RTI Project Number
0212177.004.002

U.S. Synthesized Population 2005–2009 Version 2.0

Quick Start Guide

August 2012

Prepared by

William D. Wheaton
wdw@rti.org
RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709

This work was supported by the Models of Infectious Disease Agency Study (MIDAS) from the National Institute of General Medical Sciences (NIGMS), grant number U24GM087704. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIGMS or the National Institutes of Health.

Table of Contents

Chapter **Page**

Appendixes

List of Tables

Number **Page**

Overview and Introduction

RTI has developed the 2005–2009 U.S. Synthesized Population dataset in two phases[1]Each phase resulted in a new version of the dataset. The phases and version numbers are as follows:

- Version 1: Generic synthesized households with no group quarters, school assignments, or workplace assignments
- Version 2: Addition of group quarters facilities, school assignments and workplace assignments and inclusion of a new table containing a record for each workplace and its location.

Version 1 was completed in January 2012. Version 2 is now available.

The data are distributed in a compressed file containing a series of ASCII files with comma-separated value. This document refers specifically to Version 2 of the 2005–2009 U.S. Synthesized Population dataset.

The 2005-2009 U.S. Synthesized Population Ver. 1 data has been removed from the download site. If you need to download any of these older "Ver. 1" data, please contact Bill Wheaton (wdw@rti.org).

Downloading and Processing

The 2005–2009 U.S. Synthesized Population data are available for download by state or by county from https://www.epimodels.org/midas/pubsyntdata1.do. Users can download any combination of states or counties.

The names of the ASCII files provided with each extract follow a naming convention that identifies the version and contents of the data. Each ASCII file in a particular distribution contains a prefix consisting of:

- American Community Survey (ACS) year range (e.g., "`2005-2009`")
- Synthesized population version number (e.g., "`ver2`")

[1] Originally, we planned on four phases. Based on timing of the completion of various production processes, it was decided to simplify to two phases. Original Phases 2-4 are now all being combined and released as Phase 2.

- Geographic identifier (e.g., FIPS state code for entire state extracts; FIPS state and county codes for county extracts)

For example, extract files for Version 2 of the 2005–2009 data have the following naming convention:

- **`2005-2009_ver2_01_synth_households.txt`** (for an extract of state FIPS "01," which is Alabama)
- **`2005-2009_ver2_01005_synth_households.txt`** (for an extract of state FIPS "01," county FIPS "005," which is Barbour County, Alabama)

If you wish to combine several county extracts into a single dataset prior to loading into your database or model, then simply remove the header line from each file, and then concatenate the records.

If you need help building a specific multi-county or multi-state study area dataset, please contact Bill Wheaton (wdw@rti.org).

Data Identification and Metadata

Because different versions of the dataset have different contents, the metadata file that accompanies each set of ASCII files extracted for end user delivery is used to help identify and track versions.

The metadata file (e.g., `2005_2009_ver2_[fips]_metadata.txt`) is an ASCII file that contains essential information about the exact contents, source, and version of any particular synthetic population download, including information on the version number, data sources, and files in the distribution of each dataset. A complete description of the contents of the metadata file can be found in the data dictionary in Appendix C.

Citing the U.S. Synthetic Population Database

RTI and its funding agency, the National Institutes of General Medical Sciences (NIGMS) request that you cite the U.S. Synthetic Population database in any publications or journal articles in which the data were used. The correct citation for the data is:

> Wheaton, W.D. (August, 2012). 2005-2009 U.S. Synthetic Population Ver. 2. RTI International. Retrieved from https://www.epimodels.org/midas/Rpubsyntdata1.do.

This Quick Start guide should be cited as:

Wheaton, W.D. 2012. "U.S. Synthetic Population Database 2005-2009: Quick Start Guide". RTI International. Retrieved from https://www.epimodels.org/midasdocs/SynthPop/2005-2009_synth_pop_ver2_quickstart_20120801.pdf

Data Sources

The following data sources were used to compile the information in the synthesized population data.

- **2005–2009 Public Use Microdata Sample:** The Public Use Microdata Sample (PUMS) files are generated from responses to the ACS and include most of the variables that are included in the survey. The smallest geographic unit for which the PUMS data are collected is the Public Use Microdata Area (PUMA). These PUMAs are defined for each decennial census and are based on minimum population thresholds of 100,000 people. This research used the 5% sample PUMA data, which reflect 5% of actual household responses used to create the dataset. This method ensures the confidentiality of respondents.
  - **Download:** The 2005–2009 PUMS data were downloaded from http://factfinder.census.gov/home/en/acs_pums_2009_5yr.html.
- **U.S. Census Bureau Topologically Integrated Geographic Encoding and Referencing (TIGER) Data Block Group Boundaries:** The TIGER 2010 version of Census 2000 block group boundaries includes all 50 states and the District of Columbia but not Puerto Rico. Water features (lakes, wide rivers, coastal water, etc.) within the block groups were removed, along with block groups that were entirely on water. Other block groups were either modified or replaced, resulting in nationwide block group data that match the ACS coding.
  - **Download:** U.S. Census Bureau 2010 Census Redistricting (P.L. 94-171) TIGER/Line Shapefiles were downloaded from the following FTP site: ftp://ftp2.census.gov/geo/pvs/tiger2010st/.
- **2005–2009 American Community Survey (ACS):** The ACS data were collected over 60 months, between January 2005 and December 2009. These are the first published ACS data that include all geographic areas, a feature made possible through the use of a sample size large enough to include estimates for smaller geographies—unlike the previous 1- and 3-year estimates. The values represent the average characteristics over the 5-year period.
  - **Download:** The 2005–2009 5-year summary files were downloaded from http://www2.census.gov/acs2009_5yr/summaryfile/2005-2009_ACSSF_All_In_2_Giant_Files(Experienced-Users-Only)/.

- **2008 LandScan USA:** This source provides gridded population data for the United States. These data, developed by Oak Ridge National Laboratory, are available at http://www.ornl.gov/sci/landscan.

- **HSIP Freedom 2011:** This data source provided some locations for nursing homes, universities, prisons, and military bases. Information on HSIP Freedom is available at: https://docs.google.com/file/d/0B3UROkg-qGTaMkpPU3FiaGNUVy1rUDRSSzJDWUVJUQ/edit?pli=1

- **2010 Census SF1:** Data on age and gender distributions in group quarters was provided by the 2010 Census SF1 files.

Data Files Contained in Each Synthesized Dataset

Synthesized datasets are provided to the user community in subsets defined by geographic area (e.g., a county, set of counties, state, or set of states). Each synthesized dataset contains several individual ASCII text files that, together, provide all the synthesized data for a particular geographic area. The individual ASCII text files are detailed in Table 1.

