MANAGING LONGITUDINAL RESEARCH STUDIES:

POPULATION MASTER FILES

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: fhop.ucsf.edu

July 2020

TABLE OF CONTENTS

Suggested Citation

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

Table of Figures

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
PROGRAM NAME	SAS program with brief description of steps

ACRONYMS

ACS	American Community Survey
AIAN	American Indian/Alaska Native
API	Asian/Pacific Islander
CDC	Centers for Disease Control
DOF	California Department of Finance
FHOP	Family Health Outcomes Project
HSA	Health Service Areas
NCHS	National Center for Health Statistics
OMB	Office of Management and Budget
OSHPD	Office of Statewide Health Planning and Development
SF1	Summary File 1 from the U.S. Census Bureau
SF3	Summary File 3 from the U.S. Census Bureau
SPA	Service Planning Area
UCSF	University of California, San Francisco
USPS	United States Postal Service
WHO	World Health Organization
ZCTA	Zone Improvement Plan (ZIP)-Code Tabulation Area
ZIP	Zone Improvement Plan

POPULATION MASTER FILES

This document describes methods the UCSF Family Health Outcomes Project (FHOP) uses to prepare longitudinal population master files to use as denominators. We assume that the user of this document has read other documents describing the foundation of our methodology [1-3].

OVERVIEW

FHOP uses four data sources to construct our population master files:

1) County-level population estimates and projections distributed by the Demographic Research Unit of the California Department of Finance (DOF)
2) Decennial multi-level population estimates distributed by the U.S. Census Bureau
3) Intercensal population projections from commercial sources such as Geolytics, Esri and others
4) Longitudinal bridged-race county-level population files distributed by the National Center for Health Statistics (NCHS), in cooperation with the U.S. Census Bureau

To support its longitudinal research, FHOP collates and maintains a series of non-confidential population files from these four sources. At the writing of this document, we have: county-level DOF files from 1970 through 2060 projections; multi-level 1970-2010 Census files; 1970-2010 and 2016 multi-level, longitudinally normalized census files from GeoLytics ; various intercensal Zone Improvement Plan (ZIP)-level population estimates from commercial vendors (e.g. small area estimates distributed by Esri as part of ArcMap); and NCHS 1990-2015 bridged race, county-level population files.

Processes and resources described here will help users of other population files. We are making this basic methodology and its associated software public to help researchers understand the nature of data management for complex longitudinal research. This also will provide a background to users of our longitudinal DataBook products and readers of FHOP studies using external population denominators.

Contracts with various departments in the State of California require us to provide funding agencies with an annual backup of all programs, logs, listings, and output files. This creates an audit trail of our work. Since we do not know where the programs and/or resulting files will be used, we try to write code that will run in any environment and provide as much documentation as possible. All work is in SAS, assisted by Microsoft Excel and Visio. Because public funding supports development of these programs, they are in the public domain.

DEPARTMENT OF FINANCE COUNTY-LEVEL POPULATION

Background

The California State Legislature designated the DOF Demographic Research Unit as the single official source of demographic data for state planning and budgeting. When FHOP contracts with state agencies to produce data products, we are required to use DOF population estimates, rather than other population data sources. These data are not confidential, and using them does not require an approved research protocol.

DOF produces total population and housing estimates as of January each year at the state, county, and city level. July population estimates are county-level with annual estimates by age, sex, and race/ethnicity. Analyses needing more granular population estimates must use other data sources. Race/ethnicity groupings have changed over time, so preparing these data for longitudinal analyses must address these changes.

Obtaining the Data

DOF distributes the 1970-1989 data as annual text files [4]. It distributes the 1990-1999 data as annual Excel files by county [5]. The 2000-2009 data was originally distributed as Excel files by year, but now it is distributed as a text file that is too large to import into Excel [6]. Population projections are currently available for 2010 through 2060 [7]. Again, the projection file is too large to import into Excel.

Be aware that DOF population estimates have sometimes been inaccurate. On occasion, we have identified problems with their estimates and they have corrected them. Sometimes DOF has updated a given sequence of years without changing the publication date on the web. Be sure to check the file date internal to the ZIP file. DOF issued three population projection updates in 2017, with the same external ZIP file name but different creation dates in the internal files. For these and other reasons, we visit the DOF website every time we need to use their population estimates, to confirm that we have the most current versions.

Preparing the Population Files

We have had to update source files numerous times since starting to work with DOF data in 1999. For example, in 2010, we updated the period 1970-2010. In 2013, DOF updated the 2000-2010 intercensal estimates and released its first post-censal projections based on the 2010 Census. Later that year, DOF removed 2010 from the intercensal estimates and moved 2010 into the projections. We now feel it is likely that the 1970-2009 population estimates will remain stable. DOF has regularly updated projections since 2014, all for the year range 2010 to 2060.

Figure 1. Preparing 1970-2007 DOF Data

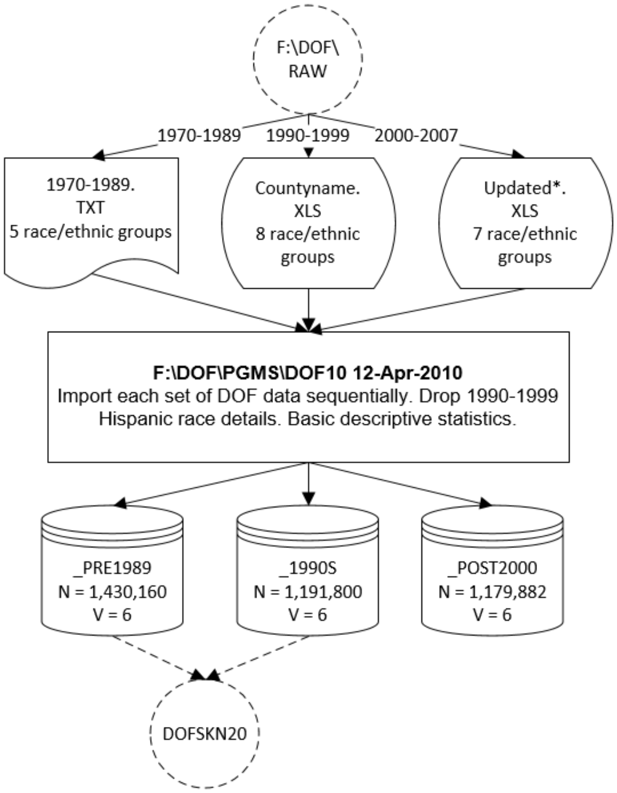
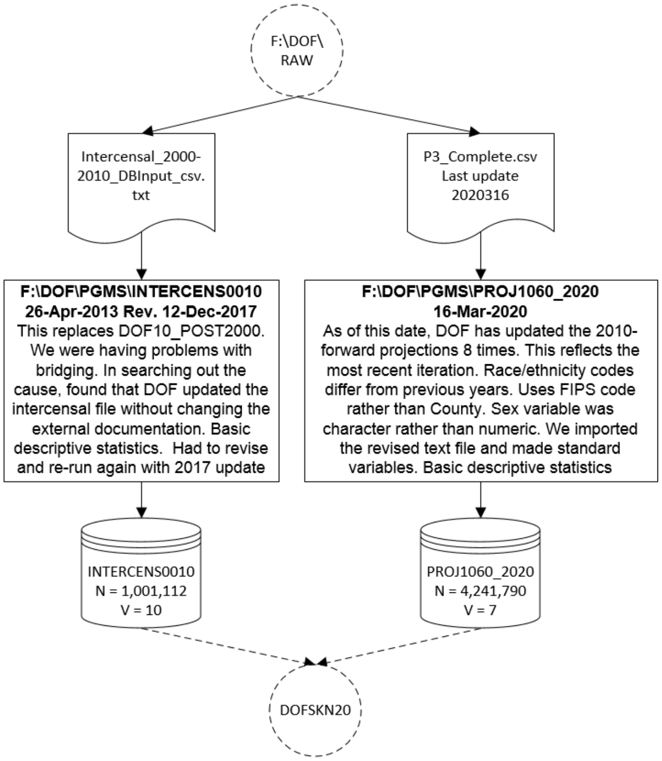


Figure 1 summarizes the program to import the 1970-2007 DOF population data. The program DOF10 (we did this in 2010) sequentially reads each set of files, creates and labels standard variables. At the end of importing each set of files, basic descriptive statistics are reported.

