Chapter 8

Spatial Analysis of Disease - Applications

B. Sue Bell
National Cancer Institute

1. INTRODUCTION

The objective of this chapter is to provide useful information for taking the spatial analysis of health data "from the lab to the clinic." The preceding chapter reviewed the history and theory of spatial statistics as applied to health data. This chapter provides examples of how this theory can be used in practice. Emphasis is placed on the tools and resources available to enable a statistical analyst to perform a spatial statistical analysis. Because the methods and software are constantly improving, the author advises the reader to review the latest literature as a first step in embarking on a spatial analysis.

Often spatial statistics is exploratory and descriptive but can be inferential. As with all statistics, first identify your objective. What question are you trying to answer? The data available and the research objective will dictate the method.

After providing general background information on software and data considerations, this chapter will present four examples of spatial analyses. The methods applied will include spatial filtering, cluster identification using the spatial scan statistic, a hierarchical analysis, and an analysis using the conditional autoregression (CAR) model.

Evaluation of a possible disease cluster around a putative source was introduced in the preceding chapter in Section 2.4.3 and Section 6.3 and the production of an Atlas was discussed in Section 2.3. This chapter does not expand on these two topics.

1.1 Software

This chapter focuses on presenting methods supported by software generally available in the fall of 2000. The programs may be commercially available or may be in the public domain and available for free download. The use of any particular program is not intended to be a recommendation for that program, but merely reflects the software available to this analyst.

Whenever alternative software is identified, the list is not intended to be exhaustive. Thanks to the competition among the developers of the statistical theory, between geoscientists and statistical scientists, and between developers of geographic information systems (GIS) and developers of statistical software, software for spatial analysis will be a dynamic area for the foreseeable future.

1.1.1 Mapping Software

When performing a spatial analysis, mapping software is essential. Consider the mapping software a critical tool in the statistician's graphical toolbox. As regression analysis often begins with a histogram to show a variable's distribution, spatial analysis begins with observing a variable's spatial distribution and the final result of the analysis is also often a map.

There are four main sources for mapping software. First, there are commercially available geographic information systems (GIS) such as MapInfo Professional® (MapInfo Corporation 2000), Maptitude (Caliper Corporation 2000) and Environmental Systems Research Institute's (ESRI) ArcView® GIS (ESRI 2000). This type of complete GIS software will be the most expensive option, but a GIS will have the most advanced facilities for displaying geographic data and will include functions based on years of geographical research. ESRI offers a Spatial Analyst extension that includes geostatistical functions such as kriging. Second, there is commercially available mapping software that is not as sophisticated as the GIS software so is less expensive. For example, ESRI currently offers ArcExplorer as a free product for GIS data viewing. Third, some statistical software vendors (e.g., SAS and S-PLUS) include basic mapping functions as either a feature or extension of their software. S-PLUS also offers S+ Spatial Stats and S-PLUS® for ArcView GIS® as two add-on products to provide an environment specialized to spatial analysis and integrated with ArcView GIS®. The fourth option is to use one of the publicly available programs. Two programs available from the U.S. government for free downloads are LandView® from the U.S. Census Bureau and Epi Map from the Centers for Disease Control and Prevention (CDC).

1.1.1.1 LandView®

LandView® is a desktop mapping system developed by the U.S. government that can be used on standard personal computers. It includes database extracts from the Environmental Protection Agency, the Bureau of the Census, The U.S. Geological Survey, the Nuclear Regulatory Commission, the Department of Transportation, and the Federal Emergency Management Agency. These databases are presented in a geographic context on maps that show jurisdictional boundaries, detailed networks of roads, rivers, and railroads, census block group and tract polygons, schools, hospitals, churches, cemeteries, airports, dams, and other landmark features. Importantly, data can be both imported into and exported from LandView®.

The U.S. Census Bureau web site http://landview.census.gov/ provides links for acquiring either LandView® IV or LandView® III. The web site provides documentation, tutorials, and purchase information. LandView® IV was released in November 2000 with a single DVD containing the data for the entire U.S. LandView® III software is available for download free on the web site. Data can also be downloaded one county at a time from the web site.

1.1.1.2 Epi Map 2000

The Epidemiology Program Office of the Centers for Disease Control and Prevention (CDC) provides free mapping software in conjunction with Epi Info 2000. Epi Map 2000 is built around the MapObjects program from ESRI, the producers of ArcView®. Epi Map is compatible with GIS data from numerous Internet sites in the ESRI formats. Epi Map is designed to show data from Epi Info 2000 files by relating data fields to SHAPE files containing the geographic boundaries. Numeric data can be displayed either as color/pattern (choropleth) maps or as dot density maps. Visit the web site http://www.cdc.gov/epiinfo/ for links to download the software and for tutorials and manuals.

1.2 Data

1.2.1 Spatial Sampling

When spatial distribution is important to the research question, then space should be considered in developing the sampling plan. The environmental and agricultural sciences have a long history of sampling in space (e.g., air quality monitoring stations, soil composition samples). Likewise, economists use systematic time points (e.g., monthly consumer price index) in their time series analyses. As we investigate the spatial distribution of cancer or other diseases, systematic spatial sampling methods

should be considered when it is important that the sample represents the population of interest across space.

Consistent with any sampling plan, a spatial sampling plan considers the population to be represented and the sampling frame of units within the population available to be sampled. The difference is that now the location in space is also important leading to a map frame. The sampling unit may be points that need to be systematically placed such as air pollution monitoring stations. The map frame may consist of preexisting areas or polygons, referred to as areal units. Examples of areal units include Census areas (e.g., block groups and tracts) and political areas (e.g., states and counties).

Often data are made available from predetermined sampling locations or from vital records, but occasionally the statistician may have the opportunity to design a spatial sampling plan and to choose the framework for data collection. Stehman and Overton (1996) present a number of approaches to spatial sampling. Likewise, Haining (1990) examines the problems that arise in sampling a surface. Beyond presenting sampling designs, Haining discusses how survey work in the social sciences traditionally involves areal stratification to control for important social and economic characteristics while the choice of units within the strata is made at random. The map, both of the individual units and the strata, is considered of minor importance; but Haining suggests that perhaps it should not be. National surveys consider it more important to adjust for the effects of a stratified or clustered sampling design than to consider problems of spatial dependency. Haining (1990 pages 192-193) concludes that although systematic sampling may prove impractical or too costly in certain cases, the theoretical evidence stresses the superiority of systematic sampling in a variety of spatial situations such as soil maps and land use. He argues that sampling of social, health, and economic data is not that different from sampling of environmental data where spatial sampling methods have evolved.

Methods for obtaining a point sample from a continuous universe that is spatially well distributed (i.e. population of interest is not tightly clustered in space) include stratified, cluster, and systematic sampling. The population can be stratified geographically and a sample taken within each strata. Spatial information may be used to form clusters and then either one-stage or two-stage cluster sampling is possible. However, when there are spatial components stated in a study's objectives, regular patterns of points or plots are usually favored.

First, prepare either a square grid with a square tessellation (like a chess board) or a triangular grid with a hexagonal tessellation (i.e., uses hexagons instead of squares for the mosaic) that will cover the study area (Stehman and Overton 1996). Second, fix the tessellation randomly on the surface to

be sampled. Finally, randomize the location of the site in the first cell and allocate points in the remaining cells to the same relative position.

