Hunting for Auxiliary Variables in the Census Planning Database Tract File

Clayton Knappenberger, Arcenis Rojas, Lucilla Tan

U.S. Bureau of Labor Statistics*

Background

Higher response rates only reduce nonresponse bias if the added respondents reduce the contrast between respondents and non-respondents. Information which is auxiliary to the survey and available on the complete sample can be used to help measure this contrast with Representivity (R) Indicators.

As part of ongoing efforts to understanding nonresponse bias in the Consumer Expenditure Surveys (CE), we explored a potential source of auxiliary information that could be used to construct R Indicators - the Census Planning Database (PDB).

CE already has paradata on contact attempts and interview methods. We hope that using the PDB as an auxiliary data source will enable use to capture "neighborhood" effects in unit nonresponse.

Goal: To learn how to use the PDB with CE and to determine which variables in the PDB might be most useful in constructing R Indicators for the CE.

The Consumer Expenditure Surveys (CE)

- Data on household (consumer unit) characteristics, expenditures, income, and taxes. See http://www.bls.gov/cex
- Two independent surveys: Quarterly Interview and Diary
  - Quarterly Interview is a multi-wave recall survey covering mainly large and/or recurring expenditures *(This is the CE data source for this study)*
  - Diary covers smaller, day-to-day expenditures over two weekly periods
- Paradata on contact attempts and other interview characteristics
- Internal geographic data exists for each household at the PSU, state, county, and tract levels
- 9,778 Households

The Census Planning Database (PDB)

- Contains housing, demographic, socioeconomic, and census operational data from census and American Community Survey (ACS) databases to be used for survey and census planning. See http://www.census.gov/research/data/planning_database
- Variable names and descriptions can be found at https://www.census.gov/research/data/planning_database/2015/docs/PDB_Tract_2015-07-28a.pdf
- Data exists at the census block group and tract levels
- 74,021 Tract level records

Census Relationship Files

- The Over Time (Comparability) files show relationships for the same type of geography over different periods of time. See http://www.census.gov/geo/maps-data/data/relationship.html
- Contains mapping files that link the 2000 Census geographies to the 2010 Census geographies.
- Relationships can be many to many

Merging the PDB and CE

The key used in this merge is an 11 digit track number made by concatenating STATE, COUNTY, and TRACT codes.

The 2015 PDB contains 5 year ACS estimates covering 2009 - 2013, so the appropriate CE equivalent are the 2013 files. However, the 2015 PDB uses 2010 Census tract numbers while the 2013 CE files continue to use 2000 tract numbers.

The 2010 Census Tract Relationship files are necessary for "translating" 2000 tract numbers in CE into 2010 tract numbers that exist in the PDB. For a summary of the steps taken see Chart 1.