Table 1. ASCII Text Files Included with Each Synthesized Dataset

File	Contents
`[prefix]_metadata.txt`	Contains metadata on the contents of the extract synthetic population data.
`[prefix]_synth_households.txt`	Contains the location and descriptive attributes for each household. Household records in the **synth_households.txt** file link to individual person records in the **synth_people.txt** table.
`[prefix]_synth_people.txt`	Contains a record for each person, along with his or her age, race, and sex. These synthetic person records link to the **synth_households.txt** file (via the **hh_id** field) and/or to the U.S. Census Public Use Microdata Sample (PUMS) attributes from **pums_p.txt** (via the **serialno** field).
`[prefix]_schools.txt`	Contains locations and descriptive attributes of each public and private school. The **school_id**s link to the **school_id** variable in the **synth_people.txt** table.
`[prefix]_workplaces.txt`	Contains locations and sizes of each workplace. The **workplace_id** links to the **workplace_id** variable in the **synth_people.txt** table.
`[prefix]_synth_gq.txt`	Contains locations of group quarters and counts of individuals in each one by group quarters type.
`[prefix]_synth_gq_people.txt`	Contains age and sex characteristics and link to group quarters type for each group quarters resident.
`[prefix]_pums_h.txt`	Contains complete PUMS household records from the original PUMS 5% data. Links to the **[prefix]_synth_households.txt** file via the **serialno** field.
`[prefix]_pums_p.txt`	Contains complete PUMS person records from the original PUMS 5% data. Links to the **[prefix]_synth_persons.txt** file the **serialno** field.

File	Contents
[prefix]_age_compare.txt	Contains data on the expected count of households and the actual count of households for each block group and each of the seven age categories (see Appendix B for codes). This file also includes ACS margin-of-error ranges for each category and is useful for assessing the accuracy of the synthetic population data as compared to the ACS source data.
[prefix]_race_compare.txt	Contains data on the expected count of households and the actual count of households for each block group and each of the five race categories (see Appendix B for codes). This file also includes ACS margin-of-error ranges for each category and is useful for assessing the accuracy of the synthetic population data as compared to the ACS source data.
[prefix]_income_compare.txt	Contains data on the expected count of households and the actual count of households for each block group and each of the seven income categories (see Appendix B for codes). This file also includes ACS margin-of-error ranges for each category and is useful for assessing the accuracy of the synthetic population data as compared to the ACS source data.
[prefix]_size_compare.txt	Contains data on the expected count of households and the actual count of households for each block group and each of the seven size categories (see Appendix B for codes). This file also includes ACS margin-of-error ranges for each category and is useful for assessing the accuracy of the synthetic population data as compared to the ACS source data.
[prefix]_summary_compare.txt	Contains summary information on the expected and final matches for each of the four matching variables and a summary value, by block group, that provides an overall measure of how closely the synthesized households for a block group match the expectations of the ACS data.

Generating Synthesized Households

To generate synthesized households, RTI used a method developed at the Los Alamos National Laboratory for use with the TranSims transportation simulation package. This method selects households from the PUMS data (the 5% sample) to fit marginal distributions of various aggregated census counts by census block group. The statistical method, called Iterative Proportional Fitting, results in household records from the PUMS 5% sample being selected and replicated so that a complete 100% household dataset is derived for each census block group. A complete description of the TranSims population generator algorithms can be found in an article by Beckman, Baggerly, and McKay (1996).

Four matching variables are used to select households from the PUMS data to match aggregated counts at the block group level. The synthetic population generator attempts to select households from the PUMS data so that the count of households in each of four categories (i.e., age of the head of household, household income, household size, and race of head of households), in each block group, equal the count of households for these same categories that are estimated in the ACS data.

These four selection variables and the detailed categories used for matching to ACS within each variable are shown in Tables 2 to 5.

Table 2. Household Income Categories

Synthetic Population Category	Range	ACS Source Fields (sequence 53)
1	<$10,000	B19001_002
2	$10,000–$15,000	B19001_003
3	$15,001–$25,000	B19001_004 + b19001_005
4	$25,001-$35,000	B19001_006 + b19001_007
5	$35,001–$50,000	B19001_008 + b19001_009 + b19001_010
6	$50,001–$100,000	B19001_011 + b19001_012 + b19001_013
7	>$100,000	B19001_014 + b19001_015 + b19001_016 + b19001_017

Table 3. Head-of-Household Age Categories

Synthetic Population Category	Range	ACS Source Fields (sequence 96)
1	15–24	B25007_003 + b25007_013
2	25–34	B25007_004 + b25007_014
3	35–44	B25007_005 + b25007_15
4	45–54	B25007_006 + b25007_016
5	55–64	B25007_007 + b25007_017 + b25007_008 + b25007_018
6	65–74	B25007_009 + b25007_019
7	>74	B25007_010 + b25007_011 + b25007_020 + b2500_021

Table 4. Household Size Categories

Synthetic Population Category	Range	ACS Source Fields (sequence 33)
1	one-person household	B11016_010
2	two-person household	B11016_003 + b11016_011
3	three-person household	B11016_004 + b11016_012
4	four-person household	B11016_005 + b11016_013
5	five-person household	B11016_006 + b11016_014
6	six-person household	B11016_007 + b11016_015
7	seven or more person household	B11016_008 + b11016_016

Table 5. Head-of-Household Race Categories

Synthetic Population Category	Values	ACS Source Fields (sequence 33)
1	White alone	B11001a_001
2	Black or African American alone	B11001b_001
3	Asian alone	B11001d_001
4	Other	B11001c_001 + b11001e_001 + b11001f_001
5	Two or more races	B11001q_001

Group Quarters and Group Quarters Residents

The group quarters data, which were not part of Version 1 of the data, are now populated in Version 2 of the 2005–2009 U.S. Synthesized Household dataset.

People who reside in group quarters (e.g., nursing homes, prisons, military barracks, college dormitories) accounted for 2.7% of the U.S. population in the 2005–2009 ACS. Because of their close living situations and frequent contact, residents of group quarters may be disproportionately important to infectious disease modeling.

Because the generic population generator provided by TranSims does not produce synthesized group quarters residents, RTI developed modules to generate locations for group quarters and synthesize persons who live in them.

Due to differences in how these group quarters are generated and because the synthesized group quarters residents do not exist in the PUMS data, these entities are provided in two separate files (i.e., the `[prefix]_synth_gq.txt` file and the `[prefix]_synth_gq_people.txt` file) instead of being incorporated directly into the household file and the persons file.

Group Quarters facility locations are derived first by using existing sources of locations from the HSIP Freedom database. Additional facilities are created (at block group centroids) in block groups when SF1 data indicates that there are group quarters residents in places where HSIP Freedom does not indicate group quarters facilities exist OR when group quarters sub-populations exist that logically would be housed in different facilities (for example, presence of juvenile prisoners and adult prisoners in a block group having only a single prison).

The following table provides a summary of the counts of group quarters facilities and of the count of synthetic persons created for them.

Type	Number of Facilities	Number of People
Nursing Home	23,760	1,502,264
Prison	19,786	2,429,326
Military Base	307	337,529
University	4,559	2,523,971

After group quarters facilities are selected or generated, census SF1 data on counts of group quarters residents by type of facility, age groups, and gender are used to create the synthetic residents housed in each facility. More specific age distribution data noted below were used to supplement SF1. Data sources used to generate the group quarters data include:

Type	Source for Age Distributions	Source for Facility Size
Nursing Homes	CDC NCHIS Demographics	CDC Census of Nursing Home Statistics 2010
Prisons	DOJ Bureau of Justice Statistics	HSIP Freedom Data
Military Bases	Dept of Defense Selected Manpower Statistics Fiscal Year 2005 (most recent)	??
Universities	American Community Survey PUMS data aggregated at the national level	HSIP Freedom Data.