Tessellation-stratified sampling is generally more efficient than unrestricted random sampling for most surfaces likely to be found in practice (Stehman and Overton 1996). Random samples will tend to over sample from the denser areas while failing to have any samples from the less dense areas. In a county with both urban and rural areas, a random sample will most likely include few subjects from the less densely populated rural areas compromising an analysis of a spatial distribution over the entire county.

Areal sampling may also be used to sample continuous populations. In the traditional areal sampling design, the continuous spatial universe is partitioned into areal units that are moderately uniform in size. A sample from this frame of areal units is taken to form an areal sample. Because of the spatial context, the units for the areal sample are identified in a systematic way. For instance, a regular point grid is overlaid on the areal units and those units that include a point from the grid become one of the sample areal units. The grid is coarse enough so that it would be unlikely for two or more points of the grid to fall in a single unit.

Spatial sampling theory depicts space as a continuously varying surface. The sample must provide coverage of the surface to ensure that the surface mean and its variance can be estimated.

1.2.2 Geocoding

Spatial analysis requires the association of each observation with its geographic location. Translating the street address to its longitude and latitude co-ordinates makes this association. Geocoding is the term for the process and is often easier said than done.

Bias can be introduced when the addresses that are unsuccessfully geocoded are not missing at random. For instance, rural addresses are less likely to be successfully geocoded so that a disproportionate number of subjects living in rural areas could be missing from the analysis. If the address file is several years old, addresses in newer subdivisions will not be matched. Post office box addresses can only be geocoded to the 5-digit zip code causing error in the point location assignment. Cross streets are sometimes used as a surrogate for street address because of privacy concerns, but these can be very imprecise in rural areas.

The analysis plan and project budget must address the issue of geocoding misclassification and include an approach that is appropriate for the study question. If the study question requires precise point estimates then one option is to use a global positioning system (GPS) to physically pinpoint locations that are not successfully address matched. Another option is to

purchase the most up to date address files from a geocoding service company. If the study question does not require pinpoint accuracy, then less precise areal geocoding (e.g., using the centroid of the zip code or the county of residence) may be acceptable.

So, how do you actually geocode? It requires a GIS that will perform geocoding, an address file for matching against, and the addresses to be matched. In addition to the commercial GIS packages, LandView® IV has some limited geocoding capabilities. Address master files may be purchased from GIS software vendors or from a service bureau such as GDT (www.geographic.com). The next issue for the do-it-yourself geocoder is assessing how clean is your address file. A considerable amount of time can be spent in reformatting addresses to fit the standard expected for matching to the address master file. If this is the case, then it may be worthwhile to purchase automatic address correction software such as ZP4 from Semaphore Corporation (www.semaphorecorp.com).

Sound like a lot of work? Another option is to use a service bureau such as Tele Atlas (www.etak.com) or GDT (www.geographic.com). You can send the service bureau your file to be address matched for processing in batch mode or you can use their interactive web-based interface to geocode the records yourself. The advantage of the latter is that you can use their address correction software and their latest address master file.

Whether done in-house or by a service company, an automated match rate of about 90% is considered typical (Zakos-Feliberti 2000). Just as with subjects who are lost to follow-up in a longitudinal study, the project must assess what effort is to be expended to complete the information for the subjects without an address match. This could range from making phone calls to verify addresses to actually going to the location and using a GPS.

1.2.3 Confidentiality

State and Federal law (e.g., Public Health Service Act (Section 308 (d))) requires that personal health information collected by Public Health Officials remain confidential. This applies to data from vital records (e.g., births, deaths) and from health surveys (e.g., National Health Interview Survey, Behavioral Risk Factor Surveillance System). The Department of Health and Human Services (DHHS) is in the process of developing medical privacy rules that may further restrict access to personal health information (DHHS 1999).

DHHS omits all direct identifiers, as well as any characteristics that might lead to identification, from their data sets. Naturally, this includes location information such as address. The National Center for Health

Statistics (NCHS) requires researchers to work onsite with its files that contain individual identifiers, and the researcher can leave with only summary statistics.

The common practice is to work with aggregated data, but even aggregated data may be suppressed. One example can be taken from the restrictions applied to data retrieved from the Compressed Mortality File (CMF) on CDC's WONDER on the Web. When the request is for a county's mortality data for a single year, the counts and rates are suppressed when the single-year count is less than or equal to five for counties with a total population that is less than 100,000. When the data is aggregated over three or more years, there is no suppression of small counts even when the count is less than five.

Assuming that the data available for spatial analysis has the subject's exact location geocoded, the presentation of that information in maps must ensure the subject's confidentiality. One approach is to add a random error to the longitude and latitude before display on a dot map. This jittering of the location is documented for users of the map and the jittering must be sufficient to ensure privacy. More traditional is the presentation of spatial statistics in the form of a choropleth or isopleth map so that individual locations are never mapped for presentation.

Armstrong, Rushton, and Zimmerman describe geographical masks that they feel, when appropriately used, protect the confidentiality of health records while permitting many important geographically-based analyses (Armstrong et al. 1999). They explore transformation-masking methods, aggregation-masking methods, nearest-neighbor masks, and the replacement of geographic identifiers with contextual information of specific interest to the data user.

2. FROM POINT DATA TO RATE MAPS

2.1 Methods

2.1.1 Areal Aggregation

Because of confidentiality concerns, the most commonly produced, reported and mapped spatial statistic is a rate for a predefined geographic area (e.g,. county, census tract). These rates could be calculated as crude rates, as standardized mortality/morbidity ratios (SMR), as direct adjusted rates, or as rates predicted by statistical models. However, it is not wise to

map either a crude rate or an SMR. Section 4 in the preceding chapter shows how these rates are calculated.

Direct adjusted rates (Pickle and White 1995) or model-based rates are preferred because these methods can adjust for important confounding variables and produce comparable estimates. Consider a choropleth map of states where the color-coding is based on the rank of a state's rate in five categories. The first thing the map-reader will usually do is look at his home state and read out his state's category. This would be fine, but his next step is to look at another state of interest and read out the category for that state for comparison with the first. It is inappropriate to compare either crude rates or SMRs from one area to another area because the disease risk of the underlying population is not comparable. For instance, if the risk of disease is associated with age, as is usually the case with cancer, then unless the two areas being compared have the exact same age distribution, then one cannot compare the crude rates for the two areas or the SMRs for the two areas.

If rates based upon aggregation to areal units (e.g., counties) are to be mapped, then the rate can be calculated using direct adjustment for at least the most important confounding variable(s) which is most likely age. The Atlas of United States Mortality (Pickle et al. 1996) (also available at NCHS web site http://www.cdc.gov/nchs/products/pubs/pubd/other/atlas/atlas.htm) includes a choropleth map of the age-adjusted death rates by Health Service Area (HSA) for each cause of death for White Male, Black Male, White Female, and Black Female. In this way, the confounding variables of age, race, and sex have been considered. The map legend includes not only the rate ranges for the categories but also the range of the ratios comparing the HSA rate and the U.S. rate. These comparative mortality ratios assist the map-reader in evaluating how different a rate for a HSA is from the U.S. overall rate.