Note that the three sets of files have different numbers of race/ethnic categories. The program standardizes the race/ethnicity variables, defining them consistently over time.

We address problems and solutions associated with longitudinal changes in race/ethnic categories elsewhere [8].

Figure 2. Revising DOF Data for years 2010 - 2060


F:\DOF\ RAW
Intercensal_2000-2010_DBInput_csv.txt
P3_Complete.csv Last update 2020316
INTERCENS0010 N = 1,001,112 V = 10
PROJ1060_2020 N = 4,241,790 V = 7
DOFSKN20

F:\DOF\PGMS\INTERCENS0010
26-Apr-2013 Rev. 12-Dec-2017
This replaces DOF10_POST2000. We were having problems with bridging. In searching out the cause, found that DOF updated the intercensal file without changing the external documentation. Basic descriptive statistics. Had to revise and re-run again with 2017 update

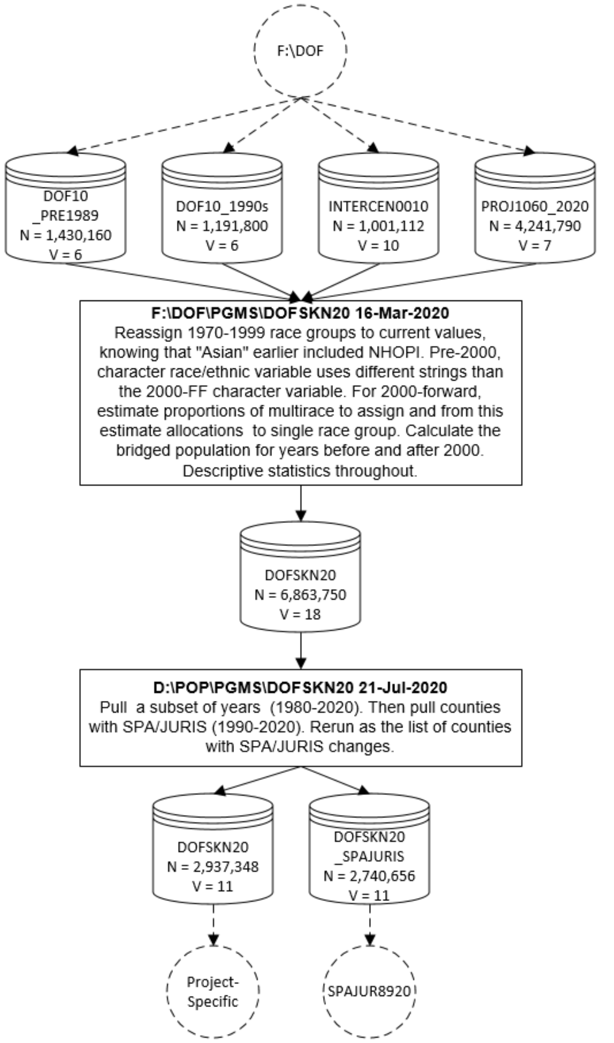
F:\DOF\PGMS\PROJ1060_2020
16-Mar-2020
As of this date, DOF has updated the 2010-forward projections 8 times. This reflects the most recent iteration. Race/ethnicity codes differ from previous years. Uses FIPS code rather than County. Sex variable was character rather than numeric. We imported the revised text file and made standard variables. Basic descriptive statistics

In December 2017, we encountered some problems while attempting to bridge race categories between DOF files. While troubleshooting, we discovered that between our last run in April 2013 and the 2017 run, DOF had updated the intercensal projections file but did not change their website's documentation. The ZIP file containing the data had same name, but the text file within the ZIP file had a different creation date and different structure for the year variable. These changes required yet another program re-write, summarized in Figure 2.

We wrote the SAS program INTERCENS0010 to re-import the DOF intercensal population data from the year 2000 onward. It reads the replaced files into SAS, creates and labels standard variables, and reports basic descriptive statistics.

Next, we use the program PROJ1060_YYYY to update the 2010 to 2016 population projection file. YYYY is the year we make the update. Since we started processing the file, we know of 8 updates, with 3 updates in 2017 alone. We recommend that analysts check the DOF website for updates regularly. This program creates and labels standard variables, and reports basic descriptive statistics.

Figure 3. Making longitudinal DOF population file


F:\DOF
DOF10 _PRE1989 N = 1,430,160 V = 6
DOF10_1990s N = 1,191,800 V = 6
INTERCEN0010 N = 1,001,112 V = 10
PROJ1060_2020 N = 4,241,790 V = 7
F:\DOF\PGMS\DOFSKN20 16-Mar-2020
DOFSKN20 N = 6,863,750 V = 18
D:\POP\PGMS\DOFSKN20 21-Jul-2020
DOFSKN20 N = 2,937,348 V = 11
DOFSKN20 _SPAJURIS N = 2,740,656 V = 11
Project-Specific
SPAJUR8920

Reassign 1970-1999 race groups to current values, knowing that "Asian" earlier included NHOPI. Pre-2000, character race/ethnic variable uses different strings than the 2000-FF character variable. For 2000-forward, estimate proportions of multirace to assign and from this estimate allocations to single race group. Calculate the bridged population for years before and after 2000. Descriptive statistics throughout.

Pull a subset of years (1980-2020). Then pull counties with SPA/JURIS (1990-2020). Rerun as the list of counties with SPA/JURIS changes.

Figure 3 summarizes the last steps to make a county-level population file with longitudinally consistent groups. Within each file, we convert incoming character variables for sex and race/ethnicity into numeric variables, cognizant of when definitions change.

From 2000 onward, we make a single race/ethnicity variable consistent with the pre-2000 race/ethnicity variable. Drawing on methods described elsewhere [9], we proportionally allocate multi-race to five defined groups (White, Black, Hispanic, Asian/Pacific Islander (API), American Indian/Alaska Native (AIAN)) that existed before 2000.

Next, we move from the drive/directory where we make the DOF population files to where we make population files that we use for our work. Here, we pull records from 1980 through 2060. 1980 is the first year we have longitudinal population health data.

We produce the dataset DOFSKNyy (yy being the year of last update) for use in specific analyses. The structure of this dataset is one row for every year, 1-year age group, sex and geographic level, with the count and distribution of each race/ethnicity group in each row.

We also produce dataset DOFSKNyy_SPAJURIS to use in sub-county level analyses. Please see the section on intercensal small-area populations later in this document for a discussion on sub-county geographies.

NATIONAL POPULATION ESTIMATES AND PROJECTIONS

Background

When we are not required to use DOF population estimates and projections for our analyses, FHOP uses several sources for national population estimates.

