeling, including its development, uses, and advantages. Next, it outlines the rise of GIS technology and investigates a variety of GIS applications. Finally, this paper concludes with a discussion of the integration of air dispersion modeling results with other spatial data on a GIS platform.

Introduction to Air Dispersion Modeling

Air dispersion modeling is a predictive method used to estimate the concentrations of pollutants in the atmosphere resulting from point or nonpoint atmospheric emissions. It takes into account various factors that can affect a substances's concentration in a plume as it migrates through the air. These influential factors include gravity, meteorological conditions, and chemical reactions. Among a variety of other purposes, air dispersion models have been used to:

- Design stacks to minimize the nuisance from pollutants at ground level
- Calculate when odors might be expected
- Determine the needed removal efficiency of air pollution control equipment
- Plan for emergency response to accidental releases
- Determine the acceptable levels of operation of air pollution generating facilities (2)

The regulatory role of air dispersion models has gained great significance because judicial rulings in 1977, 1978, and 1980 have upheld the Clean Air Act Amendment of 1977, which states that a modeled air pollution violation is just as valid as a sampled violation (2).

The Creation of Air Models

Air dispersion models are semiempirical—they are in part derived from basic

principles and in part derived from measured data. Specifically, an air dispersion model is developed by measuring both emissions and meteorological conditions and monitoring contaminant concentrations at various points downwind of the source. These data are then correlated to find a mathematical model equation that provides the best fit. The resulting model is validated through rigorous testing at various combinations of known emissions and meteorological conditions. Although the mathematics of the model greatly affect its successful prediction of a substance's concentration, the quality of the input data also has a great impact on the results (2).

Advantages of Air Dispersion Modeling

The use of air modeling is advantageous for a variety of reasons, including the following (3):

- Models can be used to estimate a substance's concentration 24 hours a day for any time period for which both emissions and meteorological data exist.
- Models can account for deposition from particulate matter settling to the ground, as well as depletion resulting from the substance's reaction with sunlight and other materials.
- Models can be used to estimate the level of various substances existing in the ambient air as a result of emissions from a single source or multiple sources.
- Models can average short-term fluctuations in emissions and meteorological conditions, resulting in a long-term average.
- Models can estimate a substance's concentration at an unlimited number of locations.

Conversely, conventional air sampling can be limiting for a variety of reasons, including the following:

- Sampling measures substances arising from many and varied sources in the area; it cannot determine the effects of a single facility.
- Sampling results are based on conditions at the time of the sampling event. These conditions could be an extreme and not represent average conditions (3).
- Sampling efforts can be very expensive.

Geographic Information System Technology

GIS technology provides an excellent platform upon which different types of spatially referenced data can be united for analysis and display purposes. Prior to the advent of GIS technology, many operations that involved the concerted utilization of datasets derived from different sources and in different formats were carried out using a "push-pin" approach in which hard-copy maps were generated and overlaid upon one another. The approach was both costly and time-consuming and often yielded substandard results. This type of spatial analysis harkens back to a map by French military leader and cartographer Louis Alexandre Berthier (1753–1815) that was composed of hinged overlays showing troop movements during the 1781 Siege of Yorktown (4). In the 1960s, Dr Roger Tomlinson of the Canada Geographic Information System (CGIS) developed the first computer-based GIS (5). CGIS was designed to store, manage,

analyze, and manipulate spatial data in an effort to assess the productivity of Canadian farmland.

The arrival of high-powered computers in the 1980s facilitated the proliferation of GIS applications in a variety of different disciplines. It has been estimated that over 80% of the world's data have a spatial component. These spatial components could include an address in a database, coordinates in sampling data, or a zip code in sales data. GIS has been used to manage county tax records, route delivery and emergency vehicles, perform site selection based on many parameters, manage utility networks, and develop strategies to address problems such as crime, urban sprawl, and environmental degradation.

Elements of a GIS

The process of developing a GIS includes data acquisition and preprocessing; data management, manipulation and analysis; and product generation (4). Data acquisition is often the most expensive and time-consuming element in utilizing GIS technology. Issues involved in this phase of development include data accuracy, scale, and metadata. Metadata is a written detailed description of the data similar to engineering specifications. Preprocessing, or the machinations necessary to convert data to a digital format and integrate it with other spatial data, often involves elements such as digitizing and quality assurance and control procedures. As with air dispersion modeling efforts, the successful use of GIS technology depends much upon the datasets used and the methods used to automate those datasets. Thus, data management, including storage and documentation, is key to a GIS project. During the manipulation and analysis phase, the data are used to get results and make decisions. It is important to remember that GIS is a tool—a flexible, easy to use tool—but, still, only a tool. The application or effective use of GIS is key to the success of a GIS enterprise. Gigabytes of spatial data are of no value to any organization if that information is not used. Finally, product generation involves the transmittal of results produced by GIS to the people who need them. These results may be delivered as a conventional paper map or digitally via diskettes, intranet, or Internet. Whereas the paper map is static, a digital map can be dynamic because it can contain a greater amount of information. This allows the map reader to define the map message by selecting different map scales or different data to display.

Integration of Modeling Results on a GIS Platform

Typically, GIS and air dispersion modeling software are separate packages often written in different languages. Therefore, the question arises as to how the air dispersion modeling results can be most effectively integrated with other spatial data on a GIS platform. Models for the integration of modeling results and GIS data include full integration, loose coupling, and tight coupling (Figure 1) (1). Full integration means that the calculations performed by the model are encoded in a high-level language packaged with the GIS software, such as ARC/INFO's Arc Macro Language (AML) or ArcView's Avenue (Environmental Systems Research Institute, Redlands, CA). Loose coupling integration consists of uniting the systems at predefined end points. For example, the data would first be processed using air modeling software, then the results would be used in a GIS package. Tight coupling integration involves the development of a user

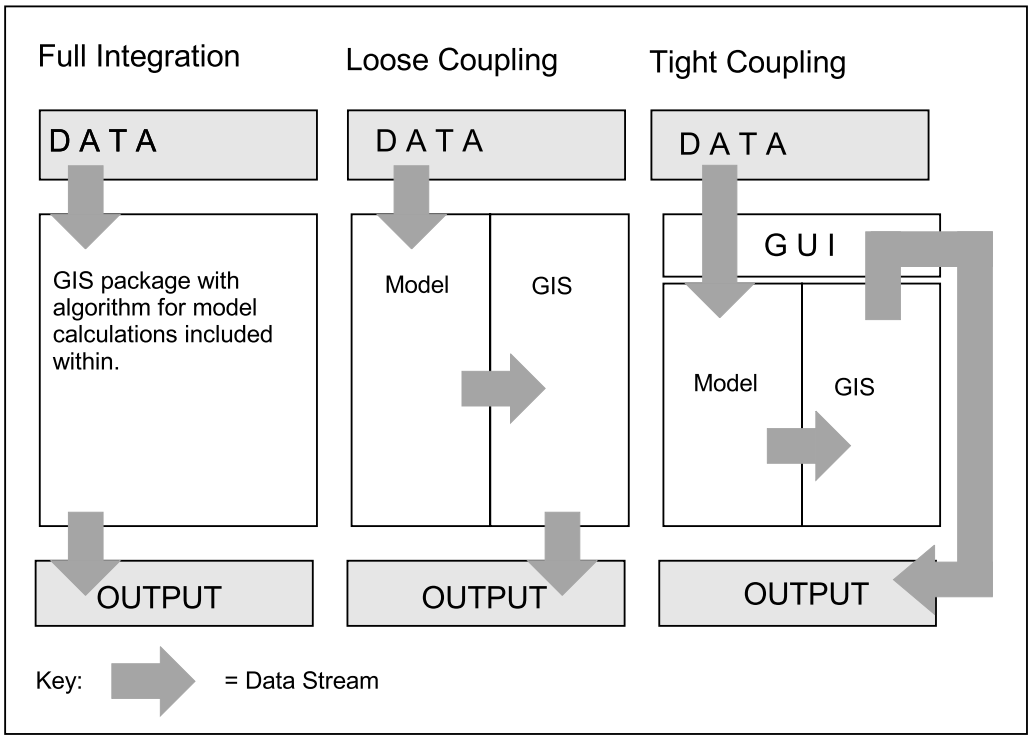


Figure 1 Integration strategies (1).

interface that allows users to access both the model software and the GIS package using one graphical user interface (GUI). An application built on the tight coupling concept gives the illusion that one software package is being used when, in reality, the GIS package and the modeling program are being used separately in concert with each other.

The work discussed in this paper uses loose coupling because loose coupling provides a faster implementation time than tight coupling and with the same results. If, in the future, this process needs to be used on a repetitive basis, it may be worth the time to develop a tight coupling integration application. In such a case, the investment in development time and effort would be offset by the speed and efficiency with which the task could then be carried out.

GIS Datasets

Use of air dispersion model results with other GIS spatial datasets on a GIS platform is at the heart of the work discussed here. This section discusses the spatial datasets that were compiled and integrated with the air dispersion model results. The conditions at the site and the goal of ATSDR in its investigation have warranted the use of various datasets. These include aerial photography, 1995 TIGER/Line files, and US Census demographic data.

Aerial Photography

Aerial photography can be gathered from a variety of data sources. Although image

processing techniques provide many ways to view and manipulate aerial photography, the authors used it here primarily for orientation purposes.

TIGER/Line Data

Topologically Integrated Geographically Encoded and Referencing (TIGER/Line) files (6) are spatial datasets compiled by the US Census Bureau that include base map features for all areas of the United States. These features include roads, rivers, water bodies, political boundaries, and cultural features (schools, parks, and hospitals). GIS data consisting of on-base features in large scale were obtained from the engineering division of the Air Force base; however, the aerial extent of the analysis to be performed required more extensive use of the TIGER/Line files.

US Census Demographic Data

ATSDR's work depends heavily on demographic information when analyzing environmental contamination and the potentially related health effects. Specifically, obtaining the raw numbers of people in an area and the numbers of people who are in high-risk groups (children, the elderly, and women of reproductive age) is critical. The US Census provides this type of data along with an abundance of other information on socioeconomic variables. For analysis at the Air Force base, block-level spatial data and demographic data were compiled. This information was dynamically integrated with the model-based results data in the GIS analysis phase of the project.

Integration Technique

ATSDR used the most recent version of the US Environmental Protection Agency's (EPA's) Industrial Source Complex Short Term air dispersion model, version 3 (ISCST3), to estimate emissions from the Air Force base. EPA developed the ISCST3 model primarily for determining if air emissions sources meet state and federal air quality standards (7). The ISCST3 model allows the refinement of results through the use of additional parameters such as building height, source temperature, particle size, and decay rate. When modeling the emissions from the Air Force base, ATSDR did not specify the values for these additional parameters. The default values were deemed appropriate because the potential improvement of the results arrived at through the setting of these parameters was considered to be negligible when compared to the uncertainty of the input parameters.

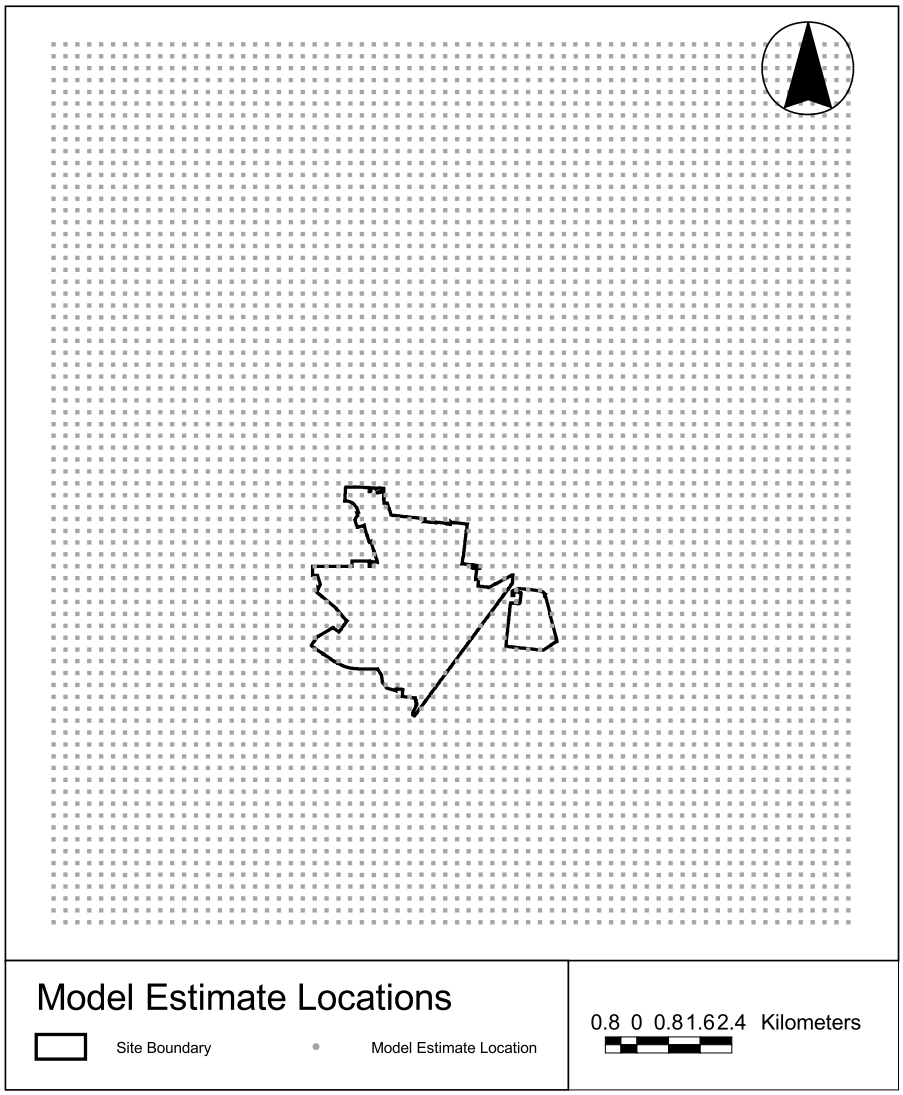
The model was run on 7,016 inputs (sources and multiple chemicals from each source) from the Air Force base emission inventory for a variety of carcinogenic compounds. This inventory, which is similar to EPA's Toxics Release Inventory (TRI), contains an emission rate in grams per second for each emitted air pollutant from each source. The model estimated concentration values for a uniform grid of 5,100 points spread evenly across the 446-square-kilometer study area at 300-meter intervals. Five years of meteorological data were averaged to form the meteorological component. The results included the coordinates for each modeled point, the average compound concentration estimate at each point, and the number of hours for which the compound concentration values were estimated.

A GIS point file for each of the modeled contaminants was created. One point in the GIS file exists for each modeled grid point in the original air model output. Thus, the

spatially referenced modeled output dataset can now be integrated with other spatial datasets (Figure 2).

Map Projections

The air dispersion model produces contaminant values at grid intersections of a Cartesian coordinate system. This coordinate system is a two dimensional representation of the earth's surface. The coordinate system is derived from a map projection that is a mathematical method for representing the features of the spherical surface of the earth on a planar map. GIS datasets can be stored using a variety of different map



Source: ATSDR, 1998.

Figure 2 Locations of points at which the model generated a concentration estimate (3).

projections. For datasets to be integrated, however, it is imperative that they are stored in the same projection. Therefore, the integration of air dispersion modeling data involves the reprojection of data from the original map projection to the projection of the remaining datasets. In this case, the GIS point file was projected from the Texas State Plane South Central coordinate system (based on the Lambert conformal conic projection) to the Plate Carree projection. After the importation and reprojection have been accomplished, the modeled datasets can be analyzed in concert with existing data.

Applicable GIS Processes

Once data have been integrated into a single digital map on a GIS platform, several spatial processes are available for analyzing the information. For instance, contours can be generated from values at point locations by a variety of interpolation techniques. Kriging is considered the optimal method of spatial linear interpolation. Kriging is a logarithmic method in which the mean is estimated from the best linear-weighted moving average (8). For the work presented in this paper, the authors used kriging to generate contours from the air dispersion modeling data because logarithmic interpolation is more consistent with the natural distribution of contamination.

Contours developed with kriging can be used to generate polygons that, in turn, can be used in additional GIS analysis, such as the area-proportion technique. The area-proportion technique is an excellent example of the use of GIS to analyze disparate types of data to get results. It is essentially a "cookie-cutter" operation that estimates values within a polygon based on values of polygons in another data layer. For instance, the number of people living within a contour generated by kriging could be estimated based on the number of people living in census blocks in the same area. For cases in which the contour polygon crosses the census blocks, a simple proportion of the area of the census blocks lying within the target polygon is used to compute population numbers (Figure 3). The area-proportion technique assumes an even distribution of population (or whatever is being estimated) and, therefore, might result in a certain amount of measurable error. For example, the area-proportion technique assumes that the population within a census block is evenly distributed throughout the block. Thus, if a contour polygon encompasses 50% of a block it is assumed that 50% of the population of that block lives within the contour polygon. However, because population is rarely evenly distributed across an area, the area-proportion technique results in some amount of error. For example, if all of the population living in a census block that the contour polygon boundary intersects actually live *outside* the contour polygon, then the

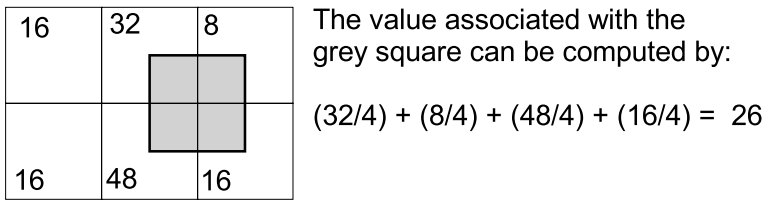


Figure 3 Area-proportion technique illustrated.

area-proportion estimate is high. Conversely, if all of the population living in a census block intersected by the contour polygon boundary actually live *inside* the buffer polygon, the area-proportion estimate is low. The size of the error generated by the area-proportion technique depends on:

- The number of reference polygons (e.g., census blocks) used to compute the estimate
- The size of the reference polygons

The application of the area-proportion technique is an excellent way to begin exploration of disparate datasets using GIS.

Public Health Applications

After all datasets have been melded in a GIS format, many avenues of data exploration are open to public health professionals. As stated earlier, ATSDR's goal is to evaluate the potential detrimental effects of air releases on the neighboring populations. This goal provides the impetus for integrating air dispersion modeling data with existing spatial datasets.

Calculation of Cancer Health Risk

The additional cancer health risk to a population can be estimated by multiplying the EPA's cancer inhalation slope factor (unit risk) by the average annual concentration predicted by the model. This estimate is considered a screen because it is a worst-case conservative estimate and indicates which segments of the population of interest can be eliminated from further analysis and which segments require more refined analysis. Segments of the population requiring further analysis can be more accurately identified using this integrated approach, not only as to physical location, but also as to demographic composition. This process would allow a more site-specific approach to account for the prevalence of more sensitive subpopulations. This refined information can also be used to identify residents for information mailings and notifications, solicitation of information, and clinical intervention or medical monitoring.

Refinement of the Exposed Population

Without the use of modeling (or sampling), public health professionals are tied to obsolete methods for estimating the exposed population. For instance, some organizations might have specified that a population living within a uniform distance of a site boundary is the "exposed population." Population estimates derived in this manner are often substantially different from the actual exposed population. Modeling allows the refinement of size, location, and demographic composition of an exposed population. In the example shown in Figure 4, air dispersion modeling has resulted in a smaller exposed population than the estimate arrived at using the uniform distance technique (Table 1). Furthermore, the population derived from air modeling is distributed across a slightly different area.

Correlation with Point-Based Health Outcome Data

Often health professionals can obtain spatially referenced health outcome data for specific populations. Such a dataset would include the coordinates of a residence or

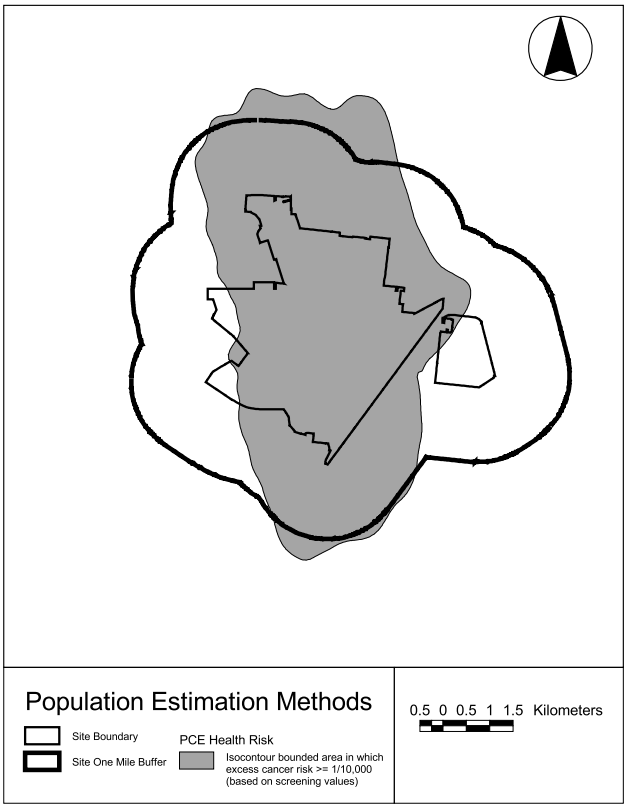


Figure 4 Exposed population estimation methods.

Table 1 Estimates of Population Exposed to Perchloroethene (Unnamed Air Force Base)

Area of Estimation	Total Population[a]
Population within 1 mile of site boundary	50,861
Population within 1/10,000 risk contour	26,033

[a] Computed by area-proportion technique.
Source: (3,6)

workplace of an individual and the health outcome exhibited by that individual. Obtaining such data is often made difficult because of confidentiality issues; however, when they can be obtained, they can be successfully integrated with GIS data (such as air dispersion modeling results).

After the coordinates are imported into a GIS package, the health outcome points can be correlated with the contours of elevated health risk. This type of analysis can indicate if there is a potential association between a modeled exposure level and a reported health outcome. However, many other issues related to the makeup of the population must be considered. These critical elements include, but are not limited to:

- Length of residence

- Potential occupational exposure
- Potential residential exposure
- Genetics
- Socioeconomic status
- Lifestyle (e.g., smoking and nutrition)

Each of these elements can have an effect on the interpretation of health outcome data. For example, length of residence is considered because cancer development usually involves a latency period of many years. Therefore, if a person in the community of interest developed a cancer, but had only lived in the community a short time, it would be unlikely that the cancer development occurred as a result of an exposure in the community of interest.

Correlation Methods

At this time, the health outcome data based on the residence location of reported cases have not yet been compiled for the project discussed in this paper. Nonetheless, the analysis will generally proceed as follows.

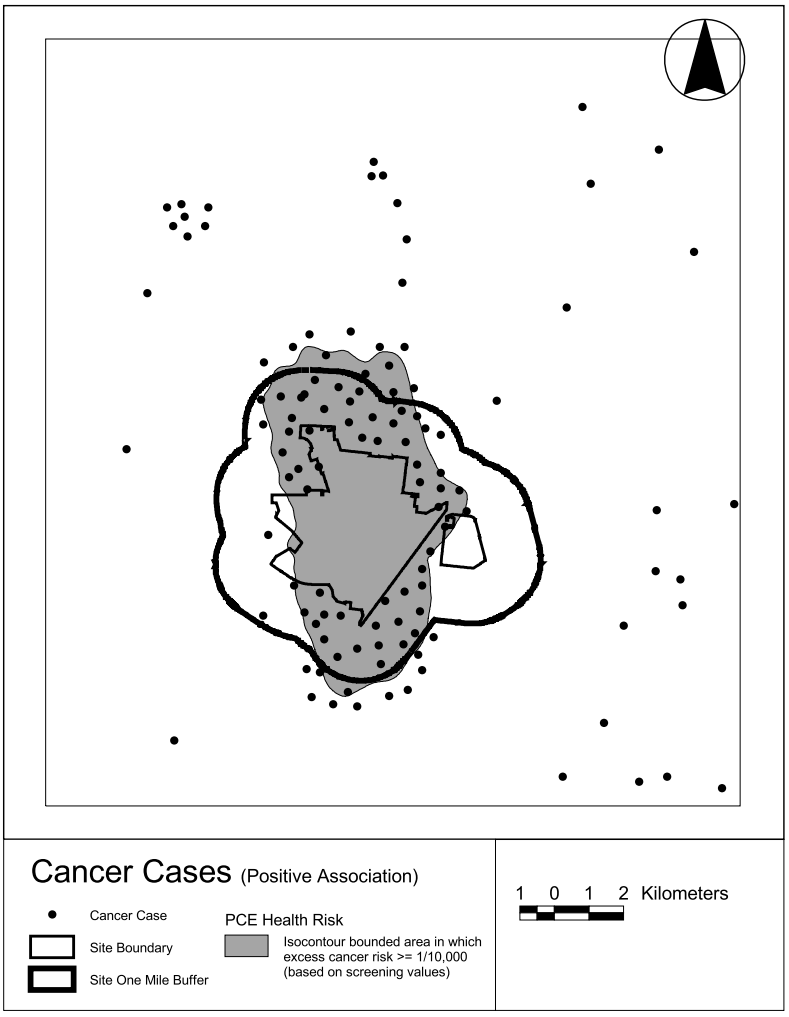
In determining the correlation of health outcome points with a cancer risk contour, five properties of the distribution of point features are pertinent. These are:

- Frequency: Number of occurrences
- Density: Number of occurrences per unit area
- Geometric center: Means of x-y coordinates
- Spatial dispersion: Standard deviation of the means of x-y coordinates
- Spatial arrangement: "Pattern" of the points, can be clustered, random, or scattered (8)

Based on the distribution of cancer case points in the vicinity of the study area, a judgement can be made on the potential association of the cases with the cancer risk contour. This judgement is just one factor in a weight-of-evidence approach that evaluates all relevant data, rather than a purely quantitative approach in which assumptions and uncertainties are often not represented (9).

The properties of point distribution can be utilized to evaluate the distribution of cases in many and varying ways. For instance, density can be used to establish association when the point density inside a critical contour is higher than the point density outside the contour. However, this measure should not be taken by itself. It must be normalized by population to yield a valid assessment. Furthermore, the spatial arrangement must be evaluated to completely assess the exposed population. For example, the clustering of cases at one or more points within the critical contour might be related to the location of a retirement or assisted-living home, not the exposure of a population to carcinogens. Finally, a comparison of the spatial arrangement of the cases inside and outside the critical contour can be enlightening. A marked difference (Figure 5), in which a scattered pattern exists inside the critical contour and a random or clustered pattern exists outside, could be an indication of a potential association of cases with chemical or substance exposure.

The distribution of cancer point cases can also be used to eliminate any association and accompanying public fears. For example, if the majority of cases lie outside the critical contour, it would be safe to infer that the polluter might not be emitting quantities



Source: ATSDR, 1998.

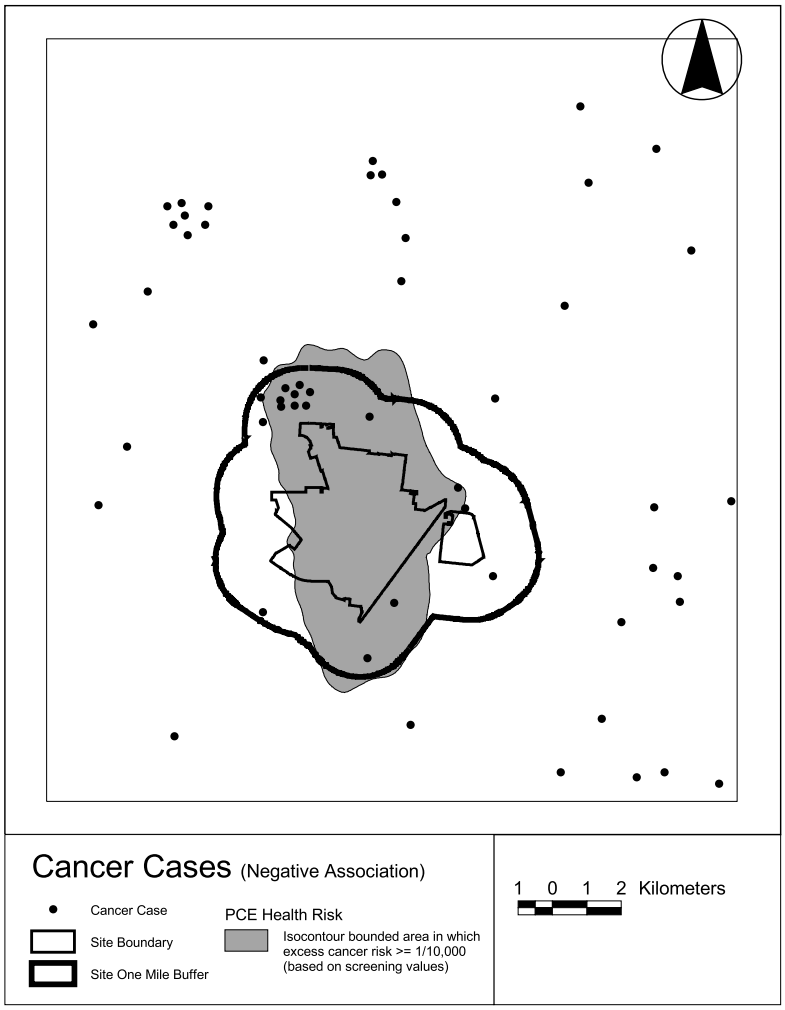
Figure 5 Positive association scenario.

of pollutants in a sufficient amount to be dangerous. In addition, as illustrated in Figure 6, a pattern that remains the same inside and outside the critical contour indicates that the cases may not be associated with exposure. Finally, a uniform density (normalized by population) can also be a reason to conclude that emissions may not be a factor in the health outcome.

The key to evaluation of case locations in the context of a critical risk contour is to consider the totality of many elements including, but not limited to, spatial distribution of points, population density, and population characteristics.

Conclusion

GIS is a fast-developing technology with an ever-increasing number of applications. Air



Source: ATSDR, 1998.

Figure 6 Negative association scenario.

dispersion modeling is a well-established discipline that can produce results in a spatial context. The marriage of these two applications is optimal because it leverages the predictive capacity of modeling and the data management, analysis, and display capabilities of GIS.

In the public health arena, exposure estimation techniques are invaluable. The use of air emission data, such as TRI data, and air dispersion modeling with GIS, enable public health professionals to identify a potentially exposed population, estimate the health risk burden of that population, and attempt to correlate point-based health outcome results with estimated health risk. As demonstrated in this paper, the GIS platform provides a means by which air dispersion modeling results can be effectively applied to public health studies.

References

1. Folgert D. 1998. An investigation of dynamic simulation: The integration of ARC/INFO with air dispersion modeling to facilitate data handling and visualization. In: *ESRI Eighteenth Annual User Conference proceedings*. Environmental Systems Research Institute, Inc., Redlands, CA.
2. Schulze RH, Turner DB. 1996. *Practical guide to atmospheric dispersion modeling.* Trinity Consultants, Inc., Dallas, TX.
3. Agency for Toxic Substances and Disease Registry. 1998. Appendix A: Air exposure pathway. In: *Draft public health assessment: Kelly Air Force Base, San Antonio, TX*. Atlanta, GA: ATSDR.
4. Star J, Estes J. 1990. *Geographic information systems: An introduction.* Englewood Cliffs, NJ: Prentice-Hall, Inc. 18.
5. Environmental Systems Research Institute. 1997. Groundbreaking video introducing the world's first GIS now available. *ARCNews* 19(2):33.
6. US Bureau of the Census. 1991. *Census of population and housing, 1990: Summary tape file 1A and summary tape file 3.* Washington, DC: US Bureau of the Census.
7. Agency for Toxic Substances and Disease Registry. 1996. *Air impact assessment during site investigation.* Health assessment training module. Atlanta, GA, June 17–19. Atlanta, GA: ATSDR.
8. Chou, Y. 1997. Surface analysis. In: *Exploring spatial analysis in geographic information systems.* Santa Fe, NM: OnWord Press.
9. Agency for Toxic Substances and Disease Registry. 1993. *ATSDR cancer policy framework.* January. Atlanta, GA: ATSDR.

Assessing the Accuracy of Geocoding Using Address Data from Birth Certificates: New Jersey, 1989 to 1996

Mark C Fulcomer (1),* Matthew M Bastardi (2), Haniya Raza (1), Michael Duffy (1), Ellen Dufficy (1), Marcia M Sass (1)
(1) New Jersey Dept. of Health and Senior Services, Center for Health Statistics, Trenton, NJ; (2) New Jersey Dept. of Treasury, Office of Telecommunications and Information Systems, Trenton, NJ

Abstract

With the widespread availability of low-cost geographic information systems (GIS) on microcomputers, there has been growing interest in linking sociodemographic variables from census and other sources to vital records for individuals (e.g., from birth and death certificates). Such linked data sets would especially assist investigations into factors contributing to adverse reproductive outcomes (e.g., very low birth weight, infant mortality) and other health events in small areas. Address standardization software with built-in geocoding features offers particular promise in appending data from locations such as census tracts. Because successful linkages of social area and individual levels of data rely on accurate geocoding information, this presentation examines the quality of address data from a large, population-based vital records system. Expanding on an earlier report that studied adverse reproductive outcomes for 1985 to 1988, this paper describes New Jersey's efforts to assess the accuracy of locational data reported on its 1989 to 1996 birth certificates (N=971,592) with that resulting from the application of an address standardization procedure (N=951,895). At the municipality level the agreement between geocoding from address standardization and the certificates was 91.68%, while for 870,149 (91.41%) of the records the census tract or block group could be identified. Before the results could be compared, preliminary work of reviewing and correcting some records was required, especially for post office boxes and rural delivery addresses. Because many records fall into areas spanning multiple municipalities, the results will affect linkages of zip code information for municipalities. Specifically, with considerable confusion between zip code and municipality boundaries, methods to minimize misclassification errors will have major implications for projecting school enrollments and estimating health outcomes.

Keywords: address, geocoding, census

Introduction

The major purpose of this paper is to describe the accuracy and utility of birth certificate mailing address data in geocoding municipalities (also known as minor civil divisions, or MCDs) compared with traditional coding of MCDs using mother's residence also listed on the same vital record. Analyses were conducted for the birth years 1989 to 1996, following the implementation of New Jersey's variant of the national standard certificate. This paper introduces a problem that came to the attention of the New

* Mark C Fulcomer, New Jersey Dept. of Health and Senior Services, Center for Health Statistics, Health and Agriculture Bldg, Room 405, Trenton, NJ 08625-0360 USA; (p) 609-984-6702; (f) 609-984-7633; E-mail: mcf@doh.state.nj.us

Jersey Department of Health and Senior Services (NJDHSS) Center for Health Statistics (CHS) early in 1991—that of confusing results arising from different methods used for assigning geocodes. The paper then describes a multi-step process followed to improve the quality of addresses and other information on its birth certificates, including the introduction in 1995 of the most ambitious electronic birth certificate (EBC) system in the entire country (1).

New Jersey, along with several other states, has a long history of reporting births and other health outcomes at the municipality level (2). Typically, these reports are based on statistical analysis of the numeric values (often referred to as geocodes) of such areas as recorded on vital records, with little attention to how accurately the coding process reflects the actual MCDs. Although there has been a growing interest in using sophisticated geographic information systems (GIS) to carefully pinpoint the residences of cases for cluster investigations and other epidemiological studies of possible environmental exposures (3), these efforts have usually focused on relatively small geographic areas such as a few zip codes or census tracts. The emerging capabilities, however, of address standardization software procedures with GIS features has made it feasible and inexpensive to expand the computerized geocoding of events to much larger areas. This report presents some results from what appears to be among the first large-scale, population-based studies (i.e., data from several years for an entire state) comparing the underlying accuracy of traditional manual geocoding of MCDs on vital records with that found automatically through address standardization. While perhaps giving initial impressions of being specific to New Jersey, some of the confusions between address data and traditional geocodes are likely to be encountered in other areas with large populations, especially as software packages for address standardization and GIS are more widely applied.

Background

New Jersey's efforts to improve the quality of its geocoded information on the state's municipalities began early in 1991, shortly after CHS was reorganized under its current director. Because the funding of New Jersey's schools relies heavily on taxes collected and administered at the local level (all 566 of its current MCDs are incorporated and there are 611 distinct school districts), there is considerable interest in projecting school enrollments to estimate future classroom construction needs using cohort survival methods. As a result, representatives of many of the districts call CHS for the latest official birth statistics for selected municipalities and counties in order to develop their projections.

In the process of responding to requests from school planners for the then newly available data from 1989 birth certificates, CHS staff encountered several instances of enormous changes between birth figures for 1988 (and earlier years) and 1989. These shifts appeared to have been due to the introduction of a new birth certificate form for 1989 births. Table 1 shows shifts in birth figures that involved an apparent 241% increase in births between 1988 and 1989 in a small town with a relatively high median age, while there was a corresponding drop of births (-86%) for the same two years in a surrounding municipality that shared the first town's zip code. Although it involved smaller percentage changes in the birth attribution between 1988 and 1989 for two adjacent municipalities (28% and -34%, respectively), the second example shown in

Table 1 Examples of Shifts in Birth Figures Associated with the Introduction of New Certificates in 1989

EXAMPLE #1													
		Number of Births Per Year											
MCD	Code Type	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
1A	Old	72	57	89	99	338	334	90	75	82	89	80	77
	New	—	—	—	—	91	99	63	65	66	72	71	68
1B	Old	130	161	218	203	28	61	274	295	273	254	258	229
	New	—	—	—	—	245	265	295	288	274	251	249	225
EXAMPLE #2													
		Number of Births Per Year											
MCD	Code Type	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
2A	Old	1,888	1,951	2,061	2,136	2,728	2,614	2,624	2,072	1,805	1,663	1,569	1,549
	New	—	—	—	—	2,112	2,015	2,056	1,926	1,777	1,633	1,518	1,521
2B	Old	1,003	1,034	1,081	1,122	740	794	737	1,010	1,103	993	968	1,004
	New	—	—	—	—	1,155	1,190	1,134	1,129	1,121	1,032	1,026	1,025

Old=Traditional vital records geocoding of the mother's residence municipality as reported on the birth certificate
New=Coding of residence municipality based on mother's mailing address as reported on the birth certificate

Table 1 represents an even more dramatic situation, because the first municipality had to account to state government for nearly $9,000,000 in excess payments it had received based on faulty enrollment projections (4).

The fundamental problem that needed to be resolved was the confusion between representing a mother's mailing address as a postal address versus representing the same address relative to the boundaries of the municipality/MCD. This confusion can be visualized by overlaying MCDs within a county on top of zip code boundaries for the postal deliveries to those same areas. Especially vivid are overlays that display zip codes overlapping both municipality and county boundaries, a situation that is not unique to New Jersey.

While participating in the 1989 nationwide implementation of a new standard birth certificate, New Jersey had contributed to the blurring of MCD and zip code boundaries by inadvertently placing the birth certificate query for the mother's mailing address before that for her municipality of residence. Although this seemingly minor reversal was soon corrected in the printed birth certificates (in mid-1991), considerable confusion persisted, in part because the collection of many additional items on a multi-part form (instead of the more compact, 5- by 8-inch version used prior to 1989) had altered the methods used by hospitals to prepare typed versions of the certificates. In particular,

the new certificate forms could not easily be copied for use by parents/informants in completing selected items as had been done with the forms for earlier years. Instead, many hospitals began to substitute the postal city of the mother's mailing address for the municipality of residence, often without questioning the mother at all.

Beyond their role in projecting school enrollments and classroom needs, birth figures have had several important health applications reported on by CHS staff and associates. Especially noteworthy are those instances in which live births provide denominators for the calculation of rates for analysis of geographic variations, including low birth weight and inadequate prenatal care, as well as other birth-related characteristics such as infant and fetal mortality (2); case-control studies of environmental exposures and adverse reproductive outcomes (3); and cluster investigations (5). Of course, an even more important consequence of any misclassification of events in adjacent areas is the possibility that the "numerator" characteristics involved in the calculation of such rates may also be affected when the records of some individuals are assigned to different municipalities. For example, those infants who were determined to be inappropriately geocoded to municipality "2A" in Table 1 had a mean birth weight that was approximately 150 grams higher than those births for whom the original geocode was unaltered (6); that is, the misallocation of cases would paint a much more optimistic picture of the health status of infants in "2A" than would be warranted.

In retrospect, given that it is the nation's most densely populated state and has a tradition of local home rule dating back to the American Revolution, New Jersey's geocoding difficulties are hardly surprising. In contrast to the generally rectangular partitioning of other parts of the country that later became part of the emerging nation (e.g., the Northwest Territory), the boundaries of New Jersey's counties and municipalities are irregular and often lack readily identifiable physical demarcations such as rivers or roads. Other confusions stem from instances of duplicate municipality names in different counties, in recognition of important Revolutionary War figures (e.g., 6 instances of Washington Townships, 4 Franklin Townships, etc.). There are also confusions that occur in about 25 pairs of adjacent municipalities (e.g., Princeton Boro and Princeton Township) in which the post office serving a central area also delivers mail to a surrounding MCD with a nearly identical name. Furthermore, with a long and colorful history, New Jersey has about 3,600 small areas that are known by local names, most of which are not municipalities and, therefore, lack any official governmental status or well-defined boundaries, despite their sometimes distinctive-sounding names. Toms River, a part of Dover Township in Ocean County and site of an ongoing childhood cancer cluster investigation, is perhaps the most well-known recent example of these local name areas.

When one considers New Jersey's municipalities and local name areas in conjunction with the postal zip codes serving them, the basis of geocoding confusion becomes even more apparent. Based on the "good" addresses employed in this report—those that met post office certification standards and could be geocoded unambiguously by census TIGER boundaries with no alterations, the state's MCDs are overlapped by 624 zip codes representing 659 different postal city names. Because postal delivery routes are not required to correspond to other geopolitical entities and are sometimes changed to improve service, a large number (392) of the state's zip codes cross municipality boundaries, sometimes even crossing counties in the process. As part of early work on this study, 360 of New Jersey's municipalities were identified as being affected by a

substantial sharing of mail delivery routes with neighboring communities, while 42 of its zip codes cross into a second county and 5 of those cross into yet a third county. Although the sharing of zip codes most often affects pairs of municipalities (152 affected), there are instances of zip codes that serve as many as twelve or thirteen different MCDs (2 and 1 affected, respectively). Finally, there were 1,932 combinations of postal cities, zip codes, and municipalities that accounted for the locations of the good records in the present study. (Note that ongoing population growth/migration and the closing of smaller post offices are among factors likely to lead to the future recognition of similar confusions between zip code and MCD boundaries in other states.)

Steps to Improve Quality

Once the scope and nature of the geocoding confusion on New Jersey birth certificates became apparent in early 1991, CHS and NJDHSS's Bureau of Vital Statistics (BVS) began coordinating a series of steps to improve the quality of the state's locational data for these records. Early steps concentrated on the traditional geocoding of municipalities prepared by BVS. Then, after CHS gained access to address data from birth certificates (also based largely on work performed by BVS), efforts shifted toward making such information more useful by geocoding with greater accuracy, not only at the municipality level but also for smaller areas such as census block groups. (Clearly, some of these steps could be replicated elsewhere.)

Hospital Visits

The hospital accounting for the vast majority of the confusion in the first example cited above was visited on three separate occasions. Working with the local registrar in that municipality, hospital records were inspected and birth records were amended to reflect the actual municipalities of residence cited (versus the apparent municipalities reflected in the postal cities listed in mailing addresses). While such a process would be labor-intensive if done on a statewide basis, these initial visits gave valuable insights into subsequent efforts to improve geocoding accuracy.

Change Order of Birth Certificate Items

Simultaneous with the hospital visits, the order of the mailing address and residence municipality items on the birth certificates was reversed. Unfortunately, this change was probably only minimally effective because it was made long after the new data collection methods had been introduced by hospitals to account for the vast increase in information collected on the new 1989 certificates. However, as a new cycle of national standard certificates are introduced in the near future, it is hoped that more attention will be paid to improving and standardizing the acquisition of some key pieces of information such as residential locations and race/ethnicity. A particularly critical change would seem to be the inclusion of direct, personalized probing of sensitive information (versus relying on mailing addresses and visual attributions of race and ethnicity).

Comparison of Statistical Results to Street Maps

By the summer of 1991, CHS had completed the process of inspecting street maps for the state's 21 counties to better understand the changes in geocoding results between 1988 and 1989 births for adjacent municipalities. In the absence of a computerized

mapping capability, this step was more difficult than initially envisioned, especially when using out-of-date or incompatible maps to inspect zip codes that crossed county boundaries. As a result of the statistical/map comparison work, however, 360 municipalities were identified as having substantial overlapping of zip codes from nearby communities.

Design and Implementation of Birth Certificate Worksheets

By the fall of 1991, CHS had designed, pilot-tested, and revised four-part worksheets to improve the quality of birth certificate data by standardizing its collection. Central to this effort was a parents' information sheet that concentrated on carefully ascertaining the mother's residence and mailing addresses as well as other items relating to race and ethnicity. The worksheets were implemented statewide through regional training sessions in the spring of 1992 and provided the basis for the eventual introduction of New Jersey's ambitious EBC system in 1995. The second example in Table 1 indicates how the worksheets had an apparent impact on improving the traditional coding of municipalities as early as 1992 and 1993, long before any work on address standardization had commenced.

Interactions with Local Registrars, Hospital Personnel, and School Officials

CHS initiated discussions with and sought feedback from local registrars and hospital personnel to improve the design of the birth certificate worksheets. Being responsible for critical components of the birth certificate process, these groups were seen as key players in a data quality improvement effort. Later, as part of attempts to explain oftentimes large variations in birth figures over time, there were also hundreds of interactions with school superintendents and planners, clearly underscoring the complexities of this project, especially with respect to predicting school enrollments.

Initial Inspections of Addresses for In-State Records

By the winter of 1992, CHS had gained access to computerized files of mothers' mailing addresses on birth certificates. Although it was quickly realized that considerable effort would be required to correct keypunching errors and parse the information into separate fields (e.g., street numbers and names) in order to perform a meaningful analysis, the ability to eventually access improved versions of these data reinforced the efforts to implement the worksheets. Later in 1992, the design of the EBC system began and included an early commitment to standardize the collection of addresses and other important information.

Discovery of Address Standardization with Census Geocoding

The major breakthrough in this project came in the winter of 1994 when, as part of a general interdepartmental discussion with the New Jersey Department of Environmental Protection on how sophisticated GIS techniques might be applied to the birth geocoding problem, it was discovered that a sister New Jersey agency (OTIS, the Office of Telecommunications and Information Systems) could support access to a state-of-the-art address standardization software package, Finalist/FinalFocus. As a tool for achieving valuable postal discounts through the assignment of zip+4 codes, the software would provide important data-cleaning and parsing features. Even more important, a little investigation soon revealed that census tract and block group

identifiers were not only employed internally as part of the standardization procedure (along with latitudes and longitudes), but they would also be returned as part of New Jersey's acquisition of the software package. This breakthrough meant that, for the first time, there was the prospect of a computerized procedure that could standardize "messy" address data and automatically assign geocodes to records.