The end result of the Group Quarters data development process is two files; one containing a list of facility locations, types, and capacities; the other containing a list of residents by age and gender for each facility.

School Assignments

Synthetic persons who, according to PUMS attend primary or secondary schools, are assigned to actual schools based on school/grade capacity.

The basic assignment methodology is to process each synthetic person age 18 or less by examining his or her PUMS school enrollment code (SCH) and school grade level attending (SCHG) and, for those who attend school, assign each one to the closest school that services that grade level. Since the schools database being used contains both public and private schools and the synthetic persons have coding to determine which students attend public or private schools, the assignments for these two types of schools are handled independently.

If the student goes to a private school, then he or she is enrolled in the closest private school less than 50 kilometers (approximately 31 miles) away that has capacity for the student's given grade category. If there are no private schools within 50 kilometers that have capacity, then the student is assigned to the closest private school servicing the students' grade category (even if already full).

For students attending a public school, the assignment method, is as follows:

- Find up to three schools (one each), within 50 kilometers of the student that also meets the following criteria

- Find the closest non-charter, non-magnet school in the same county as the student with capacity for the appropriate grade
- Find the closest charter school in the same state as the student with capacity for the appropriate grade
- Find the closest magnet school in the same county as the student with capacity for the appropriate grade

- Assign the student to the school (from the set found above) that has the smallest ratio of currently enrolled students to grade capacity. In other words, try to fill schools that are less full first.
- If no schools meeting the above criteria are found, then use the same criteria above except include schools that are already at capacity
  - For those schools found using this criteria, assign the student to the school that is least 'overfilled' based on enrolled/capacity ratio
- If no schools exist in the county (within 50 kilometers), then relax the criteria to include any schools in the state that are within 50 kilometers and repeat all the above steps.

School assignment data sources include:

- School locations: HSIP Freedom 2011.
- Enrollment data: National Center for Educational Statistics (NCES)
- School attendance status and grade: PUMS SCH and SCHG variables. The SCH variable contains data on school enrollment. SCH codes are:

Code	Description
B	N/A (less than 3 years old)
1	No, has not attended in the last 3 months
2	Yes, public school or public college
3	Yes, private school or private college

The SCHG variable contains data on school grade level for those attending school. The SCHG codes are:

Code	Description
B	N/A (not attending school)
1	Nursery school/preschool
2	Kindergarten
3	Grade 1 to 4
4	Grade 5 to 8
5	Grade 9 to 12
6	College undergraduate
7	Graduate or professional school

Workplace Assignments

Synthetic persons who are in the age range for the U.S. workforce are assigned to synthesized workplaces based on workplace sizes and locations.

The basic methodology is comprised of three steps. The first step is to join to the extended PUMS person data to each synthetic person and access the PUMS field "ESR" (Employment Status Recode) to identify only those persons who are either employed in the civilian or military workforce. The second step is to use the calibrated 2000 STP64 table that contains the number of workers living in one census tract, and working in the same or another census tract. This selection provides the total number of workers working in all the tracts associated with a given tract of residence. Based on this, we calculate the tract-of-residence to tract-of-work probabilities. The third step is to assign individual workers to workplaces based on tract-of-work selected and business size.

No attempt was made to match the synthetic persons to workplaces based on occupation or industry. Synthetic persons are assigned to workplaces solely based on commuting patterns and business size (number of employees), not on occupation or industry.

Data sources used in the workplace assignment process include:

- InfoUSA businesses as provided by ESRI Business Analyst in 2005 and 2010
- Synthetic persons and PUMS person table

- STP64 containing estimates of workers counted by census tract of residence and census tract of workplace for 2000.
- U.S. census tract boundaries for 2000 and 2009.

Since the U.S. Census Bureau has not updated the STP64 commuting data file since 2000, RTI calibrated the number of persons living in one tract and working in the same or another tract using the percent difference in employee counts by census tract as reported by InfoUSA between 2005 and 2010.

The basic algorithm to define the tract-of-residence to tract-of-work probabilities took the following steps:

- by tract-of-work, sum the number of employees in 2005, and 2010.
- calculate the percent difference between 2005 and 2010 employee sums
- multiply this percent difference by the number of workers working in any given tract.
- sum the new adjusted number of workers by tract-of-residence.
- join this sum of adjusted number of workers by tract-of-residence.
- calculate the new probability by dividing the new adjusted number of workers by the sum of the new adjusted number of workers.

The result of this process is a probability, for each census tract, that a worker residing in that tract works in any given workplace in the tract-of-work.

The InfoUSA data contains the location and size (number of employees) of over 14,000,000 businesses in the United States. Once the tract-of-residence to tract-of-work probabilities were generated, then the process of determining which persons to assign to which business was performed. The basic operation involved, for each working synthetic person, determining which tract he or she should work in (based on the tract-of-residence to tract-of work probabilities) and then assigning the person to a specific workplace in the specified census tract. The workers are allocated to businesses based on the ratio of how many workers have already been assigned to the total size of the business—the worker/capacity ratio. In effect, each worker is assigned to the business that has the lowest worker/capacity ratio in the assigned tract. Each time a worker is assigned to a workplace, the worker/capacity ratio grows larger for that business. The effect of this is that larger businesses will tend to get a higher priority for worker assignments than small businesses. For example, a business with 10 workers, after having been assigned a single worker, will have a worker/capacity ration of .1 whereas a business with 2 workers, after having been assigned a single worker will have a worker/capacity ratio of .5. Therefore, the 10-worker business will get priority for additional workers until its worker/capacity ratio is above .5.

If there are more workers to be assigned to a tract than there are total worker capacity of businesses, then the algorithm continues assigning workers (this time choosing the workplaces that are the least overfilled) until all workers have been assigned somewhere.

Note: InfoUSA business data records with a count of zero employees were not used in the assignment process.

Household Spatial Distributions

RTI developed a process of placing each synthesized household at appropriate locations across the landscape to ensure that counts of synthetic persons within a census block group matched the aggregated census counts of persons from the ACS and that the distribution of households and people reflected the best, highest precision population distribution available nationwide.

The LandScan dataset was used as the "gold standard" for population distribution in the United States. (For more information about the LandScan dataset, please see http://www.ornl.gov/sci/landscan.) LandScan data are calibrated so that the populations of 90-meter gridded cells, when summarized at the block group level, match census counts. Use of the LandScan population data results in a distribution of households that better reflects the actual distribution of a population than would be possible by simply placing synthesized households randomly within each block group.[2]

The placement method selects all of the synthetic households that are defined for a block group and distributes those households within the LandScan 90-meter gridded cells so that the total count of synthetic persons matches the population within each 90-meter gridded cell. A post-processing method was then used to distribute the households within the 90-meter gridded cells to which they were assigned.