The first source we use is decennial census files from the U.S. Census Bureau. FHOP has decennial census files from 1970 to 2010 for California, and in some years for additional states. For this monograph, we focus on how we process California census data from 1990 to 2010. We use the census Summary File 1 (SF1) from all four decennial censuses, which contains 100% coverage of questions administered to all households. Information include: sex, age, race/ethnicity, household composition, and various housing variables [10]. We summarize population counts for at the state, county, place, census tract levels for all three SF1 files, with some years having additional U.S. Postal Service (USPS) ZIP-related geographies. We also use the Summary File 3 (SF3), which contain data from a sample of households asked to complete the 1990 and 2000 decennial census long form [11]. For the 2010 Census, the long form and SF3 were replaced by the American Community Survey, which began in 2005 and samples a proportion of the U.S. population monthly, collecting information on education, employment, internet access and transportation [12].

The second data source we employ is from GeoLytics, a company that creates data products using census data. From them, we purchased 1970 to 2010 multi-level census files with geography standardized longitudinally, and their 2016 intercensal population estimates. We also have earlier intercensal estimates from other commercial vendors.

Lastly, we use the county-level NCHS bridged-race population estimates for 1990 through 2015 and the NCHS county-level crosswalk file to bridge race/ethnicity categories across multiple decennial census years. In these estimates, NCHS bridges the 31 race categories used in 2000 and 2010 Censuses to the four race categories specified in the 1977 Office of Management and Budget (OMB) standards [14].

Obtaining Census Data

In the early 1990s, we purchased 1990 California Census data directly from the Census Bureau, distributed at that time on a CD as dBase files. These data now are available on the web [15,16]. We used the SF1 and SF3 files for our population counts. For the 1990 Census, the Bureau distributed ZIP-level data only in the SF3 file, with 100 percent population estimates and samples for the other variable sets.

The 2000 Census data are available on the web as text files [17,18]. Programs to read the 2000 data into SAS also are available [19,20], which we modified for our purposes.

For 2010, the Census Bureau only distributed the SF1 file, and switched the equivalent of the SF3 file to the American Community Survey (ACS). We did not use data from the ACS to construct our population estimates for this census year. The 2010 SF1 data is available on the web [21], and we found a program to read it into SAS [22], again modified for our purposes.

We concentrate this discussion on reading the SF1 files into SAS, subsetting the population variables, and preparing the population variables for further use. The Census Bureau distributes data as a series of upwards of 80 distinct files, depending on the year. When we read census files into SAS, we first pull in all files and all geographic summary levels, and do not label variables. When we subset population variables, we only keep the state, county, place, tract, and ZIP-related levels.

Preparing the Census Population

Figure 4. Preparing the 1990 Census Population

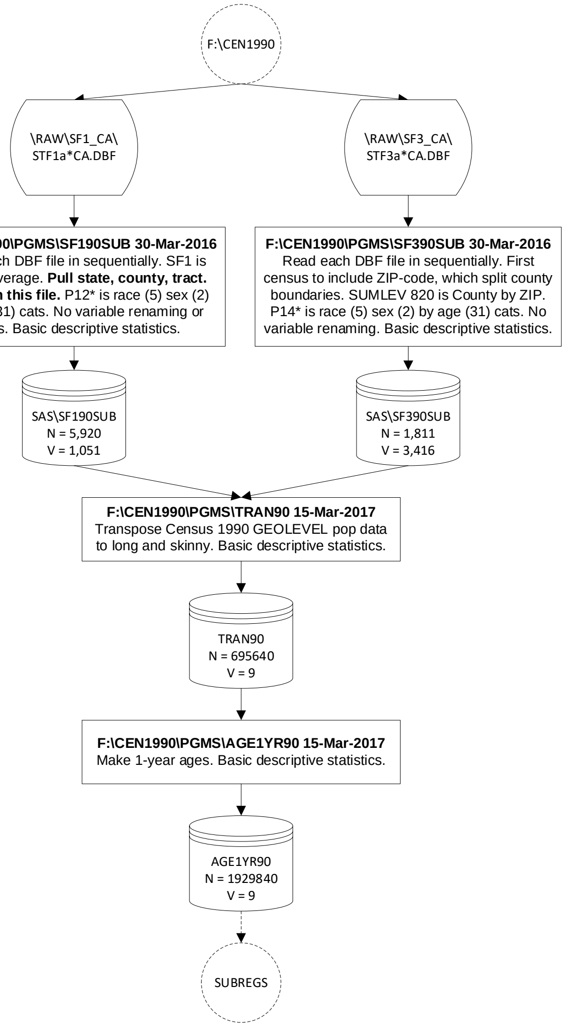

F:\CEN1990
\RAW\SF1_CA\ STF1a*CA.DBF
\RAW\SF3_CA\ STF3a*CA.DBF
F:\CEN1990\PGMS\SF190SUB 30-Mar-2016
Read each DBF file in sequentially. SF1 is 100% coverage. Pull state, county, tract. ZIP not in this file. P12* is race (5) sex (2) by age (31) cats. No variable renaming or labels. Basic descriptive statistics.
F:\CEN1990\PGMS\SF390SUB 30-Mar-2016
Read each DBF file in sequentially. First census to include ZIP-code, which split county boundaries. SUMLEV 820 is County by ZIP. P14* is race (5) sex (2) by age (31) cats. No variable renaming. Basic descriptive statistics.
SAS\SF190SUB N = 5,920 V = 1,051
SAS\SF390SUB N = 1,811 V = 3,416
F:\CEN1990\PGMS\TRAN90 15-Mar-2017
Transpose Census 1990 GEOLEVEL pop data to long and skinny. Basic descriptive statistics.
TRAN90 N = 695640 V = 9
F:\CEN1990\PGMS\AGE1YR90 15-Mar-2017
Make 1-year ages. Basic descriptive statistics.
AGE1YR90 N = 1929840 V = 9
SUBREGS

Figure 4 summarizes steps to prepare a multi-level population file, using the 1990 Census as an example.

The 1990 data we received from the Census Bureau only included the geographic levels we requested. In this instance, we pulled all files directly into our “subset” file, without a first step of reading all geographic summary levels into SAS. In the 1990 Census, ZIP-level data was included only in the SF3, not SF1, file.

The program TRANyy (e.g., TRAN90) pulls each census set of race by age by sex variables and transposes the structure to coincide with the structure of the DOF files: one record per each year, geography level, sex, age, and race/ethnicity combination.

The program AGE1YRyy then converts multi-age groups into one-year estimates. For example, if the age group is 1 to 4, the program assigns an equal proportion of each age to the one-year estimate.

We repeat this sequence of programs for each census period from 1970 onward, using SF3 files as needed for specific projects. We do not detail processing the other census years in this document, as it is similar to what we have just described.

Intercensal Population Estimates

About mid-way through each census period, we obtain commercial intercensal population estimates to estimate sub-county population changes. For the current intercensal period, we turned to GeoLytics [23]. We purchased two national files, one structured like the standard SF1, and the other like the P5 table, which summarizes race and ethnicity over all age groups.

