A GEOCODING CASE STUDY:

CALIFORNIA DEATH CERTIFICATES 2005 and 2007

By

Linda L Remy, MSW, PhD
Research Director
UCSF Family Health Outcomes Project
Email: lremy@well.com

June 2009

Family Health Outcomes Project
Family and Community Medicine
University of California, San Francisco
500 Parnassus Ave. Room MU-337
San Francisco CA 94143-0900
Phone: 415-476-5283
Fax: 415- 476-6051
Email: FHOP@fcm.ucsf.edu

TABLE OF CONTENTS

Funding by the California Department of Public Health, Center for Health Statistics under CDPH Contract No. 07-65212.

This was a cooperative effort involving the following individuals:

UCSF Family Health Outcomes Project: Gerry Oliva MD, Director; Linda Remy PhD, Research Director and analyst for this project.

California Department of Public Health:Gwen Doebbert, Chief Health Informatics Policy & Planning; Jane McKendry, Chief OHIR; Scott Fujimoto MD, Asst Chief OHIR. Special thanks to Anne Millington for extracting the death certificate data and Matthew Stone for geocoding the datasets.

DEFINITION OF TERMS

CDIC	Chronic Disease and Injury Control
CHS	Center for Health Statistics
County	San Diego County, California
DHS	Department of Health Services
DSM	Death Statistical Master
EDRS	Electronic Death Registry System
FHOP	UCSF Family Health Outcomes Project
Geocoding	The process of assigning geographic identifiers (e.g., codes or geographic coordinates expressed as latitude-longitude) to map features and other data records, such as street addresses.
GIS	Geographic information system
OHIR	Office of Health Information and Research
SDGE	San Diego County death records geocoded
SDGIS	San Diego Geographic Information System address records
SFN	The State File Number (SFN) is the sequential number assigned by the State Office of Vital Records
Standardized Address	A standardized address is fully spelled out, abbreviated using standard Postal Service abbreviations.
ZP4	ZP4 is a DVD-ROM available from Semaphor.com that contains United States addresses, ZIP + 4® codes, mail carrier route numbers, and other supplemental postal databases (all maintained by the Postal Service™), plus complete automatic CASS Certified™ address correction software for Windows computers.

A GEOCODING CASE STUDY:

CALIFORNIA DEATH CERTIFICATES 2005 and 2007

BACKGROUND

The death certificate is the official source of information about deceased people. It provides *prima facie* evidence of a person's death and is relied upon for important legal proceedings as well as for the creation of important public health and regulatory data. California processes about 250,000 death certificates per year (1 in 10 deaths in the U.S.). For over a century, the State used a costly paper-based death registry process that had many deficiencies. In 2002, the legislature passed AB2550 which provided funding for the California Department of Health Services (DHS) to build an Electronic Death Registration System (EDRS).[1]

Training for CA-EDRS began in December 2005, with different counties trained monthly.[1] By the end of 2007, CA-EDRS was mandatory in all but the smallest California counties. Today, coroners, funeral directors, doctors, and hospitals can use the web-based CA-EDRS system to submit electronic death certificates for registration 24 hours a day. With the implementation of CA-EDRS, California became the second state in the nation to do electronic death registration.

This study has two aims:

1. Evaluate the quality of address data on death certificates in 2005 (pre-EDRS) and 2007 (post-EDRS). More specifically, this analysis focuses on information relevant to geocoding: address, city, ZIP-code. Both years of data are referred to as EDRS.

2. Compare accuracy of geocoding for two systems now used at the State level: CENTRUS and ESRI.

San Diego County (hereinafter "County") was selected for in-depth analysis because it has a history of submitting high quality death certificate data. The County began to submit death certificate data electronically in March 2006.

The policy-setting document *Healthy People 2010* highlighted potential benefits of geocoding: "The capacity to achieve national goals is related to the ability to target strategies to geographic areas. Extension of geocoding capacities throughout health data systems will facilitate this ability."[2] Researchers rely on geocoded address data to understand the prevalence and timing of public health data which often are inherently spatial.[3] For example, a study of premature death due to chemical exposures might consider residence and work addresses. The problem is how to get these locations on a map.

Preparing address data for a map is called geocoding. In this process, geographic coordinates expressed as latitude-longitude are assigned to street addresses which can be assigned to map features. This paper evaluates the accuracy of exact address geocoding using one software package to clean addresses for County deaths and two different software packages to assign coordinates.

PREPARE DEATH CERTIFICATE DATA

METHOD

In preparing the death data to test geocoding quality, our goal was to have one record for each address where at least one person died. Multiple records for the same person or address can occur for various reasons. Some occur when amendments are submitted. Condominiums, nursing homes, and apartment buildings often have multiple dwellings with the same address. Others occur because multiple people can die in the same house over time.

In mid-2008, staff at the California Center for Health Statistics (CHS) obtained a set of raw EDRS records for calendar years 2005 and 2007. They then extracted County records, standardized addresses using ZP4[4] and geocoded the addresses. We refer to geocoded data files as SDGE. For testing, CHS staff geocoded SDGE data using CENTRUS[5] and ESRI[6] software. Both firms are well-known and respected geographic information providers.

CHS staff sent EDRS and SDGE data to the UCSF Family Health Outcomes Project (FHOP) where records were unduplicated as described in **Appendix A. Unduplicate Death Records**.

RESULTS

Multiple records per SFN. Duplicate records for the same SFN occur when death certificates are amended. Here we focus on the impact of EDRS implementation using number of records as a workload proxy. Table 1 compares the number of records for the State and County in 2005 (pre-EDRS) and 2007 (post-EDRS). The number of records are fewer in 2007 because submission was incomplete when data were pulled.

Table 1. Death registry records, State and County

	State		County	
	Year		Year	
Records	2005	2007	2005	2007
Total records	262,532	187,660	20,479	18,044
Multiple records per SFN	47,898	45,890	1,285	918
Single record SFN	214,634	141,770	19,194	17,126
Single record SFN %	82	76	94	95

Statewide, SFN with single records dropped from 82% of submissions for 2005 to 76% for the incomplete 2007. This is interpreted as a measure of burden associated with EDRS implementation because more amendments were needed.

By comparison, the County resubmission rate was much lower and slightly improved post-EDRS. It submitted 1% fewer corrections after EDRS than before.

