# MANAGING LONGITUDINAL RESEARCH STUDIES:

# THE GEOGRAPHY MASTER

By

Linda L Remy, MSW PhD

Ted Clay, MS

Rita Shiau, MPH

Geraldine Oliva, MD MPH, Director
Jennifer Rienks, PhD, Associate Director
Linda L Remy, MSW PhD, Research Director

UCSF Family Health Outcomes Project
500 Parnassus Ave. Room MU-337
San Francisco, California 94143-0900
Phone: 415-476-5283
Fax: 415-476-6051
Web: https://fhop.ucsf.edu/

July 2020

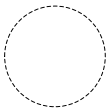
# Table of Contents

# Table of Figures

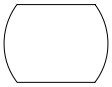
## Table of Tables



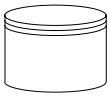
## Table of Legends



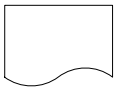
From program creating incoming file or name of next program using file



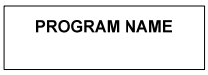
Excel file input or output



SAS file input or output



Flat or text file input or output




SAS program with brief description of steps.

# Table of Abbreviations

| | |
|---|---|
| ACLAIMS | Automated Certification and Licensing Administrative Information and Management Systems |
| ASC | Ambulatory Surgery Center dataset, CA OSHPD |
| BSMF | Birth Statistical Master File, California Vital Statistics |
| CDP | Census Designated Place |
| CDPH | California Department of Public Health |
| DSMF | Death Certificate, California Vital Statistics |
| ED | Emergency Department dataset, CA OSHPD |
| FDSMF | Fetal Death Statistical Master File, California Vital Statistics |
| FIPS | Federal Information Processing Series |
| HADR | Hospital Annual Discloser Report, California Department of Public Health |
| HFPA | Health Facility Planning Area |
| HSA | Health Service Area |
| OSHPD | Office of Statewide Health Planning and Development |
| PDD | Patient Discharge Data, CA OSHPD |
| POB | Post Office Box |
| SPA | Service Planning Area |
| USPS | United States Postal Service |
| ZCTA | Zone Improvement Plan Code Tabulation Areas |
| ZIP | Zone Improvement Plan (Code) |

## Suggested Citation

Remy LL, Clay T, Shiau R. (2020) Managing Longitudinal Research Studies: Methods to Prepare the Geography Master. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

# THE GEOGRAPHY MASTER

## THE NEED FOR LONGITUDINAL GEOGRAPHIC DATASETS

The various datasets FHOP uses in its research encompass key developmental events for generations of California families, now over a span of four decades. A central piece of FHOP's research agenda is to examine population health outcomes for these families and the communities where they live, both longitudinally and geographically.

In this context, a fundamental task is to define where a given person lives at a given point in time, in order to calculate population-based rates. Although every residence is located at a specific latitude and longitude, few population databases permit such exquisite precision, to protect the identity of people who live(d) there. Further, a given person may live at multiple places over time, and multiple people may live at the same place over time. Finally, although residences remain stationary, administrative boundaries superimposed over collections of addresses (United States Postal Service (USPS) Zone Improvement Plan Codes (ZIP), census tracts, congressional districts) may retain the same designation in name, yet reflect the same or completely different geographic boundaries over time. Hospitals may change addresses from a block on one side of the street to a block on another side in a completely different ZIP, or move to a new site in the same or another city, or to a new county. Because of its complexity, an entire sector of commercial vendors exists to track this geographic environment to help researchers understand these changes.

With data entry errors in population datasets, changes in USPS ZIPs, and changes in commercial files, maintaining geographic longitudinal consistency has been a challenging task. The purpose of this monograph is to describe important types of geography-related variables, and then focus on the methods FHOP uses to develop and maintain our Geography Master.

Table 1 summarizes the types of geographic variables available from different data sources, and the years for which these variables are available. Column 1 identifies the data source, column 2 the dataset, and column 3 the period the variable is available. In the remaining columns, “X” denotes that the variable is present and “C” denotes that it is calculable from other information. ZIP type indicates whether the ZIP is a post-office (PO) box, unique address, or other. Parent indicates whether another ZIP (the “parent” ZIP) encloses the ZIP of interest, for example, a PO-box-only ZIP enclosed by a larger geographic area. Lat(itude) and Lon(gitude) for ZIP centroids are available in commercial data and can be calculated for other data.

The next section describes various geographic variables, what we know about their history, and caveats about using them. Then we present detailed descriptions of our methods to create our Geography Master.

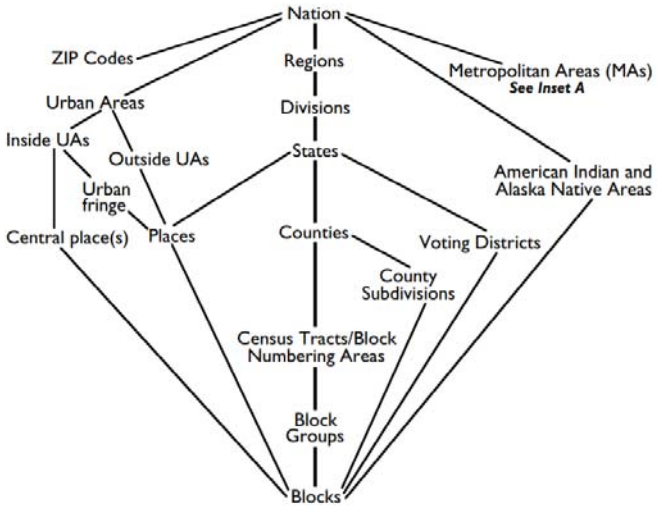
## Table 1. Geography variables available (X) or calculable (C) in FHOP datasets

| | | | ZIP | | | | Local | | | NCHS | | US Census | | | | | | | Health Planning | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Source | Dataset | Period | Code | Type | Parent | Lat/Lon | Address | City | County | State | Nation | Block | Group | Tract | Place | Reg | Div | World | HSA | HFPA | MCAH | SPA |
| Office of Statewide Health Planning and Development | Discharge | 1983-2000 | X | | | C | | C | X | | | | | | | | | | X | X | C | C |
| | | 2001-Present | X | | | C | | C | X | | | | | | | | | | C | C | C | C |
| | Emergency | 2005-Present | X | | | C | | C | X | | | | | | | | | | C | C | C | C |
| | Ambulatory | 2005-Present | X | | | C | | C | X | | | | | | | | | | C | C | C | C |
| | Annual Disclosure | 1981-Present | X | | | C | X | X | X | | | | | | | | | | C | C | C | C |
| California Department of Public Health | Birth Hospital Codes | 1970-Present | X | | | C | X | X | X | | | | | | | | | | | | | |
| | Birth Master | 1989-1996 | X | | | C | | C | X | X | X | | | | X | C | C | C | C | C | C | C |
| | | 1997-Present | X | | | C | X | X | X | X | X | | | | X | C | C | C | C | C | C | C |
| | Death Master | 1989-2004 | X | | | C | | C | X | X | X | | | | X | C | C | C | C | C | C | C |
| | | 2005-Present | X | | | C | X | X | X | X | X | | | | X | C | C | C | C | C | C | C |
| | Fetal Death Master | 1989-2012 | X | | | C | | C | X | X | X | | | | X | C | C | C | C | C | C | C |
| Commercial | SAS Institute | 2004-Present | X | | | X | | C | X | | | | | | | | | | | | | |
| | ZIPInfo | 2005, 2009, 2011, 2013 2015, 2017 2020 | X | X | X | X | | X | X | | | | | | | C | C | C | C | C | C | C |
| | ESRI | 2009, 2012, 2015, 2019 | X | X | | X | | X | X | X | X | X | X | X | X | X | C | X | C | C | C | C |
| No longer available | Douglas Boynton Quine | 1963-2005 | X | | | | | X | X | X | | | | | | C | C | | C | C | C | |
| | Claritas | 1998, 2009 | X | X | X | | | C | X | | | | | | | | | | C | C | C | C |
| | Western Economic Res | 1990-2012 | X | X | | | | C | C | X | X | | | | | C | C | C | C | C | C | C |

# STANDARD ADMINISTRATIVE BOUNDARIES

Standard administrative boundaries are a limit or border of a geographic area under the jurisdiction of a governmental or managerial entity. Governmental or quasi-governmental agencies assign these boundaries. States and counties are the major legally-defined political and administrative units of the United States and are the primary geographic units for which the Census Bureau reports data [1]. Other geographic variables include census tracts; block groups or blocks assigned by the U.S. Census; ZIP codes and post office boxes assigned by USPS; and city, town or special agency boundaries that change over time due to annexation or partition. As Figure 1 shows, most geographic entities have a superior/subordinate structure derived from their legal, administrative, or areal relationships [2].

Figure 1. Standard hierarchy of census geographic entities




An example structure is the standard census geographic hierarchy within the United States from lowest to highest granularity: block, block group, tract, county, state, division, region. Special districts such as congressional or school districts consist of different combinations of standard census boundaries that often do not coincide with the standard census hierarchy. Note that in the standard hierarchy, ZIP codes and metropolitan areas are not related because they may cross state and county boundaries.

In addition to geographies designated by the Census Bureau, the National Institute of Standards and Technology issues a set of numeric or alphabetic codes called Federal Information Processing Series (FIPS) codes [3] to ensure uniform identification of geographic entities through all federal government agencies. FIPS codes cover “states and statistically equivalent entities, counties and statistically equivalent entities, named populated and related location entities (such as, places and county subdivisions), and American Indian and Alaska Native areas, and foreign nations”. For example, Puerto Rico and the Outlying Areas are considered statistical equivalents of states. FIPS codes are used in many private, industry and public datasets, and sometimes provide different coverage compared to Census Bureau geographies [4]. We include formats for FIPS codes in our geography format library.

## Census Regions and Divisions

Figure 2. Census regions and divisions


**CENSUS REGIONS** are groupings of states and the District of Columbia that subdivide the United States for the presentation of census data. There are four census regions: Northeast (1), Midwest (2), South (3), and West (4). California is in Region 4. Each region is divided into two or more census divisions [5].

**CENSUS DIVISIONS** are groupings of states and the District of Columbia subdivided under the four census regions. There are nine census divisions.

Figure 2 shows the United States divided into regions, divisions, and states. California is located in Region 4 (Western), Division 9 (Pacific) [6]. California's abbreviation is “CA”. Its state FIPS code is 06. The geography format library includes formats for divisions and regions.

## County

States and counties are the major legally-defined political and administrative units of the United States [7]. California has 58 counties, with one encompassing both a county and city (San Francisco). The federal government assigns 3-digit county FIPS codes based on alphabetical order within a state. With the first two digits representing the state, Marin County has FIPS code 06041. The State of California also assigns county codes using the underlying alphabetical order in its administrative datasets. In alphabetical order, Marin County is California’s twenty-first county. The geography format library includes formats for various ways to report national and California county and FIPS codes.

When crosswalking between federal and California administrative datasets such as hospital discharges or births, it is possible to identify the California administrative number using this equation:

California County Number = (3-digit Federal County FIPS Number + 1) / 2
Marin County = (041 + 1)/2 = 21

## ZIP Code

ZIP Code is an acronym for the 5-digit Zone Improvement Plan (ZIP) Code. The USPS developed it to facilitate sorting mail for delivery, assigning a five-digit code to every address in the nation. It went into effect on July 1, 1963 [8].

The basic ZIP consists of five numeric digits. The first digit with values from 0 to 9 represents a U.S. region: “0” for the northeastern U.S. through “9” for the western states. The second and third digits together represent a region, a large city, or an area accessible to transportation networks. The fourth and fifth digits represent more specific areas, such as small towns or postal zones in large cities. The main town in a region (if applicable) typically was assigned the first ZIP for the region; then the numerical order often follows the alphabetical order within the region.

