POLIDATA® Political Data Analysis

DATABASE DEVELOPMENT, ANALYSIS AND PUBLICATION;
POLITICAL AND CENSUS DATA; LITIGATION SUPPORT

CLARK BENSEN

POLIDATA · 1303 HAYWARD ROAD, P. O. BOX 530 · CORINTH, VT 05039
Tel: 703-690-4066 · Fax: 202-318-0793 · email: clark@polidata.org

PUBLISHER OF THE POLIDATA ® DEMOGRAPHIC AND POLITICAL GUIDES AND ATLASES

Notes on the PL94-171 Prototype for the 2010 Census

NCSL Redistricting Seminar #1, San Francisco, June 2009

The Bureau of the Census recently released prototype files with test data for San Joaquin, California, the site of a Dress Rehearsal in 2008. The following are notes that might be useful in preparing to use the final 2010 files when they are released in the winter-spring of 2011. The San Joaquin data files are accompanied by a technical documentation booklet of 179 pages. Selected text from that document has been included here for reference in the boxed text.

Subject Matter: The 2010 Census is a Census of Population and Housing and the PL files[1] will contain some housing and population data for the lowest level of census geography, the census block. The housing data are limited to: a) the number of housing units and b) occupancy status, a count of occupied versus vacant units. The population data are limited to race and Hispanic origin categories, broken out for the total population and for the voting age population.

Even though the census questionnaire[2] asks a few other questions (relationship to householder; sex; and age) the PL data files will not contain any direct information related to these questions, though some of this information will be available in the Summary File 1 release planned for the summer of 2011.

Nor will the PL data files contain any information as to socio-economic status. This type of information will be available via the American Community Survey (ACS), which has replaced the long form for the collection of data from a sample of the population. Data from the ACS, e.g., citizenship status, will not be available by census block but should be available at the higher levels of geography of the block group and/or census tract. The ACS is an ongoing rolling survey and the first release that will report information

[1] The data files are often referred to as the "PL" files for a simple reason: it was Public Law 94-171 that established the legislative apportionment data program in the 94th Congress. It was introduced as H.R. 1753 in January of 1975 and signed into law by President Gerald R. Ford on December 23, 1975.

[2] Remember, there is only one questionnaire for Census 2010, what we would have called the "short form" in previous years.

for all sizes of communities is scheduled for release in December of 2010, though based upon the 2000 census geography and reflecting the five-year survey period of 2005 through 2009.

Race and Hispanic Origin: The census form for 2010 will continue the form's format from 2000 by asking race and Hispanic origin as separate questions[3]. The tables will report these data in two basic modes: a) each question separately, i.e., race in one table and Hispanic origin in another separate table; and b) a combination of both questions, i.e., a count of persons of Hispanic origin and the racial breaks for Non-Hispanic persons as part of the same table. In addition, as with the 2000 Census, the population data are also provided for the six race categories and the 57 combinations of race that tabulates up to six responses to the race question.

These basic modes of reporting the population data are represented in four tables that allow for a break of the population by voting age. The tables first report the two modes for the total population and then for the voting age population:

Universe: Total population
P1. RACE [71 cells]
P2. HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE [73 cells]

Universe: Total population 18 years and over
P3. RACE FOR THE POPULATION 18 YEARS AND OVER [71 cells]
P4. HISPANIC OR LATINO, AND NOT HISPANIC OR LATINO BY RACE FOR THE POPULATION 18 YEARS AND OVER [73 cells]

The mode of reporting provides two basic options for the racial and Hispanic origin data: whether the sum of all the population subgroups should equal the population total or not. This is not a very likely scenario because the total for each PL record does equal the total of all the appropriate cells listed therein. Yet such a problem could arise if the researcher tries to report the race from one table and the Hispanic origin from another.

The important factors to remember here are a) that Hispanics can be of any race and many of them will be tabulated as White; and, b) the census form asks this information in two questions (Hispanic origin and then race). Tables P1 and P3 detail the racial breakdowns from the race question only and persons of Hispanic, Latino or Spanish origin may be included in any racial subgroup[4]. Tables P2 and P4 combine the results

[3] See the Appendix for the Questionnaire.

[4] The race responses of Hispanics in the 2000 Census were 48% White, 42% Other, 6% Multi, 2% Black, 1% Am. Indian, 0.3% Asian and 0.1% Pac. Islander. The proportion of race responses in 2000 who were

from both questions and summarize the Hispanic and Non-Hispanic population first and then provide the detail for the Non-Hispanic persons by racial subgroups. Tables P2 and P4 thus provide information for the majority/minority subgroups that are most relevant to the political environment.

Multiple Race Responses: The 2000 Census was the first census to allow respondents to choose more than one race by multiple responses to the race question.[5] This option has been continued for the 2010 Census and remains the major reason why there are so many data cells (291) in the files[6].

The data for each of the four tables are reported in a similar manner in the sense that information on the standard race categories will appear before the combination of races for the multiple race responses. The standard race groups (White; Black or African American; American Indian or Alaska Native; Asian; Native Hawaiian and Other Pacific Islander; and Some Other Race) are all followed by the term “alone”. The remainder of the data cells represent the combinations, e.g., from White; Black or African American [2 responses] to White; Black; American Indian and Alaska Native; Asian; Native Hawaiian and Other Pacific Islander; Some other race [6 responses].

> Two or more races. People may choose to provide two or more races either by checking two or more race response check boxes, by providing multiple responses, or by some combination of check boxes and other responses. The race response categories shown on the questionnaire are collapsed into the five minimum race groups identified by OMB, and the Census Bureau’s ‘‘Some other race’’ category. For data product purposes, ‘‘Two or more races’’ refers to combinations of two or more of the following race categories:
>
> 1. White
> 2. Black or African American
> 3. American Indian or Alaska Native
> 4. Asian
> 5. Native Hawaiian or Other Pacific Islander
> 6. Some other race
>
> There are 57 possible combinations … involving the race categories shown above. Thus, according to this approach, a response of ‘‘White’’ and ‘‘Asian’’ was tallied as two or more races, while a response of ‘‘Japanese’’ and ‘‘Chinese’’ was not because ‘‘Japanese’’ and ‘‘Chinese’’ are both Asian responses.

Hispanic was 97% of Other; 33% of Multi; 16% of Am. Indian; 11% of Pac. Islander; 8% of White; 2% of Black, and 1% of Asian. These values are based upon total population.

[5] It was also the first to include Native Hawaiian and Other Pacific Islander as a separate race group.

[6] See the Appendix for the 1990 Table Matrix which included 24 data cells.

Summary Levels: The files contain information for each census block[7] as well as higher levels of geography in the census hierarchy. For example, the counts for each block are summed to derive the totals for the block group, census tract, county or state. Ultimately, it is the sum of the census blocks that determine how many persons reside in the nation as of the April 1, 2010 record date. The sum of all these records represents the number used to determine the apportionment counts for seats in the U.S. House in December 2010.

Most of the summary levels provided for 2010 are similar to those provided for 2000 except that there has been a large increase in the number of overall summary fields due to the addition of additional levels for three types of school districts and congressional and state legislative districts. Of course, the district information for Congress and the State Legislatures reflect districts established following the 2000 census[8].

The summary levels for the most common geographic areas remain the same, e.g.: census block (SL 750); VTD (SL 700); county (SL 050); and state (SL 040). A count of the number of records for the San Joaquin data is included in the appendix. There are 15,762 records on the file, 12,096 of which have housing units and 12,094 of which have population. Over 2/3 of the records are block records, though about 25% of them have no population. The file includes records for 390 block groups (SL 150), 121 census tracts (SL 140), 20 census places (SL 160), and 9 county subdivisions (SL 060).

A key level of geography for redistricting users is the VTD, or Voting District. Most of us consider these to be precincts but this isn't always the case.

Voting district is the generic name for areas, such as precincts and wards, that are established by state and local governments for the purpose of elections. States participating in the 2010 Census Redistricting Data Program provide the Census Bureau with boundaries, codes, and optionally names, for their voting districts. The voting district code is a six-character field (position 162) that may contain any ASCII character.

Voting District Indicator—States participating in the 2010 Census Redistricting Data Program have the option to indicate whether the voting district boundaries they submit to the Census Bureau are actual or pseudo. Pseudo voting districts are those that were approximated. These approximated voting districts are represented in the voting district indicator field (position 168) as a P. Actual voting districts are represented in this field as an A. Actual voting districts are additionally identified by an asterisk (*)

[7] Note that some blocks are actually "split" by other levels in the census hierarchy, e.g., VTD, county subdivision, or place: SL 750 is block within BG within TR within PL within CS within CY within VTD.