One additional consideration in mapping rates is the stability of the area rates. Some areas may have unusually high or low rates based on very small numbers. Several approaches have been used to inform the map-reader about the reliability of an area's rate. The Atlas of United States Mortality (Pickle et al. 1996) employed double hatching with parallel white and black lines to identify areas with sparse data. The Atlas of Cancer Mortality in the United States: 1950-94 (Devesa et al. 1999) used a separate category for counties with sparse data and used the color gray on the maps thereby suppressing all rate information for those counties.

2.1.2 Spatial Filtering

When point data is available, but it is aggregated to political or administratively defined geographic areas (i.e. areal units), spatial

information is lost. This spatial information could be particularly important for local public health officials concerned with identifying local "hot spots".

As discussed in Section 5 of the preceding chapter, spatial filtering (smoothing) is a non-parametric analysis method within the field of exploratory spatial analysis. Spatial filtering produces spatial density estimates based on health events that have been observed at individual locations. Examples of health events include births, deaths, and incident cases. A form of data smoothing, spatial filters reduce variability in a data set while retaining the local features. By varying the size of the filter, features in the data that vary at different spatial scales can be differentially removed. Spatial filtering is a useful technique for identifying areas that have higher or lower values than generally occur.

Rushton drew from his background in geography to develop a spatial filtering method and software for application to spatial analysis of health events. He demonstrated the application on birth defect rates (Rushton and Lolonis 1996) and infant mortality rates (Rushton et al. 1996) in Des Moines, Iowa.

The process consists of the following steps:

a) geocode the numerator events and denominator events.
b) use a grid to locate grid points uniformly across the study area.
c) calculate the distance from each grid point to all health events within a maximum filter area represented by a circle around the grid point.
d) choose a filter size less than or equal to the maximum filter size and using events whose distance to the grid point is less than the radius of the filter calculate observed rates for each grid point by dividing the count of numerator events in the filter area by the count of denominator events in the filter area.
e) use GIS software to create contour or spatial density maps of the rates.
f) recalculate the rates and reproduce the map using a larger filter size for a smoother surface or smaller filter size for a surface with more texture.

In addition, Monte Carlo simulations are used to test the observed rates for statistical significance so statistical significance can also be mapped. A weight can be provided for each event. If data has been aggregated to a geographical unit, then the weight could be the count of the events within the aggregation unit and the location could be the centroid for the geographic unit.

The Distance Mapping and Analysis Program (DMAP) will create the grid points and will calculate rates and statistical significance. It is available for downloading free from the University of Iowa's GIS in Public Health web site (http://www.uiowa.edu/~geog/health/). DMAP can also be

requested on CD-ROM along with other valuable information provided on the web site. GIS software is needed for geocoding and for displaying the results as a map.

2.2 Example

2.2.1 Data

The primary data file that will be used in this chapter to demonstrate the spatial analysis processes contains birth events rather than cancer events. However, the types of processes and methods are not health event specific but relate to the research question to be answered and the type of data available to address the question and can therefore be applied to cancer as well as other health events.

The source of the health event data was the Community Health Information System (CHIS) developed in partnership by the University of Texas-Houston School of Public Health and the St. Luke's Episcopal Health Charities. The CHIS integrates health and well-being indicators and measures on an interactive web site that allows users to examine community health profiles, conduct population health assessments, and link to community resources. Their web site is http://www.slehc.org. In support of CHIS, the project team is geocoding all the vital birth and death records for Texas from 1990 forward.

The research question was "Does the birth outcome of intrauterine growth retardation (IUGR) co-vary ecologically with characteristics at the neighborhood level?" IUGR (Usher and McLean 1969; Frisbie et al. 1997) is an outcome measure to categorize births that are small for gestational age. The project team analyzed geocoded birth records for Harris County (Houston), Texas for 1991.

The maps in the following sections all focus on the central urban area of the county. The major highways are included to provide reference points for comparing the maps across sections and methods.

2.2.2 Methods

The first objective was to assess whether there were geographic areas of unusually high rates of IUGR births in Harris County. The methods used were the same as were used for analyzing birth defects (Rushton and Lolonis 1996) and infant mortality (Rushton et al. 1996) in Des Moines, Iowa.

First, the data files were created in the formats required by DMAP for each of the numerator events, denominator events, and the probability file.

The data should be viewed and verified in a text editor before trying to import it into DMAP.

Next, a regular lattice of grid points was arbitrarily located at approximately half-mile intervals. DMAP produced the grid file to cover a square area defined by longitude and latitude coordinates in decimal degrees for the upper left and lower right corners of the square. ArcView® GIS (ESRI 2000) was used to identify these corners for entry into DMAP. Because Harris County is not square, there were points in the grid outside of the county. Also, there are large industrial complexes, flood control reservoirs, and other uninhabited areas that technically should not be included. Grid points in these areas could be removed by using the capability of a GIS to use one layer as a pattern to select features in another layer. For example, by overlaying a layer with surface water on the layer with the grid points, then only grid points not in the water could be selected.

After importing the three data files into DMAP, everything is in place to start the process of computing rates for each grid point. Selecting a filter size that will sufficiently smooth the data to reveal patterns without over smoothing is an iterative process of trying a filter size and then increasing or decreasing the filter size to adjust the amount of smoothing. In DMAP, the first step is to specify a radius for a maximum filter size of interest so that the computer intensive process of calculating and storing the distance from each grid point to all events within that maximum filter distance can be done once. Two additional entries are required before actually calculating rates. One is a filter size for this iteration of smoothing. The other is a minimum number of denominator events required before a rate is calculated for a grid point. Because the distances from grid points to events have already been calculated, scenarios of differing filter sizes and differing minimum denominator counts can be processed quickly.

Processing DMAP for all of Harris County was pushing the capacity of a powerful personal computer. In particular, calculation of the distances and probabilities took hours. As discussed above, there are sizeable areas in Harris County that are not inhabited. The project team decided to focus on the central metropolitan area.

The last step in DMAP was to calculate the statistical significance of the rate at each grid point using 1000 Monte Carlo simulations. Again, the processing time was reduced considerably (from hours to minutes) when the number of grid points was reduced.

Finally, ArcView® and ArcView® Spatial Analyst were used to map the DMAP results. In working with ArcView Spatial Analyst, it is important that the analyst be aware of any projections in use for any of the feature themes or grid themes. Projections lead to different views of the same area. For example, one common projection of the continental U.S. preserves the area

in each state resulting in the curved boundary with Canada while a different projection has the boundary as a straight line. If a grid in decimal degrees is mixed incorrectly with a theme projection, the data will not align properly and integrated analysis cannot be performed.

In the example, the coordinates for both the geocoded birth events and the grid points created in DMAP were in decimal degrees. Further, the analysis was conducted on a large-scale, that is one that covers a small area (Meade et al. 1988), making theme projections unnecessary.

2.2.3 Results

Figure 1 was produced in basically two steps. The first step was to use DMAP to calculate rates using spatial filtering and the second step was to use ArcView® GIS and Spatial Analyst to prepare a surface map of those rates. The specifications in DMAP were to use grid points at one-half mile intervals, a spatial filter of one-half mile, and a minimum of 40 births for a rate to be calculated at a grid point. After importing the DMAP results into ArcView®, a smoothed surface of the rates was created using ArcView Spatial Analyst's inverse distance weighted interpolator. This smoothing function only assumes that the influence of one point on another diminishes with distance.

The filtering and mapping were really an iterative process that drew upon local knowledge to produce results that made sense for the local topology. Because of the large area covered by Harris County, there was an attempt to increase the grid and filter to one-mile instead of the half-mile used for Des Moines. Those results were so smoothed that there appeared to be births in large industrial complexes and in green space reserves. Returning to the half-mile grid and filter size retained some of these local features. For instance, the white areas on the east is a major industrial complex, on the north-east is a reservoir, and on the near west side bordering a highway is a large park.

These maps show that the higher rates of IUGR births in 1991 were occurring along a north-south corridor and in an area in the northwest. Figure 2 is a map of areas with the most economic disadvantaged census tracts shown in black. Note the obvious similarities with the patterns of Figure 1.

Figure 1. Map based on spatial filtering using a half-mile grid and a half-mile filter

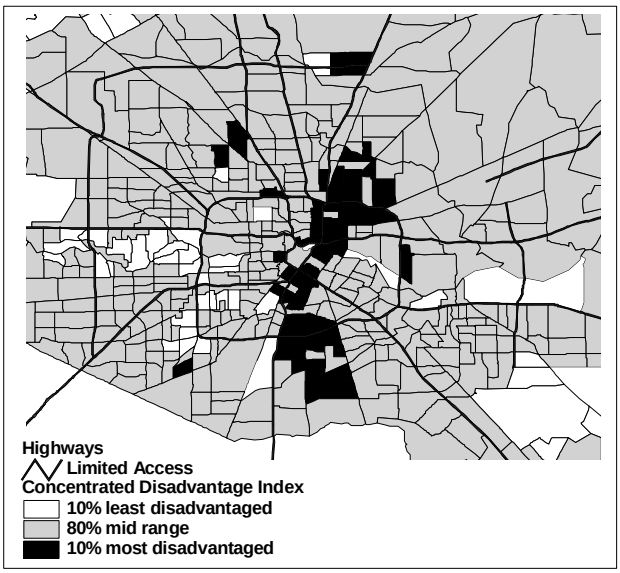

Figure 2. Map highlighting economically disadvantaged census tracts

2.2.4 Discussion

The maps reflecting the results of spatial filtering with DMAP are consistent with our hypothesis that IUGR births are co-varying with neighborhood level characteristics. Local characteristics of the data were retained by working with the individually geocoded records and by adjusting the smoothing parameters. Confidentiality is maintained by only calculating rates in areas with sufficient denominator data and by presenting only the smoothed maps.

Spatial filtering can be a valuable tool for exploring the spatial distribution of cases in relation to persons at risk. For Public Health, this can be particularly valuable in justifying potential target locations for focusing scarce intervention resources. However, because spatial filtering does not include the capacity to adjust for potential confounding variables, its application is limited to producing descriptive maps that must be prepared and interpreted with care.

3. DETECTING AREAS OF SIGNIFICANTLY HIGH OR LOW RATES

3.1 Objective

Epidemiologists and Public Health professionals often need to identify clusters of areas of unusually high or low risk for further study or possible intervention. Because comparisons will be made between the areas, is it imperative that the rates or risk estimates be adjusted for such potential confounding variables as age, race/ethnicity, and sex.

3.2 Methods

3.2.1 Spatial Scan Statistic

The spatial scan statistic (Kulldorff and Nagarwalla 1995; Kulldorff 1997) was developed to support a method of detection and inference for spatial clusters of disease. The proposed test can detect clusters of any size, located anywhere in the study region.

The SaTScan software implements the spatial scan statistic and can be used to analyze spatial, temporal and space-time point data. It is designed for any of the following interrelated purposes:

- To evaluate reported spatial or space-time disease clusters, to determine if they are statistically significant.
- To test whether a disease is randomly distributed over space or over time or over space and time.
- To perform geographical surveillance of disease, to detect areas of significantly high or low rates.

The outcome variable for common health events is often either dichotomous such as in case-control data or a count such as the number of cases among a population at risk in a geographic area. SaTScan allows the user to specify either a binomial distribution as would be appropriate for case-control data or a Poisson distribution as would be appropriate for count data. In addition, the program adjusts for the underlying heterogeneity of a background population. With the Poisson model, SaTScan can also adjust for any number of categorical covariates.

The software and documentation for SaTScan are available for downloading free from within the National Cancer Institute's (NCI) web site. SaTScan is currently sponsored by the Statistical Research and Applications Branch within the NCI Division of Cancer Control and Population Science's Surveillance Research Program. The current web site is http://srab.cancer.gov/othersoft.html where there are links to software developed specifically for cancer surveillance. The programs run in a Microsoft Windows® environment.

3.3 Example

Continuing with the data on IUGR births, SaTScan is used to identify clusters of census block groups with statistically high risk of IUGR births.

3.3.1 Data

The birth records included information on the mother's age and her race/ethnicity that can be important predictors of low birth weight (Showstack et al. 1984; Frisbie et al. 1997; O'Campo et al. 1997). The mother's age was categorized into the following three levels: less than 20 years of age, 20 to 34 years of age, and 35 years of age or older. The race/ethnicity was categorized as Anglo, African American, Mexican American, and Asian American. Births to other races were excluded from this cluster analysis because there were very sparse.

3.3.2 Methods

Although SaTScan will work with individual observations, we aggregated the data into block group counts to ensure the confidentiality of the subjects and also for efficiency of processing. There were over 50,000 births compared with 2,015 block groups. As with DMAP, the records must be formatted exactly as the SaTScan program requires. The cases were the IUGR births aggregated to Census block group and for age group and race/ethnic group within the block groups. The population file for the Poisson model was the corresponding aggregation of all births. Also, a coordinates file with the longitude and latitude of the centroids for the block groups was exported from LandView® III.

The analysis assumed that the IUGR births were Poisson distributed in space and 4,999 Monte Carlo replications were conducted. Although the maximum spatial cluster size can be as large as 50% of the population, the maximum spatial cluster size for this analysis was limited to 10% of the population consistent with an objective of identifying neighborhoods that are at higher risk. The cluster results were imported into ArcView® GIS for geographic display.

3.3.3 Results

Figure 3 illustrates the impact on the identification of potential clusters when adjustment for potentially confounding variables is included in the analysis. Figure 3a presents the results from SaTScan when no adjustment for potential confounding variables was made in the analysis. Note that the pattern is very similar to the pattern from spatial filtering in Figure 1 and also the pattern of the economically disadvantaged census tracts in Figure 2. The cluster with the darkest shading in Figure 3a is the primary cluster. The primary cluster has a relative risk estimate of 2.0 and is highly statistically significant ($p = 0.0002$). The secondary clusters in Figure 3a are also highly statistically significant and have relative risk estimates ranging from 1.5 to 1.9. It has been shown that simulated p values for secondary clusters are conservative, i.e., they overestimate their true values (Kulldorff 1997).

Figure 3b and 3c show the impact on the analysis of including the race/ethnicity of the mother and of including the age group of the mother in addition to her race/ethnicity, respectively. The relative risk estimates shrink toward the null value of 1.0 and the significance level is reduced ($p < 0.10$). The two clusters that appear in Figure 3b are located in economically depressed areas consistent with our research question of whether IUGR co-varies ecologically with characteristics at the neighborhood level. The inner city area cluster persists after adjustment for the age of the mother while the

secondary cluster to the east in Figure 3c that surrounds a large industrial complex is no longer statistically significant.

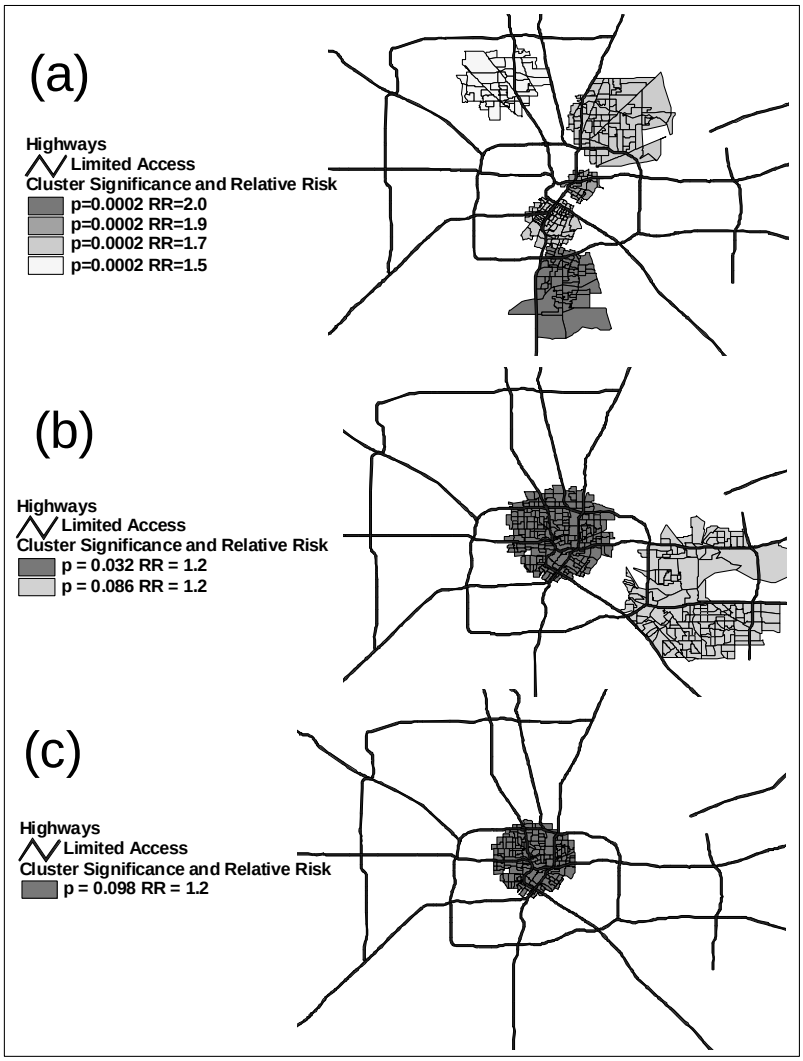

Figure 3. Comparison of clustering results from SaTScan with (a) no adjustment for confounding variables, (b) adjustment for race/ethnicity of the mother, and (c) adjustment for race/ethnicity and age of mother. (Note: first cluster in each group with the darkest shading is the primary cluster.)

3.3.4 Discussion

The spatial scan statistic is preferred over spatial filtering for identifying areas of high or low risk when there are important confounding variables for which the risk estimate must be adjusted. Also, the spatial scan statistic preserves the overall statistical significance in its test for the primary spatial cluster.

While the application of the spatial scan statistic produced results that allowed us to explore the spatial variability that remained after adjusting for individual level confounding variables, it did not answer the question of whether IUGR has spatial variability beyond these individual characteristics. To answer such questions requires the next step of spatial modeling.

4. SPATIAL MODELING

4.1 Objective

The preceding methods, spatial filtering and the spatial scan statistic, aid in exploring potential spatial patterns in disease. To answer the question of what is influencing the spatial pattern or to better predict rates requires the next step, spatial modeling.

Unlike the observations in a traditional statistical analysis that are assumed to be independent and randomly distributed, the observations in a spatial analysis are assumed to have spatial dependency and to be systematically distributed.

Time-series modeling extended statistical modeling by considering observations that are dependent along the single dimension of time. Spatial modeling further expands statistical modeling by considering observations that are dependent in the two dimensions of space. Research into spatio-temporal models is extending the methods even further (Waller et al. 1997; Xia and Carlin 1998).

For this chapter, two approaches to spatial modeling will be explored. The first is based on the hierarchical model (a.k.a. multilevel model, random effects model, or mixed model) and the second is based on the conditional autoregression (CAR) model.

4.2 Hierarchical modeling

4.2.1 Background

Section 7.3 in the preceding chapter introduces the theory involved in adding a random effect to a fixed effects model. The addition of the random effect leads to a hierarchical model. Briefly, hierarchical models have been used to stabilize rate estimates for mapping purposes (Clayton and Kaldor 1987; Manton et al. 1989). The spatial associations are defined by a hierarchy often based on political boundaries such as counties within states. The objective is to "borrow strength" from the parent level and from neighbors in the same region to achieve local rate stabilization without losing geographic resolution. As discussed in the preceding chapter, a hierarchical random effects model was used to model rates for health service areas (HSA) within regions for *The Atlas of United States Mortality* (Pickle et al. 1996).

Initially, theory and software for hierarchical models focused on continuous outcome variables. Parameter estimation in hierarchical generalized linear models is more complicated than in hierarchical linear models. Nevertheless, the theory and software have been developed to support the fitting of additional types of outcome variables enabling hierarchical logistic regression and hierarchical Poisson regression. Refer to journal articles by Breslow and Clayton (Breslow and Clayton 1993) and by Goldstein and Rasbash (Goldstein and Rasbash 1996) for more on the marginal quasi-likelihood (MQL) and the penalized or predictive quasi-likelihood (PQL) methods of estimation. PQL has been implemented in MLwiN, HLM, VARCL, and SAS® GLIMMIX macro (Littell et al. 1996) software. SAS® version 8 introduced the NLMIXED Procedure for fitting nonlinear and generalized mixed models. In addition, MLwiN includes approximations by Gibbs sampling and HLM includes Laplace approximation (Raudenbush et al. 2000).

4.2.2 Example

Continuing with the data on IUGR births, MLwiN will be used to fit a hierarchical binary response model. The modeling objective is to assess the influence of neighborhood level factors on the occurrence of IUGR births after controlling for individual characteristics of the mother.