Figure 5. Prepare GeoLytics 2016 estimates

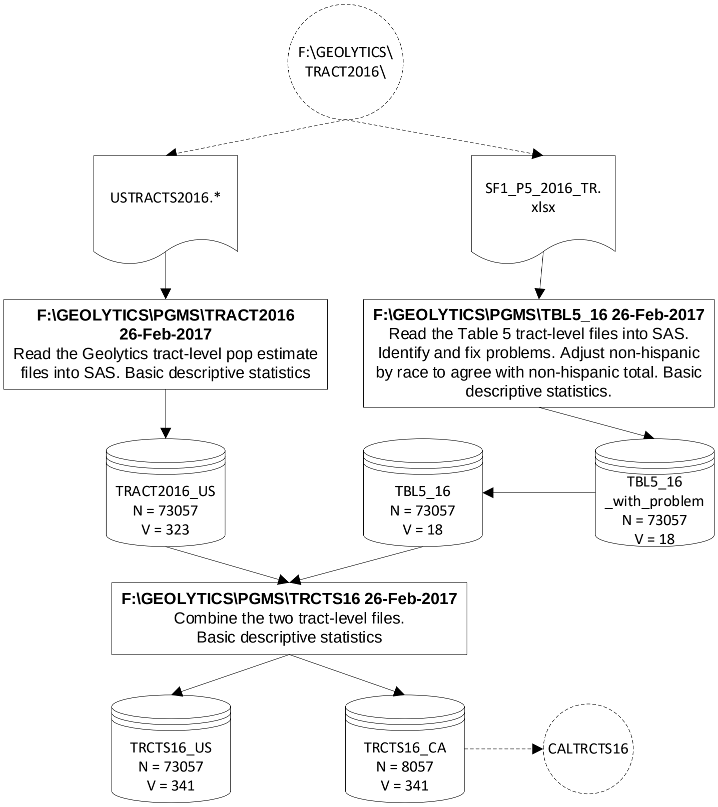

F:\GEOLYTICS\ TRACT2016\
USTRACTS2016.*
SF1_P5_2016_TR. xlsx
F:\GEOLYTICS\PGMS\TRACT2016 26-Feb-2017
Read the Geolytics tract-level pop estimate files into SAS. Basic descriptive statistics
F:\GEOLYTICS\PGMS\TBL5_16 26-Feb-2017
Read the Table 5 tract-level files into SAS. Identify and fix problems. Adjust non-hispanic by race to agree with non-hispanic total. Basic descriptive statistics.
TRACT2016_US N = 73057 V = 323
TBL5_16 N = 73057 V = 18
TBL5_16 _with_problem N = 73057 V = 18
F:\GEOLYTICS\PGMS\TRCTS16 26-Feb-2017
Combine the two tract-level files. Basic descriptive statistics
TRCTS16_US N = 73057 V = 341
TRCTS16_CA N = 8057 V = 341
CALTRCTS16

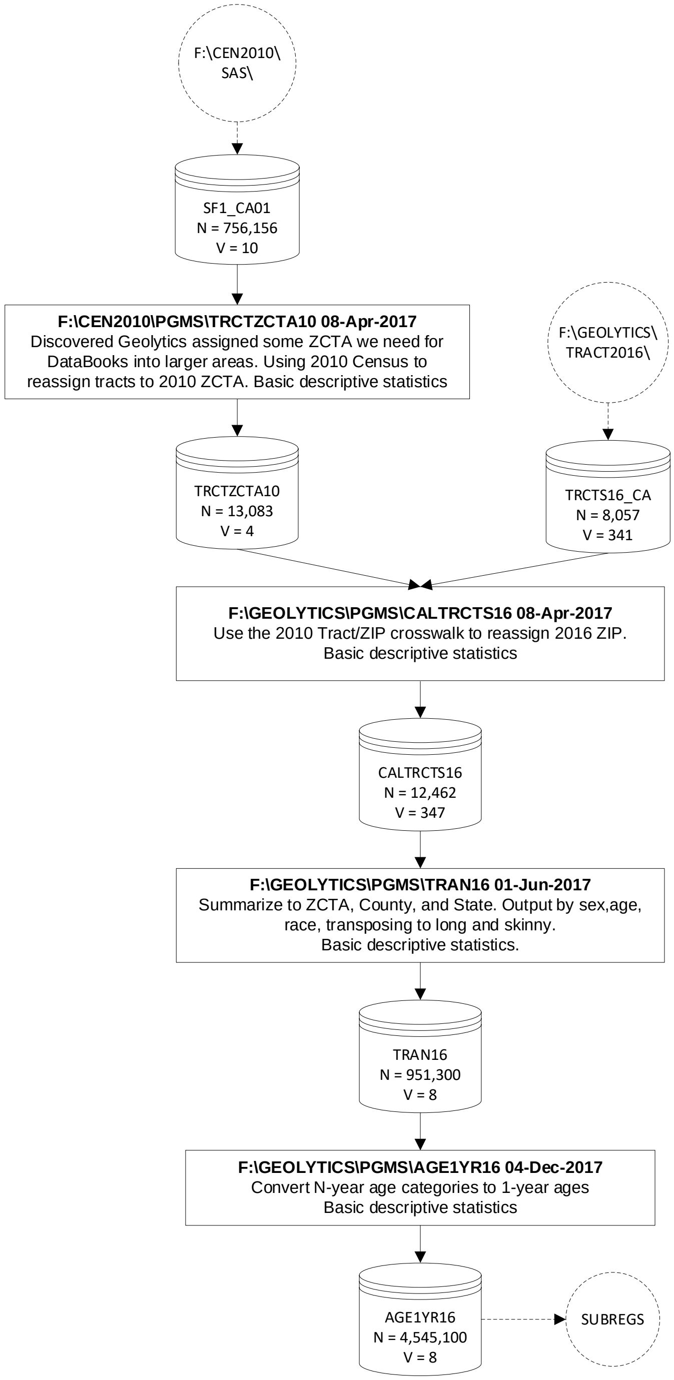
Figure 5 summarizes steps to prepare GeoLytics intercensal estimates, starting with the SF1 file, USTRACTS2016. The large number of fields (V = 323) reflects that race by sex by ethnicity by age are arrayed across the file rather than vertically. GeoLytics also has a set of fields to identify other geographic levels (county, ZIP, place, etc.) for data summary.

The SF1 P5 file has counts for total population, Hispanic ethnicity by race, and geography variables.

The program TRCTS16 puts the two files together, saves the national file then subsets out the California records.

In 1990, the Census Bureau added the ZIP-Code Tabulation Area (ZCTA), a new geography summary level that approximates ZIP-Codes issued by the USPS. Prior to ZCTA creation, many analysts used ZIP-Codes to summarize population size. However, because ZIP-Codes were originally created by USPS to map delivery routes, they do not necessarily represent defined geographic areas. Their boundaries changed periodically to facilitate mail delivery and could cross county boundaries. For these reasons, the Census Bureau created ZCTAs as a more stable, geographically-based alternative to ZIP-Code summary levels. ZCTAs do not change between decennial census years. For an in-depth discussion of this issue, we suggest the monograph that describes how we construct our Geography Master [24].

Figure 6. Bring back lost ZIPS


F:\CEN2010\ SAS\
SF1_CA01 N = 756,156 V = 10
F:\CEN2010\PGMS\TRCTZCTA10 08-Apr-2017
Discovered Geolytics assigned some ZCTA we need for DataBooks into larger areas. Using 2010 Census to reassign tracts to 2010 ZCTA. Basic descriptive statistics
F:\GEOLYTICS\ TRACT2016\
TRCTZCTA10 N = 13,083 V = 4
TRCTS16_CA N = 8,057 V = 341
F:\GEOLYTICS\PGMS\CALTRCTS16 08-Apr-2017
Use the 2010 Tract/ZIP crosswalk to reassign 2016 ZIP. Basic descriptive statistics
CALTRCTS16 N = 12,462 V = 347
F:\GEOLYTICS\PGMS\TRAN16 01-Jun-2017
Summarize to ZCTA, County, and State. Output by sex,age, race, transposing to long and skinny. Basic descriptive statistics.
TRAN16 N = 951,300 V = 8
F:\GEOLYTICS\PGMS\AGE1YR16 04-Dec-2017
Convert N-year age categories to 1-year ages Basic descriptive statistics
AGE1YR16 N = 4,545,100 V = 8
SUBREGS

When we first ran programs to estimate or project small area using the 2016 intercensal population, we discovered that GeoLytics “disappeared” some ZCTAs in the 2010 census, assigning them to larger areas. Figure 6 summarizes steps to recover lost ZCTA.

