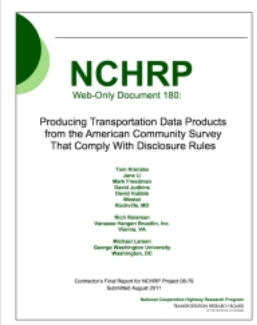


Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

DETAILS

386 pages | null | PAPERBACK
ISBN 978-0-309-43105-7 | DOI 10.17226/18160

AUTHORS

ACKNOWLEDGMENT

This work was sponsored by the American Association of State Highway and Transportation Officials (AASHTO), in cooperation with the Federal Highway Administration, and was conducted in the National Cooperative Highway Research Program (NCHRP), which is administered by the Transportation Research Board (TRB) of the National Academies.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

The **Transportation Research Board** is one of six major divisions of the National Research Council. The mission of the Transportation Research Board is to provide leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation. **www.TRB.org**

www.national-academies.org

Table of Contents

CONTENTS PAGE

Appendices

List of Figures and Tables

FIGURE PAGE

Author Acknowledgments

The research reported herein was performed under NCHRP Project 08-79 by Westat with subcontractor Vanasse Hangen Brustlin and consultant Michael Larsen of George Washington University. Due to data security requirements relating to the American Community Survey data, much of the work was done at the Census Bureau. The authors gratefully acknowledge the many individuals who contributed to the preparation of this report.

First, our thanks to Guy Rousseau, the Panel chair, and the Panel members (listed below) for their review and valuable comments and suggestions:

Guy Rousseau (chair), Jonette Kreideweis, Haoqiang Fu, Bruce Griesenbeck, Dmitry Messen, Jean Opsomer, Robert Santos, Aaron Westcott, Kevin Tierney, Ed Christopher, Elaine Murakami, Penelope Weinberger, Michael Cohen, Alison Fields, Asoka Ramanayake, Laura Zayatz, and Thomas Palmerlee.

At the Census Bureau, special thanks to Laura Zayatz, our Census Bureau contact, for her special attention to this project and to the members of the Disclosure Review Board for facilitating timely discussions. The authors are indebted to Chad Russell for accommodating our various systems needs, to Alison Fields of the operations staff, and to David Raglin for helpful arrangements for accessing the input data files.

At Vanasse Hangen Brustlin, the authors express thanks to Maggie Qi for her help with the analysis on the comparison to travel model outputs, and to Frank Spielberg for additional guidance on the travel forecasting process.

At Westat, our thanks to the senior statistical advisory group for their invaluable comments and advice. The group was led by David Judkins, with the following participating members: Graham Kalton, Mike Brick, Bob Fay, and David Morganstein.

Lastly, our thanks goes to Nanda Srinivasan, the project officer for NCHRP 08-79 who, among other management tasks, facilitated the correspondence between the research team and the National Academies of Science Transportation Research Board Panel.

Abstract

With the migration to the American Community Survey (ACS) and current Disclosure Review Board (DRB) data suppression rules, Census Transportation Planning Products (CTPP) would be severely compromised. Therefore, the National Highway Cooperative Research Program (NCHRP) investigated approaches to apply data perturbation techniques that will provide CTPP data users complete tables that are accurate enough to support transportation planning applications, but that also are modified enough that the DRB is satisfied that they prevent effective data snooping. The research team reviewed a significant amount of previous statistical community research on data disclosure research and some previous NCHRP and Federal Highway Administration (FHWA) analyses of CTPP, and settled on a small number of promising data perturbation strategies. These strategies were developed into specific procedures using Census Bureau ACS tables and microdata for four development sites, and the outputs of the different procedures were evaluated and compared in terms of data disclosure limitation and data table utility. An optimal approach that used a combination of the tested procedures was forwarded and then validated on two test sites. During the validation, the procedures were further enhanced and coded. The full procedures performed well on the validation site data, so the research team worked with Census Bureau staff to develop an operational set of computer programs that will enable the perturbation to be applied nationally. The CTPP tables that can be derived from the application of the developed procedures will enable transportation planners to make significantly better use of the ACS-based CTPP tables than they could otherwise do.

Executive Summary

The Census Transportation Planning Products (CTPP) will contain tables for about 150,000 traffic analysis zones (TAZs), other geographies, and journey to work flows, which will result in millions of tables involving dozens of variables. The main disclosure avoidance practice that had been used on certain CTPP tabulations to accomplish this objective was cell suppression. In this method, small cells were identified and suppressed, and then other related table cells that would allow the primary small cell's value to be logically deduced from the table's margins also were suppressed. The small cells were defined using the "Rule of 3," which certainly reduced the disclosure risk. Unfortunately, suppression would result in suppressed data in an estimated 80 percent or more of places in the nation, using a 10-level Means of Transportation (MOT) variable (Miller 2008) for three-year American Community Survey (ACS) tabulations. With the underlying data for the CTPP moving to an ACS five-year combined sample about half the size of the Census Long Form data, it is clear that the data loss due to Census Bureau Disclosure Review Board (DRB) disclosure rules at finer geographic areas would have been substantial. In fact, the results of the initial risk assessment on the national sample for the 2005–2009 ACS, which identified data values at most risk of disclosure, determined that about 90 percent of the TAZs were affected by DRB rules for at least one table, which would have triggered the cell suppression.

The main goal of the project was to arrive at an operationally practical data perturbation approach that satisfies the transportation data user community's analytical needs while simultaneously satisfying the disclosure rules set by the DRB. For this reason, efforts were focused on ways to generate a complete set of data containing a mix of perturbed values and real values that strive to retain the usability of the data while being acceptable to the DRB.

Provided the opportunity to address this issue, the research team, consisting of Westat, its subcontractor, Vanasse Hangen Brustlin, Inc. (VHB), and analysis consultant Dr. Michael D. Larsen, along with the Westat Senior Statistical Advisory Group, divided the research tasks into the following four phases:

1. Initial investigatory research (Chapter 1);
2. Development (Chapter 2);
3. Validation (Chapter 3); and
4. National test and transition of programs (Chapter 4).

This report documents the entire project process from start through the nationwide testing and the documentation of programs. In doing so, it describes the genesis of the recommended method and the logic used at each step in arriving at the final choice. This report provides a historical record of the decisions made along the way, so that future users will have a better understanding of the entire development, including the evaluation process that involved three perturbation approaches: parametric model-based, semi-parametric model-assisted, and constrained hot deck.

1. INITIAL INVESTIGATORY RESEARCH

The initial investigatory research provided the research parameters for this study. It advanced the authors' understanding of the tables and variables involved in the CTPP, the disclosure rule thresholds and the risk elements, the transportation user needs, operational needs, and statistical disclosure control (SDC) approaches.

Tables and variables. As discussions commenced with the Census Bureau's Disclosure Review Board (DRB), transportation experts, Census operations staff, and the statistical advisory group, it was readily apparent that establishing the set of tables and variables for the purposes of this research was needed to facilitate concrete discussion, decisions, and efficient use of resources. The research tables are provided in Appendix A. In general, tables were derived from the 2006–2010 ACS combined sample microdata by tabulating counts of individuals in cells determined by the cross-classification of values on one or more variables. There are three parts to the CTPP tables: residence-based (Part 1), workplace-based (Part 2), and residence-to-workplace flows (Part 3). Most of the three table parts contain estimates of total workers. Some tables include cell aggregates, means, and medians. There are also household-based tabulations on, for example, household income and other tabulations. The most important variable for the transportation community is the Means of Transportation (MOT), especially when it comes to the flows. Parts 1, 2 and 3 each include a table on MOT which consists of 18 categories. For small areas (smaller than counties), the MOT variable is crossed pairwise to generate cell estimates of workers with 5 variables each in Part 1 and Part 2.

Disclosure risk. In working with the Census Bureau's DRB, the authors sought to obtain clarification of disclosure thresholds, in order to meet the standards set forth through DRB disclosure rules for this special set of tabulations. One particular threat of disclosure, as recognized by the Census Bureau DRB, arises in the CTPP tables when sample uniques (singletons) exist in the marginals of MOT. When a single sample unit appears in the marginals of several tables, say, MOT* A, MOT* B,… MOT* P, tables can be linked together to define a microdata record for the sample unit consisting of MOT, A, B, … P. That is, even though the CTPP are in tabular form, tables can be linked together to form a string of identifying characteristics (referred to as a "key"). In some cases, the key could be matched to external databases, such as the ACS Public Use Microdata sample (PUMS), potentially leading to a disclosure of an individual's identity. In addition, if there is a count of two in the marginals, and a sample case can be identified in the marginal, then that case can piece together the other sample case accordingly. Therefore, in Parts 1 and 2, for pairwise cross-tabs involving MOT, the Rule of 3 was applied by the DRB to MOT marginals. In general, the "Rule of 3" based on the concept of k-anonymity specifies that at least three individuals must be represented (a count of at least three). In addition, there are a few tables that have cell aggregates and means. For these tables, the DRB applies the Rule of 3 on every cell. For Part 3, the Rule of 3 was applied by the DRB for any one-way table, other than MOT. For cross-tabs involving MOT, the Rule of 3 was applied to MOT marginals. As in the Part 1 and 2 tables, the Rule of 3 was applied to each cell of a table that involves cell aggregates and means. That is, in CTPP tables, means and aggregates must be based on at least three unweighted records for every cell. These DRB disclosure rules were used in the perturbation approach to identify high risk cells. Given that the perturbation approach, at a minimum, would target the underlying microdata contributing to those high risk cells, the DRB agreed that there would be no DRB threshold rules applied to the tables.

In addition, the risk elements associated with the delivery of the tables were identified. The risk elements pertaining to the CTPP tables included small geography, small ACS sample sizes, flow tables, and outlier trip scenarios. Through matching keys, identity disclosure and matchability to the ACS PUMS) data records was an issue. Neighbors and workmates who may have the motivation to obtain sensitive information about their acquaintances were also considered as risk elements.

Transportation user needs. In on-going discussions with VHB about the use of transportation/CTPP data in the design, development, and use in travel demand models, the authors sought to determine the variables most important in the development of travel demand models, to gain further understanding about the needs of the transportation community, and to work toward the involvement of transportation planners in the validation of the resulting perturbed data.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Operational needs. In collaboration with the Census Bureau's special tabulation group, discussions in the spring and fall of 2010 identified the datasets that serve as the basis for the research, and established relationships to help each other understand the respective needs toward assimilating a final product from this research.

Statistical disclosure control (SDC) approaches. An important step in moving toward the goal of this effort was a critical assessment of a set of promising data perturbation approaches in order to identify the most credible among the approaches, so that a small number of approaches were programmed and evaluated. The information gained in initial discussions about the CTPP tables, ACS transportation and other variables, DRB disclosure rules, transportation users' needs, and Census Bureau operational constraints and needs, established a concrete foundation on which to base these discussions and decisions. Initial suggestions resulted from a sequence of meetings between members of the research group and members of Westat's Senior Statistical Advisory Group. Table ES-1 provides a list of SDC treatments that were initially considered. The SDC approaches were evaluated based on criteria relating to impact on disclosure risk and data utility, operational practicality, applicability and flexibility to a variety of types of variables, ability to facilitate variance estimation, and ability to provide consistent results within the set of CTPP tables. Three main perturbation approaches were selected for the development phase: parametric model-based, semi-parametric model-assisted, and a constrained hot deck.

Table ES-1. Statistical Disclosure Control Treatment Options Evaluated for CTPP

Type of approach	Level of application	Approach
Deterministic	Variables	Coarsening
	TAZ	TAZ redefinition
Perturbed	Table modifications	Small area estimation
		OnTheMap approach
		Bayesian/iterative proportionate fitting
	Microdata modifications	Semi-parametric*
		Parametric modeling*
		Data swapping (later evolved into a constrained hot deck)*
		Super-sampling

Note: Terms are explained in Chapter 1.
*Selected for development phase of project.

Set A and B Tables

Discussions with the DRB lead to the decision to use perturbed data where tables are subject to DRB disclosure rules, and to use the ACS five-year data for tables where there are no disclosure thresholds. This decision was motivated by trying to retain as much observed ACS data as possible. The end result can be thought of as dividing the current CTPP tables into two sets:

- CTPP Set A (ACS five-year data tabs) based on real data and ACS weights, where the DRB agrees to release data in table format fully, without suppression, but with rounding.
- CTPP Set B (perturbed part) based on perturbed (postdisclosure proofing) data and CTPP adjusted weights, where the DRB has concerns. The Set B tables were identified and the list is provided in Appendix C.

The benefit of this setup is that data are not touched unless needed, perhaps providing better data utility to the users. However, the main disadvantage of the approach will be that different marginal totals for the same variable may exist in both Sets A and B; that is, the marginal totals may not be consistent across the set of CTPP tables for the same variable.

Operationally, the table generator will need to call the correct version of the variables, and each table will need to be checked carefully before release. The DRB has found the Set A and Set B table approach acceptable. It is important to note that they have specified that the usual rounding rules will apply to the Set A tables, as they do for the other special tabulations from ACS data. The rounding rules are applied to interior cells while fixing the marginal totals. If the marginal totals are the summation of the interior cells after rounding, this in effect can cause the marginal totals to differ for the same variable across tables. Further clarification is summarized as follows:

1. There will be two underlying microdata files as input into the Census Bureau CTPP table generator program. The first microdata file will contain all original data and the second file will contain perturbed microdata for the variables in the Set B tables.

2. The perturbed microdata file resulting from the initial risk analysis for the Set B tables on TAZ level Part 1, 2, and 3 tables will be used for all localities for the Set B tables. That is, the tables will be generated from the same perturbed microdata for all geographies including TAZs, Block Groups, Tracts, Transportation Analysis Districts (TADs), Places, Counties, States and PUMAs.

3. The perturbed microdata file will be used for TAZs where there are no violations as determined by the initial risk analysis. Even if the values of variables are unchanged, the raked weights may differ from the ACS weights, and therefore the CTPP estimates will be different from the ACS estimates.

4. The list of tables (Appendix A) contains several collapsed tables: 12201C, 12201C2 and 12201C3, for example. The collapsed versions of tables will be generated from the same perturbed microdata.

5. In Appendix C, there is reference to “Large Geography Only” for some of the tables. Large geography means county, PUMA, and state.

6. The disclosure proofing process in the research used the most detailed table in the table series (e.g., 12201 was used in the risk analysis for the series 12201, 12201C, 12201C2, and 12201C3).

7. Having more detailed tables (e.g., all based on MOT(18)) would increase the amount of perturbation in the microdata. It would also impact the DRB decisions and the perturbation rates assigned they would assign. It would necessitate a reassessment of the impact on data utility.

8. On data consistency, suppose you have residence TAZ A. All flows for Table 33204 in Part 3 involving residence TAZ A, if added together, will produce the same results as Table 13204 for residence TAZ A from Part 1. All tables will be consistent with one another within the set of tables referred to as Set B since they are all generated from the same perturbed microdata file.

2. DEVELOPMENT PHASE

As mentioned above, one of the main results of the initial research activities was arriving at three main perturbation approaches to evaluate during the development phase. The approaches were evaluated to determine the best approach for moving forward to the validation phase of the research. The evaluation had the following structure:

- Four test sites (Atlanta, Iowa, Madison, St. Louis):

- Three data perturbation approaches (semi-parametric model-assisted, parametric model-based, and constrained hot deck);

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- Two perturbation amounts (partial replacement, full replacement); and
- Five runs each.

The treatment combinations resulted in 120 total runs (4 times 3 times 2 times 5). The five runs for each evaluation combination were done to gauge the replicate variability in the data perturbation results. Evaluation measures were developed for measuring disclosure risk and data utility.

Perturbation Approaches in the Development Phase

As one of the three approaches developed for the evaluation, the semi-parametric procedure is a model-assisted approach that follows closely to Judkins et al. (2007). The process in general used model predictions to form hot deck cells in which a donor for a case with missing data was selected by a random draw without replacement from the complete cases and the missing value was filled in with the donor's original value. The replacement process was done variable-by-variable, using previously replaced data in the model selection and estimation process, as well as in the prediction equation. The process proceeded sequentially through all variables to be perturbed. The approach was adapted to handle highly variable weights, as well as incorporate the small area geographic units to bring in features that may be special to that area.

The second approach, the parametric procedure, is a model-based approach that generated perturbed data through parametric models. The process involved modeling the multivariate relationships in the observed data and generating perturbed values based on the estimated model parameters. Compared to the semi-parametric procedure, for which models were used as an instrument to assist the data perturbation, the parametric procedure had modeling as its core. The gains from the parametric procedure critically relied on the validity of the models.

The third method is a constrained hot deck approach. This approach was motivated by rank-based proximity swapping as summarized in Moore (1996), which is applicable to ordinal variables only. The modification to rank-based swapping was done with the objective of replacing the continuous version of the variable while increasing the proportion of records that changed values for the categorized versions of the same variable. The constrained hot deck approach constrained the range of the replacement values by forming bins on the target variable. The bins were used with other variables to form hot deck cells from which a donor's value was drawn without replacement from the set of all sample cases in the hot deck cell.

The constrained hot deck is applicable only for ordinal variables, so it was not applicable to unordered or binary variables such as industry and minority status. To address these two variables (industry and minority status) for the development phase evaluation, a controlled random swapping approach was used, which was based on the algorithm developed by Kaufman et al. (2005) and the underlying methodology for the U.S. Department of Education's Institute of Education Sciences *DataSwap* software. The constrained nature of the swapping approach caused problems and exponentially increased run time when attempting to find swapping partners with an increased number of targeted swapping cases. Due to these difficulties, the swapping was done only for partial replacement in the development phase evaluation: it was decided to not carry out the swapping approach for full replacement. It was also decided not to pursue the controlled random swapping approach in the validation phase evaluation and beyond.

After each perturbation approach was completed, a weight calibration process, called raking, was applied to bring consistency between ACS estimates and estimates based on perturbed data. The

calibration for the development phase was done at the person and household level for totals at the PUMA level using key variables to the CTPP.

Utility Checks: First Set

Two sets of data utility checks were conducted to determine the best approach to move forward into the validation phase. The first set of checks measured differences between raw ACS and perturbed data for cell means, weighted cell counts, standard errors, Cramer's V values, pairwise associations, and multivariate associations. The best approach resulting from the first set of checks was scrutinized further by a second set of checks comparing the perturbed data with travel model outputs. The utility checks for both the development and validation phase paid more attention to the preservation of utility when there were enough data so that the original utility was at least moderate. This means that the procedures were not graded on how they handled extremely sparse tables. They had little-to-no utility originally and will have less after perturbation. Also, showing data utility for estimates based on one or two cases, for example, is in direct conflict with the need to mask these small cells due to disclosure concerns.

Bubble plots, as in Figure ES-1 for Iowa TAZs, were generated at the county level and TAZ level for the four test sites to compare mean travel time from ACS data (x-axis) and from perturbed data (y-axis) under partial replacement. The dots in the bubble plots represent the estimate population, and are shown only for localities with 30 or more ACS sample cases. In Figure ES-1, the constrained hot deck displays a much tighter line along the 45-degree angle relative to the other two approaches, which means that the perturbed and raw ACS estimates were closer in agreement in this example.

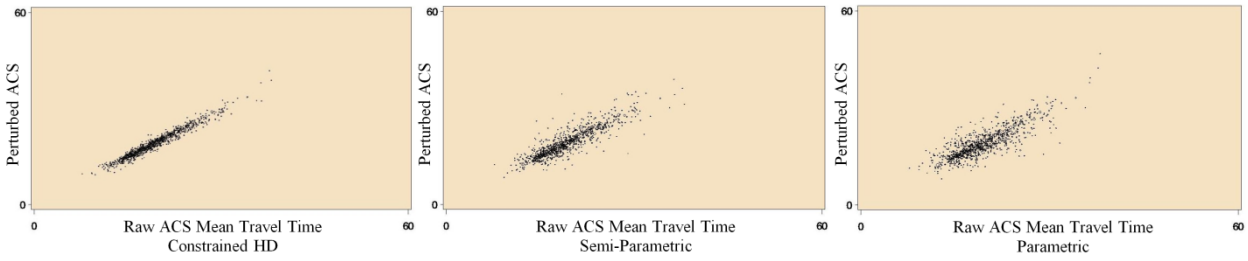


Figure ES-1. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Iowa's TAZs: Left: Constrained ot deck; Middle: Semi-Parametric; Right: Parametric (n ≥30)

The conclusions from the first set of checks in the development phase evaluation were as follows:

- The constrained hot deck approach impacted the resulting cell means the least, followed by the semi-parametric approach and then the parametric approach.
- The constrained hot deck approach and semi-parametric results each showed the least impact on the weighted cell counts, with the parametric approach resulting in the greatest amount of dispersion from the ACS estimates.
- The constrained hot deck approach had the least impact on perturbation error variance measures, with the parametric approach resulting in the largest impact.
- The constrained hot deck approach had the least impact on the Cramer's V measure on two-way tables, especially at the TAZ level. The results showed the parametric approach having the greatest impact.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- The constrained hot deck approach retained the pairwise correlations the best, with semi-parametric doing quite well, and the parametric approach doing well in many instances.
- Lastly, the constrained hot deck approach had the lowest values of the multivariate association measure (U), with mixed results from the semi-parametric approach and the parametric approach.

Therefore, given the above conclusions from the first set of checks, the constrained hot deck approach was chosen for the home-based work output comparisons (the second set of utility checks).

Disclosure Risk Review

The development phase evaluation results of the disclosure risk measures were shared with the DRB. A summary of the DRB meeting is provided in this report. The DRB accepted the risk levels, using the partial replacement rates that had been reported to them for each of the three approaches. The DRB requested further investigation into the remaining high risk cases. Subsequently, further results from the additional investigation were considered acceptable.

Utility Checks: Second Set

With the DRB acceptance of the results associated with partial replacement, the comparison with travel model outputs used the constrained hot deck results from partial replacement. The stated purpose of these comparison tests with travel model data was to conduct a reasonableness check to determine if the performance of the perturbed ACS CTPP tabulations was no worse than the raw tabulations when compared against typical model outputs.

In the second set of data utility checks, the perturbed ACS data were compared directly with travel model outputs from the four test sites. The comparisons looked at residence-based tabulations that included age categories and number of workers in households (HHs). In the comparisons, county and subcounty estimates were compared. The finding of the development phase was that the performance of the perturbed ACS tabulations was equal to that of the raw ACS when compared to model output.

Despite taking the steps to make the tests manageable, the research team encountered issues that required the elimination or reduction of several planned comparisons. Some test site models did not include the identified variable (age, income, etc.), in other cases, the variable was included in the test site's model, but the categories specified differed from those used in the ACS. Geographic compatibility for TAZs also presented an additional challenge.

A broader question was raised on how well ACS-based (either raw or perturbed) CTPP tabulation compares on its own to model output. Clearly, there were levels of difference between the model output and ACS for certain individual counties, districts, or TAZs (or pairs of each, in the case of flow data). Were these comparisons being conducted as part of a full model development or revalidation, further investigation into both the reliability of the ACS-based estimates and the uncertainty of the model estimates would have been required.

Conclusion from the Development Phase

The conclusion from the development phase was that the constrained hot deck approach was the best approach for ordinal variables. Since the constrained hot deck is only applicable to ordinal variables,

the semi-parametric approach was selected for the unordered categorical and binary variables during the validation phase. The semi-parametric approach has the benefit of combining the model predictions with the special characteristics of the localities of interest with regards to such variables as industry and minority status.

3. VALIDATION PHASE

The main goals for the validation phase were to confirm the following for the proposed data perturbation approach on disclosure avoidance:

- Compliance with the Census Bureau's Disclosure Review Board (DRB) disclosure rules for the CTPP;
- Preservation of the properties of the original CTPP data; and
- Operational feasibility of the data replacement approach in the CTPP application.

After the Interim Report meeting on October 25, 2010, the research team was focused on several aspects of the proposed data perturbation approach. The activities were centered on implementing the data perturbation technique on two test sites using the five-year accumulation of 2005–2009 ACS data. The research team made the necessary modifications to the software to incorporate the approach into CTPP data production. The research team then verified the retention of data utility of the disclosure-proofed data (as compared to raw ACS data) in the same manner as in the evaluation conducted in the development phase. The following list provides a more detailed summary of the efforts relating to the validation phase:

- Additional CTPP tables. The three new variables were added for Part 1 and Part 2 tables are workers in households (0, 1, 2+), minority status (Y/N), and presence of children 17 and under in household (Y/N).
- National implementation. With an eye moving forward, changes were made to the processing in order to improve the operational feasibility of the production process.
- Composite two perturbation approaches. Computer programs were prepared to combine the constrained hot deck and the semi-parametric approaches into one data replacement step.
- Raking (weight calibration procedure). A raking dimension was added at the combined TAZ level, where the combined TAZs have at least 300 ACS sample cases (CTAZ300). Three hundred (300) ACS records represent about 4,000 workers. Table ES-2 and ES-3 provide the raking dimensions for the household and person weight calibration adjustments, respectively.

Table ES-2. Validation phase: Raking dimensions for the Household File

Dimension	ByVar1	ByVar2
1	PUMA	Vehicles available (6)
2	PUMA	Number of workers in HH (6)
3	PUMA	HH income (5)
4	Residence CTAZ300	--

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Table ES-3. Validation phase: Raking dimensions for the Person File

Dimension	ByVar1	ByVar2
1	PUMA	Vehicles available (6)
2	PUMA	Number of workers in HH (6)
3	PUMA	HH income (5)
4	Place of work PUMA	HH income (5)
5	PUMA	Travel time (4)
6	PUMA	MOT(6)
7	Place of work PUMA	MOT(6)
8	Residence CTAZ300	--

- Utility measures. The research team provided the variance formula to the Census Bureau's DRB, and the DRB approved the formula for production. The research team also described to the Census Bureau's ACS operations staff how the variances for the resulting perturbed data are to be calculated and provided justification of its use to the ACS Statistical Design group. In terms of the utility validation, a comparison of medians and 75th percentiles between the perturbed data vs. raw data was included.

- Travel model outputs. The development phase approach for comparing to travel model outputs was applied to Olympia, WA, and Atlanta, GA for the validation phase.

- Simulation. The Panel requested further investigation into any potential there could be for bias. Therefore, the research team conducted a simulation to identify any potential bias and impact on variance due to perturbation on drive-alone travel times in Olympia, WA.

- Population synthesis. Efforts were made to have Atlanta's population synthesizer (which consists of Java programs) processed at the Census Bureau for testing the impact of the perturbation on population synthesis. Such efforts proved onerous, and these attempts were later dropped.

- Census staff operations needs. More discussion occurred with the ACS operations staff regarding implementation of the Set A and Set B tables, as well as the variance estimation approach, leading up to the implementation of the approach in the production run on ACS 2006–2010 data.

- Data users' needs. The research team gave a presentation at the Transportation Research Board annual meeting on January 23, 2011, giving data users the opportunity to raise concerns about the impact of perturbation procedures on data quality. The research team also participated in the CTPP table subcommittee meeting held on May 2, 2011.

- Statistical methodology. The research team conducted a presentation on January 18, 2011, for the Washington Statistical Society. The presentation covered the basic contents of the Interim Report, including results from the development phase evaluation. The research team fielded questions, such as how weights were used, how far the approach could be generalized, and how travel model outputs were considered in the evaluation.

Validation Phase Results

The processing for the validation phase began with a nationwide initial risk analysis. The initial risk analysis was used to identify the data values at most risk, and there were some key results to report. The initial risk analysis on the five-year ACS (2005–2009) revealed that 60 percent of the TAZ flows (using Census 2000 definitions of TAZs) were singletons (contain just one ACS sample record), while 90 percent of the TAZ flows are singletons or doubletons (contain just one or two ACS sample records).

After the initial risk analysis, which identified data values at high risk, the processing continued with data replacement on two test sites (Atlanta and Olympia). Processing used one combined data perturbation approach (semi-parametric for unordered categorical and binary variables, and the constrained hot deck for ordinal variables), one perturbation amount (partial replacement), and five runs each. After data replacement, a raking procedure was conducted to bring consistency between raw ACS estimates and perturbed estimates.

As in the development phase, two sets of data utility checks were conducted, as well as a set of disclosure risk checks. With regard to disclosure risk, before the validation phase processing, the partial replacement rates were modified in consultation with the Census Bureau's DRB. After processing the data replacement and raking process, summary tables showing disclosure risk measures were produced and provided to the Census Bureau's DRB for review on February 3, 2011. The indications of disclosure risk were found to be at an acceptable level. Therefore, the Census DRB provided approval to move ahead to the nationwide testing and production run with the partial replacement rates used in the validation phase.

The first set of data utility checks explored differences between raw ACS and perturbed data. Reported were cell means, medians and 75th percentiles, weighted cell counts, standard errors, Cramer's V values, and pairwise and multivariate associations. The validation phase results were compared with the results from the development phase. Several bubble plots were generated at the county level and TAZ level to compare mean travel time from ACS data (x-axis) and from perturbed data (y-axis) under partial replacement. To illustrate, Figure ES-2 shows results for Atlanta TAZs for the development phase and for the validation phase. The figure shows some minimal deviations, although the deviations are slight as determined by the tightness to the 45-degree line and consistent with the development phase plot (left plot in figure).

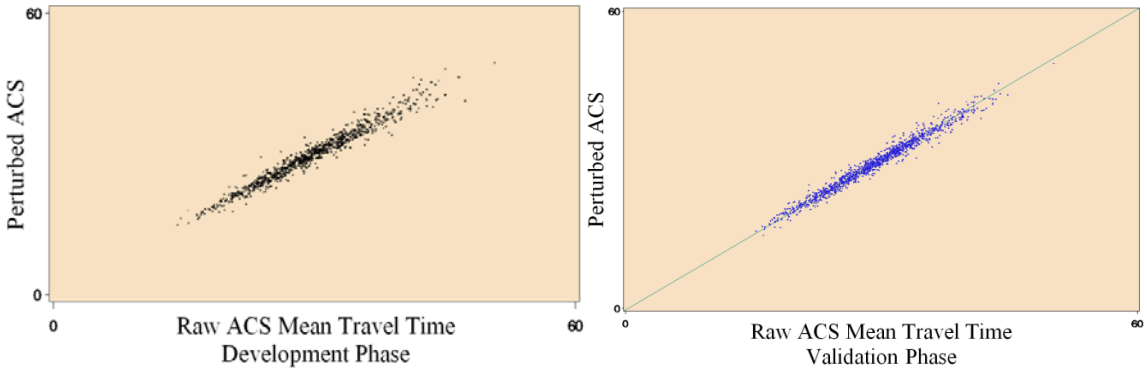


Figure ES-2. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's TAZs: Left: Development Phase, Right: Validation Phase (n≥30)

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Conclusions and Limitations

The conclusions from the validation phase's first set of checks were as follows:

- The cell means, medians, and 75th percentile analysis clearly supported the favorable results given by the constrained hot deck in the development phase. That is, the impact of the perturbation approach was at an acceptable level and there was little indication of bias introduced by the perturbation approach, as also seen by simulation results. There was some indication that the results had improved since the development phase due to changes to the application of the constrained hot deck.

- The analysis on weighted cell counts revealed the same acceptable level of impact from the perturbation approach as seen in the development phase. In general, there was minimal impact at the county level, and more impact at the TAZ level. Additional checks that were conducted on tables involving industry showed favorable results as well.

- The analysis on the perturbation's impact on standard error showed about the same level of impact as determined acceptable in the development phase. The simulation results helped to give an indication that the perturbation impact on the variances at a combined TAZ level (populations of about 4,000 workers) for mean travel time was not significant. In general, one can expect the perturbation to increase standard errors in most cases by 3 to 10 percent for areas of that size.

- The analyses involving Cramer's V, pairwise associations, and multivariate associations all showed results similar to the development phase. Some results give minor indications of correlations being attenuated relating to age. A couple of simple adjustments were later made to the specification of the constrained hot deck as it related to perturbing age in order to reduce the attenuation of such correlations.

The second set of checks involved the comparison of the perturbed data with travel model outputs for Atlanta and Olympia. The results for the validation phase largely replicated those of the development phase: the perturbed ACS CTPP tabulations performed equally well as the raw ACS CTPP tabulations when compared against typical model outputs. Testing for both Atlanta and Olympia indicated that ACS cell sample size will create data usability issues for transportation planners at fine levels of geography (e.g., TAZs) for cross-tabulations of key variables with means of transportation.

The research team made slight modifications to existing programs to get ready for the production run. The main focus of the changes was to have the five-year ACS data processed through six main programming components of the procedure without human intervention. The steps for the approach are organized as follows:

1. Initial risk analysis;

2. Data replacement approach, which includes partial replacement using the semi-parametric for unordered categorical and binary variables, and the constrained hot deck for ordinal variables;

3. Weight calibration—raking, which includes generating control totals;

4. Data utility measures;

5. Risk measures; and

6. Cleanup.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

The research team conducted a dry run of these steps, made further adjustments, and documented the process and programs. The research team collaborated with the ACS operations staff to ensure there was an understanding about the files necessary from the ACS to serve as input to the perturbation program, and the output from the perturbation so that it would ensure a smooth transition into the ACS CTPP table generator. Through the collaboration, the two main modifications to the CTPP processing, which included the Set A and Set B tables, and the variance estimation process, were highlighted. The research team continued to monitor the processing time for the nationwide test run. The trial on the national data ran successfully and took a total of about 24 hours.

Limitations of Approach

The perturbation approach was developed specifically for the purpose of generating CTPP data tables. Certainly careful attention is necessary for any project needing statistical disclosure control. Each dataset is unique in that it has its own complex data structures, as well as different emphasis on various uses of the data and analyses that are expected.

The CTPP tables are generally one- or two-way tables for a given geography or flow, which greatly reduces the disclosure risks that would exist under a microdata release. Essentially a microdata release of 20 CTPP variables provide a 20-way table for every TAZ flow, which would have a high risk of disclosure. A microdata release would require the retention of unit-level correlations and multivariate associations among the 20 variables. Even though the perturbation approach developed in this research creates underlying microdata for the creation of the CTPP tables, the approach in its current form would not retain three-way and several two-way interactions among the 20 variables. Splitting the CTPP into Set A and Set B tables made it possible to focus on the important interactions to retain, as explicitly given in the table structures.

The semi-parametric approach, as applied here, considered retaining all pairwise relationships, and some three-way relationships. However, due to weak associations, the point estimates were not as good (noisier) as when using the constrained hot deck for some specific analyses. The constrained hot deck was developed to limit the change in the detailed variable while considering changes necessary at the coarsened version of the variable. With some limited control on retaining multivariate relationships, and with the motivation for applying “change-as-necessary,” the associations between variables are generally retained at an acceptable level.

If a microdata file was being considered as the data product, more perturbation would be necessary and more work needed to retain relationships between variables. If more variables are released, the semi-parametric approach is more generalizable, whereas the application of the constrained hot deck requires a bit more attention. More careful attention is necessary when complex questionnaire skip patterns exist.

In general, the resulting CTPP tables have less impact from sampling and perturbation error for larger areas. Both types of error are generally driven by the sample size of the ACS. For smaller geographic areas, the sample size is spread very thin (e.g., 90 percent of TAZ flows under the 2000 Census definition have just one or two ACS sample records), and therefore more perturbation is necessary. The resulting standard errors, which include the perturbation error component, will provide the basis for judging whether or not the data are reliable.

Future Research

The procedures developed under NCHRP 08-79 are fully operational, and the next step would be the implementation of the procedures by the Census Bureau. As the procedures are applied and CTPP data users develop analyses with the data, futher refinements or research needs may be identified. Some specific possibilities for further research that might improve upon the present approach include the following:

Compare multiple dataset variance estimation approaches. This would expand the research team's variance estimation approaches that were conducted for the NCHRP 08-79 project. The research team developed an approach based on a single dataset and compared it with results from a multiple dataset formula (Reiter 2003) and other approaches. Further research could provide more stability to the new variance estimation approach through multiple perturbations while modifying the current variance formula. Other new approaches could be considered, developed, and compared. For example, a limited bootstrap could be done by conducting the perturbation of the original data multiple times (perhaps the same number as the number of replicate weights and full sample weight), producing the replicate estimates and subsequently the variance among the estimates. Another idea is a modest adjustment to the newly developed variance approach utilizing shrinkage estimation.

Compare approaches for identifying high risk values in microdata. It is important to identify high risk data values in the data in order to help determine recodes, variable suppression, and values to target in perturbation. Further research could be done to compare an exhaustive tabulation approach to an approach by Elliot et al. (2002), and to other possible approaches that may be developed. Evaluation measures would be developed for the comparison. The results would lead to a possible standard approach to identifying high risk data values.

Develop and compare approaches to perturb spatial outliers, and non-spatial outliers. Spatial outliers are more apparent as maps are used more and more when analyzing data. The spatial outliers may be considered a disclosure risk in mapping journey to work. The approach would consider characteristics of individuals as well as constraints on the movement of the data points. An evaluation could be designed to gauge the impact of the masking approaches.

Evaluate scenarios to balance the use of weights, with model predictions, and size of locality. During the development of the semi-parametric and the constrained hot deck approaches in the NCHRP 08-79 research, there was a need to form hot deck cells from groups of weights, groups of model predictions or covariates, and the locality of the target records. If the weights varied greatly, then the weights would have had more influence on the perturbation. If the model was good, then the model predictions would have had more influence on perturbation. If the localities were very different from each other, then locality would have had more influence. An evaluation could be conducted on the semi-parametric and constrained hot deck, using scenarios relating to the variation in the weights, quality of models, and variation among localities. The results would show the impact on various data utility measures, and would help guide decisions on applying the perturbation approaches.

Evaluate the potential combination of multiple data sources. This research would look into the possibility of borrowing strength from other data sources related to the CTPP. For example, integrating the OnTheMap estimates with the ACS Journey to Work estimates for the CTPP five years could be examined. The first step would be to do a comparison of estimates from the public version of the CTPP data and what OnTheMap produces. Basic questions about alignment in variable definitions, geography, and national coverage need to be investigated. Then one must determine if estimates agree across a broad spectrum of interests and places. Collaboration and input from transportation analysts would be necessary. The investigatory research could lead to using the OnTheMap estimates as predictors

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

in the semi-parametric approach, or as hot deck cell variables in the constrained hot deck, or to possibly creating the composite of the two estimates from OnTheMap with ACS.

1. Introduction to Research Investigations

The main goal of the project was to arrive at an operationally practical data perturbation approach that will satisfy the transportation data user community's analytical needs and satisfy the disclosure rules set by the Census Bureau Disclosure Review Board (DRB). The main disclosure avoidance practice that was used on certain Census Transportation Planning Products (CTPP) tabulations to accomplish this objective was cell suppression. First, small cells were identified and suppressed, and then other related table cells that would allow the primary cell's value to be logically deduced from the table's margins also were suppressed. The small cells were defined using the "Rule of 3," which reduced the disclosure risk, although this would result in suppressed data in an estimated 80 percent or more of places in the nation, using a 10-level Means of Transportation (*MOT)* variable (Miller 2008) for three-year American Community Survey (ACS) tabulations.

With the underlying data for the CTPP moving from the Census Long Form data to the smaller ACS five-year combined sample, it was clear that the data loss at finer geographic areas, such as planned Traffic Analysis Zones (TAZs) would be substantial on five-year American Community Survey (ACS) data due to DRB disclosure rules. For this reason, efforts were focused on ways to generate a complete set of data containing perturbed values that strived to retain the usability of the data.

Provided the opportunity to address this issue, the project team, consisting of Westat, its subcontractor, Vanasse Hangen Brustlin, Inc. (VHB), and analysis consultant Dr. Michael Larsen, along with the Westat Senior Statistical Advisory Group, conducted the initial tasks as laid out in the working plan document. As part of the kickoff meeting on January 26, 2010, the panel (Transportation Research Board panel members for NCHRP 08-79) provisionally approved the initial plans discussed in the working plan document and provided recommendations related to travel model validation that are discussed later in this document.

The research was divided into the following four phases:

1. Research investigations (Research Tasks 1 and 2, Chapter 1);
2. Development (Research Tasks 3, 4, and 5; Chapter 2);
3. Validation (Research Task 6, Chapter 3); and
4. National test and transition of programs (Research Task 7, Chapter 4).

The main result of the research activities was arriving at three main perturbation approaches to evaluate. The approaches were evaluated in the development phase to determine the best approach for moving forward to the validation phase of the research. Also described are the data utility and disclosure risk measures that have been developed in order to assess the performance of the perturbation approaches. The methodology is described in Chapters 2, 3 and 4, since the approaches changed slightly during each phase.

After the development phase and validation phase results were generated, the DRB was provided the documentation that discussed the variables perturbed in the development phase and the percentage of values that were replaced. The variables to be perturbed will not be known to the public and the nondisclosure of the list is considered vital in importance. All variables mentioned in this document are discussed as a subset of preliminary variables focused on for the research.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

The results of the research investigations provided the research parameters, which included the following:

- **Tables and variables.** As discussions commenced with the DRB, transportation experts, Census operations staff, and the statistical advisory group, it was readily apparent that establishing the set of tables and variables for the purposes of this research was needed to facilitate concrete discussion, decisions, and efficient use of resources.

- **Disclosure thresholds**. In working with the DRB, the research team sought to obtain clarification of disclosure thresholds, in order to ensure meeting the standards set forth through DRB disclosure rules for this special set of tabulations.

- **Transportation user needs**. In on-going discussions with VHB in the use of transportation/CTPP data in the design, development, and use in travel demand models, the research team sought to determine the variables most important in the development of travel demand models, to gain further understanding of the needs of the transportation community, and to work toward the involvement of transportation planners in the validation of the resulting perturbed data.

- **Operational needs**. In collaboration with the Census Bureau's special tabulation group, the team identified the datasets that served as the basis for the evaluations in the spring and fall of 2010, and established relationships to develop mutual understanding of the requirements for assimilating a final product from this research.

An important step in moving toward the goal of this effort was a critical assessment of a set of promising data perturbation approaches in order to identify the most credible among the approaches, so that a small number of approaches needed to be programmed and evaluated. The information gained about the tables, variables, DRB rules, transportation users needs, and operational needs worked toward establishing a concrete foundation on which to base these discussions and decisions, which resulted from a sequence of meetings between members of the research group (Mark Freedman, Tom Krenzke, Jane Li, David Hubble, Michael Larsen), and members of the Senior Statistical Advisory Group (David Judkins, Graham Kalton, Mike Brick, Bob Fay, David Morganstein). Three main perturbation approaches were selected for the development phase.

Section 1.1 provides more details of the work conducted in the initial research phase, relating to the tables, variables, and disclosure thresholds. Section 1.2 discusses the involvement of transportation planners, while Section 1.3 discusses the involvement of the ACS operations staff. An overview of the critical assessment of data perturbation approaches considered for the CTPP is provided in Section 1.4.

1.1 CENSUS DRB RULES ON ACS FIVE-YEAR TABULATIONS AND DISCLOSURE RISK ELEMENTS IN CTPP TABLES

The research began by reviewing the Census Bureau DRB rules on ACS five-year tabulations and disclosure risk elements (i.e., factors that affect disclosure risk) in CTPP tables. The general structure of the CTPP tables are described first; then the risk elements are discussed.

1.1.1 General Structure of CTPP Tables

In general, tables are derived from microdata by tabulating counts of individuals in cells determined by the cross-classification of one or more variables. In sample surveys, the survey weights for

individuals in a table cell are added to produce a weighted count in the cell. In applications such as business, establishment, or transportation studies, one can summarize information on a quantitative variable (total, mean, standard deviation, median, or percentiles) for individuals within table cells defined by other variables. Categorical variables with nominal or ordinal values and discrete quantitative variables with relatively few values can be directly used for classification of subjects.

The new CTPP product will be processed from the 2006–2010 ACS combined sample. There are three parts to the CTPP tables: residence-based (Part 1), workplace-based (Part 2), and residence-to-workplace flows (Part 3). In February 2010, a set of CTPP tables for this research was approved by the CTPP Advisory Board. The initial set of tables can be found in the Technical Memorandum for Tasks 1 and 2 (Westat 2010). In June 2010, AASHTO reduced the number of tables and proposed a new draft set of tables, which are presented in Appendix A of this report. Within the tables is a one-way flow table approved by the DRB for 18 categories of Means of Transportation (MOT). Most of the Part 1 tables, Part 2 tables, and Part 3 tables contain estimates of total workers, although some tables include cell aggregates, means, and medians. There are also household-based tabulations on household income, for example, as well as other universe differences between the tables.

Among the most important variables for the transportation community is the MOT, especially when it comes to the flows. Parts 1, 2 and 3 each include a table on MOT that consists of 18 categories. For small areas (smaller than counties), the MOT variable is compressed into fewer categories and crossed pairwise to generate cell estimates of workers with other variables in Part 1 and Part 2.

In Part 1, eight variables are crossed with MOT as follows (the number in parentheses refers to the number of categories in the variable, including the total):

- Crossed with MOT(11): Age of Worker(8), Travel Time(12), Household Income(26), Vehicles Available (6), Under 18 (3), Minority status (3), Number of workers in household (3); and
- Crossed with MOT(7): Time Leaving Home (10).

For Part 2, Time Arriving(17) is substituted in the above list for Time Leaving Home(10).

For Part 3, for small areas (defined below), MOT is crossed pairwise with only four variables, to obtain cell estimates of the number of workers as follows:

- Crossed with MOT(7): Time Leaving Home(5), Household Income(5), Vehicles Available(4); and
- Crossed with MOT(4): Travel Time(12).

Other variables involved in the small area flows are Age of Worker (8), Industry(8)[1]—overall and excluding self-employed, Time Leaving Home(17), Minority Status(3), Travel Time(12), Household Income(9), and Poverty Status(4).

The tables that include cell medians, aggregates, and means are shown in Table 1-1.

Small areas. There is a distinction in Appendix A between tables slated for large areas and tables for small areas. Small areas involve areas smaller than county, for example, block groups, tracts, places. For flows, this is transparent in the TAZ variable in the Census datasets; that is, TAZ may be defined as

[1] Industry(8) has been proposed by the research team and agreed to by AASHTO. Appendix A tables still mention Industry(15).

NOTE: R = residence based; W = workplace based; F = Flows.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

block groups, tracts, places by the Metropolitan Planning Organization (MPO), or defined as the default (tracts) for areas where TAZs are not explicitly defined. Large areas are defined as counties or areas larger than a county, such as states.

Table 1-1. CTPP Tables with Cell Aggregates, Means, and Medians

Part	Table	Variable for Cell Aggregates, Means and Medians	Subgroup
1(R)	TimeLeavingHome(5)	Workers per carpools (no median)	Workers 16 years and over using carpools
	TimeLeavingHome(5)	Workers per car, truck or van (no median)	Workers 16 years and over who used car, truck or van
	TimeLeavingHome(5)	VehiclesUsed (no median)	Workers 16 years and over using car, truck or van
	MOT(18)	TravelTime	Workers 16 years and over who did not work at home
	MOT(11)*TimeLeavingHome(17)	TravelTime	Workers 16 years and over who did not work at home
	NumWorkers(6)	HHIncome	Households
	VehAvail(6)	HHIncome	Households
2(W)	TimeLeavingHome(5)	Total workers in carpools (no median)	Workers 16 years and over using carpools
	TimeLeavingHome(5)	Workers per car, truck or van (no median)	Workers 16 years and over who used car, truck or van
	TimeLeavingHome(5)	VehiclesUsed (no median)	Workers 16 years and over using car, truck or van
	MOT(18)	TravelTime	Workers 16 years and over who did not work at home
	MOT(11)*TimeArriving(17)	TravelTime	Workers 16 years and over who did not work at home
3(F)	MOT(7)	TravelTime	Workers 16 years and over who did not work at home
	MOT(7)*TimeLeavingHome(5)	TravelTime	Workers 16 years and over who did not work at home

NOTE: R: Residence-based; W: Workplace-based; F: Flows.