Discrepancies between EDRS and SDGE Files. We merged the EDRS and SDGE files to be certain that the SDGE records matched records in the source files. We identified 5,983 records in the SDGE file that were not in the EDRS file which we thought was their source. Fewer were missing in 2007 than in 2005.

Table 2. Final selection of records by year

	Year		% of
State File Numbers	2005	2007	2005
Total single SFN	19,834	17,126	86
SFN not in EDRS	3,314	2,669	81
SFN in both files	16,520	14,457	88
Multiple SFN, same address	4,945	3,086	62
Single SFN single address	11,575	11,371	98
Usable %	70	79	

Many SDGE addresses had more than one death record. This situation arises because people die in nursing homes, senior communities, or other types of multi-dwelling addresses. Because we needed a unique address, we removed all but the first address record without regard for SFN (N = 22,946).

In the end, we had about the same number of records available for linkage in both 2005 and 2007. After removing multiple SFN at the same address, 70% of SFN were at unique addresses in 2005, and 79% in 2007.

Standardized addresses disagree. A standardized address is fully spelled out, using Postal Service standard abbreviations.[7] Software such as ZP4 validate, correct, and standardize addresses retrospectively, after they have been entered.[4] Other software is available to enter an address prospectively, using a few keystrokes which automatically find and supply the verified, complete and standardized address into an existing database application.[8] The problem with retrospective address standardization software is that they do not return the same string for two obviously similar addresses. The geocoding problem is compounded when different strings are used to signify the same address.[3]

Table 3. Original and standardized addresses

SFN	ADDRESS IN	ADDRESS OUT
1	401 "E" AVENUE, # 7 401 "I" AEVE. #7	401 E AVE # 7 401 I AVE APT 7
2	1138 AVOCADO AVE. 1138 EAST AVACADO AVE	1138 AVOCADO AVE 1138 AVOCADO AVE
3	2743 PAYSON DR. 2941 PAYSON DR.	2743 PAYSON DR 2941 PAYSON DR.
4	404 ENCINITAS BOULEVARD #405 4040 ENCINITAS BOULEVARD, #405	404 ENCINITAS BLVD APT 405 4040 ENCINITAS BOULEVARD, #405
5	11588 VIA RANCHO SAN DIEGO #F1065 1588 VIA RANCHO SAN DIEGO #F1065	11588 VIA RANCHO SAN DIEGO # F1065 1588 VIA RANCHO SAN DIEGO #F1065
6	11 ORLANDO STREET APT. 22 379 ORLANDO STREET APT. 22	11 ORLANDO STREET APT. 22 379 ORLANDO ST APT 22
7	4650 DULIN RD., SP. #54 4680 DUBLIN RD., SP. #54	4650 DULIN RD SPC 54 4680 DULIN RD
8	7501 BUCKBOARD TERRACE 7105 BUCKBOARD TRAIL	7501 BUCKBOARD TERRACE 7105 BUCKBOARD TRL
9	3185 BILENE LANE 3185 BRILENE LANE	3185 BILENE LANE 3185 BRILENE LN

Table 3 compares examples of multiple address records for a given SFN. Similar problems existed in the County parcel files.

- Similar street types resolve to different types: Street or ST; Lane or LN; Terrace or Trail.
- Units resolve inconsistently, with and without internal punctuation: Blvd Apt 405 or Boulevard, #405.
- House numbers can be very far apart: 404 or 4040, or 401 E AVE or 401 I AVE.

These types of problems on both sides of a geocoding linkage magnify problems in making accurate linkages and assignments of centroid points.[3]

ASSESS GEOCODING QUALITY

METHOD

In preparing County address data, our goal was to have one record for each parcel that had a street address. Condominiums, nursing homes, and apartment buildings often have multiple dwellings with the same address. Undeveloped parcels often have no address. We downloaded publicly available address data (parcel and street) from the San Diego County Geographic Information System (SDGIS) website and unduplicated them. This work is summarized in **Appendix A. Unduplicate County Parcels**.

In preparing the SDGIS data to merge with SDGE data, we encountered many problems and had to devise strategies to address them. This work is summarized in **Appendix A. Reconcile Differences between SDGE and SDGIS.**

Creating a file we could use to assess geocoding quality involved three final steps: (1) linking SDGE and SDGIS data, (2) adding SDGIS Census geography (tract, block group, block) variables onto the SDGE file, and (3) calculating distance from the SDGIS centroids to the SDGE centroids. From this we were able to evaluate the quality of the geocoding processes. Our linkage algorithm is based on a deterministic method.[3] This work is summarized in **Appendix A. Prepare to Assess Geocoding Quality.**

RESULTS

Table 4. Linkage algorithm results

Linkage Criteria	Number	Percent
CITY ADDRESS	20,339	88.64
CITY2 ADDRESS	143	0.62
CITY3 ADDRESS	2	0.01
ZIPC5 ADDRESS	202	0.88
ZIPC53 ADDRESS	35	0.15
CITY HNUM PREDIR STREET	32	0.14
CITY2 HNUM PREDIR STREET	32	0.14
CITY HNUM STREET STRTYPE	86	0.37
CITY2 HNUM STREET STRTYPE	19	0.08
ZIPC5 HNUM STREET STRTYPE	2	0.01
ZIPC53 HNUM STREET STRTYPE	1	0.00
CITY2 HNUM STREET	21	0.09
ADDRESS	3	0.01
NOMATCH	2,029	8.84

Linkage. Linkage involves a one-to-one match when the same variables are in two files. SDGIS data often had a city (or alternate) or ZIP-code (or alternate) but not both. This meant that we had to try a number of ways to link records. These methods and their results are summarized in Table 4.

The most important finding is that we were able to match 88.6% of records to the parcel level using a "guestimated" city and address, ignoring ZIP-code and apartment numbers. We were unable to make a parcel-level linkage for 2,029 addresses (8.84%).

With a total of 91.16% of linkages based on variations of address, this exceeds the attainable geocoding match rates of 70 to 85% reported by others and approximates the highest address matching rate reported by the New Jersey Cancer Registry.[3]

Table 5. Census geography agreement

Level	Description	Number of Addresses	Percent
	Total	22,946	
	No assignments	584	
	Usable	22,362	100.0
Tract	Disagree	232	
	Agree	22,130	0.990
Block Group	Disagree	296	
	Agree	22,066	0.987
Block	Disagree	1,700	
	Agree	20,662	0.924

Census Geography. Census variables were unavailable for 584 addresses (2.5%). Table 5 summarizes agreement between the SDGIS and SDGE Census variables from CENTRUS.