We returned to the multi-level 2010 SF1 file, and subset out records with both tract and ZIPs, keeping only the geography variables. We merged this with the GeoLytics file to reassign ZIPs, enabling us to proceed with making the files we needed to calculate small area estimates and projections. Lastly, we used our standard steps of transforming the CALTRCTS16 file into long and skinny (TRAN16) and then the one-year age file (AGE1YR16).

NCHS Bridged Race Population

After processing data from the decennial census and intercensal estimates, we standardize race/ethnicity categories using bridged-race data from NCHS. In a collaborative arrangement with the Census Bureau, NCHS releases county-level bridged-race population that it re-estimates annually for the United States to use in calculating national vital rates. Bridged data are available in single-year age from 1990 forward for all counties in the U.S. [26]. Where possible, we prefer to use these rather than DOF files because of their national comparability.

According to NCHS, "race bridging refers to making data collected using one set of race categories consistent with data collected using a different set of race categories, in order to permit estimation and comparison of race-specific statistics at a point in time or over time. More specifically, race bridging is a method used to make multiple-race and single-race data collection systems sufficiently comparable to permit longitudinal estimation and analysis of race-specific statistics." The goal of bridging is to approximate the size of single-race groups rather than to approximate how each individual would have responded to the traditional single-race question [27,28]. The bridging methodology is described elsewhere [9].

NCHS estimates result from bridging the 31 race categories used in Census 2000 and Census 2010, as specified in the 1997 OMB standards for the collection of data on race and ethnicity [29], to the four race categories specified in the 1977 OMB standards (White, Black, API, AIAN) by Hispanic and non-Hispanic ethnicity [30].

Figure 7. Prepare NCHS 1990-1999 files

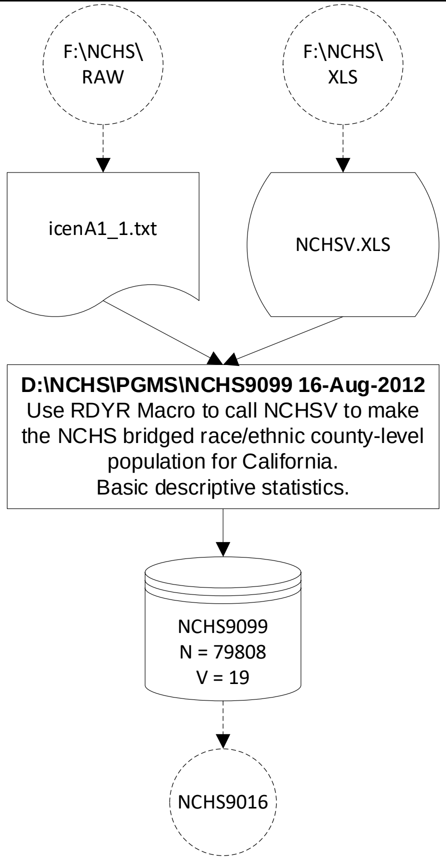
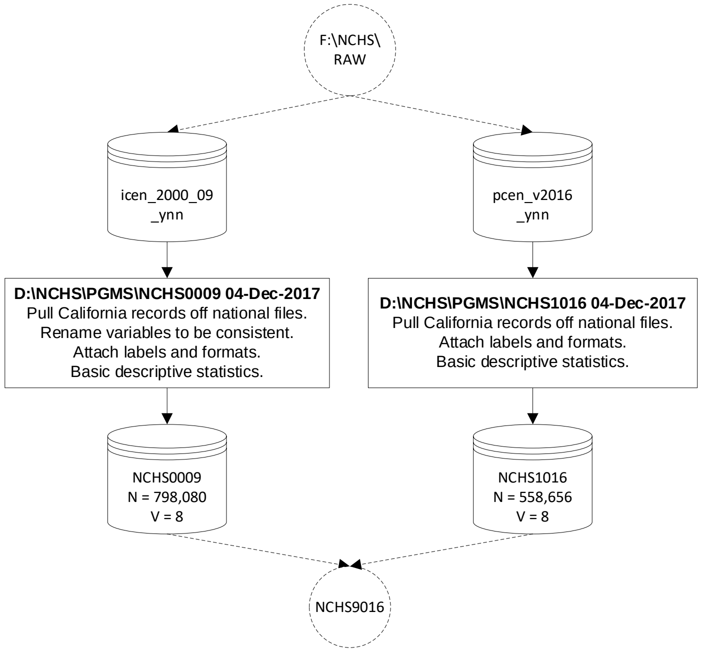

F:\NCHS\ RAW
F:\NCHS\ XLS
icenA1_1.txt
NCHSV.XLS
D:\NCHS\PGMS\NCHS9099 16-Aug-2012
Use RDYR Macro to call NCHSV to make the NCHS bridged race/ethnic county-level population for California.
Basic descriptive statistics.
NCHS9099
N = 79808
V = 19
NCHS9016

Figure 7 summarizes steps to prepare 1990-1999 NCHS data. NCHS distributes these years as zipped text files by sets of states. Thereafter, all years are available as annual SAS files with all states included. On our working drive, in this case D:\NCHS, we subset and store California data from the Non-Confidential (F) Drive.

To prepare the 1990-1999 data, we set up the RDYR macro to give control to the NCHSV spreadsheet. We describe this process elsewhere [3].

Figure 8. Prepare NCHS 2000-2016 files


F:\NCHS\ RAW
icen_2000_09 _ynn
pcen_v2016 _ynn
D:\NCHS\PGMS\NCHS0009 04-Dec-2017
Pull California records off national files.
Rename variables to be consistent.
Attach labels and formats.
Basic descriptive statistics.
D:\NCHS\PGMS\NCHS1016 04-Dec-2017
Pull California records off national files.
Attach labels and formats.
Basic descriptive statistics.
NCHS0009 N = 798,080 V = 8
NCHS1016 N = 558,656 V = 8
NCHS9016

NCHS distributes years 2000-2016 data as text or SAS files. We downloaded the annual SAS files. The variables are slightly different in the two sets. Note that the outgoing file NCHS1016 becomes obsolete annually. As NCHS updates the post-censal projections, a new program is run for each annual revision of the series. This work is shown in Figure 8.

Figure 9. The NCHS population files

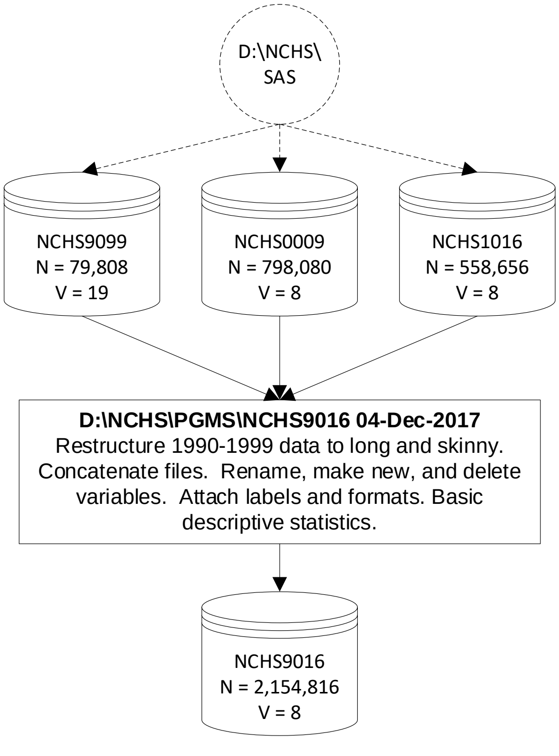

D:\NCHS\ SAS
NCHS9099 N = 79,808 V = 19
NCHS0009 N = 798,080 V = 8
NCHS1016 N = 558,656 V = 8
D:\NCHS\PGMS\NCHS9016 04-Dec-2017
Restructure 1990-1999 data to long and skinny.
Concatenate files. Rename, make new, and delete variables. Attach labels and formats. Basic descriptive statistics.
NCHS9016 N = 2,154,816 V = 8

Figure 9 summarizes our work to integrate the various NCHS population files. At this point, we put the three sets of intercensal estimates together. Note that the 1990-1999 file is “wide”, with annual population having a variable name “POPyyyy”. To be consistent with our methods for the California DOF population files, we make this file “long and skinny” before concatenating the other files. The resulting product has one record per year, county, single-year age up to 85, race category, and Hispanic ethnicity.