1.1.2 Disclosure Rules

Through frequent discussions between the research team and Census Bureau DRB chair Laura Zayatz, and a pivotal meeting with the Census Bureau DRB in January 2010, an understanding of the DRB disclosure rules was established. The rules and how they relate exactly to the CTPP tabulations were clarified. Table 1-2 provides a summary of the DRB disclosure rules as discussed and confirmed by the Census DRB.

Table 1-2. Disclosure Rules for CTPP Tables Based on the Five-Year ACS

#	Part	Table type	Example Table	Initial Risk Rule	Post-Synthetic Rule
1	1 (R)	--	Total workers(1)	No threshold	No threshold
2	1 (R)	1-way, 2-ways, etc (no MOT)	R* Vehicles available(6)	No threshold	No threshold
3	1 (R)	MOT 1-way	R* MOT(18)	No threshold	No threshold
4	1 (R)	MOT* X	R* MOT(11)* Age of worker(8)	MOT(11) marginals must have at least 3 unweighted records.	No threshold
5	1 (R)	Means, aggregates	R* MOT(18) – Aggregate Travel Time	Means and Aggregates must be based on at least 3 unweighted records for every cell	No threshold
6	1 (R)	Medians	R* MOT(18) – Median Travel Time	Medians are an interpolation from a frequency distribution of unrounded data (not subject to rounding), or as a point quantile rounded to two significant digits with at least 5 cases on either side of the quantile point.	No threshold
7	2 (W)	Same as above	Same as above	Same as Part 1 Residence tables	No threshold
8	3 (F)	Total	Total workers(1)	No threshold	No threshold
9	3 (F)	1-way	F* Poverty(4)	Must have at least 3 unweighted records in flow (F).	No threshold
10	3 (F)	MOT 1-way	F* MOT(18)	No threshold	No threshold
11	3 (F)	MOT*X	F* MOT(7)* HH Income(5)	MOT(7) marginals must have at least 3 unweighted records.	No threshold
12	3 (F)	Means, aggregates	F* MOT(7)* Time Leaving Home (5)– Mean Travel Time	Means and Aggregates must be based on at least 3 unweighted records for every cell	No threshold
13	3 (F)	Medians	F* MOT(7)* Time Leaving Home (5)– Median Travel Time	Medians are an interpolation from a frequency distribution of unrounded data (not subject to rounding), or as a point quantile rounded to two significant digits with at least 5 cases on either side of the quantile point.	No threshold

NOTE: R: Residence-based; W: Workplace-based; F: Flows.

The motivation and rationale for the DRB disclosure rules are as follows:

- One particular threat of disclosure, as recognized by the Census DRB, arises in the CTPP tables when sample uniques (singletons) exist in the marginals of MOT. When a single sample unit appears in the marginals of several tables, for example, MOT* A, MOT * B,… MOT * P, tables can be linked together to define a microdata record for the sample unit consisting of MOT, A, B, … P. The resulting microdata record then reveals a lot of information about a certain individual; that is, even though the CTPP are in tabular form, tables can be linked together to form a string of identifying characteristics (referred to as a "key"). In some cases, the key could be matched to external databases, such as the ACS Public Use Microdata Sample (PUMS) in the context of CTPP tables. Matching to microdata in an external source would further compromise the confidentiality of information.

- In addition, if there is a count of two in the marginals, and a sample case can be identified in the marginal, then that case can piece together the other sample case accordingly.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- Therefore, in Parts 1 and 2, for pairwise cross-tabs involving MOT, the Rule of 3 is applied to MOT marginals. In general, the Rule of 3, based on the concept of k-anonymity, specifies that at least three individuals must be represented (a count of at least three). If there is only one, then it is a unique person in the sample. If there are only two, then each person in the sample from that cell knows that there is only one other person in the sample in that cell. If there are three or more, one person cannot make a statement about someone being unique in the sample without additional information.

- In addition, there are a few tables that have cell aggregates and means. For these tables, the Rule of 3 is applied on every cell.

- For Part 3, the Rule of 3 is applied for any one-way table, other than MOT.

  - For cross-tabs involving MOT, the Rule of 3 is applied to MOT marginals.

  - As in the Part 1 and 2 tables, the Rule of 3 is applied to each cell of a table that involves cell aggregates and means; that is, in CTPP tables, means, and aggregates must be based on at least three unweighted records for every cell. As with counts, if two people contribute data to a total (or mean), then one of those people can determine the other person's value by subtraction.

- Medians are computed whenever means are computed. Medians will likely be computed as an interpolation from a frequency distribution of unrounded data (not subject to rounding).

It was also recognized that the DRB disclosure rules may be used to identify high risk cells. Given that the perturbation approach, at a minimum, would target the underlying microdata contributing to those high risk cells, there would be no DRB threshold rules applied to the tables.

1.1.3 Risk Elements

Neighbors, extended kin, friends, and workmates may have the motivation to obtain sensitive information about their acquaintances. If obtained, the disclosure could be of three types, as discussed in Federal Committee on Statistical Methodology (FCSM) (2005): "identity," attribute, or inferential. Identity disclosure occurs if a data snooper can identify a person from the highly identifiable released data, such as residence, workplace, MOT, age, earnings, industry, length of U.S. residence, sex, occupation, and household income. Attribute disclosure occurs when sensitive information about a person, such as earnings, household income and poverty status, is revealed. Inferential disclosure happens when data can be inferred with high confidence from statistical properties of the released data. The DRB disclosure rules established are an attempt to alleviate concerns about identity and attribute types of disclosure, which may arise through various risk elements. The risk elements pertaining to the CTPP tables include the following:

Small geography. With 166,000 TAZs from Census 2000, the size of TAZs is roughly similar to block groups. The smaller the geography, the more a data snooper can reduce the universe of possibilities.

Small ACS sample sizes. As illustrated in Table 1-3, by design, the ACS five-year sample size is expected to only be about 44 percent of historical Census Long Form design sample sizes. Even with an expected 11 percent growth in the number of housing units between 2000 and 2008 (the middle year of the 2006–2010 five-year period to be used for the first ACS-based CTPP), the ACS sample size is expected to still only be about 49 percent of the Census 2000 Long Form sample size. Essentially, the smallest TAZs will have just 20–25 ACS sample workers in TAZs with a total population of about 600 people.

Table 1-3. Comparison of ACS and Census 2000 Long Form Sample Sizes

Survey		Sampling Rate Ratio	Sampling Rate Percent	Sampling Rate After Nonresponse	ACS Design as Percent of Long Form Design	Estimated ACS Sample Size (2006-2010) as Percent of Census 2000 Long Form Sample Size Assuming Growth in Number of Housing Units
Census 2000 Long Form		1 in 6	16.7	16.7		
ACS	1 yr	1 in 45	2.2	1.5	9	
	3 yr	1 in 15	6.7	4.4	27	
	5 yr	1 in 9	11.1	7.4	44	**49**

Flow tables. With the TAZ size thresholds remaining the same and with a smaller underlying sample size, the set of flows for each TAZ is likely to result in a majority with sample uniques.

Outlier trip scenarios. Population uniques are likely for scenarios such as long distance bicycle/walker commuter from known point A to known point B.

Identity disclosure and matchability to the ACS PUMS data records. In other words, a risk in the set of CTPP tables is the ability to link the tables to build a microdata record, and then using the CTPP variables, match to the ACS PUMS to obtain about 150 variables for the record. Census Bureau rules are to not show microdata for small geographies. Given a match, there is a high probability of a true record match success.

Neighbors. Extended kin, friends, and workmates may have the motivation to leverage their knowledge of specific people's attributes to obtain sensitive information about their acquaintance.

1.1.4 Addressing the Risk Elements

Investigate the Impact of TAZ Sizes on Risk

Certain disclosure risk elements are associated with the population size of TAZs. The largest fundamental shift in these risk elements relates to the transitioning from the Census Long Form serving as the CTPP data source to the American Community Survey (ACS) five-year data files. As shown in Table 1-3, the number of ACS sample cases is only expected to be about 50 percent of what was realized from the Census 2000 Long Form. A clear consequence of this reduced sample size is an increase in the number of TAZs that would present a disclosure risk as defined by the DRB and discussed in Section 1.1.2.

Consider the following investigation into the tradeoffs associated with "small" TAZs:

- Let A = a TAZ formed as the minimum size under the current rules.
- Let B = regular size TAZ.
- Suppose A needs a lot of masking due to very many small cell sizes.
- Suppose the degree of masking is represented by A".
- Suppose B needs some masking due to some small cell sizes.
- Suppose the degree of masking is represented by B', but less masking than for A, as represented by one apostrophe instead of two used for A.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- Let AB = a TAZ that is needed to exceed a new rule of twice the current minimum threshold.
- Suppose AB needs some masking due to some small cell sizes.
- Suppose the degree of masking is represented by A'B'.

Scenario 1 has two TAZs: represented by A" and B'. Scenario 2 has just one TAZ represented by A'B'. The general tradeoff between Scenarios 1 and 2 is understood: Scenario 1 provides more unique (smaller) TAZs, but with more synthesized data, while Scenario 2 has fewer (larger) TAZs, but less perturbed data. But how sensitive the tradeoff is for "small" TAZs (TAZs near the minimum population size threshold) was unknown. The concept of this research was to assess the tradeoff between Scenario 1 and 2 by comparing the measured impact on data usability for both scenarios.

The DRB is responsible for reviewing the disclosure risks inherent in the CTPP. Disclosure risks are in part associated with the population size of the TAZs, which are the localities of interest used in transportation modeling and planning. The smallest TAZs in the ACS five-year sample would have a total population of about 600 people, and their ACS samples would have on average just 20–25 workers. This will not only provide unstable results, but will also lead to sparse TAZ to TAZ flows resulting in a substantial proportion of flows with only one or two sample cases, thereby causing concern about table linking.

The DRB rules that are placed on the tables are based on the Rule of 3. In effect, the DRB defines the riskiest cases where there are less than three sample cases in the following ways:

- Categories of (MOT) when crossed with another variable. This is because MOT is a common thread in the tables which leaves the tables susceptible to table linking.
- Cell means and aggregates. This is because a cell mean based on one case reveals the original value of the response. Also, a cell mean or aggregate based on only two cases reveals the original value of both cases if the value of one case is known, such as when a respondent to the ACS classified in a particular cell is looking at the reported value.
- Flow tables involving a table variable other than MOT. This is due to table linking risks explained above.

With many sparse tables, clearly an alternative to the traditional cell suppression was needed. The alternative was to perturb the ACS data before generating the CTPP tables. This can be thought of as adding noise to the data in order to add uncertainty to the identification of individuals. The goal of the perturbation approach is to retain as much of the ACS original data as possible, while targeting the riskiest data values, which are generally associated with small TAZs and TAZ flows. With the perturbed data, the DRB has agreed to drop the threshold rules.

The purpose of this section is to express the relationship between TAZ size and disclosure risk, which essentially is strongly related to the amount of perturbation applied. Every point estimate is associated with a measure of uncertainty, which has a sampling error component and a perturbation error component. Point estimates are produced to estimate totals or proportions in cell categories, and means or aggregate values within cells. For small TAZ estimates, the perturbation error and sampling error would both be very large in relation to the magnitude of the point estimate of interest. For larger areas, the perturbation error is smaller in relation to the sampling error.

To illustrate the impact of TAZ size on the number of records that violate the DRB disclosure rules, the proportion of records in TAZs, from the ACS 2005 to 2009 combined sample, involved in at

least one table with a DRB rule violation, was computed. In order to study the impact of TAZ size, TAZs with fewer than 50 ACS sample cases were then collapsed with nearby TAZs (the resulting collapsed TAZs are referred to here as CTAZ50). Further, TAZs with fewer than 300 ACS sample cases were collapsed with nearby TAZs (referred to as CTAZ300). As can be seen for the variables listed in Table 1–4, the proportion of records from housing units in TAZs below the DRB threshold can be substantially reduced through the collapsing of those TAZs into TAZs with at least 50 or 300 ACS sample cases. For example, for the poverty variables, around 30 percent of persons in the housing unit population are in TAZs that have a DRB rule violation in at least one of the tables involving poverty. This percentage dropped to 20 percent of persons in CTAZ50 below the DRB thresholds, and 10 percent for CTAZ300. Similar results are seen for the group quarters population (Table 1-5), though the TAZ percents are smaller. The TAZ percents are smaller most likely due to the group quarters sample being more clustered than the housing unit sample and possibly more concentrated in TAZs with larger populations.

Table 1-4. Comparison of Percent of Records from Housing Units in TAZs and Collapsed TAZs that Contain a DRB Rules Violation in at Least One Table: ACS 2005-2009

Variable	TAZ	CTAZ50	CTAZ300
Time Leaving Home	50	35	20
Travel Time	50	35	20
Age	40	25	10
Minority status	40	25	10
Poverty status	30	20	10
Industry	40	25	10

* NOTE: Values are rounded to the nearest five percent value.

Table 1-5. Comparison of Percent of Records from Group Quarters in TAZs and Collapsed TAZs that Contain a DRB Rules Violation in at Least One Table: ACS 2005-2009

Variable	TAZ	CTAZ50	CTAZ300
Time Leaving Home	25	25	15
Travel Time	30	25	15
Age	20	15	5
Minority status	20	10	5
Poverty status	5	5	5
Industry	15	10	5

* NOTE: Values are rounded to the nearest 5 percent value.

The state-level scatter plots in Figures 1-1 through 1-3 further illustrate the risk, for TAZ, CTAZ50, and CTAZ300, respectively. The plots use the variable “travel time” in determining the percent of records in TAZs that contain a DRB rule violation in at least one table involving the travel time to work variable. Each dot in the scatter plot represents a state. The x-axis is the ratio of the number of TAZs to the number of block groups by state (provided in the last column from Appendix B). Therefore, the far right on the plot means smaller TAZ sizes. The y-axis shows the percentage of ACS records in TAZs with threshold violations. Therefore, values higher on the plots have more DRB violations. This plot shows a relationship between TAZ size and disclosure risk—the smaller the TAZ size, the greater the risk and therefore, the more perturbation needed in the estimates.

As can be seen, in general, states with ratios of TAZ to block group counts less than 1.0 (i.e., the average TAZ size is larger than the average block group size within their state) have a lower percentage of their records subject to the data perturbation process. However, as seen in Figure 1-2, the percentage of records that is subject to the data perturbation process drops substantially when using CTAZ50 instead of TAZ, and further yet with CTAZ300 in Figure 1-3.

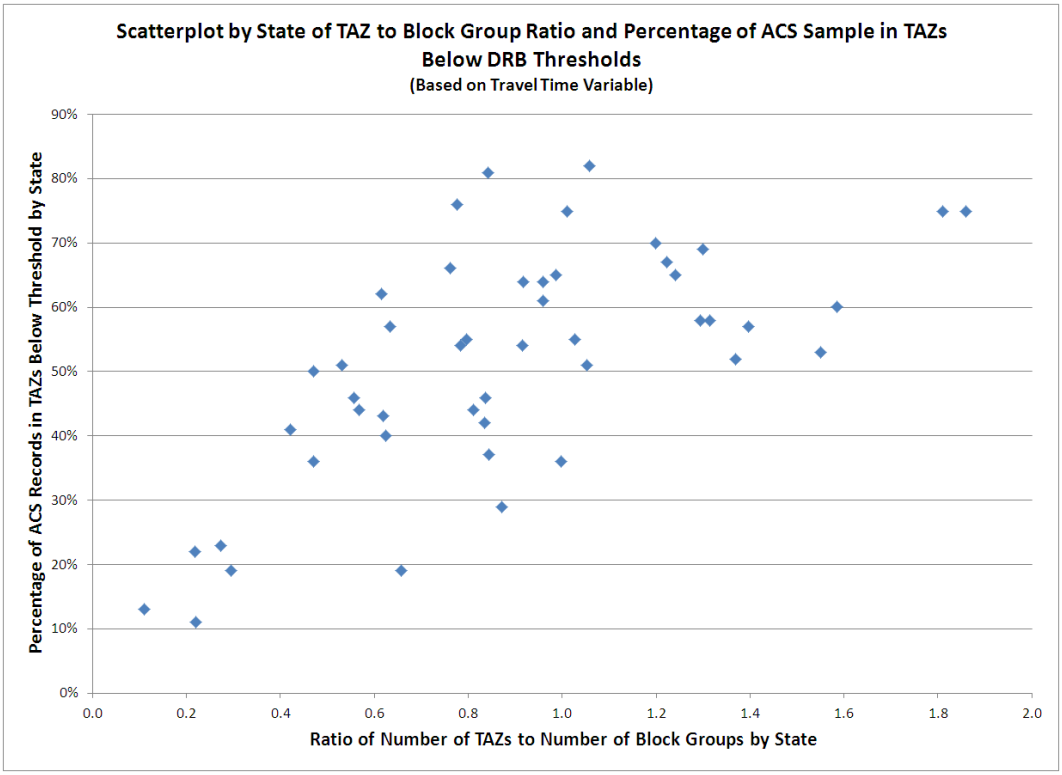


Figure 1-1. Scatter Plot by State of TAZ to Block Group Ratio and Percentage of ACS Sample in TAZs Below DRB Thresholds

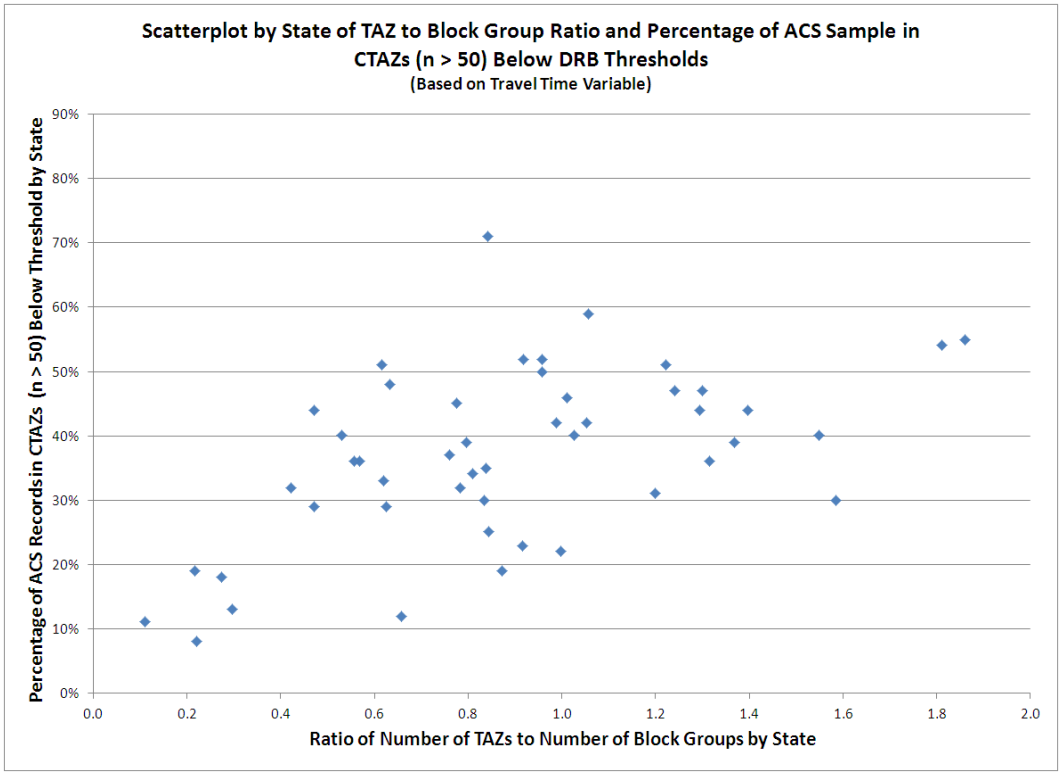


Figure 1-2. Scatter Plot by State of TAZ to Block Group Ratio and Percentage of ACS Sample in CTAZ50 Below DRB Thresholds

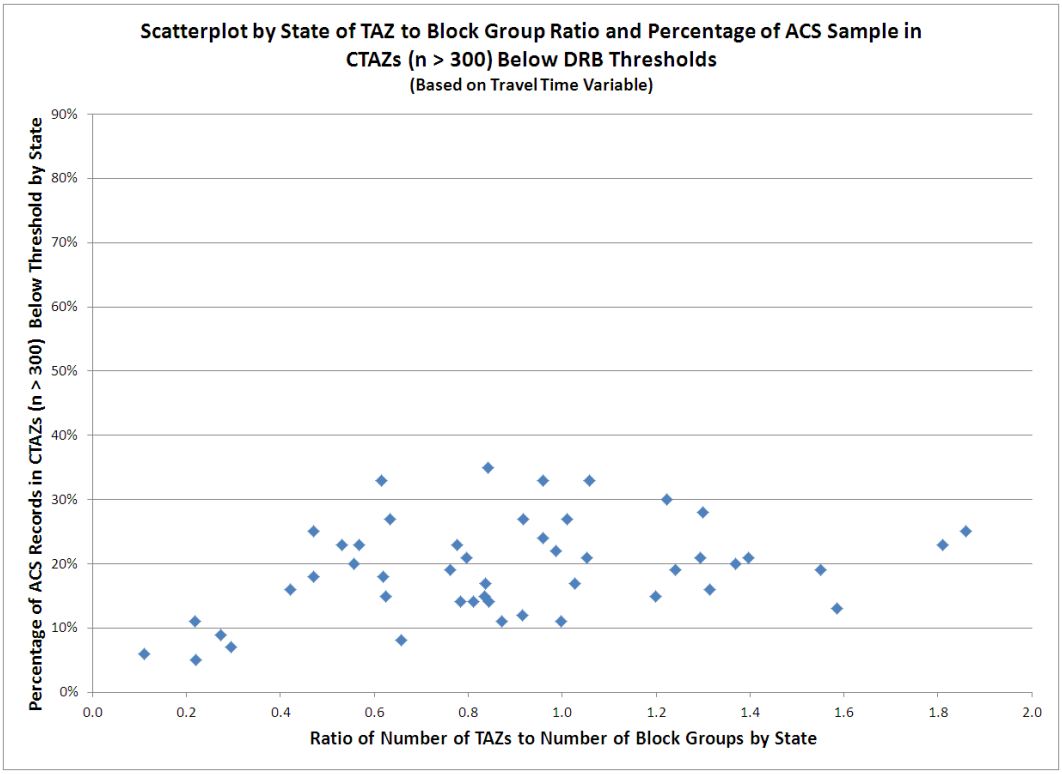


Figure 1-3. Scatter Plot by State of TAZ to Block Group Ratio and Percentage of ACS Sample in CTAZ300 Below DRB Thresholds

In summary, the above tables and figures have demonstrated the sensitivity of the proportion of the ACS sample that would be subject to the perturbation process as a function of TAZ size. The basic understanding is that if less of the ACS sample was subject to perturbation then the impact on the CTPP estimates would most likely be reduced. Another way of looking at this is that small TAZs will generally have high sampling error and will also need more perturbation to reduce disclosure risk. The greater the need for perturbation, then the greater the error in the estimates. So it is important to choose TAZ sizes carefully to balance the need for precision in modeling flows against error (perturbation and sampling error) in estimating flows.

Without any real mechanism to measure the tradeoff of lost utility from increasing TAZ sizes to 50 and 300 ACS sample cases, it has been shown that the percentage of ACS records subject to the perturbation process can be substantially reduced by forming larger TAZs (Tables 1-4 and 1-5) and that those states that on average form larger TAZs (as measured by having many fewer TAZs than block groups) have substantially fewer ACS sample records subject to the perturbation process (Figures 1-1 through 1-3). This was offered as guidance to the states as they approached the process of TAZ formation, especially those states collapsing the current TAZs as part of their CTPP data use process. States that form larger TAZs would be subject to less sampling and perturbation error, thereby reducing loss in data utility.

Identify Outlier Trip Scenarios

To alleviate concerns over outlier trip scenarios, a system was developed to detect outlier trip scenarios based on the flow, MOT, and travel time. Discussions with the DRB chair ensured that these

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

outlier detection procedures are similar to those used at the Census Bureau and acceptable. State-level travel time distributions were processed by MOT, to detect possible outliers.

Classify Risk Levels for Each Variable

When it comes to identifiability and matchability to the ACS PUMS, identifiable characteristics are subject to replacement under the perturbation approach brought forward by this research. Conceptually, variables that are highly identifiable but have low usability are prime candidates for perturbation. In order to protect the ACS PUMS data from identity disclosure, perturbing a subset of variables may only be necessary to break the link in the table-linking effort, assuming that the list of perturbed variables is kept secret. This helps to reduce the risk of attribute disclosure if a data snooper is in pursuit of coworker attributes. To this end, the research team worked with the DRB to determine the variables considered highly identifiable, and with transportation specialists to determine the variables considered highly usable for their means.

The high, medium, and low classifications for the identifiability of each variable are provided in Table 1-6. The identifiability levels shown in the table are illustrative in order to not provide too much information. These classifications do not necessarily determine a list of variables to be perturbed and a list that will not be touched. Such classifications help to understand DRB concerns. Transportation planning specialists provided the usability ratings in Table 1-6 (5 = most useful) as they relate to their use by transportation planners. Although the transportation group is most concerned with residence, workplace, and MOT, there may be cases for which this data must be modified to protect confidentiality.

Table 1-6. CTPP Variables and Their Usability and Identifiability Levels

CTPP Variable	Usability Rating	Illustrative Identifiability Level
Age	3	high
Class of worker	2	mid
Earnings	5	high
Industry	2	high
Length of US residence	1	high
Minority	3	mid
Occupation	1	high
Sex	1	high
Time leaving home	4	mid
Travel time	5	low
Age of youngest child	2	mid
HH income	5	high
Poverty	4	mid
Vehicles available	5	low
# of workers in HH	4	low
Time arriving (Part 2 only)	3	low

Identify and Target High Risk Data Values

The DRB was aware of the risk elements inherent in the ACS data, and the threshold rules were a reflection of this realization. Therefore, the primary definition of initial disclosure risk was based on the DRB disclosure rules (before applying the perturbation approach). The initial risk assessment involved processing the CTPP tabulations to identify cell violations using the DRB disclosure rules. In production,

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

data values associated with such violations will be flagged as high risk. Data values flagged as high risk will be targeted for replacement by the data perturbation approach.

Identify Risk Reducing Elements

Several sources of data protection have been identified in the CTPP based on the ACS sample data. For a given microdata record formed through table linking, there is a chance of the data being protected due to the following:

- Sampling reduces the risk of disclosure as compared with a census of individuals. As shown in Table 1-3, the sampling rate for the five-year ACS is about 7.4 percent, after nonresponse is taken into account.
- Swapping is used to reduce the risk of disclosure in ACS data products. The swapping rate and list of swapped variables is withheld by the DRB.
- Moving or changing job locations over a five-year period is non-negligible. For example, about 46 percent of the population age 5+ moved their residential address between 1995 and 2000 according to the 2000 Census, and the percentage that changed workplaces is thought to be about the same.
- Imputation due to item nonresponse is inherent in the data. The national imputation rate varies from near 0 percent for sex to about 13 percent for income, earnings, and poverty.
- In general, there is an underlying uncertainty or divergence of variables, such as response errors, that reduces the factual (re-identifiable) nature of the variable over time.

1.2 TRANSPORTATION PLANNING CONSIDERATIONS

Historically, CTPP has been used in the transportation community by travel demand forecasters as a comparative observed dataset for model validation (in some cases certain tables/variables have been used in model estimation and model calibration, but this is less frequent). CTPP data are also used by transportation planners as a base to create separate, quick-response analysis tools independent of the traditional travel demand forecasting (“four-step”) process. The Aggregate Rail Ridership Forecasting (ARRF) model developed by the Federal Transit Administration (FTA) is a good example of such a tool. Finally, planners use CTPP for historical travel trend analysis, special studies, and to assist with local travel surveys. All of these uses will continue with the ACS-based CTPP data products, and transportation data users in turn expect data products that will permit continuation of existing uses. The key to these research efforts is balancing transportation user needs with the requirements for disclosure avoidance.

As noted above, the general desire of the transportation planning community is for smaller units of analysis (TAZs) for CTPP, although in practical terms, the average TAZ size varies greatly from state to state, as shown in Appendix B. The desire for smaller TAZs is becoming more of a need in many areas where agencies are moving to a more disaggregate level of travel modeling and analysis, but those areas still represent a small fraction of Metropolitan Planning Organizations (MPOs) nationally (although a larger fraction of population, since nearly all of these tools are being applied in complex, major urban environments). More discussion of the approaches to TAZ size and the surrounding issues can be found in Section 1.1.4. Regardless of the geographic unit of analysis, the key CTPP variables needed at the microdata level for transportation planners are place of residence, place of work, and MOT. A quick

assessment of the usability of other CTPP variables in travel demand forecasting may be found in Table 1-6.

The use of microdata (PUMS) as a seed for population synthesizers as part of travel demand models is a special case that is growing into a larger transportation user need. How perturbed ACS data works in such a process was addressed as part of the research and testing, particularly for those variables where synthesis is employed at the microdata level. The effects of a "dual" synthesis, that is, running the population synthesizer on perturbed microdata to create a base year population and using that as the travel demand model input, was considered for the research.

Transportation planners need to understand the effect that new disclosure-avoidance techniques will have on their analytical tools and any limitations on the use of the resulting data, particularly when cross-tabulating (for example, crossing a raw variable with disclosure-proofed ones that have been perturbed through differing techniques). This is independent of potential error propagation from the use of perturbed data through the travel demand model chain, which will not be explicitly considered in the research. There may be a need to re-explain the use and validity of perturbed data, which, although largely accepted in the transportation planning community, will still face pockets of suspicion. All this education also makes transportation planners' jobs easier when explaining their analytical process to the local elected officials to whom they are accountable. Transportation planners devoted significant resources to understanding (and in some cases elucidating) the limitations of the 2000 CTPP for certain types of analysis, and to the extent that some of the documented and tested issues associated with perturbed ACS data for CTPP can be alerted to planners during this research, data user needs will be further satisfied.

As discussed in the January 2010 kickoff meeting notes from the Executive Session, given the range of procedures used by MPOs in model development and the range of ways that CTPP data are used in the modeling process, the Panel was unsure if the use of one model would provide any useful performance measure. Therefore the Panel provided some key recommendations during the kickoff meeting as follows:

ITEM V - Panel's Recommendations to the Research Agency

> The Panel would like some feedback from the research team on how travel model validation would be done. The Panel would like the research team to explain how the procedures applied to disclosure protection would affect model validation. The Panel hoped that this could be accomplished comparing the model-based home-based work (HBW) outputs (for models developed around the country during the same time period) with ACS raw and ACS post-disclosure based outputs. This approach would allow multiple runs to figure out the data variability, especially with respect to mode choice. The Panel hoped that the research team does not get into investigating the ripple effect of different ACS-based datasets on a model from trip generation all the way to assignment. The approach described by the Panel might not require proprietary software to be installed at the Census Bureau, but instead would comprise collecting MPO-based HBW model outputs to compare against ACS-raw, and ACS-post disclosure datasets.
>
> The research team should also pay attention to the effect on the use of the data as a source of controls for population and household-based synthesizer models.

The research team accepted the Panel's recommended approach and agreed that comparing the ACS CTPP data against home-based work (HBW) model outputs is the best approach, since HBW remains the dominant trip purpose in all metropolitan areas and is the only purpose directly comparable against the questions posed by the ACS (which does not explicitly cover non-work travel). The approach

also ameliorates the need to install, run, and support proprietary travel modeling software at Census Bureau as part of the research.

The confirmed development phase and potential validation phase test sites (discussed in the next section) attempted to use models developed and validated during the 2005–2009 period to correspond directly with the three-year and five-year ACS data used for testing. Because of different ACS-based data sets, error propagation through the model chain was not directly addressed. However, the relationship between the test results of individual model components outputs (generation, distribution, mode choice) and potential interactive effects of different ACS-based data (for example, the effect of perturbed time of departure/arrival for comparison against time-of-day models for those agencies that employ them) may be considered.

The team collected HBW model outputs from the test sites in order to conduct comparison tests between the model output and ACS raw and disclosure-proofed ACS data. The team paid close attention to the effect on the use of ACS data as a source for controls of population and household-based synthesizer models. The team specifically identified Atlanta (development and validation phase) as a test site that creates synthetic populations for use as model inputs to examine this issue and compared the ACS and perturbed CTPP data against the model-synthesized population across different levels of geography.

1.2.1 Test Sites

The evaluation for the project was partitioned into two phases. In the development phase, the selected approaches were developed and evaluated for four test sites. In the validation phase, the most credible data perturbation approach was tested further for two test sites. The selection of the test sites took into consideration the planning organization's travel modeling experiences and sought out sites from across the nation to the extent possible, rather than focus on one part of the country. The test sites are shown in Table 1-7.

Table 1-7. Model and ACS Comparison Test Sites

Phase / Model Type	Agency (Region)	Year of Model Output (Base or Forecast)	Most Recent Validation (Base Year)
Development phase using 3 year (2006-2008) ACS			
Tour/Activity-Based Model / Population Synthesizer	Atlanta Regional Commission (Atlanta, GA)	2007 (Forecast)	2008 (2000)*
Large MPO Trip-Based Model	East-West Gateway Council of Governments (St. Louis, MO)	2007 (Forecast)	2002 (2000)
Medium/Small MPO Trip-Based Model	Madison	2005 (Forecast)**	2006 (2000)
Statewide Model	Iowa Department of Transportation (State of Iowa)	2005 (Base)**	2009 (2005)
Validation phase using 5-year (2005-2009) ACS			
Tour/Activity-Based Model / Population Synthesizer	Atlanta Regional Commission (Atlanta, GA)	2007 (Forecast)	2008 (2000)*
Medium/Small MPO Trip-Based Model	Olympia		

* Many of the Atlanta submodels have been refined and recalibrated during subsequent years.

** Modelers at Madison and Iowa have confirmed that growth from 2005 to the period of the ACS is negligible and thus the 2005 data is acceptable for our tests.

1.2.2 Usability Ratings

An arm-chair assessment of the usability of CTPP variables in travel demand models was provided in Table 1-6, where CTPP variables were rated on a scale from one to five (most important for transportation purposes). Note that although earnings and household (HH) income potentially could serve as proxies for one another, the ratings reflect a strong preference for and historical use by planners of HH income (rather than earnings).

1.2.3 Data Consistency Issues

The topic of ACS five-year production tabulations relates to the issue of consistency of weighted 5-year production block group estimates and TAZ estimates coming out of the CTPP tabulations. (In many MPOs, block group geographic definitions are directly used in defining TAZ geographic boundaries.)

The Transportation Research Board Panel provided guidance at the kickoff meeting that the issue of potential inconsistency of totals and demographic distributions between ACS production block group level tables and CTPP TAZ level tabulations was a secondary consideration. They cited that "rounding" had previously created differences that users understood and were able to address. In general, the Census Bureau staff concurred with this position.

That said, the merits of calibrating at higher levels of geography were considered. TAZ definitions can be within a block group, cross block group, and tract boundaries, but must be nested within a county. Given feedback from the Panel, the research team planned to calibrate the perturbed data at the Public Use Microdata Area (PUMA) level, which are areas with at least 100,000 in population, by key variables. Based on a request by transportation planners, the research team also calibrated the weighted totals to sub-PUMA levels so that the total workers would add to the ACS total for combined TAZs that were formed to have about 4,000 workers.

1.3 ACS OPERATIONS CONSIDERATIONS

The successful implementation of a process to produce perturbed CTPP tables required comprehensive discussions and a close working relationship with the ACS operational staff, to ensure the viability of the CTPP process to work within the fairly constrained annual ACS production and special tabulation processes. Several meetings were held with key personnel related to the CTPP special tabulations, covering such key issues as data files, data file access, timeline, and identifying other tabulations other than CTPP.

1.3.1 Input Datasets

The first CTPP tabulations will be produced from the ACS five-year 2006–2010 data file. Under ideal conditions, research efforts for NCHRP Tasks 3 and 4 (February–July 2010) would have been conducted with a fully processed ACS five-year production file. However, such a file did not exist at the time that research efforts began. After discussions with the Census Bureau, the research team concluded that using the three-year production file for 2006–2008 was the best alternative. One implication of this decision was that the research file was "sparser" than a true five-year file. To partially address this limitation, the team considered that a TAZ with a population of 600 persons from a 5-year file (600 * 7.4% = 45 ACS sample persons) would have about as many ACS sample persons as a 1,000 person TAZ

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

from the three-year file (1000 * 4.5% = 45 ACS sample persons). While there could still be meaningful differences between such sized TAZs, the patterns of sparseness and potential disclosure issues should be somewhat similar.

The first 5-year production data file was made available for use in time for the validation phase (Task 6). The file included the first five years of ACS at full production levels: 2005 through 2009. Public release of 5-year tabulations did not occur until the end of 2010/early 2011. The file reflected complete and final review of weights and content.

The file reflected the 2009 geography which was still Census 2000 based, including the TAZ definitions. The first CTPP tabulations from the ACS 2006–2010 5-year file are expected to reflect the 2010 Census geography and updated TAZ definitions.

The research team and the DRB agreed that the imputation and swapping flags could be used to identify certain situations as not a disclosure risk that otherwise would have required some action to reduce the risk of disclosure. These flags were placed on this file. The degree of swapping could expand under the revised process for the ACS 2006–2010 5-year data files.

1.3.2 Operations Timeline/Resources

The plans call for the CTPP package tabulation to be generated once every five years with the first coming from the ACS 2006–2010 five-year data. Scaling up for a large once-every-five years special process has its own set of potential resource challenges. With that in mind, the research team asked that the Census Bureau members of the Special Tabulations Group review the general class of methods under consideration, and provide feedback to the research team on the pros and cons as they relate to the Census Bureau needing ultimately to take ownership of the developed method and apply it to produce, review, and approve the 2006–2010 CTPP tabulation package. Initial feedback was that, though some methods may be more “transparent” than others, no particular method is a “show stopper” from a Census Bureau production perspective when it comes to overly onerous resource requirements associated with validating/verifying the perturbed CTPP tables. The research team continued to discuss these issues with Census Bureau staff during the initial stages of research.

1.3.3 Workplace Allocation

Nationally, extended workplace allocation is necessary for about 23 percent of records missing workplace geography below the place level. This is a procedure conducted by the Census Bureau, although the timing is such that the allocations were not provided in the three-year and five-year research files. The implication of this is that Part 2 (Workplace) and Part 3 (Worker Flow) tables included on average non-missing block- and TAZ-level values of workplace allocation for only 77 percent of all the worker records. As with a three-year research file (instead of a five-year file), this caused these tables to be even “sparser” than they would be otherwise. In addition, the ACS 2005–2009 five-year data file used for implementation in Task 6 was subject to the same limitation. However, the planned five-year CTPP, based on 2006–2010 data, is expected to have the extended allocation process applied post-hoc to all records. This process is expected to code 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. The research team continued to correspond with ACS operations staff so that appropriate methods could be used to account for the workplace allocation status.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

To address the layer of added complexity due to missing workplace TAZs in the research-based files, the team considered the effect on the research, as well as the handling of such cases in the critical production run. Given Panel feedback on proposed options discussed in Westat (2010), the team decided that for this research, cases would be assigned with missing workplace TAZ to the remainder of the state for the development phase (remainder of the county for the validation phase) associated with the test site planning area.

1.4 INITIAL CRITICAL ASSESSMENT OF PROMISING DISCLOSURE AVOIDANCE TECHNIQUES

A review of disclosure prevention techniques was conducted to identify the most credible to the transportation planning process while satisfying the DRB, meeting the needs of transportation analysts, and designing a system that will allow the Census Bureau to implement it with available resources and on the promised schedule. These goals were in serious tension with each other. Given this challenge, a variety of techniques for initial examination were undertaken. At the end of the assessment, recommendations were made to reduce the number of approaches.

Section 1.4.1 provides a discussion of the set of criteria for the assessment. Section 1.4.2 continues with an overview of the perturbation approaches that were considered, presenting the pros and cons of each approach as they relate to the set of criteria.

1.4.1 Set of Criteria

The set of disclosure avoidance techniques were measured against the following criteria:

- **Disclosure risk and rules.** The approach will adequately reduce disclosure risk, and satisfy the disclosure rules that the DRB has developed.
- **Data utility**. The approach will satisfy the needs of transportation planner analysts by minimizing the effect of statistical disclosure control (SDC) on data utility.
- **Operations**. It will be necessary to implement the approach on a critical production path so that the CTPP tables can be produced in a timely manner. Sensitivity to Census operations staff processing and checking over the results needs to be a consideration.
- **Applicability and flexibility**. The approach adopted should be flexible;, the CTPP offers a variety of tables that involve different types of variables, such as unordered categorical, ordered categorical and continuous numeric, and the approach should handle all such variables The approach adopted should also be applicable to the generation of tables that require cell means, aggregates, and medians. Also, flexibility could be measured when it comes to being able to add new variables to the tables.
- **Variance estimation.** The approaches should facilitate variance estimation, not only retaining the sampling error variation, but where possible incorporating the error added due to the data perturbation approach.
- **Data consistency.** The approaches should provide consistent data within the set of CTPP tables. For example, Part 3 tables should align with Part 1 and Part 2 tables; the marginal totals for a variable must match to the marginal totals for the same variable in another table.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Consideration was given to the interplay between the resources consumed by the different approaches and the benefits obtained from each. With the development of two to four approaches, the efficiency of developing one approach may overshadow the development of other approaches. Several discussions among the Senior Statistical Advisory Group brought forth critical assessments of each approach. The approaches were presented to the group, and discussion of their advantages and disadvantages occurred. The next section provides a brief description of each of the initial approaches considered and provides a summary of the discussion.

1.4.2 Assessment of Initial Approaches

In general, there are two major types of SDC approaches. The first type is deterministic, such that there is no random error introduced into the process. The second type (data perturbation approaches) includes random error as part of the process. There are two major types of data perturbation approaches, one in which table-level modifications to already-generated tables occur, and one in which unit-level modifications occur and then the tables are processed from the modified dataset. Table 1-8 provides a list of SDC treatments that were initially considered.

Table 1-8. List of Statistical Disclosure Control Treatments Initially Planned

Type of approach	Level of application	Approach
Deterministic	Variables	Coarsening
	TAZ	TAZ redefinition
Perturbed	Table modifications	Small area estimation
		OnTheMap approach
		Bayesian/IPF
	Microdata modifications	Semi-parametric
		Parametric modeling
		Data swapping
		Super-sampling

1.4.2.1 Deterministic Approaches

Procedures for protecting tables of counts without using perturbation methods are described in Willenborg and de Waal (1996, 2001). They describe ideas for redesign of tables (collapsing) to avoid sensitive cells, suppression of cells, rounding of cells to fixed points, such as multiples of 5, and reporting of feasible intervals. Little attention, besides providing illustration, was given in 1996 to concerns that arise when tables are linked, or when multiple tables sharing some of the same margins are published. More attention was paid to a linked table example in 2001, but methods were not studied extensively. Transforming a variable by coarsening, bounding, or rounding (practices used in the ACS sample data), removes some of the information content in the values, but makes identifying a unique individual based on the data values less likely. Topcoding variables such as income or commute time would be examples of commonly used transformations.

For the CTPP, two deterministic approaches were initially considered: re-defining TAZs to be larger or having a larger minimum threshold (described earlier in Section 1.1.4), and collapsing categories of CTPP variables (coarsening). The implementation of these deterministic approaches would reduce disclosure risk and help to retain data usability, because they allow more sample records to contribute to the subgroups.

One of the first steps is to process tabulations that will detect variables that are the causes of disclosure rule violations. If certain variables are the usual culprits and the variables are relatively non-important to travel demand analyses, then collapsing categories of such variables is an option. From the initial investigations, the research team recommended that industry be collapsed to seven categories for the flow tables because of two categories in the 14 category version with very small sample size counts.

As determined by the Census Bureau, deterministic approaches by themselves will not be sufficient for the CTPP, and data perturbation approaches are still necessary.

1.4.2.2 Perturbation Approaches

Data perturbation basically refers to a disclosure control strategy that generates perturbed data from one or more statistical models, and uses the generated data for release in lieu of raw data. Perturbation approaches introduce a random component to produce perturbed data. Several meetings of the Senior Statistical Advisory Group at Westat brought forth critical assessments of each perturbation approach under initial consideration. One by one, the approaches were presented to the group, and discussion of their advantages and disadvantages occurred. The discussion of each of the approaches that follow includes an abbreviated summary of the critical assessment (refer to Westat 2010 for more discussion).

Tabular Approaches

Traditionally for the CTPP, the tabular approach of cell suppression has been used to reduce the risk of disclosure. Issues involved in theory and practice of cell suppression for tabular data are presented in Giessing (2001), Domingo-Ferrer and Franconi (2006), and Domingo-Ferrer and Saygin (2008), and the first two sections of Domingo-Ferrer and Torra (2004). If cell suppression was used under current DRB rules with the ACS sample sizes, the practice would lead to the deterioration of the usability of the produced tables.

Outside of cell suppression, the literature related for tabular approaches for tables of counts has been limited to adding noise to the counts in tables, such as in Willenborg and de Waal (2001). Doing so will create inconsistencies across multiple tables. See Fischetti and Salazar-González (1998) for more discussion of controlled rounding of tables of counts for disclosure protection.

Three additional methods for perturbing an existing table have been recently reported in the literature and were reviewed by the research team. They include small area estimation, a methodology used in OnTheMap, and a version of Bayesian iterative proportional fitting. Although of intellectual interest, they are not being pursued in this application for reasons given below. All would require substantial investment of resources beyond what is feasible in the given time frame and have uncertain outcomes given the serious limitations presented.

Small Area Estimation. Small area estimation is a statistical modeling approach that produces model-dependent estimates, called “indirect” estimates, to distinguish them from standard survey or “direct’ estimates that are derived directly from responses of sampled individuals who live in an area included in the assessment. Rao (2003) and Jiang and Lahiri (2006) provide comprehensive current overviews and comparisons of models and methods for small area estimation. The indirect estimates are produced using small area estimation techniques that rely on the direct estimates from the current area, estimates from other geographic areas included in the planning area, and other variables and other

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

geographical characteristics related to the variable of concern. Area-level small area estimation models are infeasible for the CTPP for the following reasons:

1. With sparse TAZ-to-TAZ flows, there very likely are not enough flows with adequate sample size to estimate model parameters, which would be used for the prediction of 90 to 95 percent of the flows with insufficient sample size.

2. Each cell of each CTPP table would need to be estimated, which is an infeasible undertaking—given 166,000 TAZs, with many more flows and with effectively over 200 tables.

OnTheMap Approach. This approach would use methods that are similar to the area-level modeling approach that generates perturbed block-to-block flows for LEHD OnTheMap, where Bayesian techniques are used to synthesize workers' place of residence, conditional on counts of workers by place of work, industry, age, and earnings categories that can be disclosed.