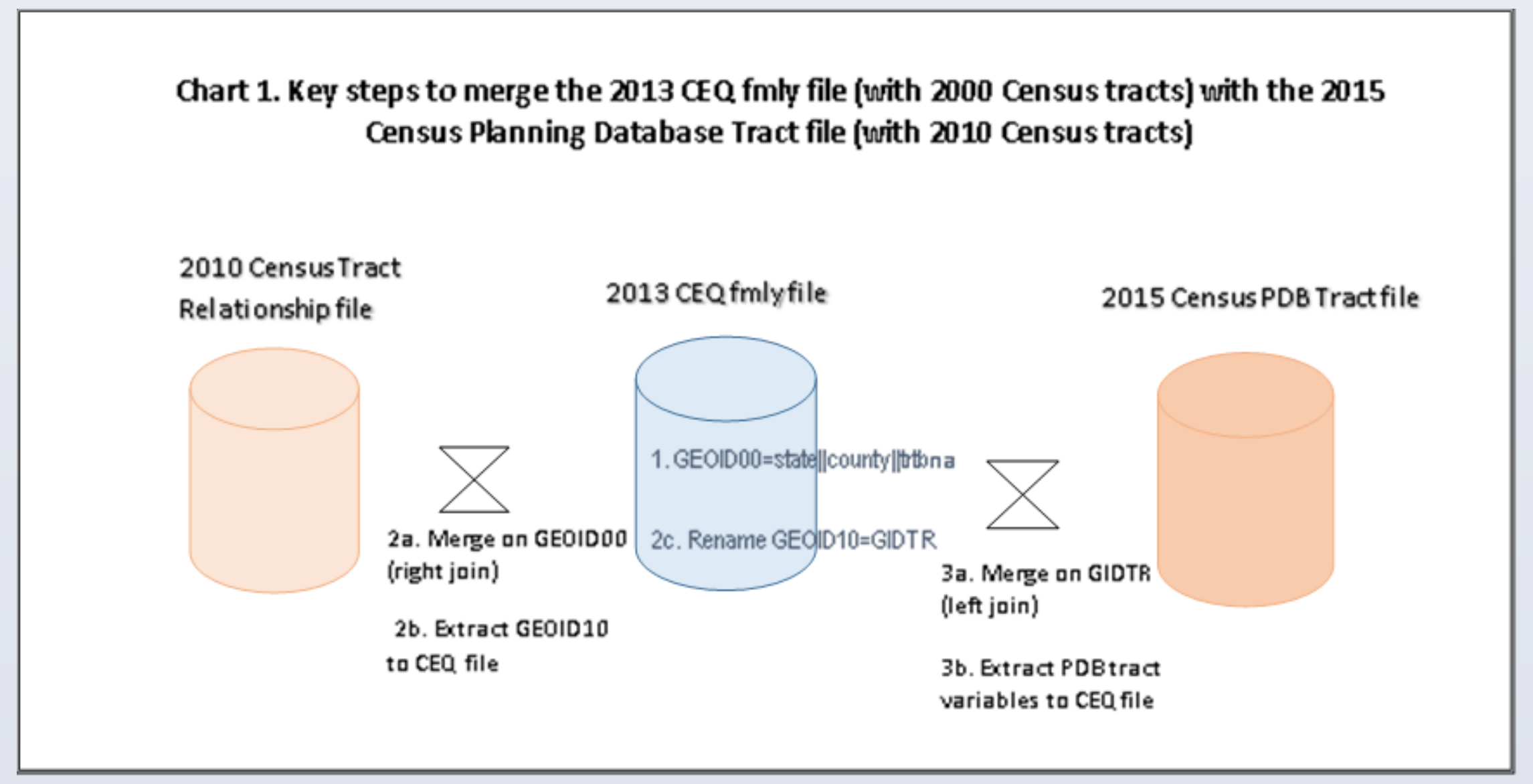


Chart 1. Key steps to merge the 2013 CEQ fmly file (with 2000 Census tracts) with the 2015 Census Planning Database Tract file (with 2010 Census tracts)

2010 Census Tract Relationship file

2013 CEQ fmly file

2015 Census PDB Tract file

1. GEOID00=state||county||trtbna

2a. Merge on GEOID00 (right join)

2b. Extract GEOID10 to CEQ file

2c. Rename GEOID10=GIDTR

3a. Merge on GIDTR (left join)

3b. Extract PDB tract variables to CEQ file

Things to Watch For

1. Almost half of the households in the 2013 CE were assigned 2000 tract numbers that mapped to more than one 2010 census tract. Those households had to be dropped from our study because they could not be reliably assigned PDB values.

Number of 2010 Tracts	Frequency	Percent
Single Tract	4,547	53.17
Multiple Tracts	4,005	46.83

2. Tracts that are in the Relationship file but not in the 2015 PDB. Census occasionally changes tract numbers in between decennial censuses - the updated tract numbers get used in the PDB, but it seems that the Relationship file is not always updated to reflect the most recent geographies. Hence there were six tracts that appeared in CE that could not be initially mapped to 2010 tract numbers that were excluded from our study.

3. Formatting issues in PDB and Relationship files:
   A. Leading zeroes in the PDB and Relationship files.
   B. Dollar signs and commas in specific PDB variables.
   C. Descriptive statistics - variable "Moved from a different household 1 year ago (%)" has max of 118,050

Variable Selection Methods

The first selection step was to extract only the variables which began with "Pct", "Med", or "Avg" - we also extracted the Census Bureau constructed Low_Response_Score variable. Variables that were not considered include margin of error variables and geography variables.

Our selection process made use of three main levels of analysis and four different methods.

1. Univariate Method
   A. Near-zero Variance: variables which have very small variances are unlikely to be useful in discriminating between groups. We identified variables that had greater than zero variance using the nearZeroVar function in the R package caret.
2. Bivariate Methods
   A. The Pearson correlation with Unit Response was calculated for each variable in the PDB and those that were significant at the 5% level were identified.
   B. Group means were computed by Unit Response for each PDB variable. A t-statistic was then calculated comparing the group means for each variable and variables for which the t-statistic was significant at the 5% level were identified.
3. Multivariate Method
   A. Using the DecisionTreeClassifier in Python's Scikit-learn package, we fitted a classification tree for Unit Response of depth 5 on the 4,387 complete cases. Those variables that were selected by the decision tree were identified.

Because the decision tree was the only multivariate method used in this analysis, we used it as the basis for our initial list of PDB variables - see Chart 2.

Programming and Package Notes

Python's Scikit-learn package uses an optimized version of the CART algorithm for classification trees and offers users the option to use either the Gini impurity or the information gain for measuring a split's quality. Our analysis uses the Gini impurity. One downside is that Scikit-learn does not provide a native Python means of viewing the tree and requires the user to install the Graphviz tool.