Data Quality Measurements

Each synthesized population dataset is delivered with a set of comparison tables that provide detailed information on the expected counts of households (based on ACS aggregated data) against the actual synthesized household counts generated by the population synthesizer. The comparison tables also contain the margin of error for each variable provided in the ACS. These comparison tables enable users to delve into measurements of how well the synthesized population household counts match expectations of the ACS data for each census block group.

There are five comparison tables: one for each of the synthesized population selection variables (`[prefix]_age_compare.txt`, `[prefix]_size_compare.txt`, `[prefix]_race_compare.txt`, `[prefix]_income_compare.txt`) and a summary comparison table (`[prefix]_summary_compare.txt`), which contains an overall measure of accuracy for each variable and the summed accuracy for all variables.

[2] The first version of the Synthesized Human Population database placed households randomly within each Census block group.

Each of the four variable comparison tables follows the same structure, containing the following fields:

- `Stcotrbg`: state, county, tract, and block group ID
- `Acs_1`: count of expected households from the ACS data for category one
- `Sp_1`: count of households generated by the synthetic population generator for category one
- `Diff_1`: difference between Sp_1 and Acs_1
- `w_diff_1`: weighted difference between Sp_1 and Acs_1. The weight is the count of ACS households in the category for the blockgroup divided by the total ACS households in the blockgroup. The weighted difference (w_diff_1) is the count of synthetic population households for the blockgroup times the weight.
- `Moe_1`: margin of error in Acs_1
- `Out_moe_1`: whether or not the Sp_1 value is outside the margin of error

For example, the `[prefix]_age_compare.txt` table would have seven sets of these Sp_x, Acs_x, Diff_x, w_diff_x, Moe_x, and Out_moe_x variables—one set for each of the seven age categories (see Appendix A) used in the IPF procedure.

The `[prefix]_summary_compare.txt` table contains an overall accuracy measure for each of the four selection variables (age, race, income, and size) and the summed total of all these weighted differences for an overall measure of the accuracy of the synthetic population households as compared to the ACS data.

To calculate the summary weighted difference for each block group, the following calculation is used:

$$a = \sum_{i=1}^{n} w_i d_i$$

where a is the weighted difference across all categories for a variable (age, size, race, or income); w_i is the weight for category i (defined as the count of ACS households in the category divided by the total ACS households in the blockgroup); d_i is the absolute value of the difference between the ACS count against the synthetic population count of households for category i. The weighted difference for each category is summed up to create the overall weighted difference a for the variable. So, for the income, age, and size variables, $n = 7$ because there are seven categories (see Tables 2 to 4), and for the race variable $n = 5$ because there are five categories (see Table 5).

The overall accuracy measure (across all variables and categories) for each block group is calculated by summing the weighted difference value (above) for the four variables as follows:

- Overall_accuracy = block_group_age_weighted_difference + block_group_race_weighted_difference + block_group_income_weighted_difference + block_group_income_weighted_difference

The margin-of-error values for each category are derived from the ACS "m" tables. Margin of error for a category (e.g., income < $10,000) that is from a single ACS variable (e.g., b19001_002) is equal to the margin of error value for that same variable (e.g., b19001_002) in the "m" table. For categories that are derived from more than one ACS variable (e.g., annual income from $15,001 to $25,000, which is the sum of variables b19001_004 and b19001_005), the margin of error is equal to the square root of the sum of the squares of the individual margins of error.

The out_moe_x variables contain a "y" if the count of households for this variable in this block group is outside of the margin of error. Otherwise, the block group is coded with an "n" to denote that it is not outside the margin of error.

Latitude/Longitude Coordinate System

If you are loading these data into a GIS, then it is important to specify the appropriate projection for the resulting GIS dataset. The coordinate system for these latitude/longitude coordinates is the World Geodetic System of 1984.

Data Relationships

- **synth_households.txt** links to original **pums_h.txt** file via **serialno** in a many-to-one relationship.
- **synth_people.txt** links to **synth_households.txt** via **hh_id** in a many-to-one relationship.
- **synth_people.txt** links to **pums_p.txt** via **serialno** and **sporder**. The **serialno** identifies a particular household in the PUMS and the **sporder** identifies each person (as a sequence from 1 to *n*) in each household. Both **serialno** and **sporder** must match when linking **synth_people.txt** to **pums_p.txt.**
- **synth_people.txt** links to **schools.txt** via **school_id** in a many-to-one relationship.
- **synth_gq_people.txt** links to **synth_gq.txt** via the **gq_id** in a many-to-one relationship.

Important Notes

- Synthetic households and persons derived from PUMS are published by U.S. Census Bureau.
- A method for constructing the synthetic households and people was developed at the Los Alamos National Laboratory for use with the TranSims transportation simulator software. The original TranSims program code was released under an open source license. (Information about continued development of TranSims is available at http://code.google.com/p/transims/) The TranSims population generator, which is a component of the TranSims transportation simulator software, used four household attributes (i.e., age of the head of household, household income, household size, and race of head of household) to construct the synthetic households and people. When synthetic households are aggregated to a block group, census tract, and county, counts for these four attributes should closely match the totals for those census geographies in the ACS tables.
- No person-level attributes are used to construct synthetic households or synthetic people; therefore, aggregated counts of synthetic people by age or sex (for example) may not closely match totals contained in ACS.

References

Beckman, R.J., K.A. Baggerly, and M.D. McKay. 1996. Creating synthetic baseline populations. *Annals of Transportation Research 30*(6):415–429.

Wheaton, W.D., J.C. Cajka, B.M. Chasteen, D.K. Wagener, P.C. Cooley, L. Ganapathi, D.J. Roberts, and J.L. Allpress. 2009. Synthesized population databases: A U.S. geospatial database for agent-based models. RTI Press paper available at http://www.rti.org/pubs/mr-0010-0905-wheaton.pdf.

Appendix A: Data Dictionary

The Data Dictionary describes the contents of each field in each of the synthesized data files.

[prefix]_pums_h.txt

- Contains complete household records from original PUMS data. For details on field contents and definitions, please see the Public Use Microdata Sample: 2000 Census of Population and Housing at http://www.census.gov/acs/www/data_documentation/public_use_microdata_sample/.

[prefix]_pums_p.txt

- Contains complete person records from original PUMS data. For details on field contents and definitions, please see complete PUMS documentation at http://www.census.gov/acs/www/data_documentation/public_use_microdata_sample/.

[prefix]_schools.txt

Field Name	Description
School_id	A unique identifier for each school.
Name	The name of the school.
Stabbr	The two-letter abbreviation of the state in which the school is located.
Address	The physical address of the school, if known.
City	The city where the school is located.
County	The name of the county where the school is located.
Zip	Five-digit zip code in which the school is located.
Zip4	The nine-digit zip code (i.e., zip code plus four digits) in which the school is located.
Nces_id	A unique identifier for each school in the National Center for Education Statistics (NCES) database.
Total	The total number of students enrolled in the school.
Prek	The total number of pre-kindergarteners enrolled in the school.
Kinder	The total number of kindergarteners enrolled in the school.
Gr01-gr12	The total number of students in grades one through twelve.
Ungraded	The total number of students enrolled in the school whose specific grade level is unknown.
Latitude	The latitude of the school, based on geocoding.
Longitude	The longitude of the school, based on geocoding.
Source	The source of the school's information (either NCES [for public schools] or schoolinformation.com [for private schools]).