Machanavajjhala et al. (2008) and Abowd et al.(2009) provide a description of the approach. Given the success of the OnTheMap system, the merits of this approach demanded initial consideration for the CTPP, although, while some investigation was done, the decision was to not develop this approach for the following reasons:

1. There is no technical documentation to truly replicate this approach and apply it to CTPP variables.

2. More time and resources would be needed to develop this approach properly and investigate questions (including consistency between marginal totals) than is allowed under the current timeline, given the need to develop other approaches to evaluate.

3. More operational resources would be needed to verify and check modifications to tables than what is needed when tabulating perturbed microdata.

4. It is unclear that the OnTheMap approach would provide any added benefits beyond the microdata approaches proposed in this document.

Bayesian/IPF. In a study by Cambridge Systematics, Inc. et al. (2009), Iterative Proportional Fitting (IPF), also known as raking, was explored to generate synthetic journey-to-work or origin-destination (OD) tract-to-tract traffic flows from five-year ACS (or long-form Census) data based on fixed marginal counts at "super-tract" or tract level. Cambridge Systematics also modified the IPF-based data synthesis by combining it with a preliminary step that used a Bayesian approach to create synthetic trip origin counts. The Bayesian/IPF is unlike the IPF only in that before fitting the model one needs to make a table of prior counts. This approach allows more variation in interior cells of the table. Our intention was to not pursue the Bayesian/IPF further, because of the following reasons:

1. The potential run time on applying this approach to make the modifications table-by-table is likely quite extensive, given 166,000 TAZs, with many more flows and with effectively over 200 tables, and also considering layers of geography.

2. With likely 90 percent of the flows having singletons or doubletons, it is not clear that this methodology will have successful results, given the sparsity of the data.

3. The assumption of applying models fit to high levels of geography to the individual TAZs inside those geographies would need more study.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

4. The modeling potential is not used to the maximum, in that a hierarchical model makes more sense rather than separate models. Likewise, the modeling potential is not used to the maximum, in that separate low dimensional tables fit separately would appear to lose some prediction power that could be gained from other auxiliary data.

5. More operational resources are needed to verify and check modifications to tables than what is needed when tabulating perturbed microdata.

The Generalized Shuttle Algorithm, discussed in Cambridge Systematics, Inc. et al. (2009), was briefly considered early on among the table modification approaches. It was dismissed because of the reported computational intensity and the increasingly widespread acceptance of perturbing microdata by the senior statistical advisory group and the Census ACS operations group.

Microdata Approaches

Besides aggregating levels of variables to increase counts above two or suppressing entries so that numbers based on small counts are not reported and cannot be derived, an option for preventing certain disclosure is to perturb the microdata that come as an input to the table before the table is created (Duncan, Fienberg et al., 2001). Doing so creates the question of whether the numbers based on small counts in the table correspond to the true composition of the sample or result from the artificial random process. It is then possible to deny that an apparent linkage to an actual individual is real.

The major advantage of this approach is that the end product is a single perturbed dataset underlying all CTPP tables; that is, the tables derived from the dataset have consistent margins while simultaneously providing disclosure protection.

Microdata are perturbed by adding noise, which can be done in a variety of ways. For example, one approach is to compute model predictions for continuous dependent variables from conditional models and add random noise from draws from the normal distribution with mean 0 and variance . For unordered categorical dependent variables, and draws from the predicted distribution resulting from conditional generalized multinomial mixed effects, models can produce the perturbed data. The methods need to be implemented in a manner appropriate for the type of variable being perturbed. Sometimes two or more methods of perturbation could be applied in conjunction.

Semi-Parametric. This methodology is influenced by Judkinset al. (2007) and the Gibbs sampler, to a certain extent. A similar approach is discussed in Bocci and Beaumont (2009). In this approach, the resulting replacement values are model-assisted, rather than model-based. Because of the less important role of the models in this approach, it is less critical to get the structure of these models exactly correct. The sequential nature of the process has the benefit of preserving multivariate associations. The model selection and estimation step would use linear regressions for variables with a small number of categories. This is done primarily to reduce processing time, and it is not as critical as long as the ordering of the predicted values is correct. Hot deck cells are formed from the predicted values and the original value from a randomly selected donor is used as the replacement value. Some consideration was given to the fact that the semi-parametric approach had to be developed and would take some resources to do so.

Parametric Modeling. The parametric modeling approach is a data perturbation approach that involves difficult work to create strong parametric models. None of the other microdata approaches suggested require nearly as much modeling work. In this approach, several variations were discussed that would model several CTPP variables. These approaches considered the Gibbs sampler (Gelfand and

Smith 1990) to link the conditional models for each variable, as well as fully Bayesian models involving Markov Chain Monte Carlo (MCMC) samples. However, a simplified approach was determined to be the most promising parametric approach. It maintains the associations between variables and may be expandable if new tables, added after the production of the perturbed dataset, involve at least one perturbed variable.

The use of Bayesian parametric algorithms for data imputation has grown in recent years (Raghunathan et al., 2002). These methods, such as discussed above, draw imputed values from a posterior predictive distribution specified by a regression model, usually with a flat or non-informative prior distribution for the regression parameters. However, they are often heavily reliant on normality assumptions and are not designed to cope well with unusually shaped distributions, such as heaping of reported income at round thousands. The simplified parametric modeling approach contained these same drawbacks, but was far less computationally intensive.

Data Swapping. Swapping values entails choosing a case for potentially changing the value, selecting a donor with which to change values, and switching values between the two units in the dataset. This could be formulated as choosing two cases to switch, but often it is applied to cases for which there is a concern about identifiability, hence the formulation as choosing a donor for a select case. Fienberg and McIntyre (2004) give an overview of microdata swapping in the context of tables for which there is a desire to preserve some marginal counts in the tables. In general, the data utility of the approach is sometimes suspect. With swapping, variables highly associated with the swapped variable can be linked to change if the swapped variable changes.

Supersampling. This approach selects a new ACS sample through a supersample frame. Suppose five copies of the ACS dataset are appended and a new ACS sample of size “n” is selected. Cell counts of one would result in counts of 0, 1, 2, 3, 4, or 5. A problem with supersampling is that it can actually create a sample unique when the cell was already acceptable to begin with. The super-sampling approach alone, as proposed, would not alleviate concerns about disclosure risk since it could result in single records given combinations of variables. A brief discussion with the Census Bureau DRB also indicated concerns.

1.4.3 Credible Approaches Selected for the Development Phase

While having good properties in reducing disclosure risk and facilitating variance estimation, the resulting data utility related to the table-level modification approaches is suspect due to the sparseness of the flow data. Also, the data consistency between tables would need to be addressed. Operationally, given the number of TAZs, flows, tables, and layers of geography, with millions of tables to be generated, the amount of maintenance and checking is understatedly non-trivial. Therefore, none of the table modification approaches were recommended for further development.

Due to the lack of disclosure protection properties, the supersampling approach was also excluded from further development. Therefore, the following three microdata perturbation approaches were selected for the development phase:

1. Parametric model-based approach;
2. Semi-parametric model-assisted approach; and
3. Data swapping (with later adaptation referred to as constrained hot deck).

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

The three perturbation approaches each retain internal consistency of marginals within the set of "threshold" Set B tables (described further below) produced in the CTPP. While the parametric modeling approach is very demanding on resources to develop, and somewhat risky when it comes to convergence issues, it has great potential in facilitating variance estimation, data usability, and in reducing disclosure risk. The semi-parametric has similar advantages as the parametric modeling approach, and includes the strength of drawing from empirical distributions from either within a small area (e.g., collapsed TAZ). The semi-parametric approach is less dependent on models and is not susceptible to convergence issues. Data swapping is appealing in that it was readily accepted by the Census Bureau, and low effort was needed to apply it to the ACS five-year data.

In addition to perturbation approaches, the following discusses approaches that were selected for further development.

Weight Calibration

After any of the perturbation approaches relating to microdata modifications are applied, the ACS weights will be calibrated to published ACS totals through raking. Raking is sometimes called iterative poststratification or iterative proportionate fitting, and was introduced by Deming and Stephan (1940), and more discussion can be found in Oh and Scheuren (1987). Raking forces the modified microdata file to have estimates for selected marginal dimensions equal to or calibrated to those from the unadjusted ACS.

Two Sets of CTPP Tables

This general approach uses perturbed data where tables would have been subjected to DRB disclosure rules, and uses the ACS five-year data for tables where there are no disclosure thresholds. It is designed to retain as much observed ACS data as possible. The end result can be thought of as dividing the current CTPP tables into two sets:

- CTPP Set A (ACS five-year data tabs) based on real data and ACS weights, where the DRB agrees to release data fully, without suppression; and
- CTPP Set B (perturbed part) based on perturbed (postdisclosure proofing) data and CTPP adjusted weights, where DRB has concerns.

The benefit of this approach is that data are not touched unless needed, perhaps providing better data utility to the users. There would be different marginal totals for the same variable; that is, the marginals will not be consistent between the Set A and Set B CTPP tables for the same variable. Operationally, a table generator would need to call the correct version of the variables, and each table would need to be checked carefully before release. The DRB has reviewed this approach and has accepted it, although they have specified that the usual rounding rules will apply to the Set A tables, as they do for the other special tabulations from ACS data. The rounding rules are applied to interior cells while fixing the marginals. Since the marginals are the summation of the interior cells, this in effect will cause the marginals to differ for the same variable across tables. Users will be alerted through the table title or a footnote that the Set B tables were generated from perturbed data. Further clarification is summarized as follows:

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

1. There will be two underlying microdata files as input into the Census Bureau CTPP table generator program. The first microdata file will contain all original data, and the second file will contain perturbed microdata for the variables in the Set B tables.

2. The perturbed microdata file resulting from the initial risk analysis for the Set B tables on TAZ level Part 1, 2, and 3 tables will be used for all localities for the Set B tables. The tables will be generated from the same perturbed microdata for all geographies including TAZs, Block Groups, Tracts, Transportation Analysis Districts (TADs), Places, Counties, States and PUMAs.

3. The perturbed microdata file will be used for TAZs where there are no violations as determined by the initial risk analysis. Even if the values of variables are unchanged, the raked weights may differ from the ACS weights, and therefore the CTPP estimates will be different from the ACS estimates.

4. The list of tables (Appendix A) contains several collapsed tables: 12201C, 12201C2 and 12201C3, for example. The collapsed versions of tables will be generated from the same perturbed microdata.

5. In Appendix C, there is reference to “Large Geography Only” for some of the tables. Large geography means county, PUMA, and state.

6. The disclosure proofing process in the research used the most detailed table in the table series (e.g., 12201 was used in the risk analysis for the series 12201, 12201C, 12201C2, and 12201C3).

7. Having more detailed tables (e.g., all based on MOT(18)) would increase the amount of perturbation in the microdata. It would also impact the DRB decisions and the perturbation rates assigned they would assign. It would necessitate a reassessment of the impact on data utility.

8. On data consistency, suppose you have residence TAZ All flows for Table 33204 in Part 3 involving residence TAZ, if added together, will produce the same results as Table 13204 for residence TAZ from Part 1. All tables will be consistent with one another within the set of tables referred to as Set B since they are all generated from the same perturbed microdata file.

Using the set of CTPP research tables provided by AASHTO and given in Appendix A, the Set B tables (ones subject to DRB rules) were identified and the list is provided in Appendix C.

Variance Estimation

A variance estimation process on the resulting single perturbed dataset was constructed to capture the sampling error in the ACS sample as well as the impact of the perturbation approach. Data utility checks compared the variances before and after the perturbation approach was applied. The research team discussed the variance estimation approach, described further in this report, with the ACS operations staff.

2. Development Phase

During the development phase, while working on site at the Census Bureau, the authors developed and evaluated the credible data perturbation techniques that were identified through the initial critical assessment of plausible approaches. An experimental design was developed, and formal specifications were written to produce preliminary data product software applications. Evaluation measures were developed for the purpose of analyzing the resulting disclosure risk and data product applicability. During the developmental phase, preliminary tests were conducted using selected Census data toward the goal of making recommendations for further testing of the most promising perturbation technique during the validation phase of the research.

The steps for the development phase were organized as follows:

1. Preliminary steps;
2. Perturbation approaches;
3. Weight calibration—raking; and
4. Data utility and risk measures.

Figure 2-1 provides the process flow of the research activities relating to the development phase (Tasks 3 and 4). It shows the general flow of tasks needed to carry out the research. The American Community Survey (ACS) three-year files from the Census Bureau contained the recodes needed for the Census Transportation Planning Products (CTPP) tables, as well as imputation flags. Swapping flags from the ACS disclosure protection process were also provided. Several preliminary steps were conducted to prepare for the processing of the perturbation approaches. Section 2.1 first discusses the design of the evaluation, and then the preliminary steps, perturbation approaches, and the raking procedure.

Two fundamental questions were considered and addressed: (1) would the tables based on the perturbed data actually be safe to release to the public? and (2) would the tables based on the perturbed data actually be useful for analysis? Section 2.2 presents data utility and disclosure risk measures to address these issues.

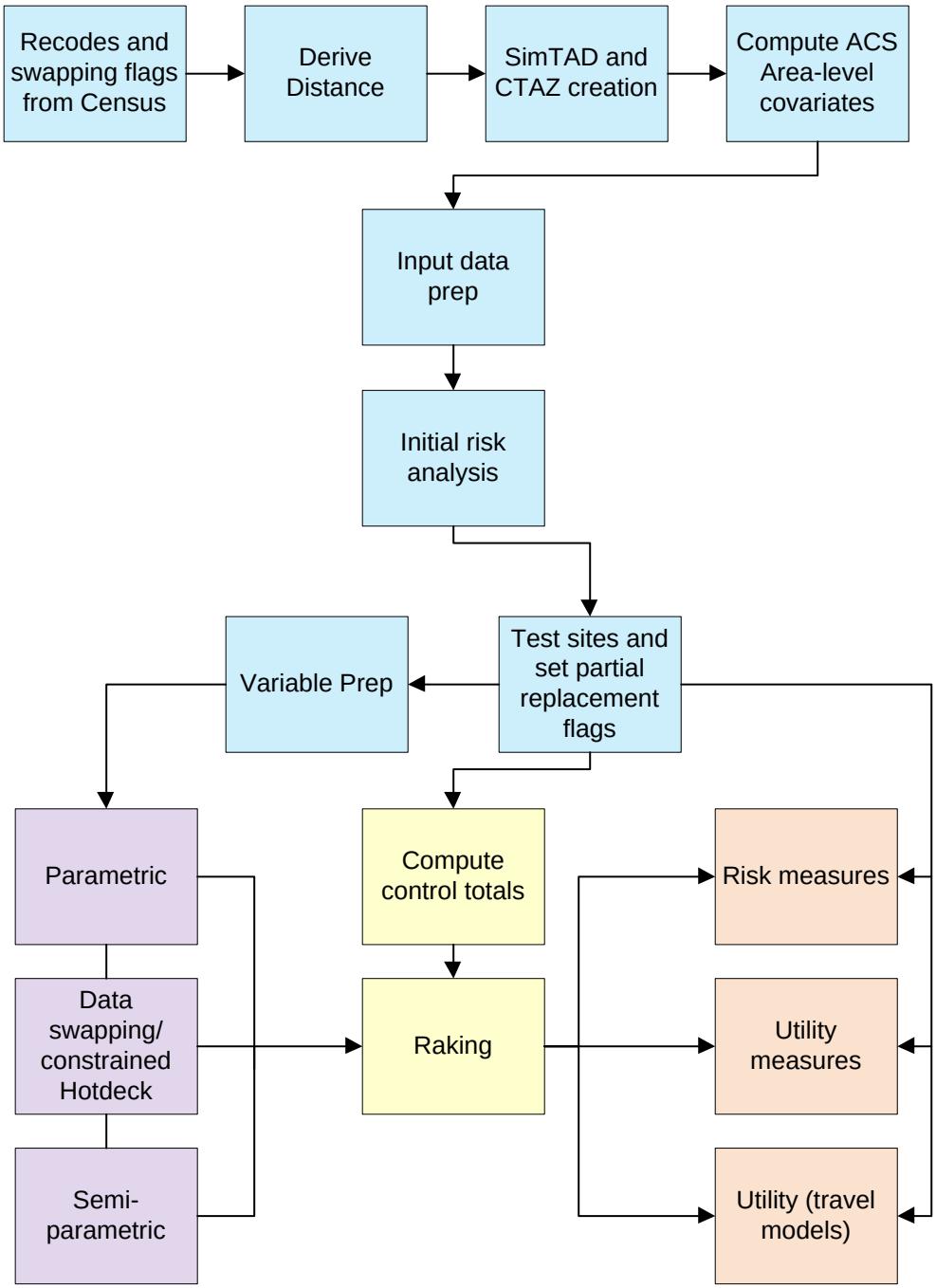


Figure 2-1. Development Phase: CTPP Research Approach

2.1 DETAILS OF THE PERTURBATION APPROACHES

2.1.1 Evaluation Design

The evaluation had the following structure:

- Four test sites (Atlanta, Iowa, Madison, St. Louis);
- Three data perturbation approaches (semi-parametric, parametric, constrained hot deck);
- Two perturbation amounts (partial replacement, full replacement); and
- Five runs each.

The treatment combinations resulted in 120 total runs (4*3*2*5). The five runs for each test site were done to gauge the replicate variability in the data perturbation results.

2.1.2 Preliminary Steps

Several preliminary steps were done to prepare for the application of the perturbation approaches. These steps were grouped into the initial processing steps, initial risk analysis, and final preparations for processing approaches.

Initial Processing Steps

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 to detect outlier commutes, and also were used as a predictor variable in the perturbation 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 latitude and longitude were not available for a block, they were imputed using the coordinates of a neighboring block. This procedure out of practicality used a straight line SAS distance, not a network distance.

SimTAD and CTAZ creation. With small ACS sample sizes in Traffic Analysis Zones (TAZs), there was some need to produce aggregates of TAZs. Therefore, two geographic variables were created by combining TAZs. One such aggregate called Census Transportation Analysis Districts (TADs) is planned for the CTPP. From the TAZ delineation business rules (draft 2009: http://www.fhwa.dot.gov/ctpp/tazddbrules.htm), Census TADs were defined as follows:

> These are aggregates of the Base TAZs and must have an estimated population lower limit of 20,000 residents. The software would issue a warning when the threshold is not respected and reject the TAD. If Base TAZs are not defined for a particular county, Census TADs can be delineated using aggregates of 2010 census tracts or block groups instead.

We were not able to obtain actual TADs formed by planning areas; however, we formed our own as a basis for the research. These TADs were called SimTADs. The SimTADs defined the area-level for

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

the computation of fixed effect area-level covariates, such as percentage in poverty, percentage minority, and so forth.

The second set of aggregates was called CTAZs, which were groupings of small TAZs. There were two different groupings formed, one such that CTAZs contained at least 50 unweighted ACS sample persons and the other such that CTAZs contained at least 500 unweighted ACS sample persons.

These SimTADs and CTAZs served different purposes. The CTAZs defined (1) the area for which random effects were based for the parametric approach, and (2) the area for which empirical distributions were computed for draws invoked in the semi-parametric approach.

ACS area-level covariates. Next, the estimated statistics (percentages, means, or medians) at the SimTAD level were created.

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

Initial Risk Analysis

There were two main approaches to identifying high risk data values, a data-driven analysis to identify disclosure risk, and a theory-driven analysis of the identifiability of the CTPP variables by the Census DRB (as discussed in Section 1.1.4). The data-driven risk analysis was a major preliminary step processed on the national database, which involved processing frequencies to detect violations of the DRB rules. The initial risk analysis was processed on the initial set of research tables as provided by the Panel in January 2010 and provided in Westat (2010). ACS variables that had already been imputed during the ACS imputation process, or swapped through the ACS disclosure process, were not replaced; that is, they were considered to have already been perturbed. This approach was acceptable to the DRB. As part of the initial risk analysis, data values were classified according to risk strata. Other useful sets of flags were the full replacement flags and the partial replacement flags, which identified the data values for which replacement values were needed from the CTPP perturbation approach.

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

- **VarName_FLG**. This 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 DRB disclosure rule.
- **VarName_FULL**. This flag was set to one for a CTPP variable (referred to generically as *VarName*) if the associated data value was not already flagged as an imputed or ACS swapped value (value swapped through ACS processing).
- **VarName_RPL**. This flag was set to one for a CTPP variable (referred to generically as *VarName*) if the associated data value was involved in a table cell that contributed to a violation of a DRB disclosure rule and it was not already flagged as an imputed or ACS swapped value (value swapped through ACS processing).

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- **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 DRB 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 DRB disclosure rule; the flag was set to 3 if the associated data values did not contribute to any violation of DRB disclosure rules and was not already flagged as an imputed or ACS swapped value; the flag was set to 4 if the associated data value was already flagged as an imputed or ACS swapped value. This flag was useful in applying the partial replacement rates, as well as in the disclosure risk measure.

- **VarName_PARTIAL**. This flag was set to one for a CTPP variable (referred to generically as *VarName*) if *VarName*_FULL was set to one and the associated data value was selected by a random process.

The data-driven analysis also identified outlier travel scenarios from travel time, estimated distance and MOT information. For such cases, the residence, workplace, travel time, and time leaving home variables were identified using the above flags. Such outliers were excluded from the data perturbation modeling process, specifically the model selection and estimation process. The travel time distributions were evaluated with the Census Bureau to determine acceptable approaches for masking the outliers. The acceptable approach was implemented and documented in a Census confidential memorandum.

The results of the initial risk assessment on the national sample identified data values at most risk of disclosure. It was conducted on three-year ACS data (five-year unavailable at that time), so therefore the results showed a bit more risk due to smaller sample sizes in the three-year ACS than in the five-year ACS. The analysis determined that over 90 percent of the TAZs were affected by DRB rules for at least one table. For most variables in the Set B “threshold” tables, about 40 to 50 percent of records contributed to a violation of a DRB rule. 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. As shown in the discussion of the impact of TAZ sizes in Section 1.1.4, small geography had a large impact on the risk levels in the tables.

Final Preparations for Processing Approaches

The final steps before processing the approaches involved subsetting to the four test sites, assigning partial replacement flags, and running an extensive variable prep module.

Test sites. The research team, assisted by subcontractor VHB, involved transportation planners in the identification of test sites. There were four test sites used in the evaluation during the development phase (Tasks 3 and 4), and two test sites for the validation phase (Task 6). The boundaries at the county level were identified, including the Federal Information Processing Standards (FIPS) code, for the four test sites.

Partial replacement rates. Two levels of perturbation rates were evaluated: full and partial. For the full replacement, all data values were replaced for a given variable, except for data values that had been imputed or swapped under the ACS processing. For the partial replacement amount, data values identified as high risk were replaced at a higher rate than other data values. The research team had consulted with the Census DRB on the partial replacement rates and agreed on a set of rates for this phase of the research. Risk strata were identified for each variable to be perturbed, and the rates were used to select and flag a sample of data values for replacement for each of the test sites.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Variable prep. After the partial replacement flags were set, predictor variables were recoded as necessary for the model selection step for the model-dependent perturbation approaches. For example, time leaving home was transformed from military time to continuous minutes (one to 1,440 in a day). 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 and select interaction terms.

A master index file (MIF) drove the process and identified the variables to be perturbed 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 to be added to the pool of candidate predictor variables were identified as well.

The predictor pools were divided into three groups:

- PredHous: Set of predictors for household level models.
- PredGQ: Set of predictors available for person-level models for persons in group quarters. This set of predictors excluded household level variables only available for persons in households such as vehicles available and household income.
- PredPers: Set of predictors available for person-level models for persons in housing units.

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.

Once the variable prep processing was completed, then the approaches could be processed.

2.1.3 Perturbation Approaches

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

Variable Types. There are different types of variables among the ones to be perturbed (continuous, circular, ordinal categorical, and unordered categorical). This presented challenges in fitting different types of models to different types of variables. The time leaving home is unique, since it has a circular aspect, as values are allowed to shift into the previous day or next day.

Variable Versions. The same variable may have multiple versions, for example, households income (HH) income (5), HH income (26), and income (continuous). The research team's approach was to use the version with the most detailed categories (or continuous) in the modeling and map to the other versions. The challenge was that if one added noise to or swapped continuous variables and then created bins to define table categories, the resulting binning might not have changed enough to protect the data from disclosure. If continuous values were perturbed and then recoded into categorical variables, then it was quite possible that a substantial fraction of cases had no effective change; that is, the replaced value may have had the same categorical value as the original value. To ensure variation from the original table,

noise addition or perturbation would have to be relatively high, which could have distorted relationships in the data. Also, a high perturbation rate could have lead to the creation of unusual data patterns for individuals with rare categorical values. For swapping, the matching calipers would have to be relatively wide, which distorts relationships in the data. The constrained hot deck was developed to address these issues.

Sparse Categories. Some variables had sparse categories, which caused problems for methods at the TAZ level given ACS sample sizes. Certain parametric models could not be estimated. For example, consider a category of industry that has very few persons in it and is highly clustered geographically. For unordered categorical dependent variables, in a random effects model, the procedure attempts to create a random intercept for each category of the dependent variable. The sparse categories interact with level of geography to produce a situation with no data for sparse categories in many geographic locations. Such situations cause problems when fitting a random effects model even when the random effect is defined at a high level of geography like Public Use Microdata Areas (PUMAs). In some cases, donors that are very similar to cases to be replaced might not be available, which could distort relationships. In some cases, potential donors with similar characteristics to cases to be replaced might not be available, forcing the use of less similar donors, which would distort relationships. To address this issue, combined TAZs were used.

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

Group Quarters. Persons in group quarters had fewer predictors than persons in HHs; therefore it required a separate model selection process. With far fewer persons in group quarters, it was necessary to fit the model at a higher geographic level. Therefore, the use of small area units at a combined TAZ level (CTAZ) in the process was not feasible.

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

Variance Estimation. The resulting variance estimates needed to account for the sampling variance from the ACS as well as the perturbation error variance. Reiter (2003) discusses the practice of creating multiple datasets and computing the variance between estimates from the multiple datasets to account for the impact of partial synthesis. An approach applied to a single dataset is presented in Section 2.1.5, and further developed and evaluated in the validation phase in Section 3.1.4.

To facilitate the discussion of the approaches that follow, Table 2-1 identifies a subset of preliminary variables that were perturbed in the development phase.

Table 2-1. Development Phase: Subset of Preliminary Variables Perturbed

Item	Variable Name	Variable Level	Type	Number of Categories
1	HH Income	HH	Ordinal categorical (OC) for semi-parametric, and Continuous (N) for parametric Continuous (N) for parametric	Continuous
2	Work-shift	Person	Unordered categorical (UC)	3
3	Time Leaving Home	Person	OC for semi-parametric, and N for parametric	Continuous
4	Travel Time	Person	OC for semi-parametric, and N for parametric	Continuous
5	Age	Person	OC	7
6	Minority status	Person	OC	2
7	Poverty status	Person	OC	3
8	Industry	Person	UC	7

2.1.3.1 Semi-Parametric

The procedure is a model-assisted approach that follows closely to Judkins et al. (2007). Initially designed for handling non-monotone (swiss cheese) missing data patterns in complex questionnaires, the 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 from 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 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.

There were two main steps involved in the process:

1. Model selection and estimation

2. Sequential prediction and perturbation

Each step in the development phase is explained in detail below. The approach has a nice property in that under full replacement, the unweighted marginal distribution for each variable is retained. The process flow for the semi-parametric approach is shown in Figure 2-2.

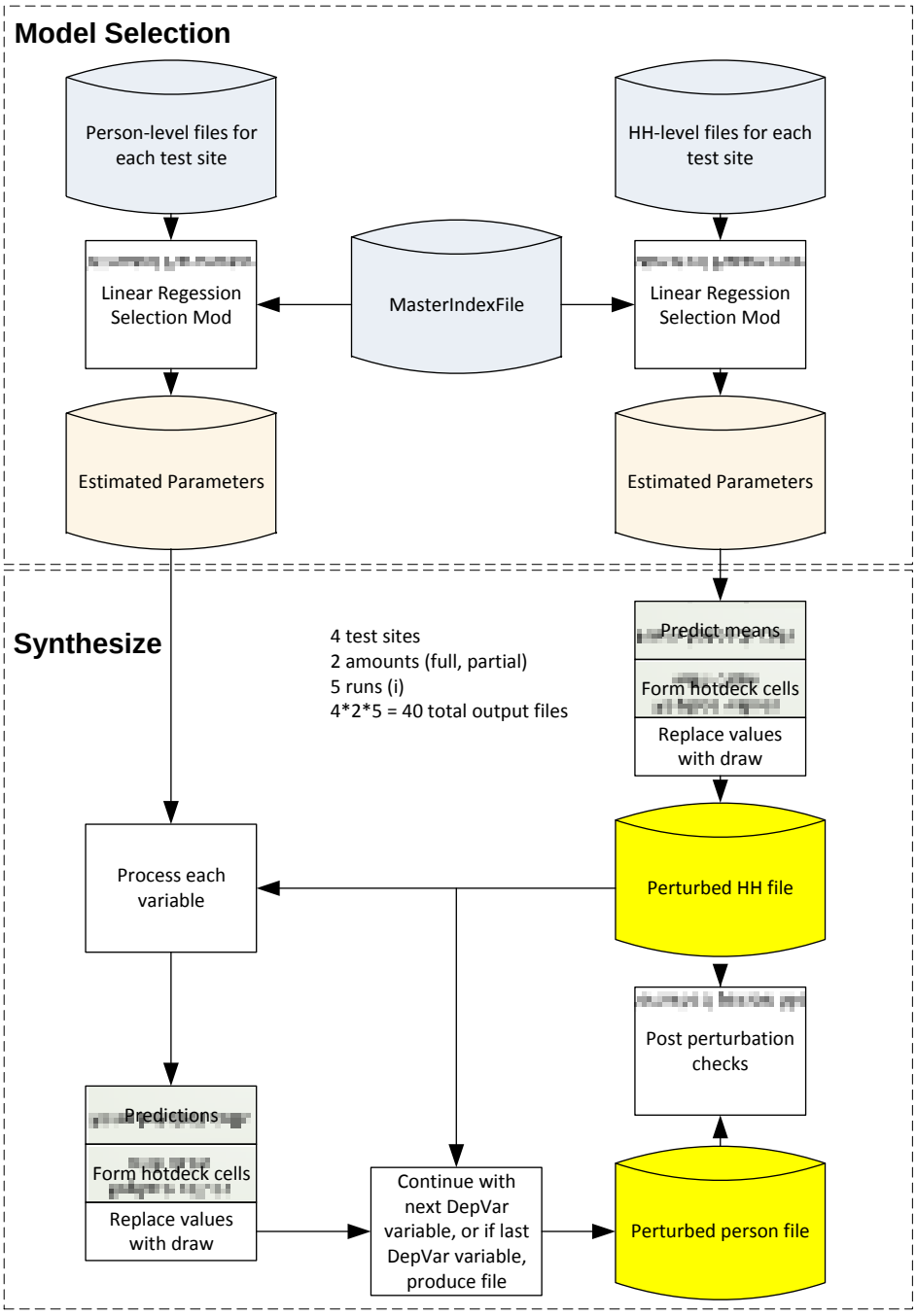


Figure 2-2. Development Phase: Semi-Parametric Approach Flowchart

Model Selection and Estimation

The model selection and estimation step was done once for each CTPP variable to be perturbed 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 perturbed 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 for HH income and at the person level for each person-level variable in Table 2-1. The modeling was done differently for variables of type OC (ordered categorical) than for 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 as hot deck cells.

Let y_{ki} denote the k^{th} variable to be perturbed for record i, where k is the item number in Table 2-1, and y represents the ACS data values. The subscript j identifies indicator variables associated with UC variables. The bolding pattern represents vectors. Therefore the model selection is essentially as follows:

$$E(y_1|\boldsymbol{y_2}, y_3, y_4, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(\boldsymbol{y_2}, y_3, y_4, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta})$$
$$E(y_{2j}|y_1, y_3, y_4, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(y_1, y_3, y_4, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta}),$$
$$\text{for } j = 1,2,3$$
$$E(y_3|y_1, \boldsymbol{y_2}, y_4, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(y_1, \boldsymbol{y_2}, y_4, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta})$$
$$E(y_4|y_1, \boldsymbol{y_2}, y_{2b}, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(y_1, \boldsymbol{y_2}, y_{2b}, y_5, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta})$$
$$E(y_5|y_1, \boldsymbol{y_2}, y_3, y_4, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(y_1, \boldsymbol{y_2}, y_3, y_4, y_6, y_7, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta})$$
$$E(y_6|y_1, \boldsymbol{y_2}, y_3, y_4, y_5, y_7, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(y_1, \boldsymbol{y_2}, y_3, y_4, y_5, y_7, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta})$$
$$E(y_7|y_1, \boldsymbol{y_2}, y_3, y_4, y_5, y_6, \boldsymbol{y_8}, \mathbf{X}) = \text{f}(y_1, \boldsymbol{y_2}, y_3, y_4, y_5, y_6, \boldsymbol{y_8}, \mathbf{X}, \boldsymbol{\beta})$$
$$E(y_{8j}|y_1, \boldsymbol{y_2}, y_3, y_4, y_5, y_6, y_7, \mathbf{X}) = \text{f}(y_1, \boldsymbol{y_2}, y_3, y_4, y_5, y_6, y_7, \mathbf{X}, \boldsymbol{\beta}),$$
$$\text{for } j = 1,2,3,4,5,6,7$$

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, earnings, age, minority status, sex, number of workers in HH, vehicles available, country of birth, travel time, and poverty status. The list of candidate predictors is given in Appendix D.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Table 2-2. Development Phase: FORCELIST Variables for Each Dependent Variable

Dependent variable	FORCELIST
Household income	MOT indicators, householder's earnings, age and minority status, vehicles available and number of workers in the household
Work-shift	MOT indicators, interaction between means of transportation and commuting distance, age, travel time, household income, minority status, poverty status, vehicles available and indicators of 24 occupations
Time leaving home	MOT indicators, interaction between means of transportation and commuting distance, age, travel time, household income, minority status, poverty status, vehicles available and the work-shift
Travel time	MOT indicators, interaction between means of transportation and commuting distance, as well as age, time leaving home, household income, minority status, poverty status, and vehicles available main effects
Age categories	MOT indicators, work-shift indicators, travel time, household income, minority status, poverty status, vehicles available, and sex
Minority status	MOT indicators, age, work-shift indicators, travel time, household income, poverty status, and vehicles available
Poverty status	MOT indicators, age, work-shift indicators, travel time, household income, minority status, and vehicles available
Industry	MOT indicators, age, work-shift indicators, travel time, household income, minority status, poverty status, and vehicles available

The model development for time leaving home, a circular variable, occurred in two phases. First, a model was constructed to predict the shift of the worker (morning start, afternoon start, and late evening start). Then work-shift was used to predict time leaving home. Also, at the person level, it was necessary to conduct the model selection separately for persons residing in group quarters (GQs) and persons in households since the predictor list for GQs was limited to predictors not associated with households.

Sequential Prediction and Perturbation

Once the model parameters were estimated for all variables, the sequential prediction and perturbations steps began for the development phase. These steps were referred to as the "Synthesize" process in Figure 2-2. Variables were perturbed, one variable at a time, beginning with the household level, transferring the perturbed household variables to the person level, and then continuing with the perturbations on person-level variables.

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 perturbed data if already available. Next, the hot deck cells were formed using highly coarsened forms of the following three contributing sources:

1. The locality;
2. The predicted values for the target variable; and
3. The sampling weights.

Within locality (e.g., CTAZ with 50 or more sample units or PUMA), the predicted values were ranked and *g1* groups were created with a close-to-equal number of sample cases within each group. We refer to the *g1* groups as prediction groups. With each prediction group, *g2* groups were formed from a ranking of the weights with an equal number of sampled cases within each group. A single set of hot deck cells was formed within each locality by cross-classifying the *g1* and *g2* groups.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Table 2-3 provides the values of *g1* and *g2* for each variable perturbed in the development phase, as well as the locality. Some discussion on the balancing of the locality, number of prediction groups, and number of weight cells is provided later in this section.

Table 2-3. Development Phase: Number of Prediction Groups and Weights Cell for Each Variable Perturbed

Dependent variable	Locality	Number of prediction groups (*g1*)	Number of weight cells (*g2*)
Poverty status and Minority status	CTAZ	6	2
Time leave home	PUMA Work-shift (day, night, graveyard shift)	100	3
Others	PUMA	100	3

NOTE: CTAZ has a minimum of 50 workers excluding those residing in GQs. Also, for industry, Workplace PUMA would be used in the production run, however, residence PUMA was used for industry due to many sparse workplaces outside the test site area.

Within each hot deck cell, a without replacement draw from the empirical distribution was conducted. The predictions and the subsequent draws from an empirical distribution occurred in a sequential manner so that perturbed values were used for the predictor variables in the model for the next variable to be perturbed.

The sequential prediction and perturbation steps are described using the items in Table 2-1 as follows. The prediction equation for OC item one (y_1) is given as (ignoring interaction terms for simplicity):

$$\hat{y}_{1i} = \beta_0 + \sum_{k=2}^{K} \beta_k y_{ki} + \sum_{l=1}^{L} \beta_l x_{li}$$

Then subsequently, as discussed above, within locality, *g1* prediction groups were formed on $\hat{y}_{1i}$ and *g2* groups were formed on the weights, within each of the *g1* groups. Let $\tilde{y}_{1i}$ represent the perturbed value drawn at random without replacement within the hot deck cell formed by the *g1*g2* groups within locality.

There were two amounts of perturbation that were conducted: full replacement and partial replacement. Under full replacement, we replaced all data values with the exception of values already imputed or swapped. Under partial replacement, the values were perturbed only if flagged for replacement; that is, high risk values were targeted as identified in the initial risk analysis described in Section 2.1.2. After each variable was perturbed, the interaction terms were recreated using perturbed values so perturbed values could be used in the prediction equation for the next dependent variables in the sequence.

Continuing sequentially for the next item in Table 2-1, there were three categories in the UC item from which three corresponding indicator variables were formed. Let the prediction equation for the j^{th} category of UC item 2 be represented as follows, using the perturbed values for item one and the ACS values for the remaining items:

$$\hat{y}_{2ji} = \beta_0 + \beta_1 \tilde{y}_{1i} + \sum_{k=3}^{K} \beta_k y_{ki} + \sum_{l=1}^{L} \beta_l x_{li}$$

For the UC variable, a clustering program (SAS Proc FastClus) was used to form *g1* clusters (prediction groups), using the three sets of predicted values $\hat{y}_{2ji}$. Then, *g2* groups were formed on the

weights, within each of the *g1* groups. Let $\tilde{y}_{2i}$ represent the perturbed value drawn within the hot deck cell formed by the *g1*g2* groups within locality. In general, after a UC variable was perturbed, indicator variables were recreated using the perturbed values.

For the OC item 3, let the prediction equation be represented as follows, using the indicator variables formed from the perturbed UC item 2:

$$\hat{y}_{3i} = \beta_0 + \beta_1 \tilde{y}_{1i} + \sum_{j=1}^{2} \beta_{2j} \tilde{y}_{2ji} + \sum_{k=4}^{K} \beta_k y_{ki} + \sum_{l=1}^{L} \beta_l x_{li}$$

The process continued sequentially until all items needing perturbation were processed. One cycle through the variables was conducted.

There were two main methodological issues to address when determining the number of groups *g1*. The first was to due to an interaction between how many groups to form and the level of geography to use for the small area. Typically, for models with high R^2 values, it would be more beneficial to rely on the predictions. The small area level of geography could be formed at a higher level, such as within PUMAs, in order to allow more prediction groups to be created for the hot deck draws. On the other hand, if models with low R^2 values or if the small area contains circumstances that were special to that area, such as industries or minority concentrations, then it may have been beneficial to rely more on the uniqueness of the specific locality, forming cells within CTAZs and having fewer prediction groups.

The second methodological issue concerned the variation in weights. Even within small areas, the ACS weight variability was high. With high weight variation, it was important to consider using the weights in the replacement process. However, it needed to be balanced with the strength of the predictions for the variable to be perturbed, and the amount of perturbation. If the models have low R^2 values and the replacement rate was high, it was preferred to have more values exchanged among records having sample weights that were of similar magnitude. If not, then the weighted estimates for the small area would be much different than the resulting ACS estimates. If the perturbations were well informed by the model, and the replacement rate was low, then concerns about the weights would be reduced.

Additive Noise for Travel Time and Time Leaving Home

Evans, et al. (1998) and Massell et al. (2006) proposed adding noise in microdata, which then are reported as totals or averages in tables. Preserving correlations and variances is a challenge when adding noise to select variables. Not adding enough noise might not adequately protect privacy. Adding too much noise severely decreases the utility of the data for analytic purposes.

The derived distance was used as a predictor for time leaving home and travel time. However, because distance was derived from residence blocks and workplace blocks, it was missing when workplace was blank (refer to Section 1.3.3 for a discussion on workplace allocation). For cases with missing derived distance (about 30 percent), noise was added to the original values. Mechanically, during the sequential prediction and perturbation step, when the predicted value was missing due to the missing distance value, additional noise was added to the original value y as follows:

$$\tilde{y}_{3i} = y_{3i}(1 + fz)$$

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Where

constant between 0 and 1
= draw from the standard normal distribution

The noise was centered at 0 with a draw from the standard normal distribution, and allowed to vary relative to the magnitude of travel time. The amount of noise also was differentiated by MOT. In a similar manner, for cases with missing derived distance, noise was added for time leaving home. For travel time, perturbed values were bottom-coded at one and top-coded at 200.

For the ACS CTPP production run, there are expected to be far fewer missing workplaces due to the workplace allocation procedures that will be implemented, which reduce the use of the additive noise approach during for travel time and time leaving home. Alternative approaches were considered, such as including distance in the sequencing to fill in for missing data and the use the completed distance as the predictor for travel time and time leaving home.

2.1.3.2 Parametric

The parametric procedure was a model-based approach which generated perturbed data through parametric models. The process involved modeling the multivariate relationships in the observed data and generating perturbed values based on the estimated model parameters. Compared to the semi-parametric procedure, for which models were used as an instrument to assist the data perturbation, the parametric procedure had modeling as its core. The gains from the parametric procedure critically relied on the validity of the models.

The parametric approach was implemented in two main steps: (1) model selection and estimation, and (2) prediction and perturbation. The modeling and perturbation process was conducted for the set of variables that were recommended by the DRB. The four test sites were combined in the modeling process to ensure that the sample size was large enough to preserve the real relationships among the variables in the data. The underlying assumption for modeling the four test sites together was that all records in the four test sites were generated from a common model. The modeling was done both at the household level and at the person level. At the person level, the workers in the households and the workers in the group quarters were modeled separately because the household characteristics, for example, number of workers in a household and number of vehicles available in a household, were not applicable for the workers in group quarters.

The set of variables which needed perturbation were classified into three categories: continuous or type N (e.g., income), ordered categorical or type OC (e.g., age group), and unordered categorical or type UC (e.g., industry). In the model selection and estimation process appropriate model structures were applied for each response variable depending on its nature. The process flow is shown in Figure 2-3. Each step is explained in detail below.

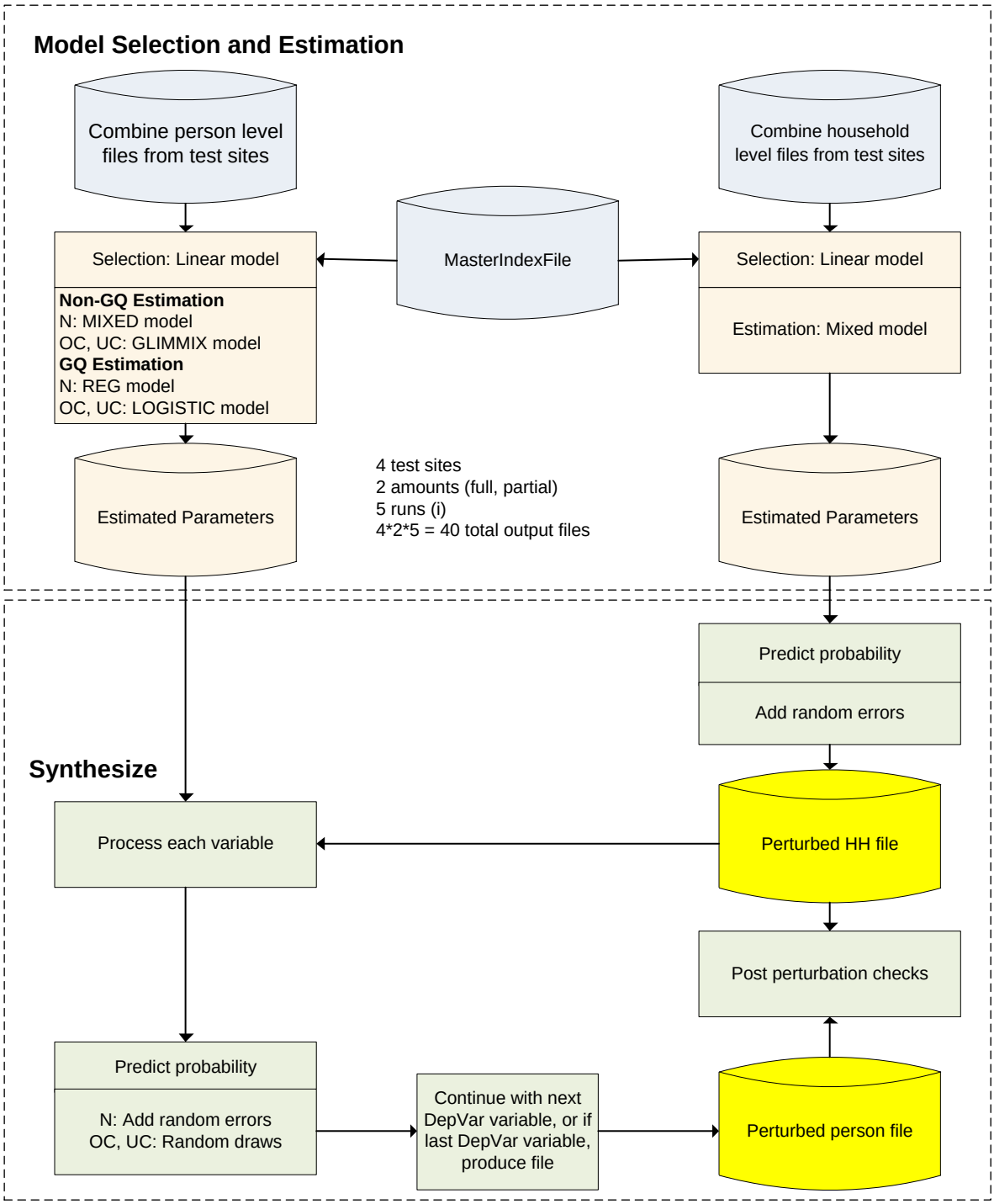


Figure 2-3. Development Phase: Parametric Modeling Approach Flowchart

Model Selection and Estimation

The first step of modeling was to select the predictors in each model. The parametric procedure used the same sets of candidate predictors and the same FORCELIST for each dependent variable as the

semi-parametric procedure did. A forward selection procedure in linear regression was used to bring in other significant predictors to improve the predictive power of the model.

