John Peterson

6/18/17

**Overview and explanation of secondary variable data release to CARE restructuring data collection group.**

Note that the data being passed along now, per request, are “marginal” or simple distributional data – not “crossed” with levels of poverty. I’m pretty unclear on the group’s analysis intentions, but they clearly don’t include an across-the-board individual level analysis of the determinants of energy vulnerability and possible ameliorative impacts of discount rearrangement, as would be possible by marrying PUMS, IPUMS, and individual utility household history data.

1. **Data sets in brief.**

Am passing along data sets at the block group level, the county level, and the county/zip7 level. The utilities/commodities included, by level, are outlined in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
|   |   | **Presence at level …** |   |
| **UTILITY** | **Commity or fuel type** | **CNTY/ZIP7** | **BLKGRP** | **CNTY** |
|   |  |  |  |   |
| ACS1115 | HH |  | Y |   |
| PGE | EG | Y | Y | Y |
| PGE | EO | Y | Y | Y |
| PGE | GO | Y | Y | Y |
| PGE | E | Y | Y | Y |
| PGE | G  | Y | Y | Y |
| PGE | T | Y | Y | Y |
| SCE | E | Y | Y | Y |
| SCG | G | Y | Y | Y |
| SDGE | E | Y | Y | Y |
| SDGE | G | Y | Y | Y |

For a given level, and a particular geography, there may be one or several “utilities” present in separate rows defined by utility and commodity.

[We inadvertently included in the release “SCX” utility records: this was a purely investigational effort in the annual estimates (Feb 2017) that dealt with sensitivity of the estimates to low usage screening (eliminating about 2% of SCE total households for very low usage in the Sept 2016 meter read). Have been scrambling today to recreate all the files without SCX. ]

Note first the “utility” ACS1115, with “fuel type” HH. These records have block group data on American Community Survey 2011-2015 household(HH), as well as total population, residential population, and civilian population. They also include, for comparison, the 2016 counts on primary variables (population, households, renter/owner, etc.) from AGS (the vendor source for “CALIFHH” records in the annual estimates produced each February). AGS is not included in this release as a separate line of estimates, in part because the distribution data mainly come from ACS directly or indirectly.

In some cases, ACS data that were only available at the Census Tract level were disaggregated to the Block Group level (there are 8057 Census Tracts in California, 23212 block groups with population). In these cases, we made the disaggregation as “informed” as was practical. For example, if we wanted a distribution of a disability by age group, and this was only available at the tract level, we would disaggregate to the block group level within specific age groups, thus mitigating/localizing error – for example, in the actual distribution of that disability within the block group.

The other utility/commodity data pertain to actual utilities, and the household universe is Autumn 2016 technical eligibility for ESA (individually metered, sub-metered, and master-metered).

Note the proliferation of PG&E “commodities.” EG (electric+gas served households), EO (electric only from PG&E), GO (gas only) are the implementation and reporting categories PG&E has used historically). E (any electric), G (any gas), and T (total) are produced for convenience in this data release.

1. **Documentation in the files**:
2. File\_Desc includes documentation for the variables found in each data table or file, including line-by-line explanatory comments.
3. Cnty\_names translates Census “FIPS CNTY” (“06” + county number per Census) to either county name or a useful concatenation of FIPS cnty and county name.
4. **Four files per aggregation level (response to “unwieldiness”), and basic layout of the files.**

It would be really great if the committee and its CPUC counterpart could simply accept SAS files, but we settled for the worst case and assumed that Excel probably would be the medium – which is a difficult way to do analysis of large data sets. In another section I describe how data is or can be alternatively packaged to avoid some of the inefficiencies of a single massive Excel file with several tables, and you can think about what you’d like among the alternatives I’ve prepared. This section is about data file layout and structure.

We wanted to make the file easily understandable, and so, among other steps, avoided making individual files very wide with lengthy sets of arrays to deal with – instead providing four files per aggregation level, with a key (variable RECORDID) allowing simple merges as wanted.