At the tract level, agreement was 99%; at the block group, 98.7%; and at the block level, 92.4%.

Our ability to make a parcel-level match affected reliability.

These results match or exceed those reported by others.[3]

Table 6. Percentile distance in feet from centroid

	County to		ESRI to
Percentile	ESRI	CENT	CENT
1%ile	0.00	372	373
5%ile	0.00	850	855
10%ile	0.00	1,213	1,218
25%ile	0.00	1,897	1,900
50%ile	0.31	2,693	2,692
75%ile	0.44	3,351	3,346
90%ile	0.54	3,856	3,837
95%ile	21.56	4,158	4,121
99%ile	639.78	4,756	4,561

Table 6 compares the percentile distribution of distance in feet from the County parcel-level centroid to the ESRI and CENTRUS centroids and from ESRI to CENTRUS.

Even at the 95%ile, the ESRI centroid varied by only about 22 feet from the County. Distances from the County to CENTRUS and from ESRI to CENTRUS were much higher. Even though SDGIS, ESRI, and CENTRUS used different projections, SDGIS and CENTRUS geocoded to the address level while CENTRUS did not. Again the results exceed those reported by others.[3]

We categorized distance in feet from the SDGIS and ESRI centroids, to examine relationships between centroid distance and agreement on Census variables between SDGIS and CENTRUS. This is shown in Table 7.

Table 7. Agreement between centroid and Census

			Distance from County centroid (Feet)					
Level	Agreement	Total	Missing	0	50	100	500	1,000
Tract	Disagree	232	174	44	0	2	3	9
	Agree	22,130	1,417	19,974	182	267	255	35
Block Group	Disagree	296	207	72	0	4	3	10
	Agree	22,066	1,384	19,946	182	265	255	34
Block	Disagree	1,700	561	970	23	50	76	20
	Agree	20,662	1,030	19,048	159	219	182	24

Even though there was agreement between SDGIS and ESRI (Table 6) to within less than 25 feet for more than 95% of addresses, most SDGIS and CENTRUS Census variables agreed exactly as to parcel centroids.

For example, there was complete agreement on parcel centroids (Distance = 0 feet) for 44 addresses disagreeing on tract, 72 disagreeing on block group, and 970 disagreeing on block.

DISCUSSION AND RECOMMENDATIONS

EDRS implementation appears to have affected data quality, at least temporarily. More records were submitted per given SFN. This was not observed in the County chosen for this exercise because it has a history of submitting high quality data.

Going into this process, FHOP thought that because the County was said to have high quality death certificate data, it would have high quality parcel data. That was not the case. Getting the databases to "talk" to each other required a great deal of tedium mixed with creativity.

Even though it was difficult to get the files to link because of non-standard addresses, geocoding variables (census, parcel centroids) agreed highly in the end after addresses were standardized using an adapted exact method.

ESRI, CENTRUS, and SDGIS report geographic location in different metrics based on different projections. ESRI uses the most commonly known projection, the Teale-Albers NAD83 that California uses for its official coordinates.[9] CENTRUS uses the NAD83 Datum that are official coordinates for the US primary geodetic network.[10] SDGIS uses NAD83 California Plane Coordinates. For the same address:

- SDGIS would be expressed as 1878406.18, 6395241.88;
- ESRI would be expressed as -571921.34, 300768.05
- CENTRUS would be expressed as 32.90, -117.06.

The obvious problem was getting these different projections to "talk" to each other so we could calculate distance agreement statistics. They had to be converted to a common metric. We used the Teale-Albers NAD83 because we had SAS software for this projection, enabling us to convert the data to distance in feet between points.

When we were able to get all projections to a common metric, the accuracy of ESRI's projections emerged. We do not know if we would have had similar results with data from another county, or what would have happened if SDGIS had not used ESRI, albeit with a different projection.

As a result of these findings we recommend:

1. **Use pulldowns at data entry.** Adding city and street within city pulldowns, plus address verification software when the full address is entered would greatly increase EDRS data entry efficiency and accuracy. Once city and ZIP are known, the list of possible streets is greatly reduced. Such a pulldown would decrease the burden on local vital statistics departments, enabling them to enter this crucial information correctly the first time. Standardizing addresses after the fact using software like ZP4 is not as accurate, efficient, or effective.

2. **Use ESRI to geocode addresses**. ESRI is based on satellite projections centered on houses. As such it is more aligned with parcel shape files such as those from SDGIS, and is a more precise geocoding product than CENTRUS.

3. **Locate parcels before overlaying Census variables.** To improve accuracy for administrative variables like Census geography, it is better to locate parcels and then overlay higher-level geographies. Parcel centroids tend to be more stable once houses are sited on

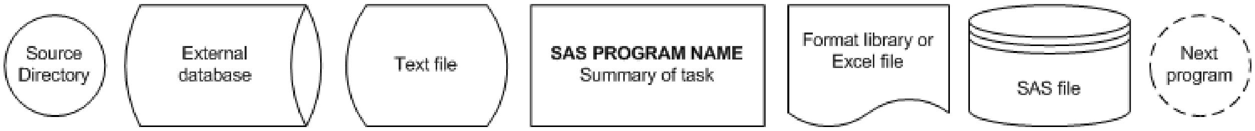
them, while administrative measures such as city boundaries or Census variables change over time. Locating the parcel first would reduce the frequency with which parcels agreeing to 0 feet on centroids are assigned to different Census variables. While this was rare at the tract level, the lower the administrative level, the greater the likelihood for discontinuity. Further, to minimize the possibility of mismatch, Census variables should be from the same company that does the geocoding.

4. Move toward a Master Address File. Once 1 to n variations on a given address have been assigned to a single geography, there is no need to go through that process twice. Use annual vital statistics datasets to create a Master Address File. Each year, only those addresses that did not link would need to be standardized and geocoded. Once a given address was processed it could be added to the Master Address File. In time, most variations of a given address would be standardized and geocoded. Rushton et al reported that 18% of cancer registries used a master address file.[3]

The Master Address file will require attention to longitudinal issues. For example, streets may exist when a military base is open that are redefined and/or realigned and/or renamed after the base closes and converts to civilian housing. This can create a situation where multiple streets are assigned to the same geography when in fact it is the same geographic location newly renamed. At the least, every variation of a given geocoded address in a Master Address File should have a date and source variable associated with it. It will be important to learn how to manage software updates to be sure that the ability to retroactively assign addresses that no longer exist is retained over time.