Linear regressions were used in model selection for all types of variables. The selection process involved repetitive model fitting and only linear regressions could do that in an efficient way. Although linear regression may not be appropriate to model categorical variables, it can be very helpful to choose a set of predictors with a strong statistical relationship with the dependent variable. The continuous and the ordered categorical response variables were directly modeled in the model selection process. For the unordered categorical variables, indicator variables were first created at each level of the outcome and each indicator was then used as the dependent variable in the model selection. After that, the selected predictors were pooled together. A subset of 30 predictors, which have been selected most frequently when modeling the indicators, was included in the model estimation for the unordered categorical variable.

In the estimation step, more sophisticated models than those in the selection step were fit using the chosen predictors. The model structure involved a random effect for every combined TAZ (CTAZ). A CTAZ was defined as a TAZ or a group of neighboring TAZs with at least 500 non-group quarter workers in the sample. This was done to capture the variation in variables of interest such as travel time to work that cannot be explained by fixed variables such as MOT and straight-line distance to work. Since the perturbed dataset will be used to generate TAZ-level tables, unique aspects of commuting characteristics in each TAZ should be preserved as best at possible. Clearly, unique aspects of commuting characteristics would be better captured by having a random effect for each TAZ, but the three-year ACS sample sizes were too small for this (meaning that modeling software is known to produce biased estimates of components of variance with such small sample sizes and that the software would frequently not even converge). More unique aspects of local commuting life could have been preserved by including random slopes in the model in addition to random intercepts (random effects for locality), but there was concern about the ability of current software and computers to handle models of that complexity. Linear mixed models were fit for the continuous variables or their transformations, and generalized linear mixed models (GLMM) were fit for categorical variables. Logarithm transformation was taken for income before modeling to adjust for its skewness. A positive number was added to all income values before taking the logarithm. The negative income values were therefore converted to positive so that the log transformation was applicable. For workers in the group quarters, linear models and logistic models were fit, instead, for continuous and categorical variables because the sample size was too small to allow the estimation of random effects. Removing all local variation from the perturbed values does, of course, undercut the objective of the CTPP of providing local information. With partial replacement instead of full replacement, this loss of local variation was not as damaging as it could have been. Moreover, the workers in group quarters were only a small portion (about one percent) of the total workers, so it was not clear that producing local information about the commuting life of this population was a realistic goal.

Mixed models for continuous normal outcomes have been extensively developed since the paper by Scheffé (1956). Mathematically, a linear mixed model can be expressed as

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij}$$
$$u_i \sim N(0, \sigma_u^2)$$
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2),$$

where *i* denotes the *i*th CTAZ and *j* denotes the *j*th observation in CTAZ *i*. The random intercept u_i and the random error ε_{ij} are assumed to be independent. The SAS procedure MIXED was used to fit linear mixed models. The estimation method for the covariance parameter was residual restricted maximum likelihood.

Generalized linear mixed model (GLMM) is an extension of generalized linear model by the inclusion of random effects in the predictor. It is suitable to analyze the data with correlations or nonconstant variability and in which the response is not necessarily normally distributed. The application of GLMMs was well described in Agresti, Booth, Hobart, and Caffo (2000). The primary assumptions underlying the analyses performed by GLMMs are (1) the distribution of the data conditional on the random effects is known; usually the distribution is a member of the exponential family and (2) the conditional expected value of the data takes the form of a linear mixed model after a monotonic transformation is applied. GLMM models used in the parametric approach can be mathematically expressed as follows:

$$E\left(Y_{ij}\middle|u_i\right) = g^{-1}\left(\mathbf{x}_{ij}\boldsymbol{\beta} + u_i\right)$$

where g( ) is a differentiable monotonic link function and g^{-1}( ) is its inverse. The random intercept u_i is assumed to be normally distributed with mean 0 and variance σ_u^2. The SAS procedure GLIMMIX was used to fit GLMM models. The link functions that were used for binary outcomes, ordered outcomes, and unordered outcomes are logit, cumulative logit, and multivariate logit, respectively. The estimation method in the GLIMMIX procedure was based on a residual pseudo-likelihood technique. The starting values for the fixed effects were fitted by a logistic model (without random effects) using the Newton-Raphson algorithm.

The model selection and estimation steps were done once based on the original ACS data. The estimated parameters, random intercepts, and predicted values were saved and used in the next step, perturbation. Poverty was an ordered variable with three levels; however, the GLIMMIX model for poverty was not estimable due to the data sparseness in two of its three levels. As a remedy, a continuous variable, poverty index, from which poverty was derived, was fit through a linear mixed model. Poverty index was then synthesized in the perturbation step and used to re-derive poverty.

Prediction and Perturbation

Once the estimated model parameters were obtained for all variables from the previous step, the prediction and perturbation step began and was conducted in a sequential operation. The household-level variables were synthesized first, and the perturbed values were then transferred to the person level. Next, the person level variables were synthesized, one at a time, until the last variable was finished. For each variable, the prediction and perturbation was conditional on the estimated model parameters and the perturbed values of its predictors if already available. The sequential feature of the perturbation step intends to maintain the multivariate relationships among the variables.

The perturbed values for the continuous variables were generated by adding random noise to the predicted values. The predicted values were calculated using $\hat{Y}_{ij} = \tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}} + \hat{u}_i$, where coefficients $\widehat{\boldsymbol{\beta}}$ and random intercept $\hat{u}_i$ were estimated from the linear mixed model in the previous step, and $\tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}} + \hat{u}_i$ is the best linear unbiased predictor. The vector $\tilde{\mathbf{x}}_{ij}$ is different from $\mathbf{x}_{ij}$ in a way that the available perturbed values were already incorporated. The random noise was generated from a normal distribution with zero mean and estimated variance $\hat{\sigma}_\varepsilon^2$.

The perturbed values for the categorical variables were generated through random draws based upon a set of predicted probabilities. Assume *Y* is an ordered response that has C categories (c = 1, 2, . . .,C.). The predicted conditional cumulative probabilities for the C categories of the outcome can be denoted as

$$\hat{p}_{ijc} = P\left(Y_{ij} \le c|\tilde{\mathbf{x}}_{ij}, \hat{u}_i\right) = \exp\left(\tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} + \hat{u}_i\right)/\left(1 + \exp\left(\tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} + \hat{u}_i\right)\right), c = 1 \ldots C - 1,$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}}$ is a vector of coefficients estimated from a GLMM model for the category *c* of Y_{ij}. The term $\tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}}_{\boldsymbol{c}} + \hat{u}_i$ is the best linear unbiased predictor, which was adjusted at each level of the random effect. The GLMM model assumes a common intercept term but differential slopes for each level of *Y*. A random number t_{ij}between zero and one was generated for each Y_{ij} and compared with $\hat{p}_{ijc}$. The synthesized Y_{ij} took the value *c* if $\hat{p}_{ij(c-1)} < t_{ij} < \hat{p}_{ijc}$., 1 if $t_{ij} < \hat{p}_{ij1}$, and C if $t_{ij} > \hat{p}_{ij(c-1)}$.

If *Y* is an unordered outcome, the probability that $Y_{ij} = c$ for a given individual *ij*, conditional on the random effect and predictors, was predicted by

$$\hat{p}_{ij1} = 1 - \sum_{h=2}^{C} \hat{p}_{ijh}$$

$$\hat{p}_{ijc} = P\left(Y_{ij} = c|\tilde{\mathbf{x}}_{ij}, \hat{u}_{ic}\right) = \exp\left(\tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}}_{c} + \hat{u}_{ic}\right)\Big/\left(1 + \sum_{h=2}^{C} \exp\left(\tilde{\mathbf{x}}_{ij}\widehat{\boldsymbol{\beta}}_{h} + \hat{u}_{ih}\right)\right), c = 2 \ldots C,$$

In this case, the GLMM model assumed differential intercepts, slopes, and random effects for each level of *Y*. The synthesized Y_{ij} was a random draw based on the predicted conditional probabilities $\hat{p}_{ijc}, c = 1 \ldots C$.

The prediction and perturbation was done separately for workers in the group quarters. The above perturbation process was simplified by eliminating the random effects.

After the poverty index was perturbed, the ACS values of poverty were mapped back to workers according to the ranks of poverty index. Perturbed income values were transformed back to the original scale through an exponential function. Other post-replacement edits listed in the semi-parametric approach were also conducted after the perturbation was done.

2.1.3.3 Constrained Hot Deck

The third method is a constrained hot deck approach. The approach was motivated by a procedure called rank-based proximity swapping, which is applicable to ordinal variables only. The original constrained rank-based proximity swap (Greenberg 1987) bounds the swap on the target variable by limiting the distance between swapped values for the target variable. The distance between the ranks in the sort order on the target variable is to be less than a pre-defined percentage difference. The approach is extended, as summarized in Moore (1996), by controlling the swap so that the correlation between two variables is attenuated by no more than a predefined proportion.

The modification was done to perturb the continuous version of the variable while increasing the proportion of records that change value in their categorized versions of the same variable. The constrained hot deck approach limits (constrains) the range of the replacement values by forming bins on the target variable. The bins were recoded categories such that more than one published category was included in the bin. The bins were used with other variables to form hot deck cells from which donors' values are drawn without replacement from the set of all sample cases in the hot deck cell.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

The constrained hot deck is applicable only for ordinal variables, so it is not applicable to unordered variables such as industry and minority status. To address these two variables for the development phase evaluation, a controlled random swapping approach was used, which was based on the algorithm developed by Kaufman et al. (2005) and the underlying methodology for the U.S. Department of Education's Institute of Education Sciences *DataSwap* software. The methodology included the use of swapping cells formed by the concatenation of coarsened variables to be swapped. A swapping partner was selected from a cell adjacent to the cell where the target record resided, where the partner was chosen based on having a similar (or identical) weight and being close in value on a key variable. The constrained nature of this swapping algorithm caused problems and exponentially increased run time when attempting to find swapping partners with increased number of targeted swapping cases. Because of these difficulties, the swapping was done only for partial replacement in the development phase evaluation, and it was decided to not pursue the controlled random swapping approach for full replacement. It was also decided not to pursue consideration for the controlled random swapping approach in the validation phase evaluation and beyond.

To summarize, for the development phase evaluation, controlled random swapping was used for the UC variables minority status and industry under partial replacement. The constrained hot deck approach was used to perturb values of the ordinal variables under both partial and full replacement: travel time, time leaving home, age, household income, and poverty status. Details are given for the constrained hot deck approach below. A limited summary of the controlled random swapping approach is provided because it was not pursued further.

Constrained Hot Deck

The ordinal variables (travel time, time leaving home, age, household income and poverty status) were perturbed using a single draw constrained hot deck approach. The approach forms 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). For example, the CTPP plans to publish tables for the categories of travel time in Table 2-4. Bins would be formed to cover at least two categories. For example, bin 1 could consist of categories 1, 2, 3; bin 2 could cover categories 4 and 5; and so forth.

Table 2-4. Development Phase: Published Categories of Travel Time (Illustrative)

Travel Time	Description	Assigned Bins
1	Less than 5 minutes	1
2	5 to 14 minutes	1
3	15 to19 minutes	1
4	20 to 29 minutes	2
5	30 to 44 minutes	2
6	45 to 59 minutes	3
7	60 to 74 minutes	3
8	75 to 89 minutes	4
9	90 minutes or more	4

The objective of the constrained hot deck procedure was 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 steps included the following:

1. Assign the bins;

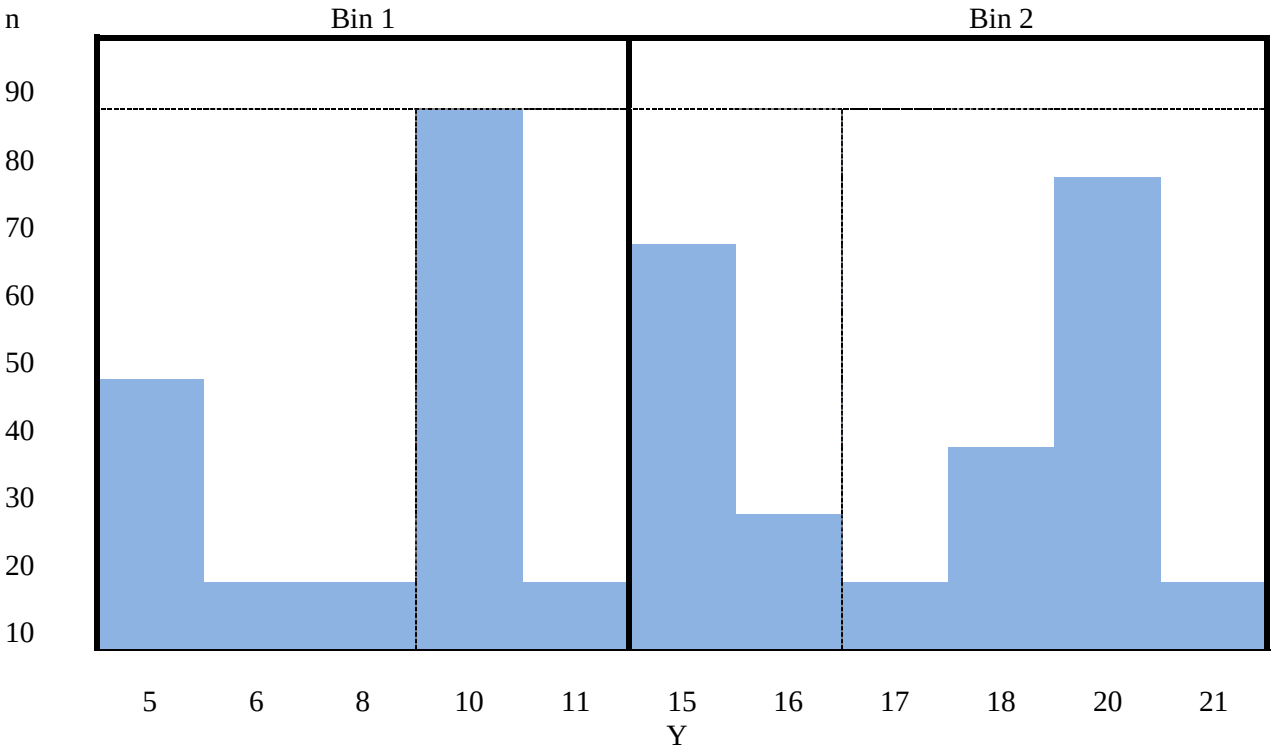
2. Form hot deck cells; and

3. Within each hot deck cell, a without replacement draw from the empirical distribution was conducted.

The hypothetical example in Figure 2-4 illustrates the assignment of bins. The figure depicts a frequency distribution, with spikes at multiples of 5. The boundaries for published categories are shown in dashed lines, while the boundaries of the bins are shown in solid bolded lines. The bins were formed with two objectives:

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

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



NOTE: Boundaries for published categories are shown in black dashed lines, and boundaries of bins shown in black solid lines.

Figure 2-4. Development Phase: Illustration of Bin Formation

Hot deck cells were formed using (1) key coarsened variables other than the target variable, (2) the bins, and (3) coarsened values of the weights. Suppose *X1* and *X2* were two key variables related to the variable to be perturbed. Within the cross-classification of *X1, X2* and the bins, *g2* groups (using the notation introduced under the semi-parametric approach) were formed from a ranking of the weights with an equal number of sampled cases within the cross-classification of *X1, X2,* bin. An SAS proc rank procedure was used to form the weight groups. For each variable to be perturbed, a single set of hot deck cells was formed by cross-classifying *X1, X2* the bins, and the weight groups.

For each data value needing replacement, a random draw without replacement was conducted within each hot deck cell. Under partial replacement, the target records were identified by their partial replacement flag (discussed in Section 3.1.2), and the replaced value was obtained through a random draw without replacement from the empirical distribution within the hot deck cell. For both full and partial replacement, all records targeted for replacement were also eligible to donate their values to others. All records not targeted for replacement were also eligible to donate their values.

The constrained hot deck approach began with replacing values of travel time. Travel time and time leaving home were linked in the process as follows. The hot deck cells for travel time were formed by MOT * time leaving home * travel time bins * weight groups. Time leaving home was coarsened to peak/non-peak when forming the hot deck cells. Peak hours were treated as between 5:00 a.m. and 8:59 a.m. For each value of travel time needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value.

Once travel time was perturbed, the bins formed for travel time were used in the formation of hot deck cells that were created for time leaving home. The hot deck cells were formed by MOT* travel time bins* time leaving home bins*weight groups. For each value of time leaving home needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value.

Next, while continuous age was not involved in the CTPP tables, it was useful for forming the bins for the categorized CTPP age variable. The hot deck cells were formed by MOT * continuous age bins * weight groups. For each value of categorized age needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value.

Household income and poverty status were linked in the process as follows: For household income, the hot deck cells were formed by number of workers in the household * vehicles available * household income bins * weight groups. Once the household income was perturbed, the ACS and the perturbed household income were merged onto the person-level file.

To perturb poverty status, a file called RAW was created with the number of workers in the household, vehicles available, ACS income and ACS poverty status. RAW was sorted by number of workers in the household, vehicles available, and ACS income. The perturbed HH income resided on the main data file. The perturbed file was then sorted by number of workers in the household, vehicles available, and perturbed income. Then the ACS poverty status from the RAW file was joined (merged) with the main data file. The ACS poverty status was replaced if flagged for replacement.

Controlled Random Swapping

For the development phase evaluation, minority status and industry were swapped through a controlled random swapping approach, similar to the *DataSwap* software mentioned earlier. The controlled random swapping was processed only for the partial replacement amount because of the constrained nature of the swapping approach that causes problems when attempting to find swapping partners. In addition, the run time increased exponentially with an increased number of targeted swapping cases. With these difficulties, it was decided not to pursue this approach in the validation phase and beyond. Therefore, only a brief summary of the approach is provided.

Based on the controlled swapping approach first introduced in Kaufman et al. (2005), this approach is similar to a common disclosure control technique used at the Census Bureau (Zayatz 2008). Westat has conducted research in collaboration with the Institute of Education Sciences on the effect on data utility. As discussed in Dohrmann et al. (2009), the swapping methodology is designed to find a

swapping partner that limits the impact on data utility. Swapping partners are selected for each target record. Swapping cells are formed by cross-classifying key categorical variables (i.e., identifiers such as MOT and minority status), henceforth referred to as swapping variables. The search for swapping partners proceeds as follows: given a selected target record in a given cell, two potential swapping partners for the target record are initially selected, one from each of the two neighboring (adjacent) cells where each potential swapping partner is chosen based on having the closest sampling weight to the target record. The search process continues by comparing the data values from the potential swapping partners with the data value of the target record, and the closest to the target record's data value becomes the swapping partner.

The first step was to select the target records for swapping. The rates for selecting targets were set to one-half the rates used for partial replacement rates so that when considering the selection of the swapping partners for each target, the resulting perturbation from the swapping would equal the partial replacement rates.

In this preliminary development, within each test site, as determined be the residence location of each household, swapping cells were formed based on the concatenation of the following:

- Risk stratum—defined by the initial risk analysis outlined in Section 2.1.2;
- Place of Work PUMA;
- MOT—defined by three categories (drive alone, carpool, other);
- HH income—defined by four categories;
- Age—defined by seven categories;
- Minority status; and
- Industry.

Values of minority status and industry were swapped between swapping partners found according to the above setup.

2.1.4 Weight Calibration

After the approaches were processed, the weight adjustment step (called raking) was done so that the weights were calibrated to reproduce select ACS estimates at the 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. 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).

The weight calibration process employed sample-based raking, meaning that the estimates for the modified estimates reflected the sampling error of the ACS control totals, rather than consider these totals to be error-free, as is often the case with calibration methods. For sample-based raking, each replicate weight for the modified file was raked to its corresponding replicate weight estimated total from the ACS.

The raking was done at the household-level to adjust household weights and at the person level to adjust the person weights. The dimensions for the household raking are given in Table 2-5 and the dimensions for person raking are given in Table 2-7. These control totals are calculated at the PUMA level. Due to the numerous place of work PUMAs (PUMAs where the ACS respondent works) with low sample counts for the test sites due to commutes outside the test site area defined by place of residence, it was decided to not process the dimensions involving place of work PUMAs for the development phase.

Table 2-5. Development Phase: Raking Dimensions for the Household File

Dimension	ByVar1	ByVar2
1	PUMA	Vehicles available (6)
2	PUMA	Number of workers in HH (6)
3	PUMA	HH income (5)

Table 2-6. Development Phase: Raking Dimensions for the Person File

Dimension	ByVar1	ByVar2
1	PUMA	Vehicles available (6)
2	PUMA	Number of workers in HH (6)
3	PUMA	HH income (5)
4	Place of work PUMA	HH income (5)
5	PUMA	Travel time (4)
6	PUMA	MOT(6)
7	Place of work PUMA	MOT(6)

NOTE: Dimensions 4 and 7 were not incorporated in the development phase due to sparse place of work PUMAs for the test sites.

Using a test file created for a comparison, programs created for this research were checked against proprietary Westat software for conducting sample-based raking on full sample and replicated weight. In addition, to ensure that it is operationally feasible to process during the production of the CTPP tables, a national level test on ACS data at the person level was conducted using the dimensions in Table 2-6. Both tests gave positive results.

Tables 2-7 and 2-8 provide percentiles of the raking adjustment factors for the person-level raking and household raking, respectively, under partial replacement from the development phase. Focusing on the range between the 10th and 90th percentiles, the range was largest for the parametric and smallest for the constrained hot deck approach. It was no surprise that the range was smallest for the constrained hot deck approach since the approach was designed to do the least amount of change possible to the values of key variables. Because the raking included HH income, then it was also not a surprise that the largest range in the factors was for the parametric approach, since the parametric approach's results on HH

income were problematic. The ranges shown in Table 2-8 for household raking were generally smaller than for person raking due to having fewer dimensions.

Table 2-7. Development Phase: Percentiles of the Raking Factors: Person Level

Approach	TSITE	Minimum	10th	25th	50th	75th	90th	Maximum
Parametric	MAD	0.64	0.84	0.95	1.03	1.06	1.08	1.41
	STL	0.57	0.89	0.95	0.99	1.05	1.11	1.98
	ATL	0.65	0.87	0.92	0.97	1.01	1.11	2.35
	IA	0.32	0.86	0.97	1.00	1.06	1.14	2.53
Semi-Parametric	MAD	0.68	0.91	0.99	1.01	1.03	1.06	1.33
	STL	0.73	0.92	0.97	1.00	1.04	1.07	1.29
	ATL	0.77	0.94	0.97	0.99	1.03	1.07	1.32
	IA	0.76	0.96	0.98	1.00	1.02	1.05	1.44
Constrained hot deck	MAD	0.91	0.97	0.99	1.00	1.02	1.05	1.08
	STL	0.71	0.93	0.96	1.00	1.04	1.08	1.24
	ATL	0.81	0.96	0.98	1.00	1.02	1.04	1.19
	IA	0.66	0.96	0.98	1.00	1.02	1.05	1.31

NOTE: Process run #1, Partial amount.

Table 2-8. Development Phase: Percentiles of the Raking Factors: Household Level

Approach	TSITE	Minimum	10th	25th	50th	75th	90th	Maximum
Parametric	MAD	0.89	0.96	0.98	1.01	1.01	1.04	1.12
	STL	0.70	0.93	0.98	1.00	1.03	1.06	1.24
	ATL	0.80	0.94	0.98	1.00	1.02	1.05	1.37
	IA	0.73	0.96	0.98	1.00	1.01	1.05	1.31
Semi-Parametric	MAD	0.88	0.95	0.99	1.00	1.03	1.04	1.14
	STL	0.85	0.96	0.98	1.00	1.02	1.06	1.18
	ATL	0.88	0.96	0.98	1.00	1.02	1.04	1.19
	IA	0.87	0.98	0.99	1.00	1.01	1.02	1.14
Constrained hot deck	MAD	0.98	0.99	0.99	1.00	1.01	1.01	1.03
	STL	0.89	0.97	0.99	1.00	1.01	1.03	1.14
	ATL	0.86	0.97	0.99	1.00	1.01	1.02	1.11
	IA	0.87	0.97	0.99	1.00	1.01	1.02	1.18

NOTE: Process run #1, Partial amount

2.1.5 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

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

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

Reiter (2003) discusses generating multiple datasets with partial synthesis to facilitate variance estimates that account for the between dataset error variance. Assume perturbations are made independently for $i = 1, \ldots, m$ to yield m different perturbed data sets. Let $\tilde{\theta}^i$ denote the CTPP perturbed

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

estimate of θ based on the *i*th perturbed data and $v(\tilde{\theta}^i)$ denote the estimated variance (computed using formula f4 below for each data set). Under certain regularity conditions, the analyst can obtain valid inferences for θ by combining $\tilde{\theta}^i$ and $v(\tilde{\theta}^i)$ as follows:

$$\bar{\tilde{\theta}}^m = \frac{1}{m}\sum_{i=1}^{m} \tilde{\theta}^i$$

$$v^m = \frac{1}{m}\frac{1}{m-1}\sum_{i=1}^{m}\left(\tilde{\theta}^i - \bar{\tilde{\theta}}^m\right)^2 + \frac{1}{m}\sum_{i=1}^{m} v(\tilde{\theta}^i). \quad \text{(f2)}$$

While multiple datasets can be generated, as in Judkins et al. (2008) for the semi-parametric approach for example, the following approach for variance estimation was developed, which is applicable to the generation of one dataset using any of the microdata approaches. The standard error of the CTPP perturbed estimate needs to account for the ACS sampling error as well as the error component due to the CTPP perturbation approach. One way of accounting for the additional variance due to data perturbation is to add a term of squared difference between the ACS and perturbed estimates as follows,

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

where the first term,

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

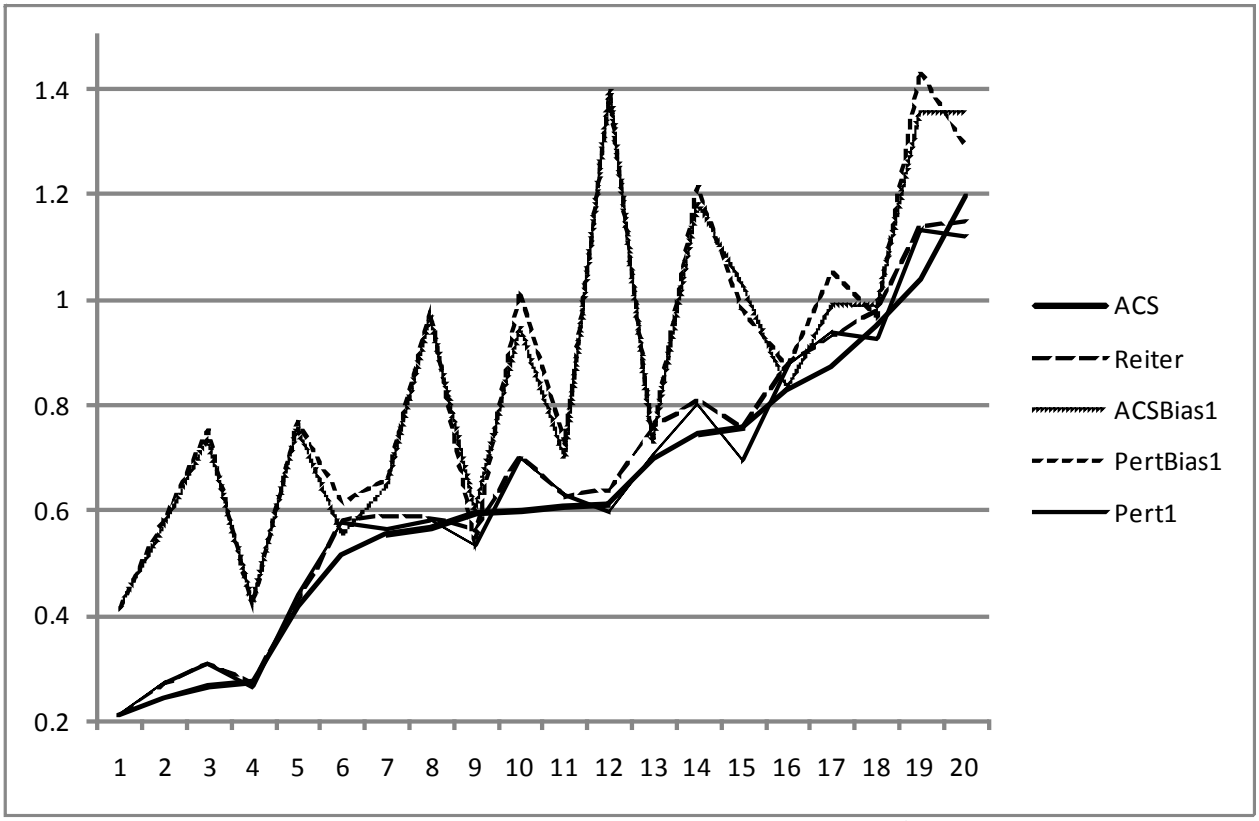
is called the naïve estimator, which results from applying the usual ACS formula directly to the perturbed data. In the formula $\tilde{\theta}_0$ represents the CTPP perturbed estimate of θ, and $\tilde{\theta}_k$ is the estimate for replicate *k*. This estimator can be biased since variance due to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

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

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

Figure 2-5 shows the estimated standard errors (SEs), the square root of the variance, of the county-level mean travel time for workers who drove alone. The computations were based on the original ACS and the perturbed datasets for the semi-parametric approach for the test site Atlanta. The horizontal axis represents the 20 counties in Atlanta. The SEs computed from (f1) (f2) and (f4) are very similar, and generally smaller than the SEs computed from (f3) and (f5). The SEs from the perturbed data (f2) are not much different from the ACS estimate (f1) because the variation in the point estimates based on the perturbed datasets from the 5 independent runs is very small. The estimated SEs computed from (f3) and (f5) account for the difference in the point estimates from the original and the perturbed data. This second term was moderate or large for some of the counties, but small or close to zero for others. This was partly because post-perturbation raking was done at the PUMA level. Although travel time was one of the raking dimensions (done at the PUMA level), the county-level estimates based on the perturbed data were not fully aligned with the estimates based on the original ACS data.

The results confirmed that more research is needed to measure the variance. Given our perturbation procedures, we would expect the impact to be minimal. One approach would be to publish SEs either from (f2) or (f4). A bootstrapping-based approach was also considered to replicate the perturbation procedures on each replicate sample (produced from the ACS). More discussion and evaluation is provided in Chapter 3.



NOTE: Source data are ATL ACS dataset and perturbed datasets using semi-parametric approach and full replacement.
ACS is the SE that uses (f1) and ACS data.
Reiter is the SE that uses (f2) and perturbed data from 5 runs.
PertBias1 is the SE that uses (f3) and perturbed data from run #1.
Pert1 is the SE that uses (f4) and perturbed data from run #1.
ACSBias1 is the SE that uses (f5) and perturbed data from run #1.

Figure 2-5. Development Phase: Estimated Standard Errors of the County-Level Mean Travel Time for Workers Who Drove Alone: ACS 2006-2008

2.2 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 data utility measures are discussed in Section 2.2.1, and the disclosure risk measures are discussed in Section 2.2.2. Each section discusses the results from the development phase evaluation and provides a recommendation for the best approach for the validation phase.

2.2.1 Data Utility Measures

The data perturbation approaches for the CTPP research 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).

While there are several techniques discussed in the literature that measure the impact of statistical disclosure control on data utility, there is no single measure that will address all planned uses of the data. For example, some measures are well suited for assessing the impact on point estimates of means or proportions, whereas others are appropriate for measuring the impact on correlations.

There were two main components to the data utility checks for the development phase. The focus of the first set of checks was to compare the ACS data with the perturbed ACS data. The comparisons checked cell means, 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. Scatter plots were used throughout to visually depict the impact of the perturbation approaches.

The best approach from the first set of utility checks was used in the second set of checks, which was to evaluate the ACS and perturbed CTPP data with travel model outputs from the four test sites. These checks were created by Rich Roisman (VHB), in consultation with Guy Rousseau (Panel chair) and Mark Freedman (Westat).

Section 2.2.1.1 discusses the CTPP and ACS data elements that affect the formulation of data utility measures. Then Section 2.2.1.2 presents a description of the data utility measures that were used to compare the CTPP perturbation approaches and provides results from the development phase. Lastly, Section 2.2.1.3 describes how the resulting data were used to compare home-based work (HBW) model outputs with the ACS data and the perturbed CTPP data from the three-year (2006–2008) ACS release and provides a summary of the results.

2.2.1.1 CTPP and ACS Data Elements That Affect the Formulation of Data Utility Measures

The set of CTPP tables and the underlying ACS data have the following characteristics that affect the formulation of data utility measures.

Tables at Various Geographies. The CTPP is a set of tables generated at different geographic levels, including, county and TAZ.

Residence and Workplace. The data utility needed to be measured for Part 1 residence tables, Part 2 workplace tables, as well as Part 3 tables showing characteristics related to the flow from residence to workplace.

Multivariate Relationships. Although the CTPP is considered a tabular product, the set of tables for a particular area can be viewed together. Therefore, it was important to measure multivariate relationships beyond the variables defining margins for a given table.

Types of Variables. There were two main types of variables that affected the formulation of the data utility measures: ordered categorical (OC, such as income) and unordered categorical (UC, such as industry). Certain measures were only applicable to certain types of variables.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Types of Estimates. The CTPP Set B tables will be based on the perturbed data, as defined in Section 1.4.3 and given in Appendix C. The tables will contain estimates of totals as well as means.

Complex Sample Design. With the complex sample design of the ACS, special variance estimation procedures are needed in order to provide the best estimates of precision.

2.2.1.2 Quantifying the Impact on Data Utility

The data utility measures for the CTPP were used to measure the impact on point estimates, variance estimates, and correlations. The main utility measures developed were the following:

- Cell mean differences;
- Weighted cell count differences;
- Standard error differences for table cells;
- Cramer's *V* differences (on HH and person-level data);
- Pairwise associations (in HH and person-level data); and
- Multivariate associations (in HH and person-level data).

Comparisons were conducted for combinations of the following:

- Four test sites;
- Three approaches; and
- Two amounts (full or partial replacement).

For each of the 24 (4*3*2) combination above, there were five perturbation runs for a total of 120 perturbed sets of data.

Cell Mean 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 $\tilde{y}$ = perturbed mean from the CTPP research
$\bar{y}$ = estimated mean from the ACS data

Cell mean differences were produced for TAZ-level and county-level residences for each of the 24 combinations (by test site, approach, and replacement amount) using the first process run (among the five runs). 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

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

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 car pool
MEANS6_4	=	car, truck, or van – in a three or more person car pool
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.

The differences were summarized in terms of the median of the differences, and the interquartile range of the differences. Bubble plots were also generated to compare the raw ACS with the perturbed CTPP TAZ and county-level means for travel time and household income.

Results

Tables E-1 through E-4 in Appendix E provide the median and interquartile range (IQR) of differences for travel time and HH income between cell means from ACS data and cell means from perturbed data. The four tables differ by geographic level of the tabulations (TAZ and county level) and by replacement amount (full, partial). In each table, results are shown for each approach and for each test site.

Table E-1 was perhaps the toughest test among the four tables since it was at the lowest level of geography (TAZ) and the highest replacement amount (full). **Among the 44 estimated differences created for each attribute (travel time and household income) and for each "by variable," with only one exception, the IQR values were the lowest for the constrained hot deck approach.** Likewise, with one exception, the parametric approach resulted in the largest IQRs of the difference. Tables E-2 through E-4 follow similarly in terms of the patterns seen in the IQRs in Table E-1, but with smaller differences between the approaches and overall lower values of IQR throughout.

In general, the medians of the differences in most cases were closer to zero for the constrained hot deck approach. This was seen more clearly at the county level in Tables E-3 and E-4. Results for county flows for mean travel time (JWMN), shown in Tables E-3 and E-4, were mixed. As discussed in Section 2.1.4, workplace-based raking dimensions were excluded from the raking process due to low sample size counts. The research team expects the dispersion in the county flows to be reduced when workplace-based raking dimensions are implemented.

Bubble plots, provided in Figures F-1 through F-8 in Appendix F, were generated at the county level and TAZ level for the four test sites to compare mean travel time from ACS data (x-axis) and from perturbed data (y-axis) under partial replacement. While the results in Tables E-1 through E-4 were from all localities no matter the ACS sample size, the dots in the bubble plots in Figures F-1 through F-8 are shown only for localities with 30 or more ACS sample cases.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Bubble plots were also generated for TAZ flows for each test site for mean travel time, for all three approaches, as shown in Figures F-9 through F-12. Each dot shown in Figures F-9 through F-12 has at least 10 ACS respondents. The size of the bubble is related to the ACS sample size in the locality.

These checks in the bubble plots paid more attention to the preservation of utility when there were enough data so that the original utility was at least moderate. The distribution of the TAZ sizes is given in Table 2-9. The table shows that the proportion of TAZs with 30 or more ACS cases (and therefore included in the bubble plots) is 19 percent, 62 percent, 55 percent and 35 percent for Madison, Atlanta, St. Louis, and Iowa, respectively.

Table 2-9. Development Phase: Proportion of TAZs with 30 or More ACS Cases, Three-Year ACS

TAZ size	MAD (%)	ATL (%)	STL (%)	IA (%)
[1,5]	25	6	7	20
[6,10]	20	7	7	13
[11,20]	24	12	16	20
[21,30]	11	12	14	12
[31,50]	12	21	22	15
[51,100]	7	32	26	12
≥100	0	9	7	8

The county-level plots in F-1 through F-4 show no apparent differential impact by approach. However, the TAZ-level plots in F-5 through F-8 and TAZ flow plots in F-9 through F-12 clearly show the constrained hot deck approach with less impact on the resulting estimates than the other two approaches.

Table 2-10 shows the median, IQR, minimum, and maximum values of the absolute relative differences for mean travel time at the TAZ level by mean travel time. TAZs with ACS mean travel time less than 5 were excluded. The data were for the partial replacement constrained hot deck approach for Atlanta, for the first of the five process runs.

Table 2-10. Development Phase: Distribution of Absolute Relative Differences for Mean Travel Time at the TAZ Level by Mean Travel Time, 3-Year ACS

ACS TAZ Mean: Travel Time (minutes)	Median (%)	IQR (%)	Min (%)	Max (%)
[5, 15)	10	19	0	100
[15, 20)	7	9	0	67
[20, 29)	4	5	0	50
[30, 45)	3	3	0	54
[45, 60)	3	6	0	28
[60, 75)	11	7	0	18
[75, 90)	4	5	0	70
≥90	50	0	50	50

NOTE: TAZs with ACS mean travel time < 5 were excluded.

Bubble plots were also produced for household income and shown in Figures G-1 through G-8. As with the mean travel time plots, the impact of approaches is indistinguishable at the county level; nevertheless, the TAZ-level plots clearly show less impact on the resulting estimates from the constrained hot deck approach. Table 2-11 shows the distribution of the absolute relative differences for mean HH income at the TAZ level by mean household income, for the partial replacement constrained hot deck approach for Atlanta, for the first of the five process runs. TAZs with absolute value of the ACS mean income less than 5000 were excluded.

Table 2-11. Development Phase: Distribution of Absolute Relative Differences for Mean HH Income at the TAZ Level by Mean HH Income, Three-Year ACS

ACS TAZ Mean: Household Income	Median (%)	IQR (%)	Min (%)	Max (%)
[$5,000, $15,000)	9	12	3	15
[$15,000, $25,000)	18	23	0	32
[$25,000, $35,000)	4	9	0	26
[$35,000, $50,000)	2	3	0	21
[$50,000, $75,000)	2	3	0	26
[$75,000, $100,000)	2	3	0	22
[$100,000, $150,000)	2	3	0	58
≥$150,000	9	12	3	15

NOTE: TAZs with the absolute value of ACS mean HH income < 5000 were excluded.

The conclusion from the results on cell means was that the constrained hot deck approach impacted the resulting cell means the least, followed by the semi-parametric approach, and then the parametric approach.

Weighted Cell Count Differences

Weighted cell counts were computed for the Set B threshold tables that produce cell counts, as listed in Appendix C, with the exception of CTPP Table numbers 32105, 32201, and 33211 because of other similar tables being generated. For each test site and for each approach, scatter plots were generated to show ACS estimates and perturbed estimates at the county level and TAZ level for both residences and workplaces, and for county flows. For the county and TAZ level plots, a dot is shown only if there were at least 10 sample cases and 10 perturbed cases in the cell. For county flow plots, a dot is shown only for flows with at least five sample cases and five perturbed cases in the cell. This was done for two reasons. First, showing data utility for estimates based on one or two cases, for example, is in direct conflict with the need to mask these small cells due to disclosure concerns. Second, more attention was paid to data utility when there were enough data such that the original utility was at least moderate. Anything less would have had little to no utility originally and would have less after perturbation.

Results

Figures H-1 through H-12 provide a visual comparison of the weighted cell count estimates before and after perturbation.

The county-level plots are given in Figures H-1, H-4, H-7 and H-10 for Madison, St. Louis, Atlanta and Iowa, respectively. While there may have been a slight edge to the constrained hot deck approach for a particular site or to the semi-parametric approach for other sites, the impact of the two approaches was virtually indistinguishable at the county level. The parametric approach had the most impact on the resulting estimates. Further investigation concluded that dispersion seen in the parametric plots are due to two variables, while the other variables were at the same reduced level of impact as the constrained hot deck and semi-parametric approaches.

The TAZ-level plots are shown in Figures H-2, H-5, H-8 and H-11. While the constrained hot deck approach had the least impact on the resulting weighted cell counts for Madison, there was virtually no difference in the impact between constrained hot deck and the semi-parametric approach for the other three test sites. The parametric approach resulted in the most deviation from the ACS estimates in general.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

The county flow scatter plots in Figures H-3, H-6, H-9 and H-12 show a slight edge to the constrained hot deck approach over the semi-parametric approach in Madison, Atlanta, and Iowa with an indistinguishable impact between constrained hot deck and semi-parametric in St. Louis. As mentioned above, we expect the dispersion in the county flows to be reduced when workplace-based raking dimensions are implemented.

Figures I-1 through I-4 provide scatter plots of ACS and constrained hot deck data for weighted counts under partial replacement. As expected, the plots under partial replacement in Appendix I show somewhat less dispersion than the corresponding plots under full replacement as given in Appendix H.

The conclusion from the results on weighted cell counts is that the constrained hot deck approach and semi-parametric approach both result in the least impact on data utility, with the parametric approach resulting in the greatest amount of dispersion from the ACS estimates.

Impact of Perturbation on Standard Errors

The following difference formula (see discussion of the research in Section 2.1.5) attempts to measure the impact on the standard error introduced by the perturbation approaches. The formula *f3* from Section 2.1.5 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 perturbation, and was computed as follows:

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

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

The standard errors were computed at the county level for mean travel time and mean HH income for each of the 24 combinations of test sites, approaches, and replacement amounts using the first process run among the five runs. 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.

Results

Table E-5 shows the comparison results under full replacement. Since Madison consists of only one county, the IQR was equal to 0 for each comparison. **Among the other three test sites, for all but two of the 33 estimates, the IQRs of the difference were the smallest under the constrained hot deck approach. The semi-parametric approach had the lowest IQRs of the difference for those two instances. The parametric approach resulted in the highest IQRs of the difference for 25 of the 33 estimates.** The median difference was closest to zero for the constrained hot deck approach in general.

Table E-6 shows the results under partial replacement. **Among the three test sites with more than one county, for all but one of the 33 estimates, the IQRs of the difference were the smallest under the constrained hot deck approach.** The semi-parametric approach had the lowest IQRs of the difference in that instance. **The parametric approach resulted in the highest IQRs of the difference for 27 of the 33 estimates.** The median difference was closest to zero for the constrained hot deck approach in general.

Therefore, the conclusion from the results was that the constrained hot deck approach had the least impact on standard errors, with the parametric approach resulting in the largest impact.

Cramer's V Ratios

As also used in Shlomo (2008), the Cramer's *V* was used to summarize the impact of the CTPP perturbation 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:

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

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 χ^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 perturbed data file, and
$V(y_i, y_j)$ denotes the Cramer's *V* on the ACS data.

Cramer's *V* differences were produced for TAZ-level and county-level residences and county-flows for each of the 24 combinations (by test site, approach, and replacement amount) using the first process run (among the five runs). 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)].

Results

Table E-7 provides the Cramer's *V* results under the full replacement amount. To summarize, the research team counted the number of times the IQRs of the differences were 0.02 higher than the other two approaches or 0.02 lower than the other two approaches.

- At the county level, among the 20 IQRs to compare (five two-way tables for four test sites), all IQRs of the differences were within 0.02.
- For county flows, the constrained hot deck approach had the lowest IQR nine times, while the semi-parametric and parametric each had the lowest IQR once. Eight times the parametric approach had the highest IQR, while the constrained hot deck had the highest IQR three times and the semi-parametric once.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- At the TAZ level, the constrained hot deck approach had the lowest IQR 12 times. Four times the parametric approach had the highest IQR, while the constrained hot deck had the highest IQR three times.

In general the full replacement amount results were best for determining the differences between the approaches. The partial replacement amounts were generally less extreme and show fewer differences. Table E-7 provides the Cramer's *V* results under the partial replacement amount.

- At the county level, the constrained hot deck approach and the semi-parametric each had the highest IQR once.
- For county flows, the constrained hot deck approach had the lowest IQR four times, while the semi-parametric and parametric each had the lowest IQR once. Four times the parametric approach had the highest IQR, while the semi-parametric had the highest IQR three times.
- At the TAZ level, the constrained hot deck approach had the lowest IQR nine times. Seven times the parametric approach had the highest IQR.

In conclusion, the constrained hot deck approach had the least impact on the Cramer's *V* measure, especially at the TAZ level and for county flows. The results show the parametric approach having the greatest impact. The results at the county level were inconclusive with respect to determining the best approach.

Pairwise Associations

Due to the sparseness of the ACS data, a majority of the TAZ 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 perturbed data.

Pearson product correlations were computed and shown in Tables E-9 and E-10 for each county between six select pairs of the following variables at the individual level: HH income, age, poverty status, time leaving home, derived distance and travel time.

Scatter plots were also generated for 11 select pairwise correlations computed for each PUMA. The 11 pairs were the following:

- Travel time with each of the following: time leaving home, HH income, derived flow distance, poverty status, and age;
- Time leaving home with: HH income, poverty status, and age;
- HH income with: age, and poverty status; and
- Poverty status with age.

Results

Table E-9 shows the results of the six pairwise comparisons for the perturbed and ACS data for each approach, for Atlanta, under full replacement. The correlations are provided for each of the process runs in the development phase in order to observe the variation in the results as the process is repeated.

The constrained hot deck approach resulted in the best retention of correlation for the following three pairs: age (AGE9) and income (AHINC), poverty status and income, and travel time (JWMN) and time leaving home (SynJWD). The semi-parametric approach did best for age and poverty status. Both the semi-parametric and the constrained hot deck approach did well for the pair of travel time and derived flow distance. The correlation between poverty status and travel time was very low and the results inconclusive as to the best approach. The same conclusions can be said for the results under partial replacement as shown in Table E-10. One interesting result was the lack of retention by the constrained hot deck approach for the pair of age and poverty status. The ACS correlation is 0.13 while the constrained hot deck approach resulted in correlation around 0.08 under the full replacement amount, suggesting a need for a modification to the approach to ensure a better linkage between the variables. However, under the partial replacement amount, the correlations from the constrained hot deck approach were between 0.12 and 0.13. There is minimal variation between the five runs in the resulting correlations for either full or partial replacement.