In this convention, California ZIPs begin with “9”. Marin County is designated by “949”, and local communities are given numbers under this prefix. The main town in a region typically gets the first ZIP for the region; then numeric order typically follows alphabetic order within the region. In Marin, when ZIPs first were applied, San Rafael was the largest city, designated “94901”. If someone lives in a multi-ZIP city, but the exact ZIP is unknown, one may find ZIPs with only the city prefix listed, such as “94900”, in population health files.

The United States has four types of ZIP Codes: unique (assigned to a single high-volume mailer), post office box-only (used only for PO boxes located at a given location, not for any other type of delivery), standard (all other U.S.-based ZIP codes), and military (assigned by the U.S. Military Postal Service to deliver mail to personnel serving overseas). Certain governmental agencies, universities, businesses, or buildings with extremely high volumes of mail have their own unique number. For example, 94143 is a unique ZIP Code for the University of California, San Francisco where FHOP is located. An example of a PO box-only ZIP is 94915, used for boxes at the main post office in San Rafael. In the area surrounding that post office, homes and businesses use ZIP 94901, a standard ZIP. This standard ZIP is sometimes called a “parent” ZIP to the PO-box-only ZIP.

Despite the geographic derivation of most ZIPs, the codes themselves are not geographic regions, but simply categories for grouping mailing addresses. ZIP boundaries change over time depending on the needs of USPS mail delivery routes. ZIP “areas” can overlap, be subsets of each other, or be artificial constructs with no geographic area. The USPS splits ZIPs as addresses increase. Sometimes USPS discontinues a ZIP and issue two new ones, other times assigning the old ZIP for a portion of the region and creating a new ZIP for the remainder. After not using a given ZIP for some time, it may reissue the number and assign it to a completely different geography. All of this makes assigning cases to a specific “geographic area” based on

ZIPs somewhat akin to an exercise in finding a moving target; the same code in 1983 may point to a different place in 2020.

ZIPs can be wholly contained within counties or span county boundaries. Consider ZIP 94952, where 90% of the human population is in Petaluma (Sonoma County) and 90% of the area is in Marin County and populated mainly by dairy cows. In areas without regular postal routes (rural route areas) or no mail delivery (undeveloped areas), ZIPs are based on sparse delivery routes, and hence the boundary between ZIP Code areas is undefined.

A given address with a ZIP and associated “city” name do not necessarily mean the address is located within the boundaries of that city. The USPS designates a single “preferred” place name, or official USPS designation, for each ZIP. This may be an actual incorporated town or city, a sub-entity of a town or city, or a “census designated place” (CDP). In addition to “preferred” names, other place names may be recognized as “acceptable” for a certain ZIP. For example, “San Anselmo” (incorporated) with ZIP 94960 is an “acceptable” place name for 94901, with a preferred name of “San Rafael” (incorporated), while Greenbrae and Kentfield (both unincorporated) are assigned to ZIP 94904 with preferred name “Greenbrae”.

“Acceptable” place names also occur where ZIP boundaries include two or more cities. For example, ZIP 94920 encompasses the incorporated cities of Belvedere and Tiburon, but primary/preferred names is “Belvedere-Tiburon”. Belvedere has precedence in the hyphenated name because it has a post office and Tiburon does not [9]. “Acceptable” names are Bel Tiburon, Belvedere, and Tiburon. Finally, many ZIPs are for villages, census places, portions of cities, or other entities that are not municipalities. Marin County examples include Woodacre, Stinson Beach, and Inverness.

Ideally, the “preferred” name would be the actual city or town where the address is located. However, many cities have incorporated since ZIPs were introduced. As a result, the legal city name is only “acceptable” and not the “preferred” place name. Many databases automatically assign the “preferred” place name for a ZIP, disregarding “acceptable” place names. For example, Centennial, Colorado, the largest city to incorporate in U.S. history, is divided between seven ZIP codes, each of which either has “Aurora”, “Englewood” or “Littleton” as its “preferred” place name. Thus, postally speaking, the city of Centennial and its 112,151 residents do not exist - they are postally part of Aurora, Englewood or Littleton. In the ZIP code directory, Centennial addresses are listed under these three names. And since it is “acceptable” to write “Centennial” in conjunction with any of the seven ZIP codes, one can write “Centennial” in an address actually in Aurora, Englewood, or Littleton, as long as it is in one of the shared ZIPs.

Like telephone area codes, the USPS sometimes discontinues, divides and changes ZIPs. In rapidly-developing suburbs, it is sometimes necessary to open a new sectional center facility. This must be allocated its own three-digit ZIP-code prefix, changing the ZIPs of all communities

associated with that center. ZIPs also change when the USPS realigns postal boundaries. Thus, a Novato address that was in ZIP 94945 in 1987 may be in 94949 in 2007.

Delivery services such as Federal Express, United Parcel Service, or DHL require ZIP Codes to route packages. ZIPs are used not only to track mail and package delivery, but also to gather geographic statistics. Many companies sell ZIP-related data with current boundaries, latitude and longitude. Others track the starting, stopping, and splitting of ZIPs.

ZIP information is used widely outside of USPS for various purposes. Point-of-sale cashiers sometimes ask consumers to identify the ZIP where they live, in order to collect purchasing pattern data. Companies analyze these data to identify a potential location of new branches or to target advertising campaigns. The insurance and banking industries have historically "redlined" certain ZIPs, making it more difficult for members of redlined communities to obtain credit or insurance; the myriad adverse effects of this practice on residents of these ZIPs are still seen today [10]. Health researchers such as FHOP use ZIPs to identify "hot spots", communities with high rates of adverse health conditions [11].

## ZIP Code Tabulation Area

Developed by the Census Bureau in 2000, ZIP Code Tabulation Areas (ZCTAs) are generalized areal representations of USPS ZIP Code service areas that approximate the delivery area for a USPS five- or three-digit ZIP [12]. Prior to ZCTA establishment, the decennial census used ZIP Codes as a level of data tabulation. ZCTA are aggregations of census blocks that have the same predominant ZIP Code associated with the residential mailing addresses in the Census Bureau Master Address File. The most commonly-found ZIP code among the aggregated blocks is assigned as the ZCTA number. In 2000, three-digit ZCTAs were assigned to large contiguous areas for which the Census Bureau does not have five-digit ZIP information in its Master Address File (e.g. national parks, uninhabited land, or bodies of water); in 2010, only five-digit ZCTAs were assigned. ZCTA do not precisely depict ZIP delivery areas, and do not include all ZIPs used for mail delivery. ZIP codes assigned to businesses only, single delivery point address, or areas with very few addresses will not necessarily be assigned ZCTAs. While the ZCTA code is typically the same as the ZIP code for an area, ZCTAs tend to be more stable over time because they may only change on decennial census years, whereas ZIP code areas may be adjusted more frequently according to USPS needs. However, there are significant differences within and between states in the extent to which ZCTA and ZIP coincide geographically [13].

We found a problem with ZCTAs when trying to construct maps in Marin County, California, an area with many PO Box-only ZIPs. We do not know how often this problem occurs, but we thought it prudent to bring it to the reader's attention, so they can be aware of this possibility. The Town of Ross is a small, incorporated community in the geographic heart of Marin County,

with a PO-Box-only ZIP of 94957. Residents need to go to the post office to retrieve their mail. It is surrounded by three cities whose ZIPs have street delivery addresses. When constructing the ZCTA for this area, the Census “disappeared” Ross as a governmental entity and apportioned its area and population into the three surrounding communities. It required a great deal of work to extract the underlying boundaries and population from the surrounding cities where the Ross population had been assigned, reform boundaries and sum counts, and recalculate the boundaries and populations for the communities into which Ross had been “disappeared”.

## Census Tracts

Census tracts are small, relatively permanent statistical subdivisions of a county [14]. The 2000 Census was the first to cover the entire United States by census tracts. These generally have between 1,500 and 8,000 people, with an optimum size of 4,000 people. Tracts were designed to be relatively homogeneous with respect to population characteristics, economic status, and living conditions. The spatial size of census tracts varies widely depending on population density. Their boundaries are intended to remain stable over decades to enable statistical comparisons from census to census. The Town of Ross has the census tract number 118100. However, changes in street patterns caused by highway construction, new developments, and so forth, require occasional boundary revisions. In addition, tracts occasionally are split due to population growth or combined due to population decline. For example, the census tract 102201 present in Novato, California, for the 2000 Census was split into two smaller census tracts (102202 and 102203) for the 2010 Census, with census tract number 102201 retired in 2010.

Some vital statistics datasets include census tracts. However, some of the census tracts assigned may be outdated, so we recommend not using census tract information in California administrative datasets. If a clean address is available, a good geocoding program can assign current tracts more accurately.

## Census Block Groups

After census tracts, the next level of granularity in the hierarchy is the census block group. Block groups consist of all census blocks having the same first digit of their four-digit identifying numbers within a census tract [15]. For example, Block Group 3 within a census tract includes all blocks numbered from 300 to 399. Block Groups generally contain between 600 and 3,000 people, with an optimum size of 1,500 people. This is the smallest geographical unit for which the Census Bureau publishes sample data. The Town of Ross, Marin County, California is divided into two Block Groups: 1181001 and 1181002. No California administrative dataset includes census block groups. If needed for analysis, one may use geocoding software to assign census block groups from latitude and longitude.

## Census Blocks

Census blocks generally are small areas bounded on all sides by visible features (streets, roads, streams, and railroad tracks) or invisible boundaries (city, town, township, and county limits, property lines, and short, imaginary extensions of streets and roads) [15]. However, in sparsely settled rural areas, they may contain many square miles of territory. Census blocks are the smallest geographic area for which the Census Bureau summarizes complete count data, and, since the 1990 Census, completely cover the U.S. geography. Census block boundaries are available and can be used as a layer for health studies based on addresses geocoded to latitude and longitude. The researcher needs to be conscious of and develop rules to address small number problems that may arise at this geographic layer.

## Census Place

A census place is a locally-recognized concentration of population that is either legally incorporated under the laws of its state, or a statistical equivalence that the Census Bureau treats as a census designated place [16]. Places are mutually exclusive and are made up of census blocks. In California, places do not cross county lines. Census places do not cover the entire U.S. geographically, and their boundaries and names change as needed between census years when municipalities incorporate, unincorporate, merge, or are renamed.

The Census Bureau assigns place codes to all places within a state in alphabetical order; gaps were present in the initial numbering scheme to maintain alphabetical order in future place designations. Through the 1990 Census, place codes were four digits long; now they are five. California Vital Statistics added place codes to birth and death files in 1984. Complicating matters further, the agency did not uniformly apply these codes to coincide with census-based intervals. The Death Statistical Master File (DSMF) used the 4-byte code from 1985 to 2002, then switched to the 5-byte code in 2003. The Birth Statistical Master File (BSMF) and Fetal Death Statistical Master File (FDSMF) used the 4-byte code from 1985-1997, a 3-byte code from 1998-2002, and the 5-byte code from 2003 onward. Because of accumulating data quality problems in these Vital Statistics files, the same place code has been assigned to different cities, ZIPs, and counties.

FHOP has an Excel file and SAS code to crosswalk longitudinally between 4- and 5-byte census place codes and ZIPs (PLACE.XLSX, available on request).

## Health Service Areas and Service Planning Areas

Health Service Areas (HSAs) were originally defined by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention, to be a single county or cluster

of contiguous counties that are relatively self-contained with respect to hospital care. HSAs have quality, accessibility, continuity, and cost containment as their major goals [17]. Originally developed in the 1970s [18], HSA boundaries are updated as underlying demographic and health service characteristics change [19,20].