APPENDIX A: SUMMARY OF DATA STEP METHODS AND RESULTS

LEGEND:



UNDUPLICATE DEATH RECORDS

Step 1. Pull EDRS data. In mid-2008, staff at the California Center for Health Statistics (CHS) obtained all available death records for calendar years 2005 and 2007 (2005 N = 262,532; 2007 N = 187,660). These data were imported into an ACCESS database and sent to the UCSF Family Health Outcomes Project (FHOP), where they were read into SAS and output into two files. The first file had unduplicated records with only one State File Number (SFN) (N = 356,404). The other had SFN with more than one record (N = 93,778). The final Death Statistical Master (DSM) will contain only one of those records. Because we did not know which was in the final DSM, we set aside all SFN with more than one record. These latter SFN have been amended in some respect and as such reflect a dimension of data quality.

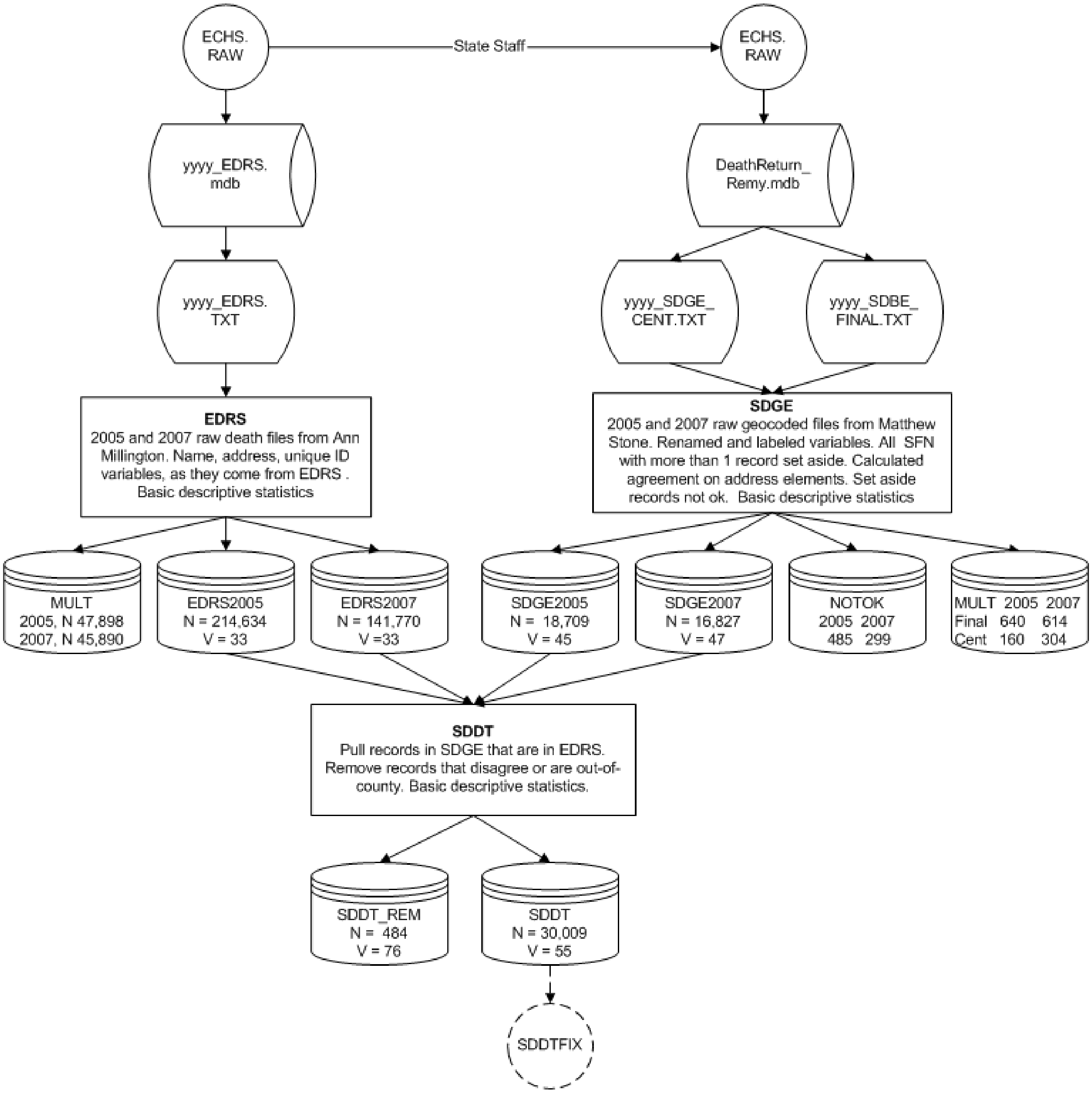
Step 2. Standardize and geocode County addresses. Staff in the DHS Division of Chronic Disease and Injury Control (CDIC) exported County death address data and standardized them using ZP4 (N = 38,523).[11] They then geocoded the standard addresses using two different coordinate systems. ESRI uses the Teale-Albers NAD83 that California uses for its official coordinates,[12] and CENTRUS uses NAD83 Datum that are official coordinates for the US primary geodetic network. [13]

Geocoded data were re-imported into ACCESS to send to FHOP. We refer to these files with the acronym SDGE (San Diego geocoded). The SDGE annual files had two tables each in the ACCESS database. One file had addresses geocoded using CENTRUS, the other using ESRI.

FHOP read these into SAS and merged the two coordinate files by SFN. Again, records were output to annual files based on the existence of either unduplicated SFN (N = 36,320) or multiple records for the same SFN (N = 2,203).

Step 3. Remove Out-of-County records. The County file contained records for people who lived in the County but may have died elsewhere. We were interested in reliability of address geocoding only for County residents whose addresses were entered by County personnel. The County geocoded file did not contain the variable identifying the decedent county of residence. To obtain this, we merged the SDGE list of unduplicated SFN with the EDRS file, and set aside people who were not County residents. This resulted in 30,037 SDGE death addresses preliminarily available for linking, with 456 cases set aside as out-of-county.