Figures J-1 through J-4, for Madison, St. Louis, Atlanta and Iowa, respectively, provide scatter plots of correlations for each PUMA for 11 select pairs of variables. The x-axis reflects the ACS correlations and the y-axis reflects the correlations from perturbed data. The plots show that the constrained hot deck approach had the best retention of the correlations, with semi-parametric second best.

The general conclusion was that the constrained hot deck approach retained the pairwise correlations the best, with semi-parametric doing quite well, and the parametric approach doing well in many instances.

Multivariate Associations

Woo et al. (2009) propose using propensity scores as a global utility measure for microdata as follows. The perturbed and ACS data files were stacked and T = 1 was assigned to the perturbed 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 perturbed data and ACS data were indistinguishable.

$$U = \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - c)^2$$

Where N = number in the stacked file
$\hat{p}_i$ = propensity score (logistic regression prediction) for record *i*
c = proportion of units from the synthetic data file (e.g., ½)

Results

Table E-11 shows the U statistic for each development phase run for each of the 24 combinations of test sites, approaches and amounts, for a model that includes main effects only. For the full replacement amount, the differences between the three approaches were much more distinguishable. The constrained hot deck approach had the lowest values of U, followed by the semi-parametric approach and then the parametric approach. Under partial replacement, the general pattern was similar with the exception that the semi-parametric approach did best for Iowa. Also there was minimal variation for both full and partial replacement between the five runs in the resulting U statistic. Table E-12 shows the U statistic for a model that included several two-way interaction terms among the perturbed variables.

While the constrained hot deck approach did best for all test sites and amounts of replacement, the parametric approach had lower values than the semi-parametric approach in general.

Summary of Results: Selection of Approach for Home Based Work Outputs Comparison

The first set of utility checks involved comparisons between the raw ACS data and the perturbed data. The constrained hot deck approach impacted the resulting cell means the least, followed by the semi-parametric approach and then the parametric approach. The constrained hot deck approach and semi-parametric approach resulted in the least impact on the weighted cell counts, with the parametric approach resulting in the greatest amount of dispersion from the ACS estimates. The constrained hot deck approach had the least impact on error measures (in terms of variance and bias), with the parametric approach resulting in the largest impact. The constrained hot deck approach had the least impact on the Cramer's *V* measure on two-way tables, especially at the TAZ level. The results showed the parametric approach having the greatest impact. The constrained hot deck approach retained the pairwise correlations the best, with semi-parametric doing quite well, and the parametric approach doing well in many instances. Lastly, the constrained hot deck approach had the lowest values of the multivariate association measure (U), followed by the semi-parametric approach and then the parametric approach. **Therefore, the constrained hot deck approach was chosen for the home-based work outputs comparison, under partial replacement.**

2.2.1.3 Comparison with Home-Based Work Outputs

Comparing data from travel demand forecasting models with both the raw and perturbed ACS-based CTPP tabulations provided an additional check on the usability of the resulting tables impacted by disclosure-avoidance procedures. With hundreds of travel forecasting models being employed in the United States and no standard model form, it would have been impossible to account for all models at all agencies. However, there were common elements to nearly all models that were used to develop effective comparisons. ACS CTPP is likely to be inherently less usable to planners and modelers than the decennial long form CTPP based solely on the reduced sample size. The purpose of these tests, then, was to conduct a reasonableness check to determine that the performance of the perturbed ACS CTPP tabulations was no worse than the raw ACS tabulations when compared against typical model outputs.

Tests

Potentially, each table of model output could be compared against seven corresponding ACS data sets—the raw data and three perturbation approaches for both full and partial replacement of disclosure-risked data. However, this level of analysis would have quickly made the resulting number of tests unmanageable; furthermore, some tests may be conducted at multiple levels of geography. Table 2-12 below proposes seven comparison tests for each of the four development phase test sites at the specified levels of geography. The following steps were implemented to reduce the number of matrix comparisons and reach meaningful and manageable tests.

Compare only one approach and replacement amount. Comparison tests between model output and ACS CTPP data were conducted for only one perturbation approach—constrained hot deck—which was the best approach based on statistical utility tests against the ACS raw data and against only partial replacement of values, and was accepted by the Census Bureau DRB. Direct comparisons between the raw ACS and perturbed ACS were also conducted using the tabulations shown in Table 2-12.

Table 2-12. Development Phase: Tests for Comparison of Travel Demand Model Output and Raw and Perturbed ACS Data (and Direct Comparison of Raw and Perturbed ACS) and Location of Data Tables in Appendix K

Test	Model Component	Model Data	ACS Data	Test Sites / Level of Geography									CTPP Part[2]	Universe for ACS Data	Number of Resulting Matrices[3]
				Atlanta		St. Louis		Madison		Iowa Statewide					
				County	Sub-County	County	Sub-County	County	Sub-County	Multi-County	County	Sub-County			
1	Population Synthesizer / Trip Generation	Population by Age Category	Age of Worker (8)	Yes (Table K-1)	TAZ	No	No	No	No	No[4]	No	No	Part 1	Workers 16+ in HHs	4
2	Population Synthesizer / Trip Generation	Households by number of workers	House-holds by number of workers (5)	Yes (Table K-2)	TAZ	Yes (Table K-3)	District (Table K-4)	Yes (Table K-5)	District (Table K-6)	No	No	No	Part 1	Households	6
3	Trip Generation / Trip Distribution	Person Trips	Total Workers (1)	Yes (Table K-7)	District (Table K-8)	Yes (Table K-9)	District (Table K-10)	No	District (Table K-11)	District (Table K-12)	No	No	Part 3	Workers 16+ in HHs	6
4	Mode Choice / Assignment	Average Travel Time by Mode	Mean TT (1) by MOT (7)	Yes	District	Yes Table K-13)	District (Table K-14)	No	District (Table K-15)	No[5]	No	No	Part 3	Workers 16+ in HHs	7 (*49*)
5	Trip Distribution / Mode Choice	Person trips by HH inc by mode	HH Inc (5) by MOT (7)	Yes	District	No[6]	No[7]	No	No	No	No	No	Part 3	Workers 16+ in HHs	4 *(28)*
6	Trip Distribution / Mode Choice	Person trips by age of worker by mode	Age of Worker (6) by MOT (7)	Yes	No	No	No	No	No	No	No	No	Part 3	Workers 16+ in HHs	1 *(7)*
7	Trip Generation / Mode Choice	Person trips by age of worker by mode	Age of worker (4) by MOT (4)	Yes	TAZ	No	No[8]	No	No	No	No	No	Part 3	Workers 16+ in HHs	4 (*16*)

[2] It is not clear if this naming convention will be retained in the new CTPP.
[3] Numbers in *italics* represent two-dimensional matrices split from three-dimensional matrices.
[4] DOT was unable to provide the requested model output.
[5] DOT was unable to provide the requested model output.
[6] Model output incorrectly tabulated.
[7] Model output incorrectly tabulated.
[8] Comparison not possible due to TAZ compatibility issues.

The level of geographic analysis should be chosen carefully. Comparisons at the TAZ level were planned to occur as follows:

1. Tests against synthetic population and households classified by number of workers must occur at the lowest possible level of geography to be meaningful, and thus plan to occur at the TAZ level. There were two such tests, one for population by age category, and one for households by number of workers.

2. The Panel expressed its desire to have at least one test that cross-tabulates means of transportation (MOT) with another key variable at the TAZ level. For this test, it was planned to collapse MOT to seven categories to avoid small or zero marginals. However, the research team was not able to compare age of worker crossed with MOT for the Atlanta Metropolitan Planning Organization (MPO), as discussed in the next section.

All other sub-county comparisons used super districts. Upon the recommendation of the state DOT, tests of the Iowa statewide model used multi-county districts instead of counties, which sharply reduced the number of cells in the statewide matrices and streamlines the analysis.

Reduce the number of tests (variables) to be compared. A unique value of the CTPP lies in the data on MOT and subsequent cross-tabulations of other variables with MOT; however, the model outputs of travel flows to compare with Part 3 tables require the most effort for production. The model outputs to be used in the comparisons must be easily available to the transportation planners at the test sites. The most complex tests (4, 5, 6, and 7) proposed in Table 2-12 had only one variable crossed with MOT. These tests would ordinarily produce three-dimensional matrices; although for ease of production and comparison, the data were separated for these tests into a series of two-dimensional matrices.

Reduction of Several Planned Comparisons

Despite taking the above steps to make the tests manageable, the study team encountered issues that required the elimination or reduction of several planned comparisons. Some test sites were unable to provide the requested model output because their model did not include the identified variable (age, income, etc.). In other cases, the variable was included in the test site's model but the categories specified differed from those used in the ACS. For households by number of workers, the test sites modeled fewer categories than the five coded in ACS; therefore, the ACS categories were collapsed (post-tabulation) to match the categories used by the model output before comparison. For MOT, none of the test sites model the full six categories used in the ACS tabulation, so those categories were also collapsed for the ACS data post-tabulation before comparison. In one case, the model output was tabulated incorrectly and could not be used; in another case, the model output was tabulated for the incorrect level of sub-county geography, but was usable for comparison with an ACS tabulation at the same level of geography.

Geographic compatibility for TAZs presented an additional challenge. ACS tabulations at the TAZ level used the Census 2000 TAZs, which were the latest available. As expected, models estimated, calibrated, and validated to or near base years between 2006 and 2008 (matching the three-year ACS data) updated their TAZ systems as part of the model improvement, so that the model's TAZ system did not match that of the ACS. Two of the four test sites provided equivalency files between their year 2000 and current TAZ systems to facilitate comparison and/or aggregation to multi-TAZ super districts before comparison. One test site was unable to provide an equivalency file between their two zone systems. An attempt to create an equivalency file by performing a spatial overlay using geographic information systems software was determined to be unreliable; therefore, the team eliminated TAZ-level comparisons for that test site. Fortunately, the changes in the TAZ system had a minimal effect on the area's super

districts, and so the team created an equivalency file between the year 2000 TAZ system and the current super districts for that test site to use for sub-county tabulations and comparisons.

Atlanta's modeled area was changed by directive of the Environmental Protection Agency (EPA) following the year 2000. ACS tabulations for Atlanta were controlled to the counties modeled in the year 2000, since the ones added later did not have TAZs defined for them. Iowa's analysis was limited to flows with both the home and work ends inside Iowa, even though the statewide model extends beyond the state boundary. Most person-travel (internal-external, or I-X and external-internal, or X-I) across the state line is more likely to be picked up by individual border MPO models (Quad Cities, Dubuque, Sioux City, Council Bluffs/Omaha) than by the statewide model. Through-trips (X-X) cannot be reliably identified from the ACS and in the model are more likely to be freight movements rather than person movements.

There were delays in receiving the model output from the Atlanta test site; therefore, some of the results of the comparison tests for Atlanta were not included in the development phase analysis. The Atlanta entries in Table 2-12 that were not included in the development phase are grayed out. Those tests that were included used year 2005 model outputs.

Comparisons

Most of the complex statistical tests on the ACS microdata were conducted and summarized as part of the overall data utility measures as discussed in Section 2.2.1.2. The tests against model output were one additional check on the usability of the resulting tables by transportation planners. For each test, for each test site, at each specified level of geography, the following comparisons were made:

- Raw ACS Minus Model;
- Perturbed ACS Minus Model; and
- Perturbed ACS Minus Raw ACS.

Both the absolute difference and percent relative difference were computed for each comparison and the following summary statistics are reported:

- Interquartile Range (75th percentile minus 25th percentile); and
- Median (50th percentile).

Each table also included the size of the matrix (number of estimates) for each test, and the ACS sample size (number of respondents) underlying the tabulation. For Tests 3, 4, 5, 6, and 7, the analysis was limited to travel flows where both the home end and work end were in the MPO area (or in the case of Iowa, within the State of Iowa), so the reported number of estimates reflected the deletion of out-of-area trips. The number of respondents was taken from a frequency distribution taken before tabulation, and so the actual number of respondents included in the comparisons may be slightly lower than reported. An example of a summary table is shown below.

Development Phase: Example of Summary Table

Raw ACS Minus Model			Synthetic ACS Minus Model		Synthetic ACS Minus Raw ACS	
	AbsDif	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR						
Median						

NOTE: Matrix Size: zero cells. ACS Sample Size: zero respondents.

Most tests also included charts showing the proportion of estimates that were within or beyond +/- 20 percent relative difference, which is a threshold often used in travel demand model validation tests. Individual cells with estimates less than or equal to 30 were not included in the comparisons, since the percent relative difference tends to exaggerate differences for cell estimates with small base values. Considering the summary statistics for both absolute and percent difference, as well as the +/- 20 percent threshold and the accompanying scatter plots for each test provided a good picture of the performance of the ACS estimates against the model data. The use of not applicable (n/a) in the pie charts means that the shown proportion of estimates was either for cell values less than or equal to 30, or cells with no (null) values, for example, a travel flow that could not be made using transit or was otherwise not reported in either the model output or the ACS estimates. Scatter plots may appear to contradict pie charts because even a relative percent difference of more than 20 percent for a given pair of values was minimized over the entire set of estimates when the values within the set ranged from less than 30 to several hundred thousand.

Results

As mentioned earlier, Table 2-12 shows the seven sets of comparison tests that were performed and the location of the corresponding summary tables in Appendix K.

Test 1: Population by Age Category (Model) vs. Age of Worker (ACS)

Atlanta

The county level results for Test 1 in Atlanta are located in Table K-1. The model output for Test 1 in Atlanta was divided by two to compensate for what appeared to be double counting. Following the division, total workers differed by less than one percent between model output and the ACS estimates. However, the distribution of total workers by county and the distribution of workers by age category differed greatly between the model outputs and ACS. Figure L-1 shows the distribution of total workers by county. Generally, the ACS estimates were higher for the more urbanized counties in the metropolitan area, and the model estimates were higher for the more outlying counties. Overall, there was a poor match between the model estimates and ACS estimates. **Most of the estimates exceed +/-20 percent relative difference between the model and ACS (see Figure L-2), and r-squared values were very low (see Figure L-3 and Figure L-4).**

Sub-county tests for Test 1 in Atlanta could not be completed in time for the development phase. Test 1 was conducted for Atlanta only.

Test 2: Households by Number of Workers (Model) vs. Households by Number of Workers (ACS)

Atlanta

The county level results for Test 2 in Atlanta were located in Table K-2. Figure L-5 shows the distribution of households by number of workers for the Atlanta area. Overall, ACS estimates were lower than model estimates; however, the pattern differed depending on the category of household. The model estimates were slightly lower than ACS for zero-worker and one-worker households, and the ACS estimates were higher than the model for two-worker and three-plus-worker households.

Within some individual counties, the differences appeared greater. Figure L-6 shows the distribution of households by number of workers for a select suburban county in the Atlanta area. There were sharp differences between the model estimates and ACS estimates in the select county, particularly

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

for zero-worker and two-worker households. However, total households for the county only differed by five percent between the model estimates and ACS estimates.

Overall, county-level results for Test 2 in Atlanta were good: nearly half of the ACS estimates were within +/- 20 percent relative difference of the model estimates (Figure L-7), and scatter plots for all household categories yielded r-squared values of 0.73 (Figure L-8 and L-9).

Sub-county comparisons for Test 2 in Atlanta could not be completed in time for the development phase.

St. Louis

The results for Test 2 in St. Louis were reported in Table K-3 at the county level, and Table K-4 at the district level. The distribution of households by number of workers was quite different between the ACS and the model output. Even though the total number of households for the MPO area only differed by two percent between the model and the ACS, the ACS showed fewer zero worker households and more 1 and 2+ worker households than the model output (see Figure L-10). As a result, all of the estimates at the county level and all except two estimates at the district level differed by more than 20 percent between the model and the ACS. **Overall differences between the raw ACS estimates and perturbed ACS estimates are minimal (see Table K-3 and Table K-4); however, Figure L-11 showed differences in the distribution of two-plus person households at the district level.**

It was not clear what caused the difference in household distribution between the model and ACS. The model was estimated on a recent local household survey of nearly 5,000 respondents, whereas the ACS sample size for the area is over 45,000 respondents. This difference in distribution of households by number of workers did not occur in Test 2 for Madison; however, the two MPO areas were very different in size (one county in Madison vs. eight counties in two states for St. Louis.

Iowa

Test 2 was not conducted for Iowa.

Madison

The results for Test 2 in Madison are shown in Table K-5 for the county level and Table K-6 for the district level. At the county level, both the perturbed and raw ACS estimates provided a good match to the model estimates; none of the estimates varied from each other by more than 15 percent. The model estimates were slightly higher than both the perturbed and raw ACS estimates for two-worker and three+ worker households, and slightly lower than both the perturbed and raw ACS estimates for zero-worker and one-worker households (see Figure L-12 below).

At the district level, there was more difference between the model estimates and both the raw and synthetic ACS estimates. Nearly 25 percent of the raw ACS estimates and 40 percent of the perturbed ACS estimates are beyond +/- 20 percent relative difference when compared with model outputs across all household categories (see Figure L-13 and Figure L-14). However, even with these differences (some are just beyond 20 percent) both sets of ACS estimates at the district level compared favorably with model output. Figure L-15 shows a scatter plot of raw ACS vs. model values, and Figure L-16 shows perturbed ACS vs. model values. **R-squared values for both comparisons were around 0.94.**

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Test 3: Person Trips (Model) vs. Total Workers (ACS)—Flows

Atlanta

Test 3 results in Atlanta for the county level are reported in Table K-7; district-level results are reported in Figure K-8. Figure L-17 shows the total home-to-work flows for the Atlanta area. The model estimates are higher than the ACS estimates. However, when looking at county-county flows, the overall difference dispersed across enough counties so that a county match received an r-squared value of 0.95 (Figure L-18 and Figure L-19). **At the district level, there was greater variation between the model estimates and ACS estimates across a larger number of districts (78 districts versus 13 counties), and as a result r-squared values decreased to 0.81 (see Figure L-21 and Figure L-22).**

St. Louis

The results for Test 3 in St. Louis are shown in Table K-9 at the county level and Figure K-10 at the district level. For the entire MPO area, the ACS estimates were slightly lower than the model estimates. The model estimates 1.4 million trips, and both the raw and perturbed ACS estimate 1.2 million trips. With this difference in the regional total, differences at the county and district level were expected. At the county level, the model estimated lower flows into the city of St. Louis than the ACS, and higher flows out of the city of St. Louis than the model (see Figure L-23 and Figure L-24). At the county level, about one-third of the raw ACS estimates and the perturbed ACS estimates were within +/- 20 percent of the model estimates (see Figure L-25 and Figure L-26). **However, the scatter plots showed an overall good match between the model estimates and ACS estimates, with r-squared values of 0.98 (see Figure L-27 and Figure L-28).**

At the district level for St. Louis, the results were not as good as the county level. Less than 15 percent of the ACS estimates were within +/- 20 percent relative difference of the model estimates (see Figure L-29 and Figure L-30) and the r-squared values from the district-level scatter plots go down to 0.93 (see Figure L-31 and Figure L-32).

Madison

Initially, the Madison model estimates were more than ten times the ACS estimates for Test 3. After further investigation, it appeared that the model outputs were for all trip purposes rather than home-based work trips only. The study team was unable to get a replacement matrix from the MPO in time for this report, so for comparison purposes 20 percent of each cell total was used, which is a reasonable factor for home-based work trips in the Madison area given the small size of the region and the increased number of university-related non-work trips due to the presence of the flagship campus of the University of Wisconsin. Even after factoring, ACS estimated only 42 percent of the total trips estimated by the model. As such, district-level differences tended to be much greater. In addition, many district-district flows estimated by the model were not estimated in ACS (that is, the flows were zero or do not exist). Only 12 percent of the ACS estimates were within +/- 20 percent relative difference of the model estimates (see Figure L-33 and Figure L-34). **R-squared values for raw ACS estimates versus model estimates were 0.66; the values were similar for perturbed ACS versus model estimates (see Figure L-35).** Results for Test 3 in Madison were reported in Table K-11.

Iowa

Test 3 was conducted for Iowa only at the (multi-county) district level and the results are shown in Table K-12. Statewide, both the raw and synthetic ACS estimates were lower than the model (see Figure L-36), and this pattern holds at the district level for both in-flows and out-flows (see Figure L-37 and Figure L-38). Unlike in St. Louis, where the difference between estimated in-flows and out-flows between the model and ACS was largely focused on travel to and from a single area, the differences in Iowa were balanced across the state. This pattern was expected from a statewide model. Districts 10 and 11 are the major urban areas of Iowa, Des Moines and Cedar Rapids, and had the highest number of trips,

as expected. Since the rest of the state was somewhat sparsely populated, district-to-district flows in those areas were somewhat low. As such, nearly two-thirds of individual cells fell below the analysis threshold of 30 estimates. **Ten percent of the ACS estimates were within +/- 20 percent of the model estimates (see Figure L-39 and Figure L-40). Overall, the scatter plots showed a good match between the model and ACS (see Figure L-41 and Figure L-42).**

Test 4: Average Travel Time by Mode (Model) vs. Mean Travel Time by MOT (ACS) -- Flows

Atlanta

This comparison could not be completed in time for the development phase.

St. Louis

Test 4 results for St. Louis may be found in Table K-13 for the county level and Table K-14 for the district level. The scatter plots provided sufficient information on these tests, and so pie charts were excluded. **There were poor matches between both raw ACS estimates and perturbed ACS estimates and the model estimates. The district-level matching was worse than the county level, and matches for transit times were poorer than matches for auto times.**

At the county level, r-squared values for both raw ACS estimates and perturbed ACS estimates versus model estimates for auto travel times were less than .5 (see Figure L-43 and Figure L-44). Comparing the raw ACS estimates to the perturbed ACS estimates yielded an r-squared value of .96, which was generally consistent with the pattern between the two sets of ACS estimates for other tests (see Figure L-45). For transit travel times at the county level, r-squared values for ACS estimates versus model estimates were no higher than .41 (see Figure L-46 and Figure L-47). Comparing raw ACS estimates to perturbed ACS estimates for transit travel times at the county level yielded an r-squared value of .82 (see Figure L-48).

At the district level, the comparisons were even less encouraging. For auto travel times, r-squared values between raw ACS and model estimates (and perturbed ACS estimates and model estimates) did not rise above .35 (see Figure L-50 and Figure L-51). For transit, there was no relationship between the ACS estimates and the model estimates (see Figure L-52 and Figure L-53).

Madison

Results for Test 4 in Madison are reported in Table K-15. **The results were somewhat contradictory but overall discouraging.** Although nearly 40 percent of both raw and perturbed ACS estimates were within +/- 20 percent relative difference of model estimates (see Figure L-55 and Figure L-56) for auto travel times, r-squared values were extremely low, under .20 (see Figure L-59 and Figure L-60). Scatter plots for transit travel times were not included due to the small number of non-null ACS cell estimates that are within +/- 20 percent relative difference (see Figure L-57 and Figure L-58).

Iowa

Test 4 was not conducted for Iowa.

Other Tests

As mentioned above, Tests 5 through 7 could not be completed in time for Atlanta for the development phase. The tests were planned for Atlanta only.

Summary of Results and Recommendations for Validation Phase Testing

The stated purpose of these comparison tests was to conduct a reasonableness check to determine if the performance of the perturbed ACS CTPP tabulations was no worse than the raw tabulations when compared against typical model outputs. Based on the results discussed above and shown in greater detail in Appendix K and Appendix L, the study team concluded that the performance of the perturbed ACS tabulations was equal to that of the raw ACS when comparing to model output. That being said, given the set of planned tests in Table 2-12, outside of Test 4, the more comprehensive Tests 5 through 7 planned for Atlanta were not done. The limitation with Tests 1 through 3 was that they were really testing the repercussions of synthesizing other variables and the impact of the resulting weight calibration process. While it was a good check, the research team expected less difference between raw ACS and perturbed than for the more comprehensive Tests 4 through 7.

The broader question remained of how well either ACS-based (either raw or perturbed) CTPP tabulation compared on its own to model output. Clearly, there were levels of difference between the model output and ACS for certain individual counties, districts, or TAZs (or pairs of each, in the case of flow data). But for some cases, the average difference between the model output and the ACS data was within +/-20 percent, which was generally acceptable in many model validation tests. As shown for each test, there were cases where the difference exceeded +/- 20 percent, and for those situations, caution must be exercised. Were these comparisons being conducted as part of a full model development or revalidation, further investigation into both the reliability of the ACS-based estimates and the uncertainty of the model estimate would be required.

In general, these limited results of comparing model output to the perturbed ACS tabulations were the same as comparing model output to the raw ACS tabulations. Put another way, there is little important difference between the raw and perturbed ACS tabulations for the comparison tests. For the comprehensive Test 4, while the comparison of mean travel times for St. Louis county flows showed that the relationship between the ACS raw and travel model output was essentially retained (r-square = .48) using ACS perturbed data (r-square = .46), there was some slight drop in r-square values when broken out by auto (r-square shifts from .40 to .31—or in terms of correlation coefficient, it shifted from .63 to .56) and transit (r-square shifts from .35 to .29). However, for lower geography, or further breakdowns by categories of variables, very sparse data would result. It is in these places where one or two sample cases existed and were prevalent, and where the objective of reducing disclosure risk conflicted with the objective of retaining data utility.

A key recommendation for the validation phase was to ensure full compatibility between the TAZ 2000 system and the current model TAZ system for the next test sites. Areas that could not provide a correspondence file between the two zone systems were not considered for testing (this includes the already selected tentative test sites for the validation phase). The study team requested feedback from the Panel regarding the continued use of mean travel time as a comparison statistic. The results for this measure in the development phase were discouraging, but the team did not feel the poor results are due to issues with the ACS *per se*, but are based on larger, well-documented issues with the reporting of travel time by survey respondents. These issues, such as respondent rounding of travel time estimates, introduced variability and “chunkiness” into the data that made comparisons problematic. The team considered dropping this test for the validation phase.

Finally, the study team also requested feedback from the Panel regarding the number and nature of the validation phase test sites. Nothing in the development phase comparisons tests suggested that the nature of the differences between model estimates and ACS estimates changes whether the model outputs come from a trip-based model or an activity-based model, from a small or large MPO, or from a statewide

model. The team's experience during the development phase suggested that focusing on more detailed geographic comparison tests for a smaller number of test sites (perhaps two) during the validation phase would provide more valuable information to all CTPP users.

2.2.2 Disclosure 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. The research team and the Census DRB recognized that combinations of just a few variables could lead to a single sample unit (sometimes referred to as a sample unique or singleton). The DRB set up rules to reduce the risks associated with small cells. For discussion about the DRB rules and the disclosure risks associated with sample unique cases, please refer to Sections 1.1.2 and 1.1.3.

As discussed in Section 1.1.2, tables can be linked together to form a string of identifying characteristics (referred to as a "key"). Perturbation of the data and/or generation of perturbed 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 perturbation rate will make more of a difference if a close acquaintance knows someone is in ACS. The perturbation replacement rate is a factor that affects both utility and risk.

Shlomo and Skinner (2009) consider measurement error in their risk assessment. Similar in concept, the research team discussed additional sources of data protection, whether it was through sampling, the realization of moving and workplace changes over time, or measurement error created through ACS swapping, ACS imputation, and the perturbed CTPP data.

The research team met with the DRB to present the following plans for computations related to disclosure risk measures. The general approach was to bring together measures of various risk elements, including a measure of the amount of changed information. **The measures were found acceptable by the DRB.** The DRB also provided some comments to consider for each of the measures.

While these risk components could be looked at separately, because there was a buildup of a series of factors, the *product* of the following risk components was therefore considered to quantify the overall risk as a score.

Matchability to the ACS Public Use Microdata Sample (PUMS). In Krenzke and Hubble (2009), a data analysis on the state of Maryland estimated the matchability to the ACS PUMS of a constructed microdata record through table linking, given a singleton in the CTPP tables. Given the outcome, and assuming the same set of CTPP variables as used in the analysis, an estimated 98 percent of CTPP singletons were identified on the ACS PUMS.

Using the three-year ACS, the risk involved in matching to the ACS PUMS was evaluated using the current set of variables involved in the set of flows tables. There were about 50 percent to 75 percent of the records (depending on the test site) that could be uniquely identified using the 10 flow attributes and Public Use Microdata Area (PUMA). About 80 percent of them are flow singletons. Therefore, about 40 percent to 60 percent were high risk exact matching singleton flows (e.g., comes from 50 percent * 80 percent). With a 2/3 subsample for the PUMS, that results in an expected match rate of about 27 percent (under 50 percent uniques). The match rate should be less for the five-year ACS since more sample cases would be available in each PUMA.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Let r_1 = the proportion of sample unique records at risk due to matching to the ACS PUMS file.
≈ 0.27

Sampling rate. Sample uniques are not necessarily population uniques due to the ACS sampling rate. Sampling reduces the risk of disclosure as compared to a census of individuals because a data snooper might not otherwise know that a case is a population unique. For five-year estimates, the sampling rate after nonresponse is about 7.4 percent. The *mu-Argus* 4.1 manual provides a discussion of how disclosure risk is measured by an approximation to the hypergeometric function in the software for the microdata using the sampling fraction when an intruder knows the unweighted cell count is one (sample unique), for example. Mainly developed at Statistics Netherlands, *mu-Argus* 4.1 is a freely downloadable software to facilitate statistical disclosure control (SDC). One of the features of the software is the estimation of initial risk, and the formulas provided in the manual were used as follows: differential sampling rates, nonresponse, and a calibration adjustment were taken into account by taking the inverse of the full sample ACS final weight (w), as $f = \frac{1}{w}$. Then the risk component (r_{3i}) for each record i was assigned as follows:

Let r_{2i}

= $-log(f_i)(f_i/(1-f_i))$, if at least one item j for person i was associated with a violation of the DRB disclosure rules as a singleton (according to the minimum value across the risk strata (*VarName*_STRT) variables),

= $(f_j)/((1-f_j)^2)(f_i\ log\ (f_i)+(1-f_i))$, if at least one item j for person i was associated with a violation of the DRB disclosure rules as a double (according to the minimum value across the risk strata (*VarName*_STRT) variables),

= f_i, otherwise (for simplicity). Through further investigation, this was modified to $f_i/2$ for the production run.

Residence and Workplace Mobility. Because the ACS selects addresses, moving residences or changing workplaces during the five-year time span introduced uncertainty. About 34 percent of householders moved within the past three years. This is an interpolation of the movement within one year (20 percent) and within five year (49 percent) change in residence, as estimated in the 2000 Census. Based on McWethy (2008), 42 percent of persons changed employers during a three-year period according to data from the Survey of Income and Program Participation. This may be a conservative estimate since it is recognized that changing locations under the same employer and changing employers in the same location were not included in the estimate. For flow tables, the union of residence and workplace mobility impacted the protection level of the flows, assuming independence between the mover and workplace change rates for simplicity.

Let r_3 ≈ 1-0.34, for Part 1 residence tables
≈ 1-0.42, for Part 2 workplace tables
≈ 1-0.62, for Part 3 flow tables

Measurement error. Imputation, swapping flags, and group quarters (GQ) synthetic data flags were available for our research. The GQ synthetic data flags identified data values for which a perturbed value exists. Since these flags would not be available with the CTPP tables, the associated values are considered masked. Therefore, before applying the perturbation approach to the ACS data, measurement error was already inherent in the ACS data through the following ways:

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

- Imputed data values were inherent in the data. The national imputation rate varies from near 0 percent for sex to about 13 percent for income, earnings and poverty.

- The Census Bureau applied targeted swapping, where swapping was applied to higher risk variables and records, and used to reduce the risk of disclosure in ACS data products. The swapping rate is kept confidential within the Disclosure Review Board (DRB). A discussion of the Census Bureau's use of target swapping is in Zayatz et al. (2009).

- The Census Bureau applies a partial synthetic data approach to the group quarter records. A discussion of the application is given in Zayatz et al. (2009).

- Other measurement error. Considerations for Census work on quantifying measurement error related to reporting and data keying. However, we did not include a quantified value to account for this type of measurement error in the data.

The motivation for the following measure was to determine the proportion of flow variables that were masked by the CTPP perturbation approach, or swapped or imputed through Census Bureau processing, among the records at most risk determined by singleton or doubleton flows. Therefore, the proportion of flow variables that changed value by the CTPP perturbation approach, or were imputed or swapped during ACS processing, was used for record *i* for each of the *J* variables to be perturbed in the flow tables.

This measure accounted for the various versions of each variable to be perturbed. For example, income has a five-category version and a nine-category version, and both are accounted for in the measure. This measure was computed only for records that were singletons or doubletons in any of the flow tables.

Let $r_{4i} = 1 - \frac{\sum_j^J k_{ij}}{J}$

where k_{ij} = 1, if flow variable *j* for record *i* has changed value or swapped or imputed
= 0, otherwise.

and where *J* is the number of flow variables being perturbed.

Overall Risk Score

The overall risk score was computed as the product of the four risk components for record *i* as:

$P1_i = r_1 * r_{2i} * r_3 * r_{4i}$

Summary tables were produced and provided to the Census DRB for review on August 16, 2010. The risk estimates for the measurement error components and the overall risk measure were provided. More detailed output was provided to show how much change occurred amongst the records at highest risk. While the DRB was acceptable to the resulting risk levels presented, the DRB requested the following:

1. Partial replacement rates be modified.

2. Investigate the characteristics of the highest risk records that remain.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

3. Investigate the magnitude of change made to each variable.

4. Investigate the risk level in matching to the ACS PUMS. This work has been done and is provided under risk component r_1.

This work was conducted and a memo was prepared for further DRB review and considered acceptable. Thus the consensus was to move ahead to the validation phase with the modified partial replacement rates and make adjustments as needed according to the above requests.

2.3 RECOMMENDATION FOR THE VALIDATION PHASE

With the Census Bureau DRB's consent on moving forward under the current risk levels presented to them in the August 16, 2010 meeting, the evaluation could focus solely on the utility results. Through the various measures given in the first set of utility checks, the constrained hot deck approach had the least impact on data utility and was chosen for the second set of utility checks, which consisted of comparisons with travel model output. The results of the travel model output comparisons showed no immediate cause for concern as to the magnitude of differences between the perturbed CTPP data and the travel model outputs.

While the constrained hot deck approach was recommended, it was only applicable to ordinal variables. Therefore, the semi-parametric approach was recommended for the small number of unordered variables (e.g., industry) since it performed well in the utility tests. The separate programs written for each approach were consolidated into a single program for use in the validation phase.

3. Validation Phase

The primary objective of the validation phase was to validate the data replacement procedures chosen from the development phase. Other goals for the validation phase were to confirm the following:

- Compliance with the Census Bureau's Disclosure Review Board (DRB) disclosure rules for the Census Transportation Planning Products (CTPP);
- Preservation of the properties of the original data; and
- Operational feasibility of the data replacement approach.

The research team implemented the data perturbation techniques identified during the development phase for two test sites using the five-year accumulation of 2005–2009 American Community Survey (ACS) data. The research team prepared software for the approach selected in the development phase. The research team tested the use of disclosure-proofed data in comparison to the use of raw ACS data in the same manner as was done in the evaluation conducted in the development phase. The impact of the approach on disclosure risk and data utility was measured. The research team was focused on several aspects of the project, as described in the following subsections.

Additional CTPP Tables

An outcome from the October 25, 2010 meeting with the Transportation Research Board (TRB) Panel was to add three new two-way tables for Parts 1 and 2 that include means of transportation. The three new variables that were examined for Part 1 and Part 2 tables were workers in households (0, 1, 2+), minority status (Y/N), and presence of children 17 and under in household (Y/N). The perturbation approaches were adapted to address the additional tables.

National Implementation

With an eye focused on moving forward, as each program was checked for its adaptability for the validation phase, changes were made to the processing in order to improve upon the operational feasibility of the production process relating to the 2006 to 2010 ACS-based CTPP tables.

Composite Two Perturbation Approaches

Programs were prepared to combine the constrained hot deck and the semi-parametric approaches into one processing step. The combination of the two approaches is considered as one perturbation approach in the validation phase. The constrained hot deck was used for ordered variables, whereas the semi-parametric was used for unordered categorical or binary variables.

Raking – New Sub-PUMA Dimension

As requested by a state representative during the AASHTO advisory board meeting, some consistency has been built in at the sub-Public Use Microdata Area (PUMA) level. A raking dimension

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

was added at the combined Traffic Analysis Zone (TAZ) level, in which the combined TAZs have at least 300 ACS sample cases (CTAZ300). Three hundred ACS records represent about 4,000 workers.

Utility Measures

Formulas were derived to approximate the variance estimates that account for the impact of the data replacement approach. The research team has provided the variance formula to the Census Bureau's DRB, and the DRB approved the formula for production. The research team described to the Census Bureau's ACS operations staff how the variances for the resulting perturbed data will need to be calculated.

In terms of reporting of utility measures, there are a couple of other items to note. As discussed in the Interim Report meeting on October 25, 2010, utility measure reports include an additional comparison of the distributions of the perturbed data versus the raw data for median values and quantiles (e.g., 75th percentile). In addition, for ease of graphical comparison, a 45-degree line (indicating equality between estimates from perturbed and raw data) was included in most comparison graphs for the validation phase.

Travel Model Outputs

The development phase approach for comparing ACS with travel model outputs was applied to Olympia, WA, and again to Atlanta. Olympia was selected to ensure the data use considerations of small and medium Metropolitan Planning Organizations (MPOs) were included as part of the comparison tests.

The Census Bureau staff attempted to work with the instructions from Guy Rousseau and the research team. The team continued attempts to provide assistance and facilitate the process of running Atlanta's population synthesizer at the Census Bureau to compare synthetic population with raw and perturbed ACS, although given complications relating to implementing the process on site at the Census Bureau, the Panel agreed to not pursue this particular analysis further.

Simulation

As discussed in the Interim Report meeting, some statistical tests could be done to examine if there was no effect due to the perturbation (e.g., null hypothesis = no effect). A simulation was conducted to identify any potential bias or impact of perturbation on variances on drive-alone travel times. To investigate the potential for bias, and to investigate the stability of variance estimates, 1000 simulated samples were drawn from the Olympia test site. The perturbation approach, including raking, was processed for each sample. Several outcomes were computed for CTAZ300 areas, which were formed by combining TAZs until there were at least 300 ACS sample cases (about 8,000 in population). Results are presented in Section 3.2.1.

TAZ Sizes

Using the initial risk analysis results from the validation phase on five-year ACS data, the research team revised the TAZ sizes in the white paper written on February 10, 2010. Section 1.1.4 provides the contents of the revised white paper, which contained updated information aimed to be useful for planners at mid- and smaller sized MPOs as they began to develop Census TAZs.

Census Staff Operations Needs

Important criteria for the approach were that it meets Census needs for confidentiality and data protection and that it is relatively easy for Census to take over with minimal disruption to their existing systems. In addition, the programs and documentation must allow for easy transfer of operational ownership to Census IT staff. Throughout the process, the Census Bureau staff who are responsible for the CTPP were informed of the approach through meetings and via email discussion. The staff also participated in the October 25, 2010 Panel meeting that reviewed the Interim Report. Further discussion with the ACS operations staff occurred from March 2011 through May 2011 regarding implementation of the Set A and Set B tables, as well as the variance estimation approach. The variance approach was also approved by the ACS Statistical Design team. The discussions were conducive to the implementation of the approach in the production run on ACS 2006–2010 data.

Data Users Needs

On August 25, 2010, the research team gave a presentation at the AASHTO advisory board meeting and responded to questions from state representatives. CTPP data users also participated in the Interim Report TRB Panel meeting on October 25, 2010, and responses were given to any questions raised. Likewise, the research team gave a presentation at the TRB annual meeting on January 23, 2011, giving data users the opportunity to raise concerns about the impact of perturbation procedures on data quality. The research team also participated in the CTPP table subcommittee meeting held on May 2, 2011, and will participate in the Using Census Data for Transportation Applications Conference in the fall of 2011 in Irvine, California.

Statistical Methodology

The research team conducted a lunchtime presentation on January 18, 2011, for the Washington Statistical Society. The presentation covered the basic contents of the Interim Report, including results from the development phase evaluation. The research team fielded questions, such as how weights are used, how generalizable the approach is, and how travel model outputs are considered in the evaluation. Future discussions with the statistical community are planned for August 2011 at the Joint Statistical Meetings. The research team will be presenting the work on the constrained hot deck as a viable perturbation approach, as well as the new research conducted on variance estimation.

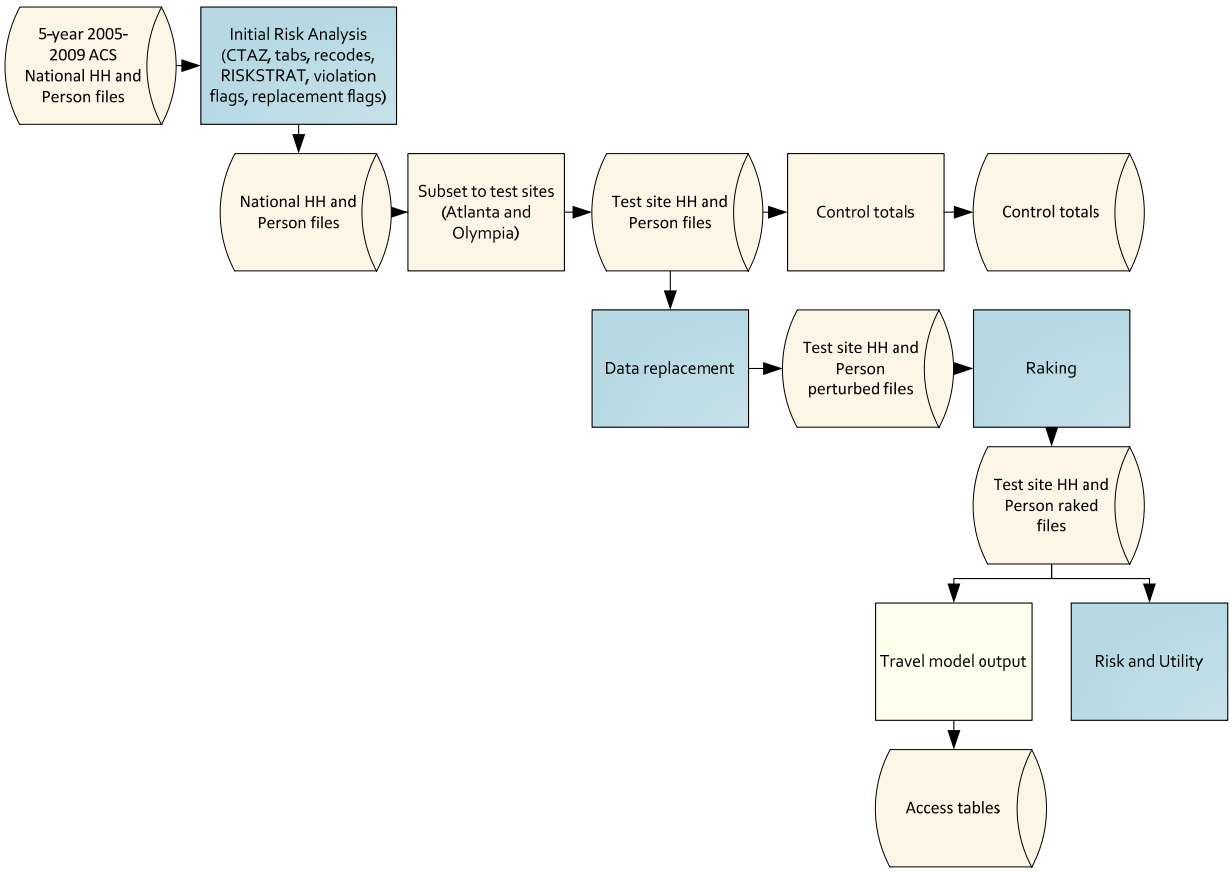
In Section 3.1 a description of the software is provided, which includes a flowchart of how the software fits into the Census operations, connecting the ACS data files and the CTPP tabulations. Details of the perturbation approach are provided, along with the metrics used to verify the impact on disclosure risk and data utility. The results from the utility and risk measures from the validation testing are included in Section 3.2.

3.1 PERTURBATION APPROACH APPLIED DURING THE VALIDATION PHASE

During the validation phase, while working on site at the Census Bureau with the 2005–2009 ACS data, the data perturbation technique that was identified during the development phase evaluation was further developed. The five-year ACS data were processed through four main steps of the procedure. The steps for the validation phase were organized as follows:

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

Figure 3-1 provides the process flow of the research activities relating to the validation phase. The ACS five-year files from the Census Bureau contain the recodes needed for the CTPP tables, as well as imputation flags. Swapping flags from the ACS disclosure protection process were also provided. Several preliminary steps were conducted to prepare for the processing of the perturbation approaches. The parameters of the validation testing are discussed first and then the initial risk analysis, data replacement approach, raking procedure, and risk and utility measures.



Note: For nationwide implementation, there will be no step for subsetting to test sites, and no travel model output generated. Also, the program for generating control totals will be incorporated into the raking process.

Figure 3-1. Validation Phase: CTPP Research Approach

The two fundamental questions addressed during the development phase, also need to be confirmed in the validation phase: (1) are the tables based on the perturbed data actually safe to release to the public, and (2) are the tables based on the perturbed data actually useful for analysis? In the development phase, the evaluation had the following structure:

- Four test sites (Atlanta, Iowa, Madison, St. Louis);
- Three data perturbation approaches (semi-parametric, parametric, constrained hot deck);
- Two perturbation amounts (partial replacement, full replacement); and
- Five runs each.

The treatment combinations resulted in 120 total runs (4 * 3 * 2 * 5). The five runs for each test site were done to gauge the replicate variability in the data perturbation results.

For the validation phase, the processing included the following:

- Two test sites (Atlanta, Olympia);
- One combined data perturbation approach (semi-parametric for unordered categorical and binary variables, the constrained hot deck for ordinal variables);
- One perturbation amount (partial replacement); and
- Five runs each.

The validation results were compared with the results from the development phase.

This chapter provides the activities and methods that were used for the validation phase. To the reader, it may appear to follow closely to the discussion in Chapter 2 for the development phase. In doing so, the validation phase is described in case there is need to replicate the development and validation phase procedures.

3.1.1 Initial Risk Analysis

The set of initial risk analysis modules were processed to generate tables. The tables were generated to flag data values that violate the DRB rules and therefore were at the highest risk of disclosure. The table generator for this step of the research was modified to incorporate the additional tables requested by the Panel. Several preliminary steps were necessary within the initial risk analysis component to prepare for the application of the perturbation approach.

Initial Risk Analysis Steps

In the initial processing steps, several variables were created to be used in the initial risk analysis and also in the implementation of the approaches.

Creation of Combined TAZs. With small ACS sample sizes in TAZs, creating aggregates of TAZs allowed for more stable variables for model predictors, allowed for some consistency in estimates at a fairly small level of geography, and allowed for more stable estimates to be evaluated for perturbation impact. In the transportation data users community, one such aggregate called Census Transportation Analysis Districts (TADs) is planned for the CTPP, as discussed in Section 2.1.2. In this research, these TADs were called CTAZ1000, since they were formed by combining TAZs until there were 1,000 sample cases (representing approximately 25,000 in population). The CTAZ1000 areas defined the area level for

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

the computation of fixed effect area-level covariates, such as percentage in poverty, percentage minority, and so forth.

There were two other levels of CTAZs. For the formation of hot deck cells in the perturbation approach, two levels of CTAZs were used; one set of CTAZs contained 50 or more sample cases (CTAZ50), and one set contained 300 or more sample cases (CTAZ300). The CTAZs defined the area for which empirical distributions were computed for draws invoked in the semi-parametric and constrained hot deck approach. In general, residence CTAZs were first created within counties, and crossed the county boundary if a county was small.

ACS Area-Level Covariates. Next, estimated statistics (percentages, means, or medians) from ACS data at the CTAZ1000 level were created. The set of area-level predictors is provided among the list of predictors in Appendix D.

Input Data Preparation. In this step, it was necessary to combine the outcomes of the prior processing steps with the household-level file. The output files from this step are a person-level (subset to workers) file and household-level file. Other recodes were needed for the creation of the pool of predictor variables in the modeling approaches.

Initial Risk Analysis

The data-driven risk analysis was a major preliminary step processed on the national database. The step involved computing frequencies in cross classified tables to detect violations of the DRB rules. As agreed to by the DRB, ACS variables that were already imputed during the ACS imputation process, or swapped through the ACS disclosure process, were not replaced; that is, they were considered as to have already been perturbed. As part of the initial risk analysis, data values were classified according to risk strata.

The following flags were created as discussed in Section 2.1.2 to assist in the perturbation process as well as in the disclosure risk measures: VarName_FLG, VarName_RPL, VarName_FULL, and VarName_STRT.

The travel time distributions were evaluated with the Census Bureau to determine acceptable approaches for coping with the outlier commuting patterns. The acceptable approach was implemented and documented in a Census confidential memorandum.

The results of the initial risk assessment on national sample identified data values at most risk of disclosure. It was conducted on five-year ACS data. The analysis determined that about 90 percent of the TAZs were affected by DRB rules for at least one table. For most variables in the Set B “threshold” tables, about 30 percent to 50 percent of records contributed to a violation of a DRB rule. In general, the risk was attributable largely to flows and cell means. When reporting flows, data were sparser and there was concern about a scenario involving an intruder linking tables together. Detailed categories in MOT and certain other variables (e.g., where cell means are computed) also contribute to the disclosure risk. As shown in the discussion of the impact of TAZ sizes in Section 1.1.4, small geography had a large impact on the risk levels in the tables.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

3.1.2 Data Replacement

In general, there were a number of methodological challenges that were addressed when implementing the perturbation approach, as discussed in Section 2.1.3. Furthermore, in the validation, the research team combined the constrained hot deck and semi-parametric approaches into one program. It also had addressed the challenge of broadening the processing so that it will be ready to process the nation. The validation phase processing flow for the data replacement step is illustrated in Figure 3-2.

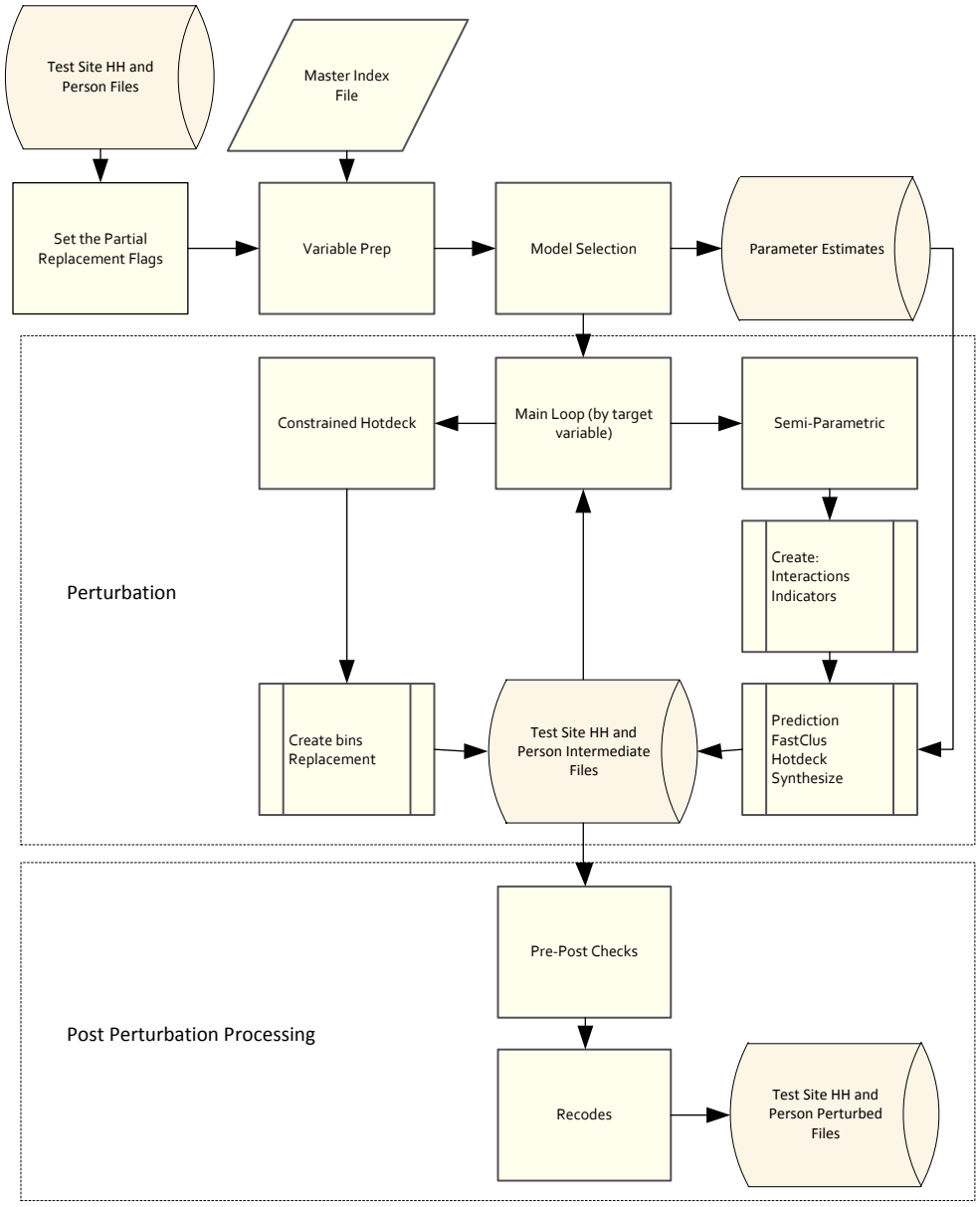


Figure 3-2. Validation Phase: Data Replacement Step Processing Flow

The initial steps before processing the approaches involved subsetting to the two test sites, 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), which is discussed further below.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Initial Steps

Test Sites. This is a very brief program that subsetted the national five-year ACS file to counties related to the Atlanta and Olympia test sites. The research team, assisted by subcontractor VHB, has involved transportation planners in the identification of test sites. There were four test sites used in the evaluation during the development phase and two test sites (Olympia, Atlanta) in the validation phase. The boundaries at the county level were identified, including the Federal Information Processing Standards (FIPS) code, for each of the test sites.

Partial Replacement Rates. For the selection of targeted records for data replacement, a stratified random sample was selected from risk strata that were formed using results from the initial risk analysis. Data values identified as high risk were replaced at a higher rate than other data values. The probability that a data value was selected for replacement was correlated with the number of data values identified as violations in the record with which the data value was associated. The research team consulted with the Census DRB on the partial replacement rates and agreed on the set of rates for this phase of the research. Risk strata were identified for each variable to be perturbed and the rates were used to select and flag a sample of data values for replacement for each of the test sites using the following flag:

> **VarName_PARTIAL**. This flag was set to one for a CTPP variable (referred to generically as *VarName*) if *VarName*_FULL was set to one and the associated data value was selected by a random process.

Variable Prep. The Variable Prep step was processed in order to prepare recodes and prepare variables as predictors for the semi-parametric approach; that is, after the partial replacement flags were set, predictor variables were recoded as necessary for the model selection step. The Variable Prep step also compiled the pool of predictor variables, as well as created and compiled indicator variables and interaction terms as predictor variables. The predictor pool was created from ACS and Census variables, including indicator variables for unordered categorical (UC) variables and select interaction terms.

A master index file (MIF) drove the process and identified the variables to be perturbed for the validation phase 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 to be added to the pool of candidate predictor variables were identified as well.

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. 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.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Model Selection and Model Areas. Once the variable prep processing was completed, then the model selection approach was processed for all variables identified in the MIF that underwent the semi-parametric approach. 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 perturbation step (explained later).

The model selection process occurs for each model area. At the person level, it was necessary to conduct the model selection separately for persons residing in group quarters (GQs) and for persons in households since the predictor list for GQs was limited to predictors not associated with households. Therefore, for the validation phase, for person-level processing, the model areas for most target variables were the same as the residence test sites, separating GQ from non-GQ records. For household-level processing, the whole test site served as the model-area.

For person-level processing during the production run for most target variables, for non-GQ records the model areas will be residence-based Core-Based Statistical Areas (CBSAs), and the rest of the state for non-CBSAs. For GQ records, the model areas will be residence-based Core-Based Statistical Areas (CBSAs), and the rest of the census division for non-CBSA. For industry, the model areas will be workplace based instead of residence based. For household-level processing, records in CBSAs will be modeled separately by CBSA, while records not in CBSAs will be modeled together in the remainder of the state.

Perturbation. One by one, the target variables were processed through the Main Loop. Either the constrained hot deck or the semi-parametric approach was processed, depending on the variable type of the target variable. First, household-level variables were perturbed, then the perturbed household variables were transferred to the person level, where the process continues with the perturbations on person-level variables.

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

To facilitate the discussion of the perturbation approach that follows, a subset of the preliminary variables to be perturbed were identified (Table 3-1). The table highlights the level (HH, person), the variable type (OC, UC), the number of categories, and the approach used. A couple of "spinoff" approaches other than the semi-parametric (SP) and constrained hot deck (CH), namely additive noise (AN) and rank linking (RL), were implemented in a limited way (described below) to add to the protection from disclosure.

Table 3-1. Validation Phase: Subset of Preliminary Variables that were Perturbed

Item	Variable Name	Variable Level	Variable Type	Number of Categories	Perturbation Approach
1	HH Income	HH	Ordinal categorical (OC)	Continuous	CH, AN
2	Number of Workers	HH	OC	5	CH
3	Children under 18	HH	OC	2	SP
4	Travel Time	Person	OC	Continuous	CH
5	Time Leaving Home	Person	OC	Continuous	CH
6	Age	Person	OC	7	CH
7	Poverty status	Person	OC	3	RL
8	Minority status	Person	OC	2	SP
9	Industry	Person	UC	7	SP

NOTE: CH = constrained hot deck, SP = semi-parametric, AN = additive noise, RL = rank linking

3.1.2.1 Semi-Parametric

Some background on the semi-parametric approach is given in Section 2.1.3.1. The semi-parametric procedure was applied to the household-level variable presence of children under 18, to the person-level variables minority status, and to industry. There were two main steps involved in the validation process:

1. Model selection and estimation; and
2. Sequential prediction and perturbation.

Each step is explained in detail below.

Model Selection and Estimation

The model selection and estimation step was done once for each CTPP variable to be perturbed 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 perturbed data since the joint distribution among the variables was already given. This joint distribution was determined by the ACS data including ACS records that had fully complete, imputed, and swapped data.

The modeling step was done separately at the household level and at the person level for each variable identified for the semi-parametric approach in Table 3-1. The modeling was done for variables of type UC (e.g., industry) and OC binary variables (e.g., presence of children, minority status). 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 as hot deck cells.

Aligning with the list of variables in Table 3-1, let y_{ki} denote the k^{th} variable to be perturbed for record *i*, where *k* is the variable number in Table 3-1, and *y* represents the American Community Survey (ACS) data values. The subscript *j* identifies indicator variables associated with UC variables (e.g., industry). The bolding pattern represents vectors. Therefore the model selection for OC binary (variables 3 and 8 in Table 3-1) and UC variables (variable 9 in Table 3-1) is essentially as follows:

$$E(y_3|y_1, y_2, \mathbf{X}) = \text{f}(y_1, y_2, \mathbf{X}, \boldsymbol{\beta}),$$
$$E(y_8|y_1, y_2, y_3, y_4, y_5, y_6, y_7, \boldsymbol{y_9}, \mathbf{X}) = \text{f}(y_1, y_2, y_3, y_4, y_5, y_6, y_7, \boldsymbol{y_9}, \mathbf{X}, \boldsymbol{\beta}),$$
$$E(y_{9j}|y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, \mathbf{X}) = \text{f}(y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8, \mathbf{X}, \boldsymbol{\beta}),$$
for j = 1,2,3,4,5,6,7 for industry categories

The modeling at the household-level for y_3 was limited to household-level predictors. As in the development phase, the models were processed to allow predictors to enter in 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. 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 3-2. Within the candidate predictor pools were select interactions with the MOT indicators. The MOT-variable interactions included interactions with household income, earnings, age, minority status, sex, number of workers in HH, vehicles available, country of birth, travel time, and poverty status. The list of candidate predictors is given in Appendix D. The processing for the model selection and model estimation was conducted within model areas.

Table 3-2. Validation Phase: FORCELIST Variables for Each Dependent Variable

Dependent variable	FORCELIST
HH status (under 18)	MOT householder indicators, household income, number of workers in the household
Minority status	MOT indicators, age, work-shift indicators, travel time, household income, poverty status, and vehicles available
Industry	MOT indicators, age, work-shift indicators, travel time, household income, minority status, poverty status, and vehicles available

Sequential Prediction and Perturbation for Semi-Parametric Processing

Once the model parameters were estimated for the UC or OC binary variables, which were designated for the semi-parametric approach, the perturbation process occurred for all target variables, one variable at a time. The process was done for all variables designated for perturbation, variable by variable, and ordered (OC) variables were perturbed (using the constrained hot deck) before the unordered (UC) or binary variables. Variables were perturbed, beginning with the household level, transferring the perturbed household variables to the person level, and then continuing with the perturbations on person-level variables. In turn, for any UC or binary variables, under the semi-parametric approach, they went down a path of sequential prediction and perturbations steps as described in this section. More discussion of the general sequential process for the semi-parametric approach is given in Section 2.1.3.1. Table 3-3 provides the values of the number of prediction groups (*g1*) and weight groups (*g2*) for each perturbed variable, as well as the locality.

Table 3-3. Validation Phase: Number of Prediction Groups and Weights Cell for Each Perturbed Variable

Dependent variable	Locality	Number of prediction groups (*g1*)	Number of weight cells (*g2*)
HH status (under 18)	Residence CTAZ300	7	3
Minority status	Residence CTAZ50	7	3
Industry	Workplace CTAZ300	7	3

NOTE: Residence CTAZ50 has a minimum of 50 persons who lived in the area. Workplace CTAZ300 will be used in the production run; in the validation run; however, residence CTAZ300 was used for industry due to many sparse workplaces outside the test site area.

Within each hot deck cell, a without replacement draw from the empirical distribution was conducted. The predictions and the subsequent draws from an empirical distribution occurred in a sequential manner so that perturbed values for the predictor variables were used in the model for the next variable to be perturbed.

The sequential prediction and perturbation steps are described using the variables in Table 3-1 as follows. Given that the first two variables in Table 3-1 were processed using the constrained hot deck, the third variable in Table 3-1 was an OC binary variable and was being processed by the semi-parametric approach. The prediction equation for OC variable 3 (y_3) in Table 3-1 is given as follows (ignoring interaction terms for simplicity), using the perturbed values for variables 1 and 2, and the ACS values for the remaining items:

$$\hat{y}_{3i} = \beta_0 + \beta_1 \tilde{y}_{1i} + \beta_2 \tilde{y}_{2i} + \sum_{l=1}^{L} \beta_l x_{li}$$

where L is the number of other predictor variables.

Then subsequently, as discussed above, within locality, *g1* prediction groups were formed on $\hat{y}_{3i}$ and *g2* groups were formed on the weights, within each of the *g1* groups. Let $\tilde{y}_{3i}$ represent the perturbed

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

value drawn at random without replacement within the hot deck cell formed by the *g1*g2* groups within locality.

Under partial replacement, the values were perturbed at this time only if flagged for replacement; that is, high risk values were targeted as identified in the initial risk analysis. After each variable was perturbed, the interaction terms were recreated using perturbed values so perturbed values could be used in the prediction equation for the next dependent variables in the sequence.

Continuing sequentially through the variables in Table 3-1, variable 8 was another OC binary variable and was handled in a similar fashion as variable 3. Let the prediction equation for variable 8 be represented as follows, using the perturbed values for the previous 7 variables:

$$\hat{y}_{8i} = \beta_0 + \beta_1 \tilde{y}_{1i} + \beta_2 \tilde{y}_{2i} + \sum_{k=3}^{7} \beta_k \tilde{y}_{ki} + \sum_{j=1}^{7} \beta_{9j} y_{9ji} + \sum_{l=1}^{L} \beta_l x_{li}$$

Next, for the 9th and last variable, there were seven categories in this UC variable from which seven corresponding indicator variables were formed. Let the prediction equation for the j^{th} category of UC variable 9 be represented as follows, using the perturbed values for the previous 8 variables:

$$\hat{y}_{9ji} = \beta_0 + \sum_{k=1}^{8} \beta_k \tilde{y}_{ki} + \sum_{l=1}^{L} \beta_l x_{li}$$

where j = 1, 2, … 7.

For the UC variable, a clustering program (SAS Proc FastClus) was used to form *g1* clusters (prediction groups), using the 7 sets of predicted values $\hat{y}_{9ji}$. Then, *g2* groups were formed on the weights, within each of the *g1* groups. Let $\tilde{y}_{9i}$ represent the perturbed value drawn within the hot deck cell formed by the *g1*g2* groups within locality. In general, after a UC variable was perturbed, indicator variables were re-created using the perturbed values.

The process ran sequentially until all items to be perturbed were processed. One cycle through the variables was conducted.

3.1.2.2 Constrained Hot deck

For clarity, we reiterate the overview of the perturbation process. Once the model parameters were estimated for the UC or OC binary variables (designated for the semi-parametric approach), the perturbation process occurred for all target variables, one variable at a time. The process was done for all variables designated for perturbation, variable by variable, and ordered variables were perturbed (using the constrained hot deck) before the unordered or binary ones. Variables were perturbed, beginning with the household level variables, transferring the perturbed household variables to the person level, and then continuing with the perturbations on person-level variables.

In turn, as given in Table 3-1, for any OC variables (non-binary), they followed the perturbation steps as described in this section. In a limited fashion, two other “spinoff” approaches were used. Rank linking (described below) was designed to closely align the household income with person-level poverty status. Additive noise (described below) was used to provide further protection to household income.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

The constrained hot deck approach was used to perturb values of the following ordinal variables under partial replacement: household income, number of workers in the household, travel time, time leaving home, age, and poverty status. Details are given for the constrained hot deck approach in Section 2.1.3.3.

Under partial replacement, 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 following describes the process for each variable. Along the way, additive noise for household income and rank linking for poverty status also are described.

Household Income. The perturbation process began with the constrained hot deck approach applied for replacing values of household income. Among all records where household income was targeted for replacement, the hot deck cells were formed by PUMA* vehicles available* household income bins * three weight groups. For each value of household income needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value.

Additive Noise for Household Income. Next, the additive noise procedure was conducted on any record where household income did not change value; that is, during the perturbation step, if left unchanged from the constrained hot deck procedure, noise was added to the original household income value y as follows:

$$\tilde{y}_{3i} = y_{3i}(1 + fz),$$

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_{3i}, which means the level of noise was allowed to vary relative to the magnitude of household income.

Number of Workers in the Household. Among all records where the number of workers in the household was targeted for replacement, the hot deck cells were formed by PUMA * vehicles available* number of workers in household bins * three weight groups. For each value of the number of workers in the household needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value.

Travel Time and Time Leaving Home. Travel time and time leaving home were linked in the process as follows. Among all records where travel time was targeted for replacement, the hot deck cells for travel time were formed by PUMA * MOT * time leaving home bins * travel time bins* two weight groups. For each value of travel time or time leaving home needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value for travel time. In most cases both variables were flagged together for replacement, and the same record was the source for replacing both variables.

Age. While continuous age is not involved in the CTPP tables, it was useful for forming the bins for the categorized CTPP age variable. The hot deck cells were formed by state* PUMA* MOT* continuous age bins* three weight groups. For each value of categorized age needing replacement, a random draw without replacement was conducted within the hot deck cell for the target data value.

Rank Linking for Person-Level Poverty Status. A variation of the constrained hot deck (we refer to internally as rank linking) was developed to link household income and the person-level poverty

status together. Once the household income was perturbed as described above, the ACS and the perturbed household income were merged onto the person level file.

To perturb poverty status, we created a file called RAW with state, Public Use Microdata Area (PUMA), the number of workers in the household, vehicles available, ACS income and ACS poverty status. We sorted RAW by a missing value indicator on poverty status, state, PUMA, number of workers in the household, vehicles available, ACS income. The perturbed household income resides on main data file. The perturbed file is then sorted by a missing value indicator for poverty status, state, PUMA, number of workers in the household, vehicles available, and perturbed income. Then the ACS poverty status from the RAW file was joined (merged) with the main data file. The ACS poverty status was replaced if flagged for replacement.

3.1.3 Weight Calibration

After the approaches were processed in the validation phase, the sample-based raking adjustment step (discussed in Section 2.1.4), was done so that the weights are calibrated to reproduce select ACS estimates at the PUMA level, which are areas formed to be greater than 100,000 in population for the purpose of releasing public use microdata. In addition, a dimension was added to calibrate to the estimated total number of workers at the CTAZ300 level, which are areas of about 8,000 in population.

For the validation phase, at the household level, the process continued until, before each dimension's last adjustment, the sums of adjusted weights were within 10 of each control total for the last dimension for the full sample, and 100 for each replicate weight. For the person level, due to lack of convergence on one dimension, the thresholds were adjusted to be 15 for Atlanta and 10 for Olympia for the full sample.

The raking was done at the household-level to adjust household weights and at the person level to adjust the person weights. The dimensions for the household raking are given in Table 3-4 and the dimensions for person raking are given in Table 3-5. Due to the numerous place of work PUMAs (PUMAs where the ACS respondent works) with low sample counts for the test sites due to commutes outside the test site area defined by place of residence, it was decided to not process the dimensions involving place of work PUMAs for the validation phase.

Table 3-4. Validation Phase: Raking Dimensions for the Household File

Dimension	ByVar1	ByVar2
1	PUMA	Vehicles available (6)
2	PUMA	Number of workers in HH (6)
3	PUMA	HH income (5)
4	Residence CTAZ300	--

Table 3-5. Validation Phase: Raking Dimensions for the Person File

Dimension	ByVar1	ByVar2
1	PUMA	Vehicles available (6)
2	PUMA	Number of workers in HH (6)
3	PUMA	HH income (5)
4	Place of work PUMA	HH income (5)
5	PUMA	Travel time (4)
6	PUMA	MOT(6)
7	Place of work PUMA	MOT(6)
8	Residence CTAZ300	--

NOTE: Dimensions 4 and 7 were not incorporated in the validation phase due to sparse place of work PUMAs for the test sites. Workplace outside of the country is combined as a PUMA.

Tables 3-6 and 3-7 provide percentiles from the validation phase of the raking adjustment factors for the person-level raking and household raking, respectively, under partial replacement. The development phase results are shown as well for the constrained hot deck. Focusing on the range between the 10th and 90th percentiles, the range has been reduced for the constrained hot deck approach between the development phase and the validation phase. This is the result of changing the approach so that the empirical draws are limited to records that are flagged for replacement. The ranges shown in Table 3-7 for household raking are generally smaller than for person raking due to having few dimensions.

Table 3-6. Validation Phase: Percentiles of the Raking Factors: Person Level

Phase	Test site	Minimum	10th	25th	50th	75th	90th	Maximum
Development	MAD	0.91	0.97	0.99	1.00	1.02	1.05	1.08
	STL	0.71	0.93	0.96	1.00	1.04	1.08	1.24
	ATL	0.81	0.96	0.98	1.00	1.02	1.04	1.19
	IA	0.66	0.96	0.98	1.00	1.02	1.05	1.31
Validation	ATL	0.87	0.98	0.99	1.00	1.01	1.02	1.15
	OLY	0.89	0.96	0.99	1.00	1.02	1.03	1.09

NOTE: Process run #1, partial replacement amount

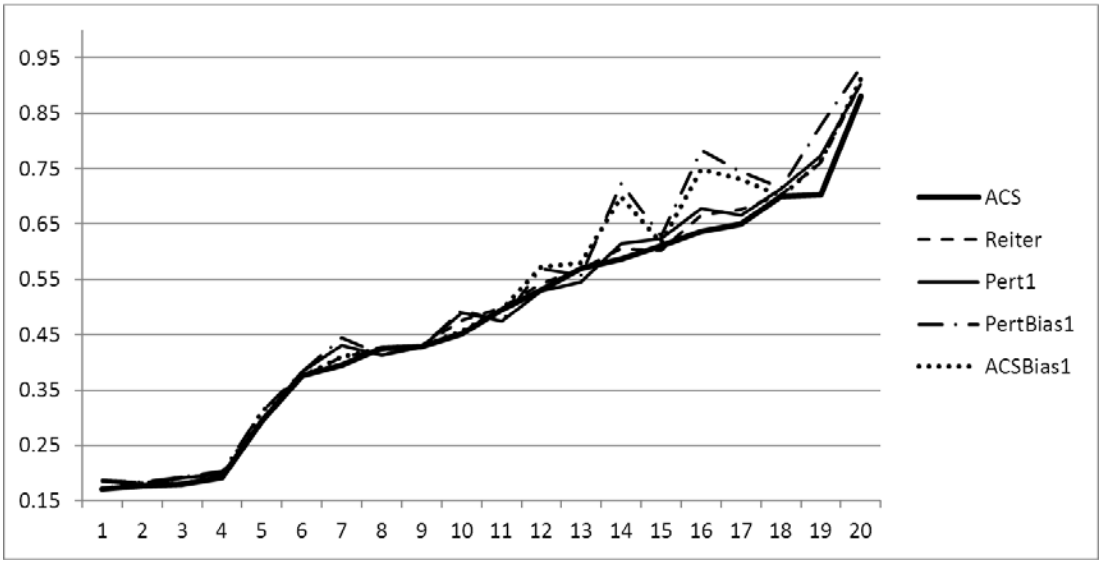
Table 3-7. Validation Phase: Percentiles of the Raking Factors: Household Level

Approach	Test site	Minimum	10th	25th	50th	75th	90th	Maximum
Development	MAD	0.98	0.99	0.99	1.00	1.01	1.01	1.03
	STL	0.89	0.97	0.99	1.00	1.01	1.03	1.14
	ATL	0.86	0.97	0.99	1.00	1.01	1.02	1.11
	IA	0.87	0.97	0.99	1.00	1.01	1.02	1.18
Validation	ATL	0.96	0.99	1.00	1.00	1.00	1.01	1.05
	OLY	0.98	0.99	0.99	1.00	1.00	1.02	1.03

NOTE: Process run #1, partial replacement amount.

3.1.4 Variance Estimation

Section 2.1.5 provides a description of the variance estimation approaches considered for the CTPP Set B tables. Figure 3-3 shows the estimated standard errors of the county-level mean travel time for workers who drove alone using the 2005–2009 ACS sample. The computations were based on the original ACS dataset and the perturbed dataset for the test site Atlanta. The horizontal axis represents the 20 counties in Atlanta. The standard errors computed from formulas (f1), (f2), and (f4) are very similar, and generally smaller than the standard errors computed from (f3) and (f5). The standard errors from the perturbed data (f2) are not much different from the ACS estimate (f1) because the variation in the point estimates based on the perturbed datasets from the five independent runs is very small. The estimated standard errors computed from (f3) and (f5) account for the difference in the point estimates from the original and the perturbed data. This second term in (f3) and (f5) is moderate for some of the counties, but small or close to zero for others. This is partly because post-perturbation raking was done at the PUMA level. Although travel time is one of the raking dimensions (done at the PUMA level), the county-level estimates based on the perturbed data are not fully aligned with the estimates based on the original ACS data.



NOTE: Source data are ATL ACS dataset and perturbed datasets using semi-parametric approach and full replacement. ACS is the SE that uses (f1) and ACS data; Reiter is the SE that uses (f2) and perturbed data from 5 runs; PertBias1 is the SE that uses (f3) and perturbed data from run #1; Pert1 is the SE that uses (f4) and perturbed data from run #1; ACSBias1 is the SE that uses (f5) and perturbed data from run #1; Counties are on the horizontal axis

Figure 3-3. Validation Phase: Estimated Standard Errors of the County-Level Mean Travel Time (in minutes) for Workers Who Drove Alone: ACS 2005-2009

Given the above results and the simulation results in Section 3.2.1, the estimator recommended by the research team and approved by the Census Bureau is to add the squared difference to the usual ACS estimate. The squared term serves for the purpose of measuring the additional variance due to perturbation. Assuming perturbation is independent of the sampling process, formula (f5) is essentially the sum of sampling variance and perturbation variance, restated here:

$$\mathrm{var}(\tilde{\theta}_0) = \mathrm{var}(\hat{\theta}_0) + (\tilde{\theta}_0 - \hat{\theta}_0)^2, \qquad \text{(f5)}$$

where $\tilde{\theta}_0$ represents the CTPP perturbed estimate of θ. Assuming that the noise introduced to the perturbed 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

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

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

3.2 IMPACT ON DATA UTILITY AND DISCLOSURE RISK

Section 2.2 provides a description of the risk and utility measures that were developed to assess the impact of the data perturbation approaches under consideration. The impact on disclosure risk from the development phase processing was determined to be at an acceptable level by the Census Bureau DRB, and the impact on data utility was determined to be at an acceptable level by the Transportation Review Board Panel. This section includes results from the validation phase, and where appropriate, the validation results are compared with the development phase results. The data utility measures are discussed in Section 3.2.1. and the disclosure risk measures in Section 3.2.2. Each section presents the results from the validation phase evaluation.

3.2.1 Impact on Data Utility

As in the development phase, there were two main components to the data utility checks. The focus of the first set of checks was to compare the ACS data with the perturbed ACS data. The second set of checks was to evaluate the ACS and perturbed CTPP data with travel model outputs from the two validation test sites. Section 2.2.1 discussed the CTPP and ACS data elements that affect the formulation of data utility measures, presented a description of the data utility measures that were used to compare the CTPP perturbation approaches, and described how the resulting data were used to compare home-based work (HBW) model outputs with the ACS data and the perturbed CTPP data from the three-year (2006–2008) ACS release.

For the validation phase, the comparisons checked cell means, 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 perturbed estimates (across estimates for geographic areas) were computed where appropriate in order to give indications of potential bias introduced by the perturbation. The interquartile range for the differences provided an indication of the variation caused by the perturbations. Scatter plots were used throughout to visually depict the impact of the perturbation approaches. Lastly, new for the validation phase, was a check on the differences for medians and 75th percentiles of travel time for table cells across estimates for geographic areas.

In general, results are shown for the development phase (for Madison and Atlanta) based on three-year ACS data and for the validation phase (Olympia and Atlanta) based on five-year ACS data. The results for Madison for the development phase were chosen since Madison was closest in size to Olympia and they each were based on a single county. That being said, Olympia was only about half the population size as Madison.

Because the validation phase was based on more tables, the disclosure risks were a bit higher than in the development phase. In addition, the DRB requested that the partial replacement rates be modified, and the resulting rates were a bit higher than in the development phase.

For the validation phase, there were two test sites (Atlanta, Olympia), one approach (described in Section 3.1), one amount (partial replacement), and five perturbation runs.

Cell Mean Differences and Quantile Differences Results

The computations for the cell mean differences are provided in Section 2.2.1.2. The cell quantile differences were computed in an analogous way. Tables M-1 and M-2 in Appendix M provide the median and interquartile range (IQR) of differences for travel time and household income between cell means

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

from ACS data and cell means from perturbed data generated by the constrained hot deck. The two tables differ by geographic level of the tabulations, where Table M-1 shows TAZ level results and Table M-2 shows county level results. In each table, results are shown for the development phase (for Madison and Atlanta) and for the validation phase (Olympia and Atlanta) under partial replacement.

Even though the perturbation rates were higher in the validation phase, the tables showed that the impact on cell mean differences has been reduced, likely due to improvements made to the constrained hot deck after the development phase. The constrained hot deck was modified so that for partial replacement, 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. For example, in Table M-1, the IQR associated with cell mean differences across TAZs for household income in Atlanta for vehicles available category 2 was 1,734 in the development phase, and is 637 in the validation phase. As another example, in Table M-2, the IQR cell mean difference across counties for travel time in Atlanta for MOT category 2 (drive alone) was 0.3 in the development phase, and is 0.1 in the validation phase. Likewise, for all other estimated cell mean differences, the IQRs for the validation phase were less than the IQRs for the development phase. The median cell mean differences tend to indicate areas where there may have been potential for bias. As given in the table most of the median cell mean differences were zero or close to zero, indicating very low potential for bias. Even for county flows on mean travel time as shown in Table M-2, the median of the cell mean differences were equal to zero.

Bubble plots, provided in Figures N-1 through N-4 in Appendix N, were generated at the county level and TAZ level to compare mean travel time from ACS data (x-axis) and from perturbed data (y-axis) under partial replacement. While the results in Tables M-1 through M-2 are from all localities no matter the ACS sample size, the dots in the bubble plots in Figures N-1 through N-4 are shown only for localities with 30 or more ACS sample cases. In each figure, results are shown for the development phase (for Madison and Atlanta) and for the validation phase (Olympia and Atlanta) under partial replacement. The county-level plots in N-1 through N-2 show no apparent impact, which is consistent with the development phase plot (top plot in figure). The TAZ-level plots in O-3 and O-4 showed some minimal deviations, although the deviations were slight as determined by the tightness to the 45 degree line and consistent with the development phase plot (top plot in figure).

Bubble plots were also generated for TAZ flows for mean travel time as shown in Figures N-5 and N-6. Each dot shown in Figures N-5 through N-6 has at least 10 ACS respondents. The size of the bubble is related to the ACS sample size in the locality. In each figure, results are shown for the development phase (for Madison and Atlanta) and for the validation phase (Olympia and Atlanta) under partial replacement. By comparing the top (development phase) and bottom (validation phase) plots for the TAZ flows in N-5 and N-6, it clearly shows the development phase results being validated for the constrained hot deck approach. Included is an additional middle plot of Figures N-5 and N-6 for county flows for the validation phase. The county flow plots showed favorable results.

The bubble plots paid attention to the preservation of utility when there were enough data so that the original utility was at least moderate. The distribution of the TAZ sizes is given in Table 3-8. The table shows that the proportion of TAZs with 30 or more ACS cases (and therefore included in the bubble plots) is 19 percent and 62 percent for Madison and Atlanta, respectively, during the development phase on the three-year ACS, and 42 percent and 76 percent for Olympia and Atlanta, respectively, for the validation phase on the five-year ACS.

Table 3-8. Validation Phase: Percentage of TAZs by ACS Sample Size Categories

	Development Phase – 3-year ACS		Validation Phase – 5-year ACS	
TAZ size	**MAD (%)**	**ATL (%)**	**OLY (%)**	**ATL (%)**
[1,5]	25	6	15	5
[6,10]	20	7	12	3
[11,20]	24	12	16	9
[21,30]	11	12	16	6
[31,50]	12	21	18	15
[51,100]	7	32	17	32
≥100	0	9	7	29

NOTE: Individual column percentages do not sum to 100% due to TAZ with zero workers and rounding.

Table 3-9 shows the median, IQR, minimum, and maximum values of the absolute relative differences for mean travel time at the TAZ level by mean travel time. TAZs with ACS mean travel time less than 5 were excluded. The data are for Atlanta, for the first of the five process runs.

Table 3-9. Validation Phase: Distribution of Absolute Relative Differences for Mean Travel Time at the TAZ Level by Mean Travel Time, Five-Year ACS

ACS TAZ Mean: Travel Time (minutes)	Median (%)	IQR (%)	Min (%)	Max (%)
[5, 15)	9	17	0	200
[15, 20)	3	5	0	45
[20, 29)	2	3	0	37
[30, 45)	2	3	0	31
[45, 60)	3	4	0	36
[60, 75)	20	30	3	33

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

Results of the simulation provide little indication of the potential for bias. As shown in Table 3-10, just two of the 22 CTAZ300 areas, namely numbers 8 and 16, showed any indication of potential bias. When conducting statistical tests for multiple comparisons, these indications were not significant.

Table 3-10. Validation Phase: Preliminary Simulation Results on Tests for Potential Bias in Drive Alone Mean Travel Time in Olympia, by CTAZ300

CTAZ300	ACSmean	Mean of Bias	Variance of Bias
1	24	0.32	0.07
2	21	0.01	0.07
3	22	0.22	0.13
4	25	0.46	0.10
5	24	-0.21	0.11
6	26	0.27	0.21
7	30	0.29	0.15
8	38	-0.93	0.18
9	30	-0.20	0.13
10	28	-0.35	0.06
11	20	0.29	0.09
12	19	0.40	0.22
13	25	0.11	0.11
14	23	-0.22	0.05
15	22	0.21	0.06
16	16	0.75	0.09
17	17	0.53	0.13
18	17	-0.22	0.06
19	17	0.12	0.12
20	20	0.12	0.19
21	24	-0.33	0.13
22	22	-0.02	0.10

Bubble plots were also produced for household income and shown in Figures O-1 and O-2 at the county level and in Figures O-3 and O-4 at the TAZ level. As seen in the mean travel time plots, the impact of the perturbation approach was negligible at the county level. At the TAZ-level, the validation phase (bottom plot) clearly resembled the results from the development phase (top plot).

Table 3-11 shows the distribution of the absolute relative differences for mean household income at the TAZ level by mean household income. The table was generated for Atlanta TAZs for the first of the five process runs. TAZs with absolute value of the ACS mean income less than $5,000 were excluded. The results showed small deviations between raw and perturbed mean household incomes, as given by the median relative difference being less than five percent across each income category.

Table 3-11. Validation Phase: Distribution of Absolute Relative Differences for Mean Household Income at the TAZ Level by Mean Household Income, Five-Year ACS

ACS TAZ Mean: Household Income	Median (%)	IQR (%)	Min (%)	Max (%)
[$5,000, $15,000)	2	14	0	121
[$15,000, $25,000)	3	6	0	36
[$25,000, $35,000)	2	3	0	21
[$35,000, $50,000)	1	2	0	22
[$50,000, $75,000)	1	2	0	30
[$75,000, $100,000)	2	2	0	28
[$100,000, $150,000)	2	4	0	107
≥$150,000	4	7	0	100

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

The conclusion was that the validation phase results on the cell means analysis clearly supported the favorable results given by the constrained hot deck in the development phase; that is, the impact of the perturbation approach on cell means was at an acceptable level and there was little indication of bias introduced by the perturbation approach.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Weighted Cell Count Differences Results

Weighted cell counts were computed for the Set B threshold tables as discussed in Section 2.2.1.2. Figures P-1 through P-7 provide a visual comparison of the weighted cell count estimates before and after perturbation. The development phase plots are shown at the top, and the validation plots are shown at the bottom. The scales differ within the pair of plots since the scale was expanded for the validation phase.

The county-level plots in Figure P-1 compare Madison from the development phase and Olympia from the validation phase, and the county-level plots in Figure P-4 compare Atlanta's development phase (3-year ACS) with its validation phase (five-year ACS). The validation plots in each figure confirmed the results from the development phase, showing just slight deviations from the 45 degree line. These results implied that the perturbation approach had minimal impact on the county-level tables.

The corresponding TAZ-level plots are provided in Figures P-2 and P-5. The validation plots in each figure confirmed the results from the development phase, in that the same magnitude of the deviations, if not less, was seen from the validation phase plots.

The corresponding county flow plots are provided in Figures P-3 and P-6. The validation plots in each figure contained about the same amount of deviation from the 45-degree line as in the development phase plots; therefore, they confirmed the results from the development phase.

Figure P-7 is included to show results for only tables that include industry. Industry was challenging to perturb. Despite this challenge, as shown in the plots for county flow tables, the deviations from the 45-degree line are about at the same level as shown in Figures P-3 and P-6 (for all tables). This, therefore, supported the conclusion that the results from applying the semi-parametric approach to perturb industry were adequate.

Figure P-8 is included to show results for areas smaller than PUMAs for tables MOT11 * AHINC(26). Geographic areas were formed by combining TAZs so that they have at least 1,000 ACS sample cases, which corresponds to about 25,000 in population. The top plot is for Olympia and the bottom is for Atlanta. Each plot showed little deviation from the 45-degree line, and therefore indicated little impact from the perturbation approach on the resulting weighted cell counts.

The conclusion from the results on weighted cell counts was that the acceptable level of impact from the perturbation approach seen in the development phase was confirmed by the validation phase results.

Results of Measuring the Impact of Perturbation on Standard Errors

The approach to measuring the impact on standard errors is described in Section 2.2.1.2. The standard errors were computed at the county level for mean travel time and mean HH income using the first process run among the five runs. 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 using formula f3.

Table M-5 shows the comparison results for the development phase (Atlanta and Madison) on 2006-2008 ACS data and for the validation phase (Atlanta and Olympia) on 2005–2009 ACS data. Since Madison and Olympia consists of only one county, the IQR is equal to 0 for each comparison. For Olympia, the median differences showed negative values for some subgroups. This means that (f3) was

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

lower than the ACS variance, which was only possible if the first term in (f3), the so-called naïve variance, was less than the ACS variance.

For Atlanta, the IQRs of the difference were less for the validation phase when compared to the development phase, which was an indication that the difference between variances did not vary as much as in the development phase. This may be due to the larger sample size in the five-year ACS. The median difference is closest to zero for the validation phase as well.

A simulation study was conducted to evaluate the variance estimators. In the simulation, the perturbation approaches developed by the research team were applied to the data from the Atlanta test site independently for 1,000 times. From each of the 1,000 independent perturbed datasets, the mean travel time for workers who drove alone within each residence CTAZ was calculated; a CTAZ contained at least 300 workers living in the area. The variances were computed using three different estimators: naïve estimator (f4), naïve estimator with adjustment (f3), and ACS estimator with adjustment (f5).

Table 3-12 shows, within each CTAZ, the relative difference between the ACS and the average of the 1,000 perturbed estimates, as well as the ratios of the average standard errors from (f5) and (f3) to the standard error from the usual ACS estimator. The perturbation noise is generally small (e.g., two percent of the ACS estimates), reaching three percent in CTAZ 17 and 5 percent in CTAZ 16. A majority of the perturbed standard errors from (f5) are 2–9 percent higher than those from the usual ACS estimator. In CTAZ 16, the standard error from (f5) is 28 percent higher. The perturbed standard errors from (f3) are similar to those from (f5) but they can sometimes be lower than the ACS estimated standard errors. It was concluded that data perturbation process only adds a small amount of noise to the ACS data for large CTAZs which makes the perturbed estimates deviate somewhat from the original estimates. The impact in small CTAZs or TAZs was not evaluated because confidentiality concerns mandated that these be substantially perturbed. The research team just wanted to be sure that they were not substantially perturbing estimates that did not need to be substantially perturbed.

Table 3-12. Validation Phase: Relative Difference Between ACS and Perturbed Estimates, and Ratios of Standard Errors from (f5) and (f3) to that from Usual ACS Estimator, by CTAZ.

CTAZ	Relative Deviation of Perturbed Estimate from ACS Estimate: Abs(Pert_Est-ACS_Est)/ACS_Est	Relative f5 Standard Error to ACS Standard Error: (f5)_SE/(f1)_SE	Relative f3 Standard Error to ACS Standard Error: (f3)_SE/(f1)_SE
1	0.01	1.06	1.03
2	0.00	1.03	0.94
3	0.01	1.05	1.00
4	0.02	1.08	1.08
5	0.01	1.03	0.93
6	0.01	1.06	1.09
7	0.01	1.04	1.04
8	0.02	1.10	1.09
9	0.01	1.04	1.09
10	0.01	1.06	1.07
11	0.01	1.08	1.09
12	0.02	1.10	1.18
13	0.00	1.02	1.08
14	0.01	1.04	1.07
15	0.01	1.06	1.21
16	0.05	1.28	1.30
17	0.03	1.09	1.12
18	0.01	1.09	1.18
19	0.01	1.03	0.88
20	0.01	1.04	1.16
21	0.01	1.06	1.04
22	0.00	1.05	1.08

Coverage rates summarize how well the constructed confidence intervals can cover the true values through independently repeated sampling and perturbation processes. However, the true values were not available in this study, and Atlanta ACS data was just one sample. Therefore, instead of drawing repeated ACS samples, the research team drew the simulated true values (mean travel time) for individual CTAZ from a normal distribution with ACS point estimate as the mean and ACS variance estimate as the variance, assuming ACS point and variance estimates for each CTAZ were approximately unbiased. The team computed, on average, how likely the confidence intervals based on the perturbed estimates contained the randomly drawn true values. The results are presented in Table 3-13. The coverage rates were very close to the nominal 95 percent when (f5) was used to estimate the variance. The performance of the confidence intervals based on (f3) was also good, but slightly less stable than those based on (f5). The coverage for the naïve estimator (f4) was always lower than the coverage based on the naïve estimator with adjustment (f3). The coverage rates from (f4) were acceptable for some CTAZs, but could be lower than the nominal rates for other CTAZs. For example, in CTAZs 16 and 19, the coverage rates even fell below 90 percent for (f4). This clearly showed that the naïve estimator (f4) did not capture the variance due to perturbation appropriately.

Table 3-13. Validation Phase: Coverage Rates of Confidence Intervals Based on Three Variance Estimators and 95% Confidence Interval Formulae.

CTAZ	Naïve estimator with adjustment: (f3)	Naïve estimator: (f4)	ACS estimator with adjustment: (f5)
1	0.95	0.94	0.96
2	0.91	0.90	0.94
3	0.94	0.93	0.96
4	0.95	0.94	0.95
5	0.92	0.91	0.95
6	0.96	0.95	0.96
7	0.95	0.94	0.95
8	0.94	0.90	0.94
9	0.97	0.96	0.96
10	0.96	0.95	0.96
11	0.95	0.93	0.95
12	0.97	0.96	0.96
13	0.97	0.97	0.95
14	0.95	0.94	0.94
15	0.98	0.98	0.96
16	0.97	0.87	0.96
17	0.93	0.92	0.93
18	0.95	0.93	0.94
19	0.90	0.88	0.95
20	0.97	0.96	0.95
21	0.93	0.92	0.94
22	0.97	0.96	0.96

Table 3-14 shows the 50th percentile (median) of the ratios of the standard errors for the weighted cell counts in the validation phase for the Atlanta test site. In the ratios, the numerators are the standard errors of the perturbed estimates and calculated based on formula (f5), and the denominators are the standard errors of the ACS estimates and calculated based on formula (f1). Weighted cell counts were computed for the tables formed by geographic level (GeoLevel) and table variables (ByVar). The ratios were computed only if both the ACS cell sizes and the perturbed cell sizes were at least three. The medians of the ratios are presented in the table when the ACS cell size was small ($3 <= n < 30$), medium ($30 <= n < 100$), or large ($n >= 100$). The median relative increase to the standard errors was generally less than 10 percent.

Table 3-14. Validation Phase: Median of Ratios of Standard Errors for Weighted Cell Counts

GeoLevel	ByVar	ACS Cell Size (n)	Median
Residence TAZ	MEANS11 HH_INC26	3<=n<30	1.03
		30<=n<100	1.04
		n>=100	1.01
Residence TAZ	MEANS7 HH_INC5	3<=n<30	1.01
		30<=n<100	1.01
		n>=100	1.01
Residence TAZ	MEANS11 TRAVEL_TM12	3<=n<30	1.06
		30<=n<100	1.07
		n>=100	1.01
Residence TAZ	AGEWRK8	3<=n<30	1.04
		30<=n<100	1.09
		n>=100	1.03
Workplace TAZ	MEANS11 TRAVEL_TM12	3<=n<30	1.06
		30<=n<100	1.08
		n>=100	1.03
Workplace TAZ	AGEWRK8	3<=n<30	1.05
		30<=n<100	1.1
		n>=100	1.05
Residence County	MEANS11 HH_INC26	3<=n<30	1.05
		30<=n<100	1.05
		n>=100	1.04
Residence County	MEANS7 HH_INC5	3<=n<30	1.01
		30<=n<100	1.03
		n>=100	1.02
Residence CTAZ	MEANS11 HH_INC26	3<=n<30	1.04
		30<=n<100	1.07
		n>=100	1.03
County flow	MEANS7 HH_INC5	3<=n<30	1.01
		30<=n<100	1.02
		n>=100	1.02
County flow	INDUSTRY8	3<=n<30	1.04
		30<=n<100	1.09
		n>=100	1.05

In conclusion, even though the perturbation process did not make the perturbed estimates so much different from the ACS estimates, an appropriate variance estimator was still needed to account for the variance resulted from perturbation. The simulation results supported the use of the formula (f5) as the variance estimator for the CTPP tabulations.

Cramer's V Ratios Results

The computation of the Cramer's *V* statistic in described in Section 2.2.1.2. Table M-6 provides the Cramer's *V* results for the development phase (Atlanta and Madison) on 2006–2008 ACS data and for the validation phase (Atlanta and Olympia) on 2005–2009 ACS data. To summarize, as seen in the development phase for these test sites, all median Cramer's *V* differences were equal to zero for the validation phase. While most IQRs were at the same magnitude or less as in the development phase, for AGE9 and Atlanta county flows, the magnitude of the IQR in the validation phase (0.06) was higher than the development phase (0.01). A similar result was seen at the TAZ level. This means that there was more variation in the Cramer's *V* differences for tables involving MOT11 and AGE9 than was seen during the development phase.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Pairwise Associations Results

For the development phase, Pearson product correlations were computed and shown in Tables M-7 for Atlanta counties between six select pairs of the following variables at the individual level: HH income, age, poverty status, time leaving home, derived distance, and travel time.

For the validation phase, the correlations were computed for Atlanta and Olympia counties for eight pairs of the following variables: HH income, age, poverty status, time leaving home, travel time, and number of workers in the household.

The scatter plots for the development phase, shown in the top plots in Figures Q-1 (Madison) and Q-2 (Atlanta) were generated for 11 select pairwise correlations computed for each PUMA. The 11 pairs were the following:

- Travel time with each of the following: time leaving home, HH income, derived flow distance, poverty status, and age;
- Time leaving home with: HH income, poverty status, and age;
- HH income with: age, and poverty status; and
- Poverty status with age.

The scatter plots for the validation phase, shown in the bottom plots in Figures Q-1 (Olympia) and Q-2 (Atlanta) were generated for 11 select pairwise correlations computed for each PUMA. The 11 pairs were the following:

- Travel time with each of the following: time leaving home and age;
- Time leaving home with age;
- HH income with each of the following: age, poverty status, number of workers in HH, number of vehicles in HH;
- Poverty status with each of the following age, number of vehicles in HH; and
- Number of workers in HH with number of vehicles in HH, and with age.

The correlations in Table M-7 are provided for each of the process runs in order to observe the variation in the results as the process is repeated. The correlations in the perturbed data tended to be retained in general. The exceptions were correlations that involved age. For example, the actual correlation between age and poverty status for Atlanta is .1312, while the perturbed correlations range from .1107 to .1150 across the five perturbation runs. Likewise for Olympia, the actual correlation is .1700, while the perturbed correlations range from .1314 to .1540. Certainly, the typical impact of perturbation is the attenuation of correlations, although, these results were improved for the national testing and production run by adding other variables to the hot deck cells formed for the constrained hot deck, as discussed in Chapter 4.

Figures Q-1 (Madison/Olympia) and Q-2 (Atlanta) provide scatter plots of correlations for each PUMA for 11 select pairs of variables. The x-axis reflects the ACS correlations, and the y-axis reflects the correlations from perturbed data. **In general, the correlations in the raw data were retained in the perturbed data, as seen by the tightness of the dots along the 45 degree line. However, the**

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

validation phase plots (bottom of each figure), showed a slight attenuation that occurs where the dots flow away from the 45-degree line. These were correlations related to age, and as discussed above, there was a remedy for this so the retention of correlations could be improved.

Multivariate Associations Results

The approach to measuring the impact on multivariate associations is explained in Section 2.2.1.2. Table M-8 shows the U statistic for each run for a model that includes main effects only. For the development phase, results from both full and partial replacement were shown for each of the four test sites. This was because the amount of perturbation was somewhat more than the partial replacement amount in the development phase, and gave two points of comparison for the validation phase results. **In general, for Atlanta, the U statistic for the validation phase fell in between the values of the full and partial replacement amounts from the development phase —closer to the partial replacement results.** With improvements to the perturbation age, the results will get closer to the development phase partial replacement result.

Summary of Results from the First Set of Checks

The conclusions from the first set of checks were as follows:

- The cell means, medians and 75th percentile analysis clearly supported the favorable results given by the constrained hot deck in the development phase; that is, the impact of the perturbation approach was at an acceptable level, and there was little indication of bias introduced by the perturbation approach, as also seen by simulation results. There was some indication that the results had improved since the development phase due to changes to the application of the constrained hot deck.

- The analysis on weighted cell counts revealed the same acceptable level of impact from the perturbation approach as seen in the development phase. In general, there was minimal impact at the county level, and more impact at the TAZ level. Additional checks that were conducted on tables involving industry showed favorable results as well.

- The analysis of the perturbation's impact on error measures showed about the same level of impact as determined acceptable in the development phase. The simulation results helped to give an indication that the perturbation impact on the variances at a combined TAZ level (populations of about 4,000 workers) for mean travel time was not significant. In general, one can expect the perturbation to increase standard errors in most cases by three to 10 percent for areas of that size.

- The analyses involving Cramer's *V*, pairwise associations, and multivariate associations all showed results similar to the development phase. The results led to a couple of simple adjustments to the specification of the constrained hot deck as it related to perturbing age.

Comparison with Home-Based Work Outputs

As mentioned above, there were two test sites for the validation phase: Atlanta, GA (carried forward from the development phase), and Olympia, WA. The testing approach was the same as the development phase, except the five-year ACS (2005–2009) data were compared with the model outputs (the development phase used the three-year ACS [2006–2008]). Olympia did not report its model results

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

by super district nor any other aggregation of TAZs, so tests were conducted only for the county and TAZ level.

Compatibility between TAZ geographies continued to create problems in the comparison tests. Conducting the comparison required a lookup table (sometimes called an equivalency table or "crosswalk") between the Census 2000 TAZ systems and the TAZ system used in the Metropolitan Planning Organization (MPO) model. These tables were created using geographic information systems (GIS) for both test sites based on information received from the MPO. In Olympia, the number of TAZs increased from 213 (Census 2000) to 766 (current model system). By further disaggregating the ACS data for comparison with the new TAZ system, the already small cell samples got even smaller. Furthermore, for the validation phase comparison tests only ACS cells with more than five sample cases were included in the comparison (the development phase tests included ACS cells with more than 30 sample cases). The resulting sparseness of data was potentially harmful to certain comparisons and overall usability for transportation planners.

For each test, for each test site, at each specified level of geography, the following comparisons were made:

- Raw ACS Minus Model;
- Perturbed ACS Minus Model; and
- Perturbed ACS Minus Raw ACS.

Both the absolute difference and percent relative difference were computed for each comparison and the following summary statistics are reported:

- Interquartile Range (75th percentile minus 25th percentile);
- Median (50th percentile).

Each table also included the size of the matrix (number of estimates) for each test, and the ACS sample size (number of respondents) underlying the tabulation. For Tests 3, 4, 5, 6, and 7, the analysis was limited to travel flows in which both the home end and work end were in the MPO area and in which there were corresponding TAZs. As noted earlier, only cells with more than five sample cases were included in the comparison tests. For the TAZ and district level tests, these steps greatly reduced the amount of ACS data available for testing.

For example, consider in the Olympia test site Test 4 (mean travel time by means of transportation [7]) at the TAZ level. The initial set of ACS data for means of transportation category 2 (drive alone) was 3,295 home-work TAZ pairs from 5,682 sample cases After subsetting to cells with more than five sample cases, and those with matching home and work TAZs within the MPO area, only 12 TAZ pairs remained, consisting of 95 sample cases. These data were not meaningful enough to analyze, and drive alone was by far the dominant mode, so other MOTs were not examined for Test 4.

For Test 5 (person trips by household income by means of transportation) in Atlanta at the district (aggregations of TAZs), the analysis was conducted both with and without the minimum of five sample cases in each cell. As noted earlier, risk analysis conducted at the TAZ level indicated that nearly 90 percent of TAZ flows in the Atlanta area consisted of singletons or doubletons. Including these flows in the district analysis greatly improved performance of both the raw and perturbed ACS compared with the

travel model outputs and compared to each other. The tests are shown by variable, test site, and level of geography in Table 3-15.

Table 3-15 shows the comparison tests that were performed and the location of the corresponding summary tables in Appendix R. The scatter plots in Appendix S were subset to flows with more than five sample cases due to disclosure concerns. With the perturbation targeted to tables and flows with a small number of cases, and most, if not all, such flows removed from the scatter plot, then the scatter plot may show perfect or near-perfect correlation between the raw and perturbed data for a small subset of all total flows.

Test 1: Population by Age Category (Model) vs. Age of Worker (ACS)

Olympia

Test 1 was not conducted for Olympia because the MPO model did not include population by age category.

Atlanta

County-level tests for Test 1 were previously completed in the development phase. TAZ level tests for Test 1 in Atlanta are included in Table R-1. Both the raw ACS and perturbed ACS estimate more total population (workers) than the model. The model estimated more younger workers and more older workers than ACS, and the reverse occurred in the "middle" age categories (see Figure S-1). **As a result, r-squared values comparing the model and ACS were very low (see scatter plots in Figures S-2 and S-3). Relative to each other and against the model, raw and perturbed ACS performed equally well (see Figure S-4).**

Test 2: Households by Number of Workers (Model) vs. Households by Number of Workers (ACS)

Olympia

Results for Test 2 in Olympia are located in Table R-2 for the county level and Table R-3 for the TAZ level. The ACS number of workers categories was collapsed to match the MPO model data. The TAZ level estimates were sufficient to allow separate computations for each category of number of workers per household. For the county, the model underestimated the number of zero-worker households compared to ACS and overestimates one, two, and three-plus worker households as well as total households (see Figure S-5). Scatter plots of raw ACS vs. model at the TAZ level were contained in Figure S-6 through Figure S-9. **The relationship between the raw ACS and the model was poor across all household categories and for total households, with r-squared values ranging from .18 for total households to .08 for two-worker households. The perturbed ACS and raw ACS performed the same against the model, so scatter plots of perturbed ACS vs. model, which are excluded from Appendix J, would be identical to the raw ACS vs. model graphs.**

Atlanta

County level comparisons for Test 2 were previously completed in the development phase. TAZ level comparisons for Test 2 in Atlanta are located in Table R-4. Both raw and perturbed ACS compared poorly to the model when considering all number of worker categories together, with r-squared values below .1 (see Figures S-22 and S-23). This performance was expected, since ACS estimates only 61 percent of the total households estimated by the model. At the individual number of worker category level, comparison of ACS and model improved, with r-squared values ranging from .26 for one-worker

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

households to .49 for three-plus worker households (see Figures S-10 through S-24). **Raw and perturbed ACS performance was virtually identical for all Test 2 comparisons.**

Table 3-15. Validation Phase: Tests for Comparison of Travel Demand Model Output and Raw and Perturbed ACS Data (and Direct Comparison of Raw and Perturbed ACS) and Location of Data Tables in Appendix R

Test	Model Component	Model Data	ACS Data	Test Sites / Level of Geography				CTPP Part[9]	Universe for ACS Data
				Atlanta		Olympia			
				County	Sub-County	County	Sub-County		
1	Population Synthesizer / Trip Generation	Population by Age Category	Age of Worker (8)	Completed in development phase	TAZ (Table R-1)	No	No	Part 1	Workers 16+ in HHs
2	Population Synthesizer / Trip Generation	Households by number of workers	Households by number of workers (5)	Completed in development phase	TAZ (Table R-4)	Yes (Table R-2)	TAZ (Table R-3)	Part 1	Households
3	Trip Generation / Trip Distribution	Person Trips	Total Workers (1)	Completed in development phase	Completed in development phase	Yes	TAZ (Table R-5)	Part 3	Workers 16+ in HHs
4	Mode Choice / Assignment	Average Travel Time by Mode	Mean TT (1) by MOT (7)	Yes (Table R-8)	District Table R-9)	Yes (Table R-6)	TAZ (Table R-7)	Part 3	Workers 16+ in HHs
5	Trip Distribution / Mode Choice	Person trips by HH inc by mode	HH Inc (5) by MOT (7)	Yes (Table R-12)	District (Table R-13)	Yes (Table R-10)	TAZ (Table R-11)	Part 3	Workers 16+ in HHs
6	Trip Distribution / Mode Choice	Person trips by age of worker by mode	Age of Worker (6) by MOT (7)	Yes (Table R-14)	No	No	No	Part 3	Workers 16+ in HHs
7	Trip Generation / Mode Choice	Person trips by age of worker by mode	Age of worker (4) by MOT (4)	Yes (Table R-15)	TAZ (Table R-16)	No	No	Part 3	Workers 16+ in HHs

[9] It is not clear if this naming convention will be retained in the new CTPP

Test 3: Person Trips (Model) vs. Total Workers (ACS) – Flows

Olympia

Test 3 results in Olympia for the TAZ level are reported in Table R-5. The relative levels of HBW person trips for raw ACS, perturbed ACS, and the model at the county level are reported in Figure S-25. The model estimates nearly doubled the trips as ACS. At the TAZ level, only a small amount of data remained for testing after subsetting to cells with more than 5 sample cases and matching and adjusting between the two TAZ systems. No relationship existed between the model data and either the raw or perturbed ACS data; regression lines were nearly horizontal. **The raw ACS and perturbed ACS matched almost exactly.**

Atlanta

Test 3 for Atlanta was previously completed as part of the development phase.

Test 4: Average Travel Time by Mode (Model) vs. Mean Travel Time by MOT (ACS) -- Flows

Olympia

Test 4 results in Olympia for the county level are reported in Table R-6. At the TAZ level, the data were too sparse for meaningful computation after subsetting to matching internal MPO TAZ pairs and cells with more than five sample cases. A relative comparison of average travel times at the county level was shown in Figure S-29. Model travel times were slightly higher than both raw and perturbed ACS for drive alone and shared ride, and sharply higher for transit and bike/walk. **Perturbed ACS travel times were slightly higher than raw ACS for drive alone and shared; for transit and bike/walk, raw and perturbed ACS travel times were identical.**

Atlanta

Test 4 results in Atlanta are reported in Table R-8 for the county level and Table R-9 for the district level. The model estimated higher average county-to-county travel times than both raw and perturbed ACS for all travel modes; the difference was greatest for public transportation (see Figure S-30). The match at the county level between the model and both raw and perturbed ACS was good, with r-squared values around .72 for raw ACS and .70 for perturbed ACS (see Figures S-31 and S-32). At the district level, where data were much more sparse after removing flows with five and fewer sample cases, the match between model and ACS was poorer, with r-squared values of .46 for raw ACS and .45 for perturbed ACS (see Figures S-34 and S-35). **There was little difference between the raw and perturbed ACS relative to each other (see Figure S-33 and S-36).**

Test 5: Person Trips by Household Income by Mode (Model) vs. Household Income [5] by Means of Transportation [7] (ACS) – Flows

Olympia

Test 5 results for Olympia are reported in Table R-10 for the county level and Table R-11 for the TAZ level. The MPO in Olympia could not provide trips by income by mode but did provide HBW trip productions by income. This provided a measure of the trip-making on the home end for different income groups but was not a true flow, so comparison with the ACS was matched only with the MPO's internal residential TAZs. ACS income categories were collapsed to match the MPO's income categories. A relative comparison of trips by income is shown in Figure S-11. The model estimates were slightly under ACS for the low income category, sharply higher than ACS for the middle income category, and higher than ACS for the high income category. Perturbed ACS was slightly lower than raw ACS for the low

income category and slightly higher than raw ACS for the middle income category. **Raw ACS was slightly higher than perturbed ACS for the high income category.**

Atlanta

Test 5 results for Atlanta are reported in Table R-12 at the county level and in Table R-13 at the district level. The model estimated more trips across all modes and income categories except for drive alone at the highest income category, where the model estimated a little more than half the travel flows of ACS (see Figure S-37). Overall, the county-level match between model and both raw and perturbed ACS was less than desirable, with r-squared values around .54 (see Figures S-38 and S-39). The county-level match between raw and perturbed ACS was lower than for other tests, with an r-squared value of .82 (see Figure S-40). The lower match was likely due to the challenges on both datasets due to the sparseness of data resulting from cross-tabulating a continuous variable with means of transportation.

The challenges created by sparse data were further illustrated by both the district level comparisons and two additional analyses performed for Test 5 in Atlanta only. At the district level, the match between the model and both raw and perturbed ACS was extremely poor, with the model estimating much higher district flow values than ACS (see Figures S-42 and S-43). R-squared values were not reported for these comparisons.

Two additional analyses were conducted for Test 5. At the county level, comparisons were made again while limiting flow cell values to 20,000 and lower. This cutoff was chosen because it is the population threshold for the aggregations of TAZs known as transportation analysis districts (TADs). While there was not a one-to-one relationship between population and HBW trips, this level did illustrate the impact of data availability for smaller areas. R-squared values at the county level dropped to .45 (see Figure S-41). **Adding singletons and doubletons to the district analysis improved the performance of raw ACS vs. perturbed ACS r-squared from .74 to .99, and improved ACS performance against the model from no relationship to an r-squared value of .40.** The scatter plots with the full range of sample cases are not shown in Appendix S. Although the inclusion of singletons and doubletons improved the performance of ACS, those flows would be suppressed according to DRB rules under all scenarios, if perturbation were not applied.

Test 6: Person Trips by Age of Worker by Mode (Model) vs. Age of Worker [6] by Means of Transportation [7] (ACS)—Flows

Olympia

Test 6 was not conducted for Olympia because the MPO model did not include age of worker as a variable.

Atlanta

Test 6 was conducted only at the county level and results are reported in Table R-14. **Perturbed ACS and raw ACS provided an equally good match to the model output, with r-squared values of about .88 (see Figures S-49 and S-50). There was a good match between raw and perturbed ACS as well (see Figure S-51).**

Test 7: Person Trips by Age of Worker by Mode (Model) vs. Age of Worker [4] by Means of Transportation [7] (ACS) – Flows

Olympia

Test 7 was not conducted for Olympia because the MPO model did not include age of worker as a variable.

Atlanta

Results for Test 7 were reported in Table R-15 at the county level and Table R-16 at the TAZ level. Test 7 differed from Test 6 in that both age of worker and means of transportation were collapsed into four categories. At the county level, raw ACS and perturbed ACS both provided a good match (r-squared of .91) to the model (see Figures S-52 and S-53). Raw and perturbed ACS matched well with each other (see Figure S-54). At the TAZ level the difference between both raw and perturbed ACS and the model were more prominent due to the sparseness of data at the TAZ level. **Both ACS tabulations compared poorly to the model (see Figures S-55 and S-56). The raw and perturbed ACS compared well to each other at the TAZ level (see Figure S-57).**

Summary of Conclusions and Recommendations

The stated purpose of these comparison tests was to conduct a reasonableness check to determine if the performance of the perturbed ACS CTPP tabulations was no worse than the raw tabulations when compared against typical model outputs. **Based on the results discussed above and shown in greater detail in Appendix R and Appendix S, the research team concluded that the validation phase test results largely replicated the development phase test results: the performance of the perturbed ACS tabulations was equal to that of the raw ACS when comparing to model output.** Again, there was little important difference between the raw and perturbed ACS tabulations for the comparison tests.

The TAZ level tests in Olympia and the district and TAZ level tests in Atlanta highlighted the sparseness of the data at fine geographies and for multivariate cross-tabulations with means of transportation. The planning community was made aware of these issues impacting data usability as they began delineating their Census 2010 TAZs.

3.2.2 Disclosure Risk Measures

Risk measures that were created during the development phase, as described in Section 2.2.2, were used for the validation phase. Of note is one difference as it relates to the risk component for the matchability to the ACS PUMS. Using the three-year ACS, the research team evaluated the risk involved in matching to the ACS PUMS using the current set of variables involved in the set of flows tables. There were about 50 percent to 75 percent of the records (depending on the test site) that could be uniquely identified using the 10 flow attributes and PUMA. About 80 percent of them were flow singletons. Therefore, about 40 percent to 60 percent were high risk exact matching singleton flows (calculation: 50 percent of 80 percent = 40 percent; 75 percent of 80 percent = 60 percent). Taking a 2/3 subsample for the Public Use Microdata set (PUMS) resulted in an expected match rate of about 27 percent to 40 percent (calculation: (2/3) 40 percent = 27 percent; (2/3) 60 percent = 40 percent), or under 50 percent uniques. Using the five-year ACS, the team evaluated the risk involved in matching to the ACS PUMS using the set of variables involved in the set of flow tables at the time of the validation phase. There were about 70 percent (Atlanta) to 74 percent (Olympia) of the records that could be uniquely identified using the 10 flow attributes and PUMA. About 50 percent of them were flow singletons. This resulted in an expected match rate of about 23 percent to 25 percent (calculation: 70 percent of 50 percent * (2/3) = 23 percent;

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

74 percent of 50 percent * (2/3) = 25 percent). Therefore, the component r_1 (the proportion of sample unique records at risk due to matching to the ACS PUMS five-year file) was set to .23 for the computation of the overall risk score that was presented to the Census Bureau DRB. Summary tables were produced and provided to the Census DRB for review on February 3, 2011. The risk estimates for the measurement error components and the overall risk score measure were provided and were found to be at an acceptable level. Therefore, approval was given by the Census Bureau DRB to move ahead to the nationwide testing and production run with the partial replacement rates used in the validation phase.

4. Production Run Processing

This chapter introduces the processing steps in Section 4.1, describes the program components in Section 4.2, and discusses other important topics in Section 4.3, including the Set A and Set B tables, variance estimation, and guidance on adding tables to Set B.

4.1 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) 2006–2010 sample data. The five-year ACS data will be processed through six main programming components of the procedure without human intervention. 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.

The technical approaches are described in Chapter 3, and further details are given in Chapter 2 where indicated as appropriate. Changes since the validation phase (Chapter 3) are indicated here as follows:

1. The variable NEW_POVERTY was incorporated as a predictor in the models.
2. AGE9 hot deck cells. Because correlations with AGE9 were attenuated, dichotomized income, poverty status, and number of workers in the household were included in the hot deck cell specification.
3. JWMN and JWD. A minor adjustment was made in the use of the partial replacement flag in the hot deck cell specification for the JWMN and JWD replacement.
4. Raking convergence criteria adjustment to include a criterion for relative difference. The raking algorithm was modified to check for convergence at the beginning of each iteration. The processing was stopped if convergence was reached.
5. Adjustments to model areas were necessary during the transition from processing test sites to processing areas of the nation. This is discussed further in Chapter 3.
6. The perturbation rates were lowered to improve the data utility while keeping the disclosure risk at an acceptable level to the Census Bureau Disclosure Review Board (DRB).

With the changes listed above, the programs were tested on the full national five-year ACS data (2005–2009). The documentation in this chapter and computer programs (residing at the Census Bureau) conform with the Census Bureau's software platform for special tabulations, and take into account the

hardware on which the Census Bureau operates for this function. The documentation in this chapter gives sufficient detail for the new system to be self-contained and autonomously operational. The research team worked in cooperation with the Census Bureau staff to build a system that is in conformity with their production schedule and computing constraints, and took into account any parameters needed to meet these.

Figure 4-1 provides the process flow of the overall program processing. The ACS five-year files from the Census Bureau contain the recodes needed for the CTPP tables, as well as imputation flags. Swapping flags from the ACS disclosure protection process were also provided. Several preliminary steps are conducted to prepare for the processing of the perturbation approaches.

The main driver for the overall program (ctpp_main_driver.sas) is found in Appendix T. Figure U-1 provides a hierarchical list of the programs by each of the six major components of the main program. Figure U-2 provides the same list of programs in alphabetical order, associating each program with its main component and giving a brief description of each program.

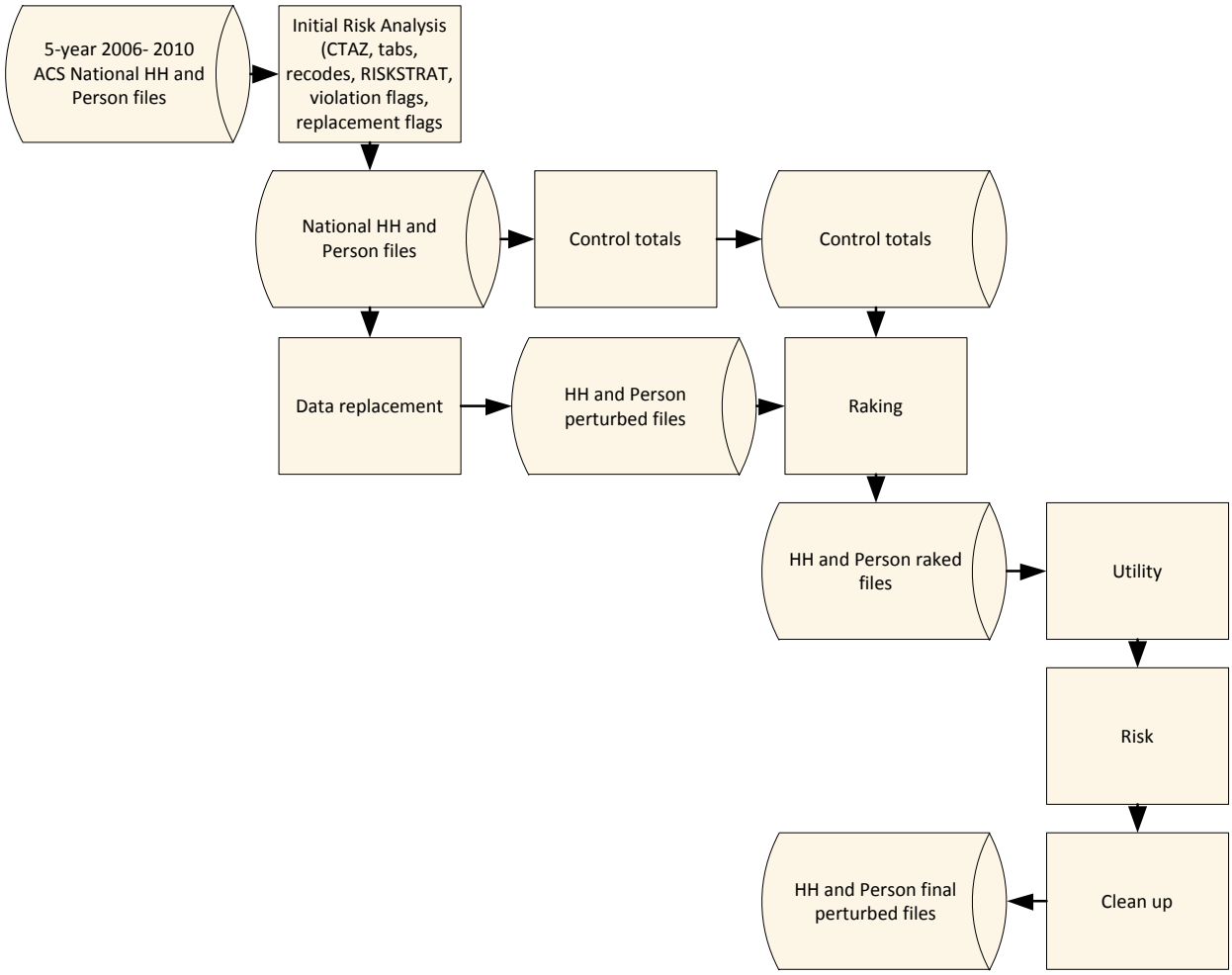


Figure 4-1. Overall Perturbation Process Flowchart

4.2 PROGRAM COMPONENTS

Each of the following sections describes a main component of the overall program. Each section contains a brief description, a table of inputs and outputs of the main datasets of ACS sample households and persons, a flowchart of the process, and a reference to the appendix containing the main driver of the program component. As mentioned above, the main driver for the overall program (ctpp_main_driver.sas) is provided in Appendix T.

4.2.1 Program Component: Initial Risk Analysis

The set of initial risk analysis modules was processed to generate the Set B tables for the purpose of flagging data values that violated the DRB rules and therefore are at the highest risk of disclosure. The table generator part of the program needed to incorporate any additional tables requested (discussed in Section 4.3). Several steps were necessary within the initial risk analysis component to prepare for the application of the perturbation approach, including the creation of Combined TAZs (CTAZs) and 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, or swapped through the ACS disclosure process, will not be replaced; that is, they will be considered to have already been perturbed. This approach is acceptable to the DRB. As part of the initial risk analysis, data values were classified according to risk strata. The following flags were created to assist in the perturbation process as well as in the disclosure risk measures: VarName_FLG, VarName_RPL, VarName_FULL, and VarName_STRT. Figures 4-2, 4-3, and 4-4 provide the flowcharts of the process. Figure 4-2 shows the creation of the CTAZs and the CTAZ-level covariates that were used in the modeling steps in data replacement. Figure 4-3 and Figure 4-4 show the initial risk analyses at the person level and at the household level, respectively. Appendix T provides the main driver of this program component (ira_main_driver.sas). Table 4-1 outlines the main differences between the input and output files (at both household level and person level) in the initial risk analysis.

Table 4-1. Initial Risk Analysis: Difference between Input and Output Datasets

Input dataset	VPERS5REC	VHOUS5REC
Number of records in input dataset	22,821,787	9,771,627
Number of variables in input dataset	568	426
Input dataset description	Original person file containing all persons	Original household file containing all housing units
Output dataset	VPERS5REC_IRA	VHOUS5REC_IRA
Number of records in output dataset	10,333,156	8,984,138
Number of variables in output dataset	308	207
Output dataset description	Subset of workers 16 and over	Subset of households
Number of variables in common	186	121
Number of variables in input dataset only	382	305
Number of variables in output dataset only	122	86
Number of variables changed types	1	1
Number of variables changed values	2	2

NOTE: Sample counts are based on ACS 2005–2009 data.

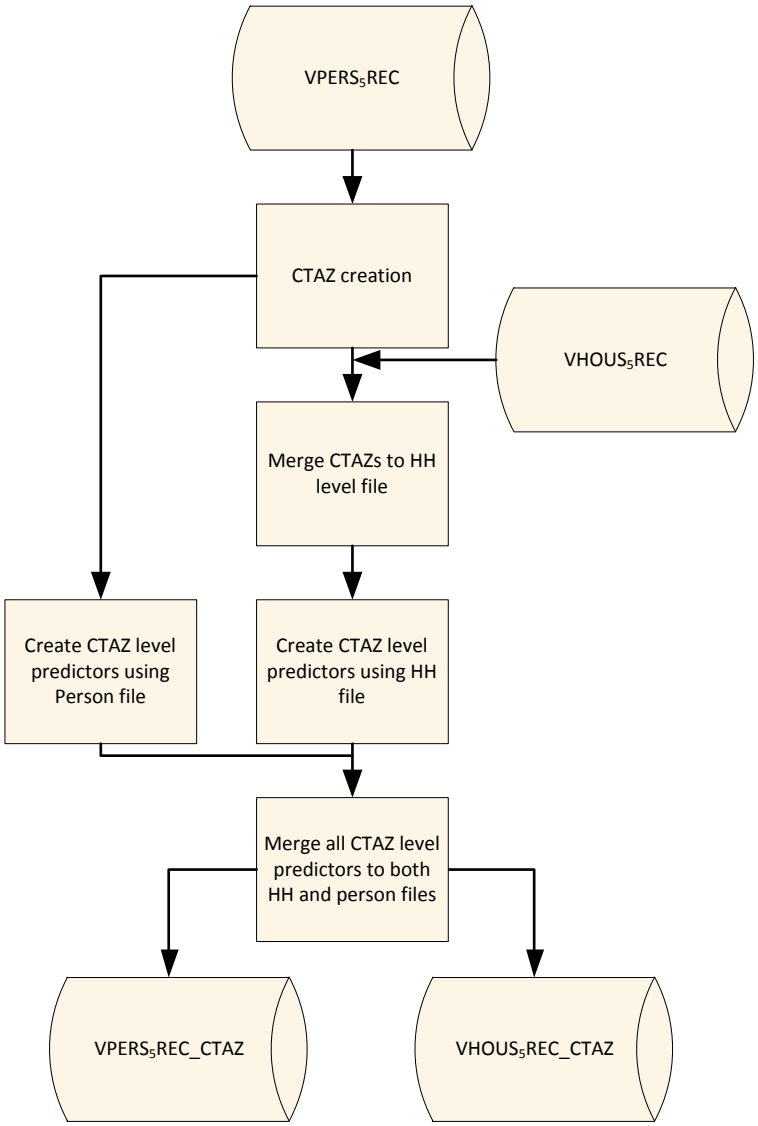


Figure 4-2. Flowchart of Creation of CTAZs and CTAZ-Level Covariates Program Component

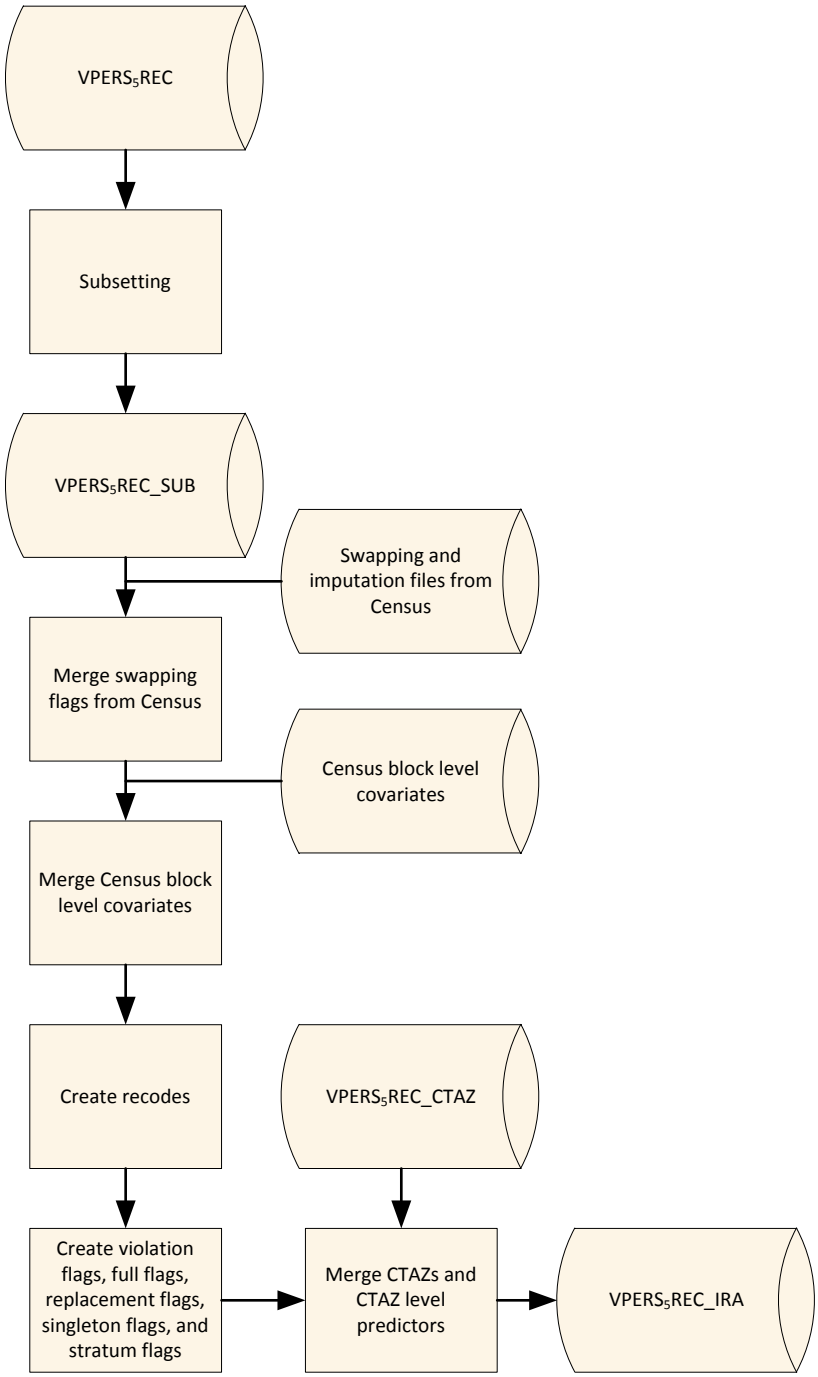


Figure 4-3. Flowchart of Person-Level Initial Risk Analysis Program Component

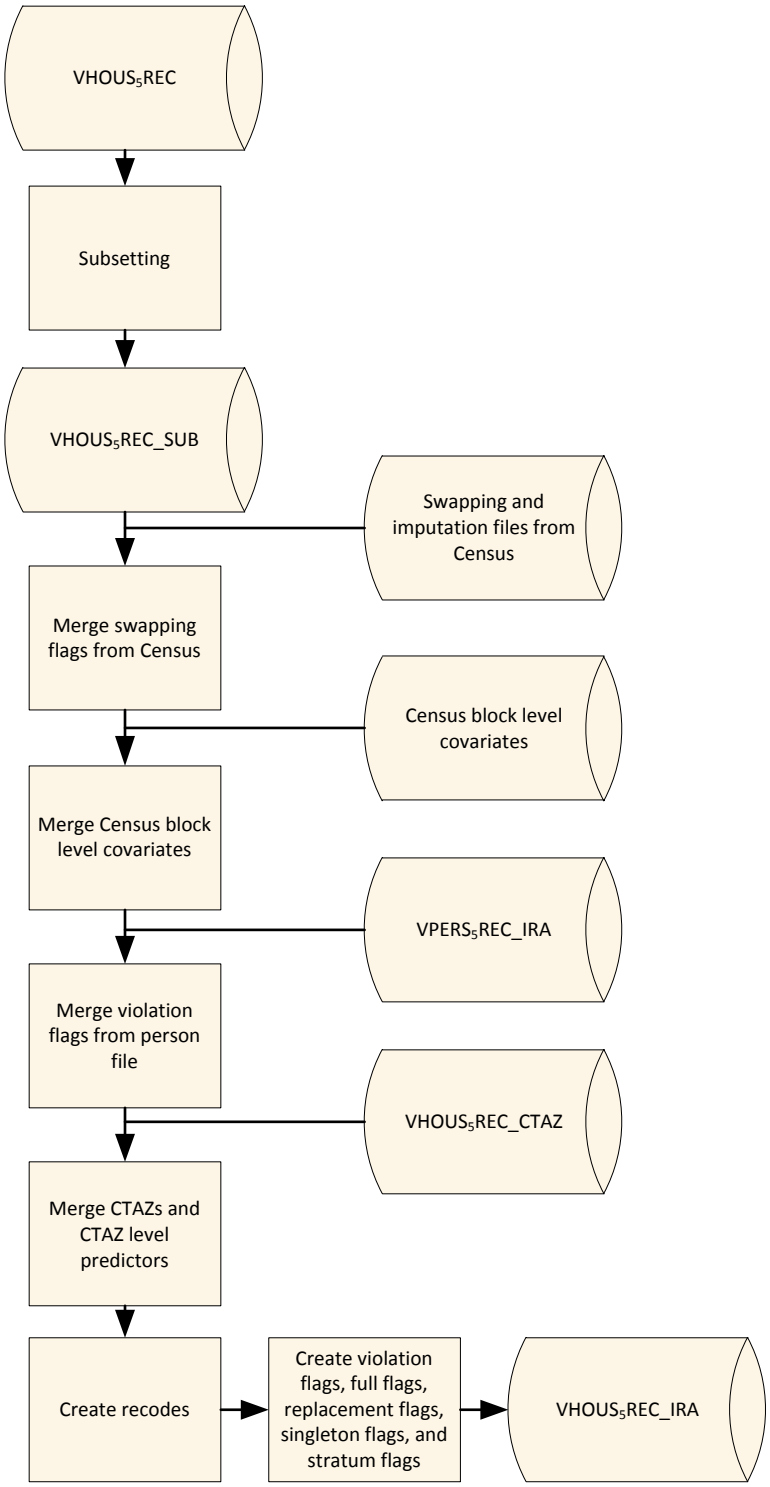


Figure 4-4. Flowchart of Household Level Initial Risk Analysis Program Component

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

4.2.2 Program Component: Data Replacement

This program combines the 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 perturbed, 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 perturbed 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
Approach	=	“CH” or “RL” or “SP” for constrained hot deck, rank linking, or semi-parametric respectively
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	=	“OC” for ordered categorical variables, “UC” for unordered categorical variables, and “N” or continuous variables
Replace	=	1/0 determines if the VarName needs to be perturbed
VarToBin	=	name of the variable to make bins for, typically same as VarName
BinVar	=	name of the variable that contains the bins
Bins	=	statements defining the bins, separated by semi-colons
NumWtCells	=	integer value of the number of weight groups to form
WtCellVar	=	name of the variable containing the weight groups
HDCellVars	=	list of variables to help define the hot deck cells (exclude WtCellVar)
LinkToVar	=	name of the variable (&VarName) used to link to via the rank linking process
TrgtVars	=	blank or list of variable(s) linked and targeted in same process
Interaction	=	1/0 determines if an interaction term needs to be created for the VarName
Predictor	=	1/0 determines if the variable (&VarName) should be included in model selection for the semi-parametric approach
ForceList	=	list of variables to force in the models for semi-parametric approach
Include	=	integer value of the number of variables in the ForceList

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 perturbation step.

One by one, the target variables are processed through the Main Loop. Either the constrained hot deck or the semi-parametric approach is processed, depending on the variable type of the target variable.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

First, household-level variables are perturbed, then the perturbed household variables are transferred to the person level, where the process continues with the perturbations on person-level variables.

After processing, pre-post checks are conducted in order to have an initial look at the impact of the perturbations. Frequencies, means, and correlations are generated before and after perturbation. Lastly, recodes are processed in order to prepare for the raking step. Figure 4-5 provides the flowchart of the process. Figure 4-6 provides an example of a MIF. Appendix T provides the main driver of this program component (data_replacement.sas). Table 4-2 outlines the main differences between the input and output files (at both household level and person level) in the data perturbation process.

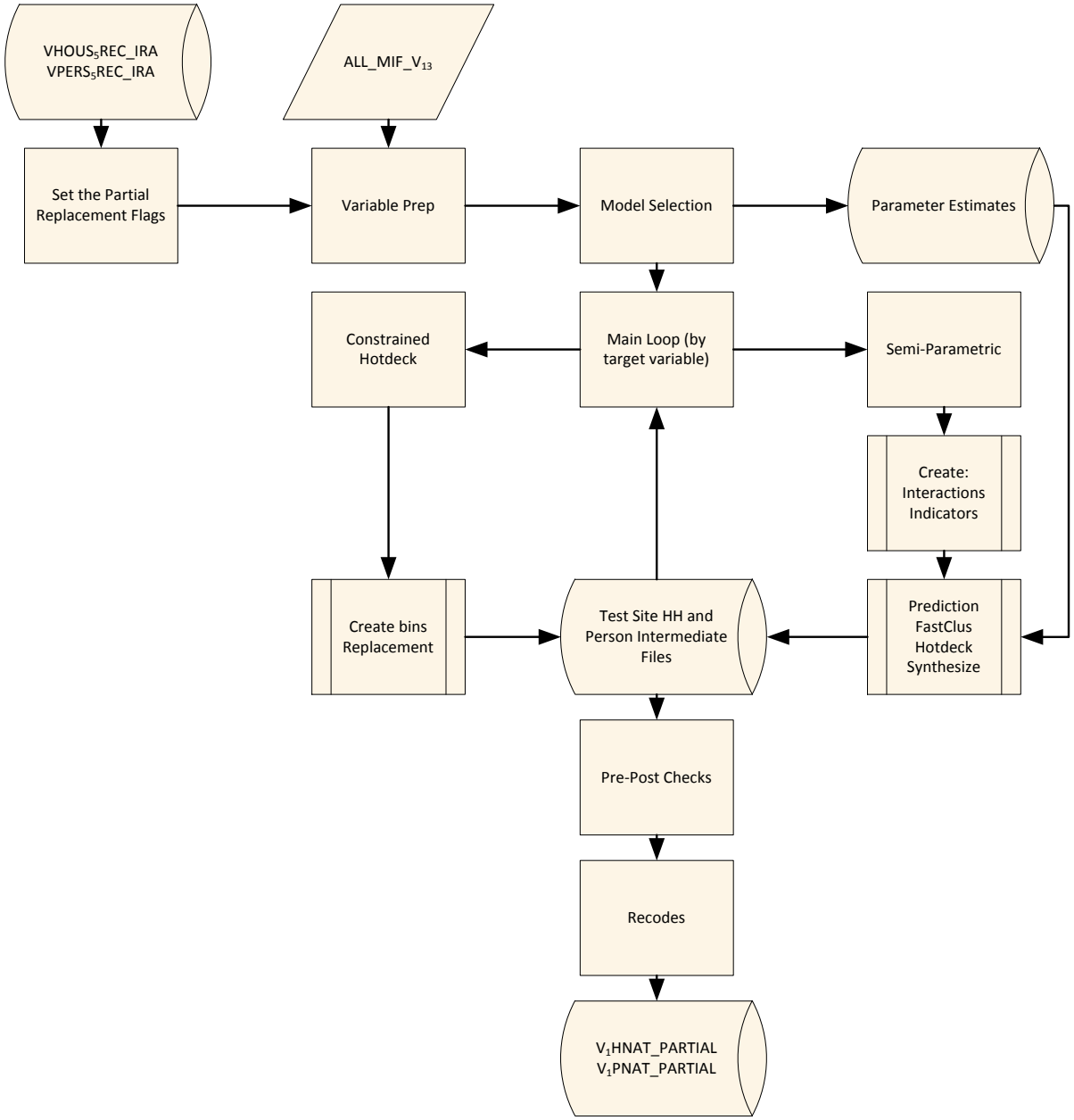


Figure 4-5. Flowchart of Data Replacement Program Component

Item	ProcessNumber	VarName	Approach	VPERS	VHOUS	Transfer	Type	Replace	VarToBin	BinVar
1	1	AHINC	CH	1	1	1	OC	1	AHINC	BIN1
…	…	…	…	…	…	…	…	…	…	…
6	5	AGE9	CH	1	0	0	OC	1	AGE	BIN5
7	6	POVERTY	RL	1	0	0	OC	1		
8	7	MINORITY	SP	1	0	0	OC	1		
…	…	…	…	…	…	…	…	…	…	…
15		AVG_HHSIZE		1	1	0	N	0		

Item	Bins	NumWgtCell	WgtCellVar	HDCellVar	LinkToVar	TrgtVars	Interaction	Predictor	ForceList	Include
1	(.,20000]; (20000,30000]; (30000,42500]; (42500,62500]; (62500,87500]; (87500,125000]; (125000, 250000]	3	WTCELL1	AHINC_&amount ST PUMA5 VEHICLES6 BIN1		AHINC	1	1		
…	…	…	…	…		…	…	…	…	…
6	(15,34]; (34,52]; (52,62]	3	WTCELL4	age9_&amount ST PUMA5 MEANS6 BIN5		AGE9	1	1		
7		0		MISSPOVERTY ST PUMA5 ACSHH_WRK6 VEHICLES6	AHINC	POVERTY	0	0		
8				RCTAZ50			1	1	I_MEANS11_1-I_MEANS11_(MaxValue-1) AGE9 I_JWD_SHIFT_1 I_JWD_SHIFT_2 I_JWD_SHIFT_3 NEW_JWMN AHINC NEW_POVERTY VEHICLES6	(MaxValue-1)+8
…	…	…	…	…		…	…	…	…	…
							1	1		

Figure 4-6. An example of a Master Index File

Table 4-2. Data Replacement: Difference between Input and Output Datasets

Input dataset	VPERS5REC_IRA	VHOUS5REC_IRA
Number of records in input dataset	10,333,156	8,984,138
Number of variables in input dataset	308	207
Input dataset description	Output from person initial risk analysis	Output from household initial risk analysis
Output dataset	V1PNAT_PARTIAL	V1HNAT_PARTIAL
Number of records in output dataset	10,333,156	8,984,138
Number of variables in output dataset	560	359
Output dataset description	Perturbed person file	Perturbed household file
Number of variables in common	308	207
Number of variables in input dataset only	0	0
Number of variables in output dataset only	252	152
Number of variables changed types	6	8
Number of variables changed values	17	5

NOTE: Sample counts are based on ACS 2005–2009 data.