Field Name	Description
Stco	State and county FIPS codes of the county and state in which the schools are located.

[prefix]_workplaces.txt

Field Name	Description
workplace_id	A unique identifier for each workplace.
Num_workers_assigned	Number of workers assigned to the workplace.
Latitude	The latitude of the workplace, based on geocoding.
Longitude	The longitude of the workplace, based on geocoding.

[prefix]_synth_gq_people.txt

Field Name	Description
P_id	A unique identifier for each synthetic group quarters agent. These identifiers are also separate and unique from the **person_ids** included for the agents residing in households.
Gq_id	A unique identifier for every group quarters facility in the United States.
Gq_type	A code indicating the type of group quarters facility (i.e., M = military, P = prison, N = nursing home, C = college).
sporder	A unique serial number assigned to persons within each group quarter.
Age	The age of this synthesized group quarters agent.
Sex	The sex of this group quarters agent (i.e., 1 = male and 2 = female).

[prefix]_synth_gq.txt

Field Name	Description
Gq_id	A unique identifier for every group quarters facility in the United States.
Gq_type	A code indicating the type of group quarters facility (i.e., M = military, P = prison, N = nursing home, C = college).
Persons	The number of synthesized persons who live in this facility.
Stcotrbg2010	The facility's census 2010 block group identifier, . Included because the group quarters data was provided by the 2010 census, not the 2005-2009 ACS
Stcotrbg2000	The facility's census 2000 block group identifier. This is the block group identifier that matches the 2005-2009 ACS data.
Latitude	The latitude of the facility, based on geocoding.
Longitude	The longitude of the facility, based on geocoding.

[prefix]_synth_households.txt

Field Name	Description
Hh_id	A unique identifier for each generated household in the United States.
Serialno	This is the PUMS standard **serialno** field, which is the PUMS unique identifier for households within states.
Stcotrbg	The state, county, tract, and block group FIPS code of the household.
Hh_race	The coded race of the householder (see Appendix B for codes).
Hh_income	The household income.
Hh_size	The number of persons in the household.
Hh_age	The age of the head of household.
Latitude	The latitude of the household, based on geocoding.
Longitude	The longitude of the household, based on geocoding.

[prefix]_synth_people.txt

Field Name	Description
P_id	A unique identifier for each person record.
Hh_id	A unique identifier for each generated household in the United States. This identifier links to the **hh_id** field in the **synth_households.txt** file.
serialno	The original PUMS serial number (unique identifier). This code is used to link persons in the synth_people.txt file to the pums_p.txt file.
Stcotrbg	The state, county, tract, and block group FIPS code of the person.
Age	The person's age.
Sex	The person's sex (duplicate of the sex attribute in the **pums_p.txt** file), where 1 = male and 2 = female.

Field Name	Description
Race	The persons coded race. See Appendix B for codes.
sporder	A unique serial number assigned to persons within each household.
Relate	The relationship of the person to the household (see Appendix B for codes).
School_id	Identifier of the school to which this person is assigned. If the person is not assigned to a school, then this field will be blank.
Workplace_id	Identifier of the workplace to which this person is assigned. If the person is not assigned to a workplace, then this field will be blank. This identifier consists of state, county, tract, and block group FIPS codes and a unique serial number added as a suffix.

[prefix]_age_compare.txt

Field Name	Description
Stcotrbg	The state, county, tract, block group ID of the block group.
Count_of_households	The total number of households in the block group as specified by ACS.
Acs_1	The estimated count of households with head of household between 15 and 24 from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_1	The count of synthetic households created for age category 1 (head of household 15–24 years old).
Diff_1	The difference between the Sp_1 synthetic population count and the Acs_1 expected count.
w_diff_1	The weighted difference between the synthetic population generator count (Sp_1) and the expected count from ACS (Acs_1) for the 15–24 age category.
Moe_1	The ACS margin of error for the Acs_1 variable.
Out_moe_1	Flag for margin of error ("y" if Sp_1 is outside the margin of error; "n" if Sp_1 is inside the margin of error).
Acs_2	The estimated count of households with head of household between 25 and 34 from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_2	The count of synthetic households created for age category 2 (head of household 25–34 years old).
Diff_2	The difference between the Sp_2 synthetic population count and the Acs_2 expected count.
w_diff_2	The weighted difference between the synthetic population generator count (Sp_2) and the expected count from ACS (Acs_2) for the 25–34 age category.
Moe_2	The ACS margin of error for the Acs_2 variable.
Out_moe_2	Flag for margin of error ("y" if Sp_2 is outside the margin of error; "n" if Sp_2 is inside the margin of error).
Acs_3	The estimated count of households with head of household between 35 and 44 from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.

Field Name	Description
Sp_3	The count of synthetic households created for age category 3 (head of household 35–44 years old).
Diff_3	The difference between the Sp_3 synthetic population count and the Acs_3 expected count.
w_diff_3	The weighted difference between the synthetic population generator count (Sp_3) and the expected count from ACS (Acs_3) for the 35–44 age category.
Moe_3	The ACS margin of error for the Acs_3 variable.
Out_moe_3	Flag for margin of error ("y" if Sp_3 is outside the margin of error; "n" if Sp_3 is inside the margin of error).
Acs_4	The estimated count of households with head of household between 45 and 54 from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_4	The count of synthetic households created for age category 4 (head of household 45–54 years old).
Diff_4	The difference between the Sp_4 synthetic population count and the Acs_4 expected count.
w_diff_4	The weighted difference between the synthetic population generator count (Sp_4) and the expected count from ACS (Acs_4) for the 45–54 age category.
Moe_4	The ACS margin of error for the Acs_4 variable.
Out_moe_4	Flag for margin of error ("y" if Sp_4 is outside the margin of error; "n" if Sp_4 is inside the margin of error).
Acs_5	The estimated count of households with head of household between 55 and 64 from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_5	The count of synthetic households created for age category 5 (head of household 55–64 years old).
Diff_5	The difference between the Sp_5 synthetic population count and the Acs_5 expected count.
w_diff_5	The weighted difference between the synthetic population generator count (Sp_5) and the expected count from ACS (Acs_5) for the 55–64 age category.
Moe_5	The ACS margin of error for the Acs_5 variable.
Out_moe_5	Flag for margin of error ("y" if Sp_5 is outside the margin of error; "n" if Sp_5 is inside the margin of error).
Acs_6	The estimated count of households with head of household between 65 and 74 from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_6	The count of synthetic households created for age category 6 (head of household 65–74 years old).
Diff_6	The difference between the Sp_6 synthetic population count and the Acs_6 expected count.
w_diff_6	The weighted difference between the synthetic population generator count (Sp_6) and the expected count from ACS (Acs_6) for the 65–74 age category.
Moe_6	The ACS margin of error for the Acs_6 variable.