Figure 1. Preliminary selection of death addresses


ECHS. RAW
State Staff
ECHS. RAW
yyyy_EDRS. mdb
DeathReturn_ Remy.mdb
yyyy_EDRS. TXT
yyyy_SDGE_ CENT.TXT
yyyy_SDBE_ FINAL.TXT
EDRS
2005 and 2007 raw death files from Ann Millington. Name, address, unique ID variables, as they come from EDRS . Basic descriptive statistics
SDGE
2005 and 2007 raw geocoded files from Matthew Stone. Renamed and labeled variables. All SFN with more than 1 record set aside. Calculated agreement on address elements. Set aside records not ok. Basic descriptive statistics
MULT 2005, N 47,898 2007, N 45,890
EDRS2005 N = 214,634 V = 33
EDRS2007 N = 141,770 V =33
SDGE2005 N = 18,709 V = 45
SDGE2007 N = 16,827 V = 47
NOTOK 2005 2007 485 299
MULT 2005 2007 Final 640 614 Cent 160 304
SDDT
Pull records in SDGE that are in EDRS. Remove records that disagree or are out-of-county. Basic descriptive statistics.
SDDT_REM N = 484 V = 76
SDDT N = 30,009 V = 55
SDDTFIX

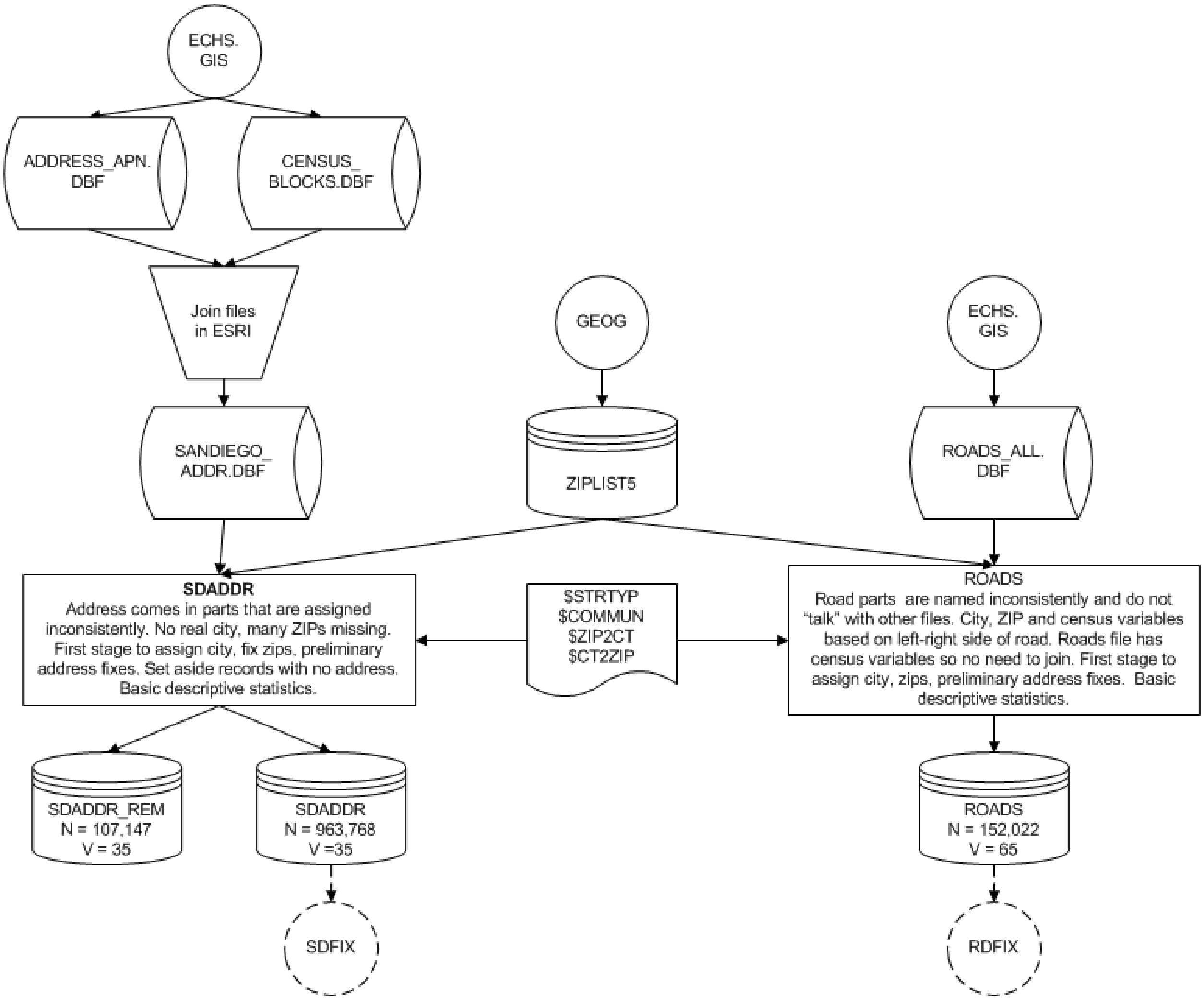
UNDUPLICATE COUNTY ADDRESSES

Step 1: Download County Address Data. Publicly available address data were downloaded from the County Geographic Information website.[14] The address data were already segmented into parts. The County geocodes to a third Projected Coordinate system that utilizes the NAD83 Datum projections based on shape files (parcels, roads).[15]

Step 2: Add 2000 Census geography to County address data. The parcel-level address file was imported into ESRI and joined with the 2000 census geography also available from the County GIS website. The roads file already had 2000 census variables.

Step 3: Preliminarily identify usable addresses. The parcel-level address file with Census geography (N = 1,070,915) was read into SAS then output to two files: preliminarily usable addresses (N = 963,768), and unusable addresses (N = 107,147). The roads file also was read into SAS. Basic edits were applied to both files to pick reference cities and ZIPs. We refer to the resulting file with the acronym SDGIS. This work is summarized in Figure 2.

Figure 2. Preliminary selection of County addresses



SDADDR
Address comes in parts that are assigned inconsistently. No real city, many ZIPs missing. First stage to assign city, fix zips, preliminary address fixes. Set aside records with no address. Basic descriptive statistics.