We support modifications that the National Cancer Institute (NCI) made to HSAs, splitting any that crossed state or Surveillance, Epidemiology, and End Results (SEER) Registry boundaries so all counties from one HSA were in one state and/or SEER Registry [21]. We follow this same rule for ZIPS that split counties, assigning them to the county with the largest population and not allowing ZIPS to split counties.

Until the early 1980s, when the state legislature disbanded them, California had Health Facility Planning Areas (HFPA) to facilitate development of regional hospital-related services. We have written extensively about the consequences of the failure of HSA/HFPA for California's mental health services programs [22] and more recently about the impact on emergency response [23]. From 1987 to 1995, California took progressive steps to deregulate and repeal health planning from its legislation, eventually abandoning the HSA/HFPA infrastructure entirely. To facilitate service delivery for maternal, child and infant health, the California Department of Public Health subsequently grouped its 58 counties into ten demographically similar super-regions [24]. These regions are frequently used in data analysis and summaries of health outcomes.

In addition to the ten super-regions, some counties also further divide into sub-regions to facilitate health service planning and delivery. For example, Los Angeles County divides itself into eight Service Planning Areas (SPA) [25]. Our geography format library has various SAS formats to associate ZIPs to SPAs and other California sub-region designations. For health jurisdictions without formally defined SPAs, we first use districts defined in any health department reports such as community health assessments to define SPAs [26, 27]. If no such districts exist, approximations of county supervisor districts are used to define SPAs. Also note that county supervisor district boundaries change with censuses, as California law requires the number of people in the county to be distributed equally across the districts.

# GEOGRAPHIC DATASETS

## Commercial

A wide variety of commercial vendors provide datasets identifying ZIPs, cities, counties, ZIP types, parent ZIPS and other features. Different vendors include different geographic features. As Table 1 identifies, some vendors went out-of-business and we had to find new vendors. In any event, we keep all historic files because we need their information to document changes in geography over time.

As described in previous sections, the USPS regularly splits ZIPS, changes their types, or assigns new numbers for essentially the same geography. ZIPS that existed in 1980 may or may not still exist in 2020. What current ZIP encompasses the homes previously assigned to the 1980 ZIP that no longer exists?

A ZIP with certain boundaries has a specific ZIP centroid latitude and longitude. Centroids change over time when the USPS splits a ZIP, retaining the original ZIP number for one portion and assigning a new number for the new second portion, or reassigns part of a ZIP to another existing ZIP. In some cases, because of changing development or population density in an area, a ZIP's shape may change dramatically, also leading to changes in centroid values over time. Boundaries that work at one time for a given ZIP number will not work at another time for the same ZIP number that the USPS assigned to different geography.

Commercial geography datasets are critical resources to begin answering such complex questions. Because of differences in file structure, SAS programs developed for one vendor at one time do not work for data from that same vendor at another time, or from other vendors. We describe how these various files are standardized for our use in later sections of this monograph.

## Population Health

The California population health datasets we use to make our geographic master file include a variety of geographic indicators. Some are available in all files and all years, others are available in some files in some years.

From the Office of Statewide Health Planning and Development (OSHPD), FHOP uses Patient Discharge Data (PDD) available from the 1983 onward, and Emergency Department (ED) and Ambulatory Surgery Center (ASC) data available from 2005 onward. These large annual files always have the hospital identifier OSHPDID and hospital county. Depending on year, they also contain various combinations of geographic variables for both patients and hospitals: ZIP, county, HSA, and HFPA.

From the California Department of Public Health (CDPH), we use birth, death, and fetal death files from 1983 onward. Depending on year, these files include street address, city and/or FIPS city/place codes, ZIP, and county. Birth and fetal death files always have the CDPH hospital identifier (HOSPCODE) and hospital county (county where infant was delivered). Death files always have county of residence and of death.

When we receive annual updates of these important data, we summarize available geography-related variables as part of our standard processing of files entering our system. A program FRQ.SAS calls a source-specific macro, for example FRQOSH.SAS, to summarize geography-related variables as well as other variables we monitor for changes and quality.

## Health Facilities

FHOP uses two main sources to define hospital geography. The California Department of Public Health (CDPH) annually updates an Excel file with the HOSPCODE that CDPH assigns to hospitals to use in recording births. This file, distributed with the annual birth datasets, includes the hospital name and address. The first HOSPCODE assignments were alphabetical based on hospitals open in 1970, the year CDPH licensing started. CDPH assigned subsequent HOSPCODE as facilities began to provide care. Unfortunately, CDPH does not update address or name changes regularly. However, we still consider it a valuable resource because it has identifiers back to the 1970s and thus is the earliest available file of hospital names and addresses. Also, unlike files from OSHPD which is our other source of hospital geography data, the CDPH Excel file includes identifiers for military hospitals and birthing centers, both of which are exempt from reporting to OSHPD.

OSHPD started assigning OSHPDIDs to hospitals in the 1970's as part of California's early health planning activities during the first term of Governor Jerry Brown [28]. Per California regulations, hospital licenses are based on a given physical location. Each major unit in the hospital is given a first-digit identifier (inpatient, emergency department, day surgery, psychiatric, rehabilitation), then a 2-digit number based on county where the facility is located, then sequential numbers within the county. Sub-units within the hospital (laboratories, delivery units, etc.) are assigned numbers through the Licensing and Certification Division, California Department of Health Services.

We use both of these data sources to document changes to hospital IDs in different datasets longitudinally. When hospitals "disappear" from various data files, the explanation is not readily apparent. We must determine if it is because the facility closed, merged, converted to consolidated reporting, or moved, resulting in a new license ID. Yet another possibility is that a new license ID was assigned to a facility at the same location. Some hospitals move without getting new IDs. We discuss these complexities and how we handle them in two related reports [29, 30].

In addition to our two primary source of hospital geography, every licensed hospital is required to file the Hospital Annual Disclosure Report (HADR), summarizing its financial and utilization characteristics by unit. Unlike encounter files (PDD, ED, ASC), released based on calendar year, HADR data is based on fiscal year (July-June). About half of hospitals have a fiscal year that coincides with calendar year. In a given year, a hospital may file one to *n* HADR, depending on changes in fiscal year, ownership, moving, closing and/or reopening. We have these from fiscal year 1981 onward [31].

The HADR does not identify all California hospitals with OSHPD licenses. Some submit consolidated data for multiple facilities under one identifier. Others, for example Kaiser, are

exempt from filing some or all pages. As a result, the number of records varies, depending on reporting requirements. Despite its complexities and deficiencies, the HADR provides the most detailed information available about overall structure (e.g. location, ownership, auspices, services, beds, staffing) and detailed financial information by licensed hospital unit.

For purposes of geography, we use HADR Page 0, which contains information about the current and previous hospital name and owner, and the physical address, mailing address, and address of the person completing the report [32]. Having this longitudinal resource of address data from OSHPD has always been a valuable resource.

One cannot know from the HADR how many hospitals operate in California at a given time. For this information, one needs the OSHPD patient data. Because of complex parent/subsidiary changes over time, we assign all subsidiaries to the current parent when we do hospital-level longitudinal studies. A related report describes the steps to resolve the OSHPD hospital identifiers and crosswalk them to the CDPH hospital identifiers [30].

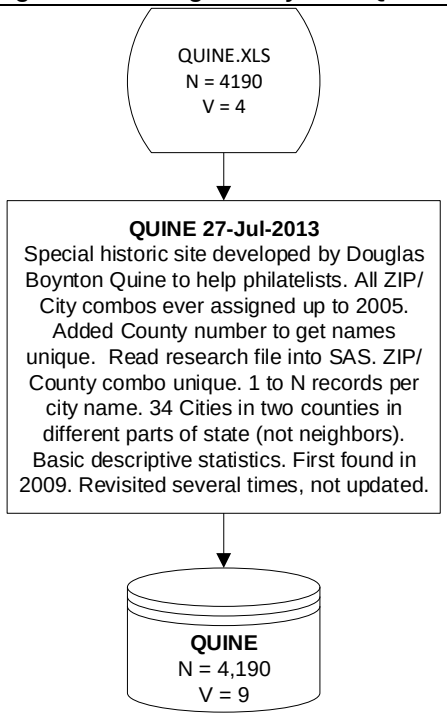
In recent years, OSHPD has been including updates on hospital moves, openings and closings when it sends the patient files. It also now sends an annual file reporting what it has identified as current names for each hospital. From these and historic sources, we maintain a longitudinal dataset that summarizes patient activity by OSHPDID, using the current (or last known) name for the facility.

# STANDARDIZE COMMERCIAL GEOGRAPHY

## Prepare Discontinued Resources

Here we describe how we processed discontinued commercial geography datasets. The figure title identifies the name of the data source. The first symbol shows the name of the incoming dataset together with the number of records and variables. The figure box summarizes major program steps. The first line in the box identifies the program name and last time we updated the program. After the description, we show the name of the outgoing file(s), usually SAS files and sometimes Excel files. Notice that we use different symbols for non-SAS and SAS files.

Figure 3: Douglas Boynton Quine




**Douglas Boynton Quine** is a well-known scientist and enthusiastic philatelist who worked at Pitney Bowes for many years. His website has a special section for philatelists that crosswalks postal codes to place names [33]. We found his site in 2009 and he apparently discontinued editing ZIPs in 2005. It purportedly contains all ZIPs from 1963 forward, making it the oldest source we have found for historic ZIPs and their locations. We revisited the site several times over the years, but it has not been updated. This clearly is not a commercial source, but we classify it as such for this monograph.

We downloaded all ZIPs in the California range 90000-96162 (N = 4,190) and imported it into Excel. The list has 2,692 ZIPs with all ZIPs in unique counties, with 1 to N ZIP-level records per city name. Thirty-four cities were assigned to two counties, with 16 assigned to two neighboring counties, and 18 to two non-neighboring counties.

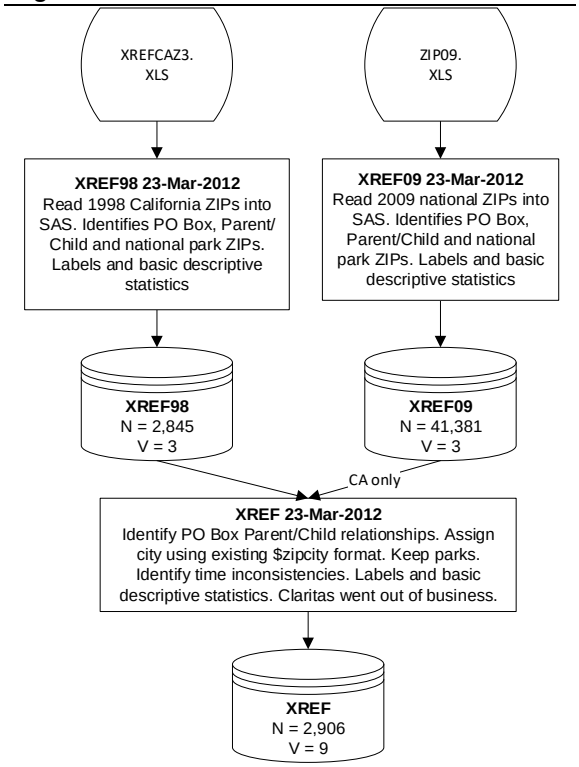
Figure 4: Western Economics Research


We purchased the Western Economics Research ZIP-change file semi-annually from 1995 until 2013, when the owners retired. This file was invaluable to track changes in ZIPs over time, and we remain sorry that Western Economics Research closed. We have not been able to locate an equivalent replacement.

The file structure is chronological from 01-Jul-1990 until 03-Mar-2012. Figure 4 describes preparation of these data. Each new record identifies the date the USPS introduced, renumbered, or split ZIPs. A comments field describes the nature of the change, including how split ZIPs are reassigned to new “replacement” ZIPs.