The basic layout for each file is available in File\_desc. In general terms, each file begins its records with standard utility/commodity/geographic variables (with geography variables blank if not relevant to the aggregation level), followed by a standard set of estimated ACS/AGS/utility counts: households, three types of population, and owner/renter household counts. Thereafter, each record is segmented into “sets” roughly numbered to correspond to my last email from Mark Alfaro. Each “set” (or set&no) consists of

1. a demarcation character variable named “set&no” that is for visual separation only for Excel users,
2. followed by a character variable indicating what the “base” or “universe” for the estimates in the segment is – e.g. “HHLDS” or “RESPOP”,
3. followed by a count or counts providing the universe value – usually simply taken from the standard set of estimate counts that begin the record,

and then with some exceptions described in a later section,

1. the distributions of the variables in the segment are expressed as percentages of the base. If it were desirable to convert these back to estimated counts, the approach would be to simply apply the array of percentages to the base or universe count immediately preceding.
2. **“Sets” by file (described more fully in File\_Desc).**

File 1:

* Set 5: Income. No percentage estimates as just described in (4) above. Mean, median and other percentiles in $2015. Also percents within the subject area that are given percentages (or below) of the State income median (100, 80, 75, 50, 40).
* Set 5A: Marginals for various percents of poverty, from 2016-17 annual estimates (Feb 2017). No percentage estimates as just described in (4) above. Follows CPUC custom of setting one and two-person limits at the two person limit. Six levels, defining seven possible categorizations of household percent of poverty: 100, 150, 200 (the ESA limit), 250, 300, and 400 percent of poverty.
* Set 6: Household size, household size by tenure.
* Set 8 (no set 7): For population age 16 or more, labor market participation and unemployment.

File 2 (one set only):

* Set 9a: Households. Percentage of household income spent on housing costs (for owners, “selected housing costs” and for renters, gross rent). Overall and by tenure. Within overall and tenure provides percentage of income within specific household income ranges.

File 3:

* Set 9b: Households. For renters and by owners/mortgage status, provide distribution of percent of household income devoted to housing costs.
* Set 10: Households. Income support indicators: social security, SSI, retirement. (Also see annual estimates data already developed on categorical program participation).
* Set 11: Households. Participation in food stamp program (SNAP).
* Set 12A: Households. Householder age distribution.
* Set 12B: Household. Age of dwelling, overall and by tenure.
* Set 13A: Civilian population. Presence of any disability.
* Set 13B: Civilian population. Presence of disability by participation in public/private health insurance.

File 4:

* Set 14: Households. indicator of public assistance.
* Set 15: Civilian population. Indicator of participation in Medicaid (Medi-Cal).
* Set 16 (included for completeness): Households. Long run weather from Energy Plus. No percentage estimates as just described in (4) above.
* Set 17 (included for completeness): Households. Tenure by dwelling type.
1. **Fractional population values (households or population), and total, res, and civ population in utility (rather than ACS) data.**

These may arise in the data due to allocation efforts in attempting to proportionally allocate households in geocoding technical eligibility data from the utilities in the Autumn extract for annual estimates – i.e., for a very small fraction of households, there is not enough accurate address information to geocode reliably to block/zip7 and for these households, correspondence tables are used to share them proportionally among small areas that are reasonable assignments given what is available on the record. Additionally, in connection with ACS data, which ought to be round numbers with respect to household and population values, we have had to solve a Los Angeles County problem involving three block groups which have vanished and then reappeared as legitimate geographies. This entailed some proportioning of ACS geography into geographies matching what we have used all along since 2010 Census for annual estimates, resulting in fractional estimates. Another source of fractional households includes disaggregating some Census Tract level ACS estimates to block group.

**5.1 Ratio approach to population from utility households.**

Finally, note that for utility estimates, the household count is the number of served households for a given commodity (technical eligibility). Obtaining utility estimates of residential population, total population, and civilian population entailed application of ACS ratios of respop/hhlds, totpop/hhlds, civpop/hhlds to the utility technical eligibility count – which of course will generate frequent fraction population estimates for utilities.

We don’t see these as problems, as they are based on efforts to provide unbiased and non-arbitrary estimates. In the annual estimates, we do provide an alternative set of datasets that include numerous rounded/sprinkled estimates (sprinkling under a weighted random plan to ensure consistency between estimates at various geographical levels). You are of course free to round the data for your own purposes.

1. **Difficulties with large Excel files: various “packaging” alternatives.**

We have written out the 12 files (4 files each, for cnty/zip7, block group, and county data), and somewhat laboriously stuffed them into a large 600 megabyte Excel spreadsheet, which is awkward and slow to use (as compared to a SAS file, for example).

As I have to recreate the data sets anyway to eliminate the experimental “SCX” utility estimates referred to earlier (page 1), there is an opportunity to create the files in somewhat more wieldy setups.