ROADS
Road parts are named inconsistently and do not "talk" with other files. City, ZIP and census variables based on left-right side of road. Roads file has census variables so no need to join. First stage to assign city, zips, preliminary address fixes. Basic descriptive statistics.

RECONCILE DIFFERENCES BETWEEN SDGE AND SDGIS DATA

We encountered a number of issues with the SDGIS parcel data. The most critical was that addresses lacked identifying cities and/or ZIP-codes. Address was entered in parts, and parts were entered inconsistently. SDGIS had a two-character field to indicate street type, and abbreviations were used inconsistently. Finally, SDGIS parcel data did not include a post-direction (Del Mar Circle West) field while the SDGE data did.

Imputing Cities and ZIP-Codes. The SDGIS parcel data had a field indicating the jurisdiction responsible for the parcel but not the actual city or neighborhood name. About a quarter of the parcels had a jurisdiction but not a ZIP-code. Referring to various sources, we had as many as four candidate "city" variables for a given address. Using a process similar to that described by others,[3] we resolved all addresses to one city and one alternate which we were able to use to link to the SDGE data. However, many cities have multiple ZIP-codes, and we lacked a ZIP-code for 239,641 County address records.

Standardizing Direction. The SDGIS had a pre-direction field (East, West, North, South), but most pre- and all post-directions (Del Mar East) were contained in the street variable. No streets with pre-directions had a post-direction. We removed all direction strings from the street variable and put them in the pre-direction variable. In the EDRS data, no streets with post-direction had a pre-direction. Again, we put the EDRS post-direction in the pre-direction field.

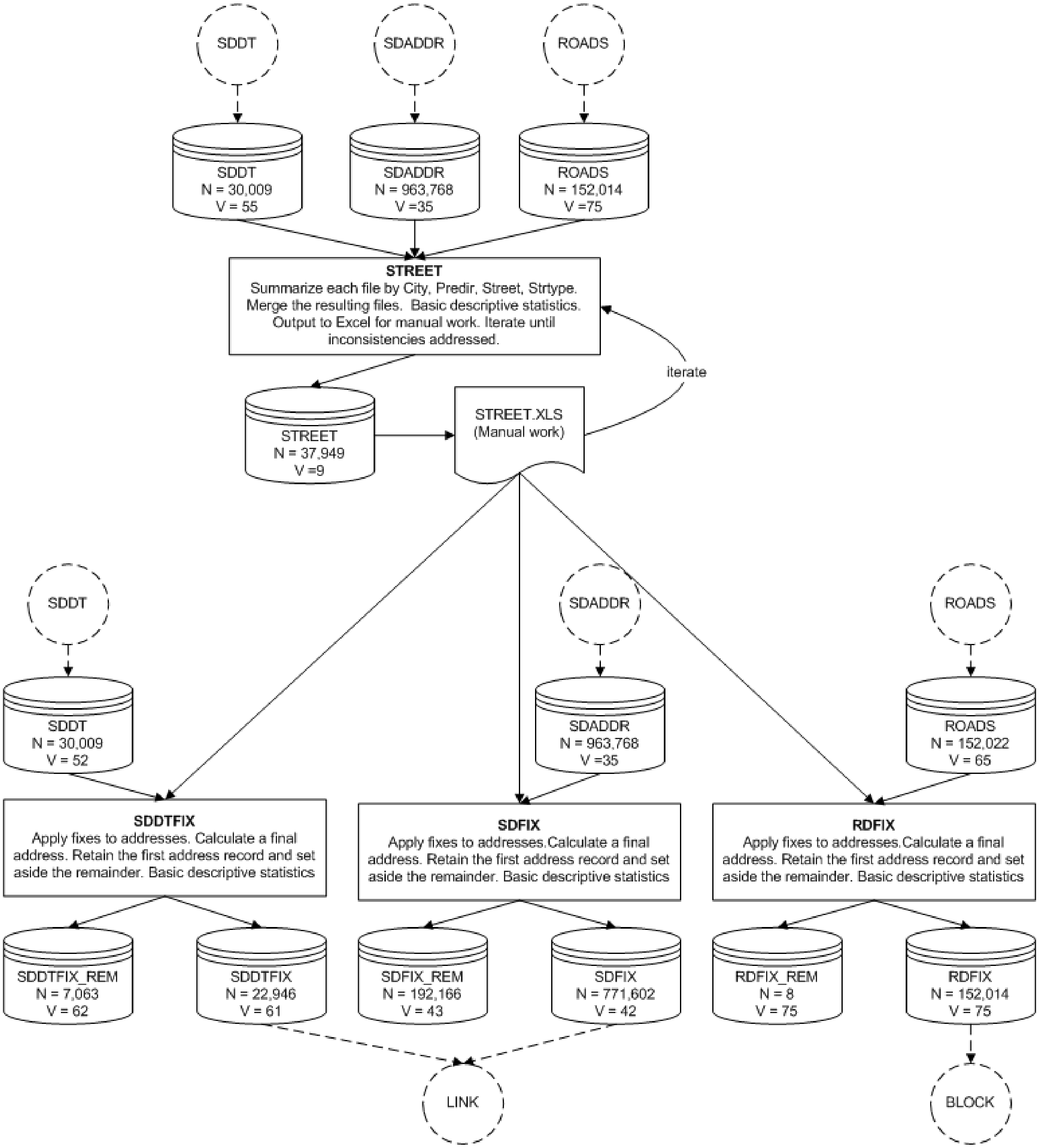
Standardizing Street. In both files, we standardized certain parts of street names to their long form. For example, we found both 'Mt Helen' and 'Mount Helen', or 'St Paul' and 'Saint Paul' in one and/or both files.

Standardizing Street Type. The county used a two-character field to indicate street type, and used abbreviations inconsistently. For example, 'TE' and 'TR' were both assigned for Terrace, and 'TR' was used for both Terrace and Trail. In many cases, street type was included in the street variable and had to be moved to the street type variable. Standardized addresses in the EDRS files used longer street type strings. We standardized street types in both files to 8-character long forms: Street, Avenue, Circle, Trail, Terrace, etc.