Acquisition of Addresses for 11,509 Out-of-State Births from Pennsylvania for 1989 to 1993

Once the address standardization procedure became available, CHS's attention turned to the acquisition of mailing address information that had previously been missing, beginning with that for out-of-state events. A special arrangement made available an electronic file of addresses for New Jersey resident events occurring in Pennsylvania for the years from 1989 to 1993, thereby eliminating the need to re-key this information.

Data Entry of Birth Addresses for Additional Births

Beginning in the summer of 1994, CHS staff began the process of keying previously missing address information, eventually accounting for the entry of such data for an additional 10,242 in-state births. Note that, prior to this step, mothers' mailing addresses were entered when a social security card was requested for a child, covering 96.28% of the births in New Jersey. For 1991 and 1992, this data entry work was concentrated on records from the 360 municipalities with identifiable geocoding confusion. Unfortunately, because interstate agreements limit how long exchanged vital events information may be retained, out-of-state records for the 1989 and 1990 birth-years were no longer available (except those already supplied by Pennsylvania, as described above), so that no additional address information was initially keyed for those two data years. More recently, all address information for New Jersey births has been computerized; CHS completed this work for the 1993 birth year, while BVS made this part of its routine activities beginning in 1994.

Because of the importance of tracking geographic variations in infant mortality, CHS staff then found and entered previously missing birth certificate address information for 754 infant deaths. Much of these data were missing due to many of these events occurring soon after birth, so that a social security card would never have been requested. In turn, this led to a decision by CHS staff to key address information for the remaining 9,277 in-state events with previously missing computerized data for which a paper certificate could be found, regardless of the geocoded municipality or data year. As a result of keying in previously missing data (much of it from larger cities that were not initially identified as being affected by geocoding confusions), addresses were available for virtually all in-state births for the entire study period from 1989 to 1996.

Initial Results and Corrections/Handling: 1989 to 1994

By the fall of 1996, the initial results for births from the 1989 to 1994 data years became available. Although the standardization process provided helpful guidance on how to inspect and improve addresses (e.g., handling rural delivery routes and post office boxes), these early results were especially encouraging in that they indicated how what had appeared to be very "messy" data could be automatically geocoded in about 90% of the cases. This hands-on experience provided an opportunity to visually inspect 26,459 records in which zip codes were changed by the software to achieve SOUNDEX

matching to the street addresses as well as the rendering of what initially appeared to be 9,633 rural delivery routes and 11,638 post office box addresses. Some suggestions on improving the parsing of addresses and the correction of a few census TIGER bugs were passed on to the software vendor and became part of the regular quarterly updates.

Subsequent Software Enhancements

Based on the initial results, major enhancements of the geocoding aspects of the address standardization process were undertaken in the winter of 1997. Instead of leading to an overall latitude/longitude assignment for a census block group to which a good address would have been assigned under the previous version, the enhanced procedure provided interpolated values (over the range in a street segment) for individual records to compare with census TIGER boundaries at the census block level. (The pair of latitude/longitude values for an entire census block group were still inserted when an address could only be matched at the zip+4 level.) To improve address matching through the use of large postal and census databases to augment the Finalist/FinalFocus results, the new procedure also (a) incorporated alternate street names; (b) accounted for street numbers beyond current TIGER file limits (e.g., for newly constructed homes); (c) corrected a "county-road problem" (i.e., by assigning rural deliveries to the county in which the residence was located in those instances where the mail box was on the opposite side of the road and, therefore, in another county); (d) provided more "return codes" to facilitate analysis of the computerized records; and (e) reduced the number of addresses falling into areas spanning multiple municipalities, because census blocks typically do not cross MCD boundaries. (Note that, beyond the initial cost to acquire the Finalist/FinalFocus package, the minimal costs for these enhancements were the only other direct costs for this entire project.) All of the 1989 to 1994 addresses were then re-processed using the new procedure. By the summer of 1998 it was also possible to complete the processing of the 1995 and 1996 certificates.

Acquisition of Addresses for 2,505 Out-of-State Births for 1996

The final quality improvement step involved an acquisition of 1996 addresses for New Jersey residents born in New York City. This step was facilitated by both jurisdictions implementing EBC systems that year using software developed by the same vendor. Much like the earlier acquisition of out-of-state data from Pennsylvania, this step also meant that the addresses for the 2,505 cases could be used directly without re-keying.

Results

This section describes the basic results obtained over the eight years of birth data covered by this entire project, beginning with simple summaries of the acquisition of the address information and efforts to standardize it. After highlighting efforts to attach locational indicators from census data to the standardized address data, the report concludes with comparisons of geocoding accuracy by the two major methods employed.

Acquisition of Address Information

Table 2 shows the results of efforts to acquire (and key) address data from the 971,592 births covered by the eight-year period from 1989 to 1996. Although the project had

Table 2 Births by Data Acquisition Source and Year

	No. of birth records obtained								
Data Source	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**Total**
BVS	115,917	117,228	116,425	115,660	115,146	114,429	117,155	113,866	925,826
PA/NYC	2,377	2,360	2,300	2,214	2,258	0	0	2,505	14,014
CHS/BVS	0	0	971	1,048	2,741	5,482	0	0	10,242
CHS/ID	199	167	202	185	1	0	0	0	754
CHS	2,757	2,733	2,083	1,632	72	0	0	0	9,277
Unavail.	3,401	3,492	2,233	2,321	24	8	0	0	11,479
Total	124,651	125,980	124,214	123,060	120,242	119,919	117,155	116,371	971,592

BVS=Bureau of Vital Statistics (NJ)
PA/NYC=State of Pennsylvania/New York City
CHS=Center for Health Statistics (NJ)
CHS/BVS=Joint effort by CHS and BVS
CHS/ID=CHS-provided statistics on infant deaths; records missing birth certificate address information

started with no address data whatsoever, Table 2 clearly demonstrates that by its conclusion such information had been attached to all except 11,479 (1.18%) of the records. Even more importantly, there is essentially a complete accounting of address data for the last four years.

Of the data sources presented in Table 2, BVS clearly accounted for the vast majority (95.29%) of the addresses. Prior to the pilot testing of the EBC (at four hospitals in 1995) and its statewide implementation beginning in 1996, the data entry of addresses by BVS was done for those infants for whom social security cards had been requested. Note that the increase in 1995 reflected BVS's keying of addresses for all births prior to the full implementation of the EBC, a process completed by April 1997, so that almost all address information is now provided directly (i.e., without any data entry by BVS) by the 71 birthing facilities in the state.

The number of addresses acquired from Pennsylvania and New York is shown in Row 2 of Table 2 and, at first glance, would appear to represent only a small portion (1.44%) of the total. However, beyond saving the effort involved in re-keying records for out-of-state events, such interstate exchanges hold great promise for improving the timeliness with which population-based vital statistics become available in the future, especially if this data sharing can be expanded to include all states and other data (e.g., death certificates).

Joint efforts by CHS and BVS to key some missing records are highlighted in Row 3 of Table 2. This work began with CHS inputting data for 1991 and 1992 certificates from those municipalities with already identified geocoding problems and was carried over to all remaining records in 1993 and beyond (by CHS and BVS, respectively).

Row 4 of Table 2 lists the number of records entered by accounting for the 754 infant deaths with previously missing birth certificate address information (i.e., in addition to those infant deaths with data already summarized in the first three rows). Because virtually all addresses were keyed after 1992, all except one of these cases

occurred between 1989 and 1992. Similarly, CHS's efforts to account for the 9,277 remaining in-state records with previously missing addresses are shown in Row 5 of Table 2.

Finally, Row 6 of Table 2 displays the number of records that were no longer available to provide any address information, primarily from those out-of-state events in 1989 to 1993 that did not occur in Pennsylvania (i.e., those records already accounted for in Row 2). Most of the unavailable addresses were from births that occurred in New York City and came from densely populated areas of Bergen and Hudson counties, where geocoding confusion was not as severe as in other areas of the state.

Address Standardization and Matching

Once data entry work was completed for the 960,113 births for which address information was available, the records were assembled into smaller files and transmitted to OTIS for initial processing. The results of the address standardization procedure were then used to separate records into three major groupings: (1) those that could be matched automatically to a known address and would require no further work (i.e., so-called "good" results); (2) those that could be matched automatically, but only after the Finalist/FinalFocus portion of the procedure changed an address using SOUNDEX matching or other five-digit codes within a three-digit zip code area; and (3) those that could not be matched to an address automatically (i.e., so-called "bad" results). CHS inspected all records in the second group and compared the address matching results after the changes were made with the original information. In those cases in which the changes were not considered acceptable, an attempt was made to edit the address on a record and mark it for resubmission, especially because rather obvious errors (e.g., the failure to join together the number and letter of an apartment such as "2C", leaving it instead as "2 C") can lead to some rather surprising matches (e.g., "2 C Street"). Although the records with bad addresses represented only a small portion of the total, they were too numerous and complex to allow editing to be completed effectively. Instead, those bad records for which there was substantial agreement between the postal city, zip code, and geocode for a given MCD (i.e., when compared with good records for the same area that could be matched to known addresses and geocoded automatically with no alterations) were separated from those that needed further inspection and editing. (Note that some records with out-of-range but consistent street numbers now had interpolated latitude/longitude values attached to them by the new procedure, making them equivalent to those with good results.) Any records marked for editing, whether initially identified by the software as changes or as bad results that could not be matched to a zip+4 area, were then prepared for resubmission to OTIS and the process was repeated with much smaller files. Any records that did not produce acceptable matching after a second attempt were then treated as bad (or problematic) results.

Overall, good matches to known addresses with no alterations were returned by the new procedure for 857,261 (88.23%) of the records. For 39,196 (4.03%) records, the software indicated that some changes were needed to match an address (not all of them were accepted by CHS), while 63,656 (6.55%) were initially identified as unmatchable. In large part, the success in achieving good matches was due to the large number of records (934,746, or 96.21%) viewed as having conventional addresses, including many of which that had been treated as rural deliveries in the initial processing done in 1996.

Rural deliveries (10,224, or 1.05%) and post office boxes (14,946, or 1.54%) accounted for all but 197 of the remaining records with address data.

Attaching Locational Indicators from Census Data

Based on the results of its standardization and matching steps, the new OTIS procedure also attached census identifiers to the 951,895 records with New Jersey addresses (or that were originally coded as in-state residents by BVS when addresses were unavailable). These records are summarized later in the first four rows of Table 3. The remaining 19,697 (2.03%) records had addresses outside of New Jersey (or were coded as out-of-state residents by BVS when addresses were unavailable) and are listed later in Rows 5 and 6 of Table 3. The census locational indicators included tracts, block groups, and blocks. Latitudes and longitudes associated with an address were also attached to the records. With the new OTIS procedure, interpolated latitude and longitude values could be found for many of these records using the census TIGER file limits based on block-level matching. The attachment of a single pair of values for an entire census block group (i.e., geocoding based on the earlier Finalist/FinalFocus procedure) now took place when a census block could not be assigned and an address could be matched only at the zip+4 level.

Table 3 Births by Level of Accuracy and Year of Birth

Level of Accuracy	1989	1990	1991	1992	1993	1994	1995	1996	Total
1 Same MCD	109,902 88.17%	111,510 88.51%	110,705 89.12%	111,489 90.60%	108,783 90.47%	108,126 90.17%	105,799 90.31%	106,400 91.43%	872,714 89.82%
2 Same county	10,169 8.16%	10,215 8.11%	9,560 7.70%	7,835 6.37%	7,917 6.58%	8,317 6.94%	7,749 6.61%	6,520 5.60%	68,282 7.03%
3 Diff. county	1,569 1.26%	1,325 1.05%	1,167 0.94%	1,173 0.95%	1,136 0.94%	1,233 1.03%	1,366 1.17%	1,252 1.08%	10,221 1.05%
4 OOS to NJ	63	68	103	56	75	74	139	100	678
5 NJ to OOS	2	3	2	1	4	8	0	0	20
6 Both OOS	2,946 2.36%	2,859 2.27%	2,677 2.16%	2,506 2.04%	2,327 1.94%	2,161 1.80%	2,102 1.79%	2,099 1.80%	19,677 2.03%

OOS=Out-of-state

After the new procedure matched an address, census locational codes at the tract, block group, block, and other levels were also linked to the records, including a county code and one or more MCD codes. The census values for counties and municipalities have a one-to-one correspondence with the BVS geocodes for the same areas and, therefore, the two sets of geocodes can be used interchangeably. In contrast, because they can cross one or more municipality boundaries (but not county boundaries), census tracts and block groups (i.e., subsets of tracts) may sometimes be shared among multiple MCDs. Fortunately, blocks are generally associated with a single municipality.

Given that it had not even been considered possible at the outset of the project, the geocoding of records to census tracts and block groups has been extraordinarily successful. Of the records for New Jersey residents, 848,189 (89.11%) could be coded to the

block group level (which includes tracts) and an additional 21,960 (2.31%) to the tract level only. This result is especially important in that it makes possible the attachment of income and other sociodemographic indicators from social areas, otherwise missing from individual-level records such as birth certificates, in "semi-ecologic" studies of adverse reproductive and other health outcomes (2) or other social area analyses (7).

The use of census locational identifiers, done in conjunction with ranges of street addresses, was also very successful in geocoding records at the municipality level, especially because all except nine of New Jersey's 566 MCDs have boundaries in the TIGER files. With respect to single municipalities, 855,286 (89.85%) of the records for New Jersey residents came from census designation areas that did not cross into another MCD and were mostly geocoded at the block level, except for 105,083 records at the block group level only. For those records that could be geocoded by census indicators but fell into block groups spanning two municipalities (10,169) or three (1,334), the original BVS geocode was relied on as much as possible. Thus, when the BVS geocode derived from the official birth certificate matched one of the possible MCD codes based on the census identifiers, regardless of its position as a member of a pair or triplet, that municipality code was used. This was done for 7,511 of the pairs and 455 of the triplets. Of the remaining 3,537 records falling into census block groups spanning multiple MCDs that did not have a matching BVS code, the municipality codes based on the census identifiers were randomly assigned (using the codes' sequential position in pairs or triplets of possible MCDs for series of similarly sorted records). Again, when contrasted with the early stages of the project, the ability of the address standardization procedure to automatically assign a municipality code to 91.06% of the records in a rational fashion is noteworthy. This result also supports the use of GIS software to display maps of birth figures at the census tract and block group levels, provided that careful attention is paid to protecting the confidentiality of individuals when the number of events in a submunicipality area is small.

Partial matches to the BVS municipal code were established for 67,822 (82.97%) of the 81,746 records for New Jersey residents that could not be automatically matched to census tracts or block groups. This was done using postal cities and zip codes for the same municipalities found in the good results coded to single municipalities. (In the future, these partial matches will be handled through AUTOMATCH [a procedure described in Jaro 1989 (8)], which should improve the geocoding process even more.) As a consequence, only a small number of records (13,924, including those for which vital events information was no longer available, as mentioned earlier) lacked similar concordance between the geocodes from traditional methods and those based on matching address information to huge postal and census databases. For this set of records, only the original BVS geocodes could be used.

Comparing the Agreement between Geocoding Methods

This section describes the agreement between the traditional coding of MCDs of mothers' residences with that based on standardization and matching from address information. Table 3 shows six levels of accuracy used to assess the agreement across the eight-year period from 1989 to 1996. Row 1 highlights the high overall accuracy (89.82% of all records; 91.68% of New Jersey residents) between the two geocoding methods in assigning records to the same municipality. In general, same-municipality agreement has improved over time, especially in 1996 as the EBC was being implemented.

Discrepancies that resulted from the two methods assigning geocodes to different municipalities within the same county are listed in Row 2 of Table 3. The percentage of within-county discrepancies has also diminished over time, perhaps reflecting both the early introduction of the worksheets as well as special features of the EBC software (e.g., pull-down lists of municipalities within counties) because the improvements were most pronounced in 1992 and 1996. An improvement in 1995-1996 data with respect to matching addresses at the census tract or block group levels (95.97% of New Jersey residents versus the overall eight-year average of 91.41%) also likely traces its origins to the better acquisition and keying methods introduced at the birthing facilities as part of the EBC. Taken together, Rows 1 and 2 of Table 3 indicate a relatively high level of accuracy of geocodes at the county level (96.85% of all records; 98.86% of New Jersey residents).

Row 3 of Table 3 shows the number of discrepancies between the two geocoding methods in assigning records to different municipalities in different counties. Unfortunately, the overall percentage of discrepancies between counties is substantial (1.05% of all records; 1.07% of New Jersey residents) and has remained essentially unchanged over the entire eight-year period. Further work, including continued emphasis on increasing the interstate exchange of data, will certainly be needed to understand how and where such errors occur and how they might be ameliorated in the future.

Row 4 of Table 3 shows the number of discrepancies that were originally given out-of-state codes by BVS but which were geocoded as in-state residents using the address data. This relatively minor level of disagreement seems to occur most frequently with births in military families, where permanent homes may be in another state while the use of schools and other resources happens in New Jersey. Row 5 shows the extremely small number of discrepancies that were geocoded as being residents of other states based on the address data but had originally been treated as New Jersey residents by BVS. Finally, Row 6 lists the agreement (2.03% of all records) between the two methods in geocoding records to other states.

Summary

Because it involved nearly one million births over the eight-year period from 1989 to 1996, the entire address standardization/analysis effort was a large and complex undertaking. The ability, however, to successfully resolve what was a great deal of initial confusion has been very gratifying, especially given that the methods can be used elsewhere and that the results have some important applications that were not envisioned at the outset. In particular, the efforts to improve data quality, link records to other data sources (e.g., income from the census), and achieve more timely and automatic geocoding hold great promise for the future.

Nonetheless, despite the clear-cut benefits of automatic geocoding based on the application of standardization and matching techniques to address data, the relatively high disagreements with traditional manual methods at the municipality level are disturbing. The fact that the discrepancies are so large (i.e., regardless of whether they occur in the same or different counties)—nearly 7% even for the most recent year (1996) with the EBC undergoing its full implementation—casts a cloud over the exclusive use of manual methods to geocode residence locations, especially in situations in which address data might be available to facilitate comparisons with results from automatic

geocoding. A frightening possibility is how easily "faulty" numerators or denominators from manual geocoding could be employed in the calculation of rates for infant mortality (or other "rare" outcomes) in small areas. Thus, while address data can clearly be helpful in assigning events to small areas, much more work on understanding and improving geocoding methods remains to be done.

References

1. Knoblauch KL, Sherel HP. 1998. *New Jersey electronic birth certificate perinatal database data dictionary*. New Jersey Department of Health and Senior Services, Center for Health Statistics, Trenton, NJ.
2. Fulcomer MC, Bove FJ, Klotz JB, Siegel B, Martin RM. 1992. Developing and utilizing "community health profiles" based on linked information on adverse reproductive outcomes. *Proceedings of the 1991 conference on records and statistics.* US Department of Health and Human Services. DHHS Publication No. (PHS) 92-1214. 168–173.
3. Bove FJ, Fulcomer MC, Klotz JB, Esmart J, Dufficy EM, Savrin JE. 1995. Public drinking water contamination and birth outcomes. *American Journal of Epidemiology* 141:850–62.
4. Bewley J. 1996. City faces $9M loss in school aid. The Trenton *Times.* 5 January. A1, A10.
5. Fulcomer MC, Ziskin LZ, France DM, Bove F. 1988. *Report on the study of Vernon Township, NJ: Study of the occurrence of chromosomal anomalies in Vernon Township between January 1, 1975 and June 30, 1987*. New Jersey Department of Health, Division of Community Health Services, Trenton, NJ.
6. Raza H. 1997. *An analysis of the effects of address standardization of live birth certificates on the health and social characteristics of mothers and babies during 1989-1994 in Trenton, New Jersey.* Unpublished master's thesis. University of Medicine and Dentistry of New Jersey, Piscataway, NJ.
7. Struening EL. 1975. Social area analysis as a method of evaluation. In *The Handbook of evaluation research, Volume I.* Ed. EL Struening and M Guttentag. Beverly Hills: Sage Publications. 519–36.
8. Jaro MA. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association* 84:414–20.

GIS Analysis of Firearm Morbidity and Mortality in Atlanta, Georgia

Dawna S Fuqua-Whitley, MA,* Kidist K Bartolomeos, MPH,
Arthur L Kellermann, MD, MPH
Emory Center for Injury Control, Rollins School of Public Health, Emory University, Atlanta, GA

Due to the sensitive nature of the datasets and their use in ongoing police investigation and enforcement activities, maps cannot be included at this time.

Abstract

Collaboration with public health researchers and use of geographic information system (GIS) analysis can help law enforcement agencies identify the times and places where most firearm homicides and aggravated assaults occur. To guide and evaluate the efforts of a local multi-agency task force developed to reduce firearm crime in the city of Atlanta, Georgia, the authors analyzed key local datasets to identify "hot spots" of firearm-related crime. While overall homicide rates declined over the past 10 years in Fulton County, Georgia (which includes most of the city of Atlanta), homicide rates for 15- to 19- and 20- to 24-year-olds increased. All of the increase in these age groups was due to a sharp increase in firearm homicides. Non-firearm homicides have remained stable. GIS analysis of county medical examiner firearm homicide records (1989–1997), City of Atlanta 911 system firearm-related calls (shots fired, person armed, person shot calls for 1997), and reports of aggravated assaults with a firearm (non-fatal shootings) indicated hot spots of firearm-related morbidity and mortality in specific police zones, beats, and census tracts within the city. Within beats and neighborhoods, high-frequency streets, intersections, public housing units, and time periods were identified. Analysis of data according to census tract indicated a higher frequency of events in tracts that were above the state, county, and citywide mean for socioeconomic status indicators such as female-headed household, percent male unemployment, and percent below poverty level. GIS data and trend analyses are reported regularly to the multi-agency task force, where they are used to assist in case investigations and target enforcement operations.

Keywords: firearms, homicide, assault, injury, prevention

Literature Review and Statement of Problem

Homicide is the second leading cause of death for Americans between 15 and 24 years old, and it is the leading cause of death among African-American youth. Among persons 25 to 44 years old, it is the sixth leading cause of death (1). Firearm homicide accounts for an increasing percentage of overall homicide in national figures. In recent years, national statistics have shown a decline in overall homicide. Among juveniles and young adults, however, we have seen a dramatic upsurge since 1985. When broken out by age group, firearm homicide accounts for nearly all homicides among 15- to

* Dawna S Fuqua-Whitley, Emory Center for Injury Control, 1518 Clifton Rd. NE, Atlanta, GA 30322 USA; (p) 404-727-3071; (f) 404-727-8744; E-mail: dfuquaw@sph.emory.edu

19-year-olds. Recent research indicates that although national rates of firearm homicide among juveniles and young adults have declined slightly since a peak in 1993, they still remain high. Non-firearm homicide has remained stable over time.

Public health and law enforcement agencies have an interest in reducing the incidence of firearm homicide and assault, mortality, and morbidity. Two strategies in criminology and law enforcement—problem-solving policing and community-oriented policing—follow tactics similar to public health research and practice. Geographic information systems (GIS) analysis can help target the problem, identify intervention points likely to have the greatest effect, guide implementation of a strategy, and evaluate its impact.

In a publication for the National Institute of Justice, Rich (2) outlines two goals for the use of GIS in the analysis of crime and victimization. The first is to further an understanding of the nature and extent of criminal and social problems in a community. The second goal is to improve the allocation of resources to combat these problems. Once "hot spots" of crime events have been identified, GIS can be used to determine if an intervention or prevention strategy suppressed new events or displaced criminal activity to other locations.

The utility of GIS to identify high-frequency areas of crime events and target law enforcement efforts has been established. Taylor and Gottfredson (3) concluded in 1986 that neighborhoods show different levels of crime across geographic boundaries, and "that there is evidence linking spatial variation in crime to the physical and social environment at the sub-neighborhood level of street blocks and multiple dwellings." Starting with the premise that crime is concentrated in specific areas that are not evenly distributed and that it is more efficient for police to concentrate their efforts on high crime areas, Sherman and Weisburd (4) conducted a one-year randomized trial in Minneapolis of an increased police presence in identified hot spots of crime. They reported that "observed disorder was only half as prevalent in experimental as in control hot spots. We conclude that a substantial increase in police patrol presence indeed causes reductions in crime and more impressive reductions in disorder within high crime locations" (4).

Weisburd and Green (5) analyzed narcotics sales arrests, drug-related emergency calls for service, and narcotics tip line information over a six-month period in Jersey City, New Jersey, to determine hot spot areas of drug activity. Using GIS, they determined that 14% of the city's intersections were the sites of most, if not all, of the drug activity in Jersey City (5). Working with the Jersey City Police Department, which had previously relied on "a series of loosely connected and unsystematic drug enforcement tactics," they designed an experimental strategy to reduce drug and drug-related activity. These investigators reported "consistent and strong effects of the experimental strategy on disorder-related emergency calls for service." They found little evidence of displacement to other areas, and in fact, reported a "diffusion of benefits" to surrounding areas (5).

History of the Project

In 1994, five counties in metropolitan Atlanta joined Project PACT (Pulling America's Communities Together), an ongoing federal violence prevention initiative intended to encourage local governments and federal agencies to work together to identify local

problems and create local solutions. Through Metro Atlanta Project PACT, area leadership and community stakeholders were asked to identify the most pressing violence problems in the project area. The participants identified youth firearm violence as a significant local problem and a top priority for the city. Although the original scope of the project covered five counties (Fulton, DeKalb, Cobb, Clayton, and Gwinnett), efforts were subsequently focused on Fulton County and the city of Atlanta.

Project Objectives

Shortly after Metro Atlanta Project PACT was initiated, a consortium of federal agencies announced their intention to fund evaluations of community-based approaches to reduce firearm violence among juveniles. The Emory Center for Injury Control (ECIC) received funding to provide a baseline assessment of the problem in the project area, provide ongoing process evaluation to help guide and refine the effort, and provide a summary evaluation to determine the effectiveness of Metro Atlanta Project PACT's efforts to reduce juvenile firearm violence in metropolitan Atlanta. The project was funded by the National Institute of Justice, the Office of Juvenile Justice and Delinquency Prevention, and the Centers for Disease Control and Prevention.

The project has three key objectives:

1. With partners, apply a problem-solving approach to developing, implementing, evaluating, and refining a comprehensive youth firearm violence prevention strategy.
2. Determine whether broad-based community action can reduce juvenile firearm violence.
3. Evaluate the utility of retrospective and prospectively collected local data to guide the development and refinement of violence prevention countermeasures.

The firearm mortality, morbidity and emergency call data presented here were collected and analyzed in support of this project.

Methods

The project area, Fulton County, lies in the northwest quadrant of the state of Georgia, claiming 338,364 acres of land. Population density has increased steadily in the past decade; in 1990 there were 1.98 persons per acre, and in 1997, there were 2.25 persons per acre. One of 159 counties in Georgia, Fulton is the most populous, with 760,100 residents in 1997. The county is roughly 50% white and 50% African-American. The population of the county more than doubled between the 1980 and 1990 censuses (6). In 1995, Fulton County was ranked first in the state for per capita personal income, well above the state and national average. Much of the wealth, however, is concentrated in the northern third of the county and the northern half of the city of Atlanta. The southern half of the city and the central and southern thirds of the county have large concentrations of working poor and unemployed citizens.

Atlanta, contained primarily in Fulton County, is home to 426,300 residents—56% of the county population in just under 25% of the total land area. Atlanta's resident population is approximately 67% African-American, which is 77% of the county's total African-American population. The downtown area is home to major business, industry,

and finance concerns, as well as host to a large convention and tourism industry. There is significant commuter traffic into the city during business hours. In the city of Atlanta, the Atlanta Police Department's (APD's) jurisdiction is divided into six zones and 56 beats.

Indicators, Datasets/Data Sources, Collection Strategy

To characterize the nature of firearm mortality and morbidity in the project area, and to determine whether or not firearm-related events cluster in identifiable high-frequency, or hot spot, places and times, the authors analyzed four key datasets.

Overall homicide and firearm homicide rates for Fulton County were obtained from the National Center for Health Statistics for 1968 to 1995. Yearly rates were broken down by gender and race.

Death records for all firearm-related deaths 1989 to 1997 were obtained from the Fulton County Medical Examiner (FCME). The FCME investigates and records all deaths occurring within the boundaries of the county. The inclusion criterion for this dataset was all individuals who died in an incident involving a firearm in Fulton County, either on the scene or from injuries resulting from a shooting. Dates of inclusion are 1/1/89 through 12/31/97. All data on race, age, sex, and resident/non-resident status were included. Each case record includes medical examiner case number, name, age, race, sex, date of birth, home address, report date and time, incident date and time, and location of incident. The data were obtained as a dBase IV file download from the FCME.

Emergency 911 computer-aided dispatch data (CAD) for a subset of firearm-related call types were obtained from Atlanta's E-911 system. This system covers only the city of Atlanta; it does not cover the remainder of Fulton County. The Atlanta 911 dataset included all calls for call types 25 (shots fired); 50, 504, 5025 (person shot); and 69 (person armed) for the time period 1/1/97 through 12/31/97. Each record includes a unique identifier number, call type, incident location, the time and date of the call, priority, zone, beat, dispatch time, arrival time, call completion time, a brief description of the event, and related police numbers. The file was received as an ASCII download from the 911 Center and translated to a dBase IV file for analysis.

Finally, data on non-fatal firearm injuries were exported from a firearm injury surveillance system developed and maintained by the authors. The surveillance system tracks shooting incidents in the five-county area of metropolitan Atlanta (Fulton, DeKalb, Cobb, Clayton, and Gwinnett counties) and links police reports of shooting incidents, emergency department records, and medical examiner records to produce a complete picture of each firearm-related injury and death in the project area.

Data Analysis

The data were imported into Paradox 7.0 for Windows for table restructuring; the designation of variables as a character or numeric was necessary for ArcView 3.0a (ESRI, Redlands, CA) analysis. The data were also imported into Microsoft Excel 97 for separation of the incident location field and cleaning to assure a high rate of successful geocoding.

Descriptive Epidemiology

The data were imported into SPSS 8.0 (SPSS, Inc., Chicago, IL) for descriptive

epidemiological analysis. Frequencies were calculated for age, race, sex of victim, and the type of incident (homicide, suicide, accident, other). Cross tabs were calculated for race and sex of victim, and histograms of the time of incident were calculated. Results were displayed graphically using Microsoft Excel 97.

Geographic and Temporal Analysis

The data were imported into ArcView 3.0a and ArcView Spatial Analyst Extension for geographic analysis. Data were geocoded based on the street address or intersection of the incident location. Geocoding match rates varied by dataset, due to the quality of the address information. City of Atlanta 911 CAD data had a geocoding match rate of 89%. The FCME data had a match rate of 92%, and the firearm injury surveillance dataset had a match rate of 64%. The geocoding match rate for this dataset is much lower due to poor address data on the police reports.

Three types of maps were produced: point maps of incident location, broken down by age, sex, crash type, and time of day; areal maps, in which data were aggregated to police zone, beat, and census tracts using a spatial join; and isoarithmic maps, obtained using the calculated density function in ArcView Spatial Analyst Extension. Density calculations were created to determine the historical center of mass of firearm homicide and assault.

These data were collected and analyzed for immediate police utility and application by a multi-agency task force that was interested in the number of incidents in particular areas. Therefore, this paper discusses frequencies only. Rate calculations will be completed in future work.

Results

Firearm Mortality

Fulton County Twenty-Five-Year Homicide Perspective Analysis of National Center for Health Statistics homicide statistics for Fulton County showed a decline in overall homicide and firearm homicide over the past 10 years. Overall homicide rates averaged 31 per year per 100,000, ranging from a low in 1983 of 22 per 100,000 to a high in 1973 of 45 per 100,000. The firearm homicide rate mirrored the overall homicide rate, while the non-firearm-related homicide remained stable, averaging 10 per year.

For ages 25 and over, overall homicide declined moderately from a peak in 1973. This age group averaged 38 homicides per year per 100,000, ranging from a low of 25 in 1984 to a high of 61 per year per 100,000 population in 1973. The firearm homicide rate mirrored the overall rate. Non-firearm homicide remained stable, averaging 13 per year.

The rates for age groups 15–19 and 20–24 were strikingly different from the older age groups. Among 20- to 24-year-olds, non-firearm homicide showed minor fluctuation but remained stable. Firearm homicide accounted for nearly all of the variation in overall rates. This age group averaged 51 per year per 100,000, ranging from a low of 26 in 1983 to a high of 71 in 1972; the rate for this group has increased since 1983.

Among 15- to 19-year-olds, the average was lower—33 per year per 100,000, ranging from a low of 16 in 1983 to a high of 70 in 1994. The pattern was similar to that of the 20- to 24-year-olds, but even more pronounced. The non-firearm homicide rates

remained stable, while firearm homicide accounted entirely for the overall increase in homicide. The rate increased dramatically beginning in 1986. The highest years, 1991 and 1994, were each followed by a precipitous drop.

For ages 0–14, the rates were comparatively low, less than 10 per year per 100,000 with wide fluctuation.

Firearm-Related Deaths, Fulton County, 1989–1997

Analysis of FCME records for 1989–1997 revealed 1,994 deaths involving a firearm. Of these, 74% (1,480) were homicides, 24% (482) were suicides, 1% (19) were ruled "accidental," and 0.6% (12) involved undetermined circumstances. (The one remaining record lacked data for this variable.) The number of firearm homicides per year has remained steady, averaging 164 per year, ranging from a low of 132 in 1997 to a high of 182 in 1989 and again in 1993. Of the 1,480 victims of firearm homicides, 88% (1,298) were male and 12% (180) were female (2 additional records lacked data on the victim's gender); 85% of the victims were African-American, 12% were white. Asians and Hispanics combined accounted for 3% of the victims. Persons aged 15–24 accounted for 36% of the victims. Center of mass calculations (calculated density) place the locus of firearm homicide (all ages) in a concentrated low-income residential neighborhood (mixed residential, abandoned business use) immediately southwest of the downtown area.

Point maps of incidence by age group were created to analyze patterns of youth and young adult (0–24 years) versus adult (24+ years) homicide. The incidents show evidence of clustering in a small number of police beats, around particular public housing complexes, and along major commercial roadways.

The data were aggregated to police zone and beat; 77% of incidents matched to a zone after spatial join. Of these, 25% of firearm homicides occurred in zone 3, and 24% in zone 1. By beat, 8 beats had over 40 firearm homicides in the nine-year study period. The high-frequency beats range in size from one of the smallest (APD beat 112) to one of the largest (APD beat 405) in area.

Finally, the data were aggregated to census tracts for Fulton County. High-frequency tracts were clustered in the city proper, with the exception of the two census tracts in the southern end of the county. Nine census tracts had over 26 homicides during the nine-year study period. These high-frequency tracts were compared with the state, county, and city average on the following indicators: percent under 18 years, percent female-headed household, percent non-family household, percent high school graduate or higher (aged 25+), percent unemployed, percent male unemployed, percent below poverty level. The highest-frequency census tracts averaged 24%, 20%, and 15% higher than the state, county, and city percentage, respectively, of female-headed households. The identified tracts averaged 25%, 22%, and 13% higher than the state, county, and city percentage, respectively, for percentage of persons living below poverty level. Certain hot spot census tracts were strikingly higher on these indicators.

Firearm Morbidity

In 1997, population-based analysis of gunshot reports from area emergency departments and local law enforcement agencies identified 226 firearm homicides and 774 non-fatal firearm assaults in the five-county metro area. Sixty-five percent (65%) of the homicide victims were 34 years old or younger; 35% were between 15 and 24 years old.

Seventy-two percent (72%) of the 774 victims of firearm assault were 34 years old or younger. Forty-four percent (44%) were 24 years old or younger. In 44% of homicide and assault cases, the age, race, and sex of the suspect was recorded. In 58% of these shootings the suspect was noted to be 24 years of age or younger; 95% of the suspects were male, and 95% were African-American. There were 3.42 cases of non-fatal firearm assault for every case of firearm homicide.

Visual inspection of the point maps of the firearm assault data indicate that the events are more diffusely spread over the city and county, although this may reflect the fact that this is only one year of data; a tighter picture emerged from the nine-year homicide dataset. Clustering was identified, perhaps not surprisingly, in the hot spot areas of firearm homicide. Calculated density maps were created to determine the locus of firearm assault in the county and city. The areas of highest frequency are clustered within the city boundaries, with minimal activity in the surrounding suburban areas to the north and south. Within the city of Atlanta, the locus of firearm assault is immediately southwest of downtown, with other high-frequency areas identified surrounding several public housing units.

Firearm-Related Emergency Calls to 911

City of Atlanta 911 computer-aided dispatch data for 1997 for firearm-related call types were obtained. Call types included were 25 (shots fired); 50, 504, 5025 (person shot, person shot/ambulance sent); and 69 (person armed). In 1997 there were 10,725 firearm-related calls, an average of 894 firearm-related calls per month. Of these, 79% were "shots fired," 17% were "person shot," and 4% were "person armed." Calls from APD zones 1 and 3 together accounted for 53% of the analyzed calls.

Data were aggregated to police beat to create areal maps, which highlighted geographic concentration within zones 1 and 3. Although the beats varied considerably in size and some variation was expected, the areal maps indicated that particular beats have a much higher frequency of firearm activity than do others of similar size. There were 15 beats with under 100 calls, 20 beats with 100–200 calls, 10 beats with 200–300 calls, 6 beats with 300–400 calls, 4 beats with 400–500 calls, and one beat with over 600 firearm-related calls in 1997. This beat was also one of the highest-frequency beats for firearm assault and homicide.

A histogram of the time the calls were received in the 911 Center indicates that 32% of the calls were received between 8:00 PM and 12:00 midnight. The high-frequency time period varies between police beats and between clusters of activity. Point maps of incidence by police shift make this clear; in certain identifiable areas, the incidents occur primarily between 11:00 PM and 7:00 AM. In other identifiable areas, incidents occur primarily between 3:00 PM and 11:00 PM.

By overlaying point maps from the three datasets—firearm homicide, assault, and 911 calls—clusters of activity were easily identifiable in specific neighborhoods and streets. Using as an example APD zone 3, a cluster was identified at a public housing unit and its surrounding neighborhood in Beat 309. This beat is consistently the highest-frequency beat for firearm crime and victimization.

Discussion

GIS analysis can help public health and law enforcement agencies identify the best time

and place to concentrate their resources, as well as measure the effectiveness of interventions to reduce crime and victimization, death, and injury in our communities. In the project area, the crime and public health problem of firearm morbidity and mortality was not evenly distributed, and GIS analysis of firearm homicide, assault, and emergency calls to 911 demonstrated this clearly. The identified high-frequency areas for the events were concentrated within two police zones and eight beats within the city. Within the beats, there was further concentration on specific streets and surrounding certain public housing units. These areas scored high on poverty indices, including percent female-headed household, unemployment, and persons living below the poverty level. Further, these areas are identified high-frequency areas for drug abuse and drug market activities.

In 1997, several local agencies joined forces in a collaborative effort to reduce overall gun violence in the city, with a special focus on juvenile gun violence. The Atlanta Police Department, the Atlanta Office of the Bureau of Alcohol, Tobacco and Firearms (ATF), the Georgia State Board of Pardons and Paroles, the Fulton County Juvenile Court, the Fulton County District Attorney, Fulton County Probation, and the Emory Center for Injury Control, as well as others, participate in this effort.

The APD Guns and Violent Crime Suppression Unit and partners are carrying out targeted law enforcement activities designed to reduce the flow of illegal weapons in the city of Atlanta (particularly those to juveniles and young adults) and reduce overall criminal firearm activity and victimization. The unit has targeted enforcement initiatives in identified hot spot areas and high-incidence time periods using the GIS data analysis provided by the authors. The unit also participates in the ATF Youth Crime Gun Interdiction Initiative, cooperative investigations with the APD Gang Task Force, pawn shop investigations, and probation enforcement. The unit works cooperatively with the ATF, the Fulton County District Attorney, and the United States Attorney's Office to develop cases appropriate for state or federal prosecution.

Future Plans

These data will be used to evaluate the impact of the multi-agency intervention to reduce firearm violence described above. In future work, the authors plan to complete analyses to determine if the geographic distribution of events is regular, random, or clustered. Visual inspection indicated clustering, but further analyses are needed to confirm or refute this finding. Finally, the authors will separate the data by year for time series analysis and views, to determine if the areas of high frequency have remained stable or shifted over time.

Acknowledgments

This work was supported by grants #NIJ-94-MU-CX-K003 (NIJ/OJJDP/CDC); #NIJ 95-IJ-CX-0025 (NIJ); #R49/CCR407419-02-3 (NCIPC/CDC). We gratefully acknowledge the support of Lois Mock (NIJ); Tomi Sampson (ECIC); Chief Harvard, Deputy Chief Gordon, Captain Arcangeli, John Carawan, Amaza Bodie, and Roblyn Rossen (APD); Dr. Zaki, Dr. Hanzlick, and Chief Investigator McGowan (FCME).

References

1. Rosenberg HM, Ventura SJ, Maurer JD, Heuser RL, Freedman MA. 1996. *Births and deaths: United States, 1995*. Monthly vital statistics report. Hyattsville, MD: National Center for Health Statistics. 45(3) supp 2:31.
2. Rich TF. 1995. *The use of computerized mapping in crime control and prevention programs.* NIJ Research in Action Series. Washington, DC: National Institute of Justice. July.
3. Taylor RB, Gottfredson S. 1986. Environmental design, crime and prevention: An examination of community dynamics. In: *Communities and crime*. Ed. AJ Reiss, M Tonry. Chicago: University of Chicago Press.
4. Sherman LW, Weisburd D. 1995. General deterrent effects of police patrol in crime "hot spots": A randomized controlled trial. *Justice Quarterly* 12:625–48.
5. Weisburd D, Green L. 1995. Policing drug hot spots: The Jersey City drug market analysis experiment. *Justice Quarterly* 12:711–35.
6. Atlanta Regional Commission. 1997. *Area population projections for metropolitan Atlanta*. Atlanta, GA: Atlanta Regional Commission.

A New GIS-Based Tool for the Assessment of Environmental Equity and Death Rates Near Superfund Sites in the Urban Counties of Washington State

Richard Hoskins PhD, MPH*
Director, GIS and Spatial Epidemiology Unit, Office of Epidemiology, Washington State Department of Health, Olympia, WA

Abstract

All the geocoded Superfund (SF) and Toxics Release Inventory (TRI) sites from the Environmental Protection Agency's Landview III database in three urban counties in Washington State were developed into a geographic information system (GIS) coverage with circular buffers. Using a census block group base coverage, a spatial overlay was used to estimate population, various socioeconomic (SES) variables, and death rates from several causes. Age-stratified, age-adjusted death rates and standard mortality ratios were calculated and adjusted using empirical Bayesian smoothing with a prior distribution developed from the whole state and one from each block group's nearest neighbors. This was done to stabilize rates where the rate sample variance was high. Using the results of the buffer overlay, a profile was developed for all sites together. To facilitate comparison with other areas, a control group coverage was developed by building similar buffers around 25,000 random points inside the study counties. Points under water and in other areas not likely to have SF/TRI sites were excluded. Similar to a Monte Carlo simulation, control points were sampled and an empirical distribution developed for each variable for statistical testing. In the buffered regions, low income, status as a minority, limited education, high population density, and a high proportion of people over 65 were associated with SF/TRI sites. For causes of death, the death rate from cancer around SF/TRI sites was marginally statistically significant. After applying Bayesian smoothing to stabilize the rates, the differences became even less.

Keywords: Bayesian smoothing, Monte Carlo, environmental equity, Superfund, disease rates

Introduction

This paper describes a methodology for assessing the characteristics of neighborhoods located around areas that contain toxic waste or facilities that emit toxic waste. We are interested in determining if the socioeconomic characteristics of people living near these sites are different from in other neighborhoods, and whether they experience different rates of disease or death. Whether rates of disease have any causal relationship to these sites, as always, remains elusive.

In Washington State, public health practitioners need ways to determine the characteristics of populations around these sites that take geography into account. It is

* Richard Hoskins, GIS and Spatial Epidemiology Unit, Office of Epidemiology, Washington State Department of Health, 1102 Quince St., Olympia, WA 98504-7812 USA (p) 360-236-4270; (f) 360-236-4245; E-mail: reh0303@doh.wa.gov

important to consider the spatial context of neighborhoods with respect to characteristics such as low income or other indicators of low socioeconomic status that may predispose them to living close to toxic sites. In addition, it is important in the assessment of outcome measures such as disease or death rates to be able to deal with small area problems that plague disease rate calculations, such as an apparently high number of cases for a low population.

Our methods could be viewed as leaning toward violating the principle in epidemiology that cautions against using the "ecologic fallacy," whereby we assign characteristics to an individual based on the group to which they belong. We believe, however, that useful assessment methods must also avoid the "atomistic fallacy," which fails to consider the social, economic, and geographic context of individuals in a public health assessment.

This paper presents a means of assessment based on a Monte Carlo method of developing a statistical distribution that can be used for statistical testing. To determine disease rates, we adjusted rates in a neighborhood using an empirical Bayesian method to help account for small area problems. This work is in its initial stages and, therefore, is not presented as a completed effort.

Methods

Monte Carlo Method

For Snohomish, King, and Pierce counties in western Washington State, we developed a geographic information system (GIS) coverage of the 257 Superfund (SF) and Toxics Release Inventory (TRI) sites from the US Environmental Protection Agency (EPA) Landview III database (1). Information in this database allows for placing the sites at street-level accuracy. For this pilot study, we did not distinguish between the type of toxic contained or being emitted at a site, or a site's status as a SF or TRI site. Around each area, we constructed concentric rings of circular buffers of 0.25-, 0.5-, 1-, and 2-kilometer radii. We then developed a GIS coverage of US Census block groups with population attribute data for age, sex, and race, and with economic data from the 1990 census and projections for intercensus years from data provided by GeoLytics (2) and the Claritas Corporation (3). Using the spatial overlay operation in the GIS software, Maptitude (Calipter Corporation, Newton, MA), we were able to estimate the population and other characteristics from the block groups (or partial block groups) in each concentric buffer (4). Taking all the sites in one area, we were able to determine a profile of the residents within each of the buffer radii for population, income, education, and other census variables for the SF/TRI sites.