The SAS program ZCC2013 reviewed the information on the incoming record (N = 519) and made a new record for each affected ZIP (N = 1,019).

Figure 5: Claritas

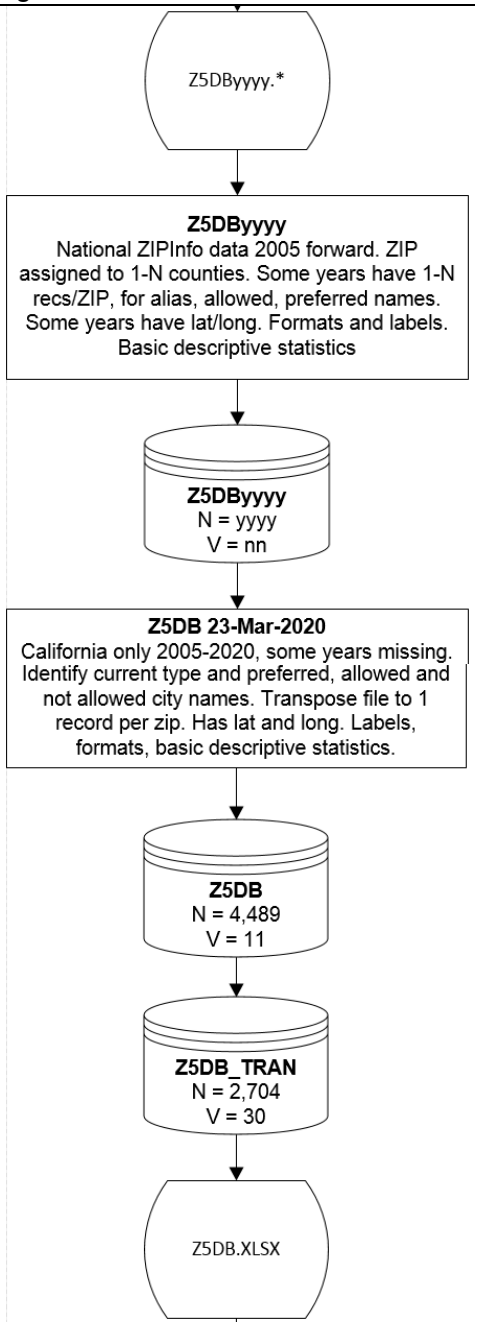



We first purchased the 1998 Claritas file for our injury study that involved linking episodes of care longitudinally and preparing small area maps at the ZIP-code level [11]. This file only had California ZIPs (N = 2,845). Figure 5 shows that when we again turned to Claritas for 2009 files, we obtained national data (N = 41,381). These files have only 3 variables: ZIP, parent ZIP, and ZIP type.

When we merged the files by ZIP, we excluded national park ZIPs and known military Army Post Office ZIPs. City names were assigned using our crosswalk format $ZIPCITY, after updating the Geography Master described later. When the ZIP existed in both files, we assigned the parent ZIP from 2009, if present. Otherwise, we used the existing parent in the year the ZIP appeared. We used similar rules to assign PO box status. The final file XREF had 2,906 records and 9 variables.

## Prepare Current Resources

Figure 6: ZIP Info



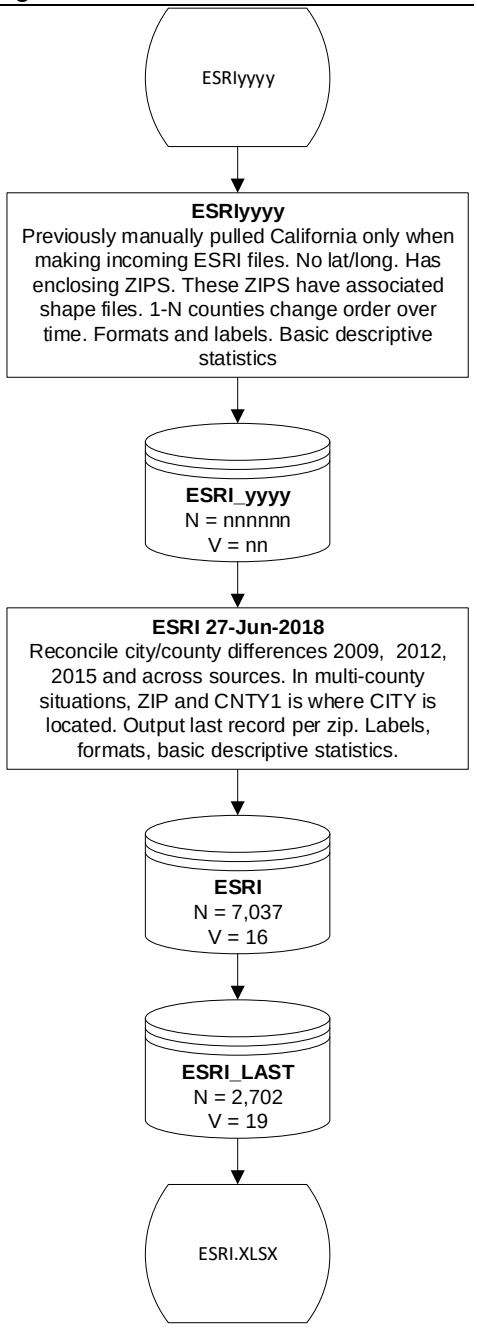

About biannually, we purchase ZipList5 files from the ZIP Info website [34]. These files download as text and we import them into Excel to prepare for SAS. Figure 6 shows the processing steps for these files.

ZipList5 has USPS preferred, alternate, and not allowed city names as reported at the time. Some years contain the latitude and longitude of the ZIP centroid, which can change as ZIP boundaries change. Because of annual structural differences, the programs Z5DByyyy read each yearly file into SAS, attaching formats and labels as needed.

The next program Z5DB pulls California data from the national files (Z5DB, N = 4,489, V = 11). It reconciles city and county differences over time, transforms the file to 1 record per ZIP, with 1 to N city names and 1 to N counties per ZIP (Z5DB_TRAN, N = 2,704, V = 30).

For each ZIP, we use the most recent preferred city name and county associated with that city. The file identifies the last year the ZIP existed. The file is output to Excel for manual review.

Figure 7: ESRI



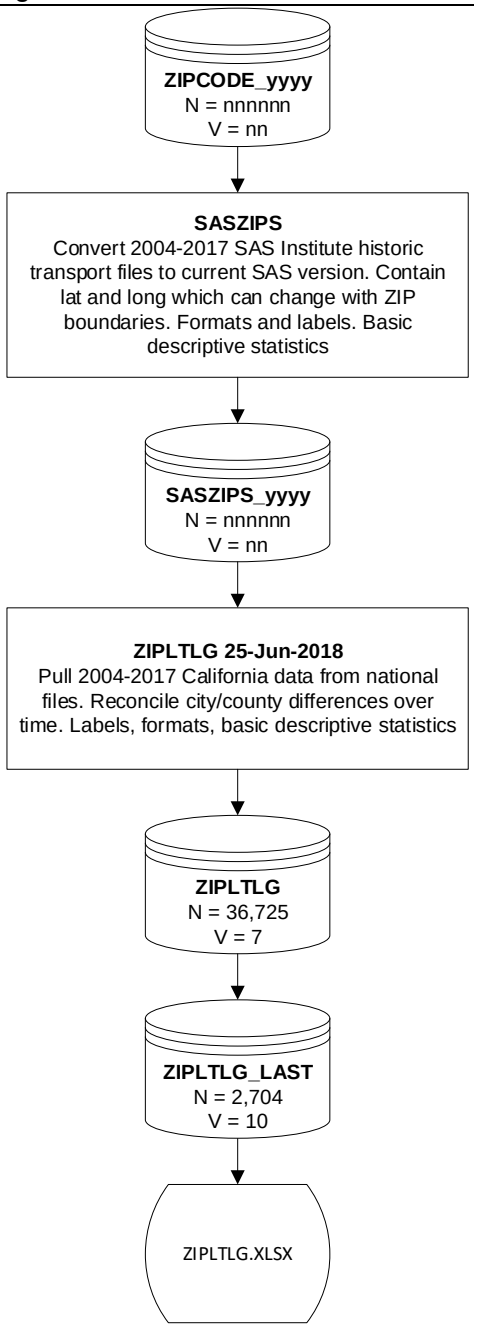

We have been processing ESRI files about every three years since 2009 [35]. In our initial processing, we only pull California records. These files have different incoming structures and different contents in different years. They do not have Lat/Lon, but they do have associated boundary files. Figure 7 shows steps for ESRI files.

In every year, ZIP is unique for each entry, but each ZIP can be assigned up to three counties (the incoming file has three county columns). Sometimes the city name assigned to the ZIP changes, going from preferred name to allowed name or the reverse. Sometimes the counties to which a ZIP is assigned are presented in a different order compared to previous files. The programs ESRIyyyy read these into SAS, add labels and attach formats as needed, with basic descriptive statistics.

The program ESRI reconciles city and county differences over time. It corrects misspelled city names. Some ZIPs span multiple counties; in such cases of “split” ZIPs, the ZIP is assigned to one city (cities do not span multiple counties), and then the county where the city is located. Where city names changed over time, the program assigns the most current name found. The resulting file ESRI has 1 record per ZIP and year (N = 7,037, V = 16).

The record for the last year the ZIP is found is output to a SAS file (ESRI_LAST, N = 2,702, V = 19) then to an Excel file for manual review.

Figure 8: SAS Institute




We annually download transport files (2004-present) from the SAS Institute website [36]. These files contain Lat/Lon for ZIP centroids, which can change as ZIP boundaries change. SASZIPS, a SAS macro program, converts them to SAS7BDAT files.

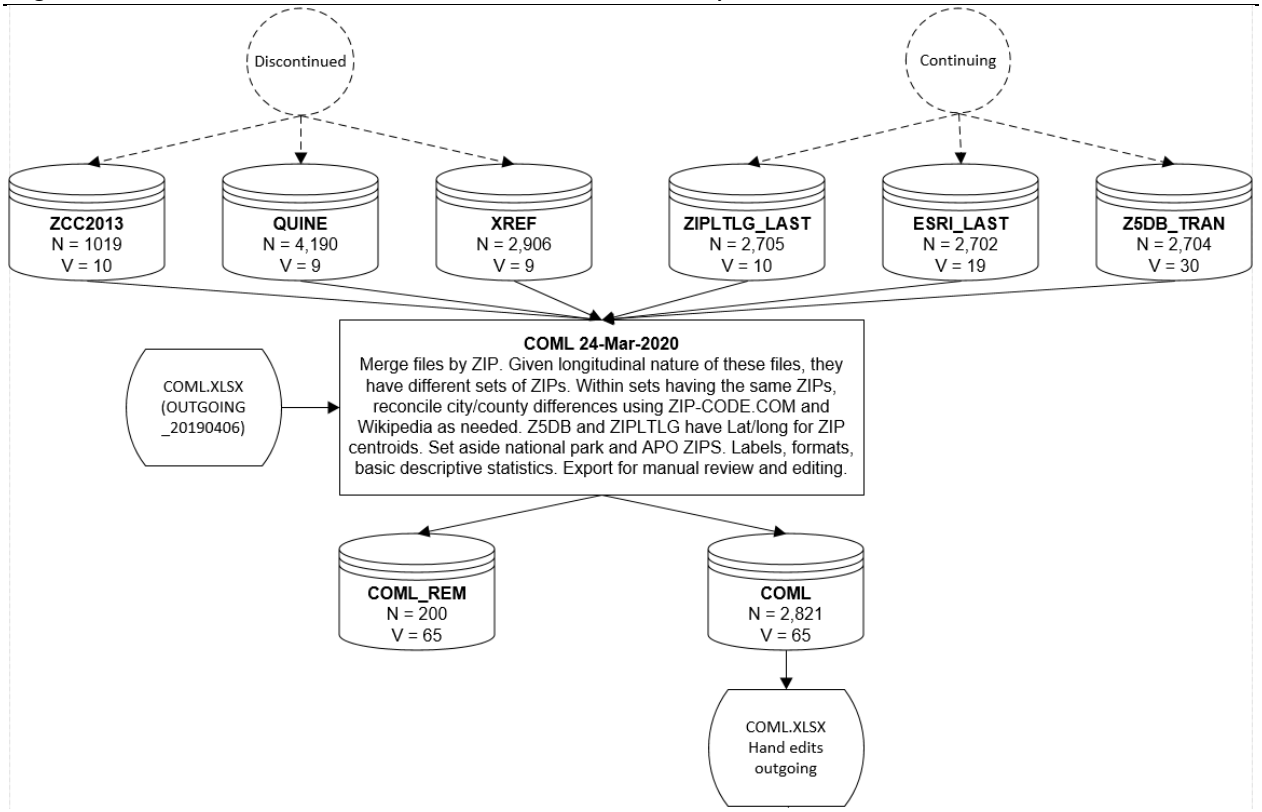
Figure 8 shows that the first step of the program ZIPLTLG pulls California data from the national files. It reconciles city and county differences over time. This includes assigning ZIPs that span multiple counties to one city, and then the county where the city is located. Where city names changed over time, the program assigns the most current name found. The file ZIPLTLG has 1 to N records per ZIP (N = 36,725, V = 7).