Field Name	Description
Out_moe_6	Flag for margin of error ("y" if Sp_6 is outside the margin of error; "n" if Sp_6 is inside the margin of error).
Acs_7	The estimated count of synthetic households with head of household greater than 74 years old from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_7	The count of synthetic households created for age category 7 (head of household older than 74 years).
Diff_7	The difference between the Sp_7 synthetic population count and the Acs_7 expected count.
w_diff_7	The weighted difference between the synthetic population generator count (Sp_7) and the expected count from ACS (Acs_7) for the >74 age category.
Moe_7	The ACS margin of error for the Acs_7 variable.
Out_moe_7	Flag for margin of error ("y" if Sp_7 is outside the margin of error; "n" if Sp_7 is inside the margin of error).

[prefix]_size_compare.txt

Field Name	Description
Stcotrbg	The state, county, tract, block group ID of the block group.
Count_of_households	The total number of households in the block group as specified by ACS.
Acs_1	The estimated count of households with one person from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_1	The count of synthetic households created for size category 1 (households with one person).
Diff_1	The difference between the Sp_1 synthetic population count and the Acs_1 expected count.
w_diff_1	The weighted difference between the synthetic population generator count (Sp_1) and the expected count from ACS (Acs_1) for the one-person household size category.
Moe_1	The ACS margin of error for the Acs_1 variable.
Out_moe_1	Flag for margin of error ("y" if Sp_1 is outside the margin of error; "n" if Sp_1 is inside the margin of error).
Acs_2	The estimated count of synthetic households with two persons from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_2	The count of synthetic households created for size category 2 (households with two persons).
Diff_2	The difference between the Sp_2 synthetic population count and the Acs_2 expected count.
w_diff_2	The weighted difference between the synthetic population generator count (Sp_2) and the expected count from ACS (Acs_2) for the two-person household size category.
Moe_2	The ACS margin of error for the Acs_2 variable.

Field Name	Description
Out_moe_2	Flag for margin of error ("y" if Sp_2 is outside the margin of error; "n" if Sp_2 is inside the margin of error).
Acs_3	The estimated count of households with three persons from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_3	The count of synthetic households created for size category 3 (households with three persons).
Diff_3	The difference between the Sp_3 synthetic population count and the Acs_3 expected count.
w_diff_3	The weighted difference between the synthetic population generator count (Sp_3) and the expected count from ACS (Acs_3) for the three-person household size category.
Moe_3	The ACS margin of error for the Acs_3 variable.
Out_moe_3	Flag for margin of error ("y" if Sp_3 is outside the margin of error; "n" if Sp_3 is inside the margin of error).
Acs_4	The estimated count of households with four persons from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_4	The count of synthetic households created for size category 4 (households with four persons).
Diff_4	The difference between the Sp_4 synthetic population count and the Acs_4 expected count.
w_diff_4	The weighted difference between the synthetic population generator count (Sp_4) and the expected count from ACS (Acs_4) for the four-person household size category.
Moe_4	The ACS margin of error for the Acs_4 variable.
Out_moe_4	Flag for margin of error ("y" if Sp_4 is outside the margin of error; "n" if Sp_4 is inside the margin of error).
Acs_5	The estimated count of households with five persons from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_5	The count of synthetic households created for size category 5 (households with five persons).
Diff_5	The difference between the Sp_5 synthetic population count and the Acs_5 expected count.
w_diff_5	The weighted difference between the synthetic population generator count (Sp_5) and the expected count from ACS (Acs_5) for the five-person household size category.
Moe_5	The ACS margin of error for the Acs_5 variable.
Out_moe_5	Flag for margin of error ("y" if Sp_5 is outside the margin of error; "n" if Sp_5 is inside the margin of error).
Acs_6	The estimated count of households with six persons from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_6	The count of synthetic households created for size category 6 (households with six persons).
Diff_6	The difference between the Sp_6 synthetic population count and the Acs_6 expected count.

Field Name	Description
w_diff_6	The weighted difference between the synthetic population generator count (Sp_6) and the expected count from ACS (Acs_6) for the six-person household size category.
Moe_6	The ACS margin of error for the Acs_1 variable.
Out_moe_6	Flag for margin of error ("y" if Sp_1 is outside the margin of error; "n" if Sp_1 is inside the margin of error).
Acs_7	The estimated count of households with 7 or more persons from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_7	The count of synthetic households created for size category 7 (households with seven or more persons).
Diff_7	The difference between the Sp_7 synthetic population count and the Acs_7 expected count.
w_diff_7	The weighted difference between the synthetic population generator count (Sp_7) and the expected count from ACS (Acs_7) for the 7 or more persons household size category.
Moe_7	The ACS margin of error for the Acs_7 variable.
Out_moe_7	Flag for margin of error ("y" if Sp_7 is outside the margin of error; "n" if Sp_7 is inside the margin of error).

[prefix]_race_compare.txt

Field Name	Description
Stcotrbg	The state, county, tract, block group ID of the block group.
Count_of_households	The total number of households in the block group as specified by ACS.
Acs_1	The estimated count of households with head of household race of "white alone" from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_1	The count of synthetic households created for race category 1 ("white alone").
Diff_1	The difference between the Sp_1 synthetic population count and the Acs_1 expected count.
w_diff_1	The weighted difference between the synthetic population generator count (Sp_1) and the expected count from ACS (Acs_1) for the race 1 ("white alone") category.
Moe_1	The ACS margin of error for the Acs_1 variable.
Out_moe_1	Flag for margin of error ("y" if Sp_1 is outside the margin of error; "n" if Sp_1 is inside the margin of error).
Acs_2	The estimated count of households with head of household race of "Black or African American alone" from ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_2	The count of synthetic households created for race category 2 ("Black or African American alone").
Diff_2	The difference between the Sp_2 synthetic population count and the Acs_2 expected count.

Field Name	Description
w_diff_2	The weighted difference between the synthetic population generator count (Sp_2) and the expected count from ACS (Acs_2) for the race 2 ("black or African American") category.
Moe_2	The ACS margin of error for the Acs_2 variable.
Out_moe_2	Flag for margin of error ("y" if Sp_2 is outside the margin of error; "n" if Sp_2 is inside the margin of error).
Acs_3	The estimated count of households with head of household race of "Asian alone" from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_3	The count of synthetic households created for race category 3 ("Asian alone").
Diff_3	The difference between the Sp_3 synthetic population count and the Acs_3 expected count.
w_diff_3	The weighted difference between the synthetic population generator count (Sp_3) and the expected count from ACS (Acs_3) for the race 3 category.
Moe_3	The ACS margin of error for the Acs_3 variable.
Out_moe_3	Flag for margin of error ("y" if Sp_3 is outside the margin of error; "n" if Sp_3 is inside the margin of error).
Acs_4	The estimated count of households with head of household race of "Other race" from the ACS data. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_4	The count of synthetic households created for race category 4 ("Other race").
Diff_4	The difference between the Sp_4 synthetic population count and the Acs_4 expected count.
w_diff_4	The weighted difference between the synthetic population generator count (Sp_4) and the expected count from ACS (Acs_4) for the race 4 category.
Moe_4	The ACS margin of error for the Acs_4 variable.
Out_moe_4	Flag for margin of error ("y" if Sp_4 is outside the margin of error; "n" if Sp_4 is inside the margin of error).
Acs_5	The estimated count of households with race category 5 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_5	The count of synthetic households created for race category 5 ("2 or more races").
Diff_5	The difference between the Sp_5 synthetic population count and the Acs_5 expected count.
w_diff_5	The weighted difference between the synthetic population generator count (Sp_5) and the expected count from ACS (Acs_5) for the "2 or more races" household category.
Moe_5	The ACS margin of error for the Acs_5 variable.
Out_moe_5	Flag for margin of error ("y" if Sp_5 is outside the margin of error; "n" if Sp_5 is inside the margin of error).