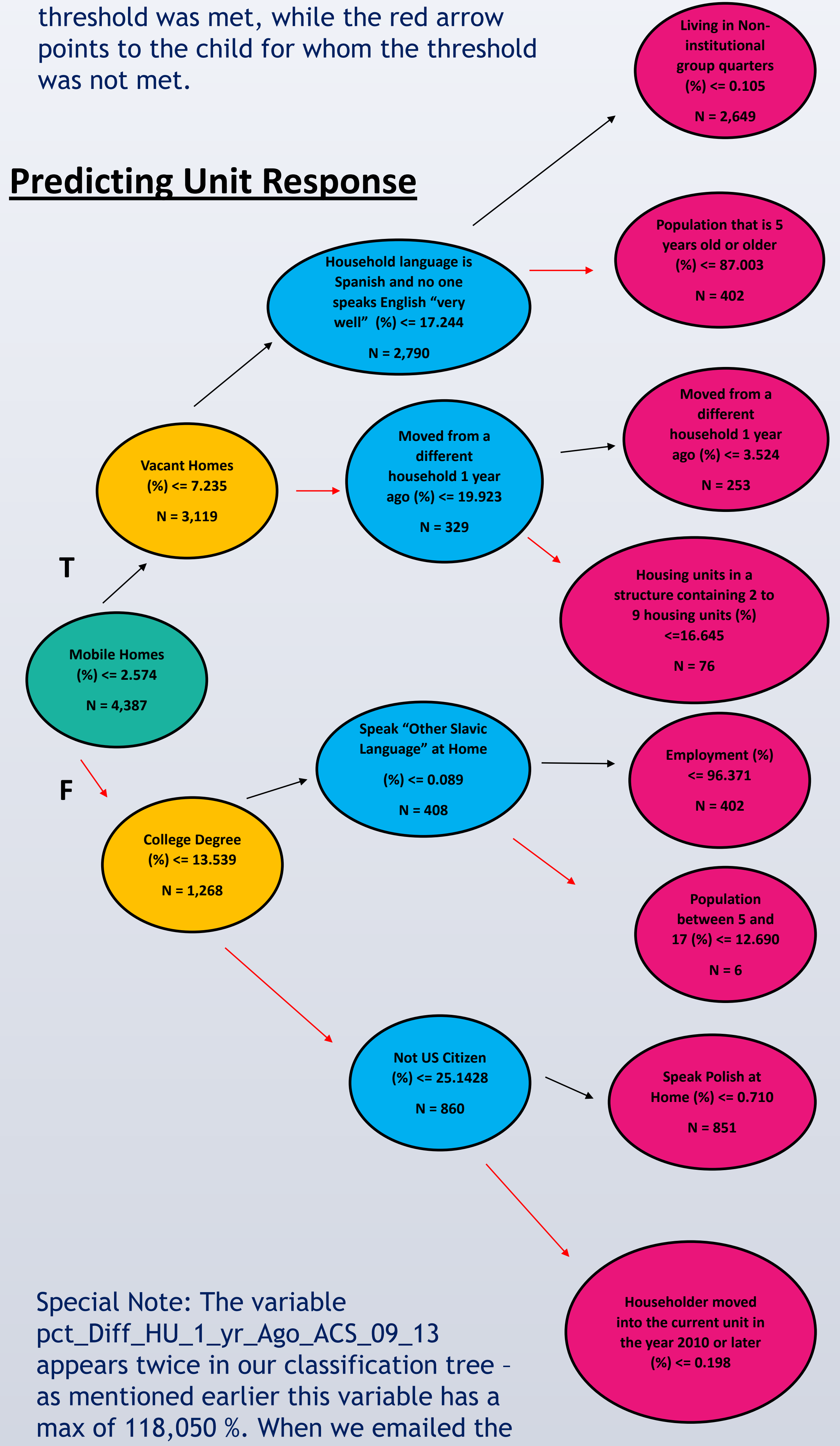
R's rpart package provides a similar level of sophistication to what is available in Scikit-learn, and allows the user to view the tree natively in R's plotting functions.

SAS Enterprise Miner provides a means of performing classification tree analysis, but this capability does not appear to be available in SAS outside of Enterprise Miner.

Chart 2: Truncated Classification Tree Results

For each split in a node, the black arrow points to the child for whom the parent's threshold was met, while the red arrow points to the child for whom the threshold was not met.