The record for the last year the ZIP is found is output to a SAS file (ZIPLTLG _LAST, N = 2,704, V = 10) then to an Excel file for manual review.

## Resolve Differences Among Commercial Sources

Now that the commercial sources were in SAS, with one record per ZIP and with differences reconciled within-source, the next task was to reconcile differences between sources. Each file contains different types of geographic information for different sets of years and ZIPs, produced by different companies using different rules. Figure 9 summarizes the work to reconcile differences.

Figure 9: Reconcile differences between commercial providers




In the program COML, we identified each incoming source. Table 2 summarizes the presence of the 2,821 ZIPs across the sources. Under the Data Source columns, 0 means the ZIP was not in the source and 1 means it was. Under the column set ZIPs, the column labeled N shows the number of common ZIPs and the column labeled Pct is the percent of total ZIPs.

Table 2. Agreement of ZIPs across sources

| Data Source | | | | | | ZIPs | | Agreement | | |
|---|---|---|---|---|---|---|---|---|---|---|
| XREF | Z5DB | ESRI | LTLG | QUINE | WER | N | Pct | No | Yes | Pct |
| 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0.04 | 0 | 1 | 100 |
| 0 | 0 | 0 | 0 | 1 | 0 | 88 | 3.12 | 0 | 88 | 100 |
| 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0.04 | 0 | 1 | 100 |
| 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0.04 | 0 | 1 | 100 |
| 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0.04 | 0 | 1 | 100 |
| 0 | 1 | 0 | 1 | 1 | 0 | 18 | 0.64 | 0 | 18 | 100 |
| 0 | 1 | 1 | 1 | 0 | 0 | 3 | 0.11 | 0 | 3 | 100 |
| 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0.04 | 0 | 1 | 100 |
| 1 | 0 | 0 | 0 | 1 | 0 | 26 | 0.92 | 0 | 26 | 100 |
| 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0.04 | 0 | 1 | 100 |
| 1 | 1 | 0 | 1 | 1 | 0 | 15 | 0.53 | 0 | 15 | 293 |
| 1 | 1 | 1 | 1 | 0 | 0 | 44 | 1.56 | 0 | 44 | 100 |
| 1 | 1 | 1 | 1 | 0 | 1 | 78 | 2.76 | 0 | 78 | 100 |
| 1 | 1 | 1 | 1 | 1 | 0 | 2,407 | 85.32 | 0 | 2,407 | 100 |
| 1 | 1 | 1 | 1 | 1 | 1 | 136 | 4.82 | 0 | 136 | 100 |
| | | | | | | 2,821 | 100 | 0 | 2,821 | 100 |

The last column set shows agreement across sources, where column labeled “No” shows the number of ZIPs that disagreed across sources and “Yes” is the number that agreed. At this time, because of our work in previous years, no combinations were discrepant for city or county across ZIP sources. When there are discrepancies, we search the internet to resolve differences and write SAS code to correct them.

Some ZIPs had Lat/Lon from both sources, some from only one. When a ZIP had Lat/Lon from both sources, we gave preference to the SAS Institute because that is the software that we use the most. Otherwise, we assigned the single Lat/Lon. In our latest run, all ZIPs had Lat/Lon. When we had two Lat/Lon sources, we compared the distance between them. We had 77 ZIPs (2.73%) with more than 10 miles distance between sources.

We output the SAS file to COML.XLSX, on a tab labeled “Incoming”. We copied that tab to one labeled “Outgoing” where we worked manually. For the 77 ZIPs with centroids more than 10 miles apart, we searched the internet to find a third resource and used that. Typically, the replacement was the same as one of those we already had or was intermediary between them. We think that these differences may be due to splitting ZIPs across counties, or when a ZIP changes shape dramatically over time, which would change the centroids from year to year or one data source to another.

When ZIPs lacking Lat/Lon had a parent ZIP, we assigned the parent values. From WER, we identified ZIPs discontinued by USPS and reassigned them to new ZIPS with Lat/Lon. In these cases, we assigned the replacement ZIP as parent to the discontinued ZIP in order to maintain documentation of the change. If the ZIP did not have a parent, we searched our existing geography master (GEOGMAS.XLSX) to find the ZIP with the same city name and the most number of records, and assigned Lat/Lon from that ZIP in COML.XLSX.

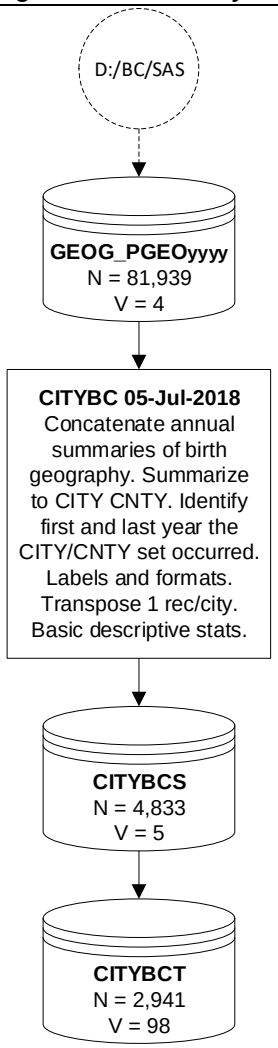
If a ZIP without Lat/Lon or a parent had a unique name, we went to the internet and found its nearest city in the same county, then returned to COML.XLSX and assigned that ZIP’s Lat/Lon to the ZIP in question. At the end of this work, for the first time, all 2,821 unique ZIPs ever found in commercial sources had centroids. In previous iterations, yellow cells on the OUTGOING tab in COML.XLSX indicated data that changed as the result of the manual work. Our hard work in earlier years paid off. Original values are shown on the INCOMING tab.

# CORRECT CITY SPELLING

We earlier described that we summarize geography-related variables in population health files as part of our routine process to standardize these resources for longitudinal research. Some files from CDPH and OSHPD include city names. These variables are notorious for spelling errors, as the law requires the agencies to distribute the data as provided, including any typographical errors. Here we describe the process to develop a format to correct city spelling.

## Summarize City Names by Source

Figure 10: Birth city summary

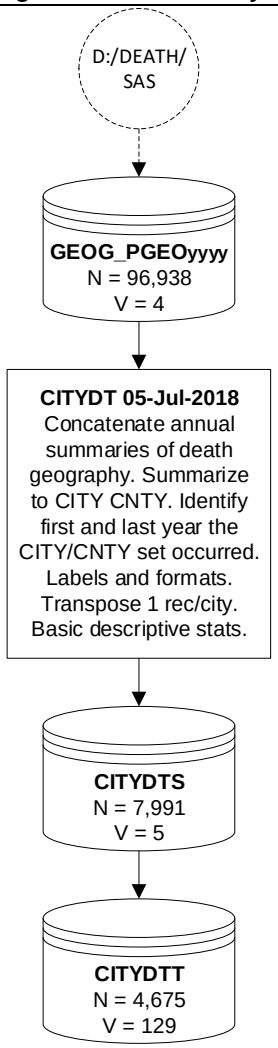



City variables available in some years of the confidential birth certificate files include mother's city of residence and her mailing address city, each with an associated county. ZIP is available from 1989 forward. For birth certificates, the GEOG macro reassigns residence and mail city and county to standard variables (CITY, CNTY), then summarizes by CNTY, ZIP, CITY, and YEAR. After summary, we also have the number of times the combination occurred in that year. The resulting file name indicates the name of the macro program that produced it (GEOG), that the macro was summarizing patient (PGEO) rather than hospital geography (HGEO), and the year (yyyy) of data that was processed.

Figure 10 shows steps to get the list of recorded city names and associated counties. In the program CITYBC, we concatenate the annual summaries of birth geography (N = 81,939). Ignoring ZIP, we re-summarize the CITY-CNTY sets, identifying the first and last year the set occurred and getting a revised count of the number of times the new set occurred (CITYBCS, N = 4,833, V = 5).

The next step transposes the file to 1 record per CITY, with what turns out to be up to 23 counties recorded for the given spelling of a city name. For example, the city San Jose in Santa Clara County has 15 (yes, 15!) associated counties in the transposed file (CITYBCT, N = 2,941, V = 98). As we said, these data are messy.

Figure 11: Death city summary

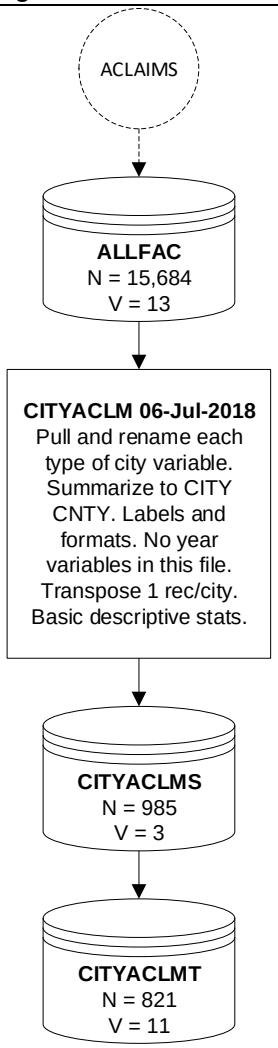



The relevant city variables available in some years of the confidential death certificate files are decedent's city, county, and ZIP of residence. CDPH added injury address in 2014, but we have not changed the macro to include this.

Figure 11 shows steps to get the list of recorded city names and associated counties. In the program CITYDT, we concatenate the annual summaries of death geography (N = 96,938). Ignoring ZIP, we re-summarize the CITY-CNTY sets, identifying the first and last year the set occurred and calculating a revised count of the number of times the new set occurred (CITYDTS, N = 7,991, V = 5).

The next step transposes the file to 1 record per CITY, with what turns out to be up to 31 counties recorded for the given spelling of a city name; nine cities had more than 15 associated counties. For example, the city Sacramento in Sacramento County has 31 associated counties in the transposed file (CITYDTT, N = 4,675, V = 129). Again, we see that these data are messy.

Figure 12: ACLAIMS city summary




California no longer maintains the Automated Certification and Licensing Administrative Information and Management Systems (ACLAIMS) data, which summarized results for quality of care investigations in California hospitals from about 1984 through 2003. Effective March 2004, California began logging complaints and investigations into the federal ASPEN Complaints Tracking System (ACTS) [37]. We have made several unsuccessful attempts to gain access to California's ACTS data.