A control set was developed against which to compare this profile. In the three counties, we randomly assigned 25,000 points in the study area. The same size buffers were used and the same spatial overlays performed for each point as was done for the toxic sites. Excluded were points that fell in water areas (e.g., Puget Sound, Lake Washington) or in other locations where SF/TRI sites were not likely to be found. The distributions of these variables were developed by a Monte Carlo simulation. Using a uniform random number generator, repeated sets of control sites of 257 points each were drawn from the control group and a determination made of the value of the variable, such as population or income, using the results from the spatial overlays around

each site. The result was a count histogram over the variable's range that indicated the likelihood that any particular value of a variable was found in the control set. A comparison of the toxic site profile with the control set was done by finding the value of the variable on the x-axis and computing the area from that point to the end of the curve. This results in a p-value for statistical comparison. An example showing the percentage of the population that is minority is presented in Figure 1, over which a Gaussian curve has been fit.

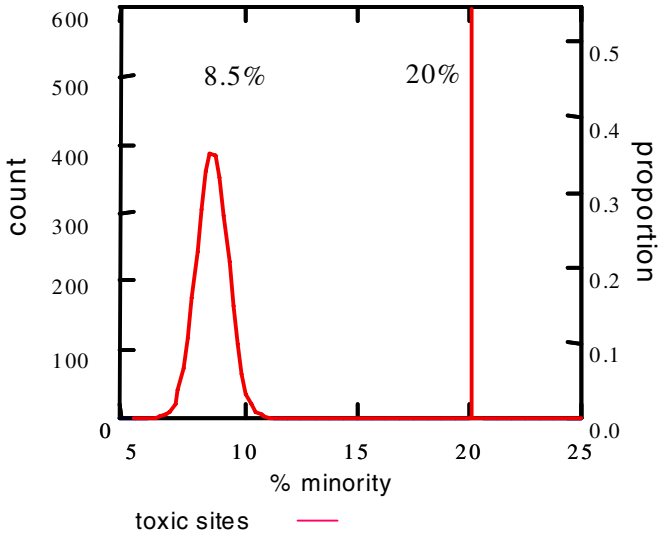


Figure 1 Percent minority race in a 1-kilometer buffer.

Empirical Bayesian Estimation of Death Rates

To determine the death rates from a variety of causes near the SF/TRI sites, and to compare them with the rates in the control areas, a spatial overlay procedure similar to the one described above was used. Death certificate data were geocoded to the street level and a point-in-polygon procedure was used to determine the number of deaths in each buffer area for a specific cause of death, age, race, and sex. This was done around each toxic area and control site. Using the buffer populations estimated by spatial overlay for the denominator, and point-in-polygon determinations for the numerator, the age-specific and age-adjusted rates were calculated for the various buffer sizes. These rates are likely to be unstable in areas where death counts are high in comparison with the underlying population. An adjustment can be made using an empirical Bayesian procedure; see Bailey and Gatrell (5) and Devine (6,7) for a good description of this method.

The adjustment of the age-specific and age-adjusted death rates is carried out according to the weighting scheme

$$\hat{r}_i^{Bayes} = \hat{w}_i r_i + (1 - \hat{w}_i)\hat{\gamma}$$

where

$$r_i = \text{observed rate for area } i$$

$$\hat{w}_i = \frac{\hat{\phi}}{(\hat{\phi} + \hat{\gamma} / n_i)} \quad \text{which is a shrinkage factor}$$

$\hat{\gamma}$ = mean death rate for the region (or nearest neighbors)

n_i = population in area i

and $\hat{\phi}$ is the weighted sample variance of the observed rates

$$\hat{\phi} = \frac{\sum n_i (r_i - \hat{\gamma})^2}{\sum n_i} - \hat{\gamma} / \bar{n}$$

We adjusted the death rates in the various buffers according to this scheme, using the mean of the prior distribution for the whole state. In addition, we calculated the weighted sample variance using only the nearest neighbor block groups. Their contribution to the variance was further weighted depending on the intercentroid distance between the nearest neighbors. This yields a spatial or local Bayesian rate estimate, which was used to calculate the standard mortality ratio (SMR). Using this Bayesian rate or SMR rather than the observed rates around the toxic waste sites and control points addresses the high variability of the rate estimates and accounts for at least some of the spatial correlation.

Results

A comparison of several census variables for the toxic sites and several for the control group is shown in Table 1.

The differences between the values for all of the toxic sites compared with the controls (Table 1) were statistically significant at the 5% level or below, except for the variable percent college graduates living in the buffered areas around the toxic sites. Buffers at the other radii had similar results. Toxic sites were more like to have minority populations living in higher population densities. The sites are also more likely to have populations with less education, lower incomes, a lower percentage of children under 5 years of age, and a higher percentage of adults over 65 years of age.

The spatial or local Bayesian smoothing of rates for a variety of diseases was determined at the block group level throughout the three county area. Figure 2 shows the empirical distribution for SMRs for all cancer deaths between 1990 and 1996.

The SMR for all the control points was 112, which indicates a slightly higher cancer rate compared with the rest of the state. The SMR with no smoothing (i.e., the observed rate) for the toxic sites was 140, corresponding to a p-value of .12. The Bayesian and spatial Bayesian rates were 118 and 127, respectively, with p-values of .41 and .22. In this case, the Bayesian adjustments that took into account the rates of the entire state or

Table 1 Selected Census Variables for a One-Kilometer Buffer Comparing SF/TRI Sites with a Three-County Empirical Distribution for Each Variable

Indicator	Toxic Site Mean	Control Site Mean
Population	8,028.5	4,617.5
Population density	2,573.2	1,480.0
Families	1,765.3	1,172.9
Households	3,583.2	1,847.4
White	6,180.5	3,896.4
Black	714.4	263.3
Asian	872.5	347.3
Indian	131.2	55.8
Hispanic	295.2	138.8
% white	78.2	85.4
% Asian	11.0	7.6
% black	9.0	5.8
% Indian	1.7	1.2
% minority	20.0	8.5
Households median income	$29,155	$40,429
Average per capita income	$16,198	$17,160
Households with earnings	2,876.4	1,550.7
Households without earnings	706.7	296.8
% households without earnings	19.7	16.1
% households with income <$15K	31.2	22.5
% households with income >$100K	3.4	4.5
Housing units	3,829.4	1,941.3
% housing units owner occupied	42.8	57.6
% without high school diploma	16.2	13.6
% college grada	29.6	27.9
% children in poverty	17.8	10.3
% children age 5 or under	7.7	8.8
% age 65 or older	12.7	10.7
% households w/o earnings	19.7	16.0
% persons in poverty	13.9	9.0
% age 5 or younger in poverty	20.8	13.5
% age 65 older in poverty	10.5	7.8

[a] Difference between values for this variable not statistically significant.

nearest neighbors, as well as the estimated sample variance of the state or nearest neighbor weights, shrunk the observed value toward the center of the distribution.

A comparison of observed versus spatial Bayesian SMR is illustrated in Figure 3. Comparing the two maps, the map of observed SMRs shows how the SMR is lowered; darker blue shades in the thematic map indicate an SMR below 100, while those above 100 tend toward darker shades of red.

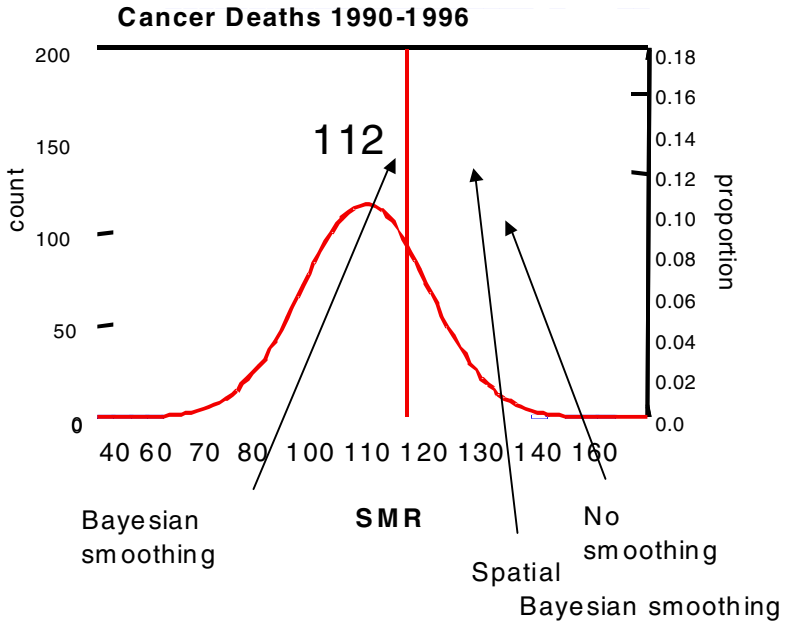


Figure 2 Standard mortality ratio (SMR) for observed, Bayesian and spatial Bayesian smoothing.

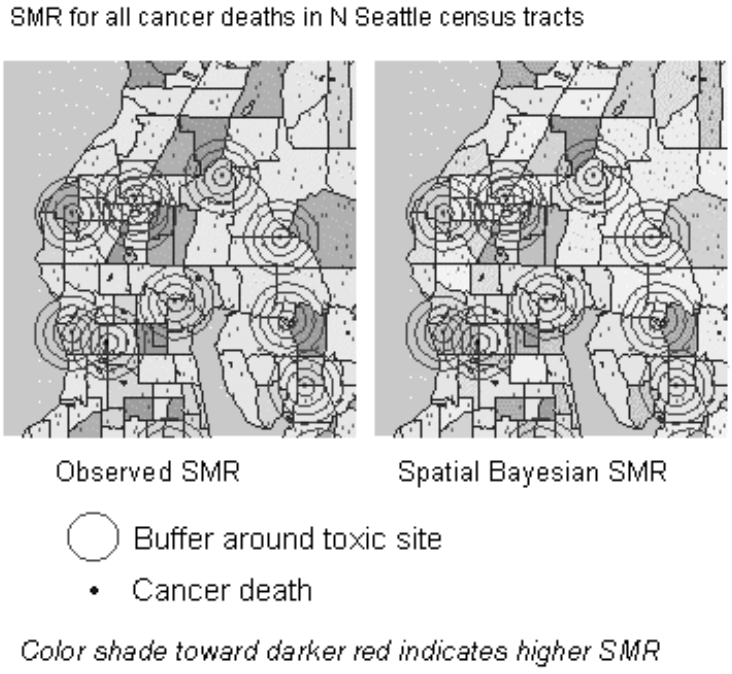


Figure 3 Thematic maps of cancer SMR before and after Bayesian smoothing.

Discussion

We have presented some initial results about a methodology for determining the characteristics of populations around SF/TRI sites. The Monte Carlo simulation builds an empirical distribution that one can use to compare a profile of one or more toxic sites with the rest of the county or region. One advantage of this method is that it requires no assumption about the shape of the distribution. It may also account for some spatial autocorrelation because the random selection of the control points is not dependent on the location of other points selected in the set. The method is simple to use and may give better results than ordinary parametric methods.

In the results presented here, it is clear that populations around the toxic sites identified from the Landview III database were more likely to be minority and lower income. In this preliminary work, however, no distinction was made between sites with respect to harmfulness or to content of known carcinogenic compounds or other putative substances known to affect health. Subsequent studies will look at neighborhoods around sites to consider what the site contains and, perhaps most importantly, its remediation status.

A similar scheme was used to assess disease rates (via death rates) around the toxic sites. There appeared to be some elevation of the death rate for all cancer deaths, but after Bayesian smoothing the differences were not important. No adjustment was made to account for the increase in death rates that is often associated with minority or lower-income populations. This might be done by calculating the SMR with race and sex stratification, as well as with the usual 5- or 10-year age groups.

As with all Monte Carlo based schemes, the quality of the results depends on the sampling protocol for the control points. Future studies need to consider adjusting the probability of selecting a particular control point based on the historical land use designation. Zoning laws that were in effect many years ago certainly influenced whether or not a potential SF/TRI site could ever be built at a particular location. In the current model, all points are equally likely to be sampled, presuming they are not under water or located in places such as cemeteries and older parks.

References

1. US Department of Commerce, US Environmental Protection Agency. 1997. *Landview III environmental mapping software.* www.census.gov/apsd/pp98/pp.html.
2. GeoLytics, Inc. 1998. *Census CD + maps, US 1990 Census (STF3A, C, and D).* East Brunswick, NJ: GeoLytics, Inc. www.Geolytics.com.
3. Claritas Corporation. 1990–1997. *Annual census data.* Arlington, VA: Claritas Corporation.
4. Caliper Corporation. 1998. *Maptitude, v 4.02.* Newton, MA: Caliper Corporation. www.caliper.com.
5. Bailey TC, Gatrell AC. 1995. Empirical Bayes estimation. In: *Interactive spatial data analysis.* Essex, England: Longman Scientific & Technical, Longman Group, Ltd. 303–98.
6. Devine OJ, Louis TA. 1994. A constrained empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine* 13(11):1119–33.
7. Devine OJ, Louis TA, Halloran ME. 1994. Empirical Bayes methods for stabilizing incidence rates before mapping. *Epidemiology* 5(6):622–30.

The Role of Geographic Information Systems in Population Health

Russell S Kirby, PhD, MS, FACE (1),* Seth L Foldy, MD (2)
(1) Department of Obstetrics and Gynecology, Milwaukee Clinical Campus, University of Wisconsin-Madison Medical School, Milwaukee, WI; (2) Commissioner, Milwaukee Department of Health and Department of Family and Community Medicine, Medical College of Wisconsin, Milwaukee, WI

Abstract

This paper explores applications of geographic information systems (GIS) in population health and the preconditions for its optimal use. Population health involves the assessment, evaluation, and optimization/improvement of health status and outcomes on a population basis. It is the ultimate pursuit of public health programs, which are more often focused reactively than actively on underserved groups or those with diseases or health needs not adequately treated by the health care delivery system. As public health broadens its focus toward the determinants of population health, GIS can perform several functions in population health informatics. GIS has hardware, software, and staffing requirements; in population health, a more important precondition for their use is a systematic, integrated approach to geocoding all population-based health data systems. With routinely geocoded databases, GIS can fulfill many roles in population health informatics. Functions include an interactive environment for the spatial display of health data; a laboratory for the development and dissemination of neighborhood/community health indicators; a tool for integrating disparate data records by location; a vehicle for displaying results of analyses from databases merged by automated record linkage; a platform for testing hypotheses concerning the epidemiologic determinants of health status, diseases/outcomes, or associations between determinants of population health and utilization of health services; and a vehicle for facilitation of public health program planning, evaluation, and community-based decision-making. As the implementation of health-oriented applications of GIS evolves, with appropriate attention to geographic, epidemiologic, and biostatistical methodological concerns and methods of map presentation, opportunities for extending GIS into population health informatics are almost limitless.

Keywords: population health, informatics, geocoding, record linkage

Introduction

While GIS has a significant role in traditional public health activities, their ability to collate, integrate, and display population-based data concerning health events, exposures, risk factors, and socio-environmental data warrants a broader, more holistic perspective on the health of populations. This brief essay explores the opportunities for population health research and practice and the central role for GIS within the emerging paradigm of population health. We have four objectives in this presentation. First, we will define population health as a distinct field of intellectual inquiry, and compare and

* Russell S Kirby, PhD, Department of Obstetrics and Gynecology, University of Wisconsin Medical School, PO Box 342, Milwaukee, WI 53201-0342 USA; (p) 414-219-5610; (f) 414-219-5201; E-mail: r-kirby@whin.net

contrast population health with the current practice of public health in most states, cities, and communities in the United States. Second, we will provide a framework for population health informatics as an operational environment for the practice of population health. Third, we will explore the potential roles and opportunities for GIS in population health. Finally, we will identify methodological issues, opportunities for multidisciplinary interaction and collaboration, and applied research in GIS-based population health.

Population Health

"Population health" refers to the health, well-being, and functioning of entire populations. It shares with public health an explicit focus on whole populations. However, the scope of population in population health may be defined flexibly—to include, for example, covered lives in a managed-care plan or a corporate workforce, rather than people within a geographic government jurisdiction. In addition, population health examines a broader set of inputs and health outcomes than are traditionally studied in public health.

What are the determinants of population health? Evans and Stoddart provide a conceptual framework for population health (1), reproduced with modifications in Figure 1. Factors like social environment and prosperity appear both as inputs and outputs, affecting each other in a reciprocal series of relationships. These extend beyond traditional host-agent-environment concerns of public health (while still incorporating genetics, behavior, and the physical environment). The model for population health also includes, but radically extends, the medical preoccupation with the relationship between disease and health care. While the traditional medical model focuses on the determinants and treatment of diseases through the provision of health care services, population health focuses on goals including general well-being and functioning, not just disease. Most importantly, outcomes of interest from the perspective of population health form a superset of the traditional public health outcomes. Societal levels of health and functioning, or general well-being, are the outcomes of greatest interest. Population health thus calls attention to the relationships between culture, polity, economy, environment, and health care utilization and quality. The model of the determinants of population health places the medical model of health care in broad societal perspective, and provides a general prevention focus for practitioners of population health.

For example, outcomes typically measured in health care delivery include appointment wait times, rehospitalization after emergency room visits, or five-year survival after cancer treatment. Typical public health outcomes include infant mortality, immunization rates, or the incidence of lead poisoning, selected from literally hundreds of measures listed in *Healthy People 2000: National Health Promotion and Disease Prevention Objectives* (2). Population health extends this view to broader health status measures like premature mortality rates (3), change in quality-of-life years (4), and the rate of limitation in activities of daily living. These are clearly influenced by, but not entirely defined by, traditional medical and public health measures.

Population health derives from a variety of intellectual traditions. These include public health (especially population-based data systems), demography, social and behavioral sciences (especially representative sample surveys of health, health status, and

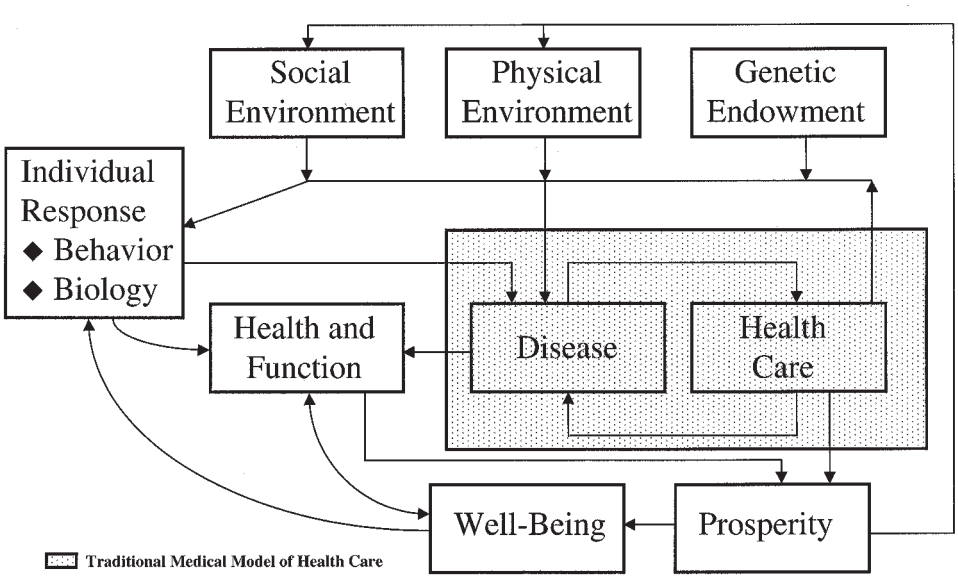


Figure 1 Conceptual framework for population health. Reprinted, with modifications, with permission from the authors, in Evans RG, Barer ML, Marmor TR, Why Are Some People Healthy and Others Not? The Determinants of Health of Populations. New York: Aldine de Gruyter, 1994. Copyright ©1994 Walter de Gruyter, Inc., New York.

well-being), environmental health, epidemiology, and health services research. This emerging discipline gains strength from the diversity of perspectives brought to bear on the issues at hand, and from the necessity for integration of staff and resources from the public, private, and academic sectors. (See Young's *Population Health: Concepts and Methods* [5] for a more thorough introduction to this field and its practice.)

Population Health Informatics

"Informatics" is defined in the American Heritage Dictionary as "information science." Recently, Friede et al. (6) described public health informatics as an emerging field that encompasses public health surveillance methodologies, data and databases, and information systems, used collectively to merge, manage, analyze, and interpret public health data. The practice of public health informatics implies a paradigm shift in the institutional arrangement, management, operationalization, and utilization of databases and information services within the public health sector. In most states and municipalities, integration of public health information services across the spectrum of programmatic activities has yet to be realized (7). The core public health functions of assessment, policy development, and assurance (8) would all be enhanced through the expansion and integration of the current health statistics/epidemiology units at the state, municipal, and local levels into seamless public health system-wide informatics environments.

Were we to accomplish this task, however, we would still fall short of the optimal

information basis for population health. We propose a new field of intellectual endeavor, population health informatics, that builds upon public health informatics but includes the following additional features:

1. *Information on entire populations (not just service users).* Information on entire populations includes public health surveillance data (vital records, reportable diseases) and other government data. It can also be gathered by assembling data from the universe of overlapping organizations that serve an entire population (for example, immunization registries or cancer registries using data assembled from multiple health care providers).
2. *Integration of databases linking public health information, environmental information, health services information, and socioeconomic information to health outcomes.* Because both the social and physical environment (factors like income, education, housing quality, and air quality) and health services (accessibility, utilization, effectiveness, efficacy) have considerable impact on health, data reflecting these conditions should be linked to health outcome data at a reasonably discrete level.
3. *Focus on broader health outcome measures including functional status, disability, and quality of life, assessed across populations.* An added goal is to associate the above information with population-level indicators of health that go beyond the traditional (mortality, prevalence) to include meaningful indicators of quality outcomes for large portions of the population (well-being, function, quality of life).
4. *The ability to define sub-populations flexibly (not only across administrative geographic units, but also by other characteristics).* Because services and policies typically affect sub-populations (e.g., members of a neighborhood or a managed-care group), a population health information system should allow the creation and analysis of sub-population information sets. Health planning is facilitated through such networks. Population health informatics can support a broader perspective on the determinants of population health; rarely does the zip code or municipality of residence correlate directly at the individual unit of analysis with likelihood of exposure to environmental hazards or utilization of health services.

Population health informatics establishes an information environment for assessing and monitoring the health, functioning, and well-being of entire populations, consistent with the Evans and Stoddart model in Figure 1. The integration of such population-based information from the many, unconnected systems that already exist would profoundly improve health planning, public health surveillance, and health services research. Integrating this information would also provide a powerful platform from which to study the impact of health and social policies on population health. Like informatics in general, population health informatics represents a fundamental transformation of the manner in which population-based health data are collected, managed, and used to support the core functions of public health—assessment, assurance, and policy development—as applied to entire populations. Although there are many barriers to the realization of such an information environment, it is not too soon to consider the technical preconditions for such a vision. In the following section, we explain why geographic information systems (GIS) are central to the conceptualization and practice

of population health informatics. The balance of this paper describes some of the potential roles for GIS in population health, as well as related issues.

Roles for GIS in Population Health Informatics

One of the central requirements for population health informatics is the integration of information systems that contain the broad range of data of interest. The integrative nature of population health requires data from systems containing health information, socioeconomic information, environmental information, and subjective or solicited information on outcomes like well-being or functioning. Existing sources include administrative databases (hospital billing records, tax files), public health databases (reportable diseases, vital statistics), programmatic databases (lead poisoning surveillance, immunization or cancer registries), census information, governmental housing and environmental databases, and representative sample surveys.

Database integration requires linkages among the various data at a discrete enough level to allow meaningful inferences about relationships, trends, co-factors, and confounding variables. The most discrete and useful form of linkage is through personal identity, represented by name or unique alphanumeric identifier. However, confidentiality and privacy concerns are very real and reasonable (9); thus, using personal identity may not be either ethically or politically feasible for many sources of data using current technology. A common but unsatisfying way to link health information is by broad categories like race (e.g., showing trends of low birth weight over time by race). Such broad linkages provide little insight into the relationships among the many variables that directly affect health outcomes, focusing instead on surrogate variables like skin color that may have little direct relationship to the outcomes of interest (10). Place, however, offers many advantages as a means of linking and then analyzing disparate data sources. Information on place is almost universally collected in health care documents (in the form of address), though it is not always entered into databases. It is often associated with a broad range of both socioeconomic and environmental factors. Using location (which can be manipulated or aggregated in various fashions) may also offer a lesser threat to personal privacy. It can serve as a definitive linkage point between two address-bearing databases, and as a categorical or a continuous two-dimensional variable along which imputation of data is possible (with appropriate care on the part of researchers and end-users).

We propose four levels of database integration in population health informatics. GIS would play important roles at each level.

- *Level 1: Surveillance of indicators.* This could include traditional surveillance, such as the incidence of communicable diseases over time. With the expanded data linkages of population health informatics, it might also include ongoing monitoring of indicators like the adequacy of prenatal care, ambulatory care sensitive hospitalizations, or arrests for illicit substance sales. While this form of unidimensional monitoring of trends can be accomplished without GIS, GIS provides the additional capability to rapidly analyze or display geographic sub-populations and to overlay geographic information over temporal information. An example might be maps to discern that increasing rates of low birth weight are occurring in a specific portion of a city.

- *Level 2: Geographic integration of multiple variables.* This integration can occur by area (data aggregated into administrative areas like census tracts, zip codes, or municipalities) or by discrete-point geocoding that displays and analyzes points or imputed spatial surfaces. To continue our example, geographic patterns of rates of low birth weight might be compared with rates and trends in premature deliveries, maternal smoking rates, prenatal care adequacy, prenatal clinic service areas, Medicaid enrollment rates, substance abuse arrests, and the incidence of sexually transmitted diseases. Because population sets are geocoded to discrete locations or very small areas, it is possible both to analyze and to display small-area information. Some information (like economic status) may be cautiously imputed from small-area census or other data. This may indicate associations between low birth weight and other features (with a cautious respect for potential fallacies of multi-level comparisons).
- *Level 3: Individual-level record linkage.* Automated linkage of individual records from multiple databases is now feasible, using probabilistic or deterministic linkage strategies. Record linkage methodologies have become standardized in recent years, and unique identifiers are not always necessary (11). By these methods, linked datasets are created (and subsequently stripped of personal identifiers). These datasets include health risks and outcomes (e.g., from birth records), participation in service programs, insurance type, and community-level data (such as income or exposure to drug sales) imputed from small-area data. To continue our example at this level, imagine that it can be shown that participants in a comprehensive prenatal care coordination program combining clinic-based and outreach-worker care have higher birth weights than individuals who live in the same area, with similar demographic and perinatal risk factors, but who do not participate in the program.
- *Level 4: Real-time, point-of-service information.* At the highest level of integration and functionality, population health databases accompany patients, health care workers, and public health workers on their daily business. Imagine that a young woman presenting for emergency care is automatically identified as receiving (or not receiving) high-risk pregnancy-related outreach services when she registers for care at the local emergency department. Although using population health databases at this level presents the greatest technical and confidentiality-related challenges, there are some existing applications that demonstrate the usefulness of population health systems for individual health services (12). Within an integrated population health information system linked to service delivery, opportunities abound for tailoring prevention, evaluation, and planning to achieve true continuous improvement in population health.

GIS is crucial at each of these levels. It greatly simplifies data management, display, and calculations for the first two levels. Also, when displaying information using commonly understood geographic boundaries, GIS helps communicate the immediate significance of information to a public who might otherwise fail to comprehend that they are at risk, inviting greater participation in planning and policy. The third and fourth levels use the specificity of point location both for linkage of records and for more discrete display and analysis (spatial representation free of arbitrary administrative polygons like zip codes or census tracts, allowing more natural visual and statistical

representations of data). This point-specificity can also facilitate linkage to a greater number of databases, and can do so in a way that may be more respectful of individual confidentiality than would use of names or other personal identifying variables. Address information could link, for example, building age and ownership status (from plat records), median census block income (from the decennial census), housing inspection and lead abatement interventions (from administrative records), and reported blood lead levels (from public health surveillance data). These data could target interventions (service planning), derive predictive models (population-based epidemiology), or evaluate the effectiveness of housing policy changes (outcome effectiveness research). For these reasons, GIS, facilitated by geocoding of health and other data records, becomes the *sine qua non* of population health informatics.

While the first and second levels of information system integration can be accomplished with lower degrees of spatial specificity, creation of spatial surfaces, automated record linkage, and point-of-service integration requires that data records be geocoded. Geocoding of locational data (address of residence, location of injury event or exposure if known, address of place of employment, location of health care service provider) is the essential element of population health informatics. Through geocoding and map generation, GIS provides an essential tool for integration of data records from various sources by location. GIS applications can also serve as laboratories for development, interpretation, and dissemination of neighborhood/community health indicators. GIS also provides an interactive environment for spatial display of health data.

In public health data systems there are numerous perceived barriers to geocoding. These include cost, timeliness and accuracy, staff and equipment needs, and privacy/confidentiality concerns. All of these perceived barriers are smoke screens. There is no valid rationale for not routinely geocoding all records in vital statistics or hospital discharge databases, cancer or birth defects registries, and all other population-based public health information systems. In fact, geocoding can improve the precision of these databases by correctly allocating cases to county, zip code, minor civil division, or census tract, and is extremely efficient when integrated into the routine, day-to-day processing of records accruing to administrative health data systems. Geocoding of population-based public health data system records also facilitates the public health mission of the agencies and programs that sponsor or support the information system, by ensuring that valid, geocoded addresses are available for every record as a precondition for filing. Further, geocoding is easily made routine. Administrators who embrace the approaches of public or population health informatics will achieve maximum value by establishing centralized geocoding centers to process all records for their agencies.

GIS can also play a role in population health as a tool for the generation of research hypotheses concerning the epidemiologic determinants of health status, well-being, health outcomes, and health service utilization. GIS has a more limited role as a platform for hypothesis testing *per se* (13). GIS also provides a supportive environment for population-based public health program planning, program evaluation, and community-based decision-making.

Methodological Issues, Concerns, and Opportunities

The application of GIS in population health provides numerous methodological

opportunities, but raises some significant issues and concerns. These have been discussed at greater length elsewhere (14–18), but include the following:

- Scale and aggregation in measuring contextual variables (i.e., the role of individual characteristics versus neighborhood variables or ecological correlates).
- Points versus areas, and rates versus numerators and denominators.
- Theoretical conceptions of space and analytical applications.
- Integration of spatial modeling and biostatistical methods with social and epidemiological theory.
- Methods of map presentation and interpretation.
- Methodological issues surrounding the quality of matching and record linkage (including geocoding).

Space does not permit a lengthy discussion of these issues here. While none of these issues has a simple solution, identification and development of a multidisciplinary working group to devise and implement GIS-based population health applications will prove beneficial in most settings.

Summary

Population health is an emerging framework for assessing and evaluating the health status and health outcomes of defined populations. It is in many ways a superset of traditional public health functions and goals. Population health informatics is the operationalization of an integrated information systems environment for the practice of population health. GIS is an integral and in many ways essential component of a comprehensive population health informatics system. GIS is, however, only a tool, not an end unto itself in the practice of population health.

References

1. Evans RG, Stoddart GL. 1994. Producing health, consuming health care. In: *Why are some people healthy and others not? The determinants of health of populations.* Ed. RG Evans, M Barer, TR Marmor. New York: Aldine de Gruyter. 27–64.
2. US Public Health Service. 1991. *Healthy people 2000: National health promotion and disease prevention objectives—full report, with commentary.* Washington, DC: Department of Health and Human Services.
3. Cohen MM, MacWilliam L. 1995. Measuring the health of the population. *Medical Care* 33:DS21–42.
4. Murray CJ. 1994. Quantifying the burden of disease: The technical basis for disability-adjusted life years. *Bulletin of the World Health Organization* 72:429–45.
5. Young TK. 1998. *Population health: Concepts and methods.* New York: Oxford University Press.
6. Friede A, Blum HL, McDonald M. 1995. Public health informatics: How information-age technology can strengthen public health. *Annual Review of Public Health* 16:239–52.
7. Lumpkin J, Atkinson D, Biery R, Cundiff D, McGlothlin M, Novick LF. 1995. The development of integrated public health information systems: A statement by the Joint Council of Governmental Public Health Agencies. *Journal of Public Health Management and Practice* 1:55–9.

8. Institute of Medicine Committee for the Study of the Future of Public Health. 1988. *The future of public health.* Washington, DC: National Academy Press.
9. Donaldson MS, Lohr KN, Eds. 1994. *Health data in the information age: Use, disclosure, and privacy.* Washington, DC: National Academy Press.
10. Krieger N. 1992. The making of public health data: Paradigms, politics, and policy. *Journal of Public Health Policy* 13:412–27.
11. Kirby RS. Manuscript in preparation. *Controlling the "urge to merge": Diagnosis and treatment of a new clinical psychosis affecting public health workers and researchers.* Unpublished paper presented at the Centers for Disease Control and Prevention 1996 Maternal, Infant, and Child Health Epidemiology Workshop: Data and Information for Planning, Prevention and Evaluation. Atlanta, GA. December 3–4, 1996.
12. Hall K, Zimmerman A, Samos J, Simon PR, Hollinshead WH. 1997. Coordinating care for children's health: A public health integrated information systems approach. *American Journal of Preventive Medicine* 13(Suppl 1):S32–6.
13. Jacquez GM. 1998. GIS as an enabling technology. In: *GIS and health.* Ed. T Gatrell, M Loytonen. London: Taylor and Francis. 17–28.
14. Kirby RS. 1996. Toward congruence between theory and practice in small area analysis and local public health data. *Statistics in Medicine* 15:1859–66.
15. Clarke KC, McLafferty SL, Tempalski BJ. 1996. On epidemiology and geographic information systems: A review and discussion of future directions. *Emerging Infectious Diseases* 2:85–92.
16. Croner CM, Sperling J, Broome FR. 1996. Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15:1961–77.
17. Briggs DJ, Elliott P. 1995. The use of geographical information systems in studies on environment and health. *World Health Statistics Quarterly* 48:85–94.
18. Waller LA, McMaster RB. 1997. Incorporating indirect standardization in tests for disease clustering in a GIS environment. *Geographical Systems* 4:327–342.

Characterizing the Environmental Features of a Region for a Community-Level Health Study of Breast Cancer

Steven J. Melly (1),[1] Yvette T. Joyce (2), Julia G. Brody (1)
(1) Silent Spring Institute, Newton, MA; (2) Applied Geographics, Inc., Boston, MA

Keywords: breast cancer, endocrine disrupters, drinking water, waste water

We have used a geographic information system (GIS) as the central management and analysis tool in the Cape Cod Breast Cancer and Environment Study. A team of public health and environmental researchers working together with GIS specialists have developed methods for using GIS data from multiple sources with different scales to study environmental factors that might be relevant to breast cancer incidence. This demonstration project will illustrate how we dealt with differences in source scales to estimate the number of households within specified distances of environmental features. Data on pesticide use on Cape Cod will be used to demonstrate how we estimated exposure categories using the GIS. We will also illustrate how researchers with limited GIS background have been able to access and analyze GIS data using ArcView. We will discuss our results from the community-level analyses in the first phase of the study, and our plans for using GIS in a case-control study beginning in the fall of 1998.

Silent Spring Institute is a nonprofit research institute dedicated to investigating the links between women's health and the environment. Since 1994 Silent Spring has been leading a multi-disciplinary team of investigators in a state-funded study of breast cancer incidence on Cape Cod. The team includes researchers from Tufts, Harvard, and Boston Universities as well as GIS specialists from Applied Geographics, Inc., a Boston-based consulting firm.

When the Cape Cod Study began in 1994, Massachusetts Cancer Registry data indicated that age-adjusted breast cancer incidence was significantly higher in a majority of Cape Cod towns than in the state as a whole. Alarmed by these statistics, citizen activists, public health officials, and researchers began sifting through possible explanations. Were high breast cancer rates due to characteristics of women who live on the Cape, or something about the environment?

In our research we focused on two environmental factors that might be different on the Cape from the rest of the state: exposure to drinking water impacted by waste water, and exposure to pesticides through air, dust, and water. Nearly all of the drinking water on the Cape comes from shallow wells, and all waste water is disposed of into the ground because the surrounding waters are a marine sanctuary. Pesticides may have been more heavily used on the Cape because of the large number of cranberry bogs and golf courses and because the native trees are more susceptible to pests.

Using ARC/INFO and ArcView on a Sun workstation, we created a comprehensive GIS for Cape Cod. We had excellent sources of data from the Cape Cod Commission, MassGIS from the Executive Office of Environmental Affairs, and the US Geological Survey. We also digitized data from paper maps that included areas treated with pesticides for tree pests. Figure 1 illustrates some of the pesticide use areas we were able to

[1] Steven J. Melly, Silent Spring Institute, 29 Crafts St., Newton, MA 02158 USA; (p) 617-332-4288; (f) 617-332-4284; E-mail: melly@silent.shore.net

map for the town of Sandwich. We wanted to answer the question: how many women might have been exposed to these pesticide use areas at different points of time? We had land use data from different points in time and parcels data from the 1990s. Because most of the residential land on the Cape is developed as single family homes, we proposed using residential parcels as a surrogate for number of households and, ultimately, number of women.

A challenge we faced was using coverages with different source scales. If we simply identified parcels that intersected residential land use polygons from a specific year, we would include many parcels that were not actually residential because of the differences in source scale. We used two approaches to eliminate the extra parcels. We used attributes in the parcels coverages to eliminate parcels we knew could not be residential, such as wetlands and undevelopable land (Figure 2). We then eliminated any parcels that intersected less than one-tenth of an acre of residential land use (Figure 3).

One possible problem with this approach involves using data sources from different points in time. Does it make sense to use 1990 parcels with 1951 land use data? We reasoned that the only situation when this would be a problem is if previously developed land were redeveloped into smaller house lots. An analysis of land use change from 1951 to 1990 indicates that the vast majority of residential development occurred in previously forested areas. The land use data were broken down according to lot size. We found that very little residential land with less than ½-acre lots in 1951 was later converted to smaller lots, and concluded that it was acceptable to work with the two coverages together.

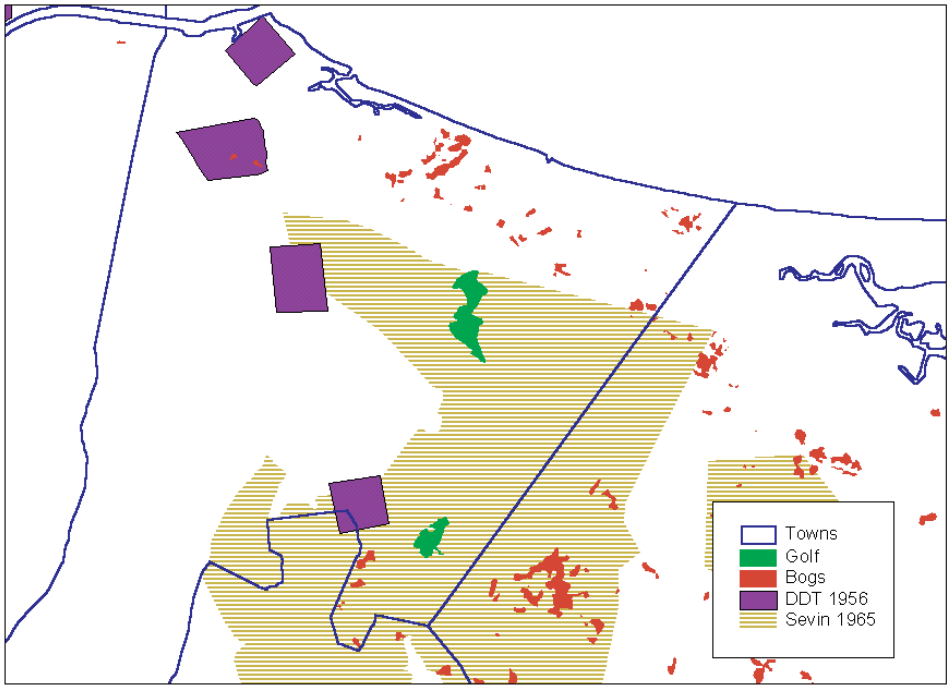


Figure 1 Pesticide use areas in Sandwich, MA. We were able to map large-scale pesticide use areas including cranberry bogs, golf courses, and areas sprayed for tree pests.

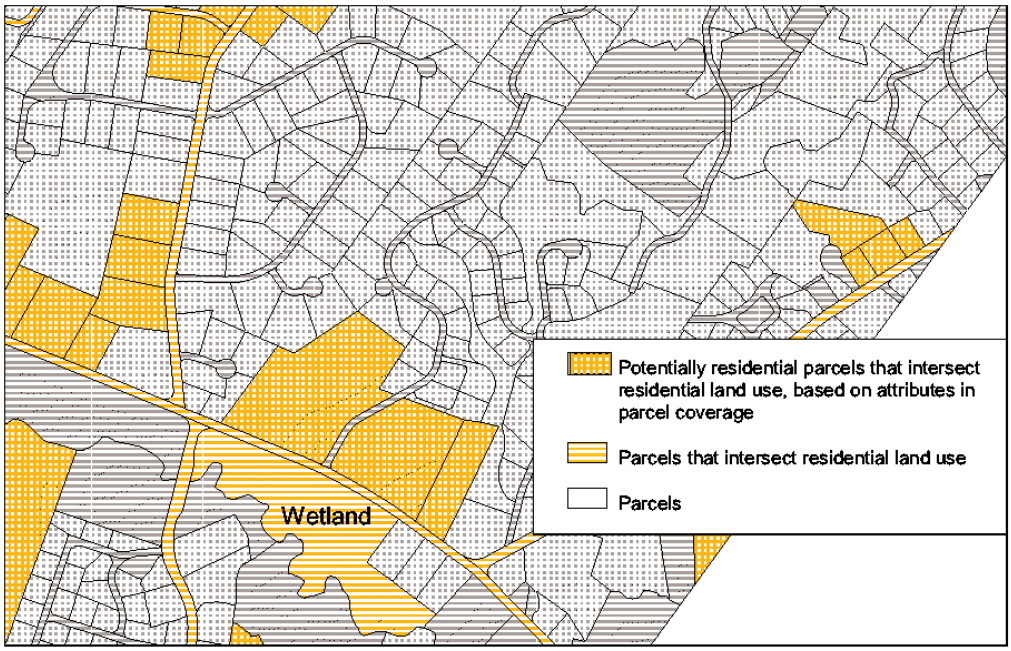


Figure 2 Estimating potentially exposed populations. In order to estimate number of households near pesticide use areas we first identified parcels that intersect residential land use, then used codes from the Department of Revenue to eliminate parcels that could not be residential (such as the wetland, shaded with horizontal stripes).

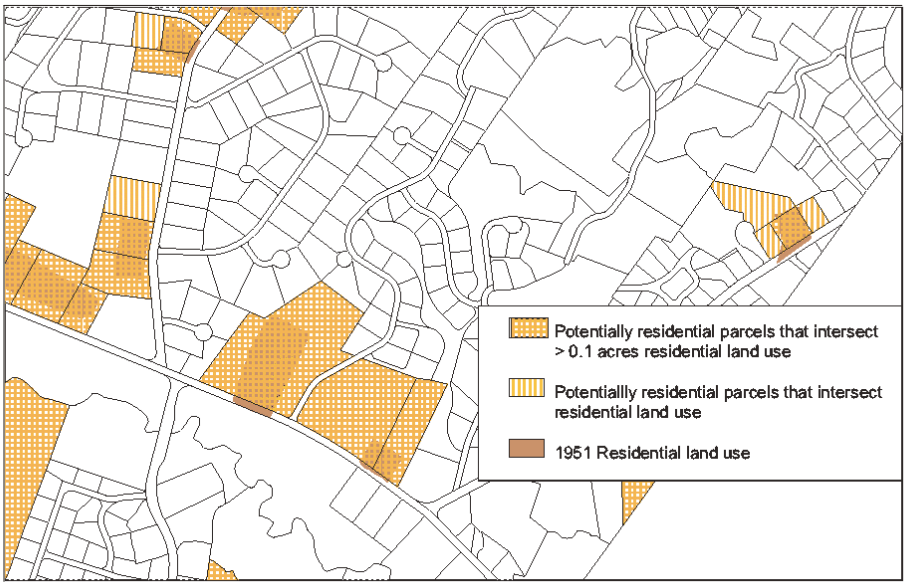


Figure 3 Refining estimates of exposed populations. We further refined our estimates of numbers of households by eliminating the parcels that intersected less than one-tenth of an acre of residential land, shown here shaded with vertical stripes.

Members of our team from Applied Geographics used Arc/Info AML (Arc Macro Language) programs to calculate the distance of residential parcels from various pesticide use areas and add the result to the attribute table. We could then identify residential parcels within a specified distance of pesticide use areas (e.g., 1,300 feet for aerial spray operations). Ultimately we used this information to classify the census block groups of the Cape into high and low exposure categories for the various types of pesticide uses we studied (Figure 4).

We developed this exposure assessment in a community-level study using cancer registry data. We did not see an association between breast cancer incidence and these exposure categories. This is perhaps not surprising given the limitations of a community-level study. We did not have information about the residential histories of the women in the cancer registry files, and we confined our analysis to block groups. The areas most exposed to pesticides are likely to be much smaller than block groups. We plan to overcome these limitations by conducting a case-control study in the near

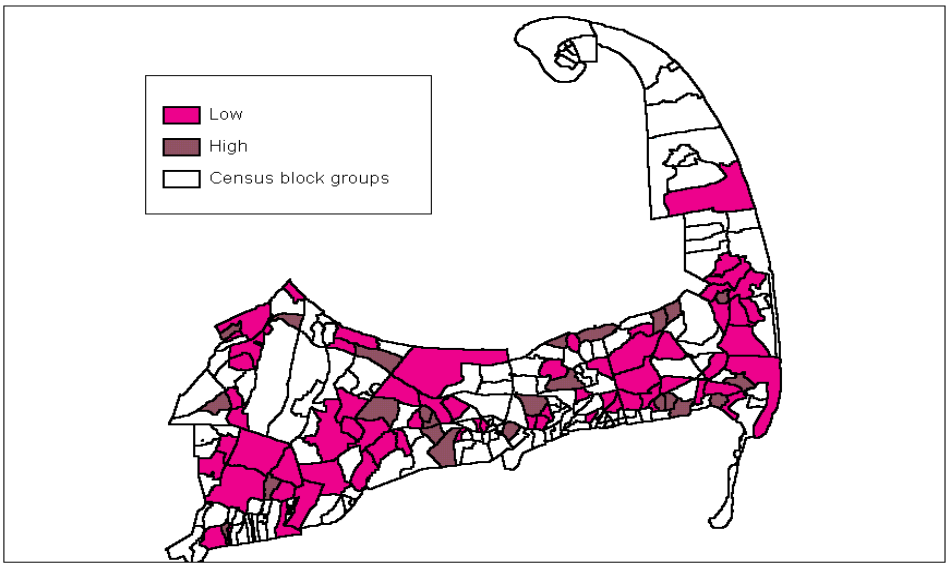


Figure 4 Exposure groups for cranberry bogs. We classified the census block groups according to the number of women potentially exposed to areas of large-scale pesticide use.

future. The study will be designed from the start to incorporate the use of GIS to develop exposure variables.

