DEPARTMENT OF FINANCE

ARNOLD SCHWARZENEGGER, GOVERNOR

915 L STREET ■ SACRAMENTO CA ■ 95814-3706 ■ WWW.DOF.CA.GOV

June 25, 2007

Honorable Wm. Lacy Clay
House of Representatives
U. S. Congress
Washington, DC 20515-6163

Dear Congressman Clay;

Thank you for the opportunity to testify before the Subcommittee on Information Policy, Census, and National Archives, Committee on Oversight and Government Reform. For my testimony I have attached an evaluation by my department of the 2007 Dress Rehearsal Local Update of Census Addresses (LUCA) program.

The Census Bureau selected San Joaquin County, California as one of the two sites in the nation to serve in 2008 as the Dress Rehearsal site for the 2010 Census. Last year the California State Data Center participated in the Dress Rehearsal LUCA program. Several problems were encountered associated with the local address file, the Census Bureau's LUCA shapefiles, and the Census Bureau maps. We would like to share what we learned and hope that our comments and suggestions for the 2010 LUCA process (see attachment) will inform your hearing.

If you have questions or need more information, please feel free to contact me.

Sincerely,

Mary Heim
Chief, Demographic Research Unit
Department of Finance
915 L Street
Sacramento, CA 95814

(916) 323-4086

San Joaquin Dress Rehearsal LUCA Evaluation

Part 1. Issues and Solutions

A) Problem: Issues with the local address file

The Data Center identified several potential sources for listings of addresses. The only source available, however, was the County Tax Assessor's file, which is used for local property tax assessment purposes.

We encountered several problems associated with using the Tax Assessor's file:

- The file includes both residential and commercial addresses; however, there is no field to identify the address as residential or commercial.
- The file reports addresses in both city-style and non-city-style formats. The Census Bureau accepts only city-style residential addresses.
- The file contains duplicate addresses.
- Some addresses that are physically located in the San Joaquin County, but are listed in an adjacent county for tax purposes, are excluded from the file.
- In order to geocode, we need a file with as complete a physical (property) address as possible. Since this address list is used for tax purposes, mailing addresses are quite complete, but physical addresses are not.
- There were more fields and information pertaining to mailing addresses than to physical addresses. For example, ZIP codes are provided for mailing addresses, but not for physical addresses.
- Fields critical to the physical address are either not reported or inconsistently reported. For example, a given physical address may inconsistently report the street type (Avenue, Road, etc.) or the street direction (North, South, etc.).

Solution: Compiling ZIP codes though Address Matching

Since it is more accurate to geocode based on ZIP code rather than place code, our first step was to try to obtain as many ZIP codes for the physical property addresses as possible. We matched the mailing address with the physical (property) address in order to get the mailing ZIP code. Any property address that did not generate a match with a mailing address was entered into the US Postal Service website in order to retrieve its corresponding ZIP code.

Due to the large number of addresses, we did not have time to fix as many street types and street directions as we would have liked before starting the geocoding process.

B) Problem: Unfinished Census Bureau shapefiles

At the onset of the project, we downloaded the revised 2005 TIGER Line file for San Joaquin County from the Census Bureau website and converted it to a road shapefile and a block shapefile. Next, we geocoded the county address list using the converted TIGER road file. We then spatially joined the results with the TIGER block file (a two-step process that allowed us to obtain tracts and blocks).

In June 2006, we received two LUCA .dbf files (AddressRanges and Roadnames) and eight shapefiles (All Lines or Roads, Area LndK, Block, County, Hydro Area, MCD, Place, and Tract) from the Census

Bureau. In order to use the shapefiles for geocoding, it was necessary to join the AllLines, AddressRanges, and Roadnames files together using the TLID. Once joined together, the resulting combination file contained the variables needed to geocode, but the number of geocoding matches was significantly lower using the LUCA road shapefile than using the TIGER. For example, the table below shows the number of addresses geocoded for the city of Tracy using the TIGER road file in comparison with the LUCA road file. Geocoding with the LUCA road file resulted in less than 45% of the addresses geocoded, compared to 66% using the TIGER file.

Score	TIGER	LUCA
Matched:		
100	11,274	887
80-99	3,836	9,396
Unmatched	7,838	12,665
Total	22,948	22,948

Another issue concerns block suffixes[ii]. The LUCA block shapefile contains split blocks with a suffix assigned to them. In comparison, the TIGER block shapefile had some, but not all, of the split blocks with suffixes. However, comparisons between the spatial join results using both the LUCA and TIGER block shapefiles showed that except for the suffixes, the 2005 revised TIGER Line file had better geocoding results than the LUCA files.