The ALLFAC file has one record for each facility, including those that submit consolidated Hospital Annual Disclosure Report (HADR) reports to OSHPD. Each record has one address, and there is no year variable. When it still existed, this was the most complete list of licensed facilities and the locations of key units. Figure 12 shows steps to get the list of recorded city names and associated counties.

The program CITYACLM outputs a long-form file of CITY and CNTY, and summarizes the CITY/CNTY sets, (CITYACLMS, N = 985, V = 3).

The last step transposes the file to 1 record per CITY, and up to 4 counties recorded for the given spelling of a city name in the transposed file (CITYACLMT, N = 821, V = 11).

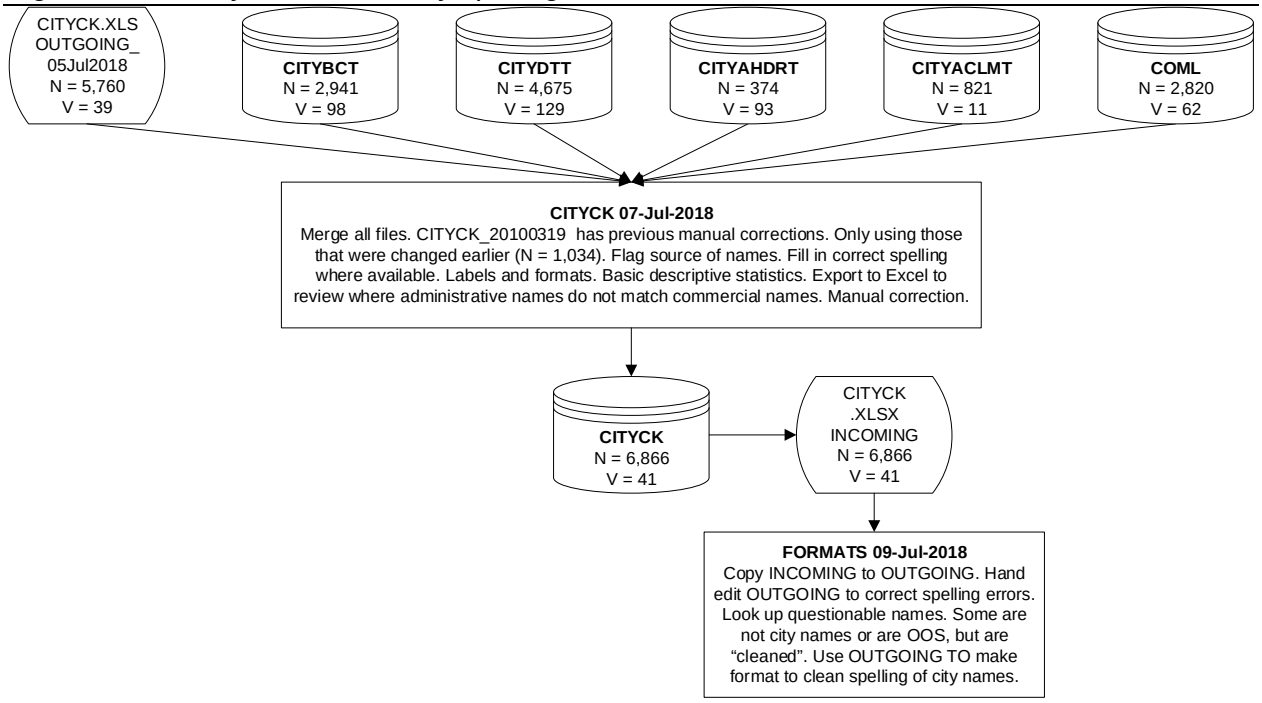
## Identify and Correct City Spelling Errors

Until this point, we have collected and summarized all city names recorded in available public datasets. Now it is time to put these files together with the results of earlier work and the most current version of the commercial sources. Figure 13 summarizes this work.

First, focus on the CITYCK.XLS (OUTGOING tab) in this figure. CITYCK.XLS is created every time city spelling information is updated from the public data sources we just described: birth certificate, death certificate, HADR and ACLAIMS. This work is done routinely in order to know where to assign newly found geographic descriptors. The last time we ran the program CITYCK (city check), we made a new INCOMING worksheet that we copied to an OUTGOING worksheet, which was manually edited; this is the OUTGOING worksheet feeding into the CITYCK program. The OUTGOING worksheet reflects the cumulative knowledge gained about city spelling errors over the years. To minimize work with each update, we bring in the old

OUTGOING sheet, which has cumulative corrections through the last update, so that we can focus on reviewing and manually editing the newest entries. The date extension on OUTGOING is the date we started work to update CITYCK.

Figure 13: Identify and correct city spelling errors




From the old CITYCK file, we keep city, corrected city (CITYC), first and last year previously found, and any notes from previous searches. From the CITYxxT files, we keep city as spelled, first and last year found, and number of times found. From COML, we keep CITY (which is the current preferred city), county, and number of times each source found the name. After merging by CITY, we make a variable summing all the times the spelling was found over all sources. We also recalculate the first and last year the spelling appeared. If CITYC is blank because COML brings a new city name, we impute CITYC with CITY, because the COML city spelling is correct. We flag records that still lacking a cleaned city name. We export the SAS file to the tab INCOMING in the file CITYCK.XLSX (note that incoming was CITYCK.**XLS,** with xlsx signifying our move to Windows 10).

Table 3. Source of city errors (N = 6,866)

| | Sources | | | | | | Total | |
|---|---|---|---|---|---|---|---|---|
| CITYCK | PREV | COML | BC | DT | AHDR | ACLM | N | Pct |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1,194 | 17.4 |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 27 | 0.4 |
| 0 | 1 | 0 | 0 | 0 | 1 | 0 | 17 | 0.2 |
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1,540 | 22.4 |
| 0 | 1 | 0 | 0 | 1 | 0 | 1 | 20 | 0.3 |
| 0 | 1 | 0 | 1 | 0 | 0 | 0 | 840 | 12.2 |
| 0 | 1 | 0 | 1 | 0 | 1 | 0 | 2 | 0.0 |
| 0 | 1 | 0 | 1 | 1 | 0 | 0 | 752 | 11.0 |
| 0 | 1 | 0 | 1 | 1 | 0 | 1 | 66 | 1.0 |
| 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0.0 |
| 0 | 1 | 0 | 1 | 1 | 1 | 1 | 8 | 0.1 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0.0 |
| 0 | 1 | 1 | 0 | 1 | 0 | 0 | 4 | 0.1 |
| 0 | 1 | 1 | 1 | 1 | 0 | 0 | 557 | 8.1 |
| 0 | 1 | 1 | 1 | 1 | 0 | 1 | 369 | 5.4 |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 2 | 0.0 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 360 | 5.2 |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0.0 |
| 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1,054 | 15.4 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 49 | 0.7 |
| 1 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 0.0 |

Table 3 summarizes source of city name errors over the years. The grayed portion at the bottom of the table identifies that 1,105 city spellings needed correction in this iteration (CITYCK = 1). As before, most errors came from the death files.

At this point, we move to Excel, copy the INCOMING sheet to OUTGOING, and make corrections. Note that we mainly are correcting spelling. Cities such as Phoenix, Milwaukee, Guadalajara, or other out-of-state cities are set aside as UNUSABLE because they are not in California. “Cities” with corporate names are also UNUSABLE. When all spellings are fixed, we use the Excel file as input to update format $CITYC, which reassigns misspelled city names into a variable we typically name CITYC(orrected).

The next time we update the program, we will rename the OUTGOING tab to OUTGOING_ddmmyyyy and repeat the process.

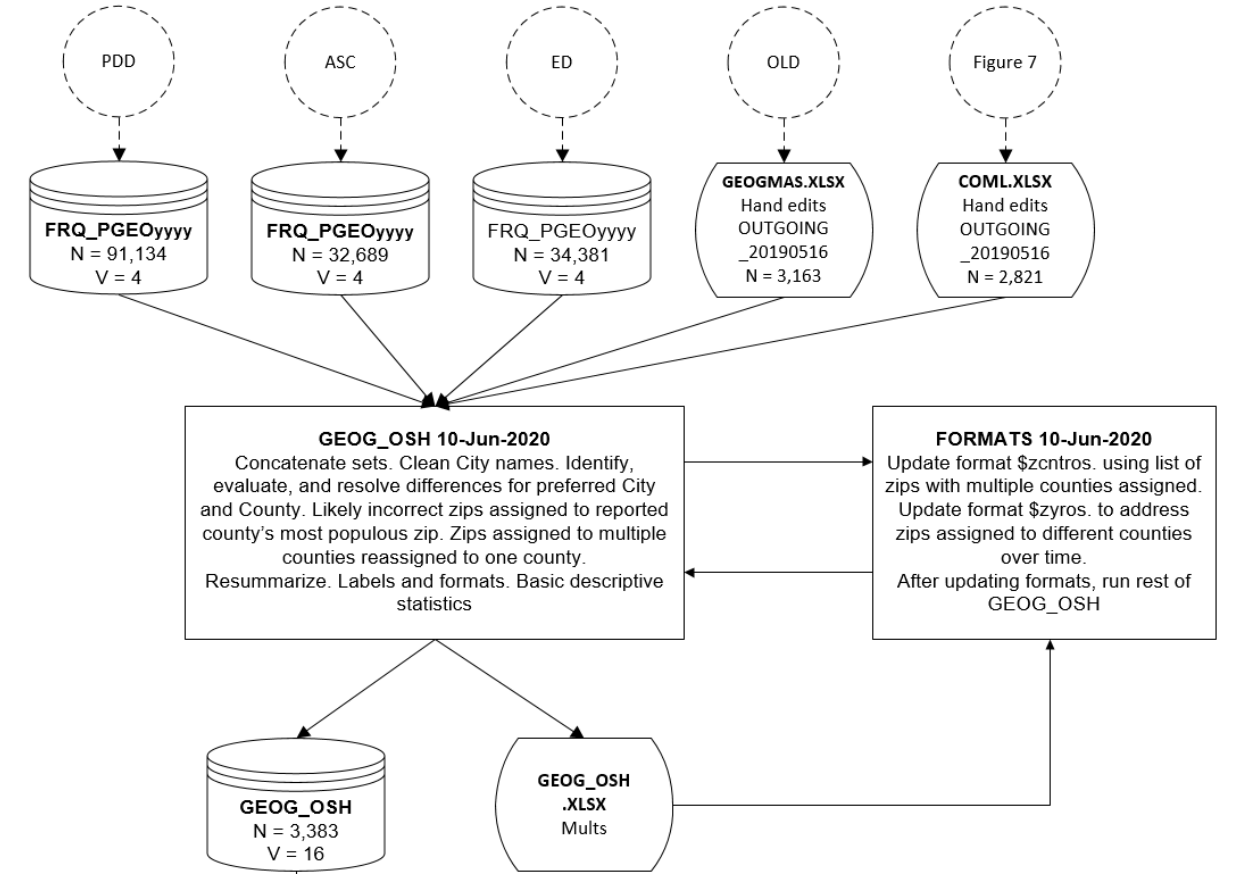
## STANDARDIZE POPULATION HEALTH GEOGRAPHY

Here we detail the steps we take to standardize population health geographic data from public administrative datasets. As described earlier in this monograph, this data includes emergency department, ambulatory surgery center and hospital discharge data from OSHPD; birth, death and fetal death files from CDPH Vital Statistics, and health facility data from HADR and ACLAIMS. The goal of this summary step is to produce, for each data source, a cleaned dataset containing the most recent city, county, health service area (HSA), service planning area (SPA), jurisdiction and Lat/Lon assignment for each ZIP in the dataset. We also resolve any occurrence where one ZIP is assigned to multiple cities and/or counties in a data source so that a ZIP occurs only once within the resultant dataset.