Acknowledgments

This work was supported by funds appropriated by the Massachusetts Legislature and administered by the Massachusetts Department of Public Health under Contract DPH79007214H11, and by the generous contributions of the Susan G. Komen Breast Cancer Foundation/Boston Race for the Cure.

GIS in Community Health Assessment and Improvement

Alan Melnick, MD, MPH (1, 2),* Nicholas Seigal, MCRP (3), Jono Hildner, MS (4), Tom Troxel, MS (2)
(1) Oregon Health Sciences University, Portland, OR; (2) Clackamas County Public Health Division, Oregon City, OR; (3) Anteus Consulting, Portland, OR; (4) Hildner and Associates, West Linn, OR

Abstract

The Clackamas County Geographic Information Systems Demonstration Project is designed to engage diverse communities in partnerships to make improvements in community well-being. An interdisciplinary team has developed a software system, the Community Health Mapping Engine (CHiME), that allows the easy incorporation of multiple datasets related to community well-being. Through a Healthy Communities partnership process, multiple agencies, both private and governmental, are beginning to share data, allowing the data to be incorporated into the CHiME. The CHiME uses readily available data obtained from vital statistics bureaus, the US Census, private sources (such as hospital discharge data), and county government collected data such as reported crime. The demonstration geographically references these datasets, allowing analysis in a geospatial format at the sub-county, community level. Interested community members and agencies can apply a user-friendly, ultimately Web-based interactive mapping function to assess a variety of health and social demographic factors and benchmarks related to community health and well-being. The demonstration is flexible and modular. As additional public and private datasets become available, the "Data Wizard" can easily incorporate them into the CHiME for use by community members. We are incorporating safeguards to protect confidentiality during small area analysis. The demonstration performs statistical analysis, including confidence intervals, allowing community members to compare their community indices with county, state, and national rates and benchmarks, and follow trends over time. Although current datasets and functionality are limited to Clackamas County, we designed the application to allow expansion to accommodate other regions and geographic scales (counties, states, and nations).

Keywords: community health planning, community well-being, health status indicators, public health administration, data collection

Introduction

The goal of the Clackamas County Geographic Information Systems Project is to increase the capacity of Clackamas County staff and Clackamas County community members by making data analysis and data presentation more accessible, localized, and community-based. By accessible, we mean that community members and interested agencies should be able to obtain relevant information about the health status of their communities at a variety of sites, including local libraries and home computers. By localized, we mean that community-level health data should be available for analysis at the neighborhood/community level. By community-based, we mean that local

* Alan L Melnick, Clackamas County Public Health Division, 710 Center St., Oregon City, OR 97045 USA; (p) 503-494-0756; (f) 503-494-4496; E-mail: melnicka@ohsu.edu

communities should be able to determine for themselves what information about their community is relevant, and that local and state governments should be a resource for these communities in providing for their data needs.

Although local and state governments routinely collect data related to community health status, the data are rarely used by local health consumers and planners for several reasons. First, the data are not timely. For example, up to two years may elapse before vital statistics data are released in hard copy form. Once the data are released, the hard copy report contains limited county level analysis and is not amenable to further data manipulation. Local planners are left to ask the responsible state agency to make specific data runs, requiring additional time and staff support. Second, a variety of health-related data is collected and maintained in different formats by many different agencies at the local, state, and federal levels and is not available in one convenient location accessible to community health planners. Third, health data are analyzed and reported only at the county, state, and national levels. Larger counties often contain many diverse and sizable communities whose borders do not necessarily coincide with other political boundaries and whose characteristics are not captured accurately by summaries based on these boundaries. Consequently, data analyzed and reported at the county level or higher are frequently not useful for many local communities in conducting health assessment and planning. Such data provide little opportunity for local public health professionals to seek dialogue and strengthen relationships with local communities.

The Clackamas County Department of Human Services Geographic Information Systems (GIS) Project is designed to address these issues. Our objectives are to improve access to data by local health consumers and planners and thereby engage our diverse communities in a partnership with us to improve community health. An interdisciplinary team has developed a prototype software application, the Community Health Mapping Engine (CHiME), that allows the easy incorporation of multiple datasets online in a timely manner. Through the Healthy Communities partnership process, multiple agencies, both private and governmental, are beginning to share data, allowing the data to be incorporated into the system. We are encouraging these partners to share datasets that include addresses. These datasets will be geographically referenced to allow analysis in a geospatial format at the local, sub-county community level. Census data (and inter-census data) will serve as the denominator for rates. The Clackamas County CHiME is intended to serve as both an enterprise GIS model and a tool to facilitate community health planning. As an enterprise GIS, the CHiME will serve as a centralized assessment tool for use by multiple county agencies and partners. The future Web-based version of the CHiME, with its help features, will be publicly available to community-based groups and consumers interested in performing community health assessments.

The CHiME will provide Clackamas County communities with a tool to help themselves in at least two ways. It will enable them to assess a variety of factors related to community well-being, and it will allow them to evaluate any actions they take in improving their health status.

Methods

We designed the system for two user skill levels: community members without formal

epidemiologic skills, and advanced epidemiologic investigators. As the prototype is further developed, an initial screen will contain text that describes the project, lists data and data sources, and provides instructions on how to use the system. An epidemiology tutorial will be built into the system for those unfamiliar with epidemiologic concepts. Besides providing instructions on how to use the system, screens will provide easily understood explanations of concepts such as incidence rates, prevalence, confidence intervals, and the need for age adjustment when evaluating mortality rates. Pop-up help screens will contain messages discussing the concept of ecologic fallacy and the need to avoid drawing conclusions when cause-and-effect relationships have not been previously established (1). Help icons and screens will be available at all times. Links to appropriate county health officials will be included, allowing users to ask questions and obtain consultation. Links to other online information sources also will be provided.

Within the current prototype application, users can analyze data at the sub-county, community level as well as at state and county levels, and can present their findings in table, chart, and polygon/map format. Census data (and estimated inter-census data) provide information about demographic characteristics and population counts. For the current prototype application, we purchased inter-census data from a private provider, Equifax National Decision Systems (ENDS) (Atlanta, GA). ENDS provides current-year estimates of demographic and population data in a variety of formats, including ArcView, and has a history of providing such data for commercial use.

The CHiME enables users to compare their community measures with countywide data, statewide data, Oregon benchmarks, and (eventually) national data. Users can compare measures for each geographical area over time and automatically calculate confidence intervals. When rates for a single year are unstable due to small numbers, users can analyze data aggregated over several years. The CHiME can display table and chart data whenever users click on a state, county, or community. The information displayed for each geographic level of analysis includes absolute numbers of events, rates, means, medians, and confidence intervals. Users can zoom in or out among the levels. Users can evaluate two variables simultaneously, so they can visualize spatial patterns and relationships. For example, users can evaluate relationships between teen birth rates and risk factors such as poverty.

A "Data Wizard" allows the project administrators to easily incorporate additional datasets into the system. This Wizard facilitates the process of geocoding and adding new data to the CHiME. Varieties of common data formats are supported. Each dataset must include an address field for purposes of geocoding. Table 1 lists types of data currently included in the CHiME. Several of these datasets currently only allow analysis at the county level or above. Datasets allowing analysis at the sub-county level will be added as address fields are completed. We envision that all health-related data eventually will include an accurate address field to enable analysis at the community level. Examples of data of special interest include mortality (so the CHiME could calculate years of potential life lost [YPLL] and age-adjusted mortality rates at the community level); immunization rates for children aged two years; cancer registry data; high school dropouts; commuting time; and, domestic abuse (including elder, child, and spouse). In addition, we plan to include data such as hospital discharge diagnoses through working partnerships with health care systems and health care providers.

Once the application is Web-based, we will ensure confidentiality in two ways.

Table 1 Data Included in the Clackamas County CHiME, 1998

Variables	County Level of Analysis	Community Level of Analysis	Years	Data Sources
Age, gender, race	X	X	Single years: 1990 to 1996	Equifax National Decision Systems
Personal income	X	X	Single years: 1990 to 1996	Equifax National Decision Systems
Births (including repeat births)	X	X	Single years: 1990 to 1996 Aggregate: 1991 to 1995	Oregon Health Division, Vital Statistics
Abortions	X		Single years: 1990 to 1996 Aggregate: 1991 to 1995	Oregon Health Division, Vital Statistics
Pregnancies	X		Single years: 1990 to 1996 Aggregate: 1991 to 1995	Oregon Health Division, Vital Statistics
Deaths	X		Single years: 1990 to 1996 Aggregate: 1991 to 1995	Oregon Health Division, Vital Statistics
Suicides	X	X	Single years: 1990 to 1996 Aggregate: 1991 to 1995	Oregon Health Division, Vital Statistics
Arrests	X	X	Single years: 1990 to 1996 Aggregate: 1991 to 1995	Clackamas County Sheriff's Department
Reported crimes	X	X	Single years: 1990 to 1996 Aggregate: 1991 to 1995	Clackamas County Sheriff's Department

First, agencies sharing data will use the Data Wizard to geocode individual records and then aggregate the records into defined geographic communities. Agencies will thus remove all individual identifiers before sharing the data with the CHiME. Not only does the Wizard help assure confidentiality, its geocoding and aggregating properties have already encouraged formerly reluctant agencies such as hospitals to share their data with us. Once the data are in the CHiME, we will further ensure confidentiality by restricting analysis, reporting, and depiction of very small numbers, especially when multiple stratification is performed.

For compatibility with population data sources (used for the denominators), we have defined communities as census block groups aggregated to approximate high school attendance areas. We chose not to use zip codes because they cross community and city boundaries and it is difficult to obtain denominator data for them. Following community input, the Data Wizard could aggregate block group data to create maps for alternative target areas such as legislative districts, elementary school attendance areas, or other user-defined small areas. We conducted several focus groups, including those with the elderly, teens, and minority populations, who concurred with our initial decision to use high school attendance areas as geographic community definitions.

In Figures 1 through 12, the CHiME has been used to generate sample maps that show teen male arrest rates, teen birth rates, and adequacy of prenatal care by high school attendance areas. Juvenile (teen) arrests, teen pregnancy rate, and adequacy of prenatal care are three of Oregon's benchmarks, measurable indicators for which data are reliably, regularly, and economically available. Oregon currently has 92 benchmarks, reduced this past year from 259. Benchmarks are developed through a public

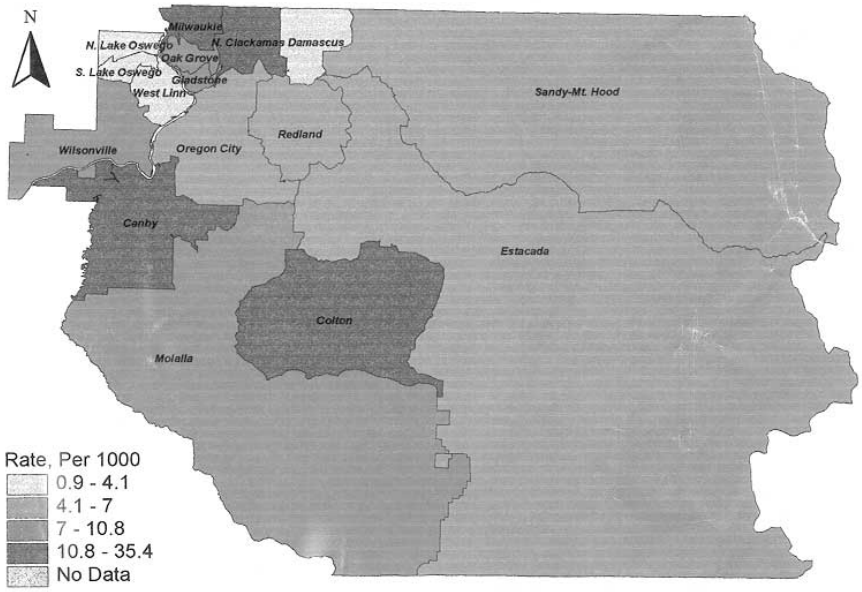


Figure 1 1996 community-level teen birth rates by quartile (CHiME).

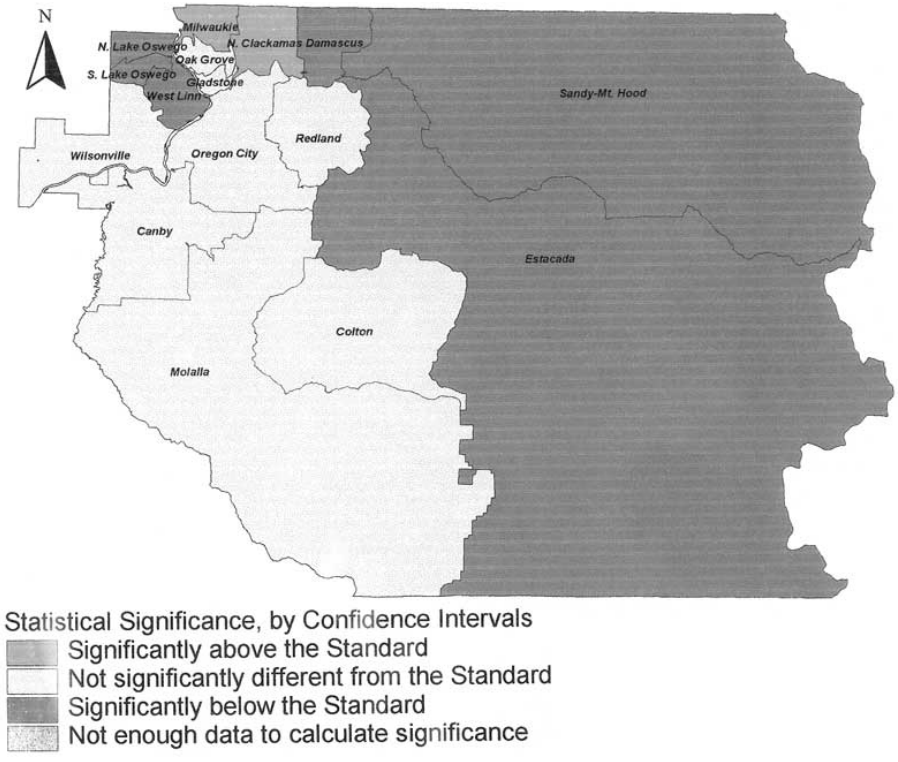


Figure 2 1996 community-level teen birth rates compared with the state rate (CHiME).

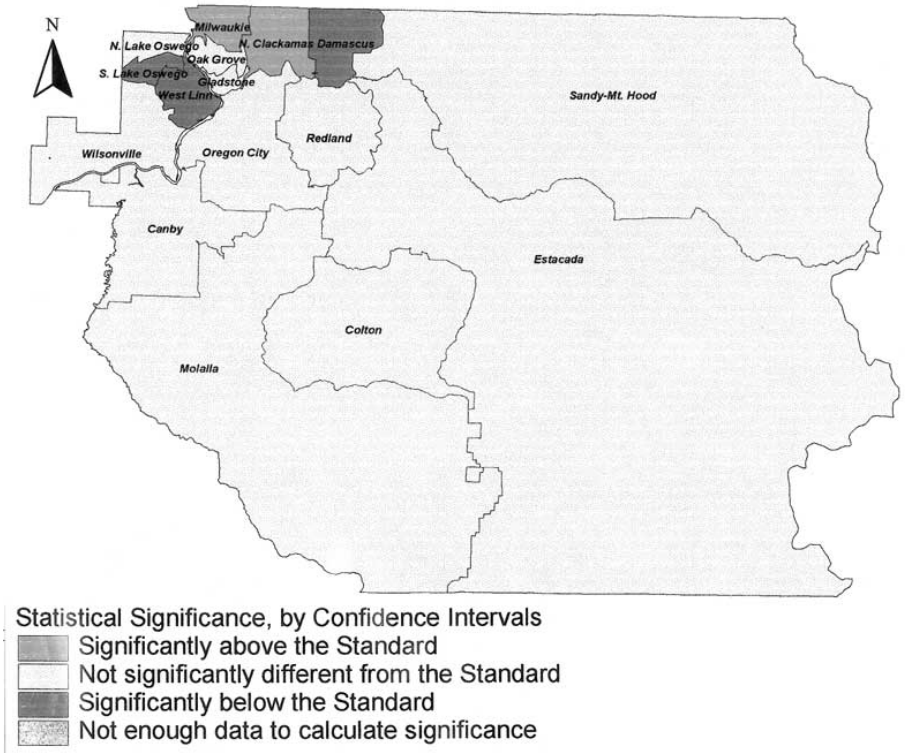


Figure 3 1996 community-level teen birth rates compared with the county rate (CHiME).

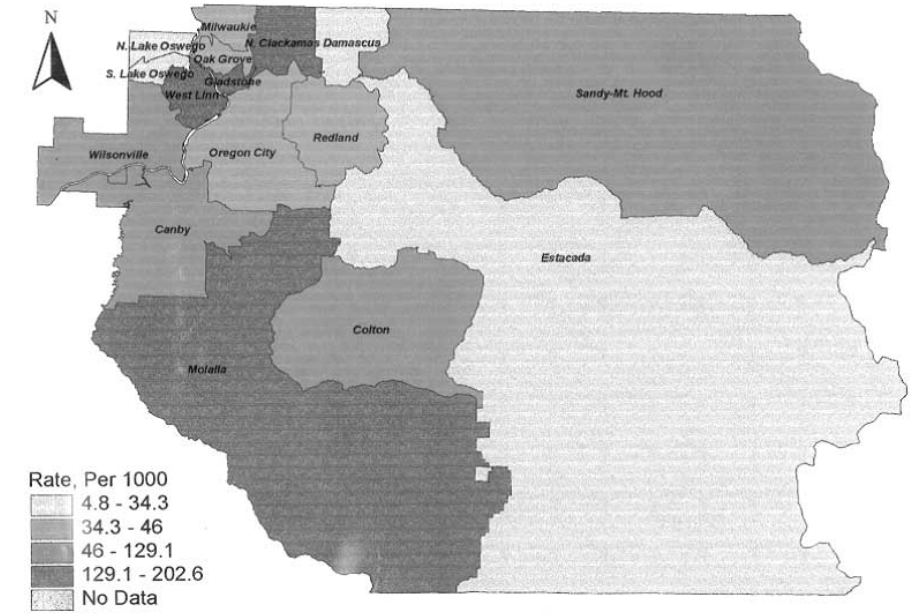


Figure 4 1995 community-level teen male arrest rates by residence by quartile (CHiME).

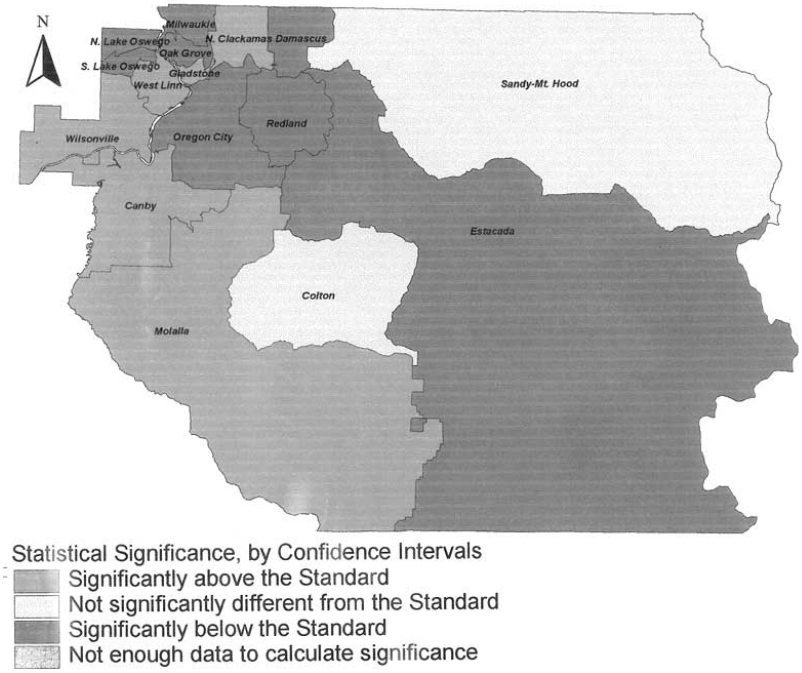


Figure 5 1995 community-level teen male arrest rates compared with the county rate (CHiME).

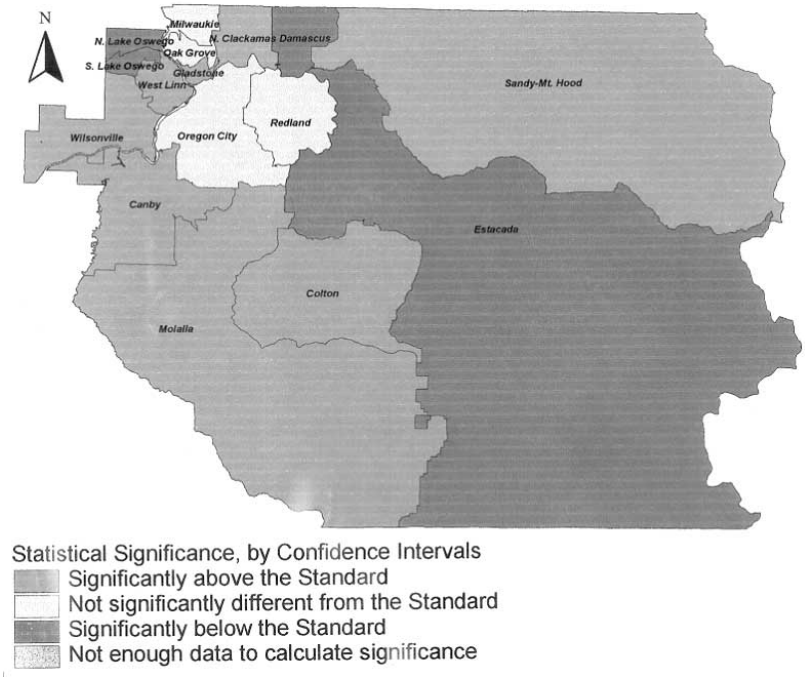


Figure 6 1995 community-level teen male arrest rates compared with the Year 2000 Oregon state juvenile arrest rate benchmark (CHiME).

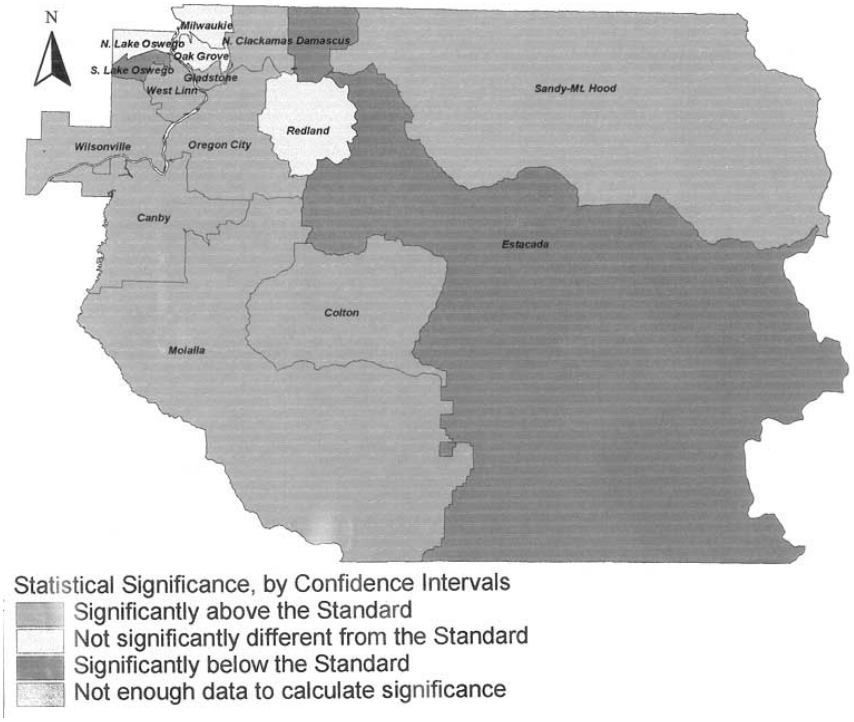


Figure 7 1995 community-level teen male arrest rates compared with the Year 2010 Oregon state juvenile arrest rate benchmark (CHiME).

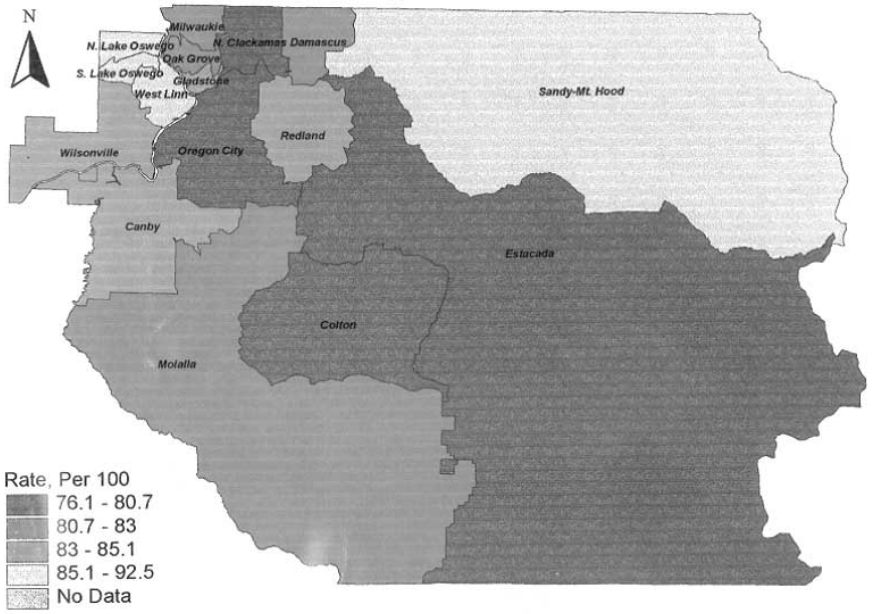


Figure 8 1996 community-level percentage first trimester care by quartile (CHiME).

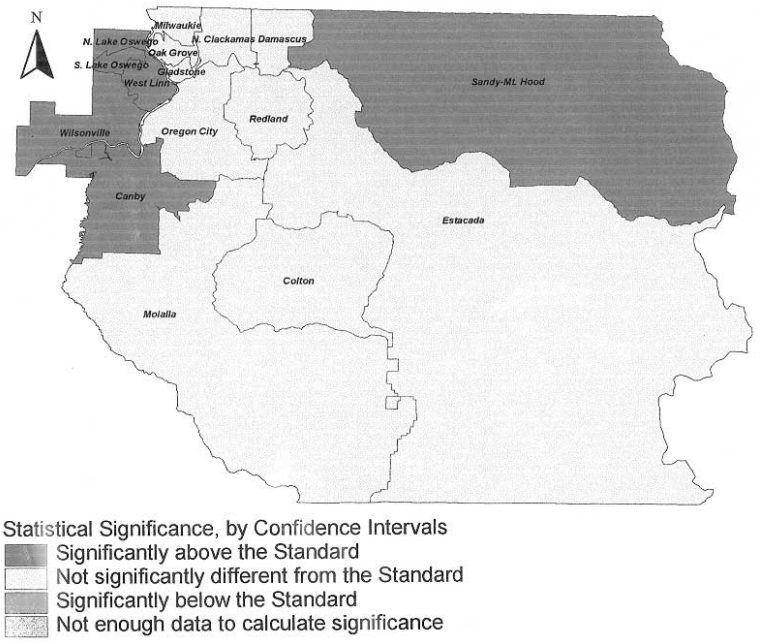


Figure 9 1996 community-level percentage first trimester care compared with state percentage (CHiME).

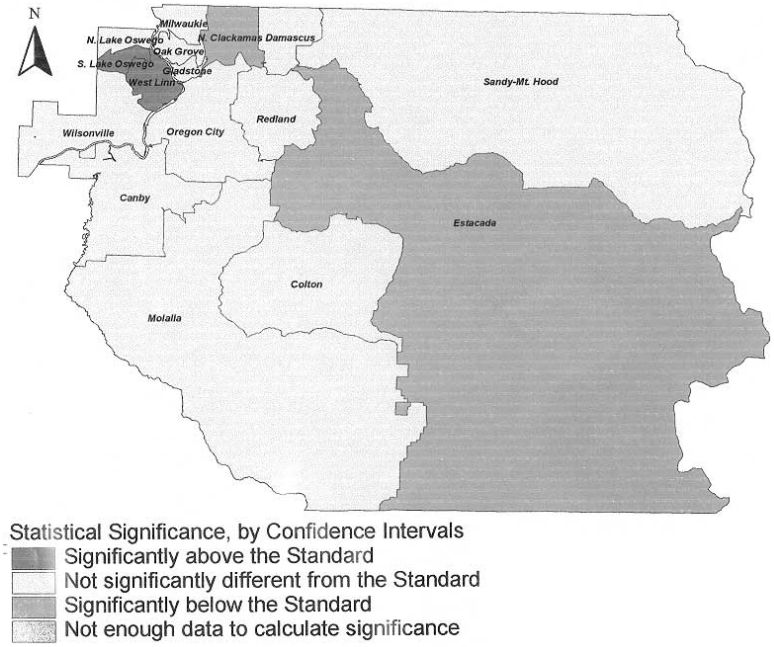


Figure 10 1996 community-level percentage first trimester care compared with county percentage (CHiME).

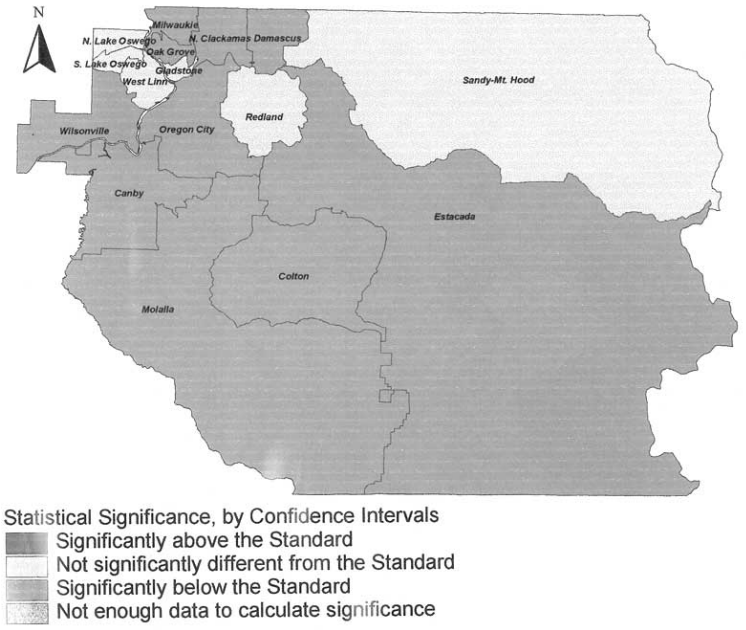


Figure 11 1996 community-level percentage first trimester care compared with the Year 2000 Oregon first trimester care benchmark (CHiME).

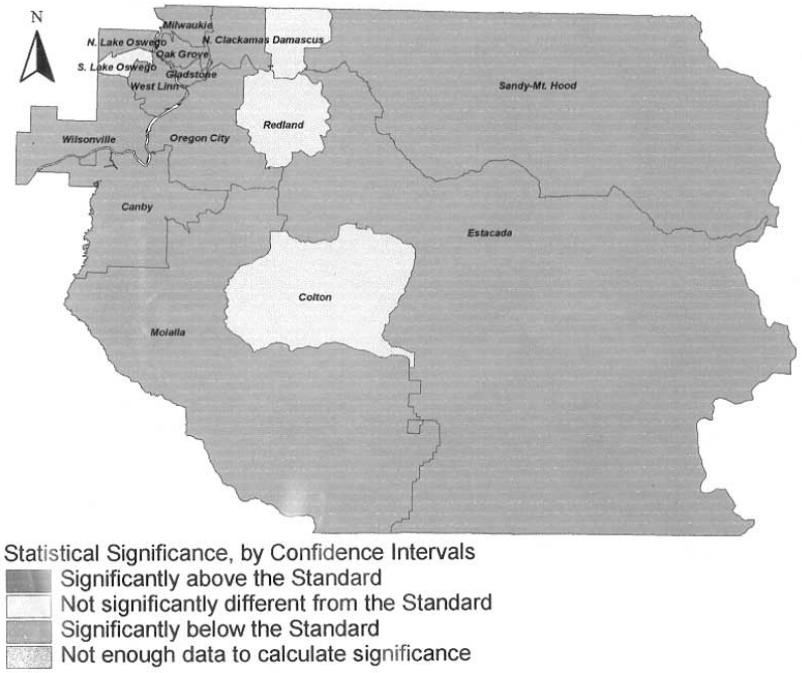


Figure 12 1992–1996 aggregate community-level percentage first trimester care for Hispanic women compared with the Year 2000 Oregon first trimester care benchmark (CHiME).

process by the Oregon Progress Board, an independent state planning and oversight agency (2). Created by the Legislature in 1989, the Progress Board is responsible for implementing the state's 20-year strategic plan, Oregon Shines. The newest version of the strategic plan, Oregon Shines II, has three major goals: quality jobs for all Oregonians; safe, caring, and engaged communities; and, healthy sustainable surroundings (3).

Ten of the current Oregon benchmarks focus on traditional public health indicators, such as infant mortality, teen pregnancy, and percentage of adequately immunized two-year-olds. However, many of the other benchmarks have public health implications. The Progress Board realizes that connections exist between all three goals and most benchmarks.

Using a public process involving thousands of Oregon residents, the Oregon Progress Board established the juvenile (teen) arrest rate as one of the benchmarks for the goal of safe, caring, and engaged communities. The Clackamas County Public Health Division views the teen arrest rate as a benchmark with public health implications, and one where CHiME potentially can play an important role in the community partnership.

Even before being placed on the Web, the Clackamas County CHiME has involved community residents through a variety of venues, including the Reduce Adolescent Pregnancy Project (RAPP), the Healthy Communities Council, the Robert Wood Johnson Turning Point Partnership, and the Local Public Safety Coordinating Council. The local RAPP group is particularly interested in looking at teen birth rates by high school attendance area and by legislative district. Both groups are interested in looking at trend data.

Healthy Communities is a partnership involving community residents, local governments, hospitals, health plans, businesses, schools, religious leaders, and other agencies in the Portland metropolitan area. Clackamas County is working with the Healthy Communities Council to expand the number and variety of datasets available for the CHiME and, ultimately, to build an infrastructure for cooperation and data sharing across organizational boundaries.

Clackamas, Multnomah, and Washington Counties (the three counties in the Portland metropolitan area), in conjunction with the Healthy Communities Council, have developed a local partnership funded through the Robert Wood Johnson Turning Point Initiative to study how public health services are delivered and to make recommendations for improvements. One goal of our Turning Point initiative is to develop an integrated data system. Healthy Communities and Turning Point have expressed an interest in using the CHiME as a way of integrating and sharing data among all of our partners.

Before the development of the CHiME, Clackamas County law enforcement agencies used the location of crime and arrest events (rather than rates) in determining where to deploy resources. Following input from its Department of Human Services member, and in consideration of the established juvenile arrest rate benchmark, the Juvenile Crime Subcommittee of the Clackamas County Local Public Safety Coordinating Council became interested in looking at juvenile arrest rates as a measure of community health and safety. Their interest increased when they found that CHiME could allow them to map and analyze juvenile arrest rates and associated risk factors at the sub-county, community level. Because the Clackamas County Sheriff provides the raw reported crime data, the Clackamas County CHiME could help the Juvenile Crime

Subcommittee visualize patterns of juvenile arrests in relation to demographic factors, specific crimes committed, and community health indicators such as the poverty rate. An example of a geographic analysis of teen male arrest rates is illustrated in several sample maps (Figures 4–7). The case definition for a teen male arrest is the arrest of a male, age 10 to 17, that is reported by law enforcement agencies. Rates were calculated based on the residence of the arrested teen. Rates also may be calculated, however, based on location of the reported crime.

As illustrated in the sample maps, when calculating arrest rates for the community, county, and state, the CHiME provides 95% confidence intervals, making it possible for communities to determine whether their arrest rates are significantly above or below the benchmark arrest rates. By adding a time trend analysis feature, CHiME eventually will enable the Public Safety Coordinating Council and other community partners to evaluate the effectiveness of neighborhood level initiatives to prevent juvenile crime.

Discussion

We have learned several lessons from our early experience with the CHiME demonstration. We learned that communities can be defined in many ways and that polygon representations of rates are frequently more useful than point representation of events for community health assessment and community health planning efforts. We learned that the public must be involved early in the process in defining community and determining what issues are addressed in a community health assessment. High school attendance area proved useful as the unit of analysis because it was meaningful as a community definition for the general public and because for two of our measures (teen arrests and teen birth) it facilitated targeting interventions and educational messages at high school teachers, students, and their parents.

We have also learned that we need to be careful when making multiple statistical comparisons when, for example, we compare multiple community teen male arrest rates with the county rate. Consequently, the CHiME can calculate Bonferroni adjustments for these comparisons. Most importantly, we learned to be vigilant to ensure that cause-and-effect conclusions are not drawn from ecologic data. These data should raise questions, not answer them.

We are not alone in learning from our experience with the CHiME. Our governmental and private partners are learning that reported data must include an address field for the data to be useful in assessing community health. Of course, we have all learned that confidentiality safeguards are essential in analyzing data at the neighborhood level.

Several technological issues remain to be addressed, including the ability of our GIS system to match addresses accurately, especially in rural areas. Even in urban areas, new roads are often constructed or the names of existing roads are changed. Interestingly, during development of the CHiME, the address of the Clackamas County Health Clinic changed when the road name changed.

Future Plans

Even in the early stages of the CHiME application, we foresee future short-term and long-term developments. In the short term, within the current prototype application,

we envision adding additional datasets, both public and private, such as hospital discharge data. A time trend analysis will enable us to evaluate the effectiveness and outcomes of our health programs over time. The upcoming Web-based application will be more accessible than our current version. Shortly, we hope to replicate the current prototype application in other jurisdictions. The CHiME prototype application is universal; for replication elsewhere, all it requires are census block group data for the denominator, local county map data, and community definitions. The Data Wizard can easily be upgraded to allow incorporation of new county templates. Within the current prototype application, we plan to add documentation, including pop-up information screens, metadata, tutorials, help windows, hyperlinks to experts, and a report on the address match rate. We plan to add additional variables for stratification, such as income.

Within the short term, we envision coordination with other community health assessment initiatives such as APEXPH'98. Clackamas County is one of a few counties nationwide piloting the use of a draft version of APEXPH'98 software for the APEXPH Community Process (4). In many of these counties, a major issue has been how to assess community health, given the scattered locations of health-related data. APEXPH provides local communities with a tool to organize the process of community health assessment. For jurisdictions containing multiple or diverse communities, GIS tools such as CHiME can facilitate the APEXPH'98 process, both for the entire jurisdiction and at the sub-county, community level. APEXPH'98 and GIS tools are complimentary. Future versions of the APEXPH Community Process tool should include a geospatial component.

Within the long term, the next generation prototype CHiME application will allow users to define community while using the CHiME. Instead of conforming to pre-selected community boundaries like high school attendance areas, users will be able to draw their own community boundaries. The only restriction to community boundaries will be that they approximate census block group boundaries. The current prototype was developed with ArcView GIS, but future versions will be developed using application-independent languages such as Visual Basic, Java, and Map Objects.

Unfortunately, we also anticipate significant barriers to further development of our application. Upgrades will be expensive, and project needs are growing beyond the scope of Clackamas County. Perhaps this is our biggest lesson: the future of using GIS for community health improvement will require a committed, collaborative partnership of governmental and private agencies and consumers.

References

1. Morgenstern H. 1998. Ecologic studies. In: *Modern epidemiology*, 2nd ed. Ed. KJ Rothman, S Greenland. Philadelphia: Lippincott-Raven Publishers. 459–80.

2. Oregon Progress Board. 1994. *Oregon benchmarks: Standards for measuring statewide progress and institutional performance. Report to the 1995 legislature*. Salem, OR: Oregon Progress Board. December.

3. Oregon Progress Board. 1997. *Oregon Shines II—Updating Oregon's strategic plan: Highlights. A report to the people of Oregon from the Oregon Progress Board and the Governor's Oregon Shines Task Force.* Salem, OR: Oregon Progress Board. May.

4. National Association of County and City Health Officials. 1991. *Assessment protocol for excellence in public health* (APEXPH). Washington, DC: National Association of County and City Health Officials.

Using a Comprehensive Community Health Information System for Public Health Planning and Program Delivery

Cordell Neudorf (1),* Nazeem Muhajarine (2)
(1) Strategic Health Information and Planning Services, Saskatoon District Health, Saskatoon, Saskatchewan, Canada; (2) Department of Community Health and Epidemiology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada

Abstract

Across Canada and elsewhere, health care systems are regionalizing and devolving local authority for health care delivery, financing, and management. Increasingly, regional health authorities find they need better tools for health care planning and decision-making to increase efficiency and effectiveness of service delivery. In order for information to be useful for this purpose, it needs to be timely, relevant, and available in sufficient detail in a flexible manner to decision-makers and planners. Information technology is being enthusiastically seen as a critical tool for addressing various needs of decentralized regional health authorities. Technological innovations combined with data are being increasingly sought for meeting various needs such as health status assessments, integrated planning, clinical and programmatic decision-making, evaluation of outcomes and targets, and dissemination and communication. Data output, whether as hard copy or via intranet or the Internet, needs to be customized according to the user (board member, administrator, manager, staff, the public) and according to the intended use (planning, evaluation, report, research). Data can be presented in tables, charts, graphs, or maps to make output more understandable and relevant to local contexts. This paper describes an initiative underway in a regional health authority (Saskatoon District Health) in Saskatchewan, Canada, to create a comprehensive community health information system (CCHIS) for the district. The purpose of the CCHIS is to provide a database system that supports and enhances a broad range of activities and functions in the district at all decision-making levels.

Keywords: health information systems, regional health authorities, health planning and evaluation

Introduction

As virtually all provinces in Canada move to either decentralized or devolved governance of the health care system, the need for reliable and relevant information at the local level for decision-making has become greater than ever. The information is needed for a variety of decision-making functions mandated to the regional boards, such as local planning, setting priorities, allocating resources, and managing and delivering services. In larger regional jurisdictions (such as capital health authorities), this information requirement may be satisfied in a number of ways, but smaller health districts have fewer options. Furthermore, from the perspective of the regional health authorities, it is more desirable if the information is generated locally, and managed or

* Cordell Neudorf, Strategic Health Information and Planning Services, Saskatoon District Health, 2nd Floor Administration, Saskatoon City Hospital, 701 Queen St., Saskatoon, Saskatchewan, Canada S7K 0M7; (p) 306-655-4415; (f) 306-655-4414; E-mail: neudorfc@sdh.sk.ca

controlled locally. This report describes an initiative underway by Saskatoon District Health (SDH) (Saskatchewan, Canada) to respond to this challenge by creating a comprehensive community health information system (CCHIS) for the district.

Health Information System Goals and Objectives

SDH's Strategic Health Information and Planning Services Department has a mandate to take information from many areas (research, service utilization, health determinants, behaviors and outcomes, and policy trends) and, by a variety of means (analysis, interpretation, dissemination in a variety of formats), turn this into knowledge for use in planning, evaluation, and policy-making. The department consists of three divisions—Research Services, District Health Services Utilization, and Population Health Surveillance Units.

The CCHIS is a key tool for all of these units to use in supporting the decision-making functions mandated to the SDH. Specifically, the CCHIS will enable strategic decision-makers to identify population health needs and prioritize these needs, configure health care delivery systems, promote cross-sectoral partnerships for health improvement, and evaluate and monitor service utilization and health outcomes. A key component of the CCHIS is the linkable (relational) database, which assembles data from multiple sources, within as well as outside the health sector, and is supported by appropriate information technologies and tools for end-user analysis, including the use of a geographic information system (GIS).

The objectives of the CCHIS are to:

- Enhance strategic decision-making by improving the quality of data used, expanding the breadth of relevant data available, and presenting information in an easily understandable manner.
- Facilitate the use of evidence in planning and delivery of services by making information directly available (e.g., online) to key staff members in the district.
- Ensure consistent, timely, and efficient data delivery by assembling and managing a single repository that incorporates a variety of sources and elements of data.
- Increase staff's skills in the analysis, interpretation, and application of information in planning by providing training in the principles of data use, attributes of the database, and analytical tools.
- Promote partnerships within the district health system and between the health system and other key sectors by sharing data, information, and other resources.
- Contribute to the research literature by conducting studies and communicating findings of original research.

Context and Concept

SDH Background

Established in 1992, SDH serves a population of 230,000. A majority of the population lives in the urban center of Saskatoon, while others reside in the surrounding small towns and rural area. Over the six years since its establishment, SDH has achieved sig-

nificant integration of health services and is now directly responsible for providing acute care, long-term care, rehabilitation, home care, mental health services, addiction services, and public health services to district residents. In addition, SDH also has affiliation agreements with several surrounding predominantly rural health districts for providing public health services.

Conceptual Model

The conceptual model we have adapted is from Roos et al. (1), originally developed by Evans and Stoddart (2). This model is helpful primarily in two ways. First, it defines the scope of what is meant by health, its determinants, and its consequences. This enables us to establish the rationale for selection of a wide array of data sources and specific measures from these sources representing each of the three dimensions—health and function, determinants of health, and consequences of threats to health. Second, this model helps us understand how determinants are linked with health status, and how health status is linked with the consequences of health problems. In other words, the conceptual model being adapted depicts how the concepts of determinants, health, and consequences of ill health may be linked with each other.

Specifically, the conceptual model combines a range of environmental factors (both external as well as internal environments) that influence and are influenced by individual responses, which in turn lead to manifestations of health and well-being. These individual responses are of many types, with two of the main types being behavior (e.g., smoking, physical activity, diet) and psychosocial factors (e.g., social support, self-esteem). Health status and, more specifically, threats to health, again are mediated by individual responses (e.g., help-seeking behavior, perception of need, and availability of services), which affect demand and use of health care services. Health care services and, in particular, their effectiveness, lead to new outcomes—restoration or rehabilitation of health status and well-being—thereby feeding back into the model in an iterative manner.

Community Health Information System

The proposed CCHIS is conceived as a dynamic system. The system will continually collect, analyze, and present information in a usable form for decision-makers. For example, using the system, we can produce detailed reports every few months, scholarly papers and articles, and timely broadsheets showing current figures and trends in selected topic areas. All these products go into making up a data-driven information system that is closely linked with an evidence-based health planning and decision-making system.