4.2.2.1 Data

At level 1 of the model are the birth observations that includes both information on the baby (e.g., gestational age and weight) and information on the mother (e.g., her age, marital status, race/ethnicity, prenatal care information, and residence at the time of the birth). At level 2 of the model is information on the neighborhood of the mother's residence at the time of the birth. The Census tract is used for level 2 and attributes of the tract include tract indices for economic disadvantage, residential stability and segregation among others.

4.2.2.2 Methods

It is always wise to start simple, so the first step was to analyze the data using standard logistic regression methods available in a familiar statistical software package. The parameter estimates for these simple, fixed effects models should not be very different when the data is fit using the hierarchical logistic model. The difference will be evident in the variance and covariance of the random effects and their effect on the statistical significance of the predictor variables.

The next step was to import the data into the hierarchical modeling software, MLwiN for this example. As with most independent software, the data had to be prepared for successful import and processing in MLwiN. With several of the hierarchical linear modeling programs it is recommended that identifiers for the observations that are long integers (e.g., social security number is nine digits) be renumbered to smaller integers and that continuous variables be centered and standardized. After several iterations of dealing with the idiosyncrasies of export results (e.g., factor variables and missing values) and import expectations, it was easier to export the data to a spreadsheet and use the spreadsheet program's capabilities to prepare the data. Then a copy from the spreadsheet data window and a paste into the MLwiN data window completed the data transfer.

The regression analysis process is then replicated in MLwiN starting with univariate analysis and building up to the full model. The hierarchical results were compared with the results from fitting the standard logistic regression model to verify that the hierarchical model was producing comparable estimates for fixed effect parameters. During the fitting process there are controls that allow the analyst to choose the fitting algorithm and to control for overdispersion. Once again, the recommendation is to start simple ensuring that there is both convergence and a reasonable fit on the data. Then the fitting process can continue using the increasingly sophisticated methods that should correct for overdispersion and improve the estimates of the variances and the covariances.

4.2.2.3 Results

With over 50,000 level 1 observations nested within over 500 level 2 observations, MLwiN successfully converged with both the MQL and PQL estimation algorithms and provided parameter estimates that were consistent with those produced using traditional logistic regression software.

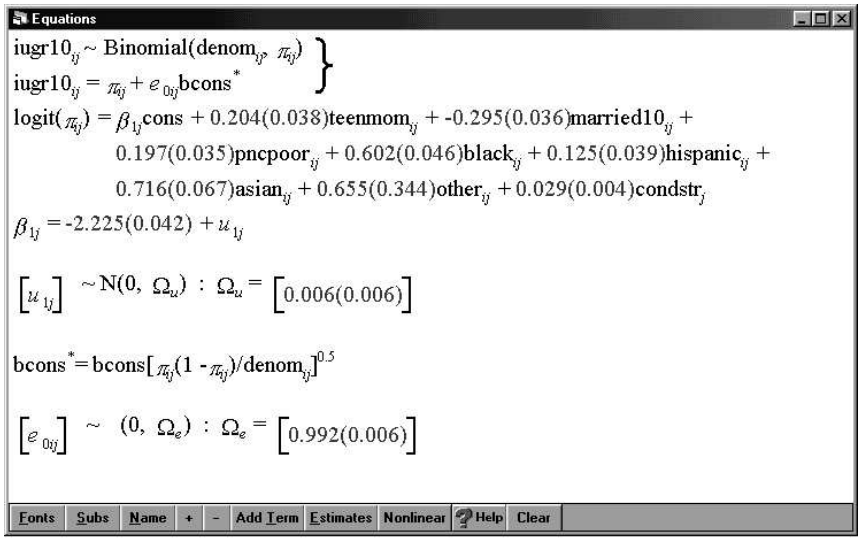

Figure 4. Hierarchical logistic regression model fit by MLwiN

Figure 4 presents the MLwiN equation screen after specifying and fitting a hierarchical logistics regression model. The outcome variable $iugr10_{ij}$ is assumed to follow a binomial distribution. It was coded as 1 if the birth was intrauterine growth retarded and was coded as 0 otherwise. A logit link function is specified with a random intercept (β_{1j}). Fixed effects at level 1 are individual characteristics of the mother and are identifiable by the ij subscripts. The mother's characteristics include her race and whether she was a teenager, was married, or had poor prenatal care. This model includes a neighborhood characteristic at level 2 that is an index of economic disadvantage for the census tract ($condstr_j$ in the model). Using a Wald test statistic computed by dividing the estimated coefficient by its standard error and assuming a normal (0,1), or Z, distribution for large samples, the neighborhood characteristic economic disadvantage is statistically significant after controlling for characteristics of the mother. The Wald test statistic for economic disadvantage is 7.25 (0.029/0.004) for a p value less than 0.0001. To better understand the syntax, refer to the MLwiN user's guide (Rasbash et al. 2000).

MLwiN provides a command interface, an interactive graphical user interface, and a macro capability. The interactive interface made it easy to expand and refit the models using the different algorithms. The windows are updated with parameter estimates and iteration counts during the process so that progress can be monitored.

4.2.3 Discussion

The algorithms and the software being used to fit generalized linear mixed models are still evolving. The methods include numerical integration as implemented in MIXOR, Laplace approximation as implemented in HLM version 5, MQL and PQL implemented in most of the available software, and Gibbs sampling implemented in MLwiN. All the methods tend to produce consistent results for the fixed parameter estimates but the estimates for the random effects can be different. Snijders and Bosker (1999) provide a general overview of these methods and discuss their relative merits in their introduction to multilevel logistic regression.

One shortcoming of both the MQL and PQL estimation methods is that their deviance statistics cannot be used to compare two nested models (Snijders and Bosker 1999 page 218) like the log-likelihood statistic is used to compare two nested logistic models. Also, there can be convergence issues depending upon the data and the complexity of the model.

4.3 Conditional Autoregression (CAR) models

4.3.1 Background

Section 7.4 of the preceding chapter discussed modeling spatial dependence where the spatial autocorrelation was modeled as a function of distance. It also introduced the conditional autoregression (CAR) model that is based on the work of Besag (Besag 1974) that has led to advances in imaging technologies. Other related terms for the CAR model include Gaussian Markov random field (MRF) (Clayton and Bernardinelli 1992), conditionally specified Gaussian (Cressie 1993), autoGaussian (Besag 1974), and Gaussian intrinsic autoregression (Besag et al. 1991).

The CAR model provides a dimension in defining the spatial autocorrelation structure that goes beyond distance based functions. It incorporates the concept of spatial neighbors where the definition of neighbor is left to the analyst. Equation 1 depicts the distribution of the single parameter CAR model

$$Z_i \mid Z_{j \neq i} \sim Gau(X'\beta, (I - \rho W)^{-1} D\sigma^2) \quad (1)$$

where W is a weighted neighbor matrix, D is a diagonal matrix used to account for nonhomogeneous variance of the marginal distributions and the parameters to be estimated are the β's, ρ, and σ^2. The analyst defines neighbors depending on the context of the problem. Options include defining neighbors as adjoining areas or as areas within a predefined distance. When the objective is to "borrow strength" to improve small area estimates, areas with similar relevant characteristics could be defined as neighbors even though they are distant geographically. This cannot be done using the models based on geostatistics that only work with distance based autocorrelation functions.

There are few software packages currently available that will estimate the parameters of the CAR model. Bayes Using Gibbs Sampling (BUGS) software (Spiegelhalter et al. 1995) can be used to apply the CAR model. An example included in their documentation fits lip cancer data from Scotland (Spiegelhalter et al. 1996). Another is S+ Spatial Stats (Kaluzny et al. 1998) that will be used for this example.

4.3.2 Example

An advantage of the hierarchical logistic model was that it allowed the individual characteristics of the mother to be included in the model where the outcome variable was dichotomous. A shortcoming was that its spatial description in MLwiN was limited to a hierarchical association (e.g., birth event within census tract). An alternative to the hierarchical association is to fit a model that incorporates a spatial autocorrelation structure. Unfortunately, the commercially available software at this time only fit models with spatial autocorrelation when the dependent variable is continuous, not dichotomous.

Before the CAR model could be fit to the IUGR birth data, the individual level birth outcome data was aggregated to areal rates resulting in loss of information about the individual births. Therefore, the analysis was ecological (i.e. at a group-level) introducing the potential for ecological bias as discussed in Section 2.4.1 of the preceding chapter. While there has been some theoretical work on conditional autoregression binomial (a.k.a. autobinomial) models (Besag 1974; Cressie 1993), this author is unaware of their implementation in practice as of this writing.

The software used for this spatial analysis included S-PLUS, S+ Spatial Stats, S-PLUS for ArcView GIS, ArcView GIS, and ArcView Spatial Analyst. The integrated environment of the statistical package and the GIS made the iterative process of modeling and mapping much less tedious than

if the data had to be continually reformatted and copied between the GIS and the statistical software.

4.3.2.1 Data

Continuing with the data for the IUGR births, the individual level data was aggregated to a census tract level for analysis. The dependent variable was the proportion of births that were IUGR. Likewise, the important individual level predictor variables were also aggregated to a proportion for the tract.

4.3.2.2 Methods

The process used to conduct a spatial analysis is not unlike the process used to analyze a time series. For a detailed example, see Cressie's analysis of sudden infant deaths (SIDS) in North Carolina in section 6.2 of his book titled *Statistics for Spatial Data* (Cressie 1993).

Before analysis, the dependent variable was transformed to remove the mean-variance dependence. Consistent with Cressie's analysis of the SIDS data, a Freeman-Tukey (FT) square-root transformation (Cressie and Chan 1989) was used to calculate a more stable dependent variable. Cressie and Read (1989) compared the FT transformation with other potential transformations and showed that the FT transformation was more stable. The FT transformation, where R_i is the number of IUGR births in the i^{th} tract and n_i is the number of births in the i^{th} tract, is calculated as follows:

$$Z_i \equiv (1000(R_i)/n_i)^{1/2} + (1000(R_i+1)/n_i)^{1/2} \quad (2)$$

After transforming the outcome variable, the first analysis step was to explore the data including mapping the crude rates and producing a probability map to see if there appeared to be clustering of areas with statistically high rates. The probabilities of observing the number of IUGR births in a Census tract given the number of births in the Census tract and the rate of IUGR births in the county were calculated using Monte Carlo simulations based on a Poisson distribution. The probabilities were then mapped using a choropleth map style.

Next, the distribution of the transformed rates was reviewed for outliers. It was decided to exclude the 15 tracts that had no births and the 15 tracts with fewer than six births because these tracts with small denominators produced some extremely skewed values even after transformation. Viewing the location of these tracts on the map showed tracts that often coincided with industrial complexes and undeveloped areas in the county.

Before testing for spatial autocorrelation or modeling, it was necessary to define the matrix of spatial neighbors. For this analysis, the assumption was

that if two tracts were adjacent (i.e. share any common boundary) then they were assumed to be neighbors. The integrated environment of ArcView GIS and S+ Spatial Stats proved its worth here. A shape file for the tracts in the county was already in ArcView®. To create the neighbor file for the 546 tracts to be included in the analysis was literally a few clicks of the mouse and a few moments of processing. The only problem arose when the shape file had a second entry for one tract. This second entry was uninformative, having zero area, so it was removed using ArcView's clipping capability. Care must be taken when there are legitimate islands or other non-contiguous areas as is the case with the North Carolina SIDS data. Because the model results can be sensitive to the selection of neighbors and neighbor weights, it is recommended that a sensitivity analysis be conducted to see if the results are robust with respect to the choice of the neighbor matrix.

Besides using the subjective reading of a map to assess whether the outcome variable has spatial autocorrelation and warrants the fitting of a spatial model, a variogram can be fit and plotted. There are also two tests for spatial autocorrelation, one using the Moran statistic and the other the Geary statistic.

Now the modeling took on a familiar process. The first model was an intercept only model. Next univariate models of possibly important explanatory variables were fit. Then a multivariate model was built by adding terms that were statistically significant in the univariate model. Nested models were compared using the same process as the likelihood ratio test (LRT) used in logistic regression. Residual plots and graphs were used to assess the fit. Maps of the residuals and the fitted values were prepared to visually assess the progress in explaining the spatial correlation (Figure 5).

4.3.2.3 Results

A plot of a variogram (see Figure 6a) based on the Freeman-Tukey transformed rates as well as statistical tests using the Moran and Geary statistics all indicated that there was spatial autocorrelation in the rates. The fitting of a CAR model was therefore warranted by the data.

Figure 5 provides map results for the modeling process. Figure 5a is a map of the Freeman-Tukey transformed rates and shows possible spatial clustering in the central and south-central parts of the county. A map of the Poisson probabilities (Figure 5b) also highlights tracts in the central and south-central parts of the county. Figure 5c maps the fitted values. Note that higher rate areas are more clustered in Figure 5c after modeling than in the map of the original rates in Figure 5a. Figure 5d is a map of the residuals. The random pattern in the residual map and the lack of apparent clustering indicates progress in modeling the spatial correlation.

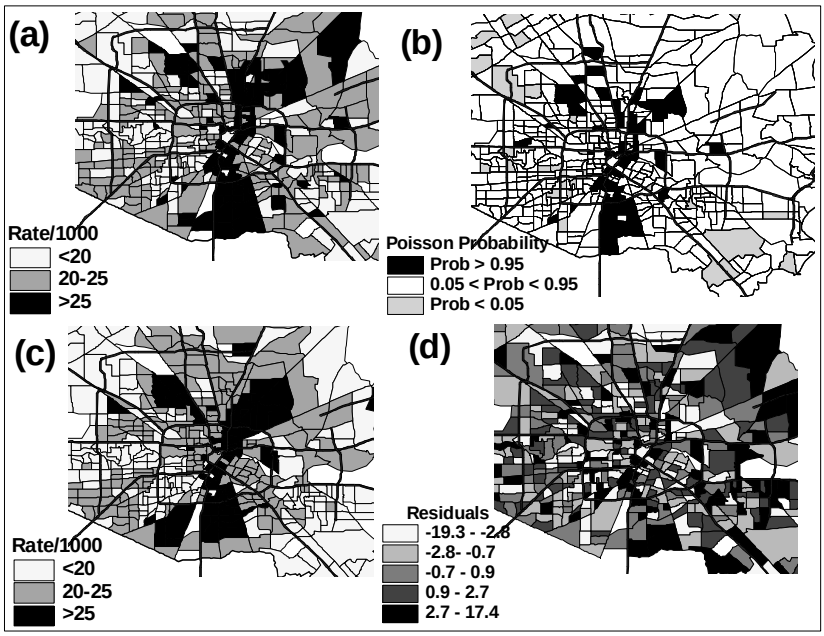

Figure 5. Maps of rates, probabilities and residuals (a) rate before spatial modeling (b) map of statistically high rates and statistically low rates assuming rates have a Poisson distribution with constant mean, (c) map of predicted rates fit by a CAR spatial model, (d) map of residuals after fit by a CAR spatial model.