## Summarize OSHPD Patient Data by ZIP, City and County

We receive new data from OSHPD annually and update our Geography Master accordingly. Figure 14 illustrates steps taken to create the GEOG_OSH dataset.

Figure 14: Summarizing OSHPD patient data by ZIP, city and county




The inputs to this program are: annual patient discharge (PDD) dataset, annual ambulatory surgery (ASC) dataset, annual emergency department (ED) dataset, the most recent version of the Geography Master (GEOGMAS.XLSX), and the final summary of ZIP data from commercial sources (COML). FRQ_* files are made during the process of importing population health files into our system [38].

First, we concatenate the three OSHPD datasets and look for instances where one ZIP has been assigned to multiple counties over time. In our latest run, there were 42 ZIPs with one or more county assignments, affecting some 2 million records. Possible reasons for this are: 1) errors in the ZIP and/or county at the point of data entry; and 2) some ZIPs span counties and OSHPD made different decisions over time about which county to assign to the ZIPs.

While our goal is to produce a summary with the most recent county assigned to each ZIP, we do not want to “lose” records that were not assigned to the currently correct county. Here, we make two assumptions: 1) the ZIP-county pair with the largest record count is the probable correct county assignment; and 2) we trust that county assignment in our source data is more likely to be correct than the ZIP, which is more prone to mistyping or transposition errors. We account for these records by applying SAS formats ($ZCNTROS and $ZYROS) to the ZIP-county pairs with lower record counts, reassigning these to the ZIP in the reported county with the highest record count. This way, the records are still accounted for when summarizing data by county, but the effect of adding these records to a ZIP with high record count will be minimal.

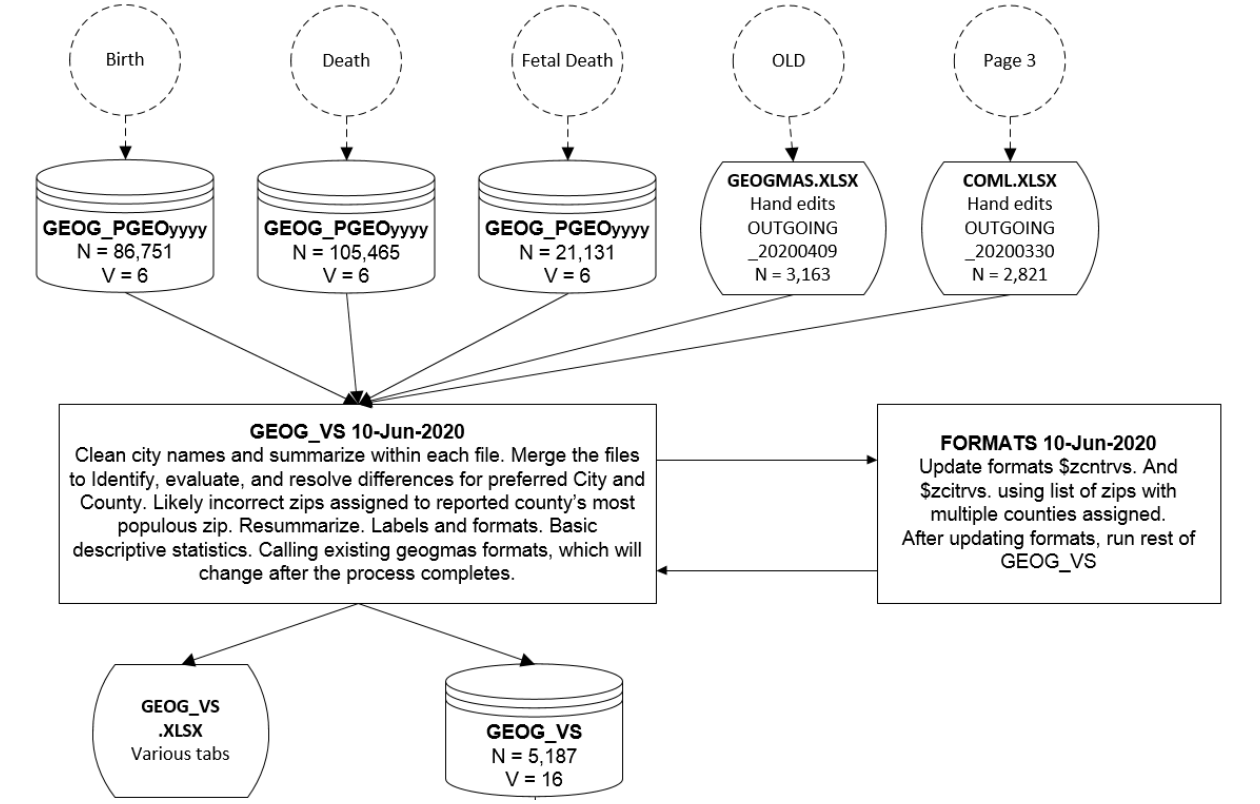
For example, within our OSHPD data sources, ZIP Code 94582 in the city of San Ramon has 84,574 records with Contra Costa as the reported county and 90 records with neighboring Alameda as the reported county. When the formats are applied, the 90 records with Alameda as the reported county are reassigned to ZIP Code 94544 in Alameda County. ZIP Code 94544 is in Hayward, Alameda County, and is the ZIP Code with the most records in Alameda County.

After reassigning ZIPs, we merge the OSHPD data with the latest Geography Master and commercial datasets (COML) to impute values for any records missing city or county. The final GEOG_OSH dataset contains 3,383 unique ZIPs, each assigned to one county, SPA, HSA, jurisdiction, and Lat/Lon coordinates.

## Summarize Vital Statistics Data by ZIP, City and County

We receive new data from CDPH Vital Statistics annually and update our Geography Master accordingly. Figure 15 illustrates the steps taken to create the GEOG_VS dataset, which are similar to steps that create GEOG_OSH.

Figure 15: Summarizing Vital Statistics data by ZIP, city and county




The inputs to create GEOG_VS are annual birth, death, and fetal death certificate data. We concatenate these files and delete any out-of-range ZIPs. Similar to the GEOG_OSH program, we look for instances where one ZIP has been assigned to multiple counties. Again, these Vital Statistics data are messy. In this last run, we identified 3,445 ZIPs with more than one county assigned (affecting nearly 27 million records), with 23 ZIPs having 10 or more counties assigned over time. We expect this because Vital Statistics is required to report data as it was originally recorded at the point of data entry, including any typographical errors.

We use the same methods to address multiple ZIP-county pairs as described in the previous OSHPD section. We assume the reported county is more accurate than reported ZIP, and reassign records with ZIP-county mismatches to the ZIP with the highest record count in the reported county. We do this with the formats $ZCNTRVS and $ZCITRVS. $ZCNTRVS specifies

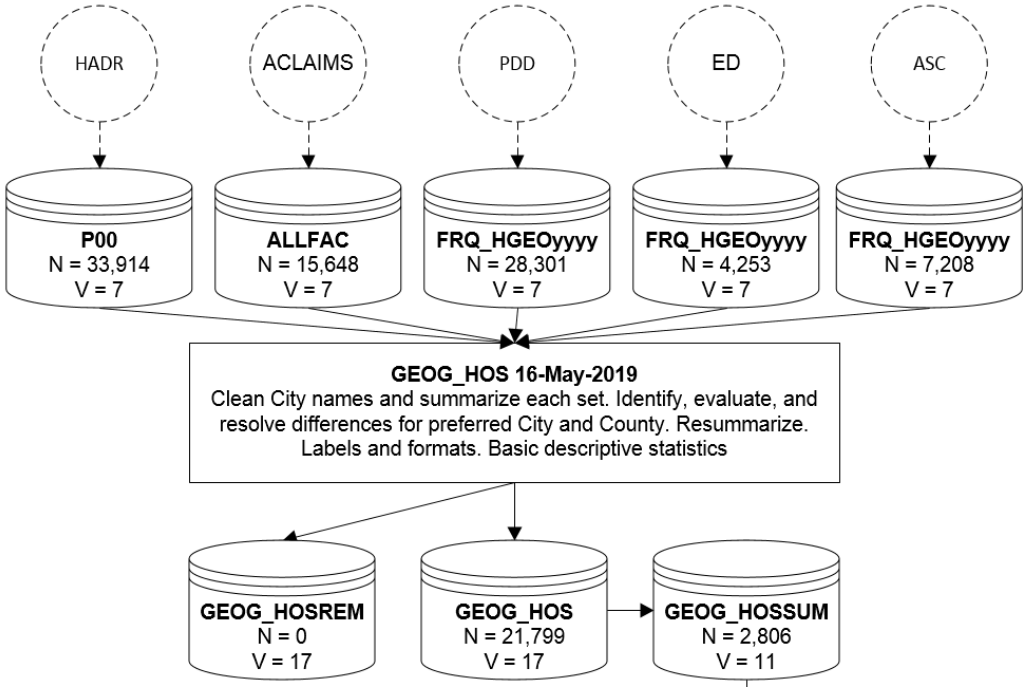
the new ZIP that the record should be reassigned to, and $ZCITRVS specifies the corresponding city name.

After reassigning ZIPs, we merged VS data with the latest commercial data and Geography Master to impute values for records missing city or county. After imputation, all ZIPs had an assigned county, but 1,201 ZIPs still did not have city names. City names for these ZIPs will be assigned in the last step that updates the Geography Master. The final GEOG_VS dataset contains 5,187 unique ZIPs, each assigned to one county, SPA, HSA, jurisdiction, and Lat/Lon.

## Summarize Health Facilities Data by ZIP, City and County

Lastly, we summarize ZIP, city, and county information for health facilities from our OSHPD (ED, ASC and PDD) and facility files (HADR and ACLAIMS). Figure 16 illustrates how the GEOG_HOS dataset is created.