The CCHIS is, technically, a network of information systems from the provincial level through to the regional and the local levels (i.e., sub-district level), each connecting "up" as needed for district-wide use. This type of information system requires organizational change and is enabled by today's information technology. It relies on the use of computers and communication technology to decentralize information and communicate it to the appropriate points of use (e.g., service delivery, monitoring, planning, and evaluation).

The CCHIS will include an extensive communication network capability to link together various entities within the system-utilization management, the providers of

public health and institutional-based services, affiliate health care organizations, and the provincial, regional, and local governments. These connections will enable the organizations to communicate with each other, exchange data, and have access to views of pre-analyzed information in a highly secure and timely manner. The communication technology will enable the networking of entities within the health care system and, in some instances, outside the system in a seamless way that helps them to function as a whole interactive system. This linkage is an important part of the underlying infrastructure.

Another important part of the infrastructure is the data repository. The CCHIS will assemble a diverse range of data on the health status and health determinants of the population. These data will be linked and available in pre-analyzed format as well as for ad hoc analysis. This data repository will serve as the basis for a wide variety of functions including regular production of health status monitoring reports, needs-based planning documents, resource allocation, evaluation, and utilization and outcome research.

The proposed CCHIS will have the following three main functions:

- Assembly of existing health data on the population, mortality, morbidity, supply of hospitals and health professionals, utilization of services, and more. Some qualitative data on public expectations and preferences will need to be collected.
- Analysis of data to produce health information on the needs, preferences, and health status of the population. Closely related is the function of synthesis and interpretation of the information to produce evidence, spelling out the implications for health planning, services, research, and policy-making.
- Dissemination of the results to decision-makers and the general public.

The CCHIS itself will not make policy decisions, but will notably enhance the factual basis of the organization and its partner agencies for decisions regarding planning, allocation of resources, what services to provide, and what research areas to pursue.

As important as it is to generate regular and timely reports for decision-makers, this function in and of itself is an incomplete conception of the CCHIS. The system is conceived to allow sufficient flexibility to enable a "participatory dialogue" between those who produce the information and those who utilize it (i.e., decision-makers, planners). This dialogue includes, as part of the operational processes, a re-channeling into the CCHIS data collection system of information gathered as a result of using the data by the decision-makers.

For example, along with fact sheets and regular reports disseminated to a wide range of individuals, we will include brief questionnaires seeking views on the usefulness of the system processes and feed them back into the design of the data collection. There are two key issues on which regular feedback would be valuable: first, how the findings are being used, and second, how the findings should be presented, and to whom. Both of these issues are at the heart of the challenge for effective dissemination, and technology provides a vital tool for its solution. The technology allows information and findings to be presented in ways ranging from informative paper-based graphics to interactive programs via intranet systems. Thus, the functions of information production and evaluation of its usefulness are built into the system.

Potential Uses of Data

The potential uses of the CCHIS are many, ranging from health status monitoring, aiding in needs assessment and all other stages of a planning cycle, to research, broadly defined. Correspondingly, the potential users of information produced by CCHIS also are many and varied. Because of this diverse audience, the system will need to be user-friendly and flexible. We plan to use both the SDH intranet as well as the Internet as distribution platforms, supplemented by hard copy reports. Tabular data will be presented in pivot tables to enable users to customize their view of the data, and allow for custom queries and reports of the database, depending on the user's level of access. A GIS will be part of the system and will allow these data to be viewed and analyzed geographically. Some users will need to have the limitations of specific data types explained and the meaning of the information interpreted, while others will be able to do their own independent analysis. Depending on the audience, a varied degree of interpretation and explanation will be provided with the data.

Currently, we are using Microsoft Excel and Access to store data, and are viewing the data with MapInfo Professional (MapInfo, Troy, NY). SDH is moving toward the creation of a single electronic client record using Oracle-based solutions. As the project progresses, this database will greatly enhance the analysis capabilities of the system. Other GIS tools will be evaluated according to their ability to be used by a wide variety of users over the Internet, with the goal of switching to a fully interactive Internet mapping tool in the future. With this capability, many different types of reports are possible depending on the intended audience and the use intended for the information.

Examples of reports that could be generated for various audiences include:

- Reports for board members and health district senior administrators:
  - Summary data, but at any geographical level found to be useful and appropriate.
  - Analysis of trends in demographics, utilization rates, and frequencies of illness, to better target and forecast health service needs.
- Reports for general managers and managers:
  - Individual access to certain views of the data (macro reports, pre-done graphs).
  - User queries to automatically generate graphical or numeric data presentations that help guide planning and evaluation.
- Reports for our partners (other health and human service agencies).
- Reports for the community:
  - For lobbying, advocacy, etc., via hard copy and the Internet (SDH home page).
- Reports/data for researchers.

SDH would be especially interested in the potential for using these data for answering applied research questions and guiding resource allocation within the health service sector.

Progress to Date

Considerable progress has already been made, both in developing and sustaining

partnerships and in acquisition and use of data. The partnership between Public Health Services (PHS) in SDH and the Corporate Information and Technology Branch (CITB) of Saskatchewan Health has been a cornerstone in this project. CITB has been working on a data warehouse project within Saskatchewan Health for several years, bringing together data from various branches within the department. Some of these data were presented to health districts for their use in health planning, aggregated at the district level as a Community Profile. This is quite useful for smaller districts, but less useful for districts with larger populations. Therefore, Saskatchewan Health is providing neighborhood-level data from their data warehouse project to SDH as a pilot project. Once SDH has found the most useful data fields and formats for health status monitoring and health planning, Saskatchewan Health hopes to roll this out to the other health districts as well.

Another key partnership has been between PHS and the University of Saskatchewan's Department of Community Health and Epidemiology. University faculty have helped PHS do background research and develop the conceptual model and evaluation plan. They have also helped in co-authoring grant proposals and articles, and in developing interest in the research community for the potential uses of the CCHIS. This partnership has been critical to helping establish the credibility of the project with our partners.

At appropriate intervals, other partner agencies have been asked to join in the discussions. These include the University of Saskatchewan's Department of Geography, the City of Saskatoon's Planning Department, the Saskatoon Tribal Council, and the Saskatoon Regional Intersectoral Committee (comprised of the provincial Departments of Health, Social Services, Education, and Justice).

Data Acquisition

Data acquisition and management are anticipated to be continuous processes that will include adding data sources and historical data. Currently we have data from the census and vital statistics, as well as data on hospital utilization, the population covered by provincial health insurance, and communicable diseases. Saskatchewan Health and SDH have provided these data primarily at the neighborhood or postal code level, which is more detailed than the district-level data routinely available from the province. Next, we are requesting data on physician services, the prescription drug plan, home care, long-term care, the cancer registry, and mental health.

Selected data from Saskatchewan Education, Saskatchewan Social Services, and Saskatchewan Justice will also be requested. We are also now in the process of adding other public health data, including immunization rates. The aim is to have access to these data at a reasonably detailed level based on certain demographic and geographic variables without it becoming identifiable at the individual level. This can be done by only acquiring the data at a certain level of aggregation, or by only releasing the data in a re-aggregated format but after data linkage has occurred. The latter option would allow for more detailed and robust analysis of associations in time and space between variables in the dataset, beyond purely ecological analysis.

Data linkage can be a time-consuming, error-ridden task. Several initiatives being undertaken in Saskatchewan make this more palatable. The data warehouse project in Saskatchewan Health is already utilizing the health insurance services number (a

unique identifier) for data linkage. Also, SDH is embarking on the task of developing a single patient registration system for use by all services in the district, which will use the same health number as the unique identifier. These initiatives will make data linkage within the health sector much easier. Linkage of data between sectors may be more difficult. We will start with ecological analysis at a relatively small geographic level and work on increased linkage over time. We also hope to incorporate local survey data being collected in the district, as well as qualitative data via a searchable index of keywords.

Roadblocks and Resolutions

We have been proactive in attempting to deal with potential difficulties in this project by meeting with agency leaders to sell the win-win features of partnerships to solve data sharing and project funding issues. We are also

- Highlighting the powerful features of the technology, showing impressive sample output, and telling about the potential for doing detailed forecasting of future needs and demographics.
- Providing reassurance regarding data security features being developed and the ability to either limit release of data when sample size gets small or publish only aggregate data.
- Planning to establish a steering committee and a technical advisory group to guide the project's activities and keep it accountable to all partners and the community via representatives on these committees.

In this manner, the leadership in SDH has been apprised of the full potential of this project and has been very supportive in its establishment. It is envisioned that PHS will continue to be a major provider of data and a major user of the system for program planning and health status reporting. The following examples show how PHS uses these data for enhanced program delivery, as well as to inform its program planning and evaluation:

1. **Low Income and Housing:** Figure 1 illustrates the relationship between poverty, housing density, and control. Areas with a higher proportion of low income tend to be more crowded and have a high proportion of rented housing. Higher income is associated with higher home ownership and less dense housing. These types of data are useful for lobbying efforts of community groups and for directing social and health programming to areas of greatest need. Within public health, our public health inspectors work with municipal fire and building inspectors to try to improve rental housing quality in the neediest areas of the city.
2. **Birth Rate and Income:** Figure 2 shows the relationship between higher teenage mother fertility and low income. This information can be used to direct further study about reasons for disparities between areas (e.g., less access or desire for abortion, less education about birth control, differences in sexual activity), and can ultimately be used to target areas in greater need of these services.
3. **Rates of Hepatitis A and Pertussis:** Figure 3 is a good example of the project's ability to combine or layer multiple data sources. Here we see that Hepatitis A rates are most closely correlated with the distribution of registered Indian

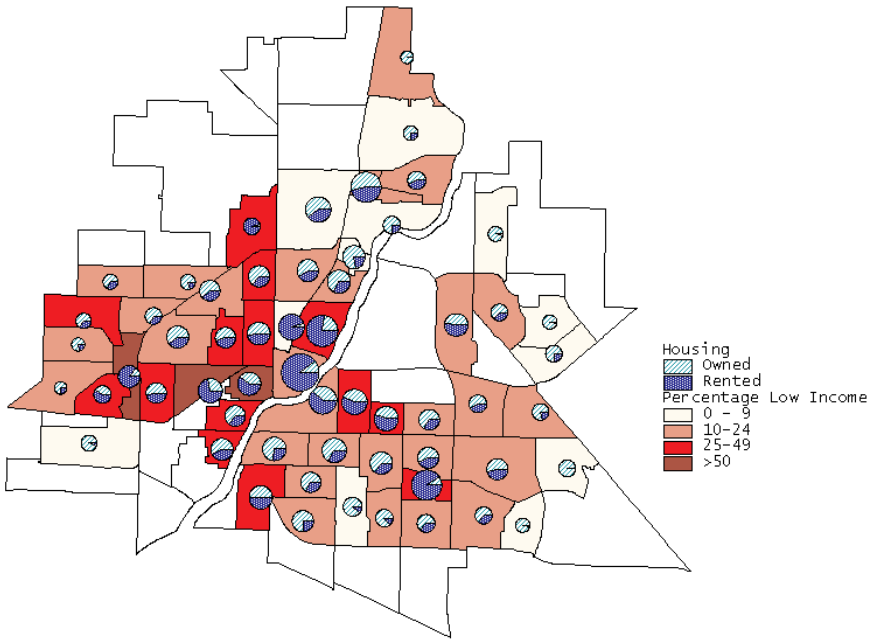


Figure 1 Percentage of low-income households and home ownership, by neighborhood; Saskatoon, 1991.

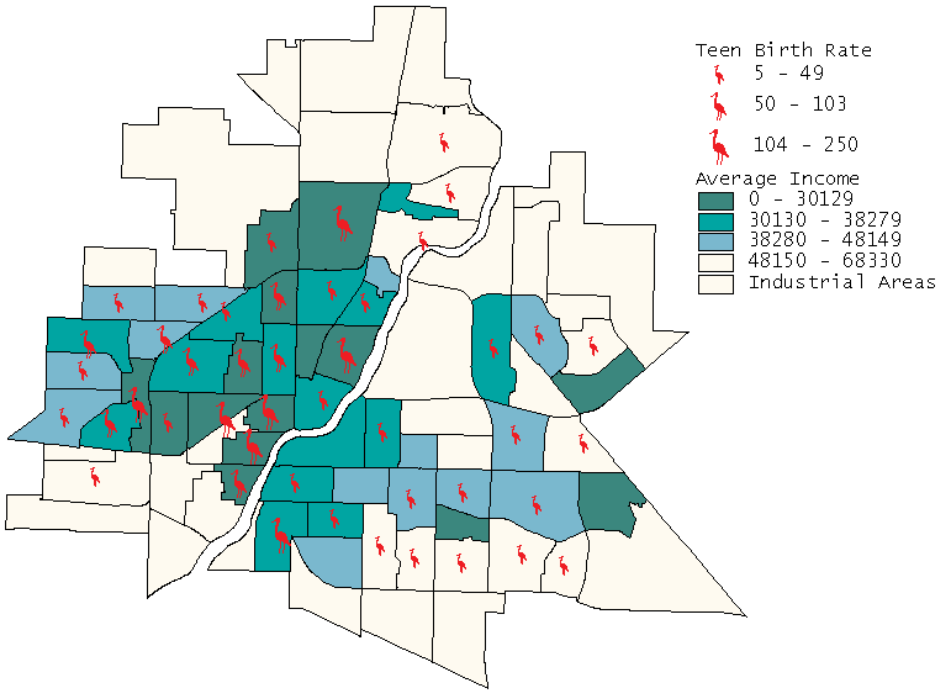


Figure 2 Birth rates for 15- to 19-year-old females, by neighborhood income; Saskatoon, 1996.

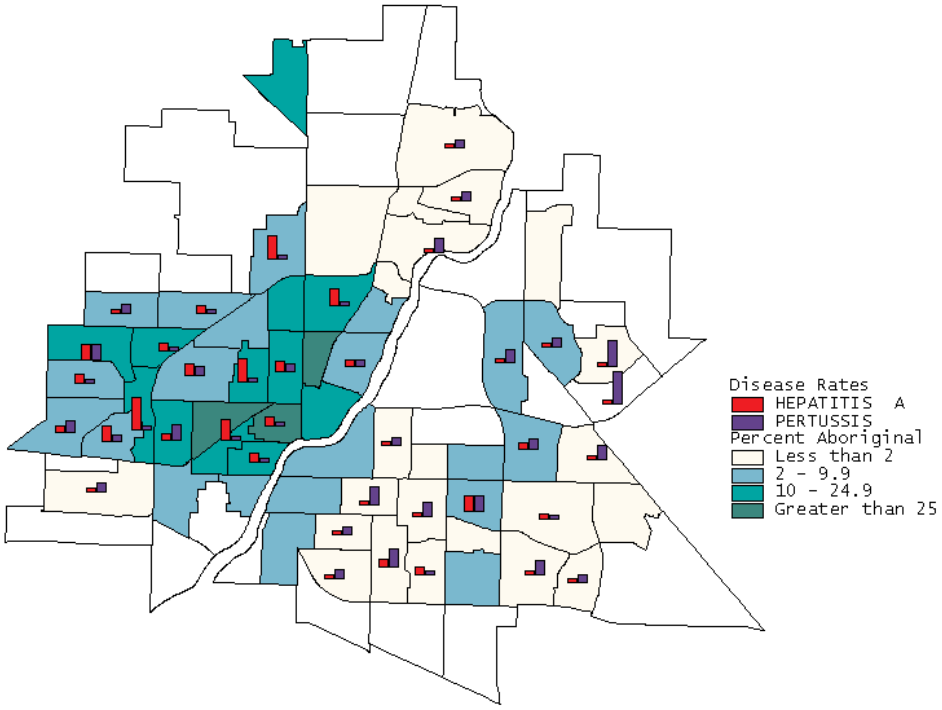


Figure 3 Rates of selected communicable diseases, by average neighborhood income; Saskatoon, 1997.

persons who are more mobile between the city and the reserves where higher Hepatitis A rates are seen. In contrast, Pertussis rates are more evenly distributed in the city. Soon, we will be incorporating street address level mapping to allow us to use this information more effectively for outbreak management.

4. **Cardiovascular Disease Morbidity and Mortality:** Figure 4 shows how hospital utilization data can be contrasted with vital statistics data for targeting prevention programs. Cardiovascular disease shows a certain distribution (after adjusting for age and sex distribution), but mortality rates due to cardiovascular disease show a slightly different pattern. One can use these data to target primary, secondary, and tertiary prevention programs as well as for further studies into what health determinants are causing the difference between mortality and morbidity rates by neighborhood (e.g., access to care, stress, exercise, diet, ethnicity).

Future Plans

It is envisioned that the CCHIS, in association with the Strategic Health Information and Planning Services Department, will maximize the potential for having an impact on regional health program planning and policy-making. Future plans for the project include

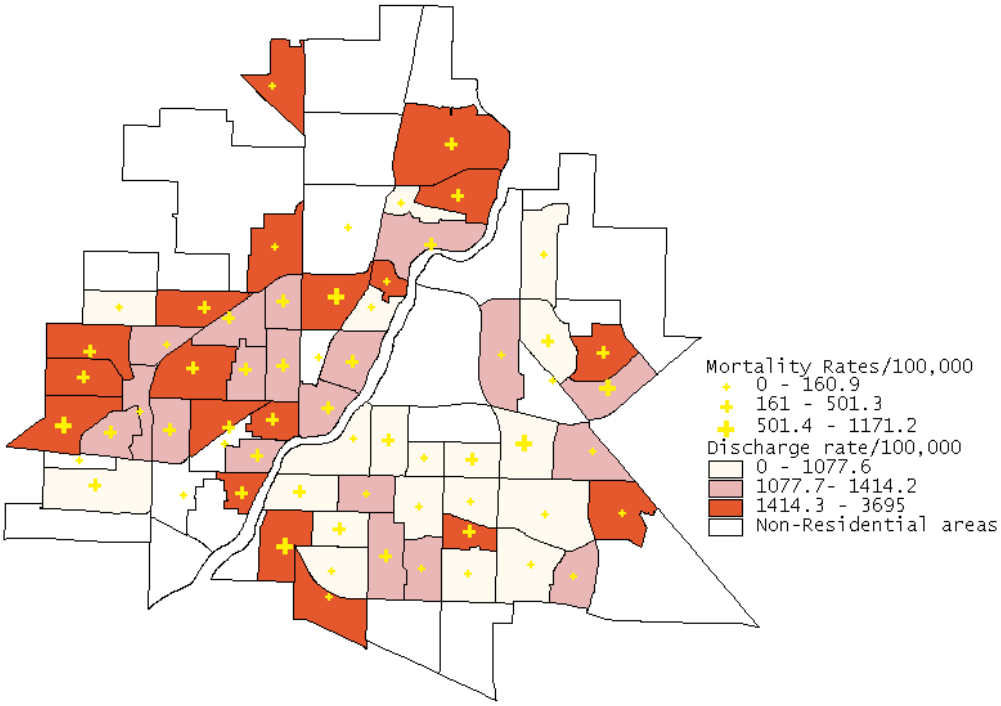


Figure 4 Cardiovascular disease, age/sex standardized rates, by neighborhood income; Saskatoon, 1996.

- Setting up a steering committee and a technical advisory committee.
- Incorporating a format to allow access to reports and data via the Internet.
- Linking databases in real time.
- Increasing data sources (including qualitative data).
- Linking with the SDH's District Client Information System (an electronic health ecord in development).
- Expanding GIS tools and resources, and making this accessible over the Internet and the SDH intranet.
- Training staff in the use of the tools and interpretation of the information.
- Encouraging use of the data and outputs at all levels of the organization and among our partners, including the community.

Summary

In summary, the CCHIS will provide a seamless dataset to a wide variety of users at various levels of aggregation. This relational data warehouse will be comprised of data from various agencies within the health sector as well as other sectors with information to share regarding health determinants. The ability to have access to this comprehensive collection of data to analyze and view using a GIS will greatly enhance health planning, evaluation, and research that will improve service delivery within the health

sector. Intersectoral planning will also be enhanced with data-sharing between various levels of government and those government sectors that have the greatest chance of making an improvement in the health of the population. Lessons learned from this project will be broadly transferable across Canada. Progress to date has been considerable, with much enthusiastic partner support.

References

1. Roos NP, Black C, Frohlich N, DeCoster C. 1996. Population health and health care use: An information system for policy makers. *Milbank Quarterly* 74(1):3–31.
2. Evans RG, Stoddart GL. 1990. Producing health, consuming health care. *Social Science and Medicine* 31:1347–63.

GIS for Community Health Planning: A Guide for Software Developers

Thomas B Richards, MD (1),* Charles M Croner, PhD (2), Carol K Brown, MS (3), Littleton Fowler, DDS (4)
(1) Public Health Practice Program Office, Centers for Disease Control and Prevention, Atlanta, GA; (2) National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD; (3) National Association of County and City Health Officials, Washington, DC; (4) Association of State and Territorial Local Health Liaison Officials, Washington, DC; Cleveland County Health Department, Norman, OK

Disclaimer: Use of trade names and commercial sources is for identification only and does not imply endorsement by the Public Health Service or by the US Department of Health and Human Services.

Abstract

Geographic information system (GIS) software products, data, and methods need to be developed to help local health departments and officials organize the process of community health assessment, identify preventable health problems, and improve public health programs and prevention effectiveness at the community level. We suggest that software developers explore the feasibility of forming private-public partnerships with innovative local health departments that have already started to apply GIS. In addition, we suggest focusing efforts on one (or a few) sentinel local public health issue(s), and developing modules that can be used separately, but that also can be nested together in a variety of different combinations, depending on a community's specific needs and priorities. The ultimate goal for local public health practice GIS product development is Web-enabled GIS with community-wide access, integrated with community planning tools such as *Assessment and Planning Excellence Through Community Partners for Health* and the *Guide to Community Preventive Services.*

Keywords: public health practice, community health planning, information systems, geography

Introduction

Geographic information system (GIS) technology can potentially offer important contributions to public health practice and management at local, state, and national levels (1–6). Software developers have started to ask, "What types of GIS software products and data methods would be useful in public health practice?"

The purpose of this paper is to help develop a dialogue on this topic by proposing types of GIS products that would be useful. In addition, this paper provides some general background about the public health marketplace for GIS products, models for organizing GIS within public health, and research challenges related to GIS software development.

* Thomas B Richards, MD, Centers for Disease Control and Prevention, 4770 Buford Highway, NE (K-55), Atlanta, GA 30341-3717 USA; (p) 770-488-3220; (f) 770-488-4639; E-mail: tbr1@cdc.gov

The Public Health Marketplace for GIS Products

Scant information exists about the current extent and types of GIS used by state and local public health agencies (7,8). Our general impression, however, is that GIS is still in its infancy in the context of public health management and practice. For example, a 1997 survey of state initiatives in geocoding vital statistics determined that only 21 of 49 responding state vital statistics registration bodies were involved in some type of automated geocoding of address data from vital records (8).

Public health practice typically involves multiple partners or collaborators. As a result, multiple public health marketplace niches, such as the following, exist:

- Federal agencies
- State health departments
- Large local health departments
- Small local health departments
- Health care organizations and providers

GIS specialty products likely will need to be developed for each of these. The focus in this paper, however, is on product development at the local level. From the perspective of community health planning, local health department (LHD) products are a logical starting point. Local level GIS offers the potential to incorporate information at the greatest level of detail and, if successful, might provide a building block for initiatives at other levels in the government hierarchy.

In addition to LHDs, a number of community health care organizations (e.g., hospitals and managed care organizations) may have considerable interest in population-based prevention programs at the local level. Thus, although the initial primary emphasis might be on developing GIS products for local health departments and officials, design features or products that have wider applicability (e.g., for use by groups such as hospitals and managed care organizations) would be beneficial.

Organizational Models for GIS in Local Public Health Practice

Four organizational models suggest how GIS might be incorporated into local public health practice:

- Model 1: Individual GIS user within a public health agency
- Model 2: GIS service unit for multiple GIS users within a public health agency
- Model 3: Enterprise-wide approach to GIS so that different programs within a public health agency can share GIS data
- Model 4: Web-enabled GIS with community-wide access

Model 1 is probably the most common at present. Models 3 and 4 are currently rare or do not exist within public health practice, but are likely to be perceived as a desirable goal by public health practitioners in the future.

Under Model 3, the LHD establishes priorities. Also, LHD spatial databases and automated systems are tailored to meet the established priorities, all while being shared among LHD programs (i.e., staff in one LHD program area are able to access spatial data in other LHD program areas in order to achieve the established priorities). Under Model 4, the LHD and its community partners join together to form a "community

enterprise" to improve public health performance, including a regional data warehouse. Shared data are on the Web, with different levels of access as needed to protect confidentiality of medical information. In addition, community groups are enabled to access and create their own maps after undergoing an educational program. Such a program would need to include discussion of potential problems in interpretation such as "lies with maps," the need to focus on comparisons where epidemiologists have already established etiologic relationships (9), and limitations in interpretation when rates are unstable because of small numbers.

GIS Research Challenges

At least eight research challenges will need to be met before the full power of GIS can be realized in community health planning:

1. Establish local public health agency enterprise-wide accessibility to local public health agency data (i.e., staff in one LHD program area are able to access spatial data in other LHD program areas in order to achieve the established priorities).
2. Establish local public health agency partnerships for integration and accessibility of georeferenced databases related to essential public health services, where the georeferenced data collected by the local public health agency can be used with georeferenced data collected by other government programs (e.g., planning, environment, or other municipal service departments) or other community health resources (e.g., hospitals, managed care organizations, and laboratories).
3. Build integrated linkage of GIS data, methods, and software with community planning tools (described below).
4. Develop local public health models for Web-enabled GIS systems with community-wide access—perhaps similar to the Community Health Mapping Engine (CHiME) Geographic Information Systems Project being developed by the Clackamas County Department of Health and Human Services, Oregon City, Oregon.[1] (see: *J Public Health Management Practice 1999*; 5(2):64–69).
5. Develop the capability to geocode, analyze, and make decisions using current georeferenced data (rather than data that are several years old).
6. Establish methods to preserve the privacy and confidentiality of medical information of individuals.
7. Document Federal Geographic Data Committee "metadata" (data about data) standards to facilitate exchange, interpretation, and analysis of public health GIS information (10).
8. Employ statistical and epidemiologic methods to GIS data related to disease surveillance and prevention decision-making by public health managers.

Community Planning Tools

GIS software, data, and methods need to be developed that build integrated linkages

[1] See Melnick A, Seigal N, Hildner J, Troxel T. 1999. Clackamas County Department of Human Services Community Health Mapping Engine (CHiME) Geographic Inormation Systems Project. *Journal of Public Health Management Practice* 5(2):64–69.

between GIS and community planning tools such as *Assessment and Planning Excellence Through Community Partners for Health* (*APEXCPH*) and the *Guide to Community Preventive Services* (11). *APEXCPH* is currently being developed by the National Association of County and City Health Officials, and builds upon the *Assessment Protocol for Excellence in Public Health* (*APEXPH*) (12). *APEXCPH* will emphasize the essential public health services (13), be available in electronic format, and explore the feasibility of incorporating GIS methods. The *Guide to Community Preventive Services* is currently being developed by a US Public Health Service Task Force and will provide evidence-based recommendations for preventive services and population-based interventions.

Building integrated linkage of GIS data, methods, and software to *APEXCPH* and the *Guide* (and to other community planning tools) provides a number of opportunities for GIS software development. The notion of linking GIS to *APEXPH* is not a new one. For example, in 1996, the Lewin Group proposed an *APEXPH*-related, GIS-based model (subsequently not fully evaluated) to aggregate data for community planning (14).

Although a single community planning tool might be the ultimate goal, given current funding constraints and the wide variety of topics in public health where research efforts might be focused, a reasonable research strategy for GIS software developers might be to focus initial efforts on developing a module for one (or a few) sentinel public health issue(s) where a small success can be demonstrated over a relatively short time period. Modules should be designed so that they can be used separately, but also so they can be nested together in a variety of different combinations, depending on the specific needs and priorities of a community.

Several examples of specific categorical program modules might include reducing the number of cases of vaccine-preventable diseases; preventing cardiovascular diseases; improving pregnancy outcomes and reducing infant mortality; preventing motor vehicle occupant injury and mortality; preventing childhood lead poisoning; and improving environmental health. Modules also could be developed for important (vertical) cross-cutting issues, such as training for beginning GIS users in LHDs.

We also suggest that GIS software developers explore the feasibility of forming private-public partnerships with innovative LHDs that have already started to apply GIS. The reasons for this are that software developers otherwise may experience considerable difficulty in obtaining access to databases to pilot test products, and insights from public health practitioners are needed to determine, for example, what constitutes a useful product and how results should be interpreted.

For those who want to learn more about GIS applications in the context of public health practice, a good source of information is the National Center for Health Statistics' free bimonthly e-mail report, *Public Health GIS News and Information.* To subscribe, send an e-mail to Dr. Charles Croner at cmc2@cdc.gov.

References

1. Croner CM, Sperling J, Broome FR. 1996. Geographic information systems (GIS): New perspectives in understanding human health and environmental relationships. *Statistics in Medicine* 15:1961–77.
2. Clarke KC, McLafferty SL, Tempalski BJ. 1996. On epidemiology and geographic information systems: A review and discussion of future directions. *Emerging Infectious Diseases* 2(2):85–92.

3. Vine MF, Degnan D, Hanchette C. 1997. Geographic information systems: Their use in environmental epidemiologic research. *Environmental Health Perspectives* 105(6):598–605.
4. Gordon A, Womersely J. 1997. The use of mapping in public health and planning health services. *Journal of Public Health Medicine* 19(2):139–47.
5. Office of Community Planning and Development. 1997. *Mapping your community: Using geographic information to strengthen community initiatives.* Washington, DC: US Department of Housing and Urban Development. HUD-1092-CPD.
6. Melnick A, Seigal N, Hildner J, Troxel T. 1999. Clackamas County Department of Human Services community health mapping engine (CHiME) geographic information system project. *Journal of Public Health Management and Practice* 5(2):51–7.
7. Warnecke L, Beattie J, Kollin C, Lyday W. 1998. *Geographic information technology in cities and counties: A nationwide assessment*. Washington, DC: American Forests.
8. MacDorman MF, Gay GA. 1999. State initiatives in geocoding vital statistics. *Journal of Public Health Management and Practice* 5(2):72–3.
9. Morgensten H. 1998. Ecologic studies. In: *Modern Epidemiology*. 2d ed. Ed. KJ Rothman, S Greenland. Philadelphia: Lippincott-Raven Publishers. 459–80.
10. Federal Geographic Data Committee. 1997. *Framework introduction and guide.* Washington, DC: Federal Geographic Data Committee.
11. Pappaionou M, Evans C. 1998. Development of the guide to community preventive services: A US public health service initiative. *Journal of Public Health Management Practice* 4(2): 48–54.
12. National Association of County and City Health Officials. 1991. *Assessment protocol for excellence in public health* (*APEXPH*). Washington, DC: National Association of County and City Health Officials.
13. Harrell JA, Baker EL. 1994. The essential services of public health. *Leadership in Public Health* 3(3):27–31.
14. Lewin Group, The. 1996. *Fulton County data aggregation system manual. Final report.* Contract #282-92-0041, Delivery Order #21. Fairfax, VA: The Lewin Group. October.

Nutrition Risk of Older Persons Participating in Home-Delivered and Congregate Meal Programs in Relationship to Demographics and Community Resources

Alice A Spangler, PhD, RD, FADA, CFCS*
Department of Family and Consumer Sciences, Ball State University, Muncie, IN

Abstract

Identifying problems that potentially compromise the nutritional status of the elderly can aid in decreasing the risk of disease and improving overall health. Several nutrition risk factors associated with health and well-being have been described by the national Nutrition Screening Initiative (NSI) and encapsulated in the Determine Your Nutritional Health (DETERMINE) Checklist. Working with the 16 planning and service areas in Indiana, we conducted a survey of all participants in congregate and home-delivered meal programs, using the NSI DETERMINE Checklist and questions about demographics and housing and living arrangements. The purpose of the initial project was to characterize the extent of potential nutrition risk among the senior citizen meal participants. Following the NSI's guidelines, meal participants were categorized as being at low, medium, or high nutrition risk potential. The percentage of those at high nutrition risk potential was computed for each county. US Census data for population demographics and community resources were used in the analysis. The maps resulting from analysis show locations of potential nutrition risks as determined by the survey data and their interaction with US Census data and other existing data. One practical application of geographic information systems in this project was the ability to strategically locate needed services for community-dwelling older Americans, who represent a composite of several ethnic groups and other groups that may have special and unique needs. Other outcomes of the project included direct impact on services provided to older persons, increased visibility of nutrition needs of older people (evidenced by media coverage), and heightened awareness by professionals regarding nutrition needs of older people.

Keywords: Nutrition Screening Initiative (NSI), DETERMINE, congregate meals, homebound, geriatric nutrition

Introduction and Purpose

Identifying problems that potentially compromise the nutritional status of the elderly can aid in decreasing their risk of disease and improving their overall health. One-third to one-half of elders' health problems are thought to be related to inadequate nutrient intake (1). Numerous factors associated with the older population can contribute to inadequate intake. Some of these include education (2,3), loneliness (4), bereavement (5), and being homebound (6,7).

* Alice Spangler, Department of Family and Consumer Sciences, Ball State University, Muncie, IN 47306 USA; (p) 765-285-1470; (f) 765-285-2314; E-mail: 00aaspangler@bsu.edu

Several nutrition risk factors associated with health and well-being have been described by the national Nutrition Screening Initiative (NSI) and encapsulated in the Determine Your Nutritional Health Checklist (a.k.a. DETERMINE) (8).

The purpose of the initial project was to characterize the extent of potential nutrition risk among the senior citizen meal program participants in Indiana and to provide education to these meal participants regarding nutrition and health. As the project unfolded, the decision was made to analyze the Indiana data using geographic information systems (GIS) because of the impact the visual map presentations would have, and because of the great potential for interfacing the nutrition survey data with US Census data and other existing data.

Materials and Methods

Data Collection and Methods of Analysis

The nutrition program coordinator of the Indiana Division on Aging and InHome Services, in conjunction with the author, a faculty member at Ball State University, conducted a survey of all participants in congregate and home-delivered meal programs in Indiana's 16 planning and service areas (PSAs). This survey used the NSI DETERMINE Checklist and a list of questions about demographics, housing and living arrangements, attitudes toward the meals provided, and zip code information. Registered dietitians, cooperative extension specialists, meal site managers, and meal delivery employees helped participants complete the surveys. Over 12,000 surveys were returned and coded for federal information processing standards (FIPS) for county data, then analyzed at Ball State University using SPSS-X (SPSS Inc, Chicago, IL) for preliminary statistical results and MGE (Intergraph Corporation, Huntsville, AL), Microsoft Excel, and Informix Online Database Engine (Informix Software, Inc, Menlo Park, CA). Excel was used to summarize data for each of the 92 counties; these data were then related to census and other data. Each county was treated as a record. Note should be made that, although use of the DETERMINE Checklist was not required at the time of this project, the National Aging Program Information System (NAPIS) (administered by the US Department of Health and Human Services' Administration on Aging) now requires that the Checklist be used to determine which clients receiving services of home-delivered meals, senior congregate meals, nutrition counseling, and case management are at high nutritional risk.

Following the NSI's guidelines, meal participants were categorized as being at low, medium, or high nutrition risk potential. The percentage of those at high nutrition risk potential was computed for each county. The 92 counties were divided into quartiles, based on the percentage of those at high nutrition risk potential. GIS was selected as the tool for analysis because it can point to geographic areas of concern, allow users to look at those areas within the context of the counties and the PSAs in which they exist, and interrelate the nutrition risk data with other demographic, health, and community resource data.

Use of Existing Data

1990 US Census data for population demographics, sampled US census data (taken from the US Census Bureau's "USA Counties 1994" CD-ROM) for health resource

information, and community meal site directory information were analyzed to characterize (at the individual-county level) the socioeconomic environment and selected resources that would potentially impact the overall well-being of the study participants. Examples of existing data categories used in the analysis include age distribution of older people, minority populations, poverty levels of older people, dependency ratios (65+/19–64 years of age), and physician and hospital availability within the counties. For each county, the number of senior meal sites was related to the number of older people within the county.

Consideration of Other Resources Available

In pursuing GIS analysis, already available campus resources at Ball State University (Muncie, IN) associated with GIS were sought and utilized. Computers and software were available, as well as faculty and staff knowledgeable about GIS. Especially critical to the success of this project was the University's Computing Services Graphics Systems Administrator. This individual has a landscape architecture background and has provided extensive knowledge to the technical aspect of using GIS; he also provided instruction and assistance as progress was made through the stages of this project.

Results

The results were summarized for the entire state and for each of the 16 PSAs, each of which included two to nine counties (Figure 1). Within the entire state, the study

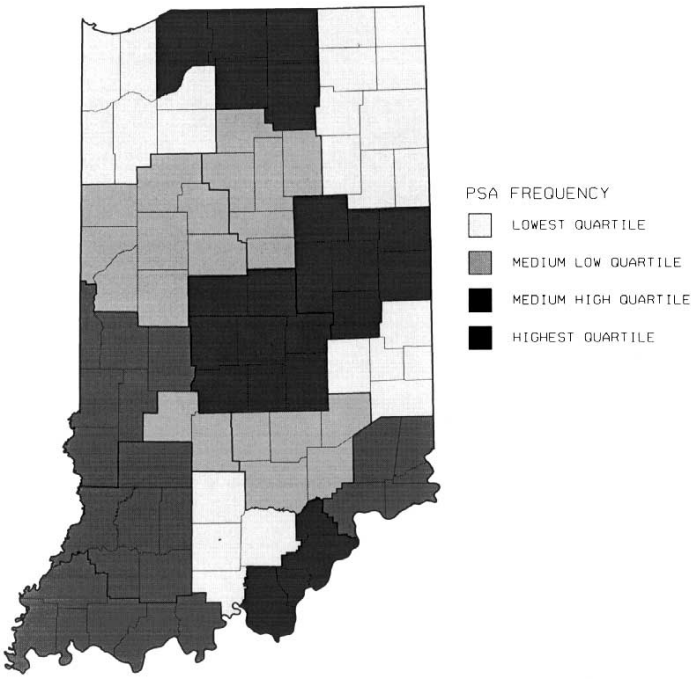


Figure 1 Distribution of high nutrition risk potential among meal program participants in Indiana's planning and service areas, 1993.

revealed that 34% of the 12,062 study participants were considered to be at high nutrition risk potential, based on the DETERMINE Checklist. Of the homebound, 52% were at high nutrition risk potential; 24% of the congregate meal group were at high risk potential. Composition of the study participants included 70% females; average age was 78, with a range from 60 to 106 years of age; 85% of the study participants were white; 8% were African-American.

The maps resulting from analysis showed in-depth geographic locations of potential nutrition risk, possible interactiveness with other data such as census demographics, meal site ratios, and possible relationships and patterns among contiguous or proximate counties. Figures 2 through 4 illustrate distribution of high nutrition risk for all participants (Figure 2), homebound meal participants only (Figure 3), and congregate meal participants only (Figure 4). Counties within a PSA vary in risk potential. Metropolitan counties tend to be in the highest or medium-high risk quartile. However, attention needs to be given also to the numerous non-metropolitan counties with high nutrition risk. The congregate program distributes services in a centralized community setting, and older people may need transportation to the site. Homebound older people need to have services delivered to the home, and lengthy travel through a remotely populated county to deliver a meal becomes a challenge to limited resources.

Figures 5 through 7 illustrate pertinent demographic distribution based on census data. In Figure 5, which represents proportions of old-old (75+ years of age)[1] residents,

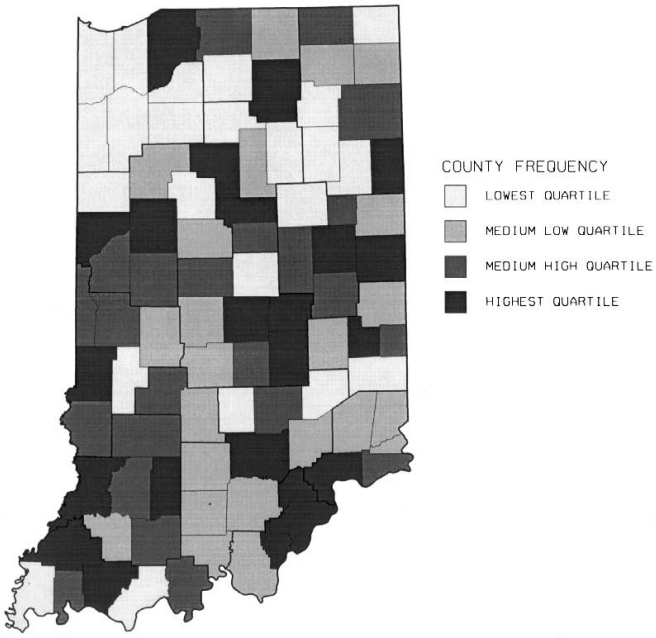


Figure 2 Distribution of high nutrition risk potential among all meal program participants, Indiana, 1993.

[1] "Old-old" is a term often used in the gerontology literature. (Generally, "young-old" describes people 60–75 or 65–75.) The old-old category reflects the probable increase in frailty and decline in functional status experienced in people over 75.

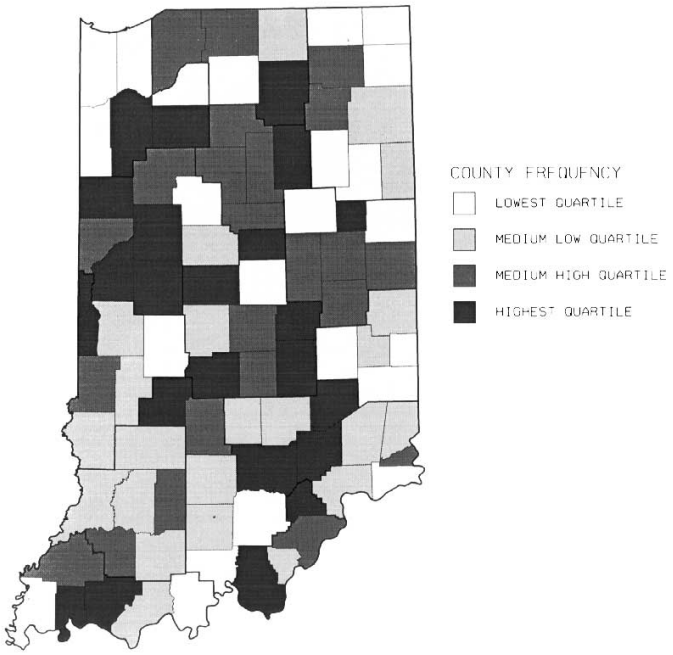


Figure 3 Distribution of high nutrition risk potential among homebound meal program participants, Indiana, 1993.

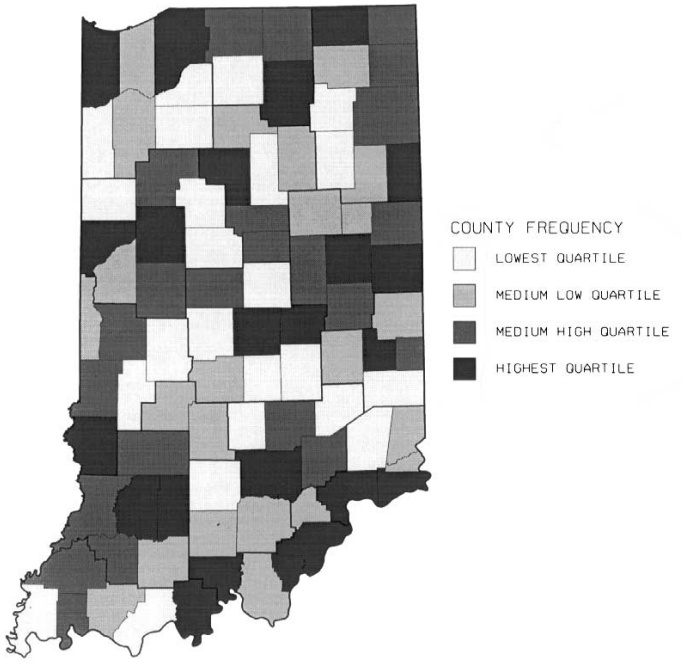


Figure 4 Distribution of high nutrition risk potential among congregate meal program participants, Indiana, 1993.

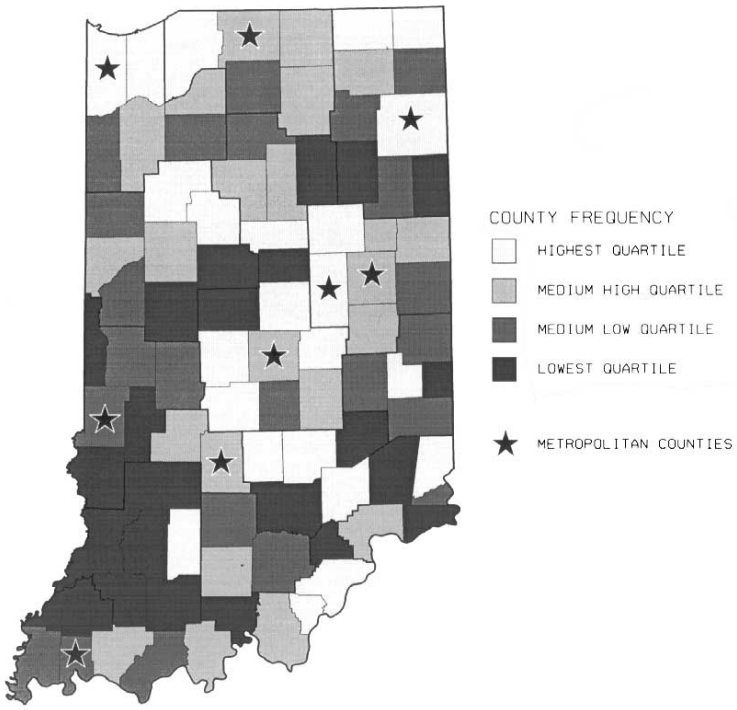


Figure 5 Quartile distribution of counties' old-old (75+) population, Indiana, 1993.