Unlike the results from the hierarchical model, the index of economic disadvantage was not statistically significant when added to a model containing the aggregated information on the characteristics of mothers residing in the tract. The index of economic disadvantage was statistically significant in its univariate model. The statistically important variables included in the model to date include the proportions of mothers in each tract who were unmarried, who were black, and who had inadequate prenatal care.

The variogram of the residuals in Figure 6b shows that the CAR model has accounted for the spatial autocorrelation. Nevertheless, a quantile-quantile plot (QQ Plot) in Figure 6d that compares the quantiles of the residuals from the full model with the quantiles of the normal cumulative probability distribution function shows that the model is still not fitting the highest and lowest values well. This lack of fits in the tails indicates a need for additional explanatory variables to improve the overall model fit.

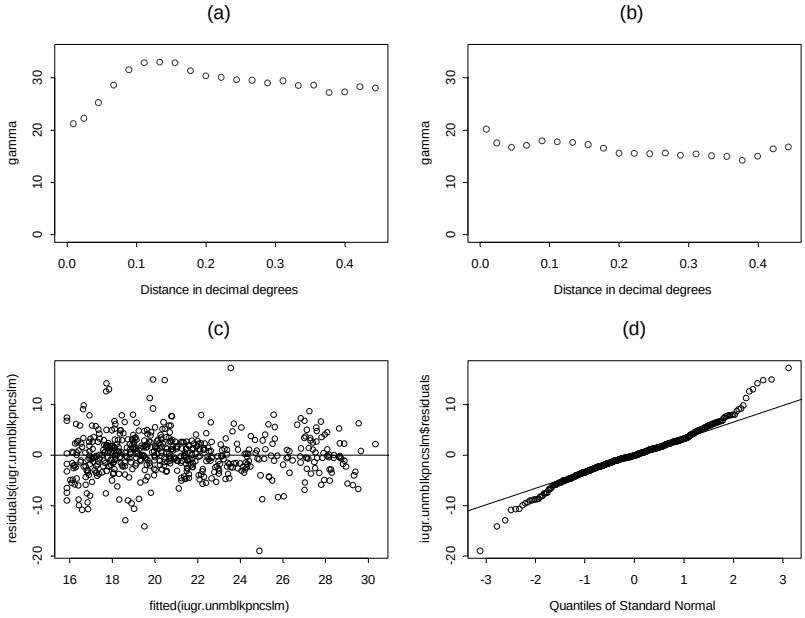

Figure 6. Plots to assess fit of the CAR model. (a) variogram of Freeman-Tukey transformed rates, (b) variogram of residuals after fitting a CAR model, (c) plot of fitted rates versus residuals, and (d) QQ plot of residuals

4.3.3 Discussion

It is now possible for the CAR model to be moved from the lab to the clinic. The tedious process of defining the neighbor matrix is now automated using GIS data. Faster computers and improved computer algorithms make the fitting process almost instantaneous. At least that was the experience on this example where 546 census tracts were included in the model.

The drawback is that to be really effective required the integrated environment of ArcView GIS, ArcView Spatial Analyst, S-PLUS and S+ Spatial Stats installed on a personal computer with the capacity to simultaneously run these applications. These are among the more expensive GIS and statistical software packages.

The hierarchical logistic model in Section 4.2 allowed modeling of individual level data but limited spatial information to the hierarchical association of observations within areal units. In contrast, the CAR model provided a means of better addressing the spatial autocorrelation but current implementations require a normally distributed dependent variable. This

meant that modeling was of ecological level data (i.e. rates and proportions by areal unit) introducing the potential for ecological fallacy.

5. SUMMARY

The application of spatial statistical analysis to health data has reached adolescence. The theory and the software are both still maturing. We are drawing upon the experiences of the geostatisticians in modeling surfaces and the econometricians in modeling time series. "New and improved" computer algorithms are constantly being provided to implement the evolving theory or to improve the processing in terms of stability, reliability, and efficiency. We will come of age when we have the theory, the software, and the process to reliably produce "generalized spatio-temporal" models suitable for health data.

In the meantime, biostatisticians need to acknowledge when their data is not independently distributed and to consider the spatial correlation in their analysis. This chapter provided examples using four available methods. The methods were spatial filtering, identifying clusters using the spatial scan statistic, hierarchical modeling, and conditional autoregression modeling.

References

Armstrong M. P., Rushton G., Zimmerman D. L. (1999). Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18(5):497-525.

Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* 36:192-236.

Besag J., York J., Mollie A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43(1):1-59.

Breslow N. E., Clayton D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 88(421):9-25.

Maptitude® Geographic Information System for Windows (2000). Version 4.2. Newton, MA:

Clayton D., Bernardinelli L. (1992). Bayesian methods for mapping disease risk. In: Elliott P., Cuzick J., English D., Stern R., editors. *Geographical*

and Environmental Epidemiology: Methods for Small-Area Studies. New York: Oxford University Press; p 205-20.

Clayton D., Kaldor J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 43:671-81.

Cressie N. A. C. (1993). *Statistics for Spatial Data*. Revised ed. New York: J. Wiley.

Cressie N. A. C., Chan N. H. (1989). Spatial modeling of regional variables. *Journal of the American Statistical Association* 84(406):393-401.

Cressie N. A. C., Read T. R. C. (1989). Spatial data-analysis of regional counts. *Biometrical Journal* 31(6):699-719.

Department of Health and Human Services. (1999). Standards for privacy of individually identifiable health information. Office of the Assistant Secretary for Planning and Evaluation, DHHS. Proposed rule. *Federal Register* 64(212):59918-60065.

Devesa S., Grauman D. J., Blot W. J., Pennello G. A., Hoover R. N., Fraumeni Jr J. F. (1999). *Atlas of Cancer Mortality in the United States: 1950-94*. Bethesda, MD: National Cancer Institute.

ArcView Spatial Analyst (2000). Version 2.0a. Redlands, CA: Environmental Systems Research Institute, Inc.

Frisbie W. P., Biegler M., de Turk P., Forbes D., Pullum S. G. (1997). Racial and ethnic differences in determinants of intrauterine growth retardation and other compromised birth outcomes. *American Journal of Public Health* 87(12):1977-83.

Goldstein H., Rasbash J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A* 159:505-13.

Haining R. P. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.

Kaluzny S. P., Vega S. C., Cardoso T. P., Shelly A. A. (1998). *S+ Spatial Stats: User's manual for Windows® and UNIX®*. New York: Springer.