Figure 16: Summarizing health facilities data by ZIP, city and county

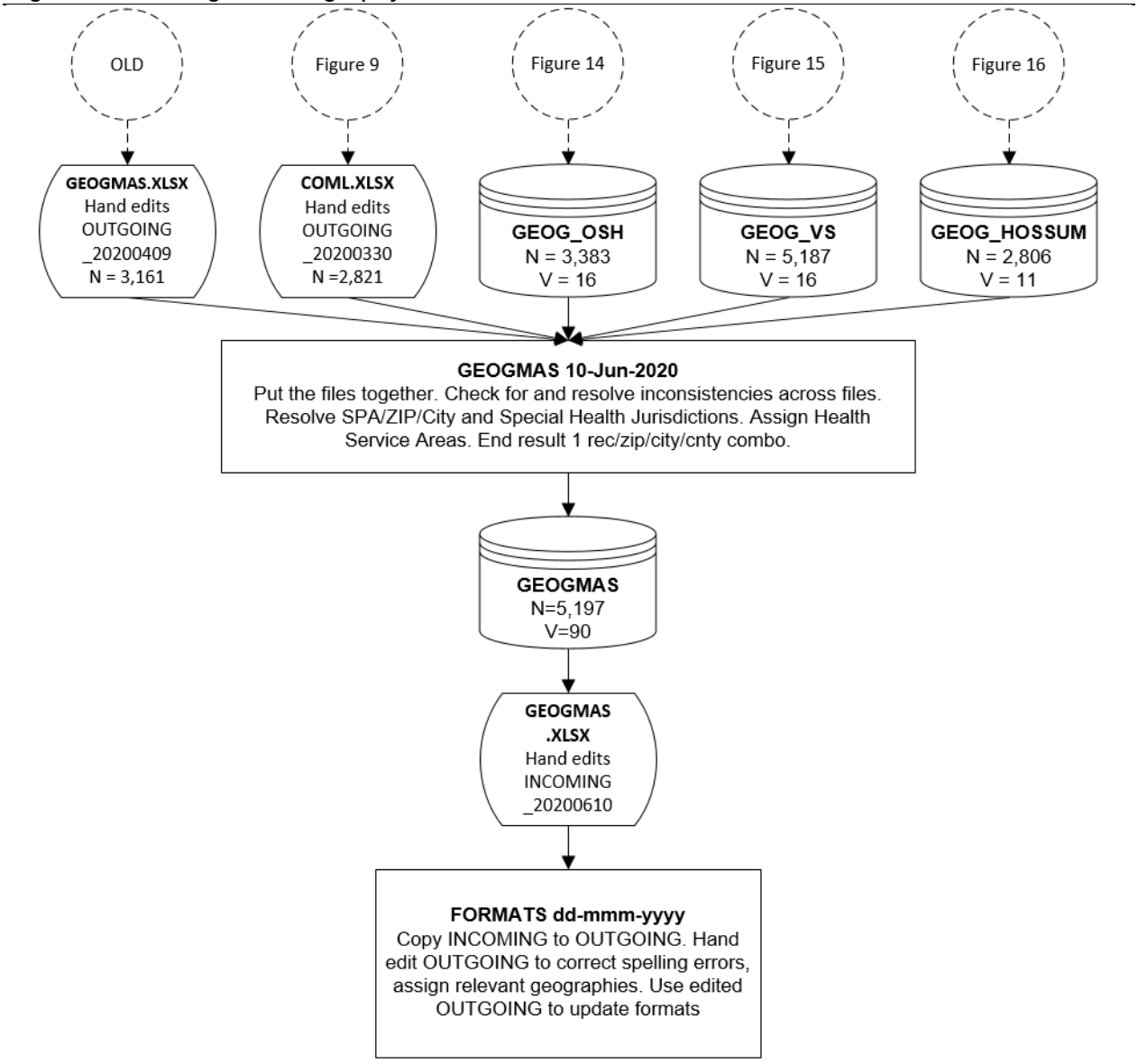



We use similar methods to extract ZIP, city and county from health facilities information. The five input datasets are merged and all out-of-range ZIPs are removed. When information from all input datasets were used, all ZIPs had an assigned county and only one did not have a city. The final GEOG_HOSSUM dataset has 2,806 unique ZIPs, each with one county assigned.

# MAKING THE GEOGRAPHY MASTER

The final step to making our updated Geography Master, shown in Figure 17, is to merge the following datasets together: GEOG_OSH, GEOG_VS, GEOG_HOSSUM, latest commercial data (COML), and the most recent Geography Master (GEOGMAS). The goal is to produce an updated GEOGMAS with one ZIP per record, and the latest available information for each ZIP such as county, city name (including preferred, alternate and not allowable names, if available), SPA, HSA, jurisdiction, and Lat/Lon values.

Figure 17: Making the Geography Master




First, we merged the public administrative datasets GEOG_OSH, GEOG_VS and GEOG_HOSSUM by ZIP. In this run, 959 ZIPs were missing city names and 116 were missing

county numbers. For these ZIPs, we imputed city and county from these sources, in this order of priority based on data quality: OSPHD, VS, and health facilities data. After this imputation, all ZIPs had counties, but 739 still did not have city names.

We then merged this dataset with the most recent commercial and Geography Master datasets. We used this hierarchy to assign final city name, county, Lat/Lon, SPA, HSA, and jurisdiction to each ZIP: commercial data, existing Geography Master, and merged administrative dataset. For the final 739 ZIPs with no city names, we assigned the ZIP with the most records in the reported county as the parent ZIP, and imputed city names and other geography characteristics from the newly assigned parent ZIP.

The final GEOGMAS dataset contained 5,197 unique ZIPs. We exported GEOGMAS to an Excel file for manual review, naming the edited version OUTGOING for use in analyses and future geography updates. After making hand edits to correct any ZIPs with missing key fields, we use the new GEOGMAS to update our geography formats to ensure they contain the most recent known values for each ZIP.

## RESOURCES

We have described the methods FHOP uses to standardize geography data and format it for use in longitudinal research. All SAS programs described here are in the public domain and are available upon request. Because staff time is limited, researchers will have to contract for more than one hour of support.

# ENDNOTES


1 Geographic Areas Reference Manual. U.S. Census Bureau. Nov 1994. Last accessed 28-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/.

2 Figure 2–1. Standard Hierarchy of Census Geographic Entities. Geographic Areas Reference Manual. Chapter 2: Geographic Overview. U.S. Census Bureau. Nov 1994. Last accessed 28-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/Ch2GARM.pdf.

3 FIPS General Information. Last accessed 28-Apr-2020 at: https://www.nist.gov/information-technology-laboratory/fips-general-information

4 Understanding Geographic Identifiers (GEOIDs). Last accessed 30-Apr-2020 at: https://www.census.gov/programs-surveys/geography/guidance/geo-identifiers.html

5 Geographic Areas Reference Manual. Chapter 6: Statistical Groupings of States and Counties. U.S. Census Bureau. Nov 1994. Last accessed 28-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/Ch6GARM.pdf.

6 Figure 6–1. Census Regions and Divisions of the United States. Geographic Areas Reference Manual. Chapter 6: Statistical Groupings of States and Counties. U.S. Census Bureau. Nov 1994. Last accessed 28-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/Ch6GARM.pdf.

7 Geographic Areas Reference Manual. Chapter 4: States, Counties, and Statistically Equivalent Entities. U.S. Census Bureau. Nov 1994. Last accessed 28-Apr-2020 at: htmlhttps://www2.census.gov/geo/pdfs/reference/GARM/Ch4GARM.pdf.

8 United States Postal Service (May 2007) The history of the United States Postal Service: An American History 1775-2006. Publication 100,2020. PSN 7610-03-000-9247. Last accessed 29-Apr-2020 at: https://about.usps.com/publications/pub100.pdf.

9 United States Zip Codes.org, ZIP Code 94920.. Last accessed 27-Jun-2020 at: https://www.unitedstateszipcodes.org/94920/

10 Zenou Y, Boccard N. (2000) Racial Discrimination and Redlining in Cities. Journal of Urban Economics. **48**, 260-285.

11 Remy L, Clay T, Oliva G. (2000) California Child and Youth Injury Hot Spots Project 1995-1997, Volume Three: Technical Guide, Sacramento, CA: California Department of Health Services, Maternal and Child Health Branch, August 2000 Available at: http://fhop.ucsf.edu/fhop-publications-injury-surveillance.

12 ZIP Code™ Tabulation Areas (ZCTAs). Last accessed 29-Apr-2020 at: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/zctas.html

13 Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R. (2002) ZIP code caveat: Bias due to spatiotemporal mismatches between ZIP codes and US Census-defined geographic areas - The Public Health Disparities Geocoding Project. AJPH, 92:7, 1100-1102.

14 Geographic Areas Reference Manual. Chapter 10: Census Tracts and Block Numbering Areas. U.S. Census Bureau. Nov 1994. Last accessed 29-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/Ch10GARM.pdf.

15 Geographic Areas Reference Manual. Chapter 11: Census Blocks and Block Groups. U.S. Census Bureau. Nov 1994. Last accessed 29-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf.

16 Geographic Areas Reference Manual. Chapter 9. Places. US Census Bureau. Nov 1994. Last accessed 29-Apr-2020 at: https://www2.census.gov/geo/pdfs/reference/GARM/Ch9GARM.pdf.

17 Luft HS, Frisvold GA. Decisionmaking in regional health planning agencies. J Health Polit Policy Law. 1979 Summer;4(2):250-72.

18 Transaction Systems Inc. Evaluation of alternative health area definition methods. DHEW Contract Number HRA 230-75-0080. 1976.

19 Makuc DM, Haglund B, Ingram DD, Kleinman JC, Feldman JJ. Health service areas for the United States. Vital Health Stat 2. 1991 Nov;(112):1-102.

20 Pickle LW, Mungiole M, Jones GK, White AA. Atlas of United States Mortality. Hyattsville, Maryland: National Center for Health Statistics. 1996. Last accessed 30-Apr-2020 at: https://www.cdc.gov/nchs/products/other/atlas/atlas.htm.

21 Health Service Areas (HSA). Last accessed 30-Apr-2020 at: https://seer.cancer.gov/seerstat/variables/countyattribs/hsa.html

22 Remy L, Oliva G, Clay T. (2007) Hospital Capacity to Treat Mental Illness 1991-2005. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: https://fhop.ucsf.edu/fhop-publications-hospitalizations-trends-and-outcomes.

23 Remy LL, Clay T, Byers V, Rosenfeld P (2019) Hospital, health, and community burden after oil refinery fires, Richmond, California 2007 and 2012. Environmental Health 18:48. https://doi.org/10.1186/s12940-019-0484-4

24 California MCAH Regions. Last accessed 30-Apr-2020 at: https://fhop.ucsf.edu/california-mcah-regions.

25 Service Planning Areas. Last accessed 30-Apr-2020 at: http://publichealth.lacounty.gov/chs/SPAMain/ServicePlanningAreas.htm

26 San Francisco Community Health Needs Assessment 2019. Page 14. San Francisco Health Improvement Parternship. 2019. Last accessed 29-Jun-2020 at: https://www.sfdph.org/dph/hc/HCAgen/2019/May%207/CHNA_2019_Report_041819_Stage%204.pdf

27 Community Indicators Report 2017. Page 4. San Bernardino County. 2017. Last accessed 29-Jun-2020 at: http://cms.sbcounty.gov/Portals/21/Resources%20Documents/CIR_2017_report.pdf?ver=2018-03-23-132312-883

28 Zaretsky HW, Foley JC. (1980) New Horizons for Better Health: 1980 California State Health Plan. Office of Statewide Health Planning and Development. Not available on the web. PDF copy available upon request to L Remy.

29 Remy L, Clay T. (2016) Managing Longitudinal Research Studies: Crosswalking Hospital Identifiers. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at http://fhop.ucsf.edu/data-management-methods.

30 Remy L, Clay T. (2015) Managing Longitudinal Research Studies: Annual Hospital Disclosure Report. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at http://fhop.ucsf.edu/data-management-methods

31 Annual Financial Disclosure Reports. State of California Office of Statewide Health Planning and Development. Last accessed 30-Apr-2020 at: https://oshpd.ca.gov/data-and-reports/cost-transparency/hospital-financials/

32 Hospital Annual Disclosure Report. Complete Hospital Annual Financial Database For Excel, 41st Year Through Subsequent Disclosure Cycles. Office of Statewide Health Planning and Development. Oct 2017. Last accessed 30-Apr-2020 at: https://data.chhs.ca.gov/dataset/9c594ceb-fe46-4e6d-9553-c138a868bdc5/resource/5e498724-7c68-4b7f-898a-56a36c4dd407/download/hadrfull-db-documentation-rpe2015-xx.pdf

33 See Douglas Boynton Quine homepage. Last access 01-May-2020 at: http://www.quine.org/

34 ZIPList5 Geocode. Last accessed 01-May-2020 at: http://www.zip-info.com/products/z5ll/z5ll.htm.

35 See https://www.esri.com/en-us/home. Last accessed 01-May-2020.

36 Maps online. SAS Institute. Last accessed 13-May-2020 at: https://support.sas.com/rnd/datavisualization/mapsonline/html/misc.html.

37 Remy, LL (2005) Complaints and Investigations: Department of Health Services Investigations of General Acute Care Hospitals. Invited testimony before California Senate Committee in support of SB 1005 (Dunn).

38 Remy L, Clay T. (2017) Managing Longitudinal Research Studies: Preparing Master Files. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods