Census Transportation Planning Products Data Synthesis Program Processing

Draft Methodology Report

Authors

Minsun Riddles
Tom Krenzke
Lin Li
Robyn Ferg
Medha Uppala

March 10, 2025

Prepared for:
U.S. Census Bureau
Washington D.C.

Prepared by:
Westat
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Westat®

COPY NO. ____

Census Transportation Planning Products

Data Synthesis Program Processing

DRAFT METHODOLOGY REPORT

CBDRB-FY25-ACS0003-B0001

Prepared for:

U.S. Census Bureau

February 27, 2025

ACKNOWLEDGMENT OF SPONSORSHIP

This work was sponsored by the U.S. Bureau of the Census.

Census Transportation Planning Products
Data Synthesis Program Processing
DRAFT METHODOLOGY REPORT

Prepared for:

U.S. Census Bureau

February 27, 2025

Minsun Riddles, Tom Krenzke, Lin Li, Robyn Ferg, Medha Uppala

Westat, Rockville, Maryland

Table of Contents

Contents (Continued)

Author Acknowledgments

The work reported herein was performed under contract with the U.S. Census Bureau. It is the result of the research carried out in the NCHRP Project 08-79 by Westat. Due to data security requirements relating to the American Community Survey data, the work for this project was done through Virtual Desktop Infrastructure (VDI). The authors gratefully acknowledge the many individuals who contributed to the accommodations for processing the data synthesis programs.

At the Census Bureau, special thanks to Nicholas Spanos, Grace Clemons, Brian McKenzie, Liam Nealon, Kristin Wimbrow, Pavina Sengkhyavong, Charlynn Burd, Xiang Li and Mike Starsinic, our Census Bureau contacts, for special attention to this project. The authors are indebted to the Census Bureau staff for accommodating our various system's needs, for helpful arrangements for accessing the input data files, and providing direction for the output files.

Executive Summary

The final report provides documentation of the statistical confidentiality data treatments and the programs used for generating synthetic American Community Survey (ACS) 2017-2021 microdata. The table generator for the Census Transportation Planning Products (CTPP) will process the synthetic microdata for the CTPP pre-specified tables. The data synthesis process was developed from the extensive research study called National Cooperative Highway Research Program (NCHRP) 08-79, which was undertaken in 2010-2011 to develop synthetic data procedures that would produce small area data (e.g., residence to workplace flows for areas approximately the size of Block Groups) that would not violate the Census Bureau's confidentiality law. During the NCHRP 08-79 research, Westat, under contract to the National Academy of Sciences, and during the production run under contract to the Census Bureau, worked closely with the Census Bureau Disclosure Review Board (DRB) and Census Bureau ACS operations staff.

The approach described in this report is the same as implemented with the 2006-2010 and 2012-2016 data with the following key differences. First, in prior applications, the CTPP tables were divided into two sets: Set A and Set B. The "Set A" tables were produced from un-perturbed data and "Set B" tables, were produced from perturbed data. However, for this application to 2017-2021 ACS microdata, all CTPP tables were generated using the synthetic microdata. Second, the lowest level of geography has switched from traffic analysis zones (TAZs), which were combinations of Census blocks, to Census tracts. Third, to satisfy the Disclosure Review Board's review, the amount of synthesis has increased in both the number of variables and the number of records synthesized. That is, the overarching rule is to ensure that 50% of all records in the ACS 2017-2021 five-year microdata are synthesized.

Prior to applying the synthetic data approach, high risk data values were identified using threshold rules as defined for the CTPP. The high-risk data values were targeted for data replacement (synthesis). Select variables and select records with high disclosure risks were synthesized, which is referred to as a "select" data synthesis approach. The main synthesis procedure conducts a model-assisted constrained hot deck (MACH), which was developed through the NCHRP 08-79 project and expanded through research conducted for the Census Bureau while implemented on 2006-2010 and 2012-2016 data. The approach constrains the amount of change in the target variable by forming hot deck cells using "bins" created on the target variable itself (bins are recoded categories such that more than one published category was included in the bin) and model predictions. Within the MACH framework is the unconstrained semi-parametric approach. Additive noise was also

applied to a small number of variables. After synthesis, a raking procedure re-calibrated weights for aggregated geography.

The usual ACS formula for variance estimation treats the ACS data as if it were reported without accounting for variance caused by the data synthesis. Therefore, we recommended the use of an approach developed under NCHRP 08-79 in computing the standard errors of the estimates. This was approved by the Census Bureau.

Lastly, the final report provides a documentation of the programs needed to process the synthetic data treatments. Flowcharts are presented to help illustrate the process flow.

Introduction 1

Prior to last decade, decennial censuses provided an invaluable source of information for decision-support for transportation planning analyses. This transportation-specific census data product, known as the Census Transportation Planning Products (CTPP), includes data derived from specific transportation questions such as commuting times, distance from home to work, and mode of travel, along with data on population and employment and their related attributes. Of greatest importance to transportation planners, however, was the unique ability of the decennial census to provide this information for cities and towns of different sizes, as well as for tracts and block groups, or combinations of these groups into traffic analysis zones (TAZs).

In the past decade or so, the Census Bureau began to collect these data on a continuing basis as part of its American Community Survey (ACS). This development allows the publication of small area detail to be provided at regular intervals throughout the decade. Because of this change in collection mode, the transportation products were based on five-year estimates, specifically using the years 2006 through 2010, and later 2012-2016. Previously, the tabulations, and especially those at the TAZ-level, were subject to suppression procedures to protect against the disclosure of identifiable information.

In order to eliminate the effects of data suppression on small-area data and retain the necessary attributes to support the desired micro modeling at the TAZ level, the research study National Cooperative Highway Research Program (NCHRP) 08-79 was undertaken in 2010-2011 (Krenzke et al., 2011; hereafter we refer to this as the "NCHRP 08-79 Final Report") to develop data synthesis procedures that would produce small area data that would not violate the Census Bureau's confidentiality law. During the research, Westat, under contract to the NCHRP and the National Academy of Sciences, worked closely with the Census Bureau Disclosure Review Board (DRB). Under this contract with the Census Bureau, Westat used the synthetic data methodologies and processed the computer programs prior to the Bureau's production of the American Association of State Highway and Transportation Officials (AASHTO) extensive transportation package called the CTPP.

Data confidentiality protection is a critical issue. In addition to the risk of disclosure from the CTPP tables themselves, there is a threat to disclosure that could result from relating the tabular data to the

ACS public use microdata files. The data synthesis procedures applied to the 2006-2010 ACS data, 2012-2016, and now 2017-2021 relevant to this report, first identify high risk data values using threshold rules as defined for the CTPP. The high risk data values are targeted for data replacement using a model-assisted constrained hot deck (MACH) approach, which constrains the amount of change in the target variable by forming hot deck cells using information from the target variable itself, model predictions, local areas, sample weights, and other key auxiliary information. Within the MACH framework is an unconstrained hot deck (semi-parametric approach) for unordered categorical variables. Additive noise is applied to select variables. Select variables and select records with high disclosure risks were synthesized, which is referred to as a "select" data synthesis approach. The process used to employ the MACH includes a software program called *SDCPert*, which is proprietary. The *SDCPert* program has been well tested and vetted. The process includes many modules and is set up to select records for data replacement, build models to help with the synthetic data generation, cycle through the variables selected to be synthesized, and generate the synthetic data. The MACH methodology has been presented in public forums, such as for the American Statistical Association's Committee on Privacy and Confidentiality (https://zenodo.org/record/4121967). The general methodology is published in Section 3 in Krenzke, et al (2017). The results show that the synthesized data will provide CTPP data users mostly complete tables that are accurate enough to support transportation planning applications, but that also are modified enough to satisfy the Disclosure Review Board's (DRB's) requirement of reducing disclosure risk.

1.1 Key Differences from 2012-2016

The approach described in Section 2 is the same as implemented with the 2006-2010 and 2012-2016 data with the following key differences. First, in prior applications, the CTPP tables were divided into two sets: Set A and Set B. The "Set A" tables were produced from un-perturbed data and "Set B" tables, were produced from perturbed data. This allowed the data treatments to focus on a smaller set of tables. Another benefit was that data were not touched unless needed, perhaps providing better data utility to the users. However, one disadvantage is that any aggregation of the Set B tables did not equal the standard Set A results. For this application to 2017-2021 ACS microdata, all CTPP tables will be generated using the synthetic microdata. This will ensure that the additivity property will be retained.

Second, the lowest level of geography has changed from TAZs, which were combinations of Census blocks, to Census tracts. Such a change reduces disclosure risk.

Third, to satisfy the Disclosure Review Board's review and reduce the disclosure risk even further, the amount of synthesis has increased in both the number of variables and the number of records synthesized. Past processing only synthesized 9 variables; however, 33 variables are synthesized in 2017-2021. In addition, the overarching rule is to ensure that 50% of all records in the ACS 2017-2021 five-year microdata are synthesized. After merging VHOUS and VPERS files together, for crosstabs including only VHOUS variables, we would expect 50% of the records to be synthesized. For crosstabs including only VPERS variables, we would expect 50% of the records to be synthesized. For crosstabs including both VHOUS and VPERS variables, we would expect more than 50% of the records to be synthesized.

Fourth, the five-year CTPP, based on 2006–2010 data, had an extended workplace allocation process applied post-hoc to all records. Nationally, extended workplace allocation was necessary for records missing workplace geography below the place level. This was a procedure conducted by the Census Bureau, and the implication was that Part 2 (Workplace) and Part 3 (Worker Flow) tables include on average non-missing block- and TAZ-level values of workplace allocation for only 77 percent of all the worker records. This process coded workplaces to the block for another 13 percent of the microdata records, resulting in about 90 percent of records with block-level workplaces and 10 percent with only place-level workplaces. Due to uncertainty in these imputations, the extended workplace allocation was discontinued for the 2012-2016 CTPP and the 2017-2021 CTPP. Therefore, about 20 to 25% of workplaces are missing in the 2017-2021 CTPP.

Lastly, among the differences between 2012-2016 and 2017-2021 CTPP is the set of tables. The set of tables for the 2017-2021 CTPP were provided to Westat from the Census Bureau under the filename "CTPP Table Shells-Part 1 Residence Part 2 POW and Part 3 Flow Tables 2017-2021_5-Year_DRB_Rules", hereafter called "CTPP Tables Spreadsheet". ***The CTPP Tables Spreadsheet contains the disclosure rules established for this process***. An example of a disclosure rule occurs to non-standard ACS variables and categories, where the Rule of 3 (three or more records) apply to a new category. In such cases, the whole table is suppressed if it fails Rule of 3 for residence and place-of-work tables using synthetic microdata.

1.2 Guidelines for the Production of CTPP Tables

The following are guidelines to produce the CTPP tables.

1. CTPP tables will be based on synthetic ACS data and CTPP adjusted weight. For variance estimation purposes, CTPP tables will need to be processed a second time using original ACS data and original ACS weights (more discussion is in Section 2.5).

2. The CTPP tables will be shown without cell suppression rules applied for standard ACS variables and categories. For non-standard variables and categories, suppression rules will be applied, as given in the CTPP Tables Spreadsheet.

3. Users should be alerted through the table title or a footnote that the CTPP tables were generated from disclosure-protection treated data.

4. The synthetic microdata file resulting from the initial risk analysis on Part 1, 2, and 3 tables will be used for all localities. The tables will be generated from the same synthetic microdata for all geographies specified, which may include Tracts, Places, Counties, States and PUMAs.

5. The synthetic microdata file will be used even where there are no violations as determined by the initial risk analysis. Even if the values of variables are unchanged, the CTPP adjusted raked weights may differ from the original ACS weights, and therefore the CTPP estimates will be different from the same tables generated from the original ACS data and weights.

6. For flows with missing workplace, no published numbers will be provided.

7. For people who work outside of the US, there is only one summary level that will include worked in Canada; worked in Mexico; worked elsewhere (not in U.S., Canada, or Mexico). Flows involving Puerto Rico will only include state and county.

8. The microdata file from which the CTPP tables are generated will be produced solely for the purpose of generating the CTPP tables. It is not intended to be used for dynamic queries for tables or microdata analyses. Household-level variables and person-level variables were treated separately and have not been adjusted to be internally consistent with each other. Among other examples, highly correlated variables, such as income, earnings and poverty have not been modified to be fully consistent in the microdata, except where they needed to be in the CTPP tables.

Methods 2

During the production stage we processed the synthesis programs on the full 2017-2021 American Community Survey (ACS) 5-year samples and all variables required for the 5-year 2017-2021 CTPP tabulation. This activity was done using the Census Bureau's information technology (IT) systems through their virtual desktop infrastructure. The resulting synthetic ACS microdata will be used in generating the CTPP tables.

Largely following two previous rounds, the general steps to the synthesis approach were to first identify high risk data values, as identified using threshold rules as defined for the CTPP. The high risk data values were targeted for data synthesis. The main procedure conducts a model-assisted constrained hot deck (MACH), which was developed through the NCHRP 08-79 project and expanded through research conducted for the Census Bureau by Westat. The approach constrains the amount of change in the target variable by forming hot deck cells using "bins" created on the target variable itself (bins are recoded categories such that more than one published category was included in the bin) and model predictions. Within the MACH framework is an unconstrained hot deck (semi-parametric approach) for unordered categorical variables. An additive noise approach is also available to use for some continuous variables.

The input files for years 2017 – 2021 were as follows:

- ACS household-level
  - vhous.sas7bdat
- ACS person-level file
  - vpers.sas7bdat

The steps involved in the synthesis process were as follows:

- Initial risk analysis;
- Data synthesis
- Weight calibration—raking;
- Data utility and risk measures; and

- Data clean up and output files.

Figure 2-1 provides the process flow of the activities relating to generating CTPP synthetic data. It shows the flow of tasks needed to carry out the data production. Section 2.1 first discusses the design of the preliminary steps and initial risk analysis. Section 2.2 describes the synthesis approaches. Section 2.3 introduces the raking procedure. Section 2.4 presents the data utility and disclosure risk measures. Section 2.5 discusses the variance estimation for synthetic data.

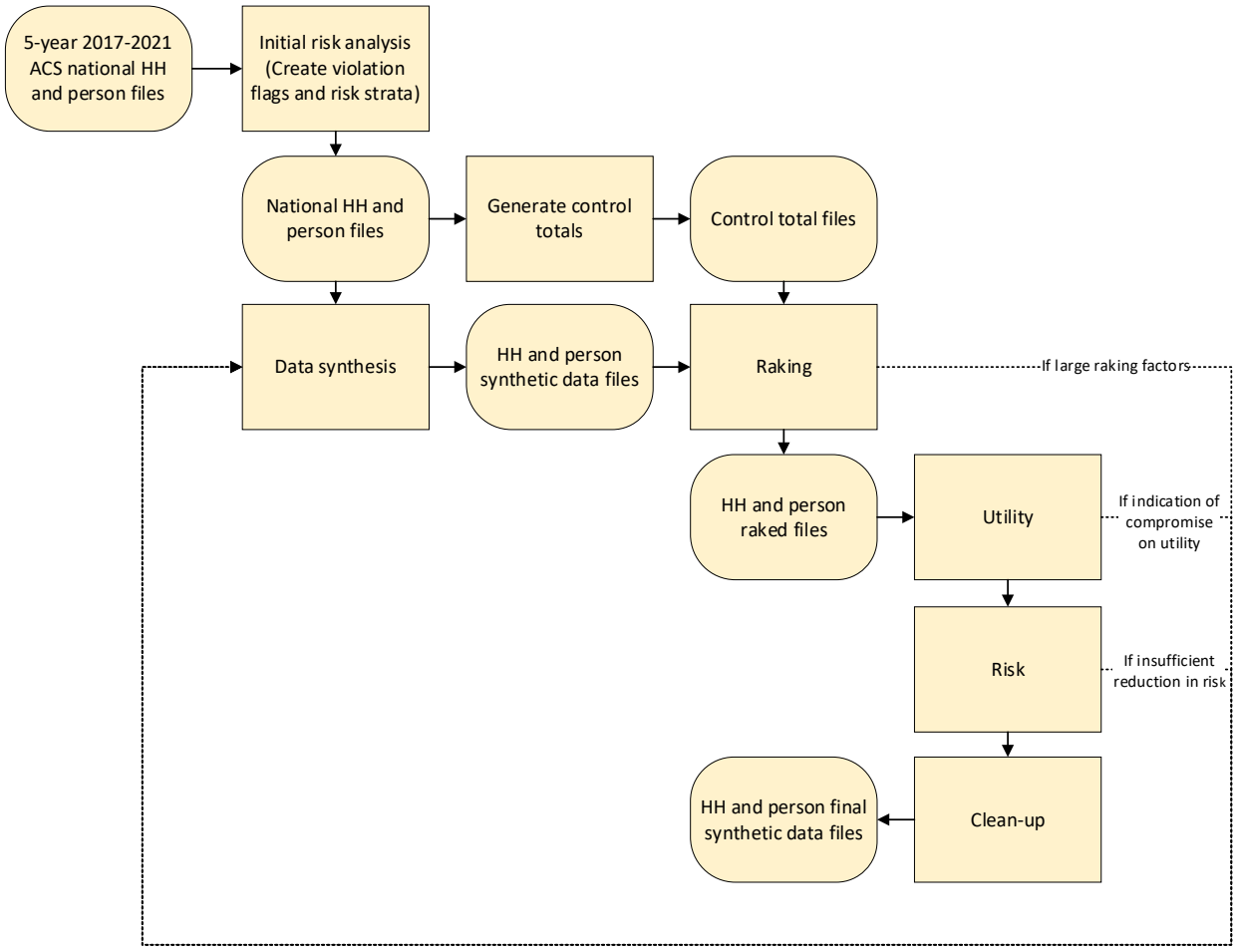

Figure 2-1 Overall Data Synthesis Process Flowchart

2.1 Initial Risk Analysis

The set of initial risk analysis modules were processed to generate tables that are subject to the process' disclosure rules. The tables were generated to flag data values that violate the rules and

therefore were at the highest risk of disclosure. Several preliminary steps were necessary within the initial risk analysis component to prepare for the application of the synthesis approach.

Several modifications to the component were necessary mainly in order to incorporate a much larger number of CTPP tables that are subject to cell suppression rules, especially in situations of non-standard variables and categories.

2.1.1 Initial Processing Steps

Firstly, we reviewed the planned list of CTPP tables to ensure that any changes (such as those mentioned in Section 1) were incorporated into the programs. In addition, the quality control checks implemented for the previous rounds of CTPP were reviewed. Initial processing commenced with test runs of the initial risk analysis component. The initial processing that was conducted on each input file provided key information about each variable. As in the 2012-2016 process, swapping flags were not used in the CTPP synthesis process, which deviates from the 2006-2010 process. This decision was made for the 2012-2016 CTPP after correspondence with the Census Bureau. The request for this information was withdrawn due to risk of impacting the timeline for DRB review of the synthesis process and the need to know about the swapping flags. The impact of avoiding such flagged values in the synthesis process is determined to be minimal. All processing was conducted on the Tabgen9 platform.

Also, we compute 2020 Census block estimates, e.g., percent of Black population, percent of Hispanic population, and percent of owner occupied households, using the 2020 geography reflected in the 2017-2021 ACS. The block estimates are key predictors in the model-assisted synthesis approach. For the purposes for processing the data synthesis, calibration and evaluation, recodes were generated for the ACS five-year files, as specified by the Census Bureau for the CTPP tables.

The technical details of the synthesis approaches are provided in Section 2 and the processing details are given in Section 3. Specifications were prepared and guidance provided to the Census Bureau to produce variances for the CTPP tables. More details about the variance estimation approach can be found in Section 2.5, as well as in Li, et al. (2011) and Krenzke, et al. (2017).

In the initial processing steps, several variables were created for use in the initial risk analysis and the processing of the approaches.

Distance. The distance between residence and workplace was computed at the Census block level as a predictor variable in the synthesis models. The GEODIST function in SAS 9.2 was used to calculate the block-to-block distance between a residence place and a workplace using the block level latitude and longitude as input. When workplace blocks were not available, distance was imputed by a nearest neighbor method with the cells formed by means of transportation and travel time.

ACS area-level covariates. At the Census tract level, the estimated statistics (percentages, means, or medians) were created and used as predictor variables in the synthesis models. The set of ACS area-level predictors is provided among the list in Appendix A.

Input data prep. This step was necessary to combine the outcomes of the prior processing steps. The output files from this step were a person-level file and household-level file. Other recodes were needed for the creation of the pool of predictor variables in the modeling approaches following specifications from the Census Bureau.

2.1.2 Processing the Initial Risk Analysis

The risk analysis was a major step processed on the national database, which involved processing frequencies to detect violations of the disclosure rules. The initial risk analysis was processed on the draft CTPP Tables Spreadsheet as provided by the Census Bureau in October 2023. ACS variables that had already been imputed during the ACS imputation process were not replaced; that is, they were considered to have already been synthesized. As part of the initial risk analysis, data values were classified according to risk strata.

The following flags were created to assist in the synthesis process as well as in the disclosure risk measures:

- **VarName_FLG**. This "violation" flag was set to one for a CTPP variable (referred to generically as *VarName*) if the associated data value was involved in a table that contributed to a violation of a disclosure rule.
- **VarName_RPL**. This "replacement" flag was set to one for a CTPP variable (referred to generically as *VarName*) regarding violations if the associated data value was involved

in a table cell that contributed to a violation of a disclosure rule and it was not already flagged as an imputed value.

- **VarName_STRT**. This flag was set to one for a CTPP variable (referred to generically as *VarName*) if the associated data value was involved in any singleton cell (cell with only one observation) that contributed to a violation of a disclosure rule; the flag was set to two if the associated data value was involved in a doubleton cell (cell with two observations) that contributed to a violation of a disclosure rule; the flag was set to 3 if the associated data values did not contribute to any violation of disclosure rules and was not already flagged as an imputed or ACS swapped value or associated with an allocated workplace; the flag was set to 4 if the associated data value was missing, was not subject to data synthesis, or already flagged as an imputed. This flag was useful in applying the partial replacement rates, as well as in the disclosure risk measure.

- **VarName_FULL**. This flag was set to one for a CTPP variable if the data needs to be replaced, without regard to violation flag.

The results of the initial risk assessment on the national sample identified data values at most risk of disclosure. It was conducted on five-year ACS data from 2017 to 2021. The analysis determined that over 87 percent of the Census tracts were affected by disclosure rules for at least one table. About 70 percent of records contributed to a violation of a disclosure rule, which is higher than the corresponding proportion for the ACS data from 2012 to 2016 due to a larger number of tables subject to disclosure rules despite no longer using TAZs in tabulations. In general, the risk is attributable to flows and cell means, due to the threat of an intruder linking tables together. Detailed categories in Means of Transportation (MOT) and certain other variables (e.g., in which cell means are computed) also contribute to the disclosure risk. Similar to what is shown in the discussion of the impact of TAZ sizes in Section 1.1.4 in the NCHRP 08-79 Final Report, small geography, such as Census tracts, has a large impact on the risk levels in the tables.

2.2 Data Synthesis

The synthesis module includes automated linking of recoded variables, overlapping bins, minimizing replacement with same data value, a mechanism to retain the unweighted distribution, and the use of model predictors to help form hot deck cells for the constrained hot deck.

When implementing the data synthesis approaches, there were a number of methodological challenges to address.

Variable Types. There are different types of variables among the ones to be synthetic (continuous, circular, ordinal categorical, and unordered categorical). This presented challenges in applying different synthesis approaches to different types of variables. The approach implemented was to use the unconstrained MACH approach (hereafter referred to as "semi-parametric") for unordered categorical variables and binary variables, and use the constrained MACH approach (hereafter referred to as "MACH") for ordered categorical variables with at least three levels. Rank linking was used for poverty. Additive noise was used for income if the synthetic value did not change.

Variable Versions. The same variable may have multiple versions, for example, categorical household income (HH_INC5, HH-INC9, HH_INC26), and continuous income. The approach implemented was to use the version with the most detailed categories (or continuous) in the modeling and map to the other versions.

Household (HH) and Person Level. Since the data included both HH and person-level data, a two-stage approach was employed. First the HH level variables (e.g., HH income) were synthesized and the values were transferred to each person within the HH. Next the person variables were synthesized.

Weights. The weights were quite variable, even within small areas, due to differential sampling rates, nonresponse follow-up sampling, and weighting adjustments. Therefore, the use of weights in the data synthesis process has the potential for reducing synthesis bias. Specifically, weights were used in the process of identifying donors for cases that need to be synthesized.

As a solution, the authors used a combined strategy with the MACH, semi-parametric, additive noise, and rank linking approaches. The processing flow for the data synthesis step is illustrated in Figure 3-3.

Prior to any data synthesis, any Census tract had fewer than 30 occupied households or fewer than 30 employed individuals aged 16 or older was collapsed with an adjacent Census tract. The process of combining tracts continued until the minimum counts were achieved. The Hilbert curve (Hilbert, 1891) was utilized to identify the adjacent tracts for this purpose. This approach was designed to ensure that data replacement occurred within neighboring tracts whenever feasible.

2.2.1 Processing Steps

The initial steps before processing the approaches involved assigning partial replacement flags, and running an extensive variable prep module. The set of data synthesis modules is driven by a Master Index File (MIF). The MIF identified the variables to be synthesized, the model areas, as well as the variables to be put into the pool of candidate predictor variables. It was used to classify the type of each variable as real numeric, ordered categorical, and unordered categorical. For the unordered categorical variables, indicator variables were created. Select interaction terms can be added to the pool of candidate predictor variables were identified as well (see Appendix A for the interaction terms included). Here are some key features relating to the MIF.

Setting the Target Selection Flags

The first step was to set the target selection flag (VarName_PARTIAL). It was set to one for a CTPP variable (referred to generically as *VarName*) if the associated data value was selected by a random process for synthesis within each risk stratum created for each variable during the initial risk analysis. As mentioned in Section 1, the partial replacement rates that was approved by the Census DRB for the 2006-2010 and 2012-2016 processes were increased for the 2017-2021 process to ensure at least 50 percent of the HH records were synthesized and 50 percent of the person records were synthesized.

Variable Prep

In the next step, referred to as 'variable prep', the predictor variables were recoded as necessary for the model selection step for the model-dependent synthesis approaches. The variable prep step also compiled the pool of predictor variables, the creation of indicator variables, and interaction terms for the predictor variables. The predictor pool was created from ACS and Census variables, including indicator variables for unordered categorical (UC) variables.

The predictor pools were divided into two groups:

- PredHous: Set of predictors for household-level models.
- PredPers: Set of predictors available for person-level models for persons in housing units and group quarters. For group quarters, the values of the household-level variables (such as vehicles available and household income) were set to zero so that they did not impact the person-level model selection and estimation process.

The MIF also identified variables to be forced into the models, called FORCELIST. These variables were forced in due to the explicit combinations of table variables in the set of CTPP tables or by

their involvement in flow tables because it was important to retain the correlation structure of the table results due to the large proportion of singletons and doubletons in flows, which essentially forms microdata.

Model Selection

Once the variable prep processing was completed, then the model selection approach was processed for all variables identified in the MIF that underwent the data synthesis. Model selection was processed for the purpose of identifying the predictors for each target variable, and to estimate the model parameters for generating predicted values, which were necessary for creating hot deck cells in the synthesis step. More details are given in Section 2.2.2.

Data Synthesis

One by one, the target variables were processed through the Main Loop. There are two main data synthesis approaches used, the semi-parametric approach and the MACH, depending on the type of variable. Both of the synthesis approaches are model-assisted in that they use the model parameters from the model selection process in order to generate predicted values to use in forming hot deck cells. First, household-level variables were synthesized, then the synthetic household variables were transferred to the person level, where the process continues with the synthesis on person-level variables. More details are provided in Section 2.2.2.

Post-Synthesis Processing

After processing, pre-post checks were conducted in order to have an initial look at the impact of the synthesis. Frequencies, means, and correlations were generated before and after synthesis. Lastly, recodes were processed in order to prepare for the raking step.

2.2.2 Details of the Synthesis Approaches

The semi-parametric and MACH procedures are model-assisted approaches that follow closely to Judkins et al. (2007). Initially designed for handling non-monotone (swiss cheese) missing data patterns in complex questionnaires, the Judkins et al. (2007) process in general uses model predictions to form hot deck cells. A donor for a case with a missing value is selected by a random draw without replacement within the hot deck cell, and the missing value is filled-in with the donor's original value. Influenced by the Gibbs sampler (an iterative method for simulating posterior distributions in Bayesian analysis through sampling from alternating conditional distributions until convergence in distribution is achieved), the imputation process is done variable-by-variable, using previously imputed data in the model selection and estimation process, as well as in the prediction

equation. The process proceeds sequentially through all variables needing imputation. Another cycle through all the variables receiving imputations is begun if the convergence criterion is not reached. The cycles after the first cycle use the completed data to form hot deck cells for the initially imputed variables. The Judkins et al. (2007) approach was adapted to replace observed data for the purpose of reducing disclosure risk. New features were added to the approach to handle highly variable weights and incorporate the small area geographic units to bring in features that may be special to that area.

The main procedure conducts the MACH, which was developed through research for the National Academies of Sciences (NCHRP 08-79 Final Report) and expanded through research conducted for the Census Bureau. The approach constrains the amount of change in the target variable by forming hot deck cells using "bins" created on the target variable itself (bins are recoded categories such that more than one published category was included in the bin). The objective of the MACH procedure is to change the value of the published categories by changing the value of the continuous version of the variable, but only by one or two categories, if possible. The basic steps are to select the target records for replacement and flag them, run models to attain predictions for the target, assign the bins, then form hot deck cells, and within each hot deck cell, a without replacement draw from the empirical distribution is used to choose the donor value. A nice feature of this approach is that it can control the amount of synthesis by changing the bin widths. Expanding or contracting the bin sizes allows the data producer the flexibility to control the distance between original and synthetic values.

Also available in the macro is a model-assisted unconstrained hot deck (semi-parametric approach), additive noise, and a rank linking approach.

The MACH has the following special features:

Use of model predictions as covariates. The MACH and semi-parametric approaches use coarsened model predictions to account for contributions from a pool of predictor variables.

Random assignment to overlapping bins. The MACH approach enables the user to form two sets of overlapping bins, and the modified algorithm would assign a set of bins at random to each record. This addresses a limitation of one set of bins formed on the target variable, which is that a target record with a value on the boundary of a bin can only have its value replaced by a lower value or an upper value, depending on if the original value is on the upper or lower boundary, respectively.

Limit replacement of the same value. A without replacement draw is conducted to address the issue that when the cell sizes became small, the procedure would be susceptible to replacing the data with the same values. Also, an automated collapsing routine combines small cells.

Link variables. As primary target variables are synthesized, so will others be replaced by the same donor to retain logical consistency.

Retain unweighted distribution. Some control over the unweighted one-way distributions is handled.

Ordering of hot deck cell variables. The order of hot deck cell variables may matter, and therefore the capability of ordering the cell variables is available as a parameter.

The MACH approach is relevant to ordinal variables with at least three levels.

The synthesis model can be expressed in general as follows:

$$\tilde{y}_{i(c)} = y_{i(c)} + \varepsilon_{i(c)},$$

Where, subscript (c) denotes the c[th] class (hot deck cell) defined from the set of factors $\{I(s), y_{g'}, \boldsymbol{x}, \hat{\mathbf{y}}_{\boldsymbol{g}''}, \boldsymbol{w}_{\boldsymbol{g}'''}\}$, where $I(s)$ is the set of indicators for being selected for synthesis, $y_{g'}$ denotes g' bins formed on the target variable y, $\boldsymbol{x}$ are the auxiliary variables, $\hat{\mathbf{y}}_{\boldsymbol{g}''}$ are the g'' groups formed from model predictions, $\boldsymbol{w}_{\boldsymbol{g}'''}$ are the g''' groups formed from the sampling weights and where $\varepsilon_{i(c)} \sim \tilde{y}_{i(c)} - y_{i(c)}$ resulting from the random error associated with case i for a random with-replacement draw within the c^{th} class. The bolding pattern represents vectors.

The main steps of the synthesis process were to:

1. Select the model and estimate its parameters (Section 2.2.2.1)
2. Form hot deck cells (Section 2.2.2.2)
3. Synthesize the data within each hot deck cell, by taking a without replacement draw from the empirical distribution. The donor's value was used to replace the target record's value. (Section 2.2.2.2)

Each step is explained in detail below.

To facilitate the discussion of the synthesis approach that follows, a subset of the variables to be synthetic was identified (Table 2-1). The table highlights the level (HH, person), the variable type (OC, UC), and the approach used. A couple of "spinoff" approaches other than the semi-parametric (SP) and MACH, namely additive noise (AN) and rank linking (RL), were implemented in a limited way (described below) to add to the protection from disclosure. The variables that share the same run number were linked.

Table 2-1. Synthesized Variables and Synthetic Data Approaches

Run	Item	Variable Name & Description	Variable Level	Type	Approach
1	Item1	AHINC (income)	HH	OC	MACH+AN
2	Item2	HHLDRAGE (householder age)	HH	OC	MACH
3	Item3	HUPAOC (age of own kids)	HH	UC	SP
4	Item4	VEH (# vehicles)	HH	OC	MACH
5	Item5	ACCESS (access to internet)	HH	OC	SP
5	Item6	BROADBND (cell data plan)	HH	OC	SP
5	Item7	DIALUP (dial up service)	HH	OC	SP
5	Item8	HISPEED (high speed internet)	HH	OC	SP
5	Item9	OTHSVCEX (other internet)	HH	OC	SP
5	Item10	SATELLITE	HH	OC	SP
6	Item11	BLD (building type)	HH	UC	SP
7	Item12	TEN (tenure status)	HH	UC	SP
8	Item13	AGE	Person	OC	MACH
9	Item14	APERN (earnings)	Person	OC	MACH
10	Item15	JWTJWNSR (means of transportation)	Person	UC	SP
11	Item16	JWRI (total rides)	Person	OC	MACH
12	Item17	JWMN (minutes to work)	Person	OC	MACH
12	Item18	JWD (time of departure)	Person	OC	MACH
12	Item19	JWA (time of arrival)	Person	OC	MACH
13	Item20	IND (industry)	Person	UC	SP
13	Item21	COW (class of worker)	Person	UC	SP
13	Item22	OCC (occupation)	Person	UC	SP
14	Item23	WKH (hours worked / week)	Person	OC	MACH
15	Item24	SCHL (education attainment)	Person	OC	MACH
15	Item25	SCHG (grade attending)	Person	OC	MACH
15	Item26	SCH (school enrollment)	Person	UC	SP
16	Item27	TOTRACE (race)	Person	UC	SP
16	Item28	HSGP (Hispanic group)	Person	UC	SP
16	Item29	LAN (language spoken)	Person	UC	SP
17	Item30	DIS (disability status)	Person	OC	SP
18	Item31	SEX	Person	OC	SP
19	Item32	POVPI (poverty index)	Person	OC	RL

2.2.2.1 Model Selection and Estimation

The model selection and estimation step was done once for each CTPP variable to be synthesized using the raw data from the ACS; that is, there was no need to re-estimate the model for each variable as vectors of variables were replaced with synthetic data since the joint distribution among the variables is already given, conditional on the fully complete ACS reported, imputed, and swapped data.

The modeling step was done separately at the household level and at the person level. Each variable to be synthesized went through the model selection step. The model selection process occurred for each model area. For person-level processing for most target variables, the model areas were residence-based counties. For industry, occupation, and class of worker, the model areas were workplace based instead of residence based. For commute related variables (For household-level processing, records were modeled separately by county.

The modeling was done differently for variables of type OC (ordered categorical, binary) than for type UC variables. For OC variables, a stepwise linear regression was processed, and the model selection forced all variables into the model that occurred with the dependent variable in any of the CTPP tables, while bringing in other significant predictors to improve the predictive power of the model. A clustering procedure was done for UC variables, which fit a separate linear regression for each category of the variable, and subsequently conducted a k-means clustering algorithm on the vector of predicted values for each level. The algorithm was run to produce g clusters to be used in the hot deck cell formation.

Aligning with the list of variables in Table 2-1, let y_{ji} denote the j-th variable to be synthesized for record i, where j is the item number in Table 2-1, and y represents the American Community Survey (ACS) data values. The subscript k identifies indicator variables associated with the k-th category of UC variables. The bolding pattern represents vectors. For illustration purposes, the model selection for OC continuous (variable 2 in Table 2-1), OC binary (variable 6 in Table 2-1) and UC variables (variable 12 in Table 2-1) is essentially as follows:

$$E(y_2|y_1, y_3, \dots, y_{13}, X) = f(y_1, y_3, \dots, y_{13}, X, \boldsymbol{\beta}),$$
$$E(y_5|y_1, \dots, y_4, y_6, \dots, y_{13}, X) = f(y_1, \dots, y_4, y_6, \dots, y_{13}, X, \boldsymbol{\beta}),$$
$$E(y_{11k}|y_1, \dots, y_{10}, y_{12}, X) = f(y_1, \dots, y_{10}, y_{12}, X, \boldsymbol{\beta}),$$
for $k = 1,2, \dots, K$ for building categories

The models were processed to allow predictors to enter the model during the stepwise modeling steps if significant at the $\alpha = .05$ level. Predictors not significant at the .05 level exited the model. The set of variables we refer to as FORCELIST, were forced into the model for two reasons: (1) the variables were explicit combinations of table variables in the set of CTPP tables, or (2) the variables were involved in flow tables. It was important to retain the correlation structure of the table results due to the large proportion of singletons and doubletons in flows, which essentially forms microdata. All models included indicators for the 10 category means of transportation (MOT). The remainder of the FORCELIST variables differed for each variable, as given below in Table 2-2. Within the candidate predictor pools were select interactions with the MOT indicators. The MOT-variable interactions included interactions with household income, number of workers in HH, presence of children, age, minority status, sex, earning, and number of vehicles available, country of birth, travel time, distance, and poverty status. The list of candidate predictors is given in Appendix A.

Table 2-2. FORCELIST Variables for Each Dependent Variable

Dependent variable	FORCELIST
Item3	HH MOT indicators, household income, and number of workers in the household
Item20,21,22	MOT indicators, age, work-shift indicators, travel time, household income, and poverty status, number of vehicles in household

2.2.2.2 Formation of Hot Deck Cells and Synthesizing Data Values

Variables were synthesized, one variable at a time, beginning with the household level, transferring the synthetic household variables to the person level, and then continuing with the synthesis on person-level variables. Hot deck cells were used as part of the synthesis process, and were formed using the following information:

0. The target selection flag to retain the unweighted empirical distribution,
1. The bins on the target record in order to control the amount of change (MACH only),
2. Key coarsened variables other than the target variable,
3. Locality,
4. Groups of sequential predictions from predictive models, and

5. Coarsened values of the sample weights. Independently, the groups of weights were created from a ranking of the weights with an equal number of sampled cases within each group.

When forming hot deck cells, small cells were identified and combined in an automated manner. The six components of the hot deck cells were sorted in a serpentine manner, which means the first hot deck variable was sorted in ascending order, the second variable in ascending order in the first level of the first variable, then descending order for the second level of the first variable, then ascending order for the third level of the first variable, and so on. Each of the remaining variables were done in the same manner. If the number of respondents in a cell was less than a pre-assigned threshold, the cell was collapsed with the prior adjacent cell. The rank order of above contributing sources can be applied by assigning values to the column RankOrderHD on the MIF. For example, RankOrderHD has values 1, 3, 4, 2, 5 for Item1, with 1 referring to bins, 3 referring to locality, 4 referring to prediction groups, 2 referring to hot deck cells, and 5 referring to weight groups. Table 2-3 provides the bin variables, the hot deck variables, locality, the number of prediction and weight groups for each synthetic variable, as well as the rank order of contributing sources to form hot deck cells.

Table 2-3. Components of Hot Deck Cells for Each Variable Synthesized

Variable Description	1. Bins	2. Hot deck variables	3. Locality	4. Number of prediction groups	5. Number of weight cells	Rank order of HD sources[1]
Item1	BIN1	VEH	Residence Census tract	4	3	1 3 4 5 2
Item2	BIN2		Residence Census tract	4	3	1 3 4 5 2
Item3	BIN3	NP	Residence Census tract	4	3	2 1 3 4 5
Item4	BIN4	NP WIH	Residence Census tract	4	3	1 2 3 4 5
Item5	NA		Residence Census tract	4	3	4 3 5 2
Item6	NA		Residence Census tract	4	3	4 3 5 2
Item7	NA		Residence Census tract	4	3	4 3 5 2
Item8	NA		Residence Census tract	4	3	4 3 5 2
Item9	NA		Residence Census tract	4	3	4 3 5 2
Item10	NA		Residence Census tract	4	3	4 3 5 2
Item11	NA		Residence Census tract	4	3	4 3 5 2
Item12	NA		Residence Census tract	4	3	4 3 5 2
Item13	BIN5	NP HHT	Residence Census tract	4	3	1 2 3 4 5
Item14	BIN6	HH_INC8 ESR	Residence Census tract	4	3	1 4 3 5 2
Item15	NA	JWRI TRAVEL_TM4	Residence Census tract	4	3	2 3 4 5
Item16	NA	MEANS10	Residence Census tract	4	3	1 2 3 4 5
Item17	BIN9	MEANS10	Residence Census tract	4	3	1 2 3 4 5
Item18	BIN10	MEANS10	Residence Census tract	4	3	1 2 3 4 5
Item19	BIN11	MEANS10	Residence Census tract	4	3	1 2 3 4 5
Item20	NA	ESR2	Workplace Census tract	4	3	1 2 3 4 5
Item21	NA	ESR2	Residence Census tract	4	3	1 2 4 3 5

Variable Description	1. Bins	2. Hot deck variables	3. Locality	4. Number of prediction groups	5. Number of weight cells	Rank order of HD sources[1]
Item22	NA	ESR2	Residence Census tract	4	3	1 2 4 3 5
Item23	BIN12	ESR2 WKL2	Residence Census tract	4	3	1 2 4 3 5
Item24	NA	AGE6 FOD1_2	Workplace Census tract	4	3	3 4 2 5
Item25	NA	AGE6 FOD1_2	Workplace Census tract	4	3	3 4 2 5
Item26	NA	AGE6 FOD1_2	Workplace Census tract	4	3	3 4 2 5
Item27	NA	LANX2	Residence Census tract	4	3	3 4 5 2
Item28	NA	LANX2	Residence Census tract	4	3	3 4 5 2
Item29	NA	LANX2	Residence Census tract	4	3	3 4 5 2
Item30	NA	AGE65PLUS	Residence Census tract	4	3	3 4 2 5
Item31	NA		Residence Census tract	4	3	3 4 5 2
Item32	NA	MISSPOVERTY ACSHH_WRK6 VEHICLES6	Residence Census tract	NA	NA	NA

If the target flag is used to retain the unweighted empirical distribution and the rank order is (1 2 3 4 5), the automated collapsing process begins by collapsing neighboring weight groups until the sample size for the resulting cells exceeds the threshold. If this is not achieved, collapsing continues in a similar manner over prediction groups, locality, hot deck variables, and bins until each cell has a sufficiently large sample size.

Within each final hot deck cell, a without replacement draw from the empirical distribution was conducted. To do so, the target records were identified by their partial replacement flag. The replaced value was obtained through a random draw without replacement from the empirical distribution within the hot deck cell among those targeted for replacement; that is, all records targeted for replacement were used to donate their values to others. All records not targeted for replacement were ineligible to donate their values. This approach retained the overall empirical distribution of the target variable.

The predictions and the subsequent draws from an empirical distribution occurred in a sequential manner so that synthetic values were used for the predictor variables in the model for the next variable to be synthesized. The process ran sequentially until all items to be synthesized. One cycle through the variables was conducted. The following describes the hot deck cell formation for each variable. For each value of an item needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value. Additive noise for variable 1 and rank linking for variable 9 also are described.

Item 1. The synthesis process began with the MACH approach applied for replacing values of Item 1. Among all records where item 1 was targeted for replacement, the hot deck cells were formed by Item 1 bins* locality (Census tract) * four prediction groups* vehicles available * two weight groups.

Additive Noise for Item 1. Next, the additive noise procedure was conducted on any target record where Item 1 did not change value; that is, during the synthesis step, if left unchanged from the MACH procedure, noise was added to the original Item 1 value y as follows:

$$\tilde{y}_{1i} = y_{1i}(1 + fz),$$

where f is a constant between 0 and 1, and z is a draw from the standard normal distribution. The noise was centered at 0 with a draw from the standard normal distribution. The standard deviation of the added noise was the product of f and y_{1i} which means the level of noise was allowed to vary relative to the magnitude of Item 1.

Item 2. Among all records where the Item 2 was targeted for synthesis by the MACH approach, the hot deck cells were formed by Item 2 bins * locality (Census tract) * four prediction groups * three weight groups. Items 3, 4, 13, 14, 17, 18 and 19 were synthesized in a similar manner but with different hotdeck cell variables and rank orders. It should be noted that Item 13 was edited to Item 2 under certain condition.

Item 5. The synthesis process for Item 3 used the semi-parametric approach, including the clustering approach described in Section 2.2.2.2. Among all records where Item 5 was targeted for replacement, the hot deck cells were formed by four prediction groups* locality (Census tract) * three weight groups. Item 11 was synthesized in a similar manner but with different hotdeck cell variables and rank orders.

Item 12. The synthesis process for Item 12 used the semi-parametric approach. The clustering approach described in Section 2.2.2.2 was not needed because Item 12 was binary. Among all records where Item 12 was targeted for replacement, the hot deck cells were formed by four prediction groups* locality (Census tract) * three weight groups. Items 30 and 31 were synthesized in a similar manner but with different hotdeck cell variables and rank orders.

Item 32. A variation of the hot deck (we refer to internally as rank linking) was developed to link Item 1 with Item 32 together. Once the Item 1 was synthesized as described above, the ACS and the synthetic Item 1 were attached to the person level file.

To synthesize Item 33, we created a file called RAW with residence Census tract identifiers, the number of workers in the household, vehicles available, ACS Item 1 and Item 33. We sorted RAW by a missing value indicator on Item 33, number of workers in the household, vehicles available, residence Census tract, and ACS Item 1. The synthetic Item 1 resides on main data file. The synthetic data file is then sorted by a missing value indicator for Item 33, number of workers in the household, vehicles available, residence Census tract and synthetic Item 1. Then, the Item 33 from the RAW file was joined (merged) with the main data file. The Item 33 from the RAW file was used for Item 33 if flagged for replacement.

Below are descriptions for the rest of items that required special treatments in data synthesis.

Item 15. Similar to Item 11, the synthesis process for Item 15 used the semi-parametric approach, including the clustering approach described in Section 2.2.2.2. Among all records where Item 15 was targeted for replacement, the hot deck cells were formed by total rides * travel time (4 groups) * locality (Census tract) * four prediction groups* three weight groups. As only a subset of categories (newly introduced and deviating from the ACS standardized categorization) were subject to data synthesis, the risk stratum 4 was assigned to records with the other categories in order to exclude these records from the process.

Items 5 through 10. Items 5 through 10 were linked in the process as follows using the semi-parametric approach. Among all records where Item 5 was targeted for replacement, the hot deck cells for Item 5 were formed by four prediction groups for Item 5 * locality (Census tract) * three weight groups. For each value of Item 6 needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value for Item 5. Items 5 through 10 were replaced together from the same donor.

Items 17 through 19. Items 17 through 19 were linked in the process as follows using the MACH approach. Among all records where Item 17 was targeted for replacement, the hot deck cells for Item 17 were formed by bins for Item 17 * means of transportation (10 groups) * bins for Item 19 * bins for Item 19 * locality (Census tract) * four prediction groups for Item 17 * three weight groups. For each value of Item 17 needing replacement, a random draw without replacement was conducted

within the hot deck cell for the target data value for Item 17. Items 17 through 19 were replaced together from the same donor.

Items 20 through 22. Items 20 through 22 were linked in the process as follows using the MACH approach. Among all records where Item 20 was targeted for replacement, the hot deck cells for Item 20 were formed by bins for Item 20 * age (6 groups) * bins for Item 22 * four prediction groups for Item 22 * locality (Census tract) * three weight groups. For each value of Item 20 needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value for Item 20. Items 20 through 22 were replaced together from the same donor.

Items 24 through 26. Items 24 through 26 were linked in the process as follows using the semi-parametric approach. Among all records where Item 24 was targeted for replacement, the hot deck cells for Item 24 were formed by locality (workplace Census tract) * four prediction groups for Item 24 * household income (8 groups) * three weight groups. For each value of Item 24 needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value for Item 21. Items 24 through 26 were replaced together from the same donor.

Items 27 through 29. Items 27 through 29 were linked in the process as follows using the semi-parametric approach. Among all records where Item 27 was targeted for replacement, the hot deck cells for Item 27 were formed by locality (Census tract) * four prediction groups for Item 27 * three weight groups. For each value of Item 27 needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value for Item 27. Items 27 through 29 were replaced together from the same donor.

2.2.2.3 Details on Bin Formation and Prediction Groups

Details on the formation of bins and the predictions groups are provided in this section.

Bin Formation

The formation of 'bins' applies only to variables synthesized through the MACH approach. The hypothetical example in Figure 2-2 illustrates the assignment of bins. The figure depicts a frequency distribution, with spikes at multiples of 5. Within Figure 2-2, below the histogram, the rows illustrate

two sets of overlapping bins (BinA and BinB) and published categories for the y variable. The bins are formed while striving to achieve the following objectives:

1. To ensure that the bins contain more than one value of the published categories.

2. To ensure that if there are spikes, then at least two spikes are included in a bin; otherwise, the approach results in values unchanged for many cases.

Using these bins allows changes in corresponding table variables while controlling the range of potential changes.

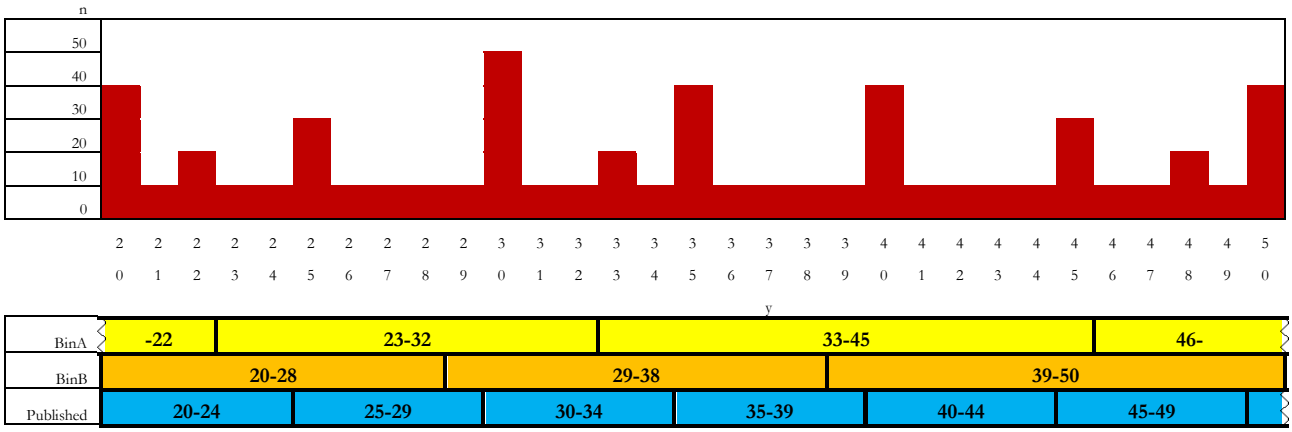

Figure 2-2. Illustration of Bin Formation

Prior to forming the hot deck cells, each record was randomly assigned with one-half chance to either BinA or BinB. This in effect split the sample in half, where one-half used set BinA and one-half used set BinB.

Prediction Group Formation

After the model parameters were estimated for all variables, and after the bin formation occurred (MACH approach only), then the sequential prediction steps occurred that led to the formation of prediction groups from prediction models. After predictions for a target variable were generated, the groups were formed from a ranking of the predictions, with an equal number of sampled cases within each group.

The general sequential process was that for each variable, a prediction equation was created from the estimated regression parameters and predictions were computed using either ACS or synthetic data if already available.

The sequential prediction and synthesis steps are described using the variables in Table 2-1 in the following example. As an example, the second variable in Table 2-1 was an OC continuous variable and was being processed by the MACH approach. The prediction equation for OC variable 2 (y_2) in Table 2-1 is given as follows (ignoring interaction terms for simplicity), using the synthetic values for variable 1, and the ACS values for the remaining items:

$$\hat{y}_{2i} = \beta_0 + \beta_1 \tilde{y}_{1i} + \beta_3 y_{3i} + \beta_4 y_{4i} + \sum_{j=5}^{13} \sum_{k=1}^{K_j} \beta_{jk} y_{jki} + \sum_{l=1}^{L} \beta_l x_{li}$$

Where K_j is the number of categories for variable $j3$, and L is the number of other predictor variables.

Then subsequently, as discussed above, seven prediction groups were formed on $\hat{y}_{2i}$. Let $\tilde{y}_{2i}$ represent the synthetic value drawn at random without replacement within the hot deck cells.

The values were synthesized at this time only if they were flagged for replacement (or VarName_PARTIAL=1). After each variable was synthesized, the interaction terms were recreated using synthetic values so synthetic values could be used in the prediction equation for the next dependent variables in the sequence.

Continuing sequentially through the variables in Table 2-1, we get to variable 6, which was an OC binary variable and was handled in a similar fashion as variable 2. Let the prediction equation for variable 6 be represented as follows, using the synthetic values for the previous 5 variables:

$$\hat{y}_{6i} = \beta_0 + \sum_{j=1}^{4} \beta_j \tilde{y}_{ji} + \sum_{k=1}^{K_5} \beta_{jk} \tilde{y}_{jki} + \sum_{j=7}^{13} \sum_{k=1}^{K_j} \beta_{jk} y_{jki} + \sum_{l=1}^{L} \beta_l x_{li}$$

Next, for the eighth variable, there were K_j categories in this UC variable from which K_j corresponding indicator variables were formed. Let the prediction equation for the l-th category of UC variable 12 be represented as follows, using the synthetic values for the previous 11 variables:

$$\hat{y}_{12li} = \beta_{0l} + \sum_{j=1}^{4} \beta_j \tilde{y}_{ji} + \sum_{j=5}^{11} \sum_{k=1}^{K_j} \beta_{jk} \tilde{y}_{jki} + \sum_{k=1}^{K_{13}} \beta_{13k} y_{13ki} + \sum_{l=1}^{L} \beta_l x_{li}$$

where $l = 1,2,\ldots,K_j$.

For the UC variable, a clustering program (SAS Proc FastClus) was used to form five clusters (prediction groups), using the K_{12} sets of predicted values $\hat{y}_{12li}$. Then, three groups were formed on the weights. Let $\tilde{y}_{12i}$ represent the synthetic value drawn within the hot deck cell. In general, after a UC variable was synthetic, indicator variables were re-created using the synthetic values.

2.3 Weight Calibration

After the data synthesis approaches were processed, the weight adjustment step, known as raking, was done so that the weights are calibrated to reproduce select ACS estimates at the following geography levels:

- Public Use Microdata Area (PUMA) level, which are areas formed to be greater than 100,000 in population for the purpose of releasing public use microdata.
- Census tract level, which are areas of about 4,000 in population;
- County level; and
- State level.

The raking procedure is commonly called iterative poststratification or calibration. In its simplest form, poststratification adjusts weights so that the weighted sample distribution for some categorical variable is the same as a known population distribution for that same variable (or a distribution based on a sample with a lower mean square error). As a result, the sums of the poststratified weights will be consistent with control totals for select subgroups of the population (i.e., the subgroups defined by the categorical variable).

Poststratification involves one dimension of population subgroups; for example, gender is one dimension with two subgroups (male, female). A dimension can be formed by combining two variables, such as, gender by MOT subgroups, which form a dimension with mutually exclusive subgroups, such as females who are bikers/walkers, or who ride in carpools, drive alone, take public transportation, and so forth, and also with males in the same MOT subgroups. Since it was desired to use several variables in the adjustment, the sample sizes associated with the resulting subgroup

categories from combining the variables were small. The solution was to create several dimensions, and apply the poststratification procedure iteratively. The process began by first postratifying using the first dimension, then using the first iteration's adjusted weights, poststratifying to the second dimension, and continuing until the maximum difference (between the sum of adjusted weights and the control totals) for each subgroup for each dimension was less than some predetermined value. The raking procedure was introduced by Deming and Stephan (1940) and more discussion can be found in Oh and Scheuren (1987).

There were two sets of input files, one for creating control totals, and one for the raking adjustment. The input files for creating control totals are the microdata output files from the initial risk analysis. The input files for adjusting the weights using raking are from the data replacement component.

The creation of control totals and the raking adjustment were both done independently at the household-level and then the person-level. Table 2-4 provides the raking dimensions at the household level. Cross-tabulations for each dimension were generated and the combinations that have less than 50 records were listed. Some combining of categories was necessary before the raking macro was processed.

Table 2-4. CTPP Production: Raking Dimensions for the Household File

Dimension	ByVar1	ByVar2
1	PUMA^	Vehicles available (5)^
2	PUMA^	Number of workers in HH (5)^
3	PUMA^	HH income (8)^
4	PUMA^	Number of children in HH (2)^
5	Census tract^	

Note: A '^' means that some combining of categories occurred.

After processing the household-level raking, the person-level raking was processed. Table 2-5 provides the raking dimensions at the person level. Similar to the household dimensions, cross-tabulations for each dimension were generated and some combining of categories was necessary before the raking macro was processed.

Table 2-5. CTPP Production: Raking Dimensions for the Person File

Dimension	ByVar1	ByVar2
1	PUMA^	Vehicles available (5)^
2	PUMA^	Number of workers in HH (5)^
3	PUMA^	HH income (8)^
4	PUMA^	Number of children in HH (2)^

5	PUMA^	Travel time (10)^
6	PUMA^	Time leaving home (13)^
7	PUMA^	Age of worker (6)^
8	PUMA^	Poverty status (4)^
9	PUMA^	Minority status (2)^
10	PUMA^	MOT(4)^
11	Place of work state^	Industry (4)^
12	Place of work state^	MOT(6)
13	County	Whether work and live in the same county (2)
14	Census tract^	
15	Place of work census tract^	

Note: A '^' means that some combining of categories occurred.

Table 2-6 provides percentiles of the raking adjustment factors for the household-level and person-level raking. Focusing on the range between the 10th and 90th percentiles, the range was considered very small.

Table 2-6. Percentiles of the Raking Factors

Level	1st	5th	10th	50th	90th	95th	99th
Household	0.90	0.95	0.99	1.00	1.01	1.03	1.11
Person	0.87	0.95	0.98	1.00	1.01	1.04	1.15

2.4 Data Utility and Disclosure Risk Measures

Gomatam and Karr (2003) and Gomatam et al. (2003, 2004), for example, have examined utility and risk in the case of data swapping. Oganian and Karr (2006) examined combining methods that perturb data for statistical disclosure control. They found that greater protection and utility can be achieved in some cases by utilizing two or more methods in less intensity than a single method. In summary, there were numerous options that could have been considered, but all have limitations and performance likely depended on the specific application. The CTPP data utility measures are discussed in Section 2.4.1, and the disclosure risk measures are discussed in Section 2.4.2.

2.4.1 Data Utility Measures

The data synthesis approaches for the CTPP production were designed to limit the impact on data utility while reducing the risk of disclosure. It is important to develop measures for the resulting data utility so that the balance between risk and utility can be understood for the CTPP tables (Drechsler and Reiter 2009; Karr et al., 2006; Duncan, Keller-McNulty et al., 2001).

The focus of the utility checks was to compare the ACS data with the synthetic data. The comparisons checked cell means, cell percentiles (medians and 75th percentiles), weighted cell counts, standard errors, Cramer's V for associations in two-way tables, pairwise associations, and multivariate associations at the county level. The median of differences between the raw and synthetic estimates (across estimates for geographic areas) were computed where appropriate in order to give indications of potential bias introduced by the synthesis. The interquartile range for the differences provided an indication of the variation caused by the synthesis. A few crosstabs for a subset of geographic areas were reviewed closely to check whether the tables based on the CTPP synthetic data are aligned with those using the ACS data.

Cell Mean Differences and Quantile Differences

Shlomo (2008) suggested computing average absolute difference in cell counts for a given variable. The research team adapted this approach for computing the difference in cell means as denoted as follows:

$$D_{\bar{y}} = \tilde{y} - \bar{y}$$

where $\tilde{y}$ = synthesized mean from the CTPP synthetic data
$\bar{y}$ = estimated mean from the ACS data

The ratio of the difference to the standard error from the ACS data was also examined. Cell mean differences were produced for tract-level and county-level residences. The differences were computed for two attributes (travel time and household income). The mean travel times were computed for two levels of time leaving home, and four levels of MOT. Mean household income was computed for five levels of vehicles available. The levels of each "by variable" are defined as follows:

VEHICLES6_2 = 0 vehicles available
VEHICLES6_3 = 1 vehicle available
VEHICLES6_4 = 2 vehicles available
VEHICLES6_5 = 3 vehicles available
VEHICLES6_6 = 4 or more vehicles available

MEANS6_2 = car, truck, or van – Drove alone

MEANS6_3 = car, truck, or van – in a two-person carpool
MEANS6_4 = car, truck, or van – in a three or more person carpool
MEANS6_5 = car, truck, or van – Public transportation, bicycle, walked, taxicab, motorcycle, or other method

TM_LEAVE5_3 = time leaving home 5:00 a.m. to 8:59 a.m.
TM_LEAVE5_4 = time leaving home 9:00 a.m. to 4:59 a.m.

Tables 2-7 and 2-8 provide the median and interquartile range (IQR) of differences for travel time and household income between cell means from ACS data and cell means from synthesized data generated by the model assisted constrained hot deck approach. The differences for all counties are summarized across the entire dataset. The tables also provide the median and interquartile range of differences for a few other variables at the county flow level, among which industry and minority were synthesized by the semi-parametric approach and poverty was synthesized by the rank linking approach (linked with household income).

Table 2-7. Median and Interquartile Range of Cell Mean Differences

Attribute	BYVAR	Geographical Level	Median	IQR	Avg Ratio of Diff to SE
JWMN	MEANS6_2	Workplace county	0.00	0.1	0.035
JWMN	MEANS6_3	Workplace county	0.00	0.2	0.024
JWMN	MEANS6_4	Workplace county	0.00	0.4	0.023
JWMN	MEANS6_5	Workplace county	0.00	0.3	0.032
AHINC	VEHICLES6_2	Residence county	2.46	133.1	0.024
AHINC	VEHICLES6_3	Residence county	4.00	97.9	0.031
AHINC	VEHICLES6_4	Residence county	2.85	449.9	0.025
AHINC	VEHICLES6_5	Residence county	5.55	645.2	0.033
AHINC	VEHICLES6_6	Residence county	-16.89	112.4	0.036
JWMN	MEANS6_2	Residence county	0.00	0.1	0.017
JWMN	MEANS6_3	Residence county	0.00	0.4	0.036
JWMN	MEANS6_4	Residence county	0.01	0.6	0.021
JWMN	MEANS6_5	Residence county	0.01	0.4	0.023
JWMN	TM_LEAVE5_3	Residence county	0.01	0.3	0.024
JWMN	TM_LEAVE5_4	Residence county	0.00	0.4	0.021
AGE9		County flow	0.00	0.0	0.030
INDUSTRY		County flow	0.00	0.0	0.033
JWD		County flow	0.00	0.8	0.015
JWMN		County flow	0.00	0.0	0.021
MINORITY		County flow	0.00	0.0	0.016
POVERTY		County flow	0.00	0.0	0.017

Table 2-7. Median and Interquartile Range of Cell Mean Differences

Attribute	BYVAR	Geographical Level	Median	IQR	Avg Ratio of Diff to SE
AHINC	VEHICLES6_2	Residence tract	1.29	25.4	0.022
AHINC	VEHICLES6_3	Residence tract	2.58	376.3	0.016
AHINC	VEHICLES6_4	Residence tract	1.56	876.8	0.026
AHINC	VEHICLES6_5	Residence tract	5.57	848.0	0.022
AHINC	VEHICLES6_6	Residence tract	4.81	463.1	0.036
JWMN	MEANS6_2	Residence tract	0.00	0.7	0.018
JWMN	MEANS6_3	Residence tract	0.00	0.3	0.018
JWMN	MEANS6_4	Residence tract	0.00	0.1	0.016
JWMN	MEANS6_5	Residence tract	0.00	0.3	0.039
JWMN	TM_LEAVE5_3	Residence tract	0.00	0.6	0.034
JWMN	TM_LEAVE5_4	Residence tract	0.00	1.4	0.043
JWMN	MEANS6_2	Workplace tract	0.00	0.7	0.037
JWMN	MEANS6_3	Workplace tract	0.00	0.1	0.041
JWMN	MEANS6_4	Workplace tract	0.00	0.0	0.017
JWMN	MEANS6_5	Workplace tract	0.00	0.2	0.043

Table 2-8. Median and Interquartile Range of Cell Quantile Differences

			Cell Median		Cell 75th Percentiles	
Attribute	BYVAR	Geographical Level	Median	IQR	Median	IQR
JWMN	MEANS6_2	Workplace county	0	0	0	0
JWMN	MEANS6_3	Workplace county	0	0	0	0
JWMN	MEANS6_4	Workplace county	0	0	0	0
JWMN	MEANS6_5	Workplace county	0	0	0	0
AHINC	VEHICLES6_2	Residence county	0	0	0	0
AHINC	VEHICLES6_3	Residence county	0	0	0	0
AHINC	VEHICLES6_4	Residence county	0	6	0	41.5
AHINC	VEHICLES6_5	Residence county	0	0	0	0
AHINC	VEHICLES6_6	Residence county	0	0	0	0
JWMN	MEANS6_2	Residence county	0	0	0	0
JWMN	MEANS6_3	Residence county	0	0	0	0
JWMN	MEANS6_4	Residence county	0	0	0	0
JWMN	MEANS6_5	Residence county	0	0	0	0
JWMN	TM_LEAVE5_3	Residence county	0	0	0	0
JWMN	TM_LEAVE5_4	Residence county	0	0	0	0
AGE9		County flow	0	0	0	0
JWMN		County flow	0	0	0	0
JWD		County flow	0	0	0	0
AHINC	VEHICLES6_2	Residence tract	0	0	0	0
AHINC	VEHICLES6_3	Residence tract	0	0	0	0
AHINC	VEHICLES6_4	Residence tract	0	0	0	0
AHINC	VEHICLES6_5	Residence tract	0	0	0	0
AHINC	VEHICLES6_6	Residence tract	0	0	0	0
JWMN	MEANS6_2	Residence tract	0	0	0	0

Table 2-8. Median and Interquartile Range of Cell Quantile Differences

			Cell Median		Cell 75th Percentiles	
Attribute	BYVAR	Geographical Level	Median	IQR	Median	IQR
JWMN	MEANS6_3	Residence tract	0	0	0	0
JWMN	MEANS6_4	Residence tract	0	0	0	0
JWMN	MEANS6_5	Residence tract	0	0	0	0
JWMN	TM_LEAVE5_3	Residence tract	0	0	0	0
JWMN	TM_LEAVE5_4	Residence tract	0	0	0	0
JWMN	MEANS6_2	Workplace tract	0	0	0	0
JWMN	MEANS6_3	Workplace tract	0	0	0	0
JWMN	MEANS6_4	Workplace tract	0	0	0	0
JWMN	MEANS6_5	Workplace tract	0	0	0	0

The median cell mean and quantile differences tend to indicate areas where there may have been potential for bias. As given in the Tables 2-7 and 2-8 most medians and IQRs of cell mean and quantile differences were zero or close to zero, indicating very low potential for bias. Even for county flows on mean travel time, the median of the cell mean differences were equal to zero. The medians and IQRs of cell mean and quantile differences for income by vehicles available in HH can deviate from 0. However, the deviations were negligible relative to the magnitude of income.

Table 2-9 shows the median, IQR, minimum, and maximum values of the absolute relative differences for mean travel time at the tract level by mean travel time. Census tracts with ACS mean travel time less than 5 were excluded. The results showed small deviations between raw and synthetic mean travel time, as given by the median relative difference being no more than 1 percent across each travel time category.

Table 2-9. Distribution of Absolute Relative Differences for Mean Travel Time at Tract Level by Mean Travel Time, ACS data 2017-2021

ACS tract Mean: Travel Time (minutes)	Median (%)	75th (%)	99th (%)
[5, 15)	0	2	15
[15, 20)	1	3	9
[20, 29)	0	2	14
[30, 45)	1	3	21
[45, 60)	1	2	25
[60, 75)	1	2	32
>=75	0	3	37

NOTE: Counties with ACS mean travel time less than 5 were excluded.

Table 2-10 shows the distribution of the absolute relative differences for mean household income at the county level by mean household income. Counties with absolute value of the ACS mean income less than $5,000 were excluded. The results showed small deviations between raw and synthetic

mean household incomes, as given by the median relative difference being no more than 2 percent across each income category.

Table 2-10. Distribution of Absolute Relative Differences for Mean Household Income at the County Level by Mean Household Income, ACS data 2017-2021

ACS tract Mean: Household Income	Median (%)	75th (%)	99th (%)
[$5,000, $15,000)	0	1	13
[$15,000, $25,000)	0	1	12
[$25,000, $35,000)	1	2	18
[$35,000, $50,000)	0	1	10
[$50,000, $75,000)	1	2	16
[$75,000, $100,000)	1	2	20
[$100,000, $150,000)	1	3	24
≥$150,000	2	5	33

NOTE: Counties with the absolute value of ACS mean HH income < $5,000 were excluded.

The conclusion was that the results on the cell means and quantiles analysis clearly supported that the impact of the synthesis approach on cell means and quantiles was at an acceptable level and there was little indication of bias introduced by the synthesis approach.

Weighted Cell Count Differences

Weighted cell counts were computed for select tables for county flows. Figure 2-4 provides a visual comparison of the weighted cell count estimates before and after synthesis for county flows. All the plots show minimal impact from the synthesis approach on these table estimates.

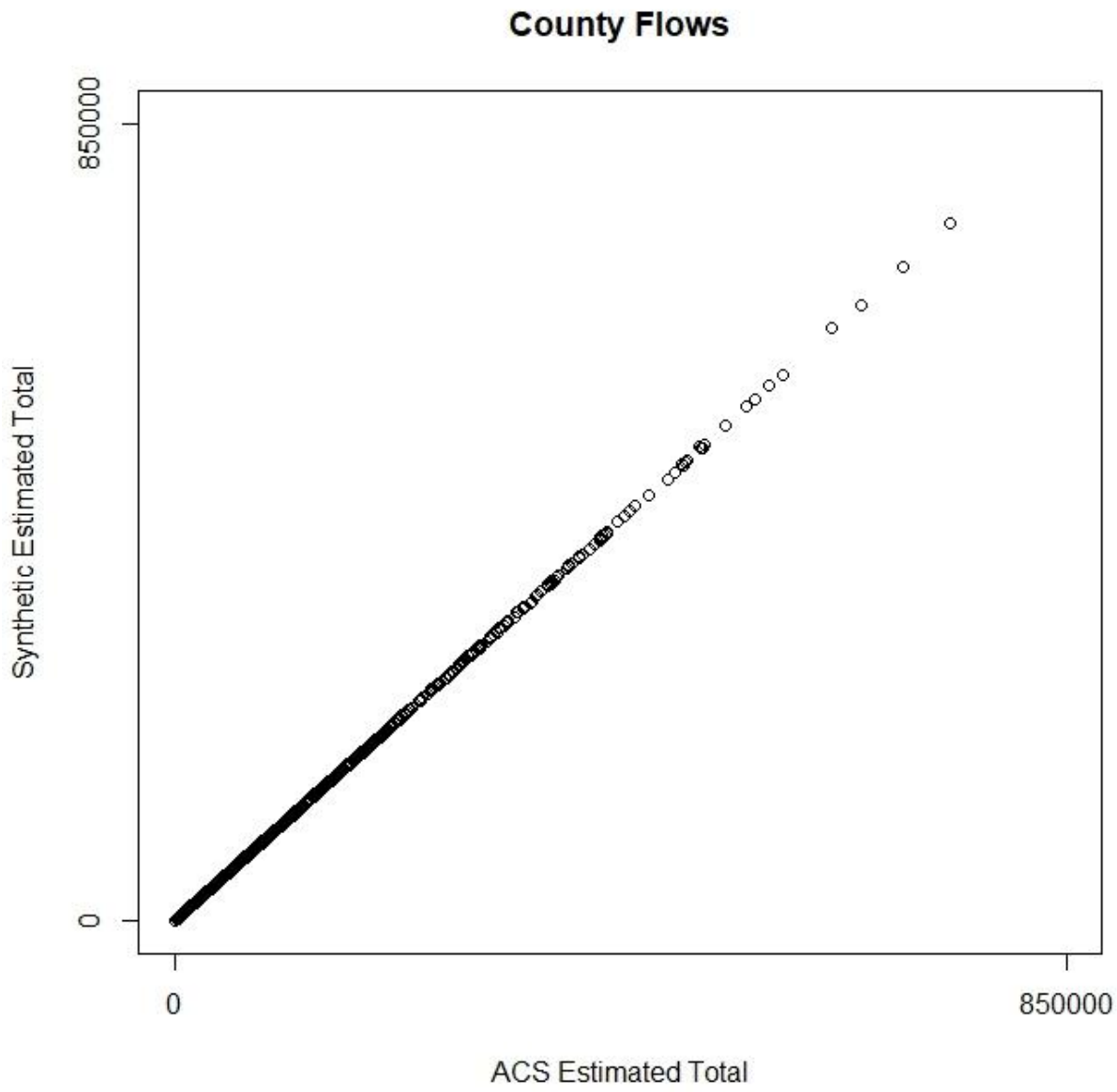

Figure 2-4. Plot of ACS and Synthetic Weighted Counts for County Flows

The conclusion from the results on weighted cell counts was that the acceptable level of impact from the synthesis approach seen in the NCHRP project was confirmed.

Impact of Synthesis on Standard Errors

The following difference formula (see discussion in Section 2.5) attempts to measure the impact on the standard error introduced by the synthetic data approaches. The formula ($f3$) from Section 2.5 was used to estimate the square root of the variance, referred to here as $se(\tilde{\theta})$. The difference

between $se(\tilde{\theta})$ and the ACS standard error is a measure of the impact of synthesis, and was computed as follows:

$$D_{se} = se(\tilde{\theta}) - se(\bar{\theta})$$

where, $se(\tilde{\theta})$= standard error of the CTPP synthetic estimate

$se(\bar{\theta})$ = standard error of the ACS estimate.

Additionally, the relative difference, $\mathrm{RD}_{\mathrm{se}} = \frac{D_{se}}{se(\bar{\theta})}$, was computed as well.

The standard errors were computed at the county level for mean travel time and mean HH income. The standard errors for mean travel times were computed for two levels of time leaving home, and four levels of MOT. The standard errors for mean household income were computed for five levels of vehicles available. The standard errors were also computed for county flows for minority and one category of industry.

Table 2-11. Median and Interquartile Range of Standard Error Differences

Attribute	BYVAR	Geographical Level	Median	IQR
AHINC	VEHICLES6_2	Residence county	0.31	1.78
AHINC	VEHICLES6_3	Residence county	0.78	2.55
AHINC	VEHICLES6_4	Residence county	4.57	22.48
AHINC	VEHICLES6_5	Residence county	8.30	45.93
AHINC	VEHICLES6_6	Residence county	11.42	75.59
JWMN	MEANS6_2	Residence county	0.00	0.02
JWMN	MEANS6_3	Residence county	0.00	0.01
JWMN	MEANS6_4	Residence county	0.00	0.02
JWMN	MEANS6_5	Residence county	0.00	0.02
JWMN	TM_LEAVE5_3	Residence county	0.00	0.03
JWMN	TM_LEAVE5_4	Residence county	0.00	0.11
INDUSTRY_5		County flow	0.00	0.00
MINORTY		County flow	0.00	0.00

Table 2-11 shows that the medians and the IQRs of the differences in standard errors were close to zero. The medians and the IQRs of the differences in standard errors of mean income by vehicles available in HH deviated from zero but the deviations were negligible relative to the magnitude of income and the standard errors of its means. Therefore, the conclusion from the results was that the synthesis approaches had very little impact on standard errors.

Cramer's V Ratios

As also used in Shlomo (2008), the Cramer's V was used to summarize the impact of the CTPP synthesis approach on two-way associations between MOT and CTPP variables. Let the Cramer's V statistic (V) (Agresti 2002) between two variables (treated as nominal) be equal to:

$$V(y_i, y_j) = \sqrt{\frac{\frac{\chi^2}{n}}{\min(k-1, l-1)}}$$

where

n = number of observations
k = number of categories for MOT (y_i), and,
l = number of categories for the other CTPP variable (y)

The range is $0 \leq V \leq 1$. The x^2 statistic, which is the Chi-squared statistic for testing independence of two nominal random variables, was weighted. Let the difference be computed as follows:

$$D_{CrV} = \tilde{V}(y_i, y_j) - V(y_i, y_j)$$

Where

$\tilde{V}(y_i, y_j)$ denotes the Cramer's V on the CTPP synthetic data file, and
$V(y_i, y_j)$ denotes the Cramer's V on the ACS data.

Cramer's V differences were produced for tract-level and county-level residences and workplaces, as well as county-flows. The differences were computed on two-way tables for MOT(11) with each of the following variables: Age [AGE(9)], HH income [HH_INC(26)], time leaving home [TM_LEAVE(10)], travel time [TRAVEL_TM(12)], and vehicles available [VEHICLES(6)].

Table 2-12. Median and Interquartile Range of Cramers V Differences

GEOAREA	MOT	Median	IQR
Workplace county	AGE9	0.00	0.00
Workplace county	HH_INC26	0.00	0.05
Workplace county	TM_LEAVE10	-0.00	0.00
Workplace county	TRAVEL_TM12	-0.00	0.00
Workplace county	VEHICLES6	-0.00	0.00

GEOAREA	MOT	Median	IQR
Workplace tract	AGE9	0.00	0.00
Workplace tract	HH_INC26	0.05	0.07
Workplace tract	TM_LEAVE10	0.00	0.02
Workplace tract	TRAVEL_TM12	0.00	0.01
Workplace tract	VEHICLES6	0.00	0.00
Residence county	AGE9	0.00	0.00
Residence county	HH_INC26	-0.01	0.05
Residence county	TM_LEAVE10	-0.00	0.00
Residence county	TRAVEL_TM12	-0.00	0.00
Residence county	VEHICLES6	0.00	0.00
County flows	AGE9	0.00	0.00
County flows	HH_INC26	-0.00	0.09
County flows	TM_LEAVE10	-0.00	0.00
County flows	TRAVEL_TM12	0.00	0.00
County flows	VEHICLES6	0.00	0.00
Residence tract	AGE9	0.00	0.01
Residence tract	HH_INC26	0.00	0.10
Residence tract	TM_LEAVE10	0.01	0.03
Residence tract	TRAVEL_TM12	0.00	0.02
Residence tract	VEHICLES6	0.00	0.01

Table 2-12 provides the Cramer's *V* results. All median Cramer's *V* differences were equal to or close to zero. The IQRs of Cramer's *V* differences were also very close to zero except for HH_INC26.

Pairwise Associations

Due to the sparseness of the ACS data, a majority of the census tract flows have one or two sample cases. The transportation planner can link together the explicit flow tables and string together several outcome tables (MOT, industry, age, income, poverty, minority status, etc) and form a microdata record. Therefore the multivariate relationships observed in the ACS data will need to be retained in the CTPP synthetic data.

Pearson product correlations were computed and shown in Tables 2-13 between select pairs of the following variables at the individual level: HH income, age, poverty status, time leaving home, travel time, and number of workers in the household.

In general, the correlations in the raw data were retained in the synthesized data. The synthesis approaches had very little impact on the pairwise correlations.

Table 2-13. Pairwise Correlations between Key Ordinal Variables

Var1	Var2	Actual	Synthetic
AHINC	AGE9	0.0803	0.0786
AHINC	POVERTY	0.2609	0.2595
AHINC	JWMN	0.0529	0.0525
AHINC	JWD	-0.0517	-0.0499
AHINC	HH_WRK6	0.2156	0.2146
AGE9	POVERTY	0.1623	0.1613
AGE9	JWMN	0.0346	0.0342
AGE9	JWD	-0.1192	-0.1100
AGE9	HH_WRK6	-0.2040	-0.2021
POVERTY	JWMN	0.0365	0.0362
POVERTY	JWD	-0.0863	-0.0808
POVERTY	HH_WRK6	0.1049	0.1038
JWMN	JWD	-0.0971	-0.0973
JWMN	HH_WRK6	-0.0132	-0.0126
JWD	HH_WRK6	0.0464	0.0432

Multivariate Associations

Woo et al. (2009) propose using propensity scores as a global utility measure for microdata as follows. The synthesized and ACS data files were stacked and T = 1 was assigned to the synthetic records and T = 0 was assigned to the ACS records. A weighted logistic regression model was processed on T using main effects, and also with interaction terms associated with synthesized variables. The following statistic *U* should be close to zero if the synthetic data and ACS data were indistinguishable.

$$U = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p_i} - c)^2$$

Where N = number in the stacked file

$\widehat{p_i}$ = propensity score (logistic regression prediction) for record i

c = proportion of units from the synthetic data file (e.g., ½)

The U statistics was computed as 0.000000142. This number is lower than the U statistics computed for the test sites in the development and validation phases of the NCHRP project (see NCHRP 08-79 Final Report) and similar to the one for the previous cycle (slightly lower even with a higher synthesis rate).

The stated purpose of these comparison tests was to conduct a reasonableness check to determine if the performance of the synthetic ACS CTPP tabulations was no worse than the raw tabulations when compared against typical model outputs. Based on the results discussed above, the authors concluded that there was little important difference between the raw and synthesized ACS tabulations for the comparison tests.

Cross-tabulation for Select Areas

As additional utility measures, we compared cross-tabulation estimates from the original data to those obtained using the synthesized data. After obtaining cross-tabulation estimates and associated confidence intervals for both the original and synthetic data, we first calculated the relative difference for each cell estimate, taken as the difference in cell estimates divided by the original estimates. This utility measure was used to ensure that the overall estimates did not change by too much. Next, we calculated interval overlap between the two confidence intervals for each table cell. If the original confidence interval is (l1, u1) and the confidence interval using the synthetic data is (l2, u2), then the interval overlap was calculated as

$$IO(l1, u1, l2, u2) = .5 * \left(\frac{\min(u1,u2)-\max(l1,l2)}{u1-l1} + \frac{\min(u1,u2)-\max(l1,l2)}{u2-l2}\right).$$

Confidence interval overlap is used to ensure the overall estimates do not change by too much when taking sampling variability into account. We also calculate the relative error, measured as the number of original standard errors the synthesized estimate is away from the original estimate. Lastly, we determined whether the synthesized estimate is within the original confidence interval. This informs us as to whether the new cross-tabulation estimates are within reason given the original data while considering the sampling variability.

We calculated the above-described utility measure for high-use tables. Specifically, we used the following tables:

1. Part 1
    a. B102101: Total workers
    b. B102106: Means of transportation
    c. B112211: Household size by Vehicles available
    d. B103203: Household income in the past 12 months by means of transportation
    e. B112202: Aggregate household income in the past 12 months by number of workers in household
2. Part 2
    a. B202100: Total workers
    b. B202105: Means of transportation

    c. B202112: Time arriving
    d. B202221: Class of worker by time arriving
3. Part 3
    a. B302100: Total workers
    b. B302103: Means of transportation

In additional to performing these cross-tabulation utility measures on the overall cross-tabulations at the national level, we also performed the same utility measure for geographical subsets. These help to ensure that the estimates using the synthetic data are also accurate at smaller geographical levels. Since the synthetic data is used by smaller localities in practice, we want the resulting estimates from smaller geographical places to be accurate as well. We used the following geographical subsets:

1. Part 1: Residence
    a. Atlanta, GA
    b. Kansas City, KS
    c. Reno, NV
2. Part 2: Place of Work
    a. Bozeman, MT
    b. Gainesville, FL
    c. Chippewa Falls, WI
3. Part 3: Flow
    a. Live in Ft. Worth, TX and work in Dallas, TX
    b. Live in Wilmington, DE and work in Philadelphia, PA
    c. Live in New Rochelle, NY and work in White Plains, NY

The confidence internal overlap was on average over 85% for each geographical subset across tables and no lower than 70%, indicating that the two sets of estimates were closely aligned .

Skip Patterns

Finally, we checked that all skip patterns hold in the synthesized data as they do in the original data. To accomplish this task, we created cross-tabulations for various subsets of variables that are used in skip patterns using both the original and synthesized data sets. Combinations of variables in the synthesized data that are not present in the original data are flagged for manual review.

2.4.2 Disclosure Risk Measures

Risk measures were developed to consider disclosure risk factors inherent in the data. These risk measures were used to identify disclosure risk with an objective to help alleviate concerns and provide assurance on the reduction of disclosure risk. As discussed in Section 1.1.2 in the NCHRP 08-79 Final Report, tables can be linked together to form a string of identifying characteristics (referred to as a "key"). Synthesis of the data and/or generation of synthetic data will mean that exact matches on the key will be unlikely and data values for an individual will not be predicted as accurately; therefore an intruder will have a harder time performing inference for an individual record's true values. The synthesis replacement rate is a factor that affects both utility and risk. The synthesis rate, and change rates (proportion of records with values that changed) are the primary measures used for the 2017-2021 CTPP process. The rates have been provided in a memorandum to the DRB for their review.

2.5 Variance Estimation

The successive difference replication approach (described in Fay and Train, 1995 and Census Bureau, 2009) was used to compute ACS variances. Suppose $\hat{\theta}_0$ represents the ACS estimate of θ, and $\hat{\theta}_k$ is the ACS estimate of θ for replicate k. Then the variance of $\hat{\theta}_0$ can be estimated as

$$\text{var}(\hat{\theta}_0) = \frac{4}{80}\sum_{k=1}^{80}(\hat{\theta}_k - \hat{\theta}_0)^2 \quad \text{(f1)}$$

This formula treats the ACS data as if it were reported without accounting for variance caused by Census Bureau's imputation and masking.

In the final report of the NCHRP project, we summarized our research results on the variance estimation for the synthetic data. Two variance estimators, (f3) and (f5) were proposed to account for the errors generated in the synthesis process. In both estimators, a term of squared difference between the ACS and data synthetic estimates is added, which serves for the purpose of measuring the additional variance due to synthesis. We recommended the use of (f5) in computing the standard errors of the estimates in CTPP tables. This has been approved by the Census Bureau. Details of the formulae are illustrated below.

$$\text{var}(\tilde{\theta}_0) = \frac{4}{80}\sum_{k=1}^{80}(\tilde{\theta}_k - \tilde{\theta}_0)^2 + \left(\tilde{\theta}_0 - \hat{\theta}_0\right)^2. \quad \text{(f3)}$$

In (f3), the first term,

$$\frac{4}{80}\sum_{k=1}^{80}(\tilde{\theta}_k - \tilde{\theta}_0)^2, \quad \text{(f4)}$$

is called the naïve estimator, which results from applying the usual ACS formula directly to the synthetic data. In the formula $\tilde{\theta}_0$ represents the CTPP synthetic estimate of θ, and $\tilde{\theta}_k$ is the estimate for replicate k. This estimator can be biased since variance due to data synthesis is not appropriately taken into consideration.

An alternative estimator to (f3) is to add the squared difference to the usual ACS estimate, $\text{var}(\hat{\theta}_0)$. Assuming data synthesis is independent of the sampling process; formula (f5) is essentially the sum of sampling variance and variance due to data synthesis.

$$\text{Var}(\tilde{\theta}_0) = \text{var}(\hat{\theta}_0) + \left(\tilde{\theta}_0 - \hat{\theta}_0\right)^2. \quad \text{(f5)}$$

Assuming that the noise introduced to the synthetic data, $\tilde{\theta}_0 - \hat{\theta}_0$, has a zero mean and constant variance σ_p^2 given the ACS estimate $\hat{\theta}_0$. Taking expectations of $(f5)$, therefore

$$\begin{aligned} E_s E_p \text{var}(\tilde{\theta}_0) &= E_s\left(\text{var}(\hat{\theta}_0) + E_p\left((\tilde{\theta}_0 - \hat{\theta}_0)^2 | \hat{\theta}_0\right)\right) \\ &= E_s\left(\text{var}(\hat{\theta}_0) + \sigma_p^2\right) \\ &\cong \text{Var}(\hat{\theta}_0) + \sigma_p^2, \end{aligned}$$

where E_s is the expectation with respect to sampling, E_p is the expectation with respect to data synthesis, and $\text{Var}(\hat{\theta}_0)$ is the true variance of $\hat{\theta}_0$. The ACS variance estimator $\text{var}(\hat{\theta}_0)$ is approximately unbiased; that is, $E_s\left(\text{var}(\hat{\theta}_0)\right) \cong \text{Var}(\hat{\theta}_0)$ (Fay and Train 1995). Hence, (f5) is approximately unbiased for the true variance of the synthetic estimate.

Computationally, formula (f5) requires the following information:

- ACS full sample and replicate weights;
- ACS data values for variables in the CTPP tables;
- CTPP full sample weight;

- Synthetic ACS data values for variables in the CTPP tables.

The processing takes the following steps:

- Generate the point estimates for all CTPP tables twice: once for ACS data, and once for the synthetic data;
- Using the successive difference replication formula (f1), generate the ACS variance estimates using ACS data and ACS full sample and replicate weights;
- Using formula (f5), compute the variances for the synthetic estimates as the sum of ACS variances and squared difference between the ACS and synthetic estimates.

The following describes the approach another way.

Let FILEA = original ACS microdata file, VHOUS and VPERS.
Let REPW0 = ACS original full sample weight, which resides on FILEA.
Let REPW1-80 = the ACS original replicate weights, which reside on FILEA.

Using FILEA, for each table cell of each CTPP table, generate the point estimate (e.g., sum of weights, weighted mean, weighted median) using REPW0, and let Z_0 represent that point estimate. Then using REPW1, in the same manner compute Z_1; using REPW2, compute Z_2; and so on. Then compute the usual ACS variance (V) as:

$$V = \frac{4}{80} \sum_{k=1}^{80} (Z_k - Z_0)^2$$

Let FILEB = synthetic ACS microdata file, VHOUS_SETB and VPERS_SETB.
Let REPW0' = CTPP adjusted full sample weight, which resides on FILEB.

Using FILEB, for each table cell of each CTPP table, generate the point estimate (e.g., sum of weights, weighted mean, weighted median) using REPW0', and let Z_0' represent that point estimate.

Then compute the final variance for Z0' as:

$$V' = V + (Z_0' - Z_0)^2.$$

The two input datasets, the original ACS data and the CTPP (synthetic ACS) data, will have the same data structure and variable names so that the CTPP production team can conveniently use

their table generating software/system. The only difference is that there are no replicate weights on the synthetic ACS dataset.

Variance for Zero-Estimated Counts

Table 2-14 shows the possible scenarios that the ACS and/or the synthetic estimates are zero counts and the estimated variances for the synthetic estimates.

Table 2-14. ACS and Synthetic Estimates and Variances for Synthetic Estimates

Scenario	Original ACS Estimate	Synthetic ACS Estimate	Variance Estimation for Synthetic Estimate
1	=0	=0	constant
2	>0	>0	(f5)
3	=0	>0	constant+$(\tilde{\theta}_0)^2$
4	>0	=0	constant

Under Scenario 1, when the original ACS estimates are zero-estimated counts of people or households, a direct application of the usual ACS variance estimator, (f1), leads to a zero variance. This is not appropriate since there may be people or households with such characteristic but were not selected in this specific ACS sample. For these cases, a model-based approach is done and the variance is represented as a constant for each state.

Under Scenario 2, when both the original and synthetic estimates are non-zero, then formula (f5*)* is used.

Under Scenario 3, when the original ACS estimated count is zero and the synthetic estimate is not, the estimated variance should be computed as the sum of a constant and $(\tilde{\theta}_0)^2$. Estimated variances strictly based on (f5*)* may increase the chance of enabling intruders to find out there are actually no real ACS cases, although this type of risk was of less concern to Census DRB (at the time of the 2006-2010 process).

Under Scenario 4, when the original ACS estimated count is not zero and the synthetic estimate is zero, the estimated variance under (f5*)* would be $\text{var}(\hat{\theta}_0)+(\hat{\theta}_0)^2$. Intruders will not be able to separate $\text{var}(\hat{\theta}_0)$ from $(\hat{\theta}_0)^2$ but they could still know that they are real original ACS cases since, most likely, $\text{var}(\hat{\theta}_0)+(\hat{\theta}_0)^2$ is different from the constant. For disclosure limitation purposes, the variances will be estimated as a constant if the synthetic estimates are zero, no matter whether the corresponding original ACS estimates are zero or not.

Variance for Estimated Ratios for Empty cells

If a table cell is empty based on the ACS microdata, its ratio estimate, such as mean travel time to work, is unavailable as well as the standard error/variance. There are chances that the corresponding table cell is not empty based on the synthetic CTPP table. Even when the CTPP ratio estimate is available, the CTPP estimate is not published as two of the three parts of the (f3) formula are missing. In addition to the unavailability of the ACS point estimate and its standard error/variance, approximating the standard error is not plausible in this scenario unlike when the ACS count or percent is zero.

3 Documentation of Programs

This chapter introduces the processing steps in Section 3.1 and describes the program components in Section 3.2.

3.1 Introduction to Processing Steps

This chapter provides a description of the production run processing for the Census Transportation Planning Products (CTPP) tables that will be processed on American Community Survey (ACS) 2017–2021 sample data. The five-year ACS data were processed through six main programming components. The steps are organized as follows:

1. Initial risk analysis;
2. Data replacement approach;
3. Weight calibration—raking (this includes generating control totals);
4. Data utility measures;
5. Risk measures; and
6. Cleanup.

3.2 Documentation of Programs

Each of the following sections describes a main component of the overall program. Each section contains a brief description and a flowchart of the process. The list of programs and modules is shown in Appendix B.

3.2.1 Program Component: Initial Risk Analysis

The set of initial risk analysis modules was processed to generate the tables for the purpose of flagging data values that violated the special tabulation disclosure rules and therefore are at the highest risk of disclosure. Several steps were necessary within the initial risk analysis component to prepare for the application of the data synthesis approach, including the creation of ACS area-level

covariates, as well as the preparation of other input data. The data-driven risk analysis is a major preliminary step processed on the national database. ACS variables that have already been imputed during the ACS imputation process were not replaced; that is, they were considered to have already been synthesized. This same approach was applied as acceptable to the DRB during the 2006-2010 and 2012-2016 processing. As part of the initial risk analysis, data values were classified according to risk strata. Figures 3-1 and 3-2 show the initial risk analyses at the person level and at the household level, respectively.

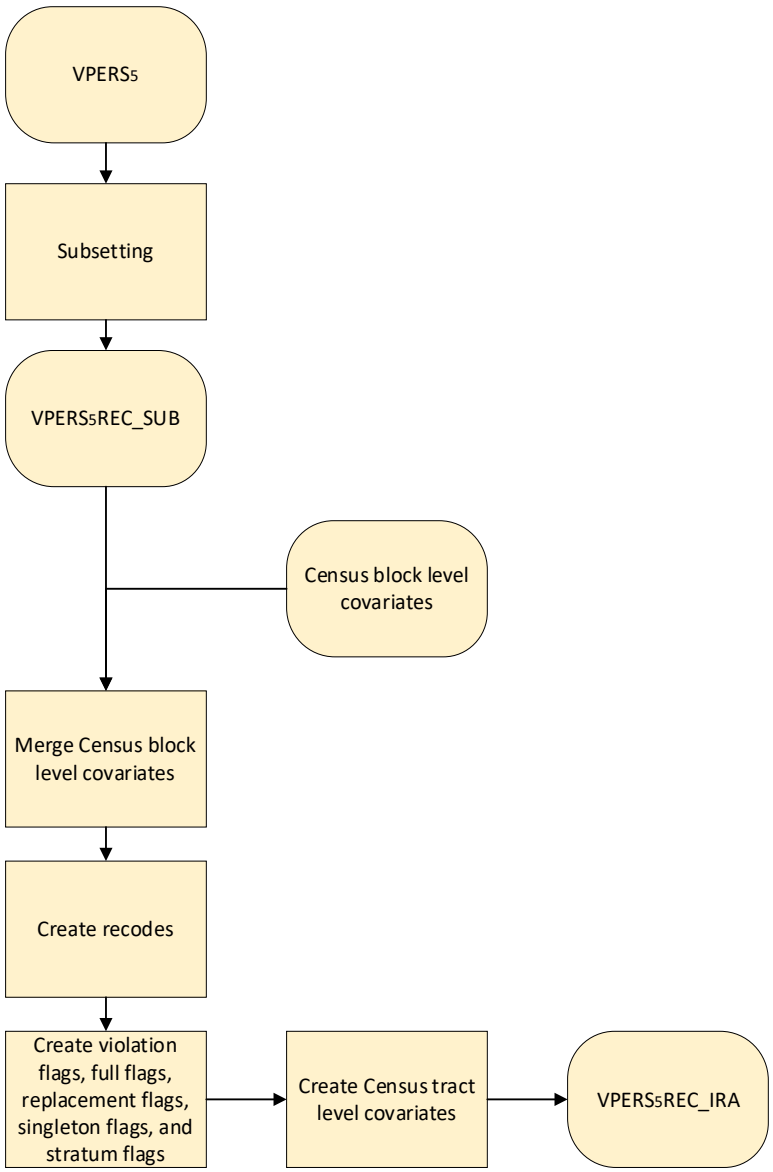

Figure 3-1. Flowchart of Person-Level Initial Risk Analysis Program Component

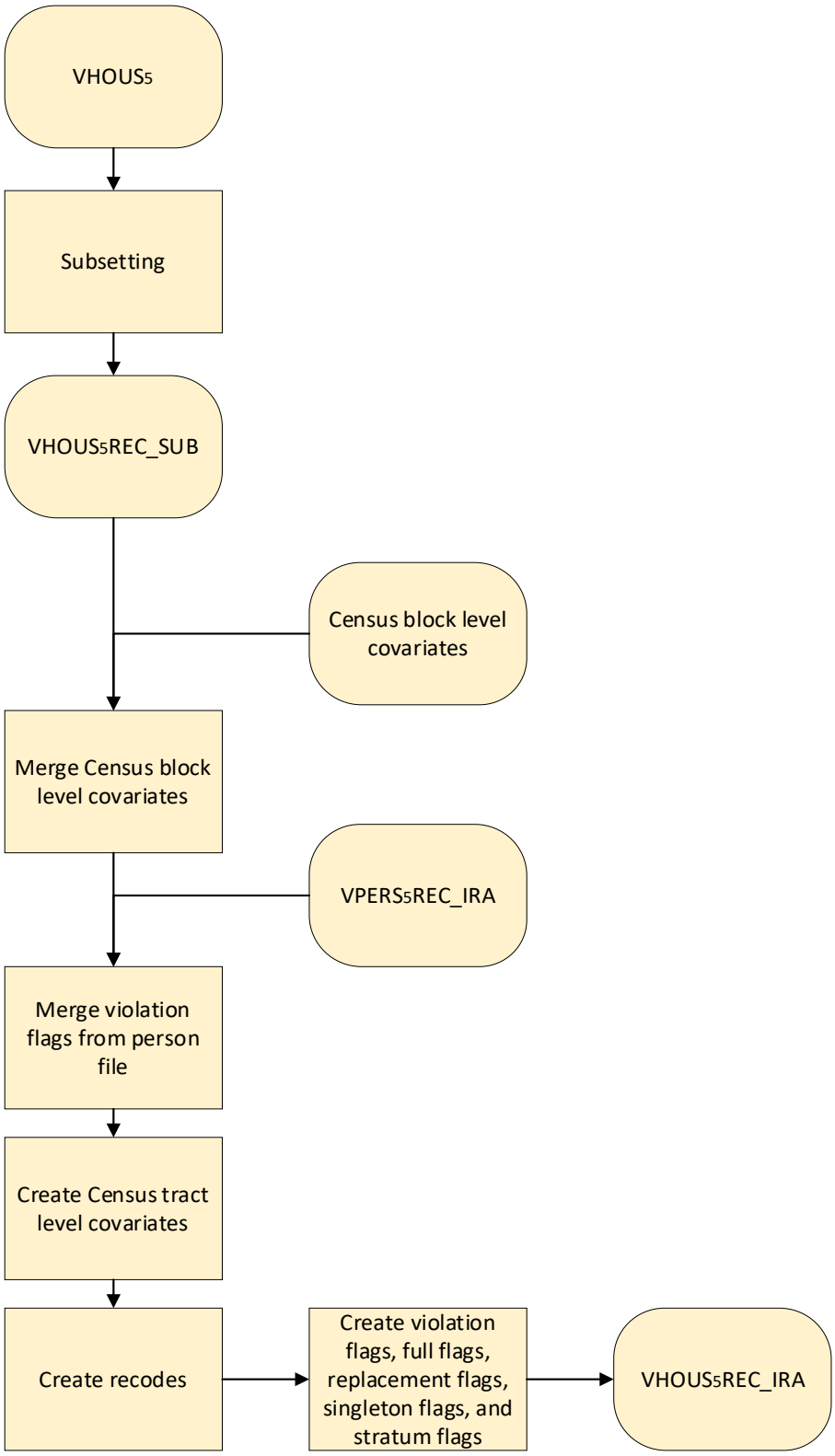

Figure 3-2. Flowchart of Household Level Initial Risk Analysis Program Component

3.2.2 Program Component: Data Synthesis

This program combines the model assisted constrained hot deck and semi-parametric approaches into one program. The initial steps before processing the approaches involve assigning partial replacement flags and running an extensive variable prep module. The set of data replacement modules is driven by a Master Index File (MIF). Risk strata were identified for each variable to be synthetic, and the rates were used to select and flag (VarName_PARTIAL) a sample of data values for replacement. The Variable Prep step is processed in order to prepare recodes and prepare variables as predictors for the semi-parametric approach. The MIF identifies the variables to be synthetic as well as the variables to be put into the pool of candidate predictor variables. It is used to classify the type of each variable as real numeric, ordered categorical, or unordered categorical. For the unordered categorical variables, indicator variables were created. Select interaction terms to be added to the pool of candidate predictor variables were identified as well.

Once the variable prep processing was completed, then the model selection approach was processed for all variables identified in the MIF that undergo the semi-parametric approach. The parameters used in the MIF are as follows:

Item = integer value that identifies the item number
ProcessNumber = blank or integer, linking together VarNames in order to process together in one step
VarName = name of the variable
SortVars = name of the sorting variables for target selection within explicit strata
RiskStrat = name of variable containing the strata for target selection
SampRate = real numbers depicting the selection rate for target selection. The rates are delimited by a space, one for each value of RiskStrat.
MOS = name of variable containing the measure of size for target selection.
Deselect = a comma-delimited set of values enclosed in parentheses. It is used in the following statement, after the PROC SURVEYSELECT: if &riskstrat in &deselect then &targselflag = 0;
Approach = 'SP', 'CH' , 'RL', or ' ' for semi-parametric, MACH, rank linking, and not-applicable, respectively. It should be non-blank if Replace = 1.
VPERS = 1/0 determines if the VarName is in the person-level file
VHOUS = 1/0 determines if the VarName is in the household-level file
Transfer = 1/0 determines if the VarName needs to be transferred from the household-level file to the person-level file

Type	=	'UC', 'OC', 'N', or ' ' for unordered categorical, ordered categorical, real numeric, and not-applicable, respectively. It should be non-blank if Replace = 1.
Replace	=	*0/1*, the value = 1 if the VarName is targeted for replacement, and it equals 0 otherwise.
VarToBin	=	name of the variable to make bins for, typically same as VarName
BinVar	=	name of the variable that contains the bins. If blank, set equal to DUMMY1, and assign DUMMY1 = 1 to all records.
BinA	=	statements defining the first set of bins, separated by semi-colons. An example is: (x2,y2]; (x3,y3] or (x4,y4]; defines two explicit groups with the remaining unspecified catch-all group, where '(' or ')'is not inclusive and '[' or ']' is inclusive. The 'or' connector is the only one allowed.
BinB	=	statements defining the second set of bins that overlap with BinA, separated by semi-colons.
TargSelFlag	=	*name* of the variable containing 0/1 values that identify the data values to be replaced for the VarName. If blank, then a target record selection process will take place.
HDCellVar	=	*list* of variables to help define the hotdeck cells (excludes BinVar). If blank, set equal to DUMMY2, and assign DUMMY2 = 1 to all records.
Locality	=	*list* of variables to help define the locality for hotdeck cells. If blank, set equal to DUMMY3, and assign DUMMY3 = 1 to all records.
ModelArea	=	*list* of variables that define the geographic areas for which the model selection takes place.
RankOrderHD	=	Series of integers from 1 to 5, delimited by a '#' sign, ordering the following, &BinVar, &HDCellVars, &Locality, &PG&DepVar, WtCell. Default = 1 # 2 # 3 # 4 # 5.The ranks need to be without duplicate integers.
NumPredGrp	=	integer value of the number of prediction groups to form
NumWtCell	=	integer value of the number of weight groups to form
LinkToVar	=	*name* of the variable (&VarName) used to link to through rank linking
TrgtVars	=	blank or list of variable(s) linked and targeted in same process 0"
AddNoise	=	*real number* greater than 0 that is used in the following formula if there is no change before and after synthesis.

$Y = y*(1+f*z)$, where f = &AddNoise, and z is a draw from the standard normal distribution. The standard deviation of the added noise is the product of f and y, which means the level of noise is allowed to vary relative to the magnitude of y.

InteractUC = *name* of variable with Type = 'UC'. The variables will have indicator variables created and each indicator will be crossed with all other variables on the MIF with Predictor = 1 and Type = 'OC' or 'N'.

Interaction = *0/1*, the value 1 says the InteractUC variable will be crossed with all other variables on the MIF with Predictor = 1 and Type = 'OC' or 'N'.

Predictor = *0/1*, the value 1 says the variables will be considered in the stepwise regression leading to the prediction model.

ForceList = *names* of variables to be forced in and kept in the resulting stepwise regression model.

Include = text entry showing the computation of the number of forced in variables, including the indicator variables.

Donors = argument provides the SAS program subsetting argument to limit the donor pool to a subset of cases. It is coded as "If &Donors;"

Drops = argument provides the SAS program statement to remove cases from the data synthesis process after targets are selected. It is coded as "If &Drops then output DropFile".

Proxy = *name* of proxy variable to be used in the model selection on the left hand side of the model.

Once the variable prep processing was completed, then the model selection approach was processed for all variables identified in the MIF. Model selection is processed for the purpose of identifying the predictors for each target variable, and to estimate the model parameters for generating predicted values, which are necessary for creating hot deck cells in the data synthesis step.

One by one, the target variables are processed through the Main Loop. Either the model assisted constrained hot deck or the semi-parametric approach is processed, depending on the variable type of the target variable. First, household-level variables are synthesized, then the synthesized household variables are transferred to the person level, where the process continues with the synthetic data approach applied to person-level variables.

After processing, pre-post checks are conducted in order to have an initial look at the impact of the data synthesis. Frequencies, means, and correlations are generated before and after data synthesis.

Lastly, recodes are processed in order to prepare for the raking step. Figure 3-3 provides the flowchart of the process.

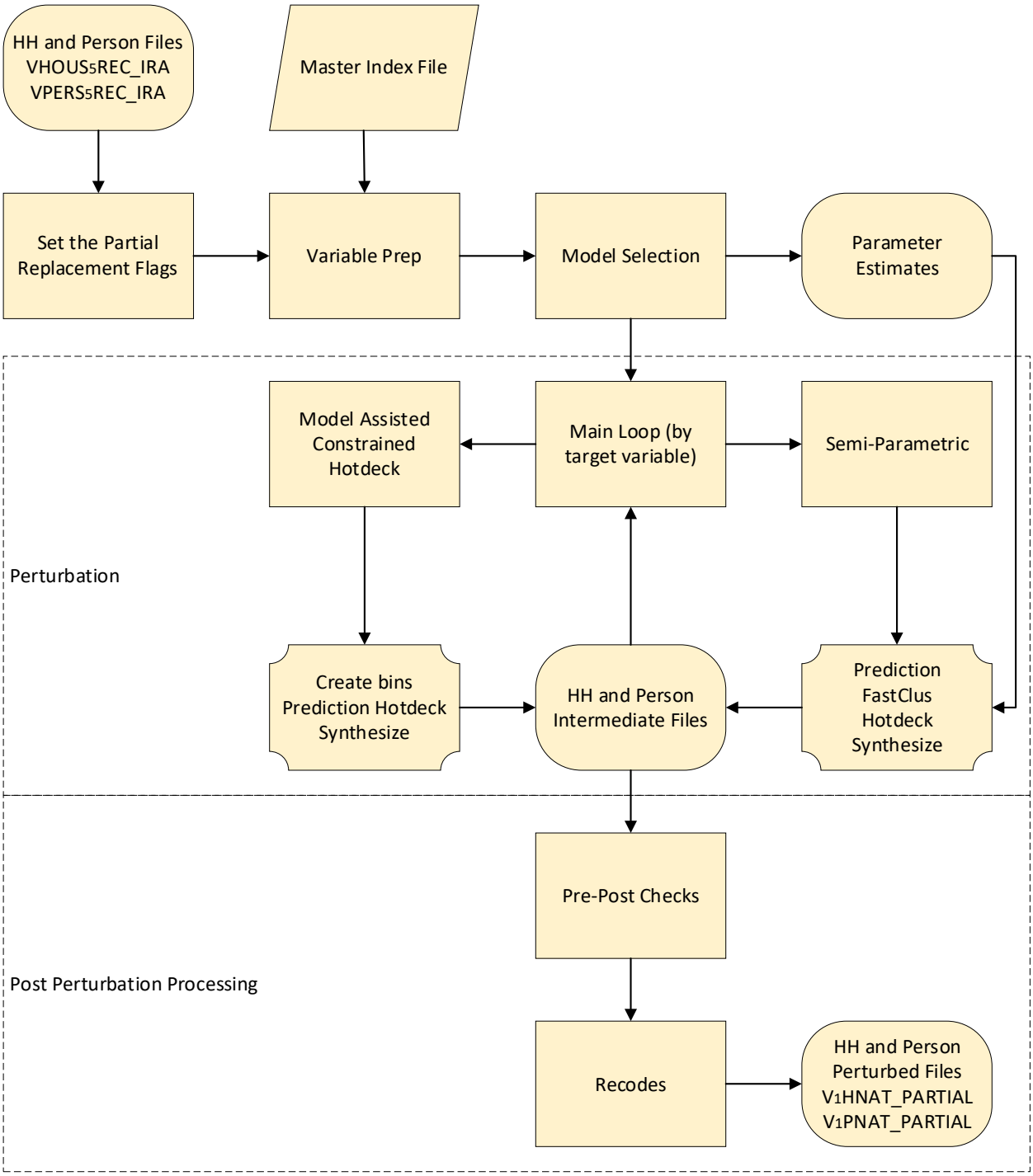

Figure 3-3. Flowchart of Data Synthesis Program Component

3.2.3 Program Component: Raking

After the approaches are processed, the weight adjustment step, known as raking, is done so that the weights are calibrated to reproduce select ACS estimates at the following geographic levels:

- Public Use Microdata Area (PUMA) level, which are areas formed to be greater than 100,000 in population for the purpose of releasing public use microdata;
- Census tract level, which are areas of about 4,000 in population;
- County level; and
- State level.

Only the full sample weight was raked to the estimated totals from the five-year ACS, as the raked replicate weights are not needed for variance estimation purpose (see section 2.5). Figure 3-4 provides the flowchart of the process.

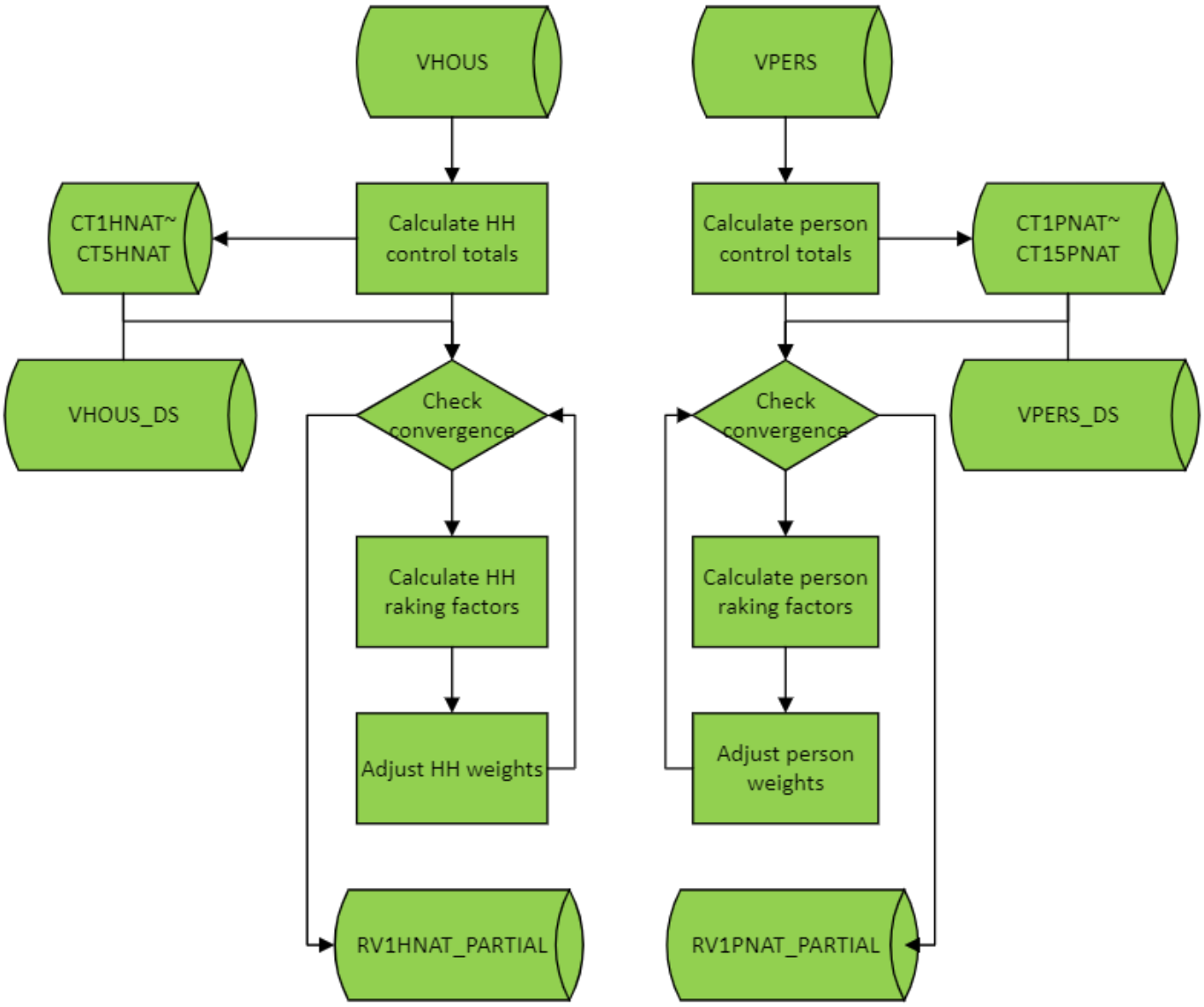

Figure 3-4. Flowchart of Household Level and Person Level Control Total Calculations and Raking Program Component

3.2.4 Program Component: Utility Measures

The data synthesis approaches for the CTPP research were designed to limit the impact on data utility while reducing the risk of disclosure. These measures were developed for the resulting data utility so that the balance between risk and utility can be understood for the CTPP tables.

The focus of the checks is to compare the ACS data with the synthetic ACS data. The comparisons check cell means and quantiles, weighted cell counts, standard errors, Cramer's V for associations in two-way tables, pairwise associations, and multivariate associations at the TAZ level and the county level. The median of differences between the raw and synthetic estimates (across estimates for geographic areas) were computed where appropriate in order to give indications of potential bias introduced by the data synthesis. The interquartile range (IQR) for the differences provided an indication of the variation caused by the data synthesis.

3.2.5 Program Component: Risk Measures

Risk measures were developed to consider disclosure risk factors inherent in the data. These risk measures were used to estimate disclosure risk with an objective to help alleviate concerns and provide assurance on the reduction of disclosure risk. At the time of the 2006-2010 and 2012-2016 processes, the research team and the Census DRB recognized that combinations of just a few variables can lead to a single sample unit (sometimes referred to as a sample unique or singleton was considered). The impact on disclosure risk reduction from sources of data protection, whether it is through sampling, the realization of moving and workplace changes over time, or measurement error created through ACS swapping, ACS imputation, and the synthetic CTPP data.

The main risk measures used for the 2017-2021 process are the synthesis and change rates. The general approach for the NCHRP 08-79, 2006-2010, and 2012-2016 production process also brought together other measures of various risk elements. The measures were found acceptable by the DRB. While these risk components could be looked at separately, with the buildup of a series of factors, the *product* of the following risk components can therefore be considered to quantify the overall risk as a score. More details can be found in Section 2.2.2 of the NCHRP 08-79 report. Figure 3-5 provides the flowchart of the process.

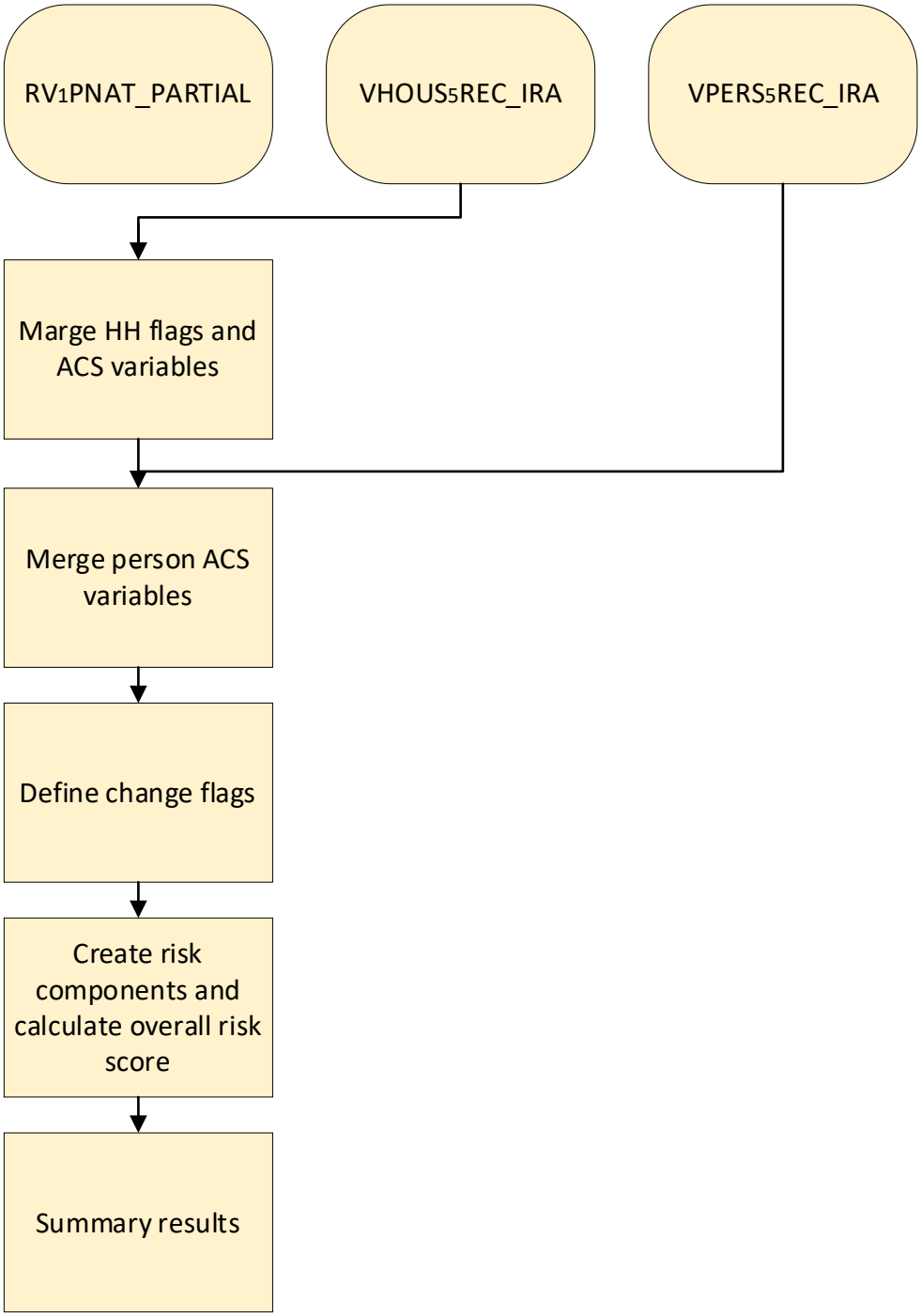

Figure 3-5. Flowchart of Disclosure Risk Measures

3.2.6 Program Component: Cleanup and Output Files

This program creates the delivery files after the processes of initial risk analysis, data synthesis, raking, risk, and utility are finished. The final files at the household and person levels will contain the same number of records and variables as in the input files, with some targeted values being replaced and full sample weight being calibrated.

The delivery files were placed in the following directory:

\\tabgen9.acs.census.gov\data\tab7\spectabs\westat_programs\prod\data\2017thru2021\07082024

The SAS datasets were named VHOUS and VPERS and contain the same information as initially provided in the original files, with the exception of the following differences.

For VHOUS, the values of synthetic variables may be different from the values in the original file. The re-calibrated weight was REPW0. The replicate weights were suppressed.

For VPERS, the values of synthesized variables may be different from the values in the original file. The re-calibrated weight was REPW0. The replicate weights were suppressed.

References

Agresti, A. (2002) *Categorical Data Analysis*, Wiley-Interscience, 2nd ed.

Deming, W.E. and F.F. Stephan (1940), "On a Least Square Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known", *Annals of Mathematical Statistics*, 11, 427–444.

Domingo-Ferrer, J. and Franconi, L., eds. (2006). Privacy in Statistical Databases. Lecture Notes in Computer Science 4302 Springer-Verlag Berlin Heidelberg.

Domingo-Ferrer, J., and Torra, V. (2004), *Privacy in Statistical Databases: CASC Project International Workshop*, Lecture Notes in Computer Science, Springer, New York.

Drechsler, J., and Reiter, J.P. (2009). Disclosure Risk and Data Utility for Partially Synthetic Data: An Empirical Study Using the German IAB Establishment Survey. *Journal of Official Statistics*, Vol. 25(4), 589–603.

Duncan, G.T., Keller-McNulty, S.A., and Stokes, S.L. (2001). Disclosure risk vs. data utility: The R-U confidentiality map. NISS technical report, 21, www.niss.org.

Fay, R. and Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods*.

Gomatam, S., and Karr, A.F. (2003). Distortion measures for categorical data swapping. *NISS Technical Report #131.*

Gomatam, S., Karr, A.F., and Sanil, A.P. (2003). A risk-utility framework for categorical data swapping. *NISS Technical Report #132.*

Gomatam, S., Karr, A.F., and Sanil, A.P. (2004). Data swapping as a decision problem. *NISS Technical Report #140.*

Hilbert, D. (1891). Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*, 38, 459–460.

Judkins, D., Piesse, A., Krenzke, T., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariance structure through semi-parametric whole-questionnaire imputation. In *Proceedings of the Joint Statistical Meetings on CD-ROM* (pp. 3211-3218). American Statistical Association.

Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., and Sanil, A.P. (2006). "A framework for evaluating the utility of data altered to protect confidentiality." *The American Statistician*, 60 (3), 224-232.

Krenzke, T., Li, J., Freedman, M., Judkins, D., Hubble, D., Roisman, R., and Larsen, M. (2011). Producing Transportation Data Products from the American Community Survey that comply with disclosure rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences.

Krenzke, T., Li, J., and Mckenna, L. (2017). Producing multiple tables for small areas with confidentiality protection. *Journal of the International Association of Official Statistics*, 33(2), 469-485. doi: 10.3233/SJI-160259

Miller, D. 2008. Critical need for data from the American Community Survey (ACS). Memo from Deb Miller (AASHTO Standing committee on planning) to Christa Jones (Census Bureau). http://trbcensus.com/drb/08052008.pdf.

Oganian, A., and Karr, A.F. (2006). Combinations of SDC methods for microdata protection. In Domingo-Ferrer, J. and Franconi, L., eds. (2006). Privacy in Statistical Databases. Lecture Notes in Computer Science 4302 Springer-Verlag Berlin Heidelberg. 102-113.

Oh, H.L. and F.J. Scheuren (1987), "Modified Raking Ratio Estimation", *Survey Methodology*, 13, 209-219.

Reiter, J. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology*, 29, 181-188.

Shlomo, N. (2008). Releasing microdata: disclosure risk estimation, data masking, and assessing utility. *Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.*

Westat. (2010). NCHRP Project 08-79: Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules. "Summary of Task 1 and 2 Results: Disclosure Rules and Initial Selection of Approaches" Technical Memorandum.

Woo, M., Reiter, J., Oganian, A. and Karr, A. (2009). Global measures of data utility for microdata masked for disclosure limitation. *The Journal of Privacy and Confidentiality,* 1(1), 111-124.

Appendix A

Set of Predictor Variables

Item	Variable Name	Variable Description	Level at Which Variable is Defined
1	AHINC	HH income	HH
2	WIH	Number of workers in HH	HH
3	CHILDREN	Presence of children	HH
4	AGE	Age	Person
5	MINORITY	Minority status	Person
6	SEX	Sex	Person
7	IND	Industry	Person
8	JWD_SHIFT	Work shift	Person
9	OCC	Occupation	Person
10	YRS_US	Years of US residence	Person
11	APERN	Person earnings	Person
12	YNGEST	Age of youngest child	HH
13	VEH	Vehicles available	HH
14	HHLDRAGE	Householder age	HH
15	MINORITY_HH	Householder minority status	HH
16	YRSUS_HH	Householder years of US residence	HH
17	COW	Class of worker	Person
18	AVG_HHSIZE	Average HH size	Tract
19	AVG_WIH	Average number of workers in HH	Tract
20	AVG_VEH	Average vehicles available	Tract
21	C_PCTBLK	Percentage of population who are Black	Block
22	C_PCTHISP	Percentage of population who are Hispanic	Block
23	C_PCTOOCC	Percentage owner occupied	Block
24	MED_APERN	Median person earnings	Tract
25	MED_HHINC	Median HH income	Tract
26	PCT_BLK	Percentage of workers who are Black	Tract
27	PCT_COLL	Percentage of workers who are college graduates	Tract
28	PCT_HIGH	Percentage of workers with a high school diploma	Tract
29	PCT_HIS	Percentage of workers who are Hispanic	Tract
30	PCT_MAR	Percentage of workers who are married	Tract
31	PCT_POV	Percentage of workers in poverty	Tract
32	PCT_RENT	Percentage of workers who rent	Tract
33	PCT_PHONE	Percentage of workers with a phone line	Tract
34	PCT_UNDER18	Percentage of workers under 18 years of age	Tract
35	PCT_WHT	Percentage of workers who are White	Tract
36	PCI	Principal city indicator	Person
37	UR	Urban rural indicator	Person
38	NEW_HHT	Household family type recode	HH
39	MODE	Mode of data collection	HH
40	NOC	Number of own children	HH
41	NEW_NPF	Number of persons in family recode	HH
42	HH_OVER16YRS	Number of persons 16 years old or older in HH	HH
43	HH_SIZE	HH size	HH
44	LNGI	Language indicator	Person
45	TEL	Telephone indicator	HH
46	BORNUSAHH	Householder country of birth	HH

Item	Variable Name	Variable Description	Level at Which Variable is Defined
47	NEW_HHLDSCHL	Householder education attainment recode	HH
48	HPOV	Household poverty index	HH
49	R60	Presence of persons > 60 years old	HH
50	RMS	Number of rooms	HH
51	HH_LIFE	Householder life cycle	HH
52	HHLDRMOT	Householder means of transportation	HH
53	STRUCTURE9	Household structure	HH
54	TEN	Tenure	HH
55	WKH	Hours worked per week	Person
56	WKW	Week worked in past 12 months	Person
57	ENROLLMENT	Enrollment status	Person
58	USUAL_HRS	Usual number of hours worked	Person
59	BORNUSA	Country of birth	Person
60	NEW_MIL	Military status recode	Person
61	MEANS11	Means of transportation	Person
62	ESR	Employment status	Person
63	NEW_MAR	Marital status recode	Person
64	NEW_MIG	Migration status recode	Person
65	NEW_POWPCI	Place of work principal city indicator recode	Person
66	NEW_VETSTAT	Veteran status recode	Person
67	NEW_JWMN	Travel time recode	Person
68	NEW_HHLDRHIS	Householder Ethnicity recode	Person
69	NEW_HHLDRACE	Householder race recode	Person
70	NEW_POVERTY	Poverty recode	Person
71	MOT*AHINC	MOT interaction with HH income	HH
72	MOT*CHILDREN	MOT interaction with presence of children	HH
73	MOT*AGE9	MOT interaction with age categories	Person
74	MOT* MINORITY	MOT interaction with minority status	Person
75	MOT*SEX	MOT interaction with sex	Person
76	MOT*HH_WRK6	MOT interaction with number of workers in HH	HH
77	MOT*VEHICLES6	MOT interaction with vehicles available	HH
78	MOT*BORNUSA	MOT interaction with country of birth	Person
79	MOT*NEW_JWMN	MOT interaction with travel time	Person
80	MOT*NEW_POVERTY	MOT interaction with poverty status	Person
81	MOT*DISTANCE	MOT interaction with distance	Person
82	MOT*HHLDRAGE	MOT interaction with householder's age	HH
83	MOT*MINORITY_HH	MOT interaction with householder's minority status	HH
84	MOT*BORNUSAHH	MOT interaction with householder's country of birth	HH
85	DISTANCE	Derived distance of flow	Block flow
86	MOT*APERN	MOT interaction with person earnings	Person

Note: MOT interactions for household-level processing uses the householder's MOT.

Appendix B

List of SAS Programs

Hierarchical List of Programs by Major Component

CTPP_MAIN_DRIVER.SAS

- **T1_IRA_MAIN_DRIVER.SAS: Driver program for Initial Risk Analysis**
  - T1_1_MAIN_PROG1.SAS: Creates TAD level covariates in person file
    - T1_1_1_CREATE_INPUTS.SAS: Create input files
    - T1_1_2_CTAZ_HOUS_LEVEL.SAS: Creates TAD level covariates in household level file
    - T1_1_3_CTAZ_PERS_LEVEL.SAS: Creates TAD level covariates in person level file
  - T1_2_M30_STEP1.SAS: Creates household and person subset files
  - T1_3_M30.SAS: Sub-Driver program for person level Initial Risk Analysis
    - T1_3_1_SWAP_FLAG.SAS: Merges swap flags and GQ change flags to person file
    - T1_3_2_NEW_CEN_MERGE.SAS: Merges census block level predictors to the person file
    - T1_3_3_M30_STEP2_1.SAS: Creates person level recode variables
      - T1_3_3_1_FLGCELL3.SAS: Calls Macro GETCELL2 to create violation flags
        - T1_M_GETCELL2.SAS: Generates violation flags
    - T1_3_4_M30_VIOLATION_PERS.SAS: Creates person level full and replacement flags
    - T1_3_5_M30_SINGLETON.SAS: Creates person level singleton and stratum flags
      - T1_3_5_1_FLGCELL4.SAS: Calls Macro GETCELL2 to create violation flags
        - T1_M_GETCELL2.SAS: Generates violation flags
  - T1_4_M30_MAIN_HOUS.SAS: Sub-Driver program for household level Initial Risk Analysis
    - T1_4_1_M30_STEP2_2.SAS: Create household level recode variables
    - T1_4_2_SWAP_FLAG_HOUS.SAS: Merges swap flags and GQ change flags to household file
    - T1_4_3_NEW_CEN_MERGE_HH2.SAS: Merges census block level predictors to the household file
    - T1_4_4_M30_VHOUS_PREP_1A.SAS: Merges person level flag variables onto household level file
    - T1_4_5_M30_VHOUS_PREP_3.SAS: Creates household level violation flags
      - T1_M_GETCELL2.SAS: Generates violation flags
    - T1_4_6_M30_VHOUS_PREP_4.SAS: Creates household level full and replacement flags
    - T1_4_7_M30_VHOUS_PREP_5.SAS: Creates household level singleton and stratum flags
      - T1_M_GETCELL2.SAS: Generates violation flags
- **T2_NEW_DATA_REPLACEMENT.SAS: Driver program for data replacement**
  - T2_1_SETZERO.SAS: Sets missing values to zero
  - T2_2_SDCPERT.SAS: Performs data replacement process
    - T2_2_1_PARTIAL_FLAGS: Creates partial flags
    - T2_2_2_FULL_FLAGS.SAS: Create full flags
    - T2_2_3_VARIABLE_PREP.SAS: Prepares the list of predictors for semi-parametric approach
      - T2_2_3_1_SETVARS.SAS: Creates ACS versions of variables
      - T2_2_M_INDICATOR.SAS: Creates indicator variables for UC variables
      - T2_2_M_INTERACTION.SAS: Creates interaction terms

- T2_2_3_4_FINALIZE_PREDPOOL.SAS: Finalizes the pool of predictors for modeling
- T2_2_4_MODELING_STEPS.SAS: Runs all of the modeling steps needed for the semi-parametric approach
  - T2_4_2_1_MODEL_SELECTION.SAS: Performs stepwise model selection for the semi-parametric method
    - T2_4_2_1_1_PREDPOOL.SAS: Creates a pool of variables as the predictors
- T2_2_5_MAIN_LOOP: Main loop to perform data replacement
  - T2_2_M_INDICATOR.SAS: Creates indicator variables for UC variables
  - T2_2_M_INTERACTION.SAS: Creates interaction terms
  - T2_2_5_1_CONSTRAINEDHOTDECK.SAS: Performs data replacement using the constrained hotdeck approach
    - T2_2_5_M_CREATEBINS.SAS: Creates bin variables for constrained hotdeck approach
    - T2_2_5_M_PREDICTION.SAS: Computes the predicted values of the dependent variables in the models
    - T2_2_5_M_HOTDECK.SAS: Creates hot deck cells
    - T2_2_5_M_CHDLLOOP.SAS: Performs main steps of the constrained hotdeck approach
    - T2_2_5_M_SERPSORT.SAS: Performs serpentine sorting
    - T2_2_5_M_COLLAPSE.SAS: Collapse hot deck cells until they contain enough cases
    - T2_2_5_M_GETSEED.SAS: Generates random numbers as seeds
    - T2_2_5_M_SYNTHESIZE_OCUC.SAS: Synthesizes OC or UC variables
      - T2_2_5_M_SERPSORT.SAS: Performs serpentine sorting
      - T2_2_5_M_COLLAPSE.SAS: Collapse hot deck cells until they contain enough cases
      - T2_2_5_M_GET_SEED.SAS: Generates random numbers as seeds
  - T2_2_5_2_SEMI_PARA.SAS: Performs data replacement using the semi-parametric approach
    - T2_2_5_M_PREDICTION.SAS: Computes the predicted values of the dependent variables in the models
    - T2_2_5_M_HOTDECK.SAS: Creates hot deck cells
    - T2_2_5_M_FASTCLUS.SAS: Creates clusters for UC variables
    - T2_2_5_M_SYNTHESIZE_OCUC.SAS: Synthesizes OC or UC variables
      - T2_2_5_M_SERPSORT.SAS: Performs serpentine sorting
      - T2_2_5_M_COLLAPSE.SAS: Collapse hot deck cells until they contain enough cases
      - T2_2_5_M_GET_SEED.SAS: Generates random numbers as seeds
- T2_2_6_CHANGE_SUMMARY.SAS: Provides a summary of changed values in the file

- **T3_RAKING_DRIVER.SAS: Driver program for raking**
  - T3_1_CONTROLTOTALS_HH.SAS: Creates control total files at household level
    - T3_1_1_RECODEH.SAS: Recoding household level data
    - T3_M_SUMM.SAS: Computes control totals or sample totals
  - T3_2_CONTROLTOTALS_PERS.SAS: Creates control total files at person level
    - T3_2_1_RECODEP.SAS: Recoding person level data
    - T3_M_SUMM.SAS: Computes control totals or sample totals
  - T3_3_RAKING_HH.SAS: Performs household level raking
    - T3_M_RAKE.SAS: Performs raking
    - T3_M_CONV_REPORT.SAS: Creates convergence report in the raking process
    - T3_M_COMPTOTALS.SAS: Checks the difference between control totals and sample totals after raking

- T3_4_RAKING_PERS.SAS: Performs person level raking
  - T3_M_RAKE.SAS: Performs raking
  - T3_M_CONV_REPORT.SAS: Creates convergence report in the raking process
  - T3_M_COMPTOTALS.SAS: Checks the difference between control totals and sample totals after raking

- **T4_UTILITY.SAS: Driver program for utility analysis**
  - T4_1_ACS_CMR.SAS: Creates ACS cell means
  - T4_2_ACS_CQR.SAS: Creates ACS cell quantiles
  - T4_3_ACS_CMSE.SAS: Creates ACS standard errors
    - T4_M_CMSE_COMPUTE.SAS: Computes cell standard errors
  - T4_4_CRV.SAS: Creates Cramer's V differences
    - T4_4_1_CRV_COMPUTE.SAS: Computes Cramer's V
  - T4_5_ACS_ASSOC.SAS: Creates multivariate associations
  - T4_6_CELL_MEAN_RATIOS.SAS: Creates cell mean differences
    - T4_6_1_CMR_COMPUTE.SAS: Computes cell means
  - T4_7_CELL_QUANTILE_RATIOS.SAS: Creates cell quantile differences
    - T4_7_1_CQR_COMPUTE.SAS: Computes cell quantiles
  - T4_8_PERT_CMSE.SAS: Creates standard error differences
    - T4_M_CMSE_COMPUTE.SAS: Computes cell standard errors
  - T4_9_PAIR_ASSOC.SAS: Creates pairwise associations
  - T4_10_MULT_ASSOC.SAS: Creates multivariate associations
- **T5_RISK.SAS: Risk analysis program**
- **T6_CLEANUP2.SAS: Cleanup and creation of delivery files**