We asked for assistance from the Census Bureau's headquarters, their Seattle Regional Office, and also from ESRI to understand why there was such a discrepancy between the TIGER and the LUCA files, but they were not able to provide an answer. Only recently, after the Dress Rehearsal LUCA program was over, we learned from ESRI that, in order to use the LUCA files for geocoding, an additional step was necessary after joining the three files together using the TLID number. ESRI told us that the road file created from joining the three LUCA files was not standardized in accordance with the formatting required by the ESRI address locator style US Streets with Zone. Therefore, we had to run ESRI's Standardize Addresses tool before using the LUCA shapefile. Using this newly-standardized shapefile, we found that LUCA results were closer to those produced using the TIGER road file (see table below).

Score	TIGER	LUCA standardized
Matched:		
100	11,274	9,439
80-99	3,836	5,414
Unmatched	7,838	8,095
Total	22,948	22,948

However, there were still discrepancies between the TIGER and the standardized LUCA geocoding results. We believe the reason for these discrepancies may be because the TIGER road file uses continuous, exhaustive street address ranges more often for the road segments, while the LUCA road file sometimes breaks address ranges into smaller segments, or into what seems like several blocks in one range. The table below shows the disparities between the two files. For example, note that the TIGER road file is missing the 400s address range on 1st Street, while the LUCA road file does not have the 400s address range for California Street.

	TIGER				LUCA			
Street Name	From Left	To Left	From Right	To Right	From Left	To Left	From Right	To Right
Baker Av	1700	1798	1701	1799	1700	1704	1701	1705

	1800	1998	1801	1999	1706	1998	1707	1999
California St	499	401	498	400	*No 400s*			
Coley Av	2000	2048	2001	2049	0	0	2001	2049
	2050	2098	2051	2099	2050	2098	2051	2099
1st St	1000	1006	1001	1033	1000	1006	1001	1033
	1018	1098	1035	1099	1018	1036	1035	1049
					1038	1114	0	0
1st St	*No 400s*				0	0	421	449
					0	0	451	465
Jackson Av	1700	1708	1701	1709				
	1710	1798	1711	1799	1710	1798	1711	1799
Jackson Av	2200	2298	2115	2299	*No 2200s*			

Solution: Merging data files and manual editing

To overcome the shortcomings of the TIGER and LUCA files, we had to merge information. In some sense, the two files were complementary: although the TIGER file had better geocoding results than the LUCA, it had incomplete block suffixes; the LUCA, on the other hand, had a more comprehensive list of block suffixes. To obtain the missing suffixes, we geocoded using both the TIGER and LUCA files and then matched by address, tract, and block. Then we appended the LUCA suffixes to the TIGER file.

For manual editing, we used with Google Earth satellite imaging to verify or estimate the location of each address that did not match. We then used both the LUCA shapefiles and paper Thomas Bros. Maps to place the unmatched address into the correct block and tract. We referred to Thomas Bros. Maps when we could not find the address location on the LUCA maps. Most of the time, this was because the LUCA maps did not have a road for our given address, or the LUCA map contained other mapping errors such as street names being placed at the wrong locations.

Had the Census Bureau provided an updated TIGER-like file, several errors could have been avoided. These unnecessary errors complicated the task and reduced the time spent identifying real problems with the local address file. From our point of view, providing the tracts and blocks for the LUCA program was very labor intensive. Tremendous amounts of time and resources were necessary to get the job done. Should a jurisdiction have neither GIS capabilities nor staff experienced with Census data, this could be a difficult, if not impossible, undertaking.

C) Problem: Errors with the Census Bureau Maps

The census maps that we received from the Census Bureau contain the following errors:

- Many streets on the LUCA maps have no name at all or are labeled with the wrong name. There are also many incorrect spellings.
- There are roads drawn on the LUCA maps where no roads actually exist.
- The maps are not always drawn to scale.
- "Non-visible boundary" lines drawn on LUCA maps make identifying the correct block nearly impossible at times.

- Some roads are inverted or reversed, which lead to addresses being placed in the wrong block or even the wrong tract.

- Many street types are not labeled properly as road, lane, street, circle, etc.

Part 2. Recommendations

a) Recommendations for the Census Bureau:

- Provide TIGER-like shapefiles that are ready to use and do not have to be joined like the LUCA Dress Rehearsal files. These files should include the most complete, up-to-date information such as street address ranges, street names, street types, street directions, ZIP codes, and place codes in the road shapefile; and tracts, and block suffixes in the block shapefile.

- Allow final submissions to be in Excel format (for jurisdictions that have less than 65,000 street addresses) or a format other than pipe-delimited ASCII files, since that option is not available in Excel.