To distribute undefined (multi-race, other, unknown) race groups, as found in our population health datasets, NCHS assigns single race randomly, using local distributions for Hispanic, sex, and age. NCHS provided FHOP with the file it uses to bridge multiple-race data into the single race format for vital statistics reporting [32]. This file and the SAS macro to carry out the reassignment is available upon request.

INTERCENSAL SMALL AREA POPULATION ESTIMATES

To estimate intercensal small area population, FHOP uses census and intercensal files, summarized to the geographic levels of interest, and either DOF or NCHS annual county-level population files. Note that we have previously standardized race/ethnicity in all sets of incoming files, with undefined race groups randomly assigned to defined groups. For our example in this document, we use DOF data for county-level population, the 1990, 2000, 2010 Census and the 2016 GeoLytics intercensal files to describe how we prepare small area files.

The Geography Master

We maintain an excel file (GEOGMAS.XLSX) with all ZIPs found in any year of every dataset we process. This includes a count of the number of times the ZIP appeared within and over all the datasets, and the earliest and most recent year the ZIP occurred. We call this our Geography Master [13].

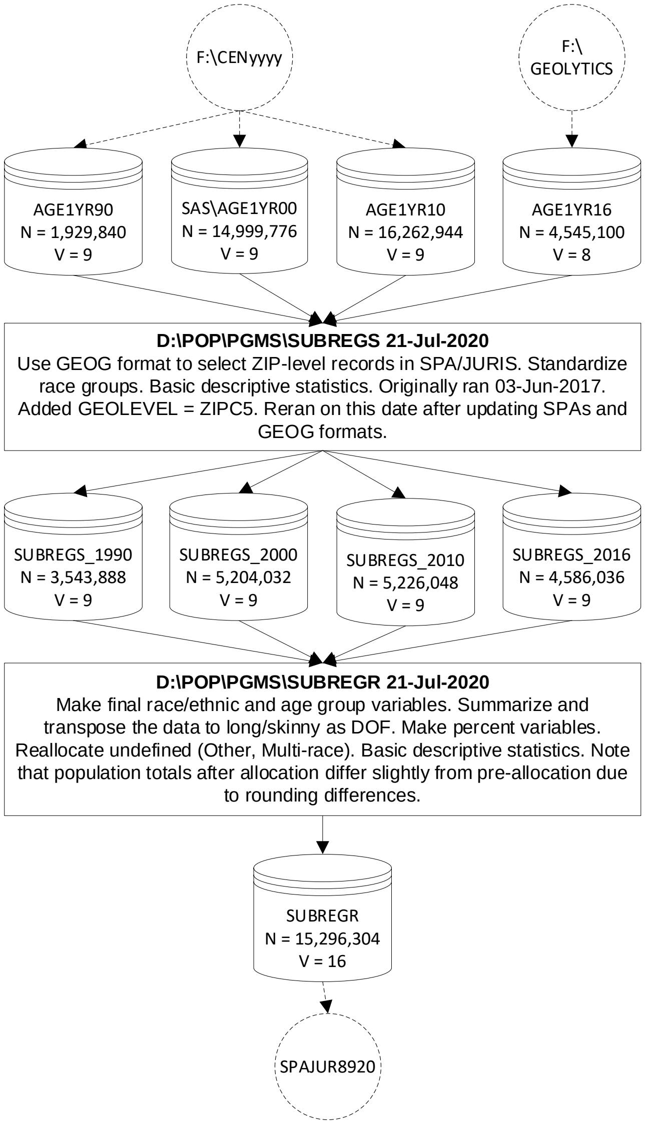
Our Geography Master contains latest known information on each ZIP-Code we have encountered in our data sources. For ZIP-Codes, we use various commercial products help us classify ZIPs as to their type (unique, military, post office, or standard), whether the USPS discontinued the ZIP, whether the ZIP split into two or more parts, or whether it is new. Following a rule from California’s Office of Statewide Health Planning and Development (OSHPD) for its hospital datasets, we do not allow ZIPs to cross county boundaries. When in doubt, we assign the ZIP to the county with the largest portion of the population rather than the largest area.

We use this file to identify areas smaller than the county but larger than the ZIP, to accommodate counties that have asked us to provide smaller-area products based on supervisor districts or other locally-defined geographies. Sub-county geographies included in the Master are: service planning areas (SPA), special jurisdictions (JURIS), or health service areas (HSA). For our Geography Master, we define SPAs as districts used by county health departments to divide their jurisdictions into meaningful sub-regions. If the local health department did not specify any SPAs, we assign county supervisor districts as SPAs. Supervisor districts are updated after every decennial census. In California, there are 58 county-based health departments and three independent health jurisdictions that are administered separately from their counties (i.e. Long Beach, Pasadena, and Berkeley); in our Geography Master, these latter three are assigned as special JURIS. For example, Los Angeles County divides itself into eight SPAs [33], and contains two JURIS (Long Beach and Pasadena).

After assigning zips to county sub-regions, we make a series of formats to make products such as our DataBooks. Here, we use formats to classify ZIPs into SPA and JURIS before summarizing files to get populations. Before running the programs to generate small area statistics, it is **critical** to confirm that ZIP-Code assignments to small areas in the Geography Master are correct and to re-build the geography format library as needed. If assignments are incorrect, small area numerators and population denominators will be much larger or smaller than reality.

Prepare Small-Area Files

Figure 10. Prepare small-area files



F:\CENyyyy
F:\ GEOLYTICS
AGE1YR90 N = 1,929,840 V = 9
SAS\AGE1YR00 N = 14,999,776 V = 9
AGE1YR10 N = 16,262,944 V = 9
AGE1YR16 N = 4,545,100 V = 8
SUBREGS_1990 N = 3,543,888 V = 9
SUBREGS_2000 N = 5,204,032 V = 9
SUBREGS_2010 N = 5,226,048 V = 9
SUBREGS_2016 N = 4,586,036 V = 9
SUBREGR N = 15,296,304 V = 16
SPAJUR8920

D:\POP\PGMS\SUBREGS 21-Jul-2020
Use GEOG format to select ZIP-level records in SPA/JURIS. Standardize race groups. Basic descriptive statistics. Originally ran 03-Jun-2017. Added GEOLEVEL = ZIPC5. Reran on this date after updating SPAs and GEOG formats.

D:\POP\PGMS\SUBREGR 21-Jul-2020
Make final race/ethnic and age group variables. Summarize and transpose the data to long/skinny as DOF. Make percent variables. Reallocate undefined (Other, Multi-race). Basic descriptive statistics. Note that population totals after allocation differ slightly from pre-allocation due to rounding differences.

Figure 10 summarizes steps to prepare census data to interpolate small-area intercensal population estimates using the 1-year age files. Note that census year populations tie exactly to the California DOF population files.

In the program SUBREGS, we use Geography Master formats to classify small areas such as SPA, JURIS, or others. Because different periods had different race/ethnicity definitions, we standardize to common groups within time. We output data for counties with sub-regions.

The program SUBREGR redefines race/ethnicity to prepare for bridging, makes a 5-year age group variable, and summarizes the data keeping both the 1-year age and 5-year age group variables.

The macro RACE_ALLOC then follows federal decision rules to allocate undefined race groups (other, multi-race) to the final race/ethnicity. A new variable POPB (Population, bridged) is the sum of the original POP variable plus the number assigned to it because of bridging.

For longitudinal consistency, all years of SUBREGR have the 5-category race/ethnic variable (White, Black, Hispanic, API, AIAN) available before 2000. We also carry along a 7-category variable (White, Black, Hispanic, Asian, Pacific Islander, AIAN, and multi-race) available from 2000 onward. We only perform small-area statistics using the 5-category race/ethnicity variable.

Interpolate Population between Censuses

Figure 11. Interpolate population

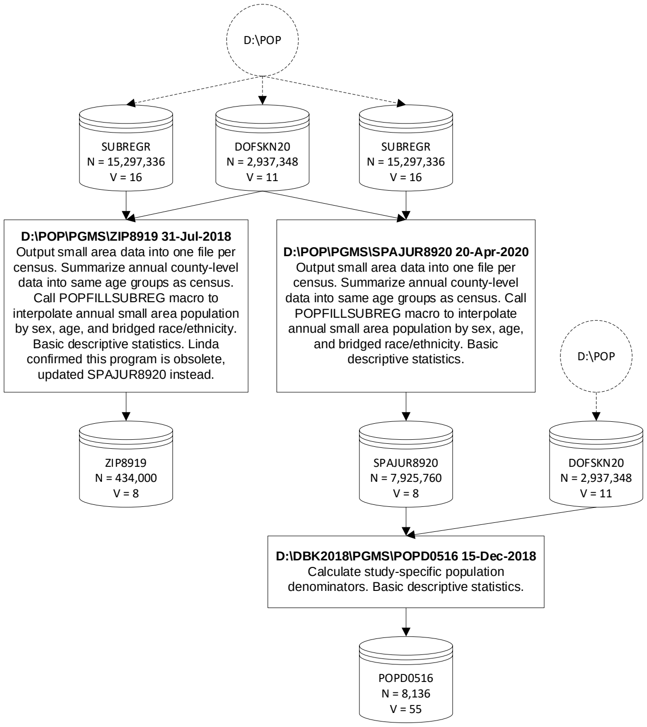

D:\POP
SUBREGR N = 15,297,336 V = 16
DOFSKN20 N = 2,937,348 V = 11
SUBREGR N = 15,297,336 V = 16
D:\POP\PGMS\ZIP8919 31-Jul-2018
Output small area data into one file per census. Summarize annual county-level data into same age groups as census. Call POPFILLSUBREG macro to interpolate annual small area population by sex, age, and bridged race/ethnicity. Basic descriptive statistics. Linda confirmed this program is obsolete, updated SPAJUR8920 instead.
D:\POP\PGMS\SPAJUR8920 20-Apr-2020
Output small area data into one file per census. Summarize annual county-level data into same age groups as census. Call POPFILLSUBREG macro to interpolate annual small area population by sex, age, and bridged race/ethnicity. Basic descriptive statistics.
D:\POP
ZIP8919 N = 434,000 V = 8
SPAJUR8920 N = 7,925,760 V = 8
DOFSKN20 N = 2,937,348 V = 11
D:\DBK2018\PGMS\POPD0516 15-Dec-2018
Calculate study-specific population denominators. Basic descriptive statistics.
POPD0516 N = 8,136 V = 55

Figure 11 summarizes the last steps to interpolate small-area intercensal estimates and prepare study-specific population denominator files. In this example, we used DOF data, which we have to use for our DataBooks since it is a product for the State of California. But we could as well have used NCHS data, as both files are similarly structured.

Recall that the file SUBREGR has small-area census data spaced a decade apart, while DOFSKN20 has annual county data in one-year age groups. We begin the program SPAJUR8920 by outputting data for each census into temporary separate files and then further splitting those by sub-regions. From DOFSKN20, we select counties with sub-regions and summarize county-level data into the same 5-year age groups as the census data.

Then SPAJUR20 calls the macro POPFILLSUBREG to interpolate annual small area population. It performs the following steps for each stratum of sex, age, and race/ethnicity category. First, for a year with data for both the county and sub-county regions (“census” years), the macro divides the two populations to obtain the percent of the county population in each sub-region. Note as here, that “census” data can be from a census year or an inter-censal year.

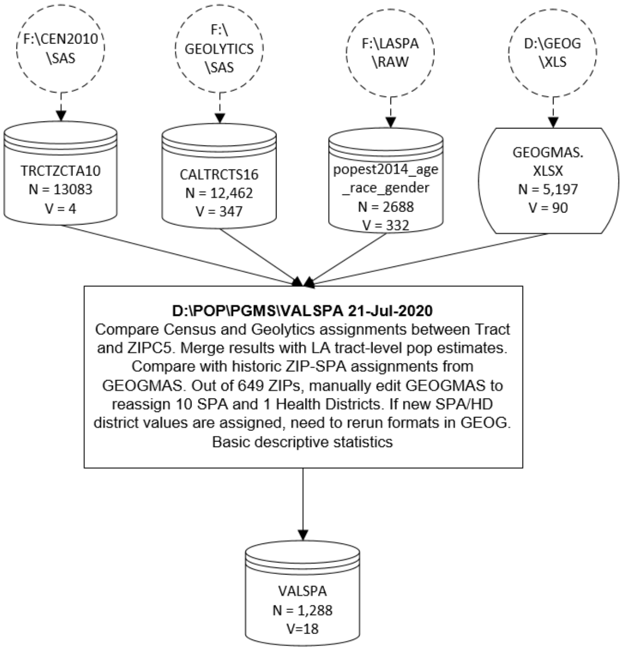
Second, the macro fills in a sub-region percent of county for all years. This step uses linear interpolation if the year is between two censuses, or uses the sub-region percent from the nearest census if not. Finally, for all years, including census years, we multiply the percent of county population with the DOF county population to obtain the sub-region population.

To make a study-specific population, we bring together the interpolated small area and county-level populations, calculate denominator variables of interest, and store the resulting file in the study-specific directory. In the example POPD0617, we are preparing 12-year annual population numbers needed to construct FHOP’s DataBook products for California counties and special jurisdictions over the interval 2006-2017. Notice that we have gone from a skinny to a wide file, with one record per geography level (county, SPA, jurisdiction) and race/ethnicity per year, and population for each age group needed (females aged 15 to 44, children aged 0 to 14, etc.).

Validating the Small Area Allocations

Before developing our current method of estimating population for county sub-regions, Los Angeles County provided us with their SPA estimates, which we used when making DataBooks. They also provided lists of ZIPs assigned to each SPA.

Figure 12. Validating Los Angeles County ZIP to SPA assignments



F:\CEN2010 \SAS
F:\ GEOLYTICS \SAS
F:\LASPA \RAW
D:\GEOG \XLS
TRCTZCTA10 N = 13083 V = 4
CALTRCTS16 N = 12,462 V = 347
popest2014_age _race_gender N = 2688 V = 332
GEOGMAS. XLSX N = 5,197 V = 90
D:\POP\PGMS\VALSPA 21-Jul-2020
VALSPA N = 1,288 V=18

Compare Census and Geolytics assignments between Tract and ZIPC5. Merge results with LA tract-level pop estimates. Compare with historic ZIP-SPA assignments from GEOGMAS. Out of 649 ZIPs, manually edit GEOGMAS to reassign 10 SPA and 1 Health Districts. If new SPA/HD district values are assigned, need to rerun formats in GEOG. Basic descriptive statistics

Figure 12 shows the steps we took to validate FHOP’s methods for producing SPA-level estimates compared to estimates from Los Angeles County. We subset out Los Angeles County census tract/ZCTAs from the 2010 Census and the Geolytics intercensal estimates, and merged the two datasets by census tract and ZIP. We then merged these census estimates with population estimates provided by Los Angeles County, again by census tract and ZIP.

.

We then applied the SPA assignments from our Geography Master to the merged dataset, and compared the SPA-level population estimates by four data sources: 2010 Census, 2016 GeoLytics estimates, DOF projections and data provided by Los Angeles County. We examine the totals to ensure that the estimates are similar to one another

RESOURCES

In this monograph, we described how we prepare non-confidential population data to use for denominators, including methods when small area data are not available consistently over long periods. Be aware of the limitations of small area geographic data. Do not push the design, analysis, or interpretation beyond what is supportable given known methodologic shortcomings.

All programs are available upon request. Users will have to contract for more than one hour of support.

ENDNOTES

[1] Remy L, Clay T. (2016) Managing Longitudinal Research Studies: The Basic Computing Environment. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

[2] Remy L, Clay T. (2016) Managing Longitudinal Research Studies: Standardizing Variables Over Time. San Francisco, CA: University of California, San Francisco, Family Health Outcomes Project. Available at: http://fhop.ucsf.edu/data-management-methods.

[3] 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

[4] State of California, Department of Finance, Race/Ethnic Population with Age and Sex Detail, 1970–1989. Sacramento, CA, December 1998. Last accessed 17-Jul-2020 at: http://www.dof.ca.gov/Forecasting/Demographics/Estimates/Race-Ethnic/1970-89/.

[5] State of California, Department of Finance, Race/Ethnic Population with Age and Sex Detail, 1990–1999. Sacramento, CA, Revised May 2009. Last accessed 17-Jul-2020 at: http://www.dof.ca.gov/Forecasting/Demographics/Estimates/Race-Ethnic/1990-99/index.html.

[6] State of California, Department of Finance, State of California, Department of Finance, Race/Hispanics Population with Age and Gender Detail, 2000–2010. Sacramento, California, September 2012. September 6, 2012 (Original Release - integer data only), September 14, 2012 (Revised to include decimal detail), March 19, 2013 (Removed data for July 2010) Last accessed 17-Jul-2020 at: http://www.dof.ca.gov/Forecasting/Demographics/Estimates/Race-Ethnic/2000-2010/

7 State of California, Department of Finance, State and County Population Projections by Race/Ethnicity, Sex, and Age 2010-2060, Sacramento, California, December 2014. Note that we use the full P-3 file. Last accessed 17-Jul-2020 at: http://www.dof.ca.gov/ Forecasting/Demographics/Projections/documents/P3_Complete. zip. The internal file date is 10-Jan-2020.

8 Remy L, Clay T, Oliva G. (2011). Issues and Decisions to be made on Collecting, Coding and Reporting Race and Ethnicity for Public Health Indicators. Family Health Outcomes Project, University of California, San Francisco. Updated January 2017. Available at: http://fhop.ucsf.edu/data-management-methods

9 Ingram DD, Parker JD, Schenker N, Weed JA, Hamilton B, Arias E, Madans JH (2003) United States Census 2000 population with bridged race categories. National Center for Health Statistics. Vital Health Stat(2)135. 2003. Last accessed 17-Jul-2020 at: http://www.cdc.gov/nchs/data/series/sr_02/sr02_135.pdf

10 Summary File 1 Dataset. U.S. Census Bureau. Last accessed 17-Jul-2020 at: https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

11 Summary File 3. 2000 Census of Population and Housing. Technical Documentation. U.S. Census Bureau. Last accessed 17-Jul-2020 at: https://www.census.gov/prod/cen2000/doc/sf3.pdf

12 About the American Community Survey. U.S. Census Bureau. Last accessed 17-Jul-2020 at: https://www.census.gov/programs-surveys/acs/about.html

14 U.S. Census Populations with Bridged Race Categories – Overview. National Center for Health Statistics. Last accessed 17-Jul-2020 at: https://www.cdc.gov/nchs/nvss/bridged_race.htm

15 1990 Census Summary File 1 ASCII files and CD-ROM discs. Last accessed 17-Jul-2020 at: http://www2.census.gov/census_1990/1990STF1.html

16 1990 Census Summary File 3 CD-ROM discs. Last accessed 17-Jul-2020 at: http://www2.census.gov/census_1990/1990STF3.html

17 2000 Census Summary File 1 ASCII files. Last accessed 17-Jul-2020 at: http://www2.census.gov/census_2000/datasets/Summary_File_1/California/

18 2000 Census Summary File 3 ASCII files. Last accessed 17-Jul-2020 at: http://www2.census.gov/census_2000/datasets/Summary_File_3/California/

19 Last accessed 17-Jul-2020 at: http://www.census.gov/support/2000/SF1/SF1SAS.zip

20 Last accessed 17-Jul-2020 at: https://www.census.gov/support/2000/SF3/SF3SAS.zip

21 Last accessed 17-Jul-2020 at: http://www2.census.gov/census_2010/04-Summary_File_1/California/

22 Last accessed 17-Jul-2020 at: http://www.sascommunity.org/wiki/Converting_2010_Census_Summary_File_1_%28SF1%29_Data_into_SAS_Data_Sets#SAS_CODE

23 Neighborhood Change Database. Geolytics. Last accessed 17-Jul-2020 at: https://GeoLytics.com/products/normalized-data/neighborhood-change-database

24 Remy L, 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.

26 See: http://www.cdc.gov/nchs/nvss/bridged_race.htm

27 Parker JD, Schenker N, Ingram DJ, Weed JA, Heck KE, Madans JH (2004) Bridging between two standards for collecting information on race and ethnicity: An application to Census 2000 and vital rates. Public Health Reports, 119:Mar-Apr, 192-205.

28 Office of Management and Budget (2000). Appendix C: The Bridge Report: Tabulation Options for Trend Analysis. Provisional guidance on the implementation of the 1997 standards for federal data on race and ethnicity. Last accessed on 21-Jul-2020 at: https://www.esd.whs.mil/Portals/54/Documents/DD/info_collect/files_public/Race%20%20Ethnicity%20Guidance.pdf?ver=2018-11-01-094407-913

29 Office of Management and Budget. Revisions to the standards for the classification of Federal data on race and ethnicity, Federal Register 62FR58781–58790, October 30, 1997. Last accessed 21-Jul-2020 at: https://obamawhitehouse.archives.gov/omb/fedreg_1997standards/.

30 Office of Management and Budget. Race and ethnic standards for Federal statistics and administrative reporting. Statistical Policy Directive 15, 1977. Last accessed 21-Jul-2020 at: http://wonder.cdc.gov/wonder/help/populations/bridged-race/Directive15.html

32 National Center for Health Statistics (2004). NCHS Procedures for Multiple-Race and Hispanic Origin Data: Collection, Coding, Editing, and Transmitting. Division of Vital Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention. May 7, 2004. Last accessed 21-Jul-2020 at: http://www.cdc.gov/nchs/data/dvs/Multiple_race_docu_5-10-04.pdf.

33 Service Planning Areas. County of Los Angeles Public Health. Last accessed 21-Jul-2020 at: http://publichealth.lacounty.gov/chs/SPAMain/ServicePlanningAreas.htm