I’m going to suggest here two alternative setups or packages to pass along, both of which I’ll prepare tonight.

1. **Big Excel file (less the SCX records):**

Two documentation tables,

4 cnty tables as described,

4 blkgrp tables as described,

4 cnty/zip7 tables as described

OR,

1. **Partial file separation:**

An Excel file including two documentation tables and 4 county tables

Separately: 4 different excel files, one table per xls, at the blk group level

Separately: 4 different excel files, one table per xls, at the blk group level

 It is possible to make other arrangements, including sending SAS files instead!

In emailing this documentation to Brett and Mark, I will include an initial

(**1 S) version of the big excel file,** in which the block group and cnty zip7 tables are limited to the first 100 records.

**(7) Frequencies from SAS file.**

The following tables provide record, household weighted, and various population-weighted counts for cnty/zip7, block group, and county, from the tables released to you.

|  |
| --- |
| **Frequency: CNTY/ZIP7 records: records, households, population (total, civ, res)** |
| util | commod | RECORDS | HHLDS | TOTPOP | RESPOP | CIVPOP |
| PGE | E | 18404 | 4578353 | 20854686 | 20561645 | 20703107 |
| PGE | EG | 14141 | 3545100 | 15891071 | 15677386 | 15779699 |
| PGE | EO | 15371 | 1033253 | 4963615 | 4884259.5 | 4923408 |
| PGE | G | 17377 | 4848702 | 21847804 | 21538229 | 21686830 |
| PGE | GO | 15780 | 1303602 | 5956733 | 5860844 | 5907131 |
| PGE | T | 21572 | 5881955 | 26811419 | 26422489 | 26610238 |
| SCE | E | 18490 | 4509581 | 22815916 | 22498754 | 22602982 |
| SCG | G | 23563 | 6117828 | 30371559 | 29959456 | 30146746 |
| SDGE | E | 4446 | 1222632 | 6505062 | 6308617.7 | 6254718 |
| SDGE | G | 3482 | 849597 | 4485575 | 4435720.1 | 4417662 |

|  |
| --- |
| **Frequency: BLK GROUP records: records, households, population (total, civ, res)** |
| util | commod | RECORDS | HHLDS | TOTPOP | RESPOP | CIVPOP |
| ACS1115 | HH | 23190 | 12717801 | 38388434 | 37572200 | 37880438 |
| PGE | E | 8176 | 4578353 | 20854686 | 20561645 | 20703107 |
| PGE | EG | 7245 | 3545100 | 15891071 | 15677386 | 15779699 |
| PGE | EO | 7778 | 1033253 | 4963615 | 4884259.5 | 4923408 |
| PGE | G | 8750 | 4848702 | 21847804 | 21538229 | 21686830 |
| PGE | GO | 8519 | 1303602 | 5956733 | 5860844 | 5907131 |
| PGE | T | 9659 | 5881955 | 26811419 | 26422489 | 26610238 |
| SCE | E | 7996 | 4509581 | 22815916 | 22498754 | 22602982 |
| SCG | G | 11305 | 6117828 | 30371559 | 29959456 | 30146746 |
| SDGE | E | 1956 | 1222632 | 6505062 | 6308617.7 | 6254718 |
| SDGE | G | 1715 | 849597 | 4485575 | 4435720.1 | 4417662 |

|  |
| --- |
| **Frequency: CNTY records: records, households, population (total, civ, res)** |
| util | commod | RECORDS | HHLDS | TOTPOP | RESPOP | CIVPOP |
| PGE | E | 47 | 4578353 | 20854686 | 20561645 | 20703107 |
| PGE | EG | 37 | 3545100 | 15891071 | 15677386 | 15779699 |
| PGE | EO | 47 | 1033253 | 4963615 | 4884259.5 | 4923408 |
| PGE | G | 38 | 4848702 | 21847804 | 21538229 | 21686830 |
| PGE | GO | 38 | 1303602 | 5956733 | 5860844 | 5907131 |
| PGE | T | 48 | 5881955 | 26811419 | 26422489 | 26610238 |
| SCE | E | 15 | 4509581 | 22815916 | 22498754 | 22602982 |
| SCG | G | 12 | 6117828 | 30371559 | 29959456 | 30146746 |
| SDGE | E | 2 | 1222632 | 6505062 | 6308617.7 | 6254718 |
| SDGE | G | 1 | 849597 | 4485575 | 4435720.1 | 4417662 |