- Have experienced, knowledgeable people available to provide technical support.

- Any address search information available online from the Census Bureau should be updated and reliable (e.g.. the Address Search feature of American FactFinder is not always reliable)

- Incorporate the latest BAS data into TIGER before printing the maps.

- The maps have to be cleaner and more topologically correct. If the TIGER file has major problems, the data cannot be geographically correct and will give inaccurate results.

- For the 2010 LUCA, local participants will need both digital PDF maps and digital shapefiles. Digital shapefiles allow participants to quickly find street names so they can get the tract and block for a given address. PDF files allow participants to print out selected map sheets as needed for review.

- It would be useful to have a computer specialist participate in the training sessions. This specialist should have a thorough understanding of the LUCA CD-ROM files, be knowledgeable about other potential software and GIS applications, and be familiar with the necessary hardware.

b) Recommendations for State Data Center Participation

- Coordinate the workshop program. The SDC will be responsible for selecting workshop locations, working with regional and county agencies to reserve workshop locations, and inviting jurisdictions to participate in the workshop.

- Encourage local jurisdictions to participate by indicating the financial benefits of an accurate population count.

- Follow-up with non-participating jurisdictions to encourage participation. Support county-wide coordination and meetings.

- Expect to provide some technical support, based on the level and promptness of support provided by the Census Bureau.

- Provide assistance to jurisdictions that want to participate but lack expertise or other resources.
    - Provide geocoding only—the local officials should do their own follow-up work for unmatched addressees. There will always be unmatched addresses due to the lag time between the production of files and their use
    - Make sure jurisdictions understand that they must provide an address list that contains addresses and ZIP codes to facilitate the geocoding process

- Provide each jurisdiction with information regarding LUCA, including strategies for participation, software, hardware, data sources, problems encountered and possible solutions, as well as sources for help.

- Focus on group quarters, employer housing, etc.
    - Start this process as soon as possible. Many of these addresses are non-city-style and it can be very time-consuming to identify census tracts and blocks.

c) Recommendations for Local Government Participation

- Start as early as possible.

- Develop a priority list of the work that needs to be performed in terms of both successful participation in LUCA as well as an accurate census count.

- Identify potential problem areas such as new housing developments, large apartment complexes, large mobile home parks, commercial areas with residential quarters, areas where addresses have changed (due to annexation, demolition, or redevelopment), and areas with significant numbers of illegal or unconventional units.

- Review LUCA maps for missing streets, address ranges, and incorrect city boundaries.

- Develop a local address file that contains addresses and ZIP codes.

- Match the unit count in the local file to the count in the LUCA file at the tract or block level to calculate the difference in the unit counts between the two files. Resolve differences between the two files starting with the areas with the largest discrepancies.
    - Street address matching can be used to understand these discrepancies. Geocoding problems (units assigned to the wrong block) may account for some of the more significant differences in a given area.

- If a jurisdiction anticipates significant building construction between June 2009 (after the Address Canvassing Operation) and April 2010, it should develop a plan to notify the Census Bureau of these new units.

- Encourage a county-wide meeting of all participants once the LUCA materials are received. Contribution from participating agencies can help lead to a more successful LUCA program.

- Participation in LUCA can potentially be very time consuming. The county coordinator should emphasize to all jurisdictions that even minimal participation (such as reviewing city boundaries and looking for discrepancies at a large geographic level) will be very useful. Some cities may feel that if they can not do all the tasks, then they should not participate at all. Any contribution, no matter how small, should be supported.

[i] Geocoding is the process of taking an address, such as those from the San Joaquin County Tax Assessor file, and converting it to x,y coordinates that can be plotted or placed on a map as a point. This process is called matching and is done using an address locator generated in ARCMap. The address locator compares the descriptive location elements of the address (i.e. street number, street name, street type, direction) to those present in the reference material (TIGER/Line road file).

Through an address locator, each address in the San Joaquin Tax Assessor's file was assigned a score, called a match score, from 0 to 100 based on how closely the elements of the address from the Tax Assessor file matched the elements in the TIGER/Line file. In general, scores are lower if address elements are misspelled (i.e. the street name is misspelled), incorrect (i.e. the street number falls outside the address range in the TIGER/Line file), or missing (i.e. a street direction is specified in the TIGER/Line file but is not present in the address file).

The address locator finds the best matches and assigns an x,y coordinate (point) to those addresses meeting or exceeding the minimum match score, as specified in the address locator. A shapefile is created showing the placement of the points on the map.

[ii] After the 2000 Census, the addition of new roads or changes in a boundary might have resulted in split blocks. In these situations, the Census Bureau adds a suffix to the new block to identify the geography where the housing unit is located.