[prefix]_Income_compare.txt

Field Name	Description
Stcotrbg	The state, county, tract, block group ID of the block group.
Count_of_households	The total number of households in the block group as specified by ACS.
Acs_1	The estimated count of households with income <$10,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_1	The count of synthetic households created for income category 1 (households with income <$10,000).
Diff_1	The difference between the Sp_1 synthetic population count and the Acs_1 expected count.
w_diff_1	The weighted difference between the synthetic population generator count (Sp_1) and the expected count from ACS (Acs_1) for the <$10,000 household income category.
Moe_1	The ACS margin of error for the Acs_1 variable.
Out_moe_1	Flag for margin of error ("y" if Sp_1 is outside the margin of error; "n" if Sp_1 is inside the margin of error).
Acs_2	The estimated count of households with income from $10,000 to $15,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_2	The count of synthetic households created for income category 2 (households with income from $10,000 to $15,000).
Diff_2	The difference between the Sp_2 synthetic population count and the Acs_2 expected count.
w_diff_2	The weighted difference between the synthetic population generator count (Sp_2) and the expected count from ACS (Acs_2) for the $10,000 to $15,000 income category.
Moe_2	The ACS margin of error for the Acs_2 variable.
Out_moe_2	Flag for margin of error ("y" if Sp_2 is outside the margin of error; "n" if Sp_2 is inside the margin of error).
Acs_3	The estimated count of households with income from $15,001 to $25,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_3	The count of synthetic households created for size category 3 (households with income from $15,001 to $25,000).
Diff_3	The difference between the Sp_3 synthetic population count and the Acs_3 expected count.
w_diff_3	The weighted difference between the synthetic population generator count (Sp_3) and the expected count from ACS (Acs_3) for the $15,001 to $25,000 household income category.
Moe_3	The ACS margin of error for the Acs_3 variable.
Out_moe_3	Flag for margin of error ("y" if Sp_3 is outside the margin of error; "n" if Sp_3 is inside the margin of error).
Acs_4	The estimated count of households with income from $25,001 to $35,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.

Field Name	Description
Sp_4	The count of synthetic households created for income category 4 (households with income from $25,001 to $35,000).
Diff_4	The difference between the Sp_4 synthetic population count and the Acs_4 expected count.
w_diff_4	The weighted difference between the synthetic population generator count (Sp_4) and the expected count from ACS (Acs_4) for the $25,001 to $35,000 household income category.
Moe_4	The ACS margin of error for the Acs_4 variable.
Out_moe_4	Flag for margin of error ("y" if Sp_4 is outside the margin of error; "n" if Sp_4 is inside the margin of error).
Acs_5	The estimated count of households with income from $35,001 to $50,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_5	The count of synthetic households created for income category 5 (households with income from $35,001 to $50,000).
Diff_5	The difference between the Sp_5 synthetic population count and the Acs_5 expected count.
w_diff_5	The weighted difference between the synthetic population generator count (Sp_5) and the expected count from ACS (Acs_5) for the $35,001 to $50,000 household income category.
Moe_5	The ACS margin of error for the Acs_5 variable.
Out_moe_5	Flag for margin of error ("y" if Sp_5 is outside the margin of error; "n" if Sp_5 is inside the margin of error).
Acs_6	The estimated count of households with income from $50,001 to $100,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_6	The count of synthetic households created for income category 6 (households with income from $50,001 to $100,000).
Diff_6	The difference between the Sp_6 synthetic population count and the Acs_6 expected count.
w_diff_6	The weighted difference between the synthetic population generator count (Sp_6) and the expected count from ACS (Acs_6) for the $50,001 to $100,000 household income category.
Moe_6	The ACS margin of error for the Acs_6 variable.
Out_moe_6	Flag for margin of error ("y" if Sp_6 is outside the margin of error; "n" if Sp_6 is inside the margin of error).
Acs_7	The estimated count of households with income >$100,000 from the ACS. This count is assumed to be the best estimate of households that the synthetic population generator attempts to match.
Sp_7	The count of synthetic households created for income category 7 (households with income greater than $100,000).
Diff_7	The difference between the Sp_7 synthetic population count and the Acs_7 expected count.
w_diff_7	The weighted difference between the synthetic population generator count (Sp_7) and the expected count from ACS (Acs_7) for the >$100,000 household income category.

Field Name	Description
Moe_7	The ACS margin of error for the Acs_7 variable.
Out_moe_7	Flag for margin of error ("y" if Sp_7 is outside the margin of error; "n" if Sp_7 is inside the margin of error).

[prefix]_summary_compare.txt

Field Name	Description
Stcotrbg	Combined state, county, census tract, and block group identifier to uniquely identify each block group.
Count_of_households	The number of households in the block group according to the ACS.
Size_accuracy	The overall weighted difference for all size groups between that expected by the ACS and that returned by the synthetic population generator.
Age_accuracy	The overall weighted difference for all age groups between that expected by the ACS and that returned by the synthetic population generator.
Income_accuracy	The overall weighted difference for all income groups between that expected by the ACS and that returned by the synthetic population generator.
Race_accuracy	The overall weighted difference for all race groups between that expected by the ACS and that returned by the synthetic population generator.
Overall_accuracy	The sum of the weighted differences across all selection variables (size, age, income, and race).

Appendix B: Codes

[prefix]_synth_households.txt

Variable: hh_income:

PUMS original variable name: HINCP

PUMS Description: Household income (past 12 months)

PUMS Codes:

Code	Meaning
Bbbbbbbb	N/A (GQ/vacant)
-59999	Loss of -$59,999 or more
1	$1 or break even
000000002-99999999	Total household income in dollars (Components are rounded)

Variable: hh_size:

PUMS original variable name: NP

PUMS Description: Number of person records following

PUMS Codes:

Code	Meaning
00	Vacant unit
01	One person record (one person in household or any person in group quarters)
02-20	Number of person records (number of persons in household)

Variable: hh_age:

PUMS original variable name: AGEP

PUMS Description: age

Notes: The AGEP variable for the householder (RELATE = 01) in each selected household is the householder age. It is attached to the [prefix]_synth_households.txt file because this field was used in selecting households in the IPF procedure.

PUMS Codes:

Code	Meaning
00	Under 1 year
01-99	1 to 99 years (Top-coded)

Variable: hh_race

PUMS original variable name: RAC1P

PUMS Description: Recoded detailed race code

Notes: The RAC1P variable for the householder (RELATE = 01) in each selected household is determined to be the race of the household. Even though this is a person-level characteristic, it was used at the household level to enable the use of race as a selection category in the IPF procedure.

PUMS Codes:

Code	Meaning
1	White alone
2	Black or African American alone
3	American Indian alone
4	Alaska Native alone
5	American Indian and Alaska Native tribes specified; or American Indian or Alaska native, not specified and no other races
6	Asian alone
7	Native Hawaiian or Other Pacific Islander alone
8	Some other race alone
9	Two or more major race groups

[prefix]_synth_persons.txt

Variable: relate

PUMS original variable name: RELATE

PUMS Description: Household relationship

PUMS Codes:

Code	Meaning
00	Reference Person
01	Husband/wife
02	Son/daughter
03	Brother/sister
04	Father/mother
05	Grandchild
06	In-law
07	Other relative
08	Roomer/boarder
09	Housemate/roommate
10	Unmarried partner
11	Foster child
12	Other nonrelative
13	Institutionalized group quarters population
14	Noninstitutionalized group quarters population

Variable: age

PUMS original variable name: AGEP

PUMS Description: age

PUMS Codes:

Code	Meaning
00	Under 1 year
01-99	1 to 99 years (Top-coded)

Variable: sex

PUMS original variable name: SEX

PUMS Description: Sex of person

PUMS Codes:

Code	Meaning
1	Male
2	Female

Variable: race

PUMS original variable name: RAC1P

PUMS Description: Recoded detailed race code

PUMS Codes:

Code	Meaning
1	White alone
2	Black or African American alone
3	American Indian alone
4	Alaska Native alone
5	American Indian and Alaska Native tribes specified; or American Indian or Alaska native, not specified and no other races
6	Asian alone
7	Native Hawaiian or Other Pacific Islander alone
8	Some other race alone
9	Two or more major race groups

Appendix C: Metadata File Contents

The listing below is an example of the metadata file contained with each synthetic population dataset.

The 'Dataset Summary Information' section provides information about the size and contents of each particular dataset including the version, geographic identifiers (state or county), the data sources, and the counts of households and persons in the dataset.

```
**************************************************************************
**              2005-2009 U.S. Synthesized Population Dataset             **
**              Copyright Research Triangle Institute  2012               **
**                           All Rights Reserved                          **
**                                                                        **
**              The development of this data was supported by             **
**               Grant Number U24GM0877044 (MIDAS) from the               **
**          National Institutes of General Medical Sciences (NIGMS)       **
**                                                                        **
**   The content is solely the responsibility of the authors and does not **
**       necessarily represent the official views of the NIGMS or the     **
**                       National Institutes of Health                    **
**                                                                        **
**************************************************************************

****************************** Contact Information ************************
**                                                                        **
**  Bill Wheaton                                                          **
**  Director, Geospatial Science and Technology Program                   **
**  RTI International                                                     **
**  P.O. Box 12194                                                        **
**  3040 Cornwallis Rd.                                                   **
**  Research Triangle Park, NC 27709                                      **
**  wdw@rti.org                                                           **
**  919-541-6158                                                          **
**                                                                        **
**************************************************************************

****************************** Citation Information ***********************
**                                                                        **
**  NIGMS and RTI request that you cite these data in any publication     **
**  or report in which they were used.  The proper citations are:         **
**                                                                        **
**  Data:                                                                 **
**        2005-2009 RTI U.S. Synthetic Population Ver. 2,                 **
**        RTI International. August, 2012. Downloaded from internet       **
**        URL: https://www.epimodels.org/midas/pubsyntdata1.do            **
**                                                                        **
**  Quick Start Guide:                                                    **
**                                                                        **
**        Wheaton, W.D., 2012. "U.S. Synthetic Population Database        **
**        2005-2009: Quick Start Guide".  RTI International. Retrieved    **
**        from http://www.epimodels.org/midasdocs/SynthPop/2005-2009_     **
**        synth_pop_ver2_quickstart_20120810.pdf                          **
**                                                                        **
**************************************************************************
```

```
************************ Dataset Summary Information ************************
               synth pop version: 2
                       geography: 01
                  geography name: Alabama
              count of households: 1819441
                 count of persons: 4389782
                 count of schools: 1896
              count of workplaces: 176187
         count of group quarters: 739
count of group quarter residents: 104436
                 source ACS year: 2005-2009 5-year sample
               source tiger year: 2008 (w/ modifications)
                   landscan year: 2008
                 date of extract: 2012-08-08

         full documentation: https://www.epimodels.org/midas/Rpubsyntdata1.do

****************** Information on Files and Field Contents ******************

-2005-2009_ver2_01_synth_households.txt
       hh_id
       serialno
       stcotrbg
       hh_race
       hh_income
       hh_size
       hh_age
       latitude
       longitude

-2005-2009_ver2_01_synth_people.txt
       p_id
       hh_id
       serialno
       stcotrbg
       age
       sex
       race
       sporder
       relate
       school_id
       workplace_id

-2005-2009_ver2_01_synth_gq.txt
       gq_id
       gq_type
       persons
       latitude
       longitude

-2005-2009_ver2_01_synth_gq_people.txt
       gq_id
       person_id
       sporder
       gq_type
       age
       sex
       latitude
       longitude

-2005-2009_ver2_01_schools.txt
       school_id
```

```
       name
       stabbr
       address
       city
       county
       zipcode
       zip4
       nces_id
       total
       prek
       kinder
       gr01-gr12
       ungraded
       latitude
       longitude
       source
       stco

-2005-2009_ver2_01_workplaces.txt
       workplace_id
       num_workers_assigned
       latitude
       longitude

-2005-2009_ver2_01_pums_h.txt
       See
http://www.census.gov/acs/www/data_documentation/public_use_microdata_sample

-2005-2009_ver2_01_pums_p.txt
       See
http://www.census.gov/acs/www/data_documentation/public_use_microdata_sample

-2005-2009_ver2_01_size_compare.txt
       stcotrbg
       acs_1
       sp_1
       diff_1
       per_diff_1
       moe_1
       out_moe_1
       [repeat for 1-7 size categories.]

-2005-2009_ver2_01_age_compare.txt
       stcotrbg
       acs_1
       sp_1
       diff_1
       per_diff_1
       moe_1
       out_moe_1
       [repeat for 1-7 age categories.]

-2005-2009_ver2_01_race_compare.txt
       stcotrbg
       acs_1
       sp_1
       diff_1
       per_diff_1
       moe_1
       out_moe_1
       [repeat for 1-5 race categories.]

-2005-2009_ver2_01_income_compare.txt
```

```
       stcotrbg
       acs_1
       sp_1
       diff_1
       per_diff_1
       moe_1
       out_moe_1
       [repeat for 1-7 income categories.]

-2005-2009_ver2_01_summary_compare.txt
       stcotrbg
       count_of_households
       size_accuracy
       age_accuracy
       income_accuracy
       race_accuracy
       overall_accuracy

*******************************************************************************
```