[8] See the Appendix for the 2010 and 2000 Summary Level Sequence Charts.

appended to the voting district name in the Area Name/LSAD Term/Part Indicator. If a state did not provide voting districts for a county, this field will be blank.

Of course, even if the VTDs do represent actual precincts[9], they may be frozen in time before the 2010 census. The actual boundaries may well change before the 2011/2012 elections due to population shifts or districting. Nevertheless, the VTD is generally the key to linking the political/election data, to the demographic/racial data.

The San Joaquin file includes records for 213 VTDs (SL 700), 196 of which have population. It includes records for 231 VTD components within state senate districts, or SLDU, (SL 630) and 227 VTD components within state assembly districts, or SLDL, (SL 635). This is one instance where the hierarchy is slightly different for the levels for congressional districts versus the levels for state legislative districts: VTDs within congressional districts are not included.

Data Structure: As detailed below, there are three disk files for each logical record for each unit of geography:
a) one with geographic header information and
b) one file for each of two segments of the logical record that contains the housing and population data.

DATA STRUCTURE AND SEGMENTATION
The data in the redistricting files and other 2010 Census summary files are segmented. This is done so that individual files will not have more than 255 fields, facilitating export into spreadsheet or database software. The segmentation information discussed here applies to the DVD files. The redistricting data and the corresponding geographic information for an individual state is known as the file set. This is the package that the DVD for a state will contain. Because of the large size of the tables, the file set will be broken into three files. These files will contain:

- Geographic Header Record file.
- File01 (Tables P1 and P2).
- File02 (Tables P3, P4, and H1).

File Structures: File structure information is provided in the DATA\REF folder[10]: a) the structure for fields in the geographic header record can be found in the GEOSTRUC.DBF; and b) the structure for fields in the data segments can be found in the SEG01.DBF and SEG02.DBF files.

[9] Remember also that even if they are actual precincts, precincts may be combined for election purposes.

[10] References here are to the DVD: Redistricting Data Prototype (P.L. 94-171) Summary File; 2010 Census of Population and Housing, 2008; released April 2009, Washington, DC. URL: http://www.census.gov/rdo/ .

These structure files contain a textual description of the field, as well as the field type and length. In addition, the field name has a maximum length of 10 characters which means it meets the .dbf standard for the first 10 characters being unique. The fields in the geographic structure file are a mix of character and numeric fields. The fields in the data segment structure fields are a mix of character and numeric fields. However, the only character fields in the data segment files are the first five that are used for identification and record location (see below); all housing and population data fields are numeric with a length of 9. Fortunately, the names of the fields and the descriptions appear to be identical to those used in Census 2000 data products.

The record location fields contained in each record are a subset of fields from the geographic header file:

1) file identification (FILEID) [C-6, e.g., PL08];
2) state/U.S. abbreviation from the USPS (STUSAB) [C-2, e.g., CA];
3) characteristic iteration (CHARITER) [C-3, e.g., 000];
4) characteristic iteration file sequence number (CIFSN) [C-2, e.g., 01];
5) a logical record number (LOGRECNO) [N-7, e.g., 0000123, or 123].

However, even though these five fields are carried forward into each record segment, not all are used for the PL94-171 datasets[11].

The geographic header record is standard across all electronic data products from the 2010 Census. Since the 2008 Redistricting Data Prototype files are quite simple, many of the fields, including some header fields that appear in all three files (geographic header, file01, and file02), are not used. For example, the characteristic iteration (CHARITER) field will be used in the 2010 Census Summary File 2. In this prototype, and in the 2010 Census redistricting data file, it is always coded as 000.

Database Considerations: Horizontal/Fields: Note also that the name of the data cell must also be used in combination with the name of the table and/or higher-level cells listed in the table matrix[12]. For example, Black, or African-American in Table P1 and Table P2 could well have different values for the same geography because the value in P1 (cell P0010004) is for all Blacks and the value in P2 (cell P0020006) is for Non-Hispanic Blacks.

[11] One may also consider using the LOGRECNO as a character field with leading zeroes for indexing operations. Also, one may wish to add some other useful fields to the data segment files, e.g., SUMLEV, or POP100, to assist in data review.

[12] See the Appendix for the 2010 and the 2000 Table Matrix sections and an example of the first few cells.

Vertical/Records: Similarly, one must look to the SUMLEV field in conjunction with the NAME field for a record of a geographic area. For example, there are six records in the geographic header file for South Woodbridge CDP, one each for the Summary Levels of 531, 614, 624, 701, 720, and 160; all of which have the same total population. Just as horizontal cells with the same description are only a subset of the appropriate table, vertical records with the same geographic name are only a subset of the appropriate summary level.

Segments/Record Links: Of paramount importance in linking the information for the three data files is to use the fields that are common to all files. For the PL files this is basically just the LOGRECNO field that indicates the record number in conjunction with the CIFSN that indicates the file segment number.

Data Formats: There are two formats of input data files provided so that users can choose the format appropriate for their software:
a) **Fixed Length:** files with the extension of ".txt" in the DATA\ top-level folder: these are fixed-width files (sometimes known as .sdf, for system-default format) [e.g., CAGEO.TXT; CA00001.TXT]; and
b) **Delimited:** data files without any extension in the DATA\ASCII_FILES folder: these are files with a delimiter between each field (sometimes known as .csv, for comma-separated values, although there is no header record with the field names included), [e.g., CA0001_PPL and CA00002_PPL]. The CAGEO_PPL is actually an .sdf file in both versions, though the mode of the CR/LF (carriage return/line feed) is different.

Creating a .dbf file in dBase-compatible software (e.g., Microsoft FoxPro) using the .dbf structure files isn't difficult[13] and the records from the .txt files can be easily appended[14]. The .csv-type files do not have field headers but can also be easily appended into the database: in their raw format the data segment files are only about 25% of the size of the .txt files.

Timing: The law requires that the PL94 data files be delivered within one year of the census, or, before April 1, 2011. During the 2000 census data release, files were released on a flow basis with over half of the states receiving their data before the middle of March.

Use of the Data: Of course, downloading and converting these files will do nothing until one can integrate them into their own data system to undertake analysis of them.

[13] One may still need to revise these structure files to create .dbf files from them because they are not in the standardized format required for the dBase "structure extended" files.

[14] Note that while .dbf and .csv-type files can be easily viewed in some spreadsheet programs, such as Microsoft Excel, there may be a limitation on the number of records that will be viewable, e.g., about 64k.

For many redistricting stakeholders, this will mean merely waiting until the software vendors convert the datasets and merge them into their proprietary databases that they will distribute state-by-state. For others, downloading and converting these datasets the day they are released is not soon enough.

The 2010 files will contain the census counts tabulated for the post-2000 districts for Congress and the State Legislatures in addition to three levels of school districts. These will be the most current and complete counts for these areas of geography and will, generally, reflect the growth patterns in the districts over the preceding decade. In order to review these trends, the user will have to undertake some very basic analysis by the calculation of net growth numbers and percentages calculated on the basis of the 2000 census numbers.

Alas, the census files do not contain any information from any earlier census and even to calculate the very basic population growth for any area means that the researcher must have that previous data available. In many states this isn't very difficult because the districts established after the 2000 census didn't change once established. In others, the districts may have changed a little, or a lot. The key issue here is to have on hand the 2000 census data for the districts in place today. These will not be the PL94 files from 2000 but are more likely to be the 110th Congressional and State Legislative District Summary Files prepared by the Bureau for the plans in effect for the 2006 elections. These can be accessed via the Census Redistricting Data home page: http://www.census.gov/rdo/ or by contacting Cathy McCully.

Calculation of Percentages for Racial Subgroups: There is sometimes confusion about which percentage is which merely based upon whether the denominator is total population or total voting age population. There may also be confusion about the percentages based upon what number is used for the numerator. For example, let us review the data for that portion of San Joaquin County that is in Congressional District 11, which is most of the county, for Asians, who are the predominant minority group.

From Table P1 we see that Asians are 106,851 in number, or 19.1% of the total population, while from Table P2 we see that Non-Hispanic Asians are 103,650 in number, or 18.6% of the total population. From Table P3 we see that Asians of voting age are 73,508 in number, or 19.0% of the total voting age population, while from Table P4 we see that Non-Hispanic Asians are 71,893 in number, or 18.5% of the total voting age population. If we believe the Race/Hispanic Origin principle is the easiest to understand, the appropriate values would be from Tables P2 and P4: Non-Hispanic Asians are 18.6% of the population and 18.5% of the voting age population[15].

[15] See the Appendix for the Full Data Record listing.

As if this isn't confusing enough already, the next option adds more considerations vis-à-vis the numerator. Let us here consider Race/Hispanic Origin as the appropriate mode for the calculation of percentages of voting age population (Table P4).

There are several ways to determine the number of Non-Hispanic Asians[16]. In fact, OMB established some guidelines for the multi-race numbers for the 2000 census as that was the first census for which these numbers were available. The methods mentioned in the March 9, 2000 Bulletin (00-02) reflect the methods listed below in a general sense[17]:
A) **Single Race:** An initial method would be to use the value for those designated as Asian alone: this is the 71,893, or 18.5% of the voting age population, as mentioned in the preceding paragraph.
B) **Two Races/Major Part:** A second method would be to take the Asian alone and only the two-race White; Asian category: this would be 75,516, or 19.5% or the voting age population.
C) **Any Part Race:** A third method would be to take the Asian alone and combine it with any other multi-race category which includes Asian: this would be 78,835, or 20.2% of the voting age population.
D) **Alternative Allocations:** A fourth method was suggested by the OMB bulletin: "If the enforcement action requires assessing disparate impact or discriminatory patterns, analyze the patterns based on alternative allocations to each of the minority groups." This would require case-specific considerations.

As can be seen from the simple example above, the percentage for Asians in this example could range from 18.5% to 20.2% of the voting age population. Moreover, using some methods may result in a sum of the population subgroups that exceeds the total population for the area. The determination of which method to use can be not only a difficult choice but can have real implications in plan drafting and review.

ACS Data Tables: As mentioned above, a release of ACS data is scheduled for December of 2010. Some of the data cells in the PL files can be found in the current structure of the ACS tables.

Table P1 (Race Only for Population) can be found in the ACS with the same 71 cells. Some of Table P2 (Race/Hispanic Origin for Population) can be found as well, though without the multi-race combinations. Tables P3 (Race Only for Voting Age Population) and P4 (Race/Hispanic Origin for Voting Age Population) do not have any direct ACS equivalent. Portions of Table P3 can be calculated, though with some difficulty. But for

[16] See the Appendix for the Alternatives for Population Subgroup Percentages listing (Asians).
[17] See the Appendix for the OMB Bulletin (00-02) Guidance on Aggregation and Allocation of Data on Race for Use in Civil Rights Monitoring and Enforcement

Table P4 information, the only ACS breakout of Hispanic origin by racial subgroup for voting age population is White into Non-Hispanic White and Hispanic.

Moreover, working with the ACS data files is a far more complex operation than joining the three record segments of the PL files. In addition, there are additional considerations in determining which data files and tables to use and the impact of suppression even at high levels of geography.

Summary: The PL94-171 block-level dataset will form the basis of your districting efforts. Regardless of the types of other data one might have available during the redistricting process, every plan will at least be analyzed on the basis of these underlying numbers for population deviation as well as the population numbers of racial subgroups. I hope this discussion has outlined some of the issues that need to be addressed as we all prepare for the release of this information in the first few months of 2011.

Attachments:

1. Census questionnaire, selected pages
2. Table Matrix, 2010 Census PL 94-171 Dataset
3. Table Matrix, 2000 Census PL 94-171 Dataset
4. First Few Cells, 2010 PL Dataset, selected record
5. Summary Level Sequence, 2010 Census
6. Summary Level Sequence, 2000 Census
7. Count of Records, 2010 PL Dataset for San Joaquin
8. Full Data Record, 2010 PL Dataset, selected record
9. Alternatives for Population Subgroup Percentages, Asians
10. OMB Bulletin (00-02) Guides for Aggregation and Allocation, March 9, 2000
11. Table Matrix, 1990 Census PL 94-171 Dataset
12. 2000 PL Summary Tables for NJ

[d:\policomm\ncsl_200906_pl94notes.docx~Tuesday, June 09, 2009]