Predicting Unit Response



Mobile Homes (%) <= 2.574 N = 4,387

T

F

Vacant Homes (%) <= 7.235 N = 3,119

College Degree (%) <= 13.539 N = 1,268

Household language is Spanish and no one speaks English "very well" (%) <= 17.244 N = 2,790

Moved from a different household 1 year ago (%) <= 19.923 N = 329

Speak "Other Slavic Language" at Home (%) <= 0.089 N = 408

Not US Citizen (%) <= 25.1428 N = 860

Living in Non-institutional group quarters (%) <= 0.105 N = 2,649

Population that is 5 years old or older (%) <= 87.003 N = 402

Moved from a different household 1 year ago (%) <= 3.524 N = 253

Housing units in a structure containing 2 to 9 housing units (%) <=16.645 N = 76

Employment (%) <= 96.371 N = 402

Population between 5 and 17 (%) <= 12.690 N = 6

Speak Polish at Home (%) <= 0.710 N = 851

Householder moved into the current unit in the year 2010 or later (%) <= 0.198

Special Note: The variable pct_Diff_HU_1_yr_Ago_ACS_09_13 appears twice in our classification tree - as mentioned earlier this variable has a max of 118,050 %. When we emailed the PDB team at Census to ask them about this in January, they confirmed that it was an error and said that they would post a user note on the PDB website.

Initial List of PDB Variables

Initial list of PDB variables associated with CE Unit response status by variable selection method

PDB variables (sorted alphabetically)	Variable Selection Method			
	Greater than zero-variance	Pearson Correlation	Group mean T-Test	Classification tree
avg_Tot_Prns_in_HHD_ACS_09_13 The average number of persons per ACS occupied housing unit.	✓			✓
pct_Age5p_OthSlavic_ACS_09_13 The percentage of the ACS population ages 5 years and over who speak English less than "very well" and speak some other Slavic language at home. Examples include Czech, Slovak, and Ukranian		✓	✓	✓
pct_Age5p_Polish_ACS_09_13 The percentage of the ACS population ages 5 years and over who speak English less than "very well" and speak Polish at home.				✓
pct_Civ_emp_16p_ACS_09_13 The percentage of ACS civilians ages 16 years and over in the labor force that are employed	✓			✓
pct_Civ_emp_65p_ACS_09_13 The percentage of ACS civilians ages 65 and over in the labor force that are employed	✓			✓
pct_College_ACS_09_13 The percentage of the ACS population aged 25 years and over that have a college degree or higher				✓
pct_Diff_HU_1yr_Ago_ACS_09_13 The percentage of the ACS population aged 1 year and over that moved from another residence in the U.S. or Puerto Rico within the last year				✓
pct_FRST_FRMS_CEN_2010 The percentage of all addresses in a 2010 Census mailback area for which the first Mailout/Mailback form mailed was completed and returned	✓			✓
pct_Females_ACS_09_13 The percentage of the ACS population that is female	✓			✓
pct_Females_CEN_2010 The percentage of the 2010 Census population that is female	✓			✓
pct_HHD_Moved_in_ACS_09_13 The percentage of all ACS occupied housing units where the householder moved into the current unit in the year 2010 or later	✓	✓	✓	✓
pct_MLT_U2_9_STRC_ACS_09_13 The percentage of all ACS housing units that are in a structure that contains two to nine housing units	✓	✓	✓	✓
pct_Mobile_Homes_ACS_09_13 The percentage of all ACS housing units that are considered mobile homes	✓	✓	✓	✓
pct_NH_Asian_alone_CEN_2010 The percentage of the 2010 Census population that indicate no Hispanic origin and their only race as "Asian Indian", "Chinese", "Filipino", "Korean", "Japanese", "Vietnamese", or "Other Asian"	✓			✓
pct_NON_US_Cit_ACS_09_13 The percentage of the ACS population who are not citizens of the United States	✓			✓
pct_Non_Inst_GQ_CEN_2010 The percentage of the 2010 Census population who live in group quarters and are primarily eligible, able, or likely to participate in labor force while residents.				✓
pct_Pop_5yrs_Over_ACS_09_13 The percentage of the ACS population who are ages 5 years and over at time of interview	✓			✓
pct_Renter_Occp_HU_ACS_09_13 The percentage of ACS occupied housing units that are not owner occupied, whether they are rented or occupied without payment of rent	✓			✓
pct_Tot_Occp_Units_CEN_2010 The percentage of all 2010 Census housing units that are classified as the usual place of residence of the individual or group living in it	✓	✓	✓	✓
pct_Vacant_CEN_2010 The percentage of all 2010 Census housing units where no one is living regularly at the time of interview; units occupied at the time of interview entirely by persons who are staying two months or less and who have a more permanent residence elsewhere are classified as vacant				✓
Note: *** Significantly different from 0 at less than 5 percent level**				

Contact

Clayton Knappenberger, Economist, Branch of Production and Control, Consumer Expenditure Surveys, U.S. Bureau of Labor Statistics

Email: Knappenberger.clayton@bls.gov

Telephone: 202-691-6236

* DISCLAIMER: The opinions expressed in this paper are those of the authors and do not represent official policy of the Bureau of Labor Statistics.