City-Street Crosswalk. Most reconciliation summarized above was done using a city-street crosswalk file. We made this by separately summarizing the SDGE and SDGIS files by city, pre-direction, street, and street type, then merging the files and exporting them into an Excel spreadsheet for manual review and editing. This required several days on the USPS website and Google maps, correcting combinations on one side or the other, so we would have the best chance to link.

The Excel file was read back into SAS and separately merged with the SDGE and SDGIS data to reassign address name parts. At this point, we made a new single address variable consisting of house number, pre-direction, street, and street type. In making the new variable, we did not use elements such as fraction (1/2) and apartment or building number (3G).[3] An important aspect of these programs was to assure unique addresses for linkage. Ignoring apartment or unit numbers, the first address was used and the others were set aside. This work is summarized in Figure 3.

Figure 3. Reconcile street name differences


SDDT
SDADDR
ROADS
SDDT
N = 30,009
V = 55
SDADDR
N = 963,768
V =35
ROADS
N = 152,014
V =75
STREET
Summarize each file by City, Predir, Street, Strtype. Merge the resulting files. Basic descriptive statistics. Output to Excel for manual work. Iterate until inconsistencies addressed.
iterate
STREET
N = 37,949
V =9
STREET.XLS
(Manual work)
SDDT
SDADDR
ROADS
SDDT
N = 30,009
V = 52
SDADDR
N = 963,768
V =35
ROADS
N = 152,022
V = 65
SDDTFIX
Apply fixes to addresses. Calculate a final address. Retain the first address record and set aside the remainder. Basic descriptive statistics
SDFIX
Apply fixes to addresses.Calculate a final address. Retain the first address record and set aside the remainder. Basic descriptive statistics
RDFIX
Apply fixes to addresses.Calculate a final address. Retain the first address record and set aside the remainder. Basic descriptive statistics
SDDTFIX_REM
N = 7,063
V = 62
SDDTFIX
N = 22,946
V = 61
SDFIX_REM
N = 192,166
V = 43
SDFIX
N = 771,602
V = 42
RDFIX_REM
N = 8
V = 75
RDFIX
N = 152,014
V = 75
LINK
BLOCK

PREPARE TO ASSESS GEOCODING QUALITY

Step 1: Link SDGE and SDGIS data. To do the linkage, we used a macro developed by FHOP that makes a unique match given certain criteria. This type of linkage is known as the deterministic method.[3] The macro matches one-to-one, and not one-to-many. If multiple records meet the criteria, no selection is made. This is why we had to assure unique addresses in both files. Among other uses, the macro has been used for FHOP's adolescent injury[16] and maternal morbidity and mortality[17] studies. It was adopted for use in OSHPD's Intensive Care Outcomes studies.[18]

Step 2. Add road-level geography for death addresses. The linkage macro made full address matches for 20,917 (91.1%) death records, but did not link 2,029 (8.9%). As a result, we lacked both Census variables (tract, block group, block) and tract-level point data for these latter addresses to compare with the SDGE results.

Virtually all unmatched addresses were on streets that existed. The list of roads in the SDGIS roads file overlapped considerably with the list of roads in the SDGIS address file, but both files had roads that were not in the other. Because of these inconsistencies, exact address matches were not possible for the unlinked residual.

The SDGIS roads file had one to n records for each road segment (e.g, a block), with variables indicating high and low numbers on the left and right side of each segment. We also had 1 to n address records on the same road in the SDGE file. To handle this many-to-many situation, we used FHOP's join macro to put road-level Census geography on the linked file.

As a result of this step, we now had two SDGIS sources for the census variables: parcel-level and street-level. Some records had census variable sets from both sources, others had them on one or the other, and a residual still had none. We calculated new SDGIS census variables as follows: if the parcel-level file had census variables we used that, otherwise we used census variables from the roads file.

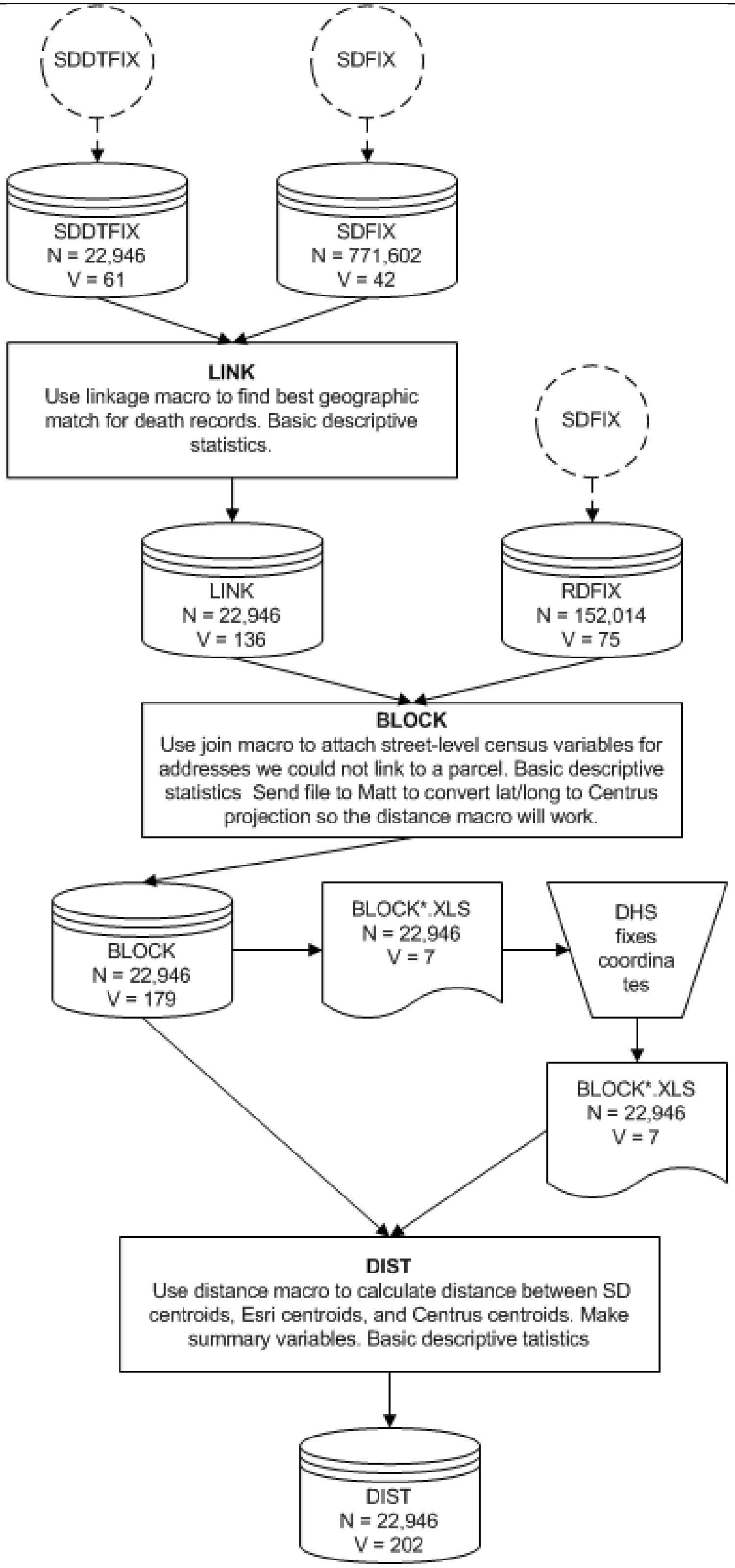
Step 3. Calculate distance from County centroids. The resulting address file had census geography variables on all records as well as latitude and longitude coordinates assigned to most parcels. SDGIS values on census and coordinate variables were used as the "gold standard" to evaluate the accuracy of geocoding by Centrus and ESRI.

However, we encountered yet another problem. Earlier we mentioned that latitude and longitude was measured in three different metrics (ESRI, CENTRUS, SDGIS). We forwarded DHS staff a subset of relevant variables from the joined file to convert all centroids to one metric.

We then used our distance macro to calculate distance between the SDGIS, ESRI, and CENTRUS centroids and calculated various agreement statistics.

Figure 4 summarizes steps to find the best geographic match for death records, join block-level geographic variables for records lacking parcel centroids, and calculate distance between the various measures.

Figure 4. Link, join, and calculate distances


SDDTFIX
SDFIX
SDDTFIX
N = 22,946
V = 61
SDFIX
N = 771,602
V = 42
LINK
Use linkage macro to find best geographic match for death records. Basic descriptive statistics.
SDFIX
LINK
N = 22,946
V = 136
RDFIX
N = 152,014
V = 75
BLOCK
Use join macro to attach street-level census variables for addresses we could not link to a parcel. Basic descriptive statistics Send file to Matt to convert lat/long to Centrus projection so the distance macro will work.
BLOCK
N = 22,946
V = 179
BLOCK*.XLS
N = 22,946
V = 7
DHS
fixes
coordina
tes
BLOCK*.XLS
N = 22,946
V = 7
DIST
Use distance macro to calculate distance between SD centroids, Esri centroids, and Centrus centroids. Make summary variables. Basic descriptive tatistics
DIST
N = 22,946
V = 202

ENDNOTES

[1] Gobar G, Hogarth M. (2007) California electronic death registration: transition from paper to electronic. Last accessed 07-Dec-2008 at: www.mendocinohre.org/rhic/200705/RHIC_5_16_2007.pdf.

[2] Healthy People 2010, Chapter 23: Public Health Infrastructure, Section 3: Use of geocoding in health data systems. Last accessed 31-Jul-2009 at: http://www.healthypeople.gov/Document/HTML/Volume2/23PHI.htm.

[3] Rushton G, Armstrong MP, Gittler J, Greene BR, Pavlik CE, West MM, Zimmerman DL. Geocoding Health Data: The Use of Geographic Codes in Cancer Prevention and Control, Research, and Practice. Boca Raton: CRC Press, 2008.

[4] ZP4: Address and postal data on DVD-ROM. See http://www.semaphorecorp.com/cgi/zp4.html

[5] See: http://www.centrus.com/

[6] See: http://www.esri.com

[7] Postal Addressing Standards. Last accessed 28-Mar-2009 at: http://pe.usps.gov/text/pub28/28c2_001.html

[8] See Experian QAS Pro, available at: http://www.qas.com/address-verification-software.htm?tid=1500&tdet=8455032&ddcid=8455032&utm_source=google&utm_medium=ppc&utm_term=address_standardized&ddcid=8455032&gclid=CMWExrHdzZkCFSMSagod7ngYuA

[9] See: California Resources Agency, Geospatial Information Office. Last accessed 12-Dec-2008 at: http://gio.resources.ca.gov/.

[10] See: NADCON - North American Datum Conversion Utility. Last accessed 12-Dec-2008 at: http://www.ngs.noaa.gov/TOOLS/Nadcon/Nadcon.html.

[11] Available from Semaphore. See: http://www.semaphorecorp.com/cgi/zp4.html. Last accessed 12-Dec-2008.

[12] See: California Resources Agency, Geospatial Information Office. Last accessed 12-Dec-2008 at: http://gio.resources.ca.gov/.

[13] See: NADCON - North American Datum Conversion Utility. Last accessed 12-Dec-2008 at: http://www.ngs.noaa.gov/TOOLS/Nadcon/Nadcon.html.

[14] San Diego Geographic Information Source. Last accessed 09-Dec-2008 at: http://www.sangis.org/Download_GIS_Data.htm

[15] San Diego County Address Points. Available at: http://files.sangis.org/fileList_categorized.aspx?dirPath=D|\sangis_fileserver\file_store\Parcels+and+Lots

[16] Remy L, Clay T, Oliva G. (Aug 2000). The California Child and Youth Injury Hot Spot Project Report for the Period 1995 to 1997, Volume Three, Technical Guide. Sacramento, CA: California Department of Health Services. See: http://www.ucsf.edu/fhop/_htm/publications/index.htm.

[17] Remy L, Oliva G, Clay T (2008) Maternal morbidity and outcomes including mortality, California 2001-2006. Family Health Outcomes Project, University of California San Francisco. Available at: http://www.ucsf.edu/publications.html

[18] Dudley RA, Kuzniewicz M, Dean M, Lane R, Rennie D, Bacchetti P, Clay T, Crane S, Luft H. (May 2007) Final Report, California Intensive Care Outcomes (CALICO) Project. Office of Statewide Health Planning and Development. Last accessed 06-Mar-2008 at: http://www.oshpd.ca.gov/HID/Products/PatDischargeData/ICUDataCALICO/index.html