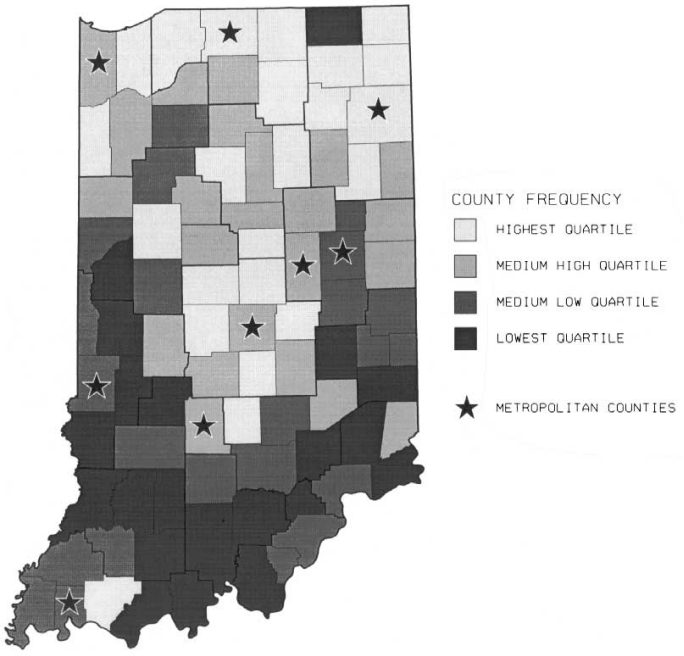


Figure 6 Distribution of counties' older population at or below poverty, Indiana, 1993.

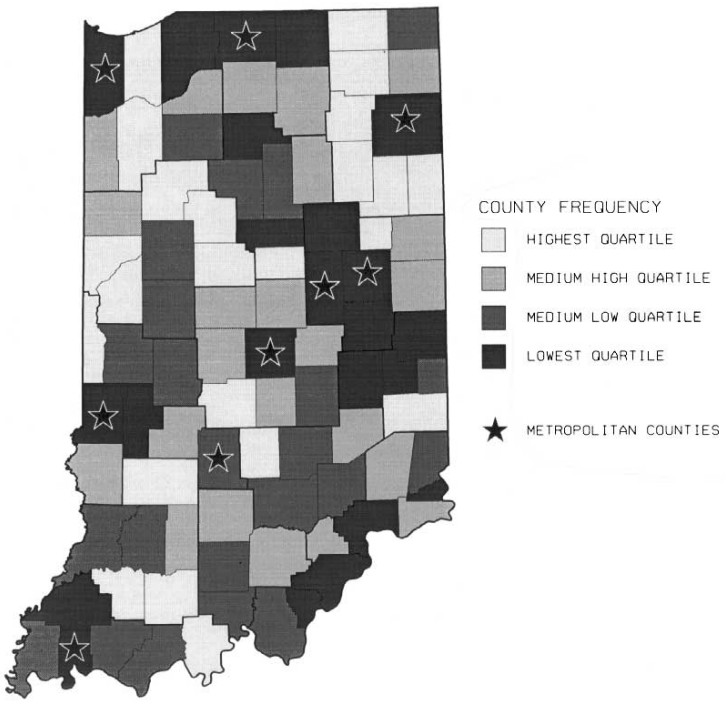


Figure 7 Distribution of counties' older African-American population, Indiana, 1993.

counties in the western and southern parts of the state have the highest proportions of those who are old-old. The metropolitan counties (starred) have a lesser proportion of old-old residents. Analysis of poverty among older persons (Figure 6) reveals that distribution of percentages of older persons at or below poverty income levels is similar to distributions of percentages of old-old residents; that is, the western and southern parts of the state have the highest percentages of older people at or below poverty income levels. Census data also provided information on locations of high percentages of various minority groups. Metropolitan counties have the highest distribution of older African-Americans (Figure 7).

Figure 8 represents the distribution of the dependency ratio (65+/19–64 year olds) and has many implications. We need to think about who is or will be taking care of older people. This can be considered at the family level as well as the service level. When we consider the demographics of aging, this ratio also has great potential in examining future trends.

The map in Figure 9 represents an index of the number of senior citizen meal sites per 1,000 older persons within each county. It is possible to identify counties that are in the highest quartile for the proportion of old-old population, but have a low number of meal sites in relation to the number of older persons. Figure 10 shows counties with the lowest meal site index and highest nutrition risk, along with counties with the highest meal site index and lowest nutrition risk. GIS can be used to identify uneven distribution of risk and allocated resources within each PSA.

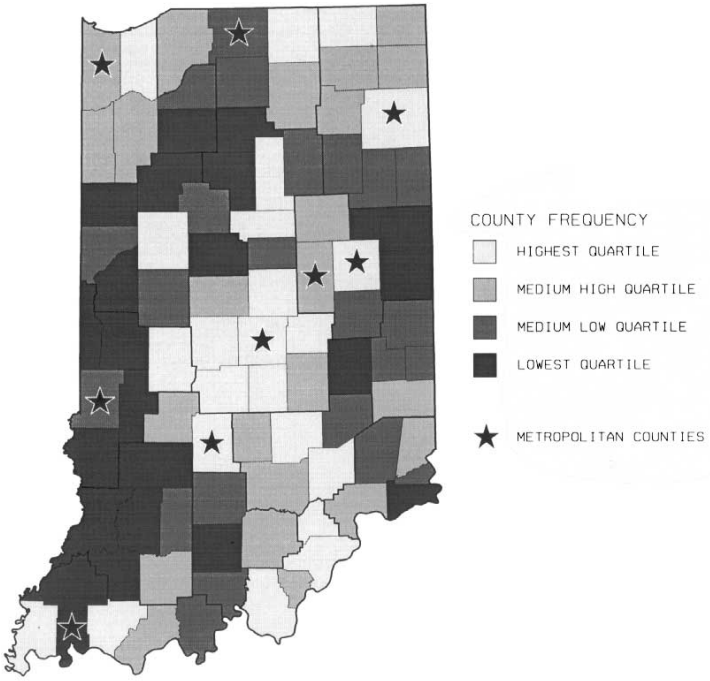


Figure 8 Distribution of counties' dependency ratio (65+/19–64), Indiana, 1993.

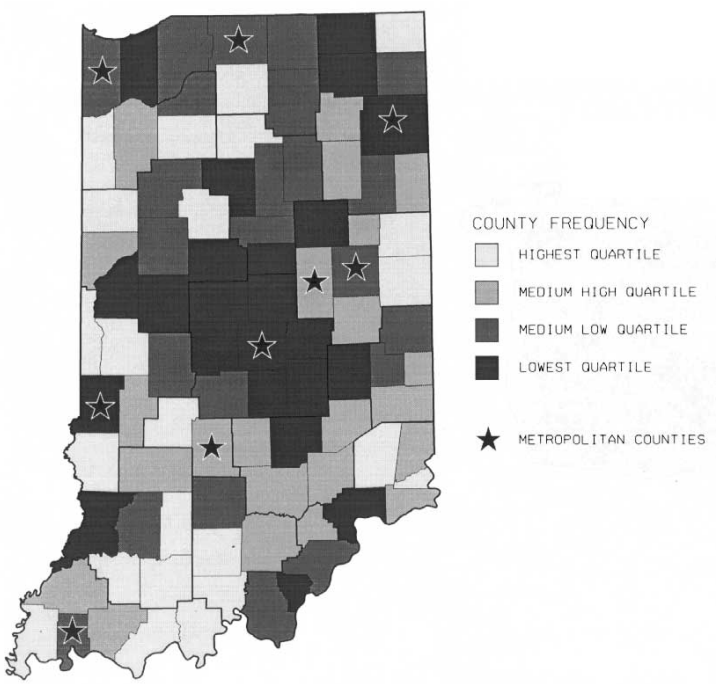


Figure 9 Distribution of meal sites per 1,000 older persons, Indiana, 1993.

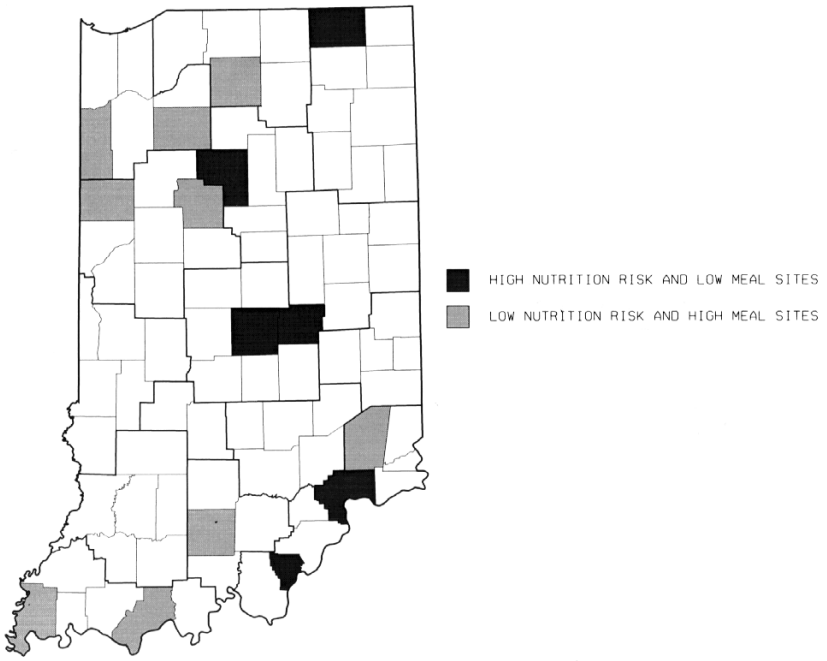


Figure 10 Example of GIS analysis revealing spatial relationship between counties with high nutrition risk/low meal sites and with low nutrition risk/high meal sites, Indiana, 1993.

Discussion and Conclusions

The limitations of the study included the 67% response rate for the entire state, with counties varying in response rate. Because of this, comparison of one county to another must be done with caution. The researcher's goal was to look at geographic patterns and clusters of counties in regard to nutrition risk potential, all within the context of other known features of the counties, such as whether the counties were metropolitan or non-metropolitan.

In addition, the surveys were completed by the older meal participants, with assistance from others if necessary. The survey was relatively simple, but relied on the respondent to accurately read and understand the questions. The DETERMINE Checklist has 10 questions that reflect behaviors or conditions associated with *potential* risk of poor nutrition status. More in-depth nutrition screening would be required to describe nutrition status more accurately. The advantage of using the DETERMINE Checklist in this study was that the checklist is used extensively throughout the United States, and health community agencies and nutrition professionals are familiar with the use and interpretation of the data. In addition, because the DETERMINE Checklist is a simple tool, intentionally developed using large print, and worded for ease of understanding by older persons, the collection of the data from a large number of individuals was possible, even with limited resources.

One time-consuming element of the project was the initial data entry into the database. This task consumed approximately a year, which lessened the timeliness of the information that was released.

Practical application of the GIS analysis in this study has included the ability to (1) strategically locate needed services for community-dwelling older Americans, who represent a composite of several ethnic groups and other groups that may have special and unique needs; (2) improve health and well being; (3) determine the need for education or other intervention; and (4) influence policy makers and decision makers who make decisions related to older Americans and their families.

Because of the extensive dissemination of its results, the project had a direct impact on services provided to older persons. For example: Programs for older people were expanded in the cooperative extension program in northeast Indiana. An additional assistant was hired to provide education and develop programs for older people. A steering committee in east central Indiana sought various types of data, including data from this project, to use in establishing guidelines and policies for future needs of senior meal programs; this steering group created, among other things, a partnership with a food bank and an innovative program to help older people in a rural remote community obtain nutritious meals at a local restaurant rather than a congregate meal site.

The data collection instrument itself provided a form of nutrition education and heightened awareness of factors that can lead to nutrition risk problems. The project also increased the visibility of older people's nutrition needs—evidenced by television, radio, and newspaper coverage—and heightened awareness by professionals of older people's nutrition needs.

Future plans for this application include further analysis of the data, using GIS, with an emphasis on examining the needs of older African-Americans in Indiana. Other goals are to automate data gathering of the DETERMINE screening tool through NAPIS and to readily utilize new census data as they become available.

References

1. Ryan VC, Bower ME. 1989. Relationship of socioeconomic status and living arrangements to nutritional intake of the older person. *Journal of the American Dietetic Association* 89:1805–7.
2. Holcomb CA. 1995. Positive influence of age and education on food consumption and nutrient intakes of older women living alone. *Journal of the American Dietetic Association* 95:1381–6.
3. Posner BM, Jette A, Smigelski C, Miller D, Mitchell P. 1994. Nutritional risk in New England elders. *Journal of Gerontology* 49:M123–32.
4. Walker D, Beauchene RE. 1991. The relationship of loneliness, social isolation, and physical health to dietary adequacy of independently living elderly. *Journal of the American Dietetic Association* 91:300–4.
5. Rosenbloom CA, Whittington FJ. 1993. The effects of bereavement on eating behaviors and nutrient intakes in elderly widowed persons. *Journal of Gerontology* 48:S223–9.
6. Gloth FM, Jordan DT, Smith CE, Meyer JN. 1996. Nutrient intakes in a frail homebound elderly population in the community vs. a nursing home population. *Journal of the American Dietetic Association* 96:605–7.
7. Coulston AM, Craig L, Voss AC. 1996. Meals-on-wheels applicants are a population at risk for poor nutritional status. *Journal of the American Dietetic Association* 96:570–3.
8. Nutrition Screening Initiative. 1991. *Nutrition screening manual for professionals caring for older Americans.* Washington, DC: Nutrition Screening Initiative.

Plotting Rural Households Where Map Details Are Insufficient: The Use of GPS in the Keokuk County Rural Health Study

ER Svendsen,* SJ Reynolds, C Zwerling, LF Burmeister, AM Stromquist, CD Taylor, JA Merchant

Keokuk County Rural Health Study, Department of Occupational and Environmental Health, University of Iowa, College of Public Health, Iowa City, IA

Abstract

The Keokuk County (Iowa) Rural Health Study (KCRHS) is a CDC-NIOSH-funded population-based, prospective cohort study, enrolling over one-fifth of the entire county population. Respiratory health and injury prevention in relation to environmental and occupational exposures are the primary focuses of this study. Health care delivery, geriatric, reproductive, and mental health are also measured. Because geographical distributions of health conditions within the study population are considered, a global positioning system (GPS) receiver has been used to geocode all rural households. The three categories of research questions that have been investigated with geographical information systems (GIS) are health care delivery, injury prevention, and health status. Within these, health care delivery questions measured the time and distance to participants' primary health care facility. Injury prevention measured crude injury rates, risk-taking behaviors, time, and distance to utilized emergency facilities. Health status measured hallmark health indicators such as tobacco use, drug and alcohol abuse, depression, and obesity. Demographic data have been collected and include age, sex, marital status, and socioeconomic status. Both crude and adjusted distributions have been performed. Medical screening and adult interviews were used as a source of GIS data. Continuous spatial distributions of the variables implicated within the study questions have been plotted in layers. Preliminary results indicate that alcohol consumption and abuse are uniformly distributed throughout the county. However, this is not the case for obesity, smoking, and reported injuries. These three seem to be clustered predominantly in the southwestern portion of the county. Further analysis is pending on the significance of these findings, and the health care delivery/injury prevention data. Through GIS analysis of the KCRHS data, the utility of GPS geocoding in rural community health and surveillance studies has been demonstrated.

Keywords: GPS, rural, Iowa

Introduction

Address matching of rural communities throughout the United States is very difficult, at best. Many addresses consist of box numbers or addresses that do not correlate with TIGER files. As a result, many rural geographic information system (GIS) applications are limited to aggregated data analyses. These methods are usually sufficient for

* Erik R Svendsen, Dept. of Occupational and Environmental Health, University of Iowa, College of Public Health, 100 Oakdale Campus, 241 IREH, Iowa City, IA 52242-5000 USA; (p) 319-335-4538; (f) 319-335-4225; E-mail: erik-svendsen@uiowa.edu

studies involving large geographic areas. For small areas such as counties, however, aggregated methods do not provide the resolution needed for small-cluster analysis. The Keokuk County (Iowa) Rural Health Study (KCRHS) has addressed this issue.

The KCRHS is a population-based, longitudinal prospective cohort study funded by the Centers for Disease Control and the National Institute for Occupational Safety and Health. The study has enrolled over one-fifth of the entire county population. Respiratory health and injury prevention are the primary focuses of this study. Health care delivery, and geriatric, reproductive, and mental health are also measured.

Method

In late 1997, the KCRHS began collecting global positioning system (GPS) coordinate data during its scheduled household site visits. This protocol was added to enable GIS analysis of the self-reported injury data collected in the study. Following its introduction, analysis of health care delivery and community health status data while controlling for confounding demographic variables including age, sex, marital status, and socioeconomic status was proposed. A nested GIS study had begun to emerge.

Within the three proposed GIS study areas of health care delivery, injury prevention, and health status, health care delivery questions measure the time and distance to participants' primary health care facility. Injury prevention questions measure crude injury rates, risk-taking behaviors, and time/distance to utilized emergency facilities. Health status measures hallmark health indicators such as tobacco use, drug and alcohol abuse, depression, and obesity. Medical screening and adult interview instruments used in the KCRHS contain all of these data.

The first round of data collection was completed in February 1998. As of this writing, only a small fraction of the homes within the study have been geocoded. The uncoded homes were scheduled for geocoding by GPS during the second round of data collection set for a two-year duration beginning September 1998. Due to the novelty of using GPS in geocoding rural households for GIS analysis on the county level, a preliminary study was undertaken for the purpose of presenting the technique at the third GIS in Public Health Conference in August 1998.

The preliminary study design incorporated a variety of plotting techniques to supplement the lack of GPS data. First, all households that were coded as residing within a town were geocoded within their town. Of the 454 town households, 326 were address matched. One hundred forty-nine (149) matches were performed directly with the Maptitude 4.03 GIS (Geonomics, Inc., Boston, MA) and 177 required manual matching assistance within the GIS. The remaining 128 were geocoded to the centroid of the town by using the locate-by-town function in Maptitude. Because the zip codes were not centered about a single town, plotting town residences at the zip code centroid was inadequate for this exercise. Rural and farm homes were not successfully address matched this way. Instead, GPS coordinates were taken by supplemental site visits for 440 households. These measures were taken at the mailbox for all measurements so that privacy was maintained. Fifty-eight (58) additional homes were plotted manually using the locate-by-pointing tool in Maptitude based on the known location. This was done using a locally produced basemap of the county homes and plats as a guide. Plotting homes manually may also have been performed by comparison with digital orthophotos now available nationwide at Microsoft's collaborative Web site, TerraServer

(http://terraserver.microsoft.com). Together, 498 (90%) of the 553 rural/farm households were plotted by GPS. When participants are considered rather than households, 904 (92%) of 987 rural/farm participants were plotted. In total, 1,563 (95%) of the 1,646 participants were plotted.

Accuracy of geocoded data is always a concern, especially when the study intends to perform continuous distribution analyses. It was approximated that the variation between the GPS recorded data and the actual household location data was within a quarter of a mile because no driveway exceeded that value. Data replication was not performed, so precision was not measured. GPS data were not updated by spatial correction factors, so the programmed error was still present. The Magellan 2000 GPS has an inherent error of within 50–100 yards, according to the manufacturer. Though not yet validated, it was assumed that the locate-by-pointing method was within the same quarter-mile accuracy constraints. Most towns were much less than a half-mile in diameter. Therefore, the majority of the town box households that were coded to the centroid are assumed to be well within the same quarter-mile accuracy range. The significance of these accuracy measures has not yet been evaluated.

Results

Once geocoded, the geographic data were analyzed using Distance Mapping and Analysis Program (DMAP) software (freeware available at http://www.uiowa.edu/~geog/health/index11.html). Obesity rates were presented for demonstration. A two-mile spatial filter was used to calculate the rates of obesity and the significance of the observed distribution. The denominator used was the observed county probability (0.42) of being obese within positive respondents. Preliminary results showed statistically significant clusters of obesity in the western portion of the county. Bands were overlaid to exhibit the two-mile spatial filter region of significance detected about the grid points. These were raw data that did not control for any confounding variables such as socioeconomic status. The intent of the demonstrated analysis was the illustration of the proposed GIS analysis of the final adjusted GPS database.

Conclusion

This initial study was not intended to produce rates or graphs for peer review publication. Rather, this study was intended to introduce the techniques available for plotting rural homes when address-matching capabilities are not available. Once a corrected and complete dataset of GPS coordinates is collected by this study in the next two years, further GIS epidemiologic studies will follow.

Preliminary results indicate that alcohol consumption and abuse are uniformly distributed throughout the county. This is not the case, however, for obesity, smoking, and reported injuries. Further analysis of the obesity data using DMAP statistical testing software has provided three regions of highly significant increased obesity rates. These three regions seem to be clustered predominantly in the western portion of the county. Through GIS analysis of the KCRHS data, the utility of GPS geocoding in rural community health and surveillance studies has been demonstrated.

Acknowledgments

G Rushton and the co-authors of the "GIS in Public Health" CD, The University of Iowa, Department of Geography

An EPA Region 2 GIS Application for Identifying Environmental Justice Areas

Daisy SY Tang, MA,* Linda Timander, MA
US Environmental Protection Agency, Region 2, New York, NY

Abstract

The US Environmental Protection Agency's (EPA's) Region 2 office (New York, New Jersey, Puerto Rico, and the US Virgin Islands), has developed a desktop geographic information system (GIS) tool for evaluating environmental justice concerns in a variety of regulatory decisions. The tool was developed according to requirements laid out in a draft regional policy that defines how demographic and environmental data should be used by EPA staff in evaluating environmental justice concerns. The policy's decision criteria define a community as an environmental justice area if (a) minority and/or low-income populations are affected significantly more than those populations in the reference areas, and (b) there is a disproportionate environmental burden on the area compared with the reference areas. The application provides three ways for the analyst to define the boundary of a community and select census block groups within the boundary for detailed examination of demographic characteristics. Boundaries can be predefined, user defined, or created by buffering around selected features. Once the boundary is defined, the percentage of minorities in the area and the percentage of population below poverty level are calculated for the block groups within the boundary, and these values are compared with values for the state and the county. If the relative difference between the community percentages and those of the state/county is greater than 25%, the community boundary is saved as a separate data layer for further comparison with other communities and for analysis of the environmental burden to determine whether it is an environmental justice area. There is no limit on adding other data layers that pertain to the analysis. Health and other available environmental data can be integrated into the application for analysis of correlation between demographic characteristics of communities, community health, and environmental exposure. This application will be widely used within the region as environmental justice concerns become integrated into the daily work of the regional employees.

Keywords: environmental justice, community of concern, relative difference, percent minority, percent population below poverty level

Introduction

Environmental justice is an issue that is of growing importance to the Clinton Administration and to the public. On February 11, 1994, the White House issued Executive Order 12898 on Federal Actions to address environmental justice in minority and low-income populations (1). The order is aimed to "focus federal attention on the environmental and human health conditions in minority communities and low-income communities with the goal of achieving environmental justice." To make achieving

* Suk Yee D Tang, US Environmental Protection Agency, Region 2, 79-05 Calamus Ave., Elmhurst, New York, NY 11373 USA; (p) 212-637-3592; (f) 212-637-4943; E-mail: tang.suk-yee-d@epamail.epa.gov

environmental justice part of its mission at the US Environmental Protection Agency (EPA) Region 2, the office established an Environmental Justice Workgroup on January 12, 1993. The workgroup was charged with providing advice and counsel on justice issues to regional management, and monitoring the progress of the region in achieving the agency's environmental justice goals. The goal of environmental justice is to identify and address unfairness and inconsistency in environmental matters. To address these, EPA Region 2 has developed a draft "Interim Policy on Identifying Environmental Justice Areas" (the "interim policy") (2).

EPA Region 2 Interim Policy

The EPA Region 2 interim policy defines terms, summarizes the steps that are to be taken in preparing for an environmental justice determination, and specifies the decision criteria that are to be used in making the actual determinations. Once an area is determined to be an environmental justice area, subsequent agency actions would be in accordance with established laws, regulations, and policies.

The process described in the policy for determining whether a specific area is subject to the agency's environmental justice program involves five steps:

1. Define the community of concern (COC).
2. Define the reference area.
3. Define the environmental burden.
4. Evaluate the demographic and burden data for the COC and reference areas.
5. Apply the decision criteria to the COC and reference areas.

The five steps center on the comparison of three factors between a COC and one or more reference areas: their respective levels of minority representation, low-income representation, and environmental burden. The screening process determines if a COC is a potential environmental justice area.

For environmental justice purposes, a COC is defined as a low-income community if the percentage of household incomes beneath the poverty level ("percent below poverty level") is significantly greater (25% or more) than in the reference area. A COC is considered a minority community if its percentage of minority residents ("percent minority") is significantly greater (25% or more) than in the reference area.

There is a two-tiered analysis used to identify environmental justice areas:

1. Screening analysis to identify potential environmental justice areas that warrant further study
2. Site-specific analysis to address environmental justice concerns

Environmental justice screening analyses are based primarily on the consideration of demographic data, and focus less on the determination of disproportionate burden. Screening analyses address the demographic characteristics of geographical units within the study area, such as census blocks or block groups, municipalities, or counties. The focus of a screening analysis is on comparison of the demographic characteristics of a discrete geographical community (the COC) with those of a reference area that encompasses the COC.

Site-specific analyses will necessarily be more in-depth than screening analyses because a number of potentially difficult determinations must be made along the way.

First, specific reference areas must be selected, their boundaries delineated, and their demographic data collected. Then, a site-specific analysis requires a detailed analysis of the environmental burden in the COC and reference communities in order to determine whether the burden is disproportionate in the COC.

Overview of the Region 2 Environmental Justice GIS Application

The environmental justice GIS application was developed to support the EPA Region 2 interim policy. In the first stage of development, only the demographic characteristics are incorporated into the application for preliminary screening of potential environmental justice areas. The analysis of disproportionate environmental burden is more complex and will be included in the second stage of application development.

This application enables the analyst to screen for potential environmental justice areas by first comparing the demographic data of the COC with that of the state and county, and then to smaller reference areas.

Two levels of analysis are involved in the screening process: a general screening analysis and a site-specific analysis. To conduct the general screening, first the boundaries of the COC must be defined. Once the boundary of a COC is defined by the user, the application selects the census block groups within the boundary for the evaluation of its demographic characteristics. The COC's percent minority and percent below poverty level are calculated and compared with those of the state and county. If the relative difference between the COC's values and the state/county values is greater than 25%, the COC is then considered a potential environmental justice area and is targeted for more detailed analysis.

In the site-specific analysis, the COC is compared with a smaller reference community. The reference community should be sufficiently close and/or comparable to the COC so that it would be reasonable to assume the presence of similar circumstances if environmental justice were not a factor. Once the boundary of the reference community is defined, the census block groups within the boundary are selected for calculation of percent minority and percent below poverty level. Similar to the general screening, the relative difference between the COC values and the reference community values is evaluated. If the relative difference is more than 25% compared with both state/county and reference areas, the COC is defined as a potential environmental justice area and the COC boundary is saved for further analysis on environmental burden.

The Application

The application starts with index maps of EPA's Region 2, which includes New York, New Jersey, Puerto Rico, and the Virgin Islands. The index map displays the counties in each state. The analyst can choose a county of interest using the "hot link tool" or the "Counties" pull-down menu. A view of the selected county will be created on the fly (Figure 1).

Data layers included in the county view are percent minority, percent below poverty level, municipal boundaries, and county boundaries. All analysis will be done in the county view (Figure 2).

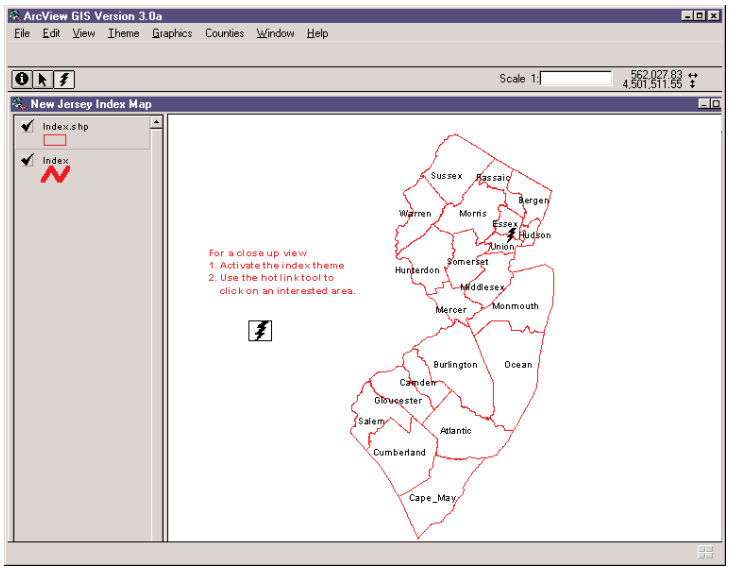


Figure 1 The New Jersey index map.

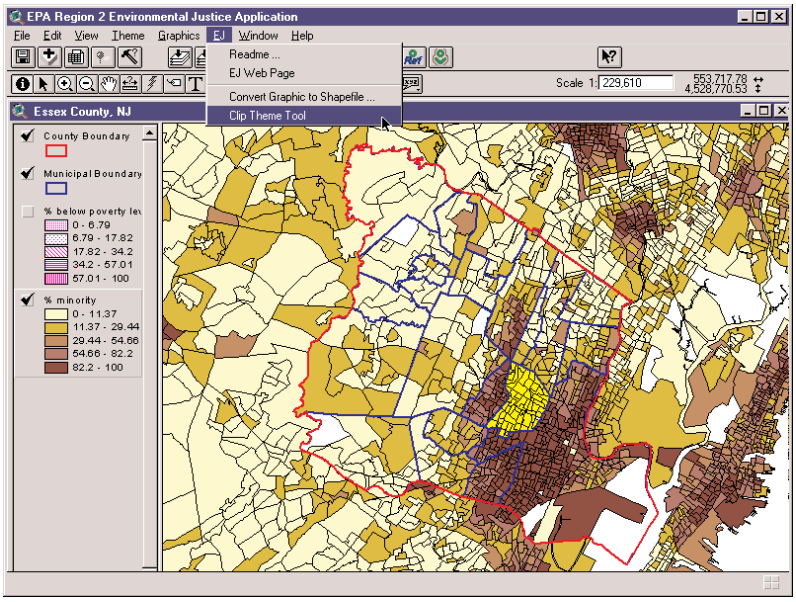


Figure 2 The county view.

The Screening Procedures

1. Define the boundary of a COC.
2. Select the census block groups within the boundary and convert into a new theme.
3. Calculate the percent minority and percent below poverty level for the COC,

and compare the relative difference between the COC's values and those of the state/county.

Defining the Boundary of a Community of Concern

There are three ways to enable the analyst to define the boundary. They are as follows:

- Using the customized "S" button to select a municipality (Figure 3).
- Using ArcView's "Graphic Tool" to draw a user-defined boundary (Figure 4).
- Using the customized "Buffer Tool" to create a buffer as the boundary (Figure 5).

The "S" button can be used to select the census block groups within a pre-defined boundary. The "S" button performs a theme-on-theme selection and for this application, the two themes used are municipality boundary and percent minority. This can be modified to be used on any pre-defined boundary.

The "Graphic Tool," when used together with the "Convert Graphic to Shapefile" and "Clip Theme Tool," will provide the flexibility to create any boundaries defined by users (Figure 6). The user can simply draw the boundary of the COC on the view with the "Graphic Tool," then convert the graphic into a theme with "Convert Graphic to Shapefile." Once a theme of the boundary is created, "Clip Theme Tool" is used to select the census block groups that fall within the boundary. Some block groups may be clipped by the COC boundary so that a portion of the block group lies within the COC boundary and a portion lies outside of the boundary. In this case, the numerical values

Figure 3 The "S" (select) button.

Figure 4 The Graphic Tool.

Figure 5 The Buffer Tool.

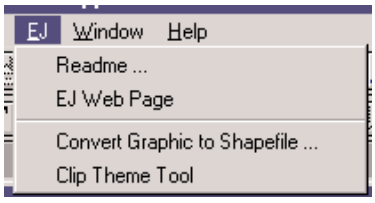


Figure 6 "Convert to Graphic" and "Clip Theme Tool" menu.

in the attribute table for block groups are updated based on the percentage of the block group area that falls within the COC boundary.

The "Buffer Tool" provides a way for users to find out the demographic characteristics around a facility. First, a buffer of one or more facilities is created and, similar to the "Graphic Tool" method, the buffer is used for clipping the census block groups. Once the census block groups within the boundary are identified, percent minority and percent below poverty level can be calculated.

Comparison of the Relative Difference between the COC and the State/County

There is a built-in function that calculates the relative difference between the COC values and the state/county values. The formulas for calculating the relative difference are as follows:

Relative difference between percent minority values:

{[(% minority in the COC) – (% minority in the reference community)] / % minority in the reference community} x 100

Relative difference between percent below poverty line values:

{[(% below poverty level in the COC) – (% below poverty level in the reference community)] / % below poverty level in the reference community} x 100

Once the census block groups within the boundary of a COC are identified, the analyst can click on the "S/C" button. The percent minority and percent below poverty level in the COC, the state, and the county and also their relative differences will be calculated and the results will be displayed. The analyst can save these statistics to a table or dismiss the window without saving (Figure 7).

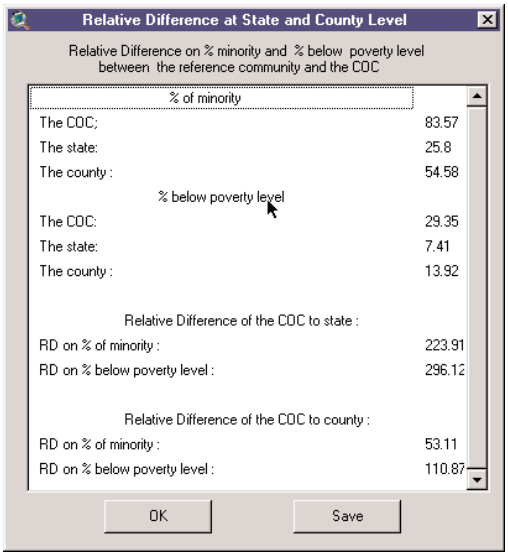


Figure 7 Relative difference between the community of concern and the state/county.

Site-Specific Analysis

Once a general screening identifies a COC as a potential environmental justice area, the community is subjected to site-specific analysis. The COC is compared with one or more reference communities. In a site-specific analysis, the reference areas should be sufficiently close and/or comparable to the COC so that it would be reasonable to assume the presence of similar circumstances if environmental justice were not a factor. For example, in our pilot study we compared Greenpoint/Williamsburg in Brooklyn, New York, the proposed site of a USA Waste transfer station, to all other New York City communities near waste transfer stations. Greenpoint/Williamsburg is the COC and the other New York City communities are the reference areas. Selection of reference communities is case specific and has to be justified by the analyst. However, once the reference community is selected, the way in which reference area boundaries are defined is similar to the way a COC is defined. Similar to the general screening, the "Ref" button will calculate and display the demographic data of the COC and the reference area. If the relative difference between the COC levels and the reference area levels is greater than 25%, the COC will be saved for further analysis on environmental burden (Figure 8).

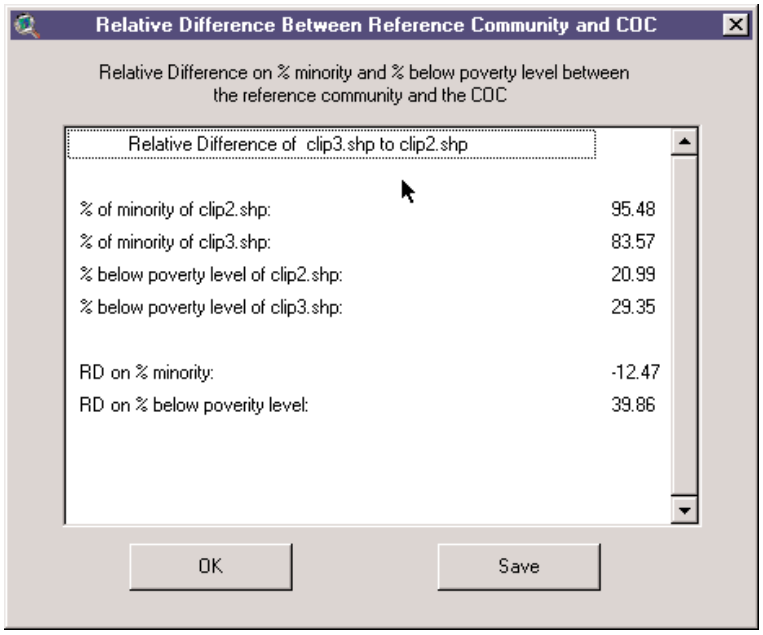


Figure 8 Relative difference between the community of concern and the reference community.

Conclusion

To achieve the goal of environmental justice, EPA Region 2 established the Environmental Justice Workgroup to provide advice and counsel on justice issues. The Environmental Justice GIS Application was developed to provide an easier way for the workgroup to identify areas of high minority or low-income representation that may suffer from disproportionate environmental burden. In this version, analysis

involves only the demographic data; environmental burden has not been addressed, as more research on methods in assessing the environmental burden is needed. Currently, EPA Region 2 is conducting a number of pilot studies using the application to screen for potential environmental justice areas. This application will be widely used within Region 2 as environmental justice becomes integrated into the daily work of the regional employees.

Although this application is developed for assessment of environmental justice, it can be easily modified for other purposes. There is no limit on adding other data layers that pertain to an environmental justice analysis. Health and other available environmental data can be integrated into the application for analysis of correlation between demographic characteristics of communities, community health, and environmental exposure.

References

1. The White House. 1994. *Executive order on federal actions to address environmental justice in minority populations and low-income populations.* Executive Order 12898. Washington, DC: The White House. 11 February.
2. USEPA Region 2. 1997. *Draft implementation guidance to the interim USEPA Region 2 policy on identifying environmental justice areas.* New York: US Environmental Protection Agency.

Understanding the Role of Geospatial Information Technologies in Environmental and Public Health: Applications and Research Directions

U Sunday Tim*
Agricultural and Biosystems Engineering Department, Iowa State University, Ames, IA

Abstract

For more than two centuries, epidemiologists, health care professionals, and medical researchers have sought to develop more refined methods of characterizing populations exposed to hazardous substances in the environment. This effort, for the most part, has involved using aspatial analysis techniques to explore the relationships between environmental quality and certain types of diseases, identify areas for the focus of public health education and community outreach programs, and delineate target and control populations for health studies. Although aspatial techniques have been quite useful in many applications, methods now exist that use rapidly emerging geospatial technologies to effectively manipulate, analyze, and display geocoded environmental and public health data on an unprecedented scale. Geographic information system (GIS) technology can be used to improve the level of understanding of environmental health problems and for exploratory data analysis to test or support hypotheses regarding disease causation. This paper examines the roles and limitations of GIS in environmental and public health research and illustrates, through an example application, the use of GIS functionality in the management and analysis of environmental and public health data. Future trends and issues in the use of GIS in environmental epidemiologic research are discussed. Given the recent advancements in GIS functionality and the widespread availability of digital public health data, it is timely to examine potential implications of geospatial technologies in this research area.

Keywords: environment, public health, epidemiology

Introduction

The need to examine and manage the health needs of a growing population has dramatically increased the demand for information systems that capture, manage, analyze, and display data. Geographic information systems (GIS) represent a powerful, new technology for integrating and manipulating large amounts of data obtained from different sources (1). Since its development in the 1960s, GIS technology has proven to be an extremely useful tool for acquiring, storing, manipulating, analyzing, and presenting georeferenced spatial data. Today, government agencies, utility companies, businesses, and researchers have invested billions of dollars in acquiring data as well as GIS hardware and software for application in such varied fields and disciplines as agriculture and natural resource management, health care, business, education, and

* U Sunday Tim, Iowa State University, 215 Davidson Hall, Ames, IA 50011 USA; (p) 515-294-0466; (f) 515-294-2552; E-mail: tim@iastate.edu

military sciences. Numerous case studies in the literature suggest that the use of GIS is definitely making significant contributions to the integration, analysis, and presentation of spatial and non-spatial data in these application areas.

Inspired by the present and future potential of GIS, epidemiologists, medical geographers, and environmental scientists are beginning to adopt the technology for integrated analysis of environmental health data. The usefulness of GIS for environmental epidemiologic research is obvious, because all relevant data can be combined, stored, queried, analyzed, and displayed within a GIS to reveal the associations between environmental exposures and the spatial distribution of disease. Somewhat reminiscent of John Snow's classic case study of the association between a cholera outbreak and the Broad Street station water pump in London in the 1840s, GIS can be used to identify the space-time distribution of disease in relation to possible environmental factors (2). Asking many of the same types of questions as before but using techniques of spatial analysis, epidemiologists, medical geographers, and biostatisticians can evaluate the spatial distribution of disease or specify locations and system interaction points that may facilitate disease control or eradication. Disease ecology is inherently integrative and spatial, and GIS provides the environment in which the biophysical, social, behavioral, and cultural worlds can be combined for a systemic understanding of health and disease.

The 1986 Chernobyl accident and the subsequent deposition of radioactivity over large areas of northern Europe focused the attention of the environmental health science community on the inadequacies of aspatial techniques for establishing relationships of disease to environmental factors (3). GIS provides the data analysis and spatial modeling functions that could be used to integrate information on radiation fallout doses with perinatal mortality rates at different geographic scales. By explicitly linking health outcome to demographic and environmental factors, GIS can facilitate a reorientation toward population-based explanations for health differentials. Other potential applications include the following:

- GIS can be used to manipulate data collected from case-control studies to estimate exposure of individuals or segments of the population to different forms of pollution and disease.
- GIS databases on the location of environmental hazards, as well as disease and demography, can be used to develop or test etiologic hypothesis.
- Using GIS for exploratory spatial analysis of health data can establish disease causation. (Because of this, epidemiologists have been evaluating the capabilities of this technology.)

In spite of this potential, though, there are substantial problems and difficulties that must be addressed before the full benefits of GIS in environmental and public health research can be derived.

This paper examines the role of GIS in environmental epidemiology. Specifically, it addresses the three most important issues related to the use of GIS in environmental health research: the benefits of GIS in environmental epidemiology, the factors that impede the use of this technology, and the emerging trends in GIS technology as they relate to environmental health research. The potential benefits of GIS are examined from two primary perspectives—GIS and environmental health research. From the GIS perspective, demand is increasing for tools and information systems that not only add

value to spatial data, but also support policy decision-making. From the environmental health research perspective, tools are needed to efficiently collect, store, manage, analyze, and display large volumes of health data that examine known or suspected associations between human health and environmental quality, establish the spatial patterns of disease etiology, or generate etiologic hypotheses.

This paper cannot do justice to the full range of issues related to the use of GIS in environmental epidemiology; it may even raise more questions than it answers. But current and emerging applications of GIS in environmental epidemiologic research make this an appropriate time to examine the role of the technology and speculate on what the future holds. The remainder of the paper is organized as follows: First, the role of GIS in environmental epidemiology is briefly examined. Next, the factors that limit the use of the technology are discussed. Finally, the future in GIS trends and challenges are discussed with emphasis on how these trends impact environmental health research.

Role of GIS in Environmental Epidemiology

Nearly all health problems related to environmental pollution have spatial dimensions that make them candidates for GIS analysis. The GIS technology provides a dynamic environment for evaluating and predicting both the short-term and long-term public health risks of environmental hazards. It provides a framework within which to analyze adverse impacts of environmental pollution and facilitates effective presentation of public health information in an easily understood manner. Douven and Scholten (4) identified several applications of GIS in environmental epidemiologic research. These include:

- Collection, storage, and organization of spatial and non-spatial data.
- Mapping of environmental health data to uncover the spatial pattern of disease.
- Spatial modeling to disclose the spatial and temporal nature of disease ecology.
- Statistical analysis to explore the association between diseases and other covariate factors (e.g., socioeconomic, demographic).
- Searching for spatially related aspects of disease etiology.

In these application areas, the benefits of GIS include:

- Rapid access to environmental, demographic, public health, and other relevant data for use in decision-making tasks.
- Easier update of surveillance data and associated geocoded databases.
- Transformation and analysis of disparate data to investigate a wide range of space-time relationships.
- Identification of geographic regions that, because of their unique physical attributes, may act as a source or sink for contaminants that are major health concerns.
- Dissemination of environmental health information in a variety of forms.

During the past several years, the number of professional and research papers and case studies documenting the relevance of GIS in environmental epidemiology has rapidly increased (2,5). Specific examples include a study of the role of environmental variables in the spread of vector-borne diseases by Glass et al. (6), a determination of community

vulnerability to hazardous materials by McMaster (7), and an evaluation of public health effects of toxic chemicals by Stockwell et al. (8). In addition to these, Geschwind et al. (9) investigated the proximity of residences of persons with congenital malformations to hazardous waste sites. Dunn and Kingham (10) combined air quality estimates with health outcome data to explore spatial variation in respiratory ill health, specifically to determine whether emissions from an industrial pollution source might be influencing health status. Kingham et al. (11) integrated statistical analysis techniques with GIS to study the environmental correlates of children's respiratory health. Collins et al. (12) combined atmospheric dispersion modeling, statistical analysis, and knowledge-based techniques with GIS to examine the relationship between exposure to nitrogen dioxide and respiratory health in children. Guthe et al. (13) combined data from various sources to map the spatial patterns of lead exposure and sensitive populations in New Jersey. Wartenberg et al. (14) used GIS to assess health risks of populations living near high-voltage power transmission lines. Stallones et al. (15) proposed a data retrieval approach based on the concepts of GIS for the surveillance of the health status of populations living near hazardous waste sites. Andes and Davis (16) manipulated the 1990 US Census TIGER/Line file data within a GIS to evaluate the geographic distribution of infant mortality in Alaska. Glass et al. (17) used GIS map overlay techniques to investigate residential environmental risks for Lyme disease in Baltimore. These studies all recognized the unique role and utility of GIS in explaining how the environment, demography, and other factors interact to determine health status and disease causation. Indeed, many of the functions and operations available in most GIS facilitate integrated analysis of environmental health data.

There are several areas of environmental epidemiologic research that could benefit from GIS analysis, including spatial epidemiology, analytical epidemiology, descriptive epidemiology, and exposure/risk assessment. Spatial epidemiology uses area-based or point-based approaches to examine differences in the frequencies of disease and health outcomes. Analytical epidemiology involves not only determining the relationship between environmental determinants and disease but also confirming hypotheses of disease causation. In descriptive epidemiology, the objective is to develop thematic, isopleth, or choropleth maps that demonstrate the spatial pattern of disease etiology. These maps can be aggregates of political units (such as census block groups) or geocoded points that express spatial clusters in the health data. Exposure/risk assessment deals primarily with the use of stochastic and deterministic modeling techniques to determine whether high levels of exposure to single or multiple environmental hazards present unreasonable risks in an area. In exposure/risk assessment, results from stochastic/deterministic models provide data on the spatio-temporal distribution of the contamination. Using GIS, for example, data from exposure/risk assessment, as well as biomarkers of human susceptibility to an environmental hazard, could be combined and analyzed to determine the spatial association between disease and environmental covariates, develop etiologic clues that facilitate public health decision-making, or provide new insight into the health risks associated with specific environmental hazards. The size and complexity of public health databases and the complexity of public health problems make the use of GIS all the more necessary. But the major limitations and hindrances of GIS must be recognized; some of those impacting most strongly are discussed below.

Limitations of GIS in Environmental Epidemiology

Until recently, most GIS users paid little attention to the issue of data quality, which is of particular significance in environmental epidemiologic research because the data are obtained from many sources. Generally, the attributes of data quality include correctness, reliability, currency, completeness, timeliness, accuracy, and accessibility. Many epidemiologists and environmental scientists take solace in the notion that public health data are reliable (i.e., the data yield the same result on repeated collection, processing, analysis, and display from the same database), current (i.e., the data are recorded at the time of the event or observation and are continually updated), and accessible (i.e., the data are available to authorized users when needed). These professionals also demand quality in the data collected, analyzed, interpreted, and reported. However, most data used in epidemiological research are incomplete, due in part to the high capital and human resources required to collect and assemble them. Using such data with GIS to explore associations between disease incidence rates and environmental, socioeconomic, and demographic factors can be problematic.

The creation of integrated databases depicting changes in disease distribution through space and time is central to many studies in environmental epidemiology. This creation requires not only the maintenance of consistent surveillance and monitoring procedures but also demands that the data be current and contemporaneous. Thus, currency and timeliness of environmental health data are another data quality issue that concerns users. A frequently cited problem is the use of incorrect point data—caused by migration across the boundaries of health reporting zones—for GIS analysis (18,19). According to Davis and Chilvers (20), currency problems resulting from migration in and out of a surveillance zone can produce a "dilution effect" in many studies that evaluate spatial variation in disease incidence rates. An issue related to currency and timeliness in health data is latency, caused by, for example, the considerable lag time between human exposure to an environmental hazard and the emergence of a disease. In many circumstances, significant problems can be introduced when attempting to discover current relationships between exposure and disease incidences. Recording a patient's history and physical examination months after patient discharge is another common form of latency.

Another data quality issue has to do with striking an appropriate balance between data accuracy and the desired scale for spatial analysis. While exploratory analysis of health data using individual case locations or census blocks can be very attractive compared to counts in aggregated regions (e.g., census block groups or census tracts), this attractiveness is lost if the data on individual locations are inaccurate or if covariate information is only available as spatial aggregates. King (21) categorized limitations of this type as "ecological fallacy," in which individual-level relationships are inferred from analysis of aggregate-level data.

Increasingly, health care professionals and epidemiologists face a dilemma: meeting the health care community's need for information while protecting patients from unauthorized, inappropriate, or unnecessary intrusion into their personal information in the database. The drive for increased use of digital health information linked together by modern networking technologies could expose sensitive health information to a variety of threats and misuse. The growing use of health data in environmental epidemiologic research demands that issues of privacy, confidentiality, and security be

adequately addressed. A report by the federal Office of Technological Assessment emphasized that current laws generally do not provide consistent, comprehensive protection of health information (22). Currently, communication between patients and their health care providers is considered confidential and health care professionals are therefore bound by legal and ethical standards to maintain confidentiality and privacy. Nonetheless, the need for more uniform and acceptable guidelines for access, use, and presentation of health information is increasing. Also, GIS programs must be equipped with improved security facilities for conducting exploratory health data analysis without disclosing confidential information. With these initiatives, an increased role of GIS in future epidemiological research is inevitable.

In geographic analysis of health data, a recurrent theme is a strong, often localized, pattern and cluster in disease ecology. Spatial heterogeneity and localized variations can present problems for conventional statistical methods that assume global relationships with few or no spatial singularities. Increased recognition of spatial heterogeneity in health data has led to a resurgence of emphasis on understanding disease ecology in a spatially explicit context. Suggesting possible environmental and behavioral factors in disease causation, identifying strong spatial relationships between environmental factors and disease, and confirming etiologic hypotheses developed from manipulation of environmental health data fall within the domain of GIS. However, these activities require the use of sophisticated statistical techniques. Thus, another factor that limits the use of GIS in environmental epidemiology is the lack of statistical analysis functions in many GIS programs. Although a few GIS programs support basic statistical summarization of data, the functions and techniques needed for exploratory analysis of environmental health data are still lacking (23). Some attempts have been made during the last few years to couple statistical programs with GIS software packages. For example, Openshaw et al. (24) described a spatial statistical analysis environment that links statistical analysis programs with GIS to search for geographical correlates of leukemia. An increasing number of case studies involving the development of interfaces between GIS and statistical software programs has been reported (25).

Yet another factor that limits the use of GIS technology in environmental epidemiology relates to the methodological problems often encountered when exploring the spatial patterns of disease etiology using spatial analysis techniques. These problems arise from the fact that a GIS-based analysis of disease patterns involves complex manipulations and overlay of data themes; many epidemiologists and health care professionals are not fully familiar with the theoretical concepts that underlie most GIS programs. Rather, these individuals are experts in the use of aspatial techniques that incorporate socioeconomic, demographic, genetic, gender, and environmental factors to explain health outcomes. Standard GIS analysis, including map overlays, cartographic modeling, and other advanced operations on spatial data, have not entered the arsenal of epidemiological analysis. Familiarity with GIS concepts is necessary to determine if the results of GIS epidemiological analysis are accurate and appropriate. Newly formed collaborations between epidemiologists, health care professionals, and GIS builders should provide opportunities for improved spatial analysis, interpretation, and presentation of environmental health data.

Example Application

Given the utility of GIS technology in many disciplines, and realizing the need for GIS in environmental health, a study was initiated to develop an integrated system for organizing, managing, analyzing, and displaying environmental and public health data collected in Iowa. The system, called EMPHASIS (for **E**nviron**M**ental and **P**ublic **H**ealth dat**A** analy**SI**s **S**ystem), used ArcView GIS (ESRI, Redlands, CA) and an Oracle (Oracle Corporation, Redwood Shores, CA) relational database management system to integrate and manipulate public health outcome data with environmental, socioeconomic, and demographic data. Specifically, EMPHASIS was developed, through a Seed Grant from the Center for Health Effects of Environmental Contamination at the University of Iowa, to provide an interactive data management and display environment. EMPHASIS could be used to (1) assemble all pertinent information on the presence of contaminants in the environment and, through GIS analysis, correlate the information with various health outcomes; (2) generate or test hypotheses regarding the spatial associations between environmental contamination and disease incidence rates; and (3) identify study populations with potential exposure to environmental hazards.

Development of EMPHASIS was set within the context of using ArcView GIS to integrate, analyze, visualize, and display large quantities of data and identify those environmental factors that covary spatially with disease indices or are concerned in disease causation. Hence, effort was focused on designing the system to facilitate determination of the spatial relationships between morbidity/mortality data from cancer surveillance activities and other relevant demographic (e.g., population) and environmental (e.g., groundwater vulnerability, chemical use factors) information. Figure 1 shows the general architecture of EMPHASIS, which was implemented on a desktop personal computer and incorporates Oracle, ArcView GIS (version 3.0), and S-PLUS (MathSoft, Cambridge, MA). The choice of these programs should not be seen as restrictive, since similarly structured programs could easily be used in their stead. However, the unique combination of these software packages facilitates identification of geographic location, data integration, data management and query processing, spatial analysis and modeling, and display of a wide variety of environmental and public health data.

A primary goal in the design of EMPHASIS was to procure a turnkey GIS environment through which large volumes of information related to environmental and public health (mainly morbidity and mortality) data could be readily accessed, efficiently analyzed, and rapidly visualized. To achieve this goal, several options for data retrieval, query, and visualization were developed. In one option, users can directly retrieve and query the data in Oracle and generate tabular reports. Figure 2 shows how a standard and interactive database query produced a tabular summary of cancer morbidity data collected in Iowa between 1973 and 1992, keyed to the respective county federal information processing standard (FIPS) code. Figure 3 shows a typical query interface and the result of a map overlay performed by using some of the spatial and attribute information in the EMPHASIS database. In Figure 3, data on groundwater vulnerability by hydrogeologic region were integrated with the water quality database obtained from the 1988–1989 Iowa Statewide Rural Water Well Survey as well as morbidity and mortality data from the State Health Registry, maintained by the Center for Health Effects of Environmental Contamination at the University of Iowa, Iowa City.

Presently, EMPHASIS is structured so that new information can be added and

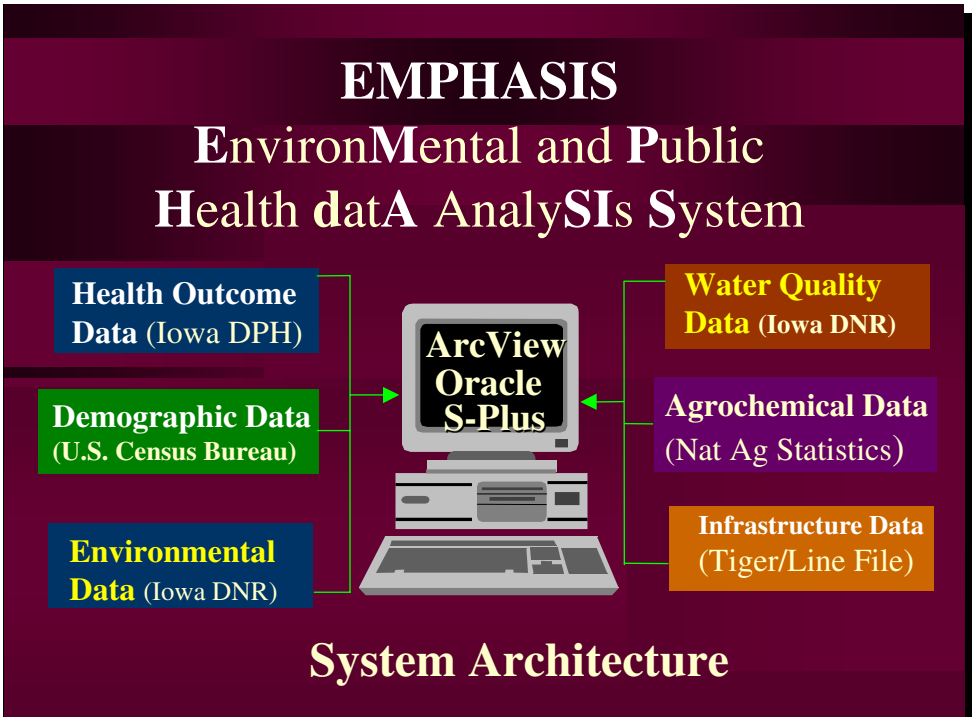


Figure 1 The conceptual structure of EMPHASIS.

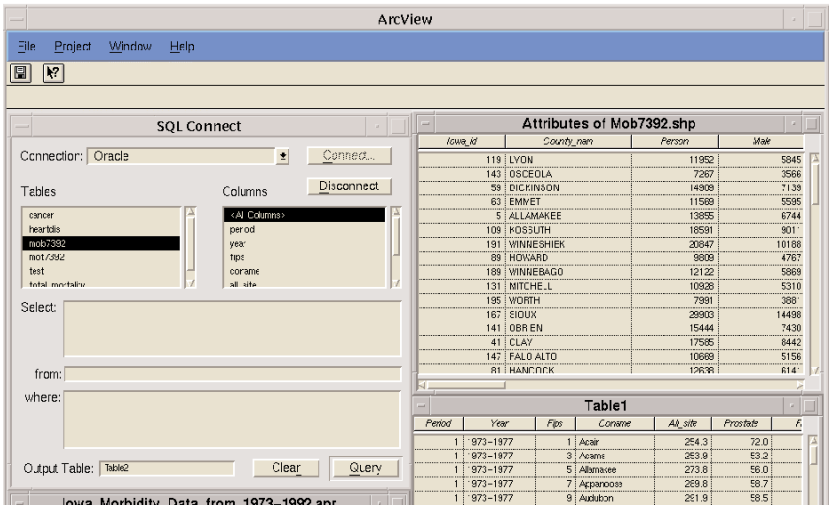


Figure 2 Typical screen display of an EMPHASIS query session.

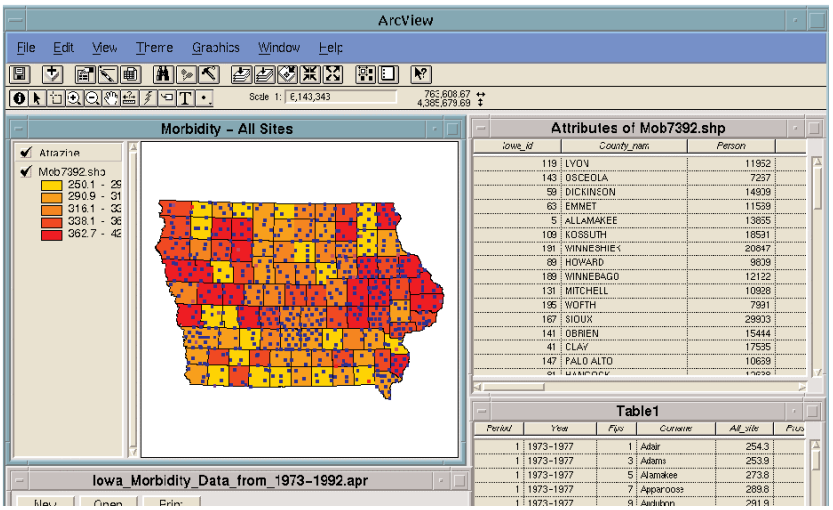


Figure 3 Screen display of environmental and public health data in EMPHASIS.

analyzed as it becomes available. It supports a user-friendly, icon-based menu with mouse interaction for selecting menu options and has two basic modules. One module supports interactive data management, analysis, and visualization, while the other supports online query and reporting. In both modules, the user has full control of selecting the attribute data for analysis and defining the geographic extent of data analysis and display by using the pan and zoom icons in ArcView GIS. Other unique features of EMPHASIS include: (1) it is designed to make optimum use of existing environmental quality data and public health information to minimize duplication of information among state agencies and institutions; (2) it can incorporate existing and future advances in information exchange (e.g., the Internet) to provide an interactive environment for efficient data access and data exchange; and (3) its data processing and display capabilities are powerful enough to facilitate integrated analysis of local or regional environmental health issues.

Future Trends

Driven by technological innovations, the methods and tools used in environmental epidemiology are changing and will continue to do so. The technology for collecting, processing, storing, and retrieving environmental health data is evolving from a paper-driven, labor-intensive process to one that employs sophisticated computers and information systems. Just as the introduction of magnetic resonance imaging provided a new technology for collecting health data, recent spatial technologies, such as GIS, are revolutionizing the way health data are analyzed and presented. In the future, two major benefits are likely to emerge from the application of GIS in environmental epidemiology. The first benefit will be the ability of health care professionals and epidemiologists to use GIS as a tool to interactively manage and disseminate public health information; search for ecological associations among health data and environmental,

socioeconomic, or demographic factors; and identify the spatial location and geographic distribution of disease outbreak to document changes in incidence and prevalence. The other benefit will be the ability to analyze disparate data interactively, emphasizing human health-environment relationships in the context of cultural and behavioral factors.

Challenges in environmental epidemiology during the past few years have ushered in a new era for integrated, spatially explicit data analyses that use GIS. While the application of GIS in this field is still in its infancy, certain observations about future trends and prospects can be made. Today, health, demographic, and socioeconomic data of various spatial scales are increasingly available on the Internet. Indeed, GIS application is entering an "information-rich" era in which large volumes of data are available through communication networks with interactive data filters and data access protocols. The ability to examine the spatial patterns of disease by integrating disparate health outcome data with other disparate information on the Internet and intranets is now within the reach of medical geographers, epidemiologists, and biostatisticians. However, maintaining the integrity of health data on the "information superhighway" will require the establishment of industry-wide standards for data access and data sharing. The ease of information transfer for multiple users without the need for human interaction will raise new concerns for health care professionals. The use of intranets and the Internet will also present new challenges.

As environmental epidemiology enters the 21st century, GIS application will become more widespread. Due to the factors discussed earlier, the full potential of GIS in environmental epidemiology has yet to be unlocked. In a number of existing applications, the need to combine environmental, social, cultural, economic, and demographic data to explore disease-environment-behavior relationships is at odds with the need to maintain security and confidentiality. Although security and confidentiality issues for demographic, socioeconomic, and health data have been established, these issues are only beginning to emerge in the integrated analysis and dissemination of environmental health data. While techniques such as encryption, security servers, user access/password authentication, and firewalls (26) have been widely implemented to control access to confidential information, the degree of concern over unauthorized, inadvertent disclosure, modification, and destruction of health data will increase in the future.

Summary

For over six decades, research activities in environmental epidemiology have focused on a series of fundamental questions: How do people and societies respond to environmental hazards and what factors influence their choice of adjustments? What relationships exist between incidence rates and socioeconomic variables? How can we model these relationships? What areas have extreme high and low disease incidence rates? Within the last decade, other questions have been added to this list, including: Are societies becoming more vulnerable to environmental contaminants? What spatial associations exist between disease incidence rates and other variables? Is there evidence of clustering in respect to specified sources or possible causes? Is there any evidence of trends, patterns, or other variation in environmental health data? To answer these questions, extensive use has been made of spatial information technologies such as GIS.

These technologies facilitate understanding of how humanity (e.g., culture, society, behavior), the physical world (e.g., topography, land use, climate), and biology (e.g., vector and pathogen ecology) interact to produce foci of disease. As discussed in this paper, GIS allows users to combine, query, transform, analyze, and present environmental health information in ways that were not previously possible.

GIS has indeed emerged as an efficient tool for understanding and characterizing the geographic, socioeconomic, demographic, and environmental variables that influence disease incidence rates. However, deriving the full benefits of GIS in environmental epidemiology will depend on how the environmental health research community approaches and resolves the issue of data quality. Also, integration of various multimedia tools to form a health care decision support system and the growing capability to link public and private databases require that issues of privacy, security, and confidentiality be fully addressed. Public perception about data privacy issues also needs to be changed. Citizens should be educated about the value of GIS and the many benefits that it offers in environmental epidemiology. Environmental scientists, medical geographers, and epidemiologists also need to understand the limitations of GIS, data, and GIS analysis.

References

1. Burrough PA. 1986. *Principles of geographical information systems for land resources assessment.* New York: Oxford University Press.
2. Vine MF, Degnan D, Hanchette C. 1997. Geographic information systems: Their use in environmental epidemiology. *Environmental Health Perspectives* 105:598–605.
3. Savchenko VK. 1995. *The ecology of the Chernobyl catastrophe.* New York: The Parthenon Publishing Group.
4. Douven W, Scholten HJ. 1995. Spatial analysis in health research. In: *The added value of geographical information systems in public and environmental health.* Ed. MJC de Lepper, HJ Scholten, RM Stein. Boston: Kluwer Academic Publishers. 117–33.
5. Mayer JD. 1983. The role of spatial analysis and geographic data in the detection of disease causation. *Social Science and Medicine* 17:1213–21.
6. Glass GE, Morgan JM, Johnson DT, Noy PM, Isreal E, Schwartz BS. 1992. Infectious disease epidemiology and GIS: A case study of Lyme disease. *Geo Info Systems* 2:65–9.
7. McMaster R. 1988. Modeling community vulnerability to hazardous materials using GIS. In: *Introductory readings in geographic information systems.* Ed. DJ Peuquet, DF Marble. London: Taylor & Francis. 183–94.
8. Stockwell JR, Sorenson JW, Eckert JW, Carrera EM. 1993. The US EPA geographic information system for mapping environmental release of toxic chemical release inventory (TRI) chemicals. *Risk Analysis* 13:155–64.
9. Geschwind SA, Stolwijk JA, Bracken M, Fitzgerald E, Stark A, Olsen C, Melius J. 1992. Risk of congenital malformations associated with proximity to hazardous waste sites. *American Journal of Epidemiology* 135:1197–1207.
10. Dunn CE, Kingham SP. 1996. Modeling air quality and the effects on health in a GIS framework. In: *Innovations in GIS 3.* Ed. D Parker. Bristol, PA: Taylor & Francis. 205–13.

11. Kingham SP, Acquilla SD, Dunn CE, Halpin JE, Foy CJ, Bhopal RS, Blain P, Pless-Mulloli T. 1995. Health in the vicinity of industry in Bishop, Auckland. Unpublished Report. University of Newcastle upon Tyne, UK.
12. Collins S, Smallbone K, Briggs D. 1995. A GIS approach to modeling small area variations in air pollution within a complex urban environment. In: *Innovations in GIS 2.* Ed. D Parker. London: Taylor & Francis. 245–53.
13. Guthe WG, Tucker RK, Murphy EA, England R, Stevenson E, Luckhart JC. 1992. Reassessment of lead exposure in New Jersey using GIS technology. *Environmental Research* 59:318–25.
14. Wartenberg D, Greenberg M, Lathrop R. 1995. Identification and characterization of populations living near high voltage transmission lines: A pilot study. *Environmental Health Perspectives* 101:626–32.
15. Stallones L, Nuckols JR, Berry JK. 1992. Surveillance around hazardous waste sites: Geographic information systems and reproductive outcomes. *Environmental Research* 59:81–92.
16. Andes N, Davis JE. 1995. Linking public health data using geographic information systems techniques: Alaskan community characteristics and infant mortality. *Statistics in Medicine* 14:481–90.
17. Glass GE, Schwartz BS, Morgan JM, Johnson DT, Noy PM, Isreal E. 1995. Environmental risk factors for Lyme disease identified with geographic information systems. *American Journal of Public Health* 85:944–8.
18. Mathews SA. 1990. Epidemiology using a GIS: The need for caution. *Computers, Environment and Urban Systems* 14:213–21.
19. Cleek RK. 1979. Cancers and the environment: The effect of scale. *Social Science and Medicine* 13D:241–7.
20. Davis JM, Chilvers C. 1980. The study of mortality variations in small administrative areas of England and Wales, with special reference to cancer. *Journal of Epidemiology and Community Health* 34:87–92.
21. King PE. 1979. Problems of spatial analysis in geographic epidemiology. *Social Science and Medicine* 13D:249–52.
22. Office of Technological Assessment. 1993. Protecting privacy in computerized medical information. Washington, DC: US Government Printing Office. OTA-TCT-576.
23. Ding Y, Fortheringham AS. 1992. The integration of spatial analysis and GIS. *Computers, Environment and Urban Systems* 16:3–9.
24. Openshaw S, Cross A, Charlton M. 1990. Building a prototype geographical correlates exploration machine. *International Journal of Geographic Information Systems* 4:297–311.
25. Fotheringham S, Rogerson P. 1994. *Spatial analysis and GIS.* Bristol, PA: Taylor & Francis.
26. Cox LH. 1996. Protecting confidentiality in small population health and environmental statistics. *Statistics in Medicine* 15:1895–1905.

Spatial Analysis of Premature Deaths among African-American Males in Fulton County (Atlanta), Georgia

Adewale Troutman, MD, MPH*
Director, Fulton County Department of Health and Wellness, Atlanta, GA

Abstract

Approximately half of all the life years lost due to premature deaths in Fulton County, Georgia, occur to African-American males. The Fulton County Department of Health and Wellness analyzed the geographical distribution of premature deaths in the county and used a geographic information system (GIS) application to map the occurrence of these deaths by census tract and by major causes of death. The spatial distribution of premature deaths was then integrated with sociodemographic data from census files to provide a geographic risk profile. Polygon overlays and queries by health center areas were provided to allow prevention interventions to be targeted. Similar polygon overlays by commission districts allow the information to be presented to elected officials and to be related to the budgetary process.

Keywords: African-American males, premature deaths, years of potential life lost (YPLL), mapping in public health

Introduction

The poor health status of African-American males is well known. Although their health status is reflected across many dimensions, it can readily be seen in their disproportionately high mortality rates. Not only do African-American males have 1.7 times the mortality rate of their white counterparts (1), they also have a very high rate of "excess" deaths. Excess deaths are defined as the number of deaths that are greater than would be expected if African-American males had the same age-specific death rates as those of US white males. It has been estimated that up to one-third of the deaths of African-American males in this country may be considered excess deaths (2).

The high death rate, as well as the high number of excess deaths, is due to six main causes: HIV/AIDS, homicides/injuries, cardiovascular/cerebrovascular diseases, diabetes, cirrhosis, and infant mortality (3,4). As pointed out by Hale (2), 70% of the excess deaths occur before age 65 and 40% occur before the age of 45. Needless to say, the life expectancy of the African-American male in this country, currently 65 years, resembles that found in a developing nation more so than in an industrialized one (5). White males in this country can expect to live over seven years longer than African-American males.

In planning health interventions, mapping the geographic variability of premature deaths is a valuable tool in understanding the distributions of the deaths. Geographically based targeting can be used to distribute health resources. However, other characteristics may also be related to the geographic pattern of disease

* Dr. Adewale Troutman, Fulton County Department of Health and Wellness, 99 Butler Street, SE, Atlanta, GA 30303 USA; (p) 404-730-1200; (f) 404-730-1294; E-mail: AT060A@dhr.state.ga.us

prevalence and premature mortality. Much information from the census as well as other databases can be related to the health status of small geographic areas.

Recent studies have focused on the contribution that structural variables play in contributing to premature mortality and excess deaths. It has been found that geographic areas characterized by a high concentration of poverty in conjunction with a high degree of segregation tend to greatly exacerbate poor health status as well as many other social and economic ills characteristic of these neighborhoods. The ill effects caused by the combination of these two conditions have been called "neighborhood" or "concentration" effects (6).

Ecological variables such as income, education, or unemployment have also been found to be important risk factors. Ecological variables are those population characteristics of the census tract as a whole as opposed to those that characterize an individual's behavior or history. Appropriate ecologic variables can be constructed from census information and related to disease patterns. For example, Wells and Horm (7) found that median education of small areas was inversely related to never having had a mammogram. Thus the ecologic characteristics of an area may be helpful in identifying underserved areas or areas with underutilization of services. Ecologic variables may also be useful in constructing estimates of screening behaviors that might be expected to occur in an area.

The purpose of this analysis was to determine the spatial distribution of the premature deaths that occur to African-American males in Fulton County, Georgia. Thus, a geographic information system (GIS) application was used to analyze the geographic patterns of premature deaths. The GIS was further used to relate the pattern of premature deaths to both structural and ecologic characteristics of small areas. In addition to the analysis, there is a practical application of GIS mapping in interacting with the community and political forces needed to turn data into information that can assist in shaping interventions and policy.

Of critical importance to health planning and program development at the local level is the ability to garner political and budgetary support for health interventions. It is necessary to work with numerous community-based organizations and partners in planning health programs geared toward the needs of African-American men. This analysis was designed to facilitate communication with the various stakeholders and partners involved in planning and implementing a preventive health program aimed toward African-American men in Fulton County. Mapping is a powerful tool for helping people visualize the geographic distribution of health problems as well as the structural and ecologic context of these problems. Such presentation makes explicit the rationale for targeting preventive interventions and contributes to the development of community as well as political support for health interventions.

Method

Population Description of Fulton County, Georgia

Fulton County has a population that is 50% African American, 48% white, and 2% composed of other races. African-American males make up approximately 24% of the population, as do white males and white females; African-American females comprise 28% of the population.

Mortality data for this analysis were from data files maintained by the Fulton County Department of Health and Wellness and reflect death certificate data that are recorded for all resident deaths in Fulton County. Mortality data were compiled for the period 1991–1995 by census tract, by sex-race, and by cause of death. In addition to analyzing the total deaths for all causes, selected cause-specific analyses were also included. The specific causes along with their *International Classification of Disease* (ICD) codes (8) are as follows: HIV/AIDS (42.1–44.9), homicide (960–969), all heart disease (390–459), heart attacks (410), hypertension (401–405), cerebrovascular disease (430–438), all cancers (140–208), cancer of the lung (162), and cancer of the prostate (185).

The number of premature deaths, defined as deaths occurring under the age of 75, was determined for all causes as well as for each of the cause-specific deaths. Also, years of potential life lost (YPLL) was calculated by subtracting the age at death from age 75 for all deaths that occurred prematurely and summing these numbers for each sex-race group by census tract.

Thematic maps were constructed using ArcView (ESRI, Redlands, CA), a GIS package that runs on a desktop PC. Maps were developed for both the number of premature deaths and the number of YPLLs by census tract, dividing the census tracts into quintiles and excluding tracts with zero mortality events.

Neighborhood, or concentration, effects were assessed using the percent minority and percent poverty for each census tract. This analysis is ecological in nature in that the population characteristics of the census tract as a whole are used rather than the individual history of decedents. Both the percent poverty and the percent minority data were obtained from summary data by census tracts compiled by the Atlanta Regional Commission (ARC), a ten-county local planning agency. The primary source of the data was the 1990 US Census Summary Tape File 3A (9). Thematic maps were also constructed dividing the census tracts into quintiles based on percent minority and percent poverty of the tracts. Altogether there are 146 census tracts in Fulton County.

The census tract was the main unit of analysis for mapping the number of premature deaths, YPLLs, percent minority, and percent poverty. However, a number of overlays were constructed for use in presenting information to policy makers, community groups, service delivery partners, and other stakeholders. Overlays consisted of commission districts for Fulton County elected commissioners, health center areas for the Department of Health and Wellness, catchment boundaries for other health care providers, boundaries for the Atlanta Empowerment Zone and for other entities as needed. Catchment area boundaries were constructed by aggregation of census tracts and block groups to the appropriate level. Fulton County has two at-large elected commissioners (known as District 1 and District 2 commissioners), and five elected commissioners representing each of five geographic areas (known as commissioners for Districts 3 through 7, respectively).

Results

Due to their small numbers, deaths to races other than African American and white were excluded from this analysis. These deaths constituted less than 1% of the total deaths. A detailed analysis of 1995 data is presented. The number of deaths, number of premature deaths, and YPLLs for each sex-race group are presented in Table 1. During 1995 the number of deaths per year in Fulton County was 6,211. Approximately 30%

Table 1 Number of Total Deaths, Premature Deaths, and Years of Potential Life Lost by Sex-Race Group for Fulton County, Georgia (1995)

	Total Deaths	Premature Deaths (<age 75)	Years of Potential Life Lost
African-American males	1,856	1,495	42,300
African-American females	1,497	863	19,868
White males	1,321	768	16,021
White females	1,537	434	6,763
Total: all sex-race groups	6,211	3,560	84,952

Source: (11,12)

were of African-American males, 24% were African-American females, 21% were white males, and 25% were white females. Of the annual total of 6,211 deaths, 3,560 were premature; that is, they occurred at an age younger than 75 years of age. Overall, 57%, or over half, of the deaths in Fulton County are premature.

Table 2 provides a profile of the population, total number of deaths, number of premature deaths, and YPLLs by sex-race groups in terms of percentages. This profile highlights the disproportionate share of disease burden suffered by African-American men in Fulton County. In 1995, African-American male deaths constituted 42% of all the premature deaths versus 24% for African-American females, 22% for white males, and 12% for white females. Thus, while African-American males make up approximately 24% of the population, they account for 42% of all the premature deaths.

Table 2 Percentage of the Population, Premature Deaths, and Years of Potential Life Lost by Sex-Race Group in Fulton County, Georgia (1995)

	Total Population[a] (%)	Total Deaths (%)	Premature Deaths (<age 75) (%)	Years of Potential Life Lost (%)
African-American males	24	30	42	50
African-American females	28	24	24	23
White males	24	21	22	19
White females	24	25	12	8

[a]Source: (9)

YPLL is another way of measuring premature death. This measure captures the impact of the age of death. This measure is higher the younger the individual at the time of death. Overall, there were 84,952 life years lost in Fulton County in 1995. Table 2 shows that deaths of African-American males accounted for 50%, or half, of all the life years lost in Fulton County due to premature deaths. Thus, while African-American males constitute 24% of the population, they account for 30% of all deaths, 42% of the premature deaths, and 50% of the YPLLs. African-American males are disproportionately represented in all three measures of mortality—total deaths, premature deaths, and YPLLs.

As can be seen in Table 3, three-quarters of the premature deaths and two-thirds of YPLLs were due to five major causes: HIV/AIDS, homicide, heart disease, stroke, and cancer.

Table 3 Number of Deaths, Premature Deaths, and Years of Potential Life Lost for All Causes and Selected Cause-Specific Mortality among African-American Males in Fulton County, Georgia (1995)

	African-American Male 1995		
	Total Number of Deaths	**Number of Premature Deaths (<age 75)**	**Years of Potential Life Lost**
All causes	1,856	1,495	42,300
HIV/AIDS	314	313	11,412
Homicides	131	129	5,565
All heart diseases	493	340	6,199
Heart attack	35	35	571
Hypertension	84	84	1,404
Stroke	43	43	795
All cancers	317	244	3,652
Lung cancer	93	76	1,079
Prostate cancer	29	29	254

Source: (12)

Notably, the geographic distribution of deaths for all age groups of African-American males, as well as for premature deaths and YPLLs, are highly correlated. This pattern of mortality reflects the underlying pattern of neighborhood, or concentration, effects characteristic of areas of high poverty and high segregation (reflected in the percent minority population of the census tract). The census tracts with high mortality, high premature mortality, and high numbers of YPLLs among African-American males also tended to be the same census tracts that had a high percentage of minority population as well as high poverty rates.

Maps for number of premature deaths for all causes (Figure 1) as well as for HIV/AIDS (Figure 2) and homicide (Figure 3) are shown to illustrate the geographic variation of premature deaths by census tract for African-American males. Number of premature deaths as opposed to rates were chosen for these maps for health planning purposes because high rates in a small population do not create as many cases as a moderate rate does in a much larger one. Overlays of commission districts illustrate the value of adding boundaries for presentations made to a political stakeholder group.

Discussion

Mapping premature deaths, as was done in this analysis, is a highly effective way of demonstrating the disproportionate disease burden borne by African-American males in Fulton County. Showing the geographic variation and clustering was effective in both demonstrating the problem and in garnering support for preventive health

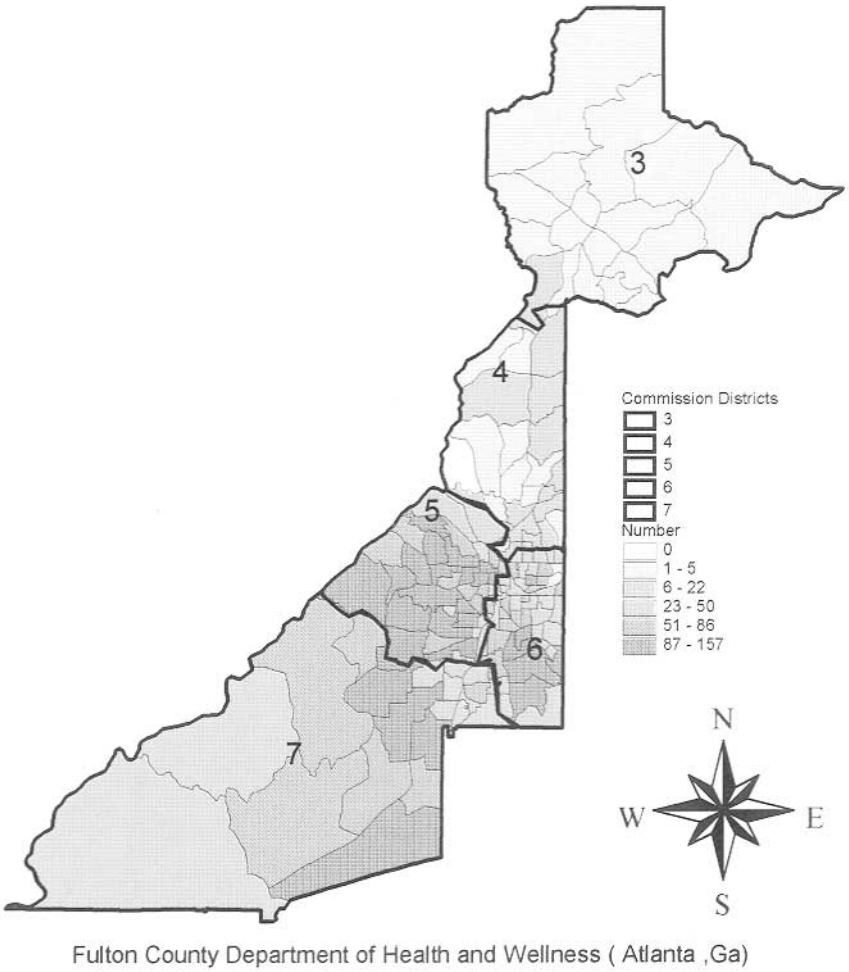


Figure 1 Number of premature deaths (age <75 years) for African-American males, all causes, 1991–1995.

programs. Relating the premature deaths to census variables allows health status to be understood in terms of its neighborhood context.

Both the spatial distribution of mortality among African-American males and the neighborhood, or concentration, effects in Fulton County are highly correlated. As found in other research, such structural variables as high poverty in conjunction with high segregation are powerful risk factors for poor health for a wide range of conditions. While public health has often focused its attention on individual health behavior and risk factors, neighborhood and environmental risk factors appear to be just as important, if not more important. In fact, some researchers have even suggested that neighborhood effects are as great as, if not greater than, such major individual risk factors as smoking (10). Clearly both individual and structural variables of neighborhoods are important and interventions are needed that target both. Preventive clinical services such as immunizations and disease screening programs have clearly been found to be

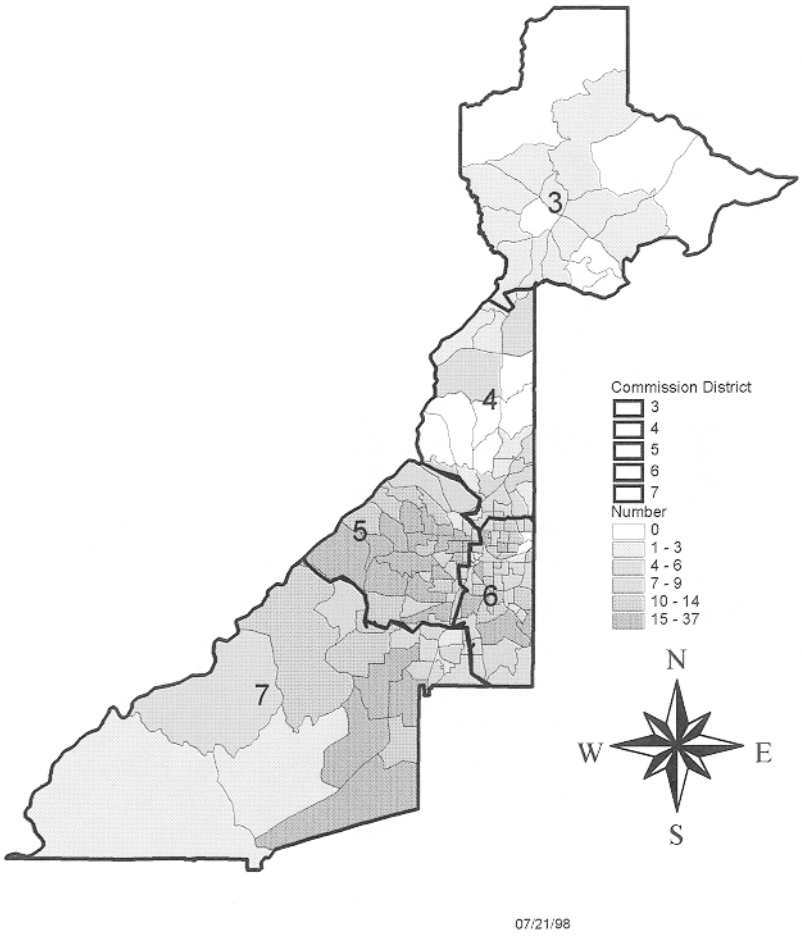


Figure 2 Number of premature deaths (age <75 years) for African-American males, HIV/AIDS, 1991–1995.

effective. So too have been strategies that focus on target groups of individuals exhibiting risk factors for various health conditions.

Much more emphasis, however, is needed on targeting interventions at the neighborhood level as opposed to focusing just on the individual level. Research has shown that behaviors are culturally or socially mediated. For example, if it is considered "cool" or "grownup" to smoke within a social group, individuals are at much greater risk of taking up the habit. Strategies such as policy development and community empowerment through partnership development and coalition building can be effective interventions at the neighborhood level.

Of course, both cultural factors and the cultural context are important for messages that seek to build individual healthful behaviors. Considerably more ethnographic research is needed in this area. This approach seeks to ameliorate, on the individual level, the high-risk circumstances of neighborhood structural variables.

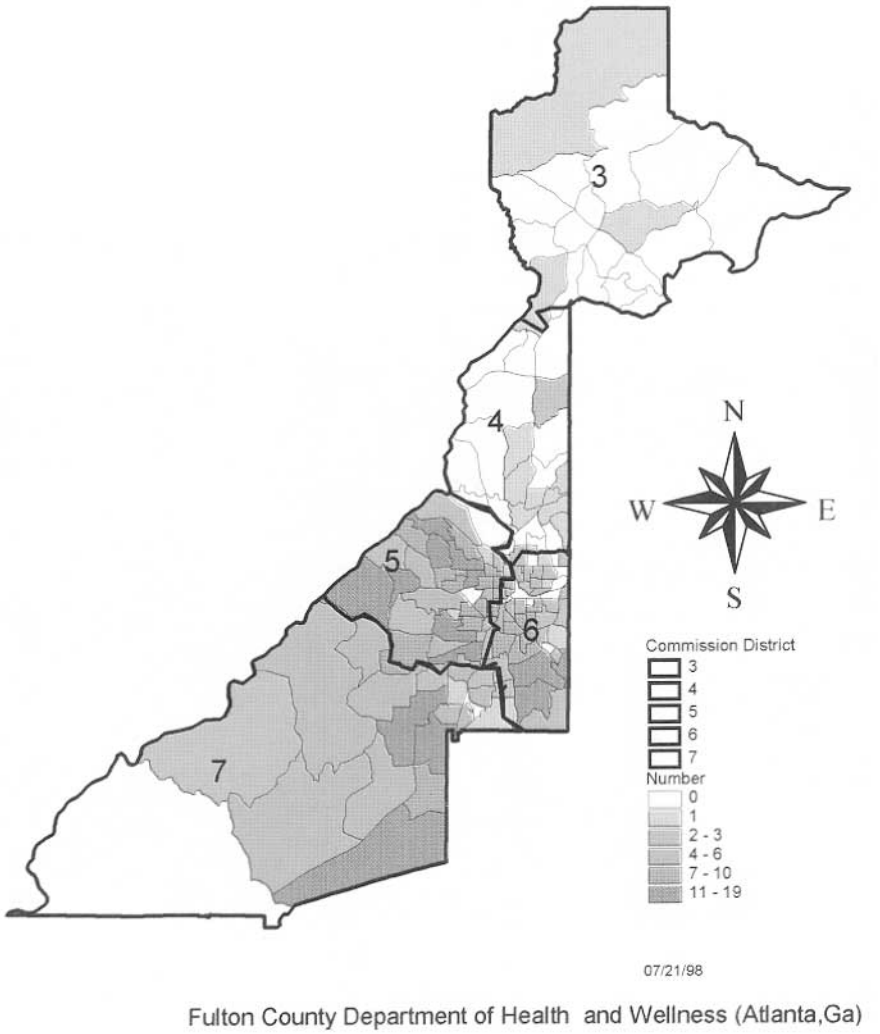


Figure 3 Number of premature deaths (age <75 years) for African-American males, homicide, 1991–1995.

Policy can also be an effective intervention as has been seen with the reductions in both mortality and serious injuries that occurred when speed limits were lowered to 55 mph. Local neighborhoods have also been successful in using zoning laws and business licensing regulations to limited the number of liquor stores that can be opened in a given neighborhood. Thus, both individual interventions and public policy, especially when combined with community empowerment, can be potent strategies for building the health of a population at the neighborhood level.

The true value of technology is in its application to the real world. Science must serve humanity to achieve its true mission. Advances in the application of geographical information mapping to focus on issues of the public's health provide us with a powerful example of that principle. In that regard, the power of GIS lies in its ability to

be used as a tool in the process of examining, evaluating, and improving the health status of communities. In this instance, we have placed GIS technology in the practical context of assisting the shaping of public policy and directing the decision-making process of local political leadership.

The value and uses of GIS technology can be illustrated in several ways. This project highlights many of them. Specifically, the practical applications of GIS technology include the following:

- As a tool in the development of stakeholder support
- Assisting in the determination of the siting of fixed resources
- Assisting in the process of targeting intervention and prevention strategies
- As a tool in providing support for budget allocation requests
- Providing a geographical focus for the development of public health advocacy groups
- Providing a framework for the building of coalitions with non-public health partners

The policy-makers at the county level depend on department heads to provide them with the data necessary to develop public policy. Frequently, their time demands and their perspective on public need are such that the presentations must be powerful, illustrative, and clear. GIS mapping fits all of these criteria.

Using color-coded GIS maps focusing on the health status of African-American men in Fulton County, we were immediately able to demonstrate to our local policy-makers where the most dramatic needs were. This will allow us to target our resources and programming to the specific areas represented on the GIS map. Progress can then be measured and its representation illustrated for the benefit of the local board of commissioners.

Much as this technique can be used with policy-makers, it is equally valuable to illustrate health issues to any segment of the population that is not oriented to "hard" epidemiological data. This includes the affected populations, other stakeholders, budget committees, and potential collaborative partners from all segments of society.

GIS mapping relates information on public health data in a clear, concise, yet dramatic manner that is consumer-friendly and almost self-explanatory. It will continue to be a mainstay in the way in which we communicate public health issues to our colleagues.

References

1. National Center for Health Statistics (NCHS). 1998. Health, United States, 1998 with socioeconomic status and health chartbook. Hyattsville, MD: NCHS.
2. Hale, CB. 1992. A demographic profile of African Americans. In: *Health issues in the black community.* Ed. RL Braithwaite, SE Taylor. San Francisco: Jossey-Bass. 6–19.
3. US Department of Health and Human Services. 1985. *Report of the Secretary's Task Force on Black and Minority Health (vol.1, executive summary).* Washington, DC: US Government Printing Office.
4. Braithwaite RL, Taylor SE, Eds. 1992. *Health issues in the black community.* San Francisco: Jossey-Bass.

5. McCord C, Freeman HP. 1990. Excess mortality in Harlem. *New England Journal of Medicine* 322:173–7.
6. Polednak, AP. 1977. *Segregation, poverty and mortality in urban African Americans.* New York: Oxford University Press.
7. Wells BL, Horm JW. 1998. Targeting the underserved for breast and cervical cancer screening: The utility of ecological analysis using the national health interview survey. *American Journal of Public Health* 88:1484–9.
8. US Department of Health and Human Services. 1989. *International Classification of Diseases, Ninth Revision.* DHHS Publication No. (PHS) 89-1260. US Department of Health and Human Services, Public Health Service, Health Care Financing Administration.
9. US Census Bureau. *1990 decennial census of population and housing, summary tape file 3A (STF3A).* Washington, DC: US Census Bureau.
10. Kaplan GA, Lynch JW. 1997. Wither studies on the socioeconomic foundations of population health. Editorial. *American Journal of Public Health* 87:1405–11.
11. Fulton County Department of Health and Wellness. 1995. Fulton County birth certificates. Atlanta: Fulton County Department of Health and Wellness.
12. Fulton County Department of Health and Wellness. 1995. Fulton County death certificates. Atlanta: Fulton County Department of Health and Wellness.

Regional Patterns of Alcohol-Specific Mortality in the United States

WF Wieczorek,* CE Hanson
Center for Health and Social Research, Buffalo State College, Buffalo, NY

Abstract

Regional patterns of health are usually determined based on areas defined by aggregations of states. A major limitation of this approach is that the regions are defined by state boundaries. Regional characteristics based on factors such as historical settlement patterns, economic activity, housing patterns, and ethnicity may not conform to state boundaries. Regional health patterns may be obscured by the artificial nature of regions defined by state boundaries. Geographic information systems (GIS) provide the capability to develop more sophisticated definitions of regions. This study examines regional patterns of alcohol-specific mortality based on complex definitions of regions not limited by state boundaries. Boundaries for 12 US regions defined by a large number of cultural factors were digitized. The digitized regional boundaries were overlaid onto all counties in the US. Counties split by regional boundaries were assigned to the region that contained the greatest amount of the area for that county. The alcohol mortality data for each county are provided by the Alcohol Epidemiologic Data System of the National Institute on Alcohol Abuse and Alcoholism. This study utilized mortality data that explicitly mention alcohol as a cause of death. Examples of alcohol-specific mortality include alcoholic cirrhosis, alcohol dependence syndrome, and alcoholic cardiomyopathy. Age-adjusted mortality rates were used. Alcohol-specific mortality was used to avoid confounding based on regional differences in the attributable fraction of alcohol-related diseases. Alcohol-specific mortality tended to be higher in the Pacific Southwest, Interior Southwest, and South. The rate in the South decreased substantially when mortality was adjusted for factors such as race and income. The Central Midwest had notably lower rates of alcohol-specific mortality. The study found significant differences in alcohol-specific mortality between regions of the United States. Regional patterns provide insight into the relationship between cultural factors, alcohol use, and alcohol-specific mortality.

Keywords: alcohol, mortality, regions

Introduction

Large areas of relative societal homogeneity are defined as cultural regions. The population of the United States is not an undifferentiated mass that is evenly distributed across the landscape. An examination of regions in the United States can provide an improved understanding of health needs and problems. Unfortunately, most regional analyses of health issues are based on aggregations of states or census statistical areas. These approaches are limited because the regions are based on state or other political boundaries, and the regions lack a strong theoretical foundation. Regional health

* William F Wieczorek, Center for Health and Social Research, HA 205E, Buffalo State College, 1300 Elmwood Ave., Buffalo, NY 14226 USA; (p) 716-878-6137; (f) 716-878-4009; E-mail: wieczowf@buffalostate.edu

patterns may be obscured by the artificial nature of the regions. Post hoc regional analyses of health data based on sub-state areas (e.g., counties, metropolitan areas) lack scientific rigor and are subject to idiosyncratic interpretations.

A geographic information system (GIS) can facilitate the utilization of more sophisticated definitions of regions that are not limited to existing political boundaries. Theoretically derived definitions of cultural regions can then be merged with public health data for analysis. Gastil (1) integrated information on historical settlement patterns, religion, economic activity, education, crime, and other factors to define cultural regions in the United States. His regional model is based on both historical and current information, which provides a stronger theoretical basis for the regions than is possible if the regions were based solely on either historical or current characteristics.

This study examines regional patterns of alcohol-specific mortality to provide an improved understanding of the variation of alcohol problems. Gastil's complexly defined regions are used to avoid problems with limited definitions and post hoc interpretations. The study is an extension of previous research on alcohol use and problems that indicates strong cultural influences on drinking (2,3), and regional differences in alcohol availability, consumption patterns, and problems (4,5).

Methods

The data used in the study were extracted from the 1990 US Census (6) and the 1986–1990 county level alcohol-related mortality tables published by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) (7). The NIAAA obtained the alcohol-related mortality data from the National Center for Health Statistics. The mortality data are based on five years of data to provide a stable estimate, especially for areas with limited total population. The age-adjusted mortality rate was calculated by dividing the number of deaths by the total population for that county and multiplying it by 100,000. For the age-adjusted rates, the number of alcohol-related deaths for each county was standardized by the county's reference age distribution to better illustrate the influence of factors other than age.

For this analysis, only the rates for causes of deaths explicitly mentioning alcohol were used to avoid confounding based on regional differences in the attributable fraction of alcohol-related diseases. Alcohol-explicit mortality includes 12 causes of death such as alcoholic psychosis, alcohol poisoning, alcohol dependence syndrome, and alcoholic cirrhosis of the liver. The census variables used in the study are total population, number of persons over 65 years of age, median household income, number of persons below the poverty line, number of males, number of blacks, and number of Hispanics. Percentages were calculated for age, poverty, male, black, and Hispanic.

The county boundaries for the continental United States were purchased as an ARC/INFO polygon coverage in unprojected geographic coordinates. The 12 regional boundaries were digitized from Gastil's *Cultural Regions of the United States* (1) and then projected into decimal degrees so that they could be overlaid successfully with the county boundaries. The regions were then overlaid with the counties so that each county obtained at least one regional identifier. For those counties that were divided by a regional boundary, the union allowed comparison of the amount of a county's land area that fell within each region, and the county was assigned to the region containing the most area.

The map displays were categorized by minimizing the sum of the variance within each grouping—a "natural breaks" method using Jenk's optimization. For the total population regional mortality map, the mean in each region was calculated by summing the population and number of age-adjusted cases for all counties assigned to that region, then dividing the total number of cases by the total population and multiplying by 100,000. Adjusted regional means were the estimated marginal means derived from an analysis of variance (ANOVA) of all the counties, using mortality rates as the dependent variable, region as the categorical factor, and four covariates (independent variables): percent over 65, median household income, percent male, and percent black.

Results

The raw alcohol mortality rates for the US counties are shown in Figure 1. Figure 1 also shows the boundaries of the cultural regions used in the study. Differences in the total geographic area of each county and in the number of counties in various regions makes it difficult to interpret the county-level data. Figure 2 shows the mean alcohol mortality rates for the various cultural regions. The rates shown in Figure 2 are based on total age-adjusted alcohol-explicit deaths and total population for each region. An ANOVA based on the counties assigned to each region indicated highly statistically significant differences between regions ($F=28.63$; $p<0.0001$). The visual impact of Figure 2 is striking, especially compared with Figure 1. Regional differences are readily apparent. Interpretation of the mortality rates shown in Figure 2 and Table 1 indicates that the Pacific Northwest, Pacific Southwest, Interior Southwest, New York metro area, and South have the highest mortality rates. The Central Midwest has notably lower alcohol mortality.

ANOVA with covariates was used to calculate adjusted means for the cultural regions. Factors such as race, age, income, and gender are known to be related to alcohol use and mortality (8). In addition, these factors also are related to cultural practices such as religion and economic activity. Figure 3 and Table 1 show the means adjusted for these factors. Age (percent over 65), race (percent black), gender (percent male), and median household income were all significant in the ANOVA. The two most influential covariates in the analysis are race and income. Race (percent black) is positively associated with alcohol mortality, while income is inversely associated with alcohol mortality. Note that additional analyses substituted percent Hispanic and percent in poverty for percent black and mean income without substantive differences in the results.

The adjusted means shown in Figure 3 and Table 1 still indicate significant and substantial differences between cultural regions. Regions high in alcohol mortality include the Pacific Northwest, Pacific Southwest, Interior Southwest, Rocky Mountain, and New York metro area. The main difference between Figures 2 and 3 is the lower adjusted mortality rate for the South. This suggests that black population and lower income population in the South accounts for a substantial portion of the mortality in this region.

Discussion

The results of this regional analysis of alcohol-explicit mortality show substantial differences between regions in the continental United States. The pattern changes when

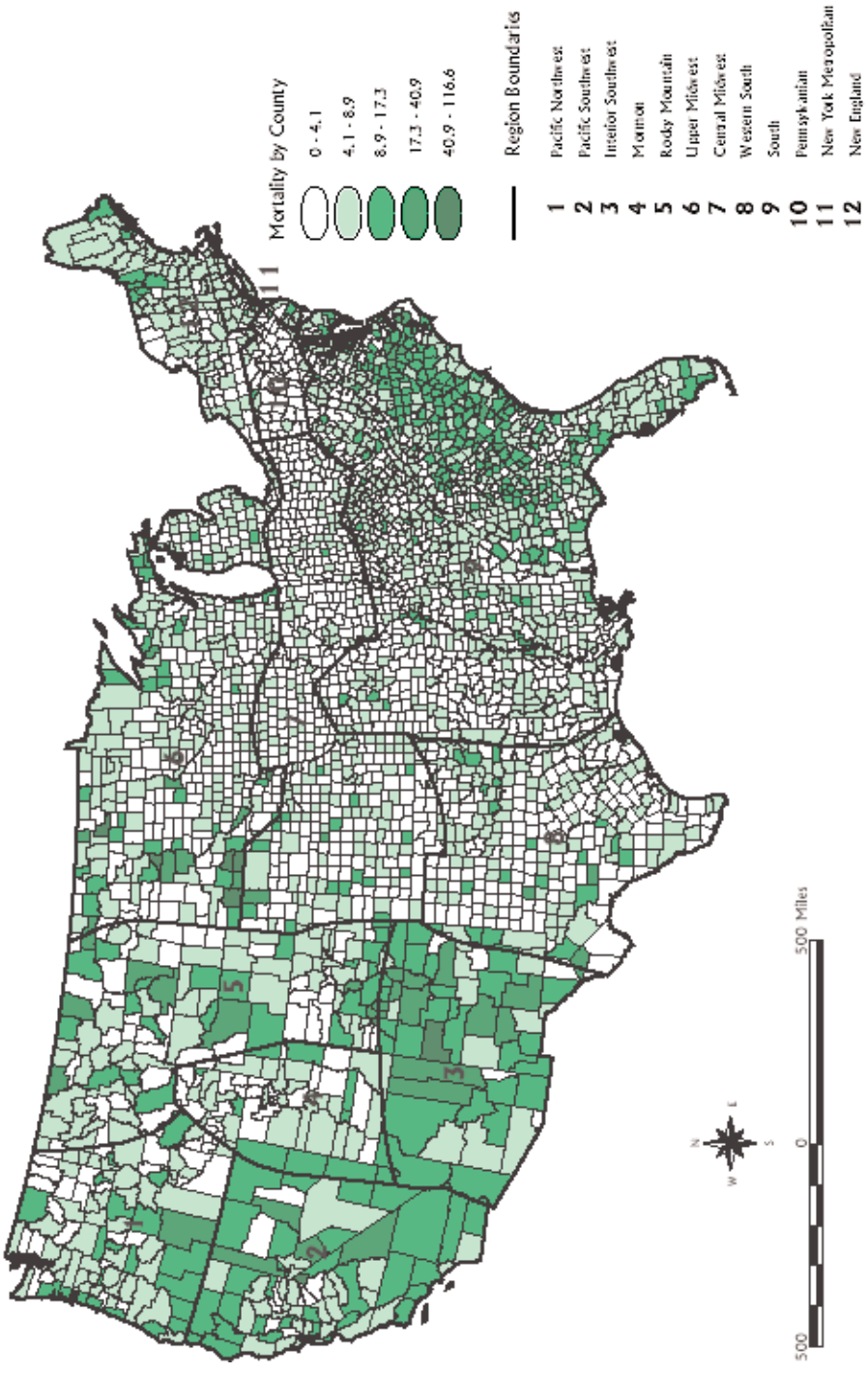


Figure 1 Cultural regions and alcohol-explicit mortality.

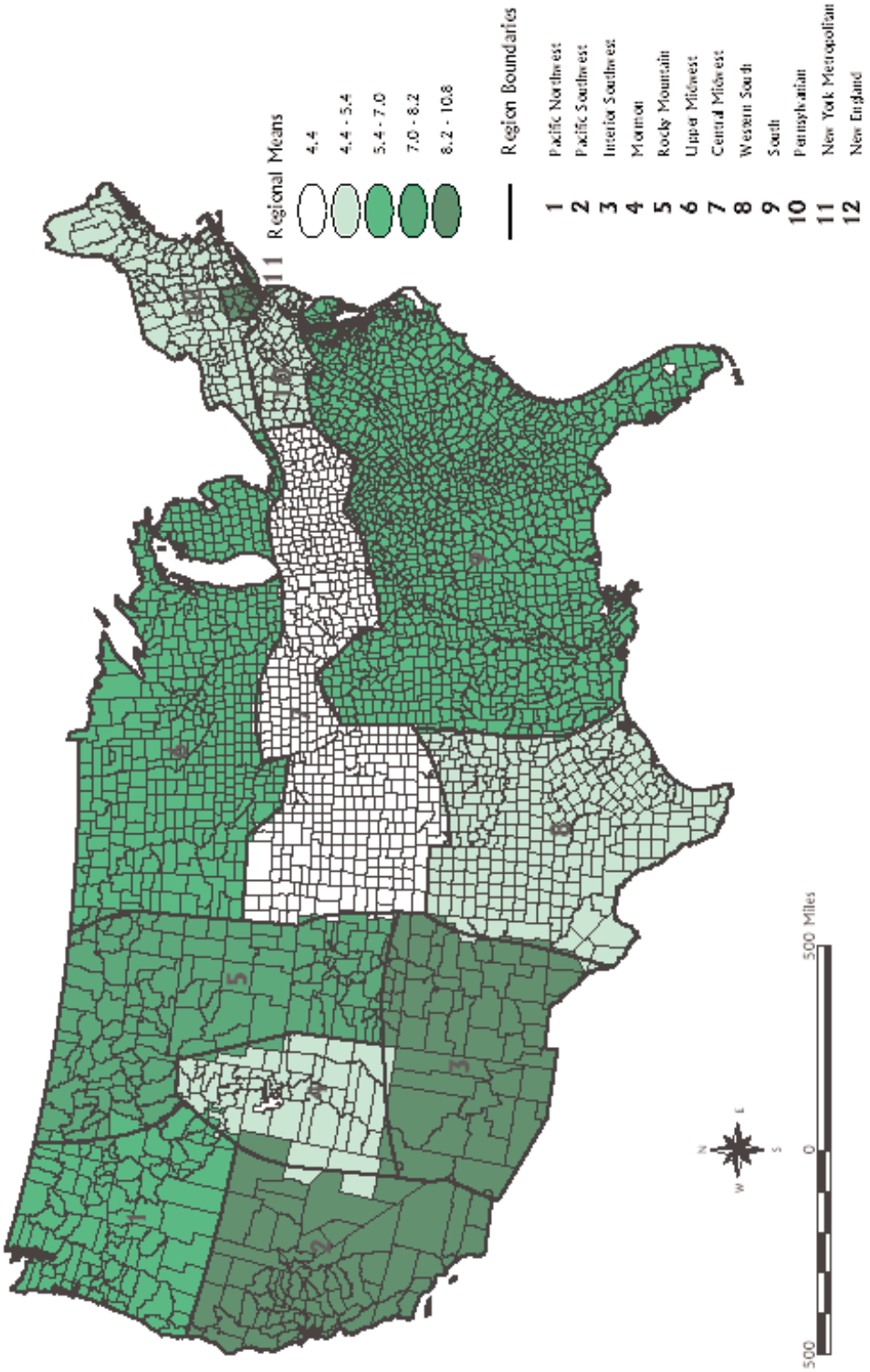


Figure 2 Regional means of alcohol-explicit mortality.

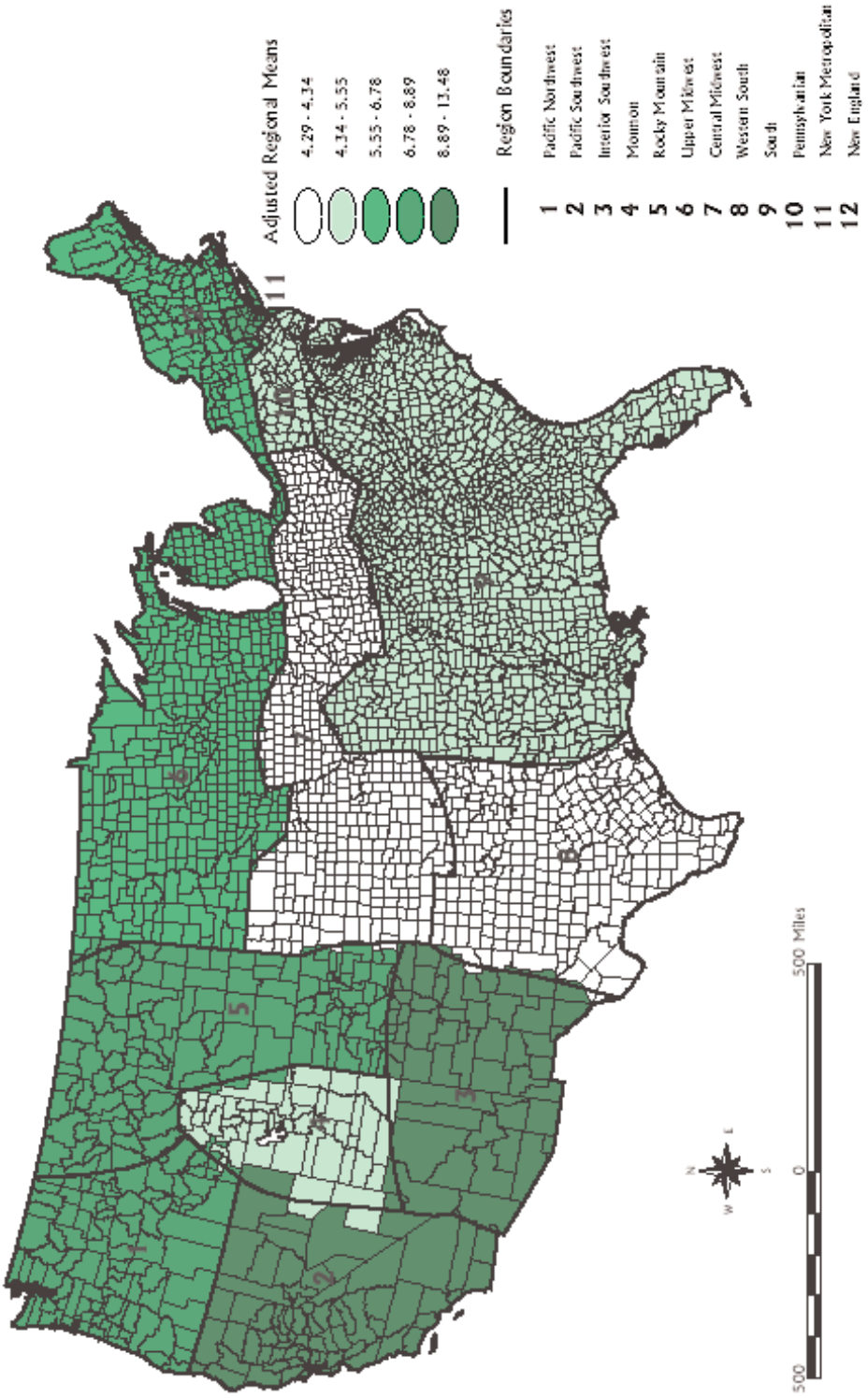


Figure 3 Adjusted regional means of alcohol-explicit mortality.

Table 1 Alcohol-Explicit Mortality Rate (per 100,000) by Region, United States

Region	Population Mortality Rate	Adjusted Mortality Rate	Adjusted Mortality (95% CI)	
			Lower	Upper
Pacific NW	7.0	7.5	6.5	8.6
Pacific SW	10.8	11.5	10.3	12.7
Interior SW	10.5	13.5	12.1	14.9
Mormon	5.1	5.5	4.1	6.9
Rocky Mt	8.2	8.0	7.1	8.9
Upper Midwest	7.0	6.8	6.3	7.3
Central Midwest	4.4	4.3	3.9	4.8
Western South	4.8	4.3	3.7	4.9
South	7.0	4.9	4.6	5.2
Pennsylvania	4.8	4.8	3.7	6.0
New York Metro	9.9	8.9	6.7	11.1
New England	5.4	6.5	5.6	7.4

covariates are controlled for, but the differences between regions remain quite notable. Clearly, the use of GIS to facilitate the regional analysis of public health data provides additional understanding and insight into public health issues, such as alcohol-related mortality.

The analysis provided evidence that alcohol mortality is particularly high in the western United States, outside of the Mormon region. These areas may require additional alcohol-focused interventions to lower the mortality. The New York metro area also appears to be in need of additional alcohol-focused public health interventions. Also notable, the initially high mortality rate in the South is explained by culturally related factors (i.e., percent black, median household income). This finding suggests that alcohol prevention and treatment efforts in the South should be targeted toward lower income populations and African Americans.

This study provides strong support for continued research on regional patterns of alcohol-related mortality. The identification of additional factors that explain regional differences may lead to further insights for interventions. Additional forms of alcohol-related mortality (e.g., specific types, attributable fractions, total alcohol mortality) should also be examined in future regional analyses.

Acknowledgments

This research was supported by grant P50 AA09802 from the National Institutes of Health, National Institute on Alcohol Abuse and Alcoholism.

References

1. Gastil RD. 1975. *Cultural regions of the United States*. Seattle: University of Washington Press.
2. McAndrew C, Edgerton R. 1969. *Drunken comportment*. Chicago: Aldine.

3. Kitano HHL, Chi I, Law CK, Lubben J, Rhee S-Y. 1988. Alcohol consumption of Japanese in Japan, Hawaii, and California. In: *Cultural influences and drinking patterns: A focus on hispanic and Japanese populations*. Research Monograph #19, US Department of Health and Human Services. 99–133.
4. Gruenewald P, Ponicki W. 1995. The relationship of alcohol sales to cirrhosis mortality. *Journal of Studies on Alcohol* 56:635–41.
5. Hilton ME. 1988. Regional diversity in United States drinking practices. *British Journal of Addiction* 83:519–32.
6. US Census Bureau. 1996. CensusCD, version 1.1. East Brunswick, NJ: Geolytics, Inc.
7. National Institute on Alcohol Abuse and Alcoholism (NIAAA). 1994. *County alcohol problem indicators, 1986–1990.* Rockville, MD: NIAAA. NIH Publication 94-3747.
8. NIAAA. 1997. *Ninth special report to Congress on alcohol and health.* Rockville, MD: US Dept. of Health and Human Services. NIH Publication 97-4017.

Warren County Landfill: Still Provocative After All These Years

PS Wittie (1,2),* B Nicholson (2)

(1) Department of Geography, University of North Carolina-Chapel Hill, Chapel Hill, NC; (2) North Carolina Superfund Section, Division of Waste Management, Department of Environment and Natural Resources, Raleigh, NC

Abstract

Protest over the burial of polychlorinated biphenyl (PCB)-contaminated soil in the Warren County PCB Landfill, a hazardous substance landfill near the small town of Afton, North Carolina, "jump-started" the environmental justice movement. A grassroots coalition of predominantly African-Americans in Warren County joined with national groups headed by the United Church of Christ's Commission for Racial Justice, the Southern Leadership Conference, and the Congressional Black Caucus to protest the development of a hazardous substance landfill and the decline of their neighborhoods. The decades-long resentment felt in minority communities over unfair siting practices, redlining practices, residential segregation, and other forms of discrimination had fueled the depth of concern in this community and galvanized it into activism. Two truck operators had illegally dumped 30,000 gallons of PCB-laced oil along a scattered 210-mile segment of roadways. (They went to prison for illegal dumping.) Roughly 32,000 cubic yards of soil contaminated with PCBs were removed from the rural roads and trucked to the landfill. After 16 years, concern within these communities that they may suffer from increased health risks continues. One question was unanswered: How do the sociodemographic profiles of the neighborhoods where the PCB-laced oil was dumped now compare with the community around the landfill? We conducted a geographic information system (GIS) study of 14 counties in north central North Carolina—the counties affected by the original dumping—to address this question. Using notes from sampling site documentation and hand-drawn maps, we created coverages of the formerly contaminated roadways and half-mile buffers to examine the demographic characteristics of these areas. The results do indeed show a higher concentration of minority population along the union of Nash, Halifax, and Warren Counties, but the other sampling sites show varying proximity to minority populations. While the roadside spill areas do show a strong concentration of minority neighborhoods, poverty was less concentrated than expected.

Keywords: environmental justice, demographics

Introduction

The Warren County (North Carolina) PCB Landfill has been controversial since its inception back in September 1982, when trucks began to deliver polychlorinated biphenyl (PCB)-contaminated soil to it.

* Peggy S Wittie, North Carolina Superfund, Division of Waste Management, Dept. of Environment and Natural Resources, 401 Oberlin Rd., Suite 150, Raleigh, NC 27605 USA; (p) 919-733-2801; (f) 919-733-4811; E-mail: pwittie@wastenot.enr.state.nc.us

Waste haulers, attempting to save the expense of legal disposal, had decided to mix their problematic chemicals with oil and spray the mixture on a 210-mile network of rural roads in 14 counties of North Carolina (Figure 1). The significance of this event is evident in its chronology. The contaminated soil and pavement was discovered on July 30, 1978, in a remote section of the Fort Bragg Military Reservation. Four days after the initial discovery, a laboratory confirmed the presence of PCBs. Sixteen days later, the governor asked the president to declare the 14 affected counties disaster areas.

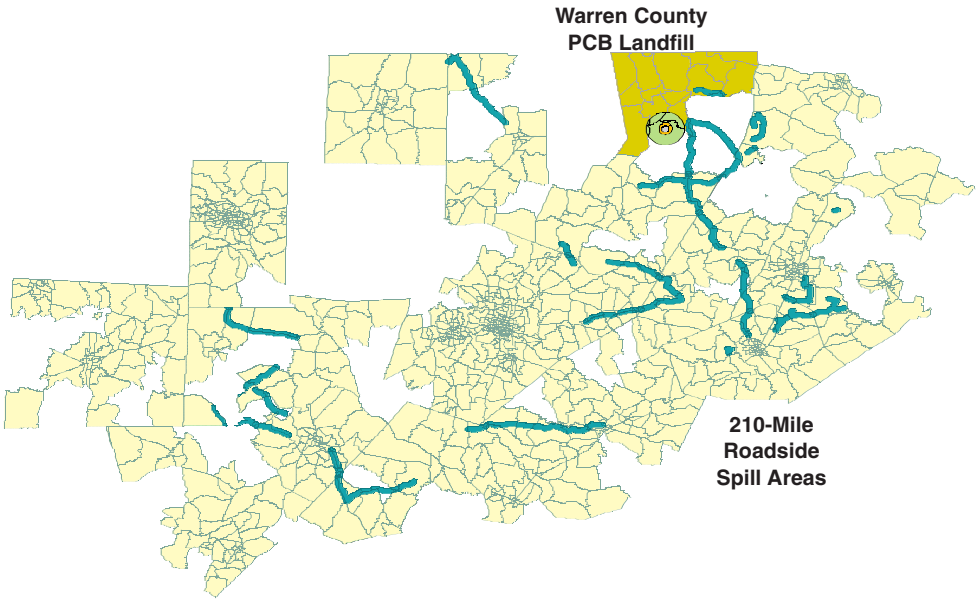


Figure 1 Buffers of the Warren County (NC) PCB Landfill and affected roadside spill areas.

A final decision to build a landfill in Warren County for the burial of the PCB-laced soil led to great controversy because the county was predominantly black. For the first time, nationally known civil rights activists and leaders joined local groups to protest the development of a hazardous substance landfill using a 1960s style of nonviolent civil disobedience.

This descriptive study is a direct reflection of the continuing controversy over the Warren County PCB Landfill. In early 1998, we conducted a demographic analysis both of the 210-mile roadside spill areas and of the landfill, using 1990 block group-level and 1980 county-level census data. Geographic information system (GIS) technology was vital to the delineation of these areas and to the examination of demographic characteristics. We sought to address three questions. First, do demographic characteristics vary when examined at the county level versus the block group level? Second, what is the minority and poverty composition of the populations in the spill areas? Third, how do the demographics of the roadside spill areas compare with those of the area surrounding the Warren County PCB Landfill?

The seminal report of the United Church of Christ's Commission for Racial Justice (UCCCRJ) (1) provides key research on the proximity of minority communities and

hazardous waste facilities across the country, as well as research on the larger environmental justice movement, which evolved from the debate and concern over the Warren County PCB Landfill. Using a zip code scale of analysis, this statistically based study finds no hazardous waste facilities in US communities with only 11% minority composition, one facility in a community with 24% minority composition, and two or more facilities in a community with 38% minority composition. Bullard's *Dumping in Dixie* (2) investigated the Warren County situation using the county as his scale of analysis, and reported that the county was predominantly black (63.7%) and poor (its median family income falls in the 92nd percentile for North Carolina). The US General Accounting Office (3) conducted a study of the socioeconomic and racial characteristics surrounding four hazardous waste facilities in the South and, because minority communities surround these facilities, found race to be a significant predictor for the presence of such facilities. Anderton et al. (4), using a census tract-level scale of analysis, examined the demographics around hazardous waste treatment, storage, and disposal facilities. Their findings show the percentage of population engaged in manufacturing to be a better predictor of hazardous waste facility presence than was race.

The probability of white migration out of urban areas increases as the proximity and percentage of minority population increases (5). Some researchers believe that the black middle class accompanies whites in their out-migration, further exacerbating the economic and spatial isolation of low-income minorities (6,7). According to Ottensmann (8), many studies have provided a threshold for the percentage of blacks in an area above which the racially mixed neighborhood would exhibit a strong tendency to transition to become predominantly black. This transition point has been discussed as being near 10% and approaching 40%. While racial segregation has decreased during the 1970s and 1980s, it still remains high (5,8,9). Likewise, economic and social disinvestment in minority communities is associated with increasing racial segregation (5,10). The aggregation intervals of 10–19%, 20–29%, and 30% and over are used to identify the effects of poverty and its isolation of minority communities (7,11,12).

The definition of a minority community varies in these studies as much as the scale of analysis used. Researchers often discuss minority communities in terms of how different neighborhoods are affected, either through poverty, isolation, or societal disinvestment. The findings of the UCCCRJ (1) and Massey and Denton (5) characterize minority communities by their proximity to potentially hazardous facilities and their increasing isolation from the amenities of the suburbs. Unlike William Julius Wilson's poverty categories (7) (which include the aggregation intervals identified in the previous paragraph), this method of describing minority communities is not well established. Consequently, we suggest merging concepts from the UCCCRJ and Massey and Denton studies to create the following categories:

1. Non-minority neighborhood: minority composition 0–14%.
2. Transitional zone: minority composition 15–24%.
3. Medium-high minority neighborhood: minority composition 25–34%; has a medium-high likelihood of white out-migration and of increasing proximity to a hazardous waste facility.
4. High minority neighborhood: minority composition 35% or higher, indicating a high propensity for two or more hazardous waste facilities.

Data and Methodology

For this project, the data consist of the 1990 Summary Tape File 3A (STF3A) from the US Census Bureau (13) and enhanced census boundary files from GDT's Dynamap/2000 (Geographic Data Technology, Lebanon, NH), which is a street network file for the state of North Carolina. Additional data include maps from sampling site information (14), paper-based USGS 7.5-minute quadrangles, and the County and City Data Book 1983 (15) for county-based 1980 census data. Windows NT-based GIS products, specifically ARC/INFO, ArcView, and ArcView Spatial Analyst (all from ESRI, Redlands, CA) were used to conduct the spatial analyses.

Using a GIS, we selected the road segments of the affected areas both through on-screen processes and through the use of structured queries. The road segments were developed as separate coverages, treated with half-mile buffers, and overlaid with block group boundaries. This process allows area calculations for the block group segments in the half-mile buffer and comparisons with the entire block group. The variables to be evaluated in these derived areas were percent nonwhite and percentage of persons living in poverty.

Two basic questions had to be addressed to evaluate the affected area. First, the characterization of minority communities needed meaningful aggregation ranges. For this study, the percentage ranges of 0–14%, 15–24%, 25–34%, and 35%+ appeared most relevant. Second, population estimates for these rural areas were needed.

Assuming a homogeneous distribution of population across each block group, the percentage of each census block group's area that was affected by roadside spills was calculated. Each block group's population count was then multiplied by this percentage-of-area number to derive estimates for the total affected population, total affected nonwhite population, and total number of affected persons living in poverty. (In the tables, the word "calculated" refers to this process of estimation.) All percentages were by dividing the calculated numerator by the calculated total population.

Two methods were used to test these estimates using regression analysis. The first was to select block groups randomly and to test the relationship between a block group's total population and its area. The second was to count each house in the half-mile buffer on USGS 7.5-minute quadrangles, whose published release dates range from 1967 through 1986. Seventeen block group segments were selected (with bias) on the basis of their high and low road density. Because most of these maps were dated closest to the 1980 census, the county-based 1980 persons-per-household statistic was multiplied by the house count in each block group segment. These multiplied counts were averaged and compared with the area contribution for each block group. These counts were also regressed against the area in each partial block group to test our surrogate measures. These measures were used in lieu of field-checking (physically counting houses within a half-mile of the affected roadside area) these segments on a random basis for house counts, which time in the initial phase of the project did not permit. To verify our use of the percentage of area as a multiplier, we required an R-square of 0.60 and a significance of $p \leq 0.20$.

Results

To characterize the segments, we sought a reasonable rural population surrogate. The

percentage of the sub-block group compared with the original block group proved to be the best indicator, with an R-square of 0.99 and a $p \leq 0.05$. The house count was not as useful, which, given the age of the maps, was not surprising.

Demographic characteristics are presented below for scale comparisons at the state, county, and block group level. Variations are evident, because political boundaries mask trends occurring at smaller aggregation units.

At the state level, the 1980 (15) and 1990 (13) decennial censuses both report the North Carolina population as 24% nonwhite. In 1980, the percentage of persons living in poverty was 12.5%; in 1990, it was 14.8%.

The county-based contributions to the overall sociodemographic characteristics show four counties—Chatham, Franklin, Nash, and Warren—each having 10% or more of the affected area (Table 1). Of these counties, Warren County has the greatest percentage of minority population (75.4%) and a medium-high poverty rate of 27.5%. Granville and Halifax Counties each have approximately 6% of the total affected area and a minority population base exceeding 50%. Franklin, Edgecombe, and Wake Counties each have minority populations exceeding 25%.

An examination of minority communities at the county level (Table 2) shows that approximately 50% of the affected area and population occur in a transitional zone (15–24.99% minority population). Communities with medium-high (25–34.99%) or high (35%+) minority composition each have an approximate 23% share of the affected area. The calculated population shows a different trend, but the highest representation (53% of the population) remains in the transitional zone. According to a county-level analysis, non-minority areas occupy less than 10% of the affected spill area, and compose approximately 10% of its population.

Table 1 Demographic Characteristics of the Roadside Spill Areas for Each County

County	Percent of Entire Affected Area (%)	Calculated Percent Nonwhite within Half-Mile Buffers (%)	Calculated Percent of Persons Living in Poverty within Half-Mile Buffers (%)
Chatham	16.4	20.77	12.52
Franklin	12.8	28.88	14.18
Nash	12.0	19.60	13.04
Warren	10.1	75.40	27.46
Harnett	7.6	20.18	14.69
Johnston	7.5	14.32	14.07
Wilson	6.6	19.40	12.00
Edgecombe	6.5	33.78	18.45
Granville	6.3	54.46	11.70
Halifax	6.0	92.03	36.78
Lee	4.2	20.77	10.54
Wake	3.9	28.88	10.04
Person	0.1	36.24	7.00
Moore	0.1	10.47	1.52
All counties	100.1		

Table 2 Minority Community Typology of Counties Affected by Roadside Spill

Minority Community Typology	Percent of Entire Affected Area (%)	Total Block Group Population[a] (n)	Calculated Total Population of Affected Areas (n)	Calculated Percent of Total Population of Affected Areas (%)	Counties
Non-minority (0–14.99%)	7.6	13,629	1,994	9.9	Johnston, Moore
Transitional (15–24.99%)	46.8	70,871	10,787	53.4	Chatham, Nash, Harnett, Wilson, Lee
Medium-high (25–34.99%)	23.2	40,527	5,347	26.5	Franklin, Edgecombe, Wake
High (35%+)	22.5	18,193	2,079	10.3	Warren, Granville, Halifax, Person
Total	100.1	143,220	20,207	100.1	

[a] 1990 US Census data (13)

When these areas are examined at the block group level, results differ (Table 3). Non-minority, transitional, and high-minority communities each occupy approximately 30% of the affected area; medium-high minority communities occupy approximately 12%. Roughly 38% of the area's total population is in the non-minority group, 27% in the transitional group, 9% in the medium-high group, and 26% in the high group.

When the same block group data are reorganized according to established poverty levels (Table 4), they follow the county-level trends. The low-poverty zone covers

Table 3 Minority Composition of Roadside Spill Area by Individual Block Group

Block Group Minority Community Typology	Percent of Entire Affected Area (%)	Calculated Total Population of Affected Areas (n)	Calculated Percent of Total Population of Affected Areas (%)	Percent Nonwhite of Entire Affected Area[a] (%)	Calculated Percent Nonwhite of Entire Affected Area (%)	Percent Below Poverty of Entire Affected Area[a] (%)	Calculated Percent Below Poverty of Entire Affected Area (%)
Non-minority (0–14.99%)	27.2	8,171	37.7	9.25	11.09	20.25	24.66
Transitional (15–24.99%)	29.2	5,879	27.2	15.64	17.75	20.03	23.46
Medium-high (25–34.99%)	11.7	1,988	9.2	9.03	10.43	10.44	9.13
High (35%+)	31.9	5,612	25.9	66.09	60.72	49.28	42.76
All block groups	100.0	21,650	100.0	100.01	99.99	100.00	100.01

[a] 1990 US Census Data (13)

Table 4 Percentage of Persons Living Below Federal Poverty Levels in Roadside Spill Area

Block Group Poverty Typology	Percent of Entire Affected Area (%)	Calculated Total Population of Entire Affected Area (n)	Calculated Percent of Total Population of Entire Affected Area (%)	Percent Nonwhite of Entire Affected Area[a] (%)	Calculated Percent Nonwhite of Entire Affected Area (%)	Percent Below Poverty of Entire Affected Area[a] (%)	Calculated Percent Below Poverty of Entire Affected Area (%)
Non-poverty (0–9.99%)	22.2	7,280	33.6	20.31	13.81	15.54	15.17
Low (10–19.99%)	54.8	10,392	48.0	34.24	40.61	42.87	47.80
Medium (20–29.99%)	10.2	1,996	9.2	24.24	20.70	21.75	15.42
High (30%+)	112.8	1,983	9.2	21.21	24.88	19.83	21.61
Total	100.0	100.0	100.0	100.00	100.00	100.00	100.00

[a] 1990 US Census Data (13)

roughly 55% of the entire spill area and contains 48% of the total population; 41% of all nonwhite persons in the spill area live in this zone. The non-poverty zone follows, with 22% of the total area and 34% of the total population.

Three points are evident when comparing the roadside spill areas with the landfill area. First, the population is sparse in the landfill's immediate area. We examined the population demographics using one-half-, one-, and three-mile buffers around the landfill (Table 5). Within the three-mile buffer, 779 people are estimated to reside, in a

Table 5 Demographic Characteristics in Buffers around Warren County PCB Landfill

Buffer	Calculated Total Population of Entire Affected Area (n)	Calculated Percent of Total Population of Entire Affected Area (%)	Percent Nonwhite of Entire Affected Area[a] (%)	Calculated Percent Nonwhite of Entire Affected Area (%)	Percent Below Poverty of Entire Affected Area[a] (%)	Calculated Percent Below Poverty of Entire Affected Area (%)
Half-mile	23	100.0	73.07	73.07 High minority	31.28	30.76 High poverty
One-mile	75	100.0	73.07	73.07 High minority	31.28	30.76 High poverty
Three-mile	779	100.0	69.60	71.14 High minority	27.16	27.68 Medium poverty

[a] 1900 US Census Data (13)

community whose population base is estimated to be 70% nonwhite and 28% living in poverty. The population estimates for the one-mile buffer and half-mile buffers are 75 and 23, respectively. These two smaller buffers both fall within a single block group that has a 73% minority population and 31% living in poverty.

Second, all the block groups around the landfill are in high-minority and medium-to-high poverty zones. The percentage of persons living in poverty and the percentage of nonwhite persons in the total population decline slightly between the half-mile and three-mile buffers.

Third, the difference in area between the landfill and the original 210-mile segment means that more communities are affected along the roadside spill areas. For instance, the half-mile buffered area around the landfill occupies only one block group, and is only 0.4% of the size of the original affected roadside spill region. The three-mile buffer crosses two block groups and is only 11% of the size of the original roadside spill region. The entire buffered roadside spill area intersects 111 block groups and is 80 times larger than the half-mile buffered landfill area.

In the affected spill areas, 32% of the block group portions have high minority populations (Table 6). An estimated 5,612 persons reside in these high-minority block group segments, which have 26% of the population of the entire affected area. These segments have 61% of all nonwhite persons and 43% of all the people living in poverty in the affected area. The correlative medium-to-high poverty groups occupy 23% of the entire affected area. In these zones, 18% of the entire area's population, or 3,979 persons, resides. The estimated percent nonwhite and percent below poverty in these zones are 46% and 37%, respectively.

Table 6 Minority and Poverty Characteristics in Buffers of the 210-Mile Roadside Spill Areas Area Typology along Affected Areas

Area Typology along Affected Areas	Percent of Entire Affected Area (%)	Calculated Total Population of Entire Affected Area (n)	Calculated Percent of Total Population of Entire Affected Area (%)	Calculated Percent Nonwhite of Entire Affected Area (%)	Calculated Percent Below Poverty of Entire Affected Area (%)
High minority	32	5,612	26	61	43
Medium-to-high poverty	23	3,979	18	46	37

Conclusions

The analysis shows a high representation of minorities and poor people along the 210-mile stretch of PCB-contaminated spill area. Not too surprisingly, differences do exist depending on the denominator or the scale chosen. While high minority representation occurs in the affected areas, poverty is not as strongly evident. The demographics of the roadside spill area show a greater variety of communities affected simply because the area involved is far greater than that of the landfill.

Central factors in any study surround the scale of analysis and the questions asked. Analysis of county-level data shows the highest proportion of the area to be in the

transitional zone. The block group-level analysis shows a different pattern, with an almost equivalent distribution in the non-minority, transitional, and high-minority groups. The removal of county boundaries provides a clearer idea of the minority representation of the affected area.

An examination of minority and poverty zones in the original roadside spill areas shows a higher correlation between spill area and minority areas than between spill area and poverty. High-minority zones account for only approximately one-third of the entire area, but they contain one-fourth of the calculated total affected population and two-thirds of the nonwhite population. The medium-to-high poverty zones show fewer people being impacted. They constitute approximately one-fourth of the affected block groups, less than one-fifth of the total population, and more than one-third of all those living in poverty in the affected area.

The landfill occupies a relatively minimal area compared with the affected roadside spill area, which makes a realistic comparison difficult. The single block group surrounding the landfill is predominantly nonwhite (73%) with medium-to-high poverty. The roadside spill cleanup area shows high-minority communities composing approximately 32% of the 111 affected block groups across an area 80 times the size of the half-mile buffer of the Warren County PCB Landfill.

A non-homogeneous population distribution within a given block group may introduce error. Population tends to follow roadways, and the density of roads within the buffer will vary. In rural areas, houses flank roadways and are found within approximately 200 feet of the road. The issues raised by non-homogeneous distribution and varying road density will be addressed in further GIS and statistical analyses. The population estimates need further evaluation, especially when areas under investigation do not conform to typical reporting units.

The delineation of minority communities using the merged theories from the urban underclass and environmental justice debate presents a useful typology. However, the urban underclass and racial segregation theories are based upon studies of urban areas, not poor rural areas in the South. Further research should examine this linkage to further test its validity.

While the primary idea of the study was to see if PCB-contaminated soil was removed from non-minority zones and dumped in high minority areas, the results show that a range of communities was affected across the 14-county area of the original spill.

The assumption of increasing isolation and decreased political power does underlie the theoretical base of environmental equity studies, but the political landscape has been changing over the years. Indeed, the controversy over the Warren County PCB Landfill helped to change the political dialogue about the siting process. Questions that previously had not been asked now have become standard. The question of who lives where is now part of the process.

References

1. United Church of Christ Commission for Racial Justice. 1987. *Toxic wastes and race in the United States.* New York: United Church of Christ.
2. Bullard RD. 1990. *Dumping in Dixie: Race, class, and environmental quality.* Boulder: Westview Press.

3. US General Accounting Office. 1986. *Hazardous waste: EPA has made limited progress in determining the wastes to be regulated.* December 1986. Washington, DC: GAO/RCED-877-27.
4. Anderton DL, Anderson AB, Oakes JM, Fraser MR. 1994. Environmental equity: The demographics of dumping. *Demography* 31(2):229–48.
5. Massey DS, Denton NA. 1993. *American apartheid: Segregation and the making of the underclass.* Cambridge, MA: Harvard University Press.
6. Auletta K. 1982. *The underclass.* New York: Random House.
7. Wilson WJ. 1987a. *The truly disadvantaged: The inner city, the underclass, and public policy.* Chicago: University of Chicago Press.
8. Ottensmann JR. 1995. Requiem for the tipping-point hypothesis. *Journal of Planning Literature* 10(2):131–42.
9. Galster GC. 1990. White flight from racially integrated neighborhoods in the 1970s: The Cleveland experience. *Urban Studies* 27(3):385–99.
10. Kasarda JD. 1990. Structural factors affecting the location and timing of underclass growth. *Urban Geography* 19:21–40.
11. Johnson JH Jr, Oliver ML. 1991. Economic restructuring and black male joblessness in the US metropolitan areas. *Urban Geography* 12:542–62.
12. Johnson JH Jr, Oliver ML. 1990. Modeling urban underclass behaviors: Theoretical considerations. *Occasional Working Paper Series* 1(2).
13. US Department of Commerce. 1990. *Decennial Census of Population and Housing, Summary Tape File 3A (STF3A).* US Census Bureau. Washington, DC: US Government Printing Office.
14. State of North Carolina. 1980. Final Environmental Impact Statement, Removal and Disposal of Soils Contaminated with PCBs along Highway Shoulders in North Carolina. November 3. Raleigh, NC: DEHNR.
15. US Department of Commerce. 1983. *County and city data book 1983.* US Census Bureau. 10th ed. Washington, DC: US Government Printing Office.