4.2.3 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 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. In addition, a dimension was added to calibrate to the estimated total number of workers at the CTAZ300 level, which are areas of about 8,000 in population. The weight calibration process employed sample-based raking, meaning that the estimates for the modified estimates reflected the sampling error of the five-year ACS control totals, rather than consider these totals to be error-free, as is often the case with calibration methods. For sample-based raking, each replicate weight for the modified file was raked to its corresponding replicate weight estimated totals from the five-year ACS. Figure 4-7 provides the flowchart of the process. Appendix T provides the main driver of this program component (raking_driver.sas). Table 4-3 outlines the main differences between the input and output files (at both household level and person level) in the raking process.

Table 4-3. Raking: Difference Between Input and Output Datasets

Input dataset	V1PNAT_PARTIAL	V1HNAT_PARTIAL
Number of records in input dataset	10,333,156	8,984,138
Number of variables in input dataset	560	359
Input dataset description	Perturbed person file	Perturbed household file
Output dataset	RV1PNAT_PARTIAL	RV1HNAT_PARTIAL
Number of records in output dataset	10,333,156	8,984,138
Number of variables in output dataset	563	360
Output dataset description	Raked perturbed person file	Raked perturbed household file
Number of variables in common	559	357
Number of variables in input dataset only	1	2
Number of variables in output dataset only	4	3
Number of variables changed types	0	0
Number of variables changed values	0	0

NOTE: Sample counts are based on ACS 2005–2009 data.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

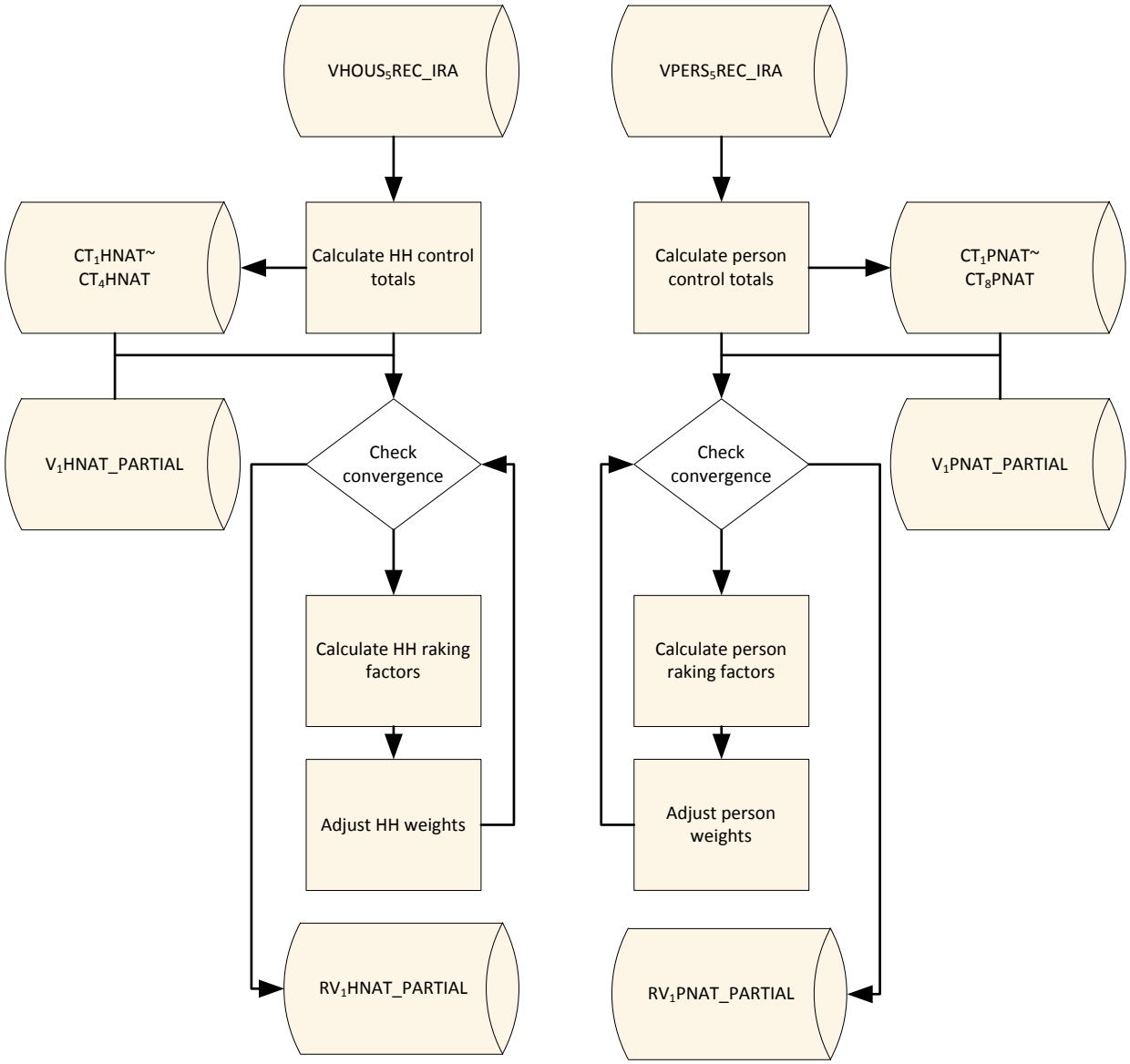


Figure 4-7. Flowchart of Household Level and Person Level Control Total Calculations and Raking Program Component

4.2.4 Program Component: Utility Measures

The data perturbation 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 perturbed ACS data. The comparisons check cell means, 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 perturbed estimates (across estimates for geographic areas) were computed where appropriate in order to give indications of potential bias introduced by the

perturbation. The interquartile range (IQR) for the differences provided an indication of the variation caused by the perturbations. Lastly, there is a check on the differences for medians and 75th percentiles of travel time for table cells across estimates for geographic areas. Figure 4-8 provides the flowchart of the process. Appendix T provides the main driver of this program component (utility.sas).

The utility measures can be calculated for the nation or by state. If the state-level utility measures are desired, the programs can be modified to loop across all the states. The developed utility measures, as illustrated above, will be generated within each state by comparing the state-level ACS data and perturbed data.

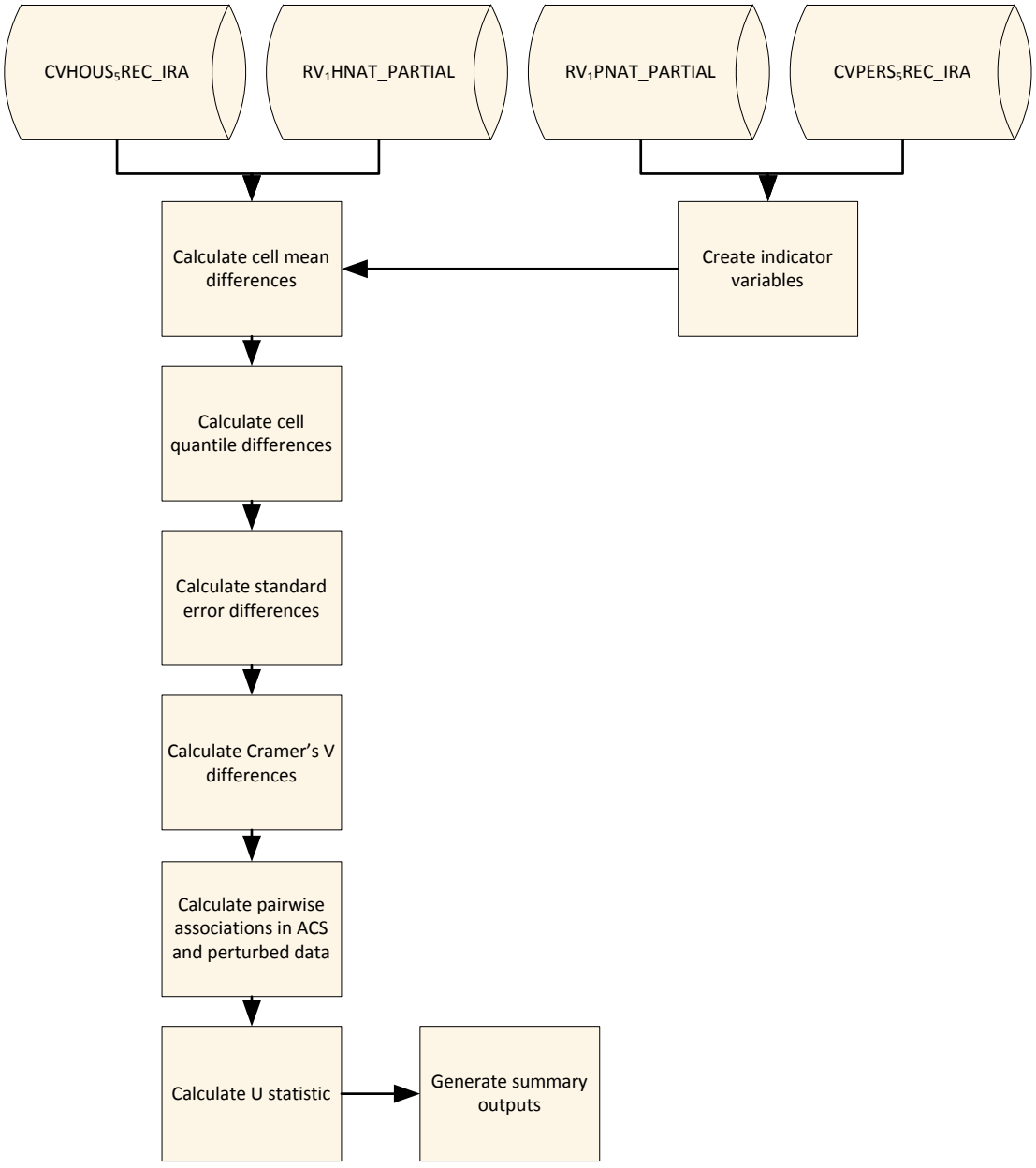


Figure 4-8. Flowchart of Data Utility Measures Program Component

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

4.2.5 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. The research team and the Census DRB recognize 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 perturbed CTPP data.

The general approach is to bring together measures of various risk elements, including a measure of the amount of changed information. The measures were found acceptable by the DRB. While these risk components can 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. Figure 4-9 provides the flowchart of the process. Appendix T provides the main driver of this program component (risk.sas).

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

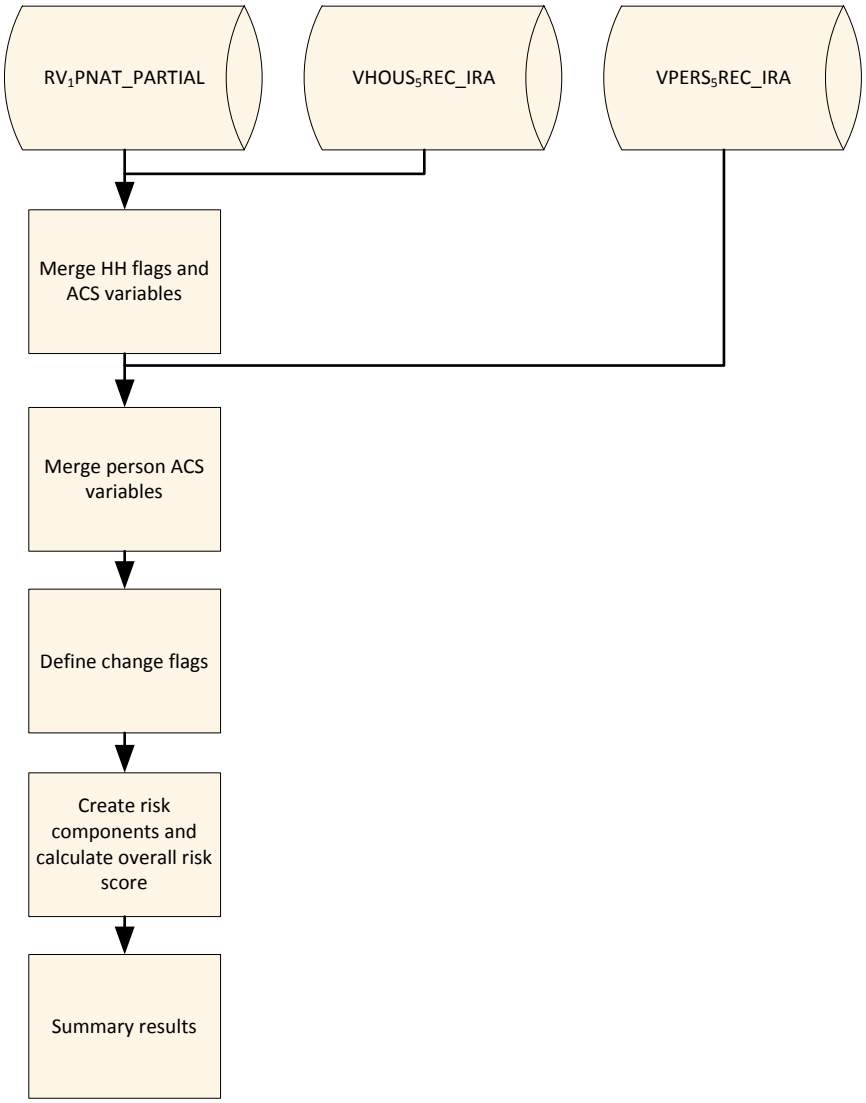


Figure 4-9. Flowchart of Disclosure Risk Measures.

4.2.6 Program Component: Cleanup

This program creates the delivery files after the processes of initial risk analysis, perturbation, raking, risk, and utility are finished. The final files at the household and person levels will contain the ID variables, the perturbed variables, and their recodes, which will be used in Set B tables. Table 4-4 outlines the main differences between the input and output files (at both household level and person level) in the cleanup process. Appendix T provides the main driver of this program component (cleanup.sas).

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Table 4-4. Cleanup: Difference between Input and Output Datasets

Input dataset	RV1PNAT_PARTIAL	RV1HNAT_PARTIAL
Number of records in input dataset	10,333,156	8,984,138
Number of variables in input dataset	563	360
Input dataset description	Raked perturbed person file	Raked perturbed household file
Output dataset	PERT_VPERS5	PERT_VHOUS5
Number of records in output dataset	10,333,156	8,984,138
Number of variables in output dataset	- -	- -
Output dataset description	Finalized perturbed person file	Finalized perturbed household file
Number of variables in common	- -	- -
Number of variables in input dataset only	544	357
Number of variables in output dataset only	0	0
Number of variables changed types	0	0
Number of variables changed values	0	0

NOTE: Sample counts are based on ACS 2005–2009 data.

4.3 OTHER TOPICS RELATING TO THE PROCESSING RUNS

At the time of the writing of this report, the Census Bureau is exploring options for the implementation of the perturbation programs described in Section 4.2. All components of the perturbation programs, as composited together in a single call, reside at the Census Bureau with the intention that the Census Bureau will implement the procedures as documented in this chapter.

To help in this transition, this section provides more information about the Set A and Set B tables in Section 4.3.1, guidance for the computation of variances in Section 4.3.2, and for adding tables to Set B in Section 4.3.3.

4.3.1 Set A and Set B tables

As approved by the Census Bureau Disclosure Review Board (DRB), the current CTPP tables will be divided into two sets: Set A and Set B. Set A tables will be produced based on real ACS five-year data, ACS full sample and replicate weights. The variances will be estimated using usual ACS formula (see formula (f5) in Section 3.1.4 for details). Set B tables, shown in Appendix C, will be based on perturbed ACS data and CTPP adjusted weights. The variance estimation for Set B tables will be elaborated upon in the next section. The American Association of State Highway and Transportation Officials representatives are working toward a final table request, which will be provided to the Census Bureau.

For each set of tables, the appropriate geography and a point estimate will be provided with an associated margin of error. The Set A tables will be shown without cell suppression rules. The usual rounding rules will apply. The Set B tables will be shown without cell suppression rules applied and the values shown in the tables will be rounded to the nearest integer. Users will see inconsistencies in the weighted marginal totals for identical variables used in both sets of tables. However, within the Set B tables, aggregations to higher levels of geography will match the result for the higher level of geography. Also, residence locality will match aggregations of flow tables for the same residence locality. The same is true for workplaces and flows to the same workplace. More discussion on what to expect from the Set A and Set B tables can be found in Section 1.4.3.

The microdata file from which the tables are generated was produced solely for the purpose of generating the tables. It is not intended to be used for dynamic queries for tables or analyses other than the Set B tables. Variables that would be directly derived from the perturbed variables, but are not in the Set B tables, have not been adjusted. Among other examples, highly correlated variables, such as income, earnings and poverty have not been fully adjusted, only to the extent necessary to process the Set B tables.

4.3.2 Variance Estimation for Set B Tables

This section provides guidance for implementing the variance estimation approach that will be used to produce the CTPP tabulations. Applying the usual ACS variance formula to the perturbed data may result in biased variance estimates because the ACS formula only accounts for the ACS sampling error, but not the variance component associated with the perturbation. Research was conducted, as outlined in Sections 2.1.5, 3.1.4, and 3.2.1, to evaluate the performance of several variance estimators. After a careful review, the Census DRB and the Census Bureau ACS Sample Design group approved the decision of using formula (f5), given in Section 3.1.4 and shown below, for variance estimation in the production process of the CTPP tables. Assuming perturbation is independent of the sampling process, formula (f5) is essentially the sum of sampling variance and perturbation variance:

$$\mathrm{var}\left(\tilde{\theta}_0\right) = \mathrm{var}\left(\hat{\theta}_0\right) + \left(\tilde{\theta}_0 - \hat{\theta}_0\right)^2, \qquad \text{(f5)}$$

where $\tilde{\theta}_0$ represents the CTPP perturbed estimate of θ.

Computationally, formula (f5) requires the following information:

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

The processing takes the following steps:

- Generate the point estimates for all Set B tables twice: once for ACS data, and once for the perturbed data;
- Using the successive difference replication formula (f1) given in Section 2.1.5, generate the ACS variance estimates using ACS data and ACS full sample and replicate weights;
- Using formula (f5), compute the variances for the perturbed estimates as the sum of ACS variances and squared difference between the ACS and perturbed estimates.

4.3.3 Impact of Adding Tables

There is potential that more Set B tables will need to be produced in the future upon transportation users' requests. In that case, some components in the overall perturbation process have to be tailored to account for the additional tables, and the whole process needs to be tested and validated before final production.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

First of all, the initial risk analysis should be modified to evaluate both the existing tables and the new tables. If any new variable in the new tables is subject to perturbation, a set of flag variables (violation flag, replacement flag, singleton flag, and full flag) and a stratum variable will need to be created for the new variable, as for other existing variables that are subject to perturbation. It is expected that the proportion of values in the highest risk strata may increase if more tables, especially flow tables, are added. The time and effort needed to modify the programs for the initial risk analysis should be minor. But the analysis results need to be carefully reviewed and justified since the risk stratum variables will have a direct impact on the determination of the perturbation rates.

If the new tables do not contain any new variables that are subject to perturbation, the rest of the components in the overall process can remain the same. Adjusting the perturbation rates may be needed, but that can be done fairly easily when the partial flags are created. If the new tables do contain some new variables that need to be perturbed, modifications of the programs become necessary, mainly in the data replacement component. The modifications include (1) adding the new variables into the MIF file once the appropriate perturbation approaches are chosen (the parameters for the existing variables may also need to be revised for the purpose of maintaining the associations between the new variables and existing variables); (2) setting up the perturbation rates and creating the partial flags for the new variables; and (3) adding adequate quality control checks on the new variables before and after perturbation. If the new variables need any special treatments such as adding random noise, additional changes can be done in the programs where appropriate. If the new variables have multiple versions/recodes in the Set B tables, the recode program should be updated to ensure that the changes in different versions are synchronized during data perturbation.

The raking process will not be affected by adding more Set B tables unless changing the raking dimensions is desired. There will be some impact on risk and utility components if the new variables need to be accounted for in the risk and utility measures. The disclosure risk results will be reviewed by the Census Bureau DRB. The cleanup component will be changed slightly to deliver the newly perturbed variables.

If the new table requests involve more detailed levels for means of transportation, for example, MOT18, the data values at risk are estimated to increase by at least 20 percent. More perturbation is necessary due to higher disclosure risk, which will greatly increase the relative change to margin of errors due to perturbation. As a conclusion, the overall process and the programs can be flexibly adjusted when there are more table requests, but the impact of adding tables on disclosure risk and data utility is worth more attention and consideration.

References

Abowd, J., Andersson, F., Graham, M., Vilhuber, L. and Wu, J. (2009). Formal Privacy Guarantees and Analytical Validity of OnTheMap Public-use Data. NSF-Census-IRS workshop on Synthetic Data and Confidentiality Protection.

Agresti, A., Booth, J.G., Hobart, J.P. and Caffo, B. (2000). Random-effects modeling of categorical response data, Sociological Methodology 30, 27–80.

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

Bocci, C. and Beaumont, J. (2009) Synthetic data creation for the cross national equivalent file. Proceedings of Statistics Canada Symposium 2009

Cambridge Systematics, Inc, Fienberg, S., and Love, T. (2009). Disclosure avoidance techniques to improve ACS data availability for transportation planners. Request by AASHTO Standing Committee on Planning. NCHRP Project 08-36, Task 71.

Census Bureau (2009). 2006-2008 PUMS accuracy of the data (Revised December 30, 2009). US Bureau of the Census (http://www.census.gov/acs/www/Downloads/2006-2008/AccuracyPUMS.pdf -- current as of May 18, 2010)

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.

Dohrmann, S., Krenzke, T. Roey, S., and Russell, N. (2009). Evaluating the impact of data swapping using global utility measures. To appear in the Proceedings of the Federal Committee on Statistical Methodology Research Conference.

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 Saygin, Y. eds. (2008). Privacy in Statistical Databases. Lecture Notes in Computer Science 5262 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, No. 4, 2009, pp. 589–603.

Duncan, G.T., Fienberg, S.E., Kishnan, R., Padman, R., and Roehrig, S.F. (2001). Disclosure limitation methods and information loss for Tabular Data. Chapter 7 in Doyle, P., Lane, J. I., Theeuwes, J. M., Zayatz, L., eds. (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. Elsevier Science BV. Amsterdam, The Netherlands. Pages 135-166.

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.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

Elliot, M. J., Manning, A., Ford, R (2002). A Computational Algorithm for Handling the Special Uniques Problem. *International Journal of Uncertainty, Fuzziness and Knowledge Based Systems* 5(10): 493-509.

Evans, T., Zayatz, L, and Slanta, J. (1998). Using noise for disclosure limitation of establishment tabular data. *Journal of Official Statistics*, 14: 537-551.

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*.

FCSM (2005). Report on Statistical Disclosure Methodology. Statistical Policy Working Paper 22 of the Federal Committee on Statistical Methodology, 2nd version. Revised by Confidentiality and Data Access Committee 2005, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget.

Fienberg, S.E., and McIntyre, J. (2004). Data swapping: Variations on a theme by Dalenius and Reiss. In Domingo-Ferrer, J., and Torra, V. (2004), *Privacy in Statistical Databases: CASC Project International Workshop*, Lecture Notes in Computer Science, Springer, New York. Pages 14-29.

Fischetti, M., and Salazar-González, J-J. (1998). Experiments with controlled rounding for statistical disclosure control in tabular data with linear constraints. *Journal of Official Statistics*, 14: 553-565.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association.*

Giessing, S. (2001). Nonperturbative disclosure control methods for tabular data. Chapter 9 in Doyle, P., Lane, J. I., Theeuwes, J. M., Zayatz, L., eds. (2001). *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies.* Elsevier Science BV. Amsterdam, The Netherlands. Pages 185-213.

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.*

Greenberg, B. (1987). Rank swapping for masking ordinal microdata, US Bureau of the Census (unpublished manuscript).

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *TEST.* An Official Journal of the Spanish Society of Statistics and Operations Research. Vol. 15, No. 1 pp. 1-96.

Judkins, D., Piesse, A., and Krenzke, T. (2008). Multiple semi-parametric imputation. Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.

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): pp. 224-232.

Kaufman, S., Seastrom, M., and Roey, S. (2005). Do disclosure controls to protect confidentiality degrade the quality of data? Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Government Statistics.

Krenzke, T. and Hubble, D. (2009). Toward quantifying disclosure risk for area-level tables when public microdata exists. To appear in the Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets Practice on the Map. IEEE 24th International Conference on Data Engineering (ICDE).

Massell, P., Zayatz, L., and Funk, J. (2006). Protecting the confidentiality of survey tabular data by adding noise to the underlying microdata: Application to the Commodity flow survey. In Domingo-Ferrer, J. and Franconi, L., eds. (2006). Privacy in Statistical Databases. Lecture Notes in Computer Science 4302 Springer-Verlag Berlin Heidelberg. Pages 304-317.

McWethy, L. 2008. SIPP Employment Data Analysis. Memo from Laura McWethy to Elaine Murakami (Federal Highway Administration), dated August 12, 2008.

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.

Moore, R. (1996). Controlled data swapping techniques for masking public use datasets. Census Bureau Statistical Research Division research report.

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. Pages 102-113.

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

Raghunathan, T.E., Solenberger, P., and van Hoewyk, J. (2002). *IVEware: Imputation and Variance Estimation Software*, available at: http://www.isr.umich.edu/src/smp/ive/.

Rao, J.N.K. (2003). *Small area estimation*. Wiley-Interscience.

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

Scheffé (1956). A "mixed model" for the analysis of variance. *The Annals of mathematical statistics*, 27, 23-36.

NCHRP Project 08-79 Final Report:
Producing Transportation Data Products from the American Community Survey That Comply With Disclosure Rules

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*.

Shlomo, N. and Skinner, C. (2009). Assessing the Protection Provided by Misclassification-Based Disclosure Limitation Methods for Survey Microdata.

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.

Willenborg, L. and de Waal, T. (1996). Statistical Disclosure Control in Practice. Springer, New York.

Willenborg, L. and de Waal, T. (2001). Elements of Statistical Disclosure Control. Springer. New York.

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, Number 1, pp. 111-124.

Zayatz, L. (2008). New ways to provide more and better data to the public while still protecting confidentiality. Proceedings of the Joint Statistical Meetings, American Statistical Association Section on Survey Research Methods.

Zayatz, L., Lucero, J., Massell, P., and Ramanayake, A. (2009), "Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products," Proceedings of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (Bilbao, Spain, 2-4 December 2009).

Appendix A

CENSUS TRANSPORTATION PLANNING PRODUCTS (CTPP)
Standard Tabulations for Research
Part 1, Residence Tables
Part 2, Workplace Tables
Part 3, Worker Flow Table

Note: The research team was provided these tables from AASHTO. Since providing these tables, the following tables were added in Part 1 and Part 2.

1. MOT(11) by Presence of children in household (3)
2. MOT(11) by Household workers (3)
3. MOT(11) by Minority (3)

The industry variable was agreed upon to have the following 7 categories in part 3, in lieu of the 14-category version given in this appendix.

- 2 – Agriculture, forestry, fishing and hunting, mining, construction, and armed forces
- 3 – Manufacturing
- 4 – Wholesale trade, retail trade, transportation and warehousing, and utilities
- 5 – Information, finance, insurance, real estate and rental and leasing, professional, scientific, management, administrative, and waste management services
- 6 – Educational, health and social services
- 7 – Arts, entertainment, recreation, accommodation and food services
- 8 – Other services including public administration

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables

					Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables			3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
1-047	1	100	11100		Total Population (1)	All persons	1	√		√	√	
NEW	1	101	11101		Age (11)	All persons	11	√		√	√	
NEW	1	102	11102		Disability (6)	All persons	6		√			
NEW	1	103	11103		Urban/Rural residence (3)	All persons	3	√		√	√	
NEW	1	104	11104		Hispanic origin (3)	All persons	3	√		√	√	
NEW	1	105	11105		Hours worked per week (7)	All persons	7		√			
NEW	1	106	11106		Length of US Residence (6)	All persons	6	√		√	√	
NEW	1	107	11107		Minority status (3)	All persons	3	√		√	√	
1-048	1	108	11108		Unweighted sample count of the population (1)	All persons	1	√		√	√	
1-049	1	109	11109		Percent of population in sample (1)	All persons	1	√		√	√	
NEW	1	110	11110		Race (5)	All persons	5	√		√	√	
NEW	1	111	11111		School Enrollment (7)	All persons	7	√		√	√	
NEW	1	112	11112		Sex (3)	All persons	3	√		√	√	
1-055	1	200	11200		Age (11) by Disability status (7)	All persons	77		√			
1-052	1	201	11201		Age (11) by Minority status (3)	All persons	33	√		√	√	
1-053	1	202	11202		Age (9) by School enrollment (7)	All persons	63	√		√	√	
1-051	1	203	11203		Age (11) by Sex (3)	All persons	33	√		√	√	
1-056	1	204	11204		Employment status (7) by Disability status (7)	All persons	49		√			
1-054	16	205	11205		Employment status (7) by Sex (3)	Persons 16 years and over	21	√		√	√	
1-050	1	206	11206		Hispanic Origin (3) by Race of Person (5)	All persons	15	√		√	√	
NEW	17	207	11207		Linguistic Isolation (3) by Language spoken at home (12)	Persons 5 years old and over in households	36	√		√	√	
NEW	2	100	12100		Total workers (1)	Workers 16 years and over	1	√		√	√	
NEW	2	101	12101		Age of worker (8)	Workers 16 years and over	8	√		√	√	
NEW	2	102	12102		Class of worker (9)	Workers 16 years and over	9	√		√	√	
NEW	2	103	12103		Earnings in the past 12 months (2008$) (11)	Workers 16 years and over	11	√		√	√	
NEW	2	104	12104		Industry (15)	Workers 16 years and over	15	√		√	√	
NEW	5	105	12105		Industry (15)	Workers 16 years and over who are not self-employed	15	√		√	√	
NEW	2	106	12106		Means of transportation (18)	Workers 16 years and over	18	√		√	√	
NEW	2	107	12107		Occupation (25)	Workers 16 years and over	25	√		√	√	
1-001	2	108	12108		Time leaving home (41)	Workers 16 years and over	41	√		√	√	
NEW	7	109	12109		Workers per car, truck, or van (1)	Workers 16 years and over who used car, truck or van	1	√		√	√	
NEW	2	112	12112		Usual Hours worked per week (7)	Workers 16 years and over	7	√		√	√	
NEW	2	118	12118		Travel time to work (12)	Workers 16 years and over	12	√		√	√	
1-033	2	200	12200		Age of Worker (8) by Earnings in the past 12 months (2008$) (11)	Workers 16 years and over	88	√		√		√
1-014	2	201	12201		Age of Worker (8) by Means of transportation (11)	Workers 16 years and over	88	√		√	√	
	2		12201C	c	Age of Worker (8) by Means of transportation (7)	Workers 16 years and over	56	√		√	√	
	2		12201C2	c2	Age of Worker (8) by Means of transportation (6)	Workers 16 years and over	48	√		√	√	
	2		12201C3	c3	Age of Worker (8) by Means of transportation (4)	Workers 16 years and over	32	√		√	√	
1-112	9	202	12202		Aggregate Carpools (1) by Time leaving home (5)	Workers 16 years and over who carpooled	5	√		√	√	
1-012	2	203	12203		Class of worker (9) by Means of transportation (11)	All workers	99		√			
1-008	2	204	12204		Disability status (7) by Means of transportation (11)	All workers	77		√			
1-013	2	205	12205		Earnings in 2007$ (11) by Means of transportation (11)	All workers	121		√			
1-019	2	206	12206		Earnings in the past 12 months (2008$) (11) by Travel time (12)	Workers 16 years and over	132	√		√		√

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables (Continued)

			Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables					3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
	2		12206C	c	Earnings in the past 12 months (2008$) (6) by Travel time (12)	Workers 16 years and over	72	√		√	√	
1-116	2	207	12207		Earnings in 2007$ [aggregate] (1) by Means of transportation (11)	All workers	11		√			
1-096	2	208	12208		Earnings in 2007$ [mean] (1) by Means of transportation (11)	All workers	11		√			
1-094	2	209	12209		Ernings in 2007$ [median] (1) by Means of transportation (11)	All workers	11		√			
NEW	2	210	12210		Hispanic Origin (3) by Race (5)	Workers 16 years and over	15	√		√	√	
1-011	2	211	12211		Industry (15) by Earnings in the past 12 months (2008$) (11)	Workers 16 years and over	165	√		√		√
	2		12211C	c	Industry (15) by Earnings in the past 12 months (2008$) (6)	Workers 16 years and over	90	√		√	√	
1-010	2	212	12212		Industry (15) by Means of transportation (11)	All workers	165		√			
1-017	2	213	12213		Industry (15) by Time leaving home (17)	Workers 16 years and over	255	√		√		√
	2		12213C	c	Industry (15) by Time leaving home (10)	Workers 16 years and over	150	√		√	√	
1-018	2	214	12214		Length of US residence (6) by Earnings in the past 12 months (2008$) (11)	Workers 16 years and over	66	√		√		√
	2		12214C	c	Length of US residence (6) by Earnings in the past 12 months (2008$) (6)	Workers 16 years and over	36	√		√	√	
1-020	2	215	12215		Length of US residence (6) by Means of transportation (11)	All workers	66		√			
1-025	2	216	12216		Minority Status (3) by Class of worker (9)	Workers 16 years and over	27	√		√	√	
1-026	2	217	12217		Minority Status (3) by Earnings in the past 12 months (2008$) (11)	Workers 16 years and over	33	√		√	√	
1-024	2	218	12218		Minority Status (3) by Industry (15)	Workers 16 years and over	45	√		√	√	
1-027	2	219	12219		Minority Status (3) by Means of transportation (11)	All workers	33		√			
1-023	2	220	12220		Minority Status (3) by Occupation (25)	Workers 16 years and over	75	√		√	√	
1-028	2	221	12221		Minority Status (3) by Travel time (18)	Workers 16 years and over	54	√		√	√	
1-015	2	222	12222		Occupation (25) by Industry (15)	Workers 16 years and over	375	√		√	√	
	2		12222C	c	Occupation (8) by Industry (15)	Workers 16 years and over	120	√		√	√	
1-009	2	223	12223		Occupation (25) by Means of transportation (11)	All workers	275		√			
1-016	2	224	12224		Occupation (25) by Time leaving home (17)	Workers 16 years and over	425	√		√		√
	2		12224C	c	Occupation (8) by Time leaving home (10)	Workers 16 years and over	80	√		√		√
1-002	2	225	12225		Sex (3) by Means of transportation (18)	All workers	54		√			
1-021	2	226	12226		Time leaving home (17) by Means of transportation (11)	Workers 16 years and over	187	√		√		√
	2		12226C	c	Time leaving home (10) by Means of transportation (7)	Workers 16 years and over	70	√		√	√	
	2		12226C2	c2	Time leaving home (10) by Means of transportation (6)	Workers 16 years and over	60	√		√	√	
	2		12226C3	c3	Time leaving home (10) by Means of transportation (4)	Workers 16 years and over	40	√		√	√	
1-022	2	227	12227		Travel time (12) by Means of transportation (11)	Workers 16 years and over	132	√		√	√	
	2		12227C	c	Travel time (12) by Means of transportation (7)	Workers 16 years and over	84	√		√	√	
	2		12227C2	c2	Travel time (12) by Means of transportation (6)	Workers 16 years and over	72	√		√	√	
	2		12227C3	c3	Travel time (12) by Means of transportation (4)	Workers 16 years and over	48	√		√	√	
1-119	6	229	12228C0	c0	Aggregate Travel time (1) by Means of transportation (18)	Workers 16 years and over who did not work at home	18	√		√	√	
1-118	6	228	12228		Aggregate Travel time (1) by Means of transportation (11)	Workers 16 years and over who did not work at home	11	√		√	√	
	6		12228C	c	Aggregate Travel time (1) by Means of transportation (7)	Workers 16 years and over who did not work at home	7	√		√	√	
	6		12228C2	c2	Aggregate Travel time (1) by Means of transportation (6)	Workers 16 years and over who did not work at home	6	√		√	√	
	6		12228C3	c3	Aggregate Travel time (1) by Means of transportation (4)	Workers 16 years and over who did not work at home	4	√		√	√	
1-119	2	229	12229		Travel time [aggregate] (1) by Means of transportation (18)	All workers	18		√			

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables (Continued)

					Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables			3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
1-103	6	231	12230C0	c0	Mean Travel time (1) by Means of transportation (18)	Workers 16 years and over who did not work at home	18	√		√	√	
1-102	6	230	12230		Mean Travel time (1) by Means of transportation (11)	Workers 16 years and over who did not work at home	11	√		√	√	
	6		12230C	c	Mean Travel time (1) by Means of transportation (7)	Workers 16 years and over who did not work at home	7	√		√	√	
	6		12230C2	c2	Mean Travel time (1) by Means of transportation (6)	Workers 16 years and over who did not work at home	6	√		√	√	
	6		12230C3	c3	Mean Travel time (1) by Means of transportation (4)	Workers 16 years and over who did not work at home	4	√		√	√	
1-103	2	231	12231		Travel time [mean] (1) by Means of transportation (18)	All workers	18		√			
1-101	6	233	12232C0	c0	Median Travel time (1) by Means of transportation (18)	Workers 16 years and over who did not work at home	18	√		√	√	
1-100	6	232	12232		Median Travel time (1) by Means of transportation (11)	Workers 16 years and over who did not work at home	11	√		√	√	
	6		12232C	c	Median Travel time (1) by Means of transportation (7)	Workers 16 years and over who did not work at home	7	√		√	√	
	6		12232C2	c2	Median Travel time (1) by Means of transportation (6)	Workers 16 years and over who did not work at home	6	√		√	√	
	6		12232C3	c3	Median Travel time (1) by Means of transportation (4)	Workers 16 years and over who did not work at home	4	√		√	√	
1-101	2	233	12233		Travel time [median] (1) by Means of transportation (18)	All workers	18		√			
1-110	7	234	12234		Aggregate Vehicles used (1) by Time leaving home (5)	Workers 16 years and over who used car, truck or van	5	√		√	√	
1-113	9	235	12235		Workers per carpool (1) by Time leaving home (5)	Workers 16 years and over who carpooled	5	√		√	√	
1-113	2	235	12235		Workers per carpool [mean] (1) by Time leaving home (4)	All workers	4		√			
1-111	7	236	12236		Workers per car, truck, or van (1) by Time leaving home (5)	Workers 16 years and over who used car, truck or van	5	√		√	√	
1-120	6	300	12300		Aggregate Travel time (1) by Means of transportation (11) by Time leaving home (17)	Workers 16 years and over who did not work at home	187	√		√	√	
	6		12300C	c	Aggregate Travel time (1) by Means of transportation (7) by Time leaving home (10)	Workers 16 years and over who did not work at home	70	√		√	√	
	6		12300C2	c2	Aggregate Travel time (1) by Means of transportation (6) by Time leaving home (10)	Workers 16 years and over who did not work at home	60	√		√	√	
	6		12300C3	c3	Aggregate Travel time (1) by Means of transportation (4) by Time leaving home (10)	Workers 16 years and over who did not work at home	40	√		√	√	
1-107	6	301	12301		Mean Travel time (1) by Means of transportation (11) by Time leaving home (17)	Workers 16 years and over who did not work at home	187	√		√	√	
	6		12301C	c	Mean Travel time (1) by Means of transportation (7) by Time leaving home (10)	Workers 16 years and over who did not work at home	70	√		√	√	
	6		12301C2	c2	Mean Travel time (1) by Means of transportation (6) by Time leaving home (10)	Workers 16 years and over who did not work at home	60	√		√	√	
	6		12301C3	c3	Mean Travel time (1) by Means of transportation (4) by Time leaving home (10)	Workers 16 years and over who did not work at home	40	√		√	√	
1-106	6	302	12302		Median Travel time (1) by Means of transportation (11) by Time leaving home (17)	Workers 16 years and over who did not work at home	187	√		√	√	

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables (Continued)

					Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables			3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
	6		12302C	c	Median Travel time (1) by Means of transportation (7) by Time leaving home (10)	Workers 16 years and over who did not work at home	70	√		√	√	
	6		12302C2	c2	Median Travel time (1) by Means of transportation (6) by Time leaving home (10)	Workers 16 years and over who did not work at home	60	√		√	√	
	6		12302C3	c3	Median Travel time (1) by Means of transportation (4) by Time leaving home (10)	Workers 16 years and over who did not work at home	40	√		√	√	
NEW	3	100	13100		Total Workers in households (1)	Workers 16 years and over in households	1	√		√	√	
1-030	3	101	13101		Household income in the past 12 months (2008$) (26)	Workers 16 years and over in households	26	√		√	√	
1-038	3	200	13200		Age group of youngest child in the household (5) by Means of transportation (11)	Workers in households	55		√			
1-117	3	201	13201		Earnings in 2007$ [aggregate] (1) by Means of transportation (11)	Workers in households	11		√			
1-097	3	202	13202		Earnings in 2007$ [mean] (1) by Means of transportation (11)	Workers in households	11		√			
1-095	3	203	13203		Earnings in 2007$ [median] (1) by Means of transportation (11)	Workers in households	11		√			
1-034	3	204	13204		Household income in the past 12 months (2008$) (26) by Means of transportation (11)	Workers 16 years and over in households	286	√		√	√	
	3		13204C	c	Household income in the past 12 months (2008$) (16) by Means of transportation (7)	Workers 16 years and over in households	112	√		√	√	
	3		13204C2	c2	Household income in the past 12 months (2008$) (16) by Means of transportation (6)	Workers 16 years and over in households	96	√		√		√
	3		13204C3	c3	Household income in the past 12 months (2008$) (16) by Means of transportation (4)	Workers 16 years and over in households	64	√		√	√	
1-039	3	205	13205		Household income in the past 12 months (2008$) (9) by Earnings in the past 12 months (2008$) (11)	Workers 16 years and over in households	99	√		√		√
	3		13205C	c	Household income in the past 12 months (2008$) (9) by Earnings in the past 12 months (2008$) (6)	Workers 16 years and over in households	54	√		√		√
1-036	3	206	13206		Poverty status (4) by Means of transportation (11)	Workers in households	44		√			
1-040	4	207	13207		Poverty status (4) by Time leaving home (17)	Workers for whom poverty status is determined	68	√		√	√	
NEW	3	208	13208		Vehicle availability [ratio] (4) by Length of US Residence (5)	Workers in households	20		√			
NEW	3	209	13209		Vehicle availability [ratio] (4) by Means of transportation (11)	Workers in households	44		√			
1-041	3	210	13210		Vehicles available (6) by Length of US residence (6)	Workers 16 years and over in households	36	√		√	√	
1-035	3	211	13211		Vehicles available (6) by Means of transportation (11)	Workers 16 years and over in households	66	√		√	√	
	3		13211C	c	Vehicles available (6) by Means of transportation (7)	Workers 16 years and over in households	42	√		√	√	
	3		13211C2	c2	Vehicles available (6) by Means of transportation (6)	Workers 16 years and over in households	36	√		√	√	
	3		13211C3	c3	Vehicles available (6) by Means of transportation (4)	Workers 16 years and over in households	24	√		√	√	
1-032	4	212	13212		Vehicles available (6) by Poverty status (4)	Workers 16 years and over in households for whom poverty status is determined	24	√		√	√	
NEW	3	213	13213		Vehicles available [ratio] (4) by Poverty status (4)	Workers in households	16		√			

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables (Continued)

					Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables			3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
1-031	3	214	13214		Number of workers in household (6) by Household size (5)	Workers 16 years and over in households	30	√		√	√	
1-037	3	215	13215		Workers in household (6) by Means of transportation (11)	Workers in households	66		√			
1-044	3	300	13300		Household income in 2007$ (5) by Minority Status (3) by Means of transportation (8)	Workers in households	120		√			
1-043	3	301	13301		Household income in the past 12 months (2008$) (5) by Number of Workers in household (6) by Age group of youngest child in the household (5)	Workers 16 years and over in households	150	√		√	√	
1-042	3	302	13302		Household income in 2007$ (5) by Vehicles available (5) by Means of transportation (8)	Workers in households	200		√			
1-046	3	303	13303		Poverty status (4) by Minority Status (3) by Means of transportation (8)	Workers in households	96		√			
NEW	3	304	13304		Vehicle availability [ratio] (4) by Minority Status (3) by Means of transportation (8)	Workers in households	96		√			
1-045	3	305	13305		Vehicles available (3) by Minority Status (3) by Means of transportation (8)	Workers in households	48		√			
1-060	12	100	14100		Total households (1)	Households	1	√		√	√	
NEW	12	101	14101		Number of Persons under 18 in household (4)	Households	4	√		√	√	
NEW	12	NEW	14102A	a	Aggregate Household income in the past 12 months (2008$) (1)	Households	1	√		√	√	
NEW	12	102	14102		Mean Household income in the past 12 months (2008$) (1)	Households	1	√		√	√	
NEW	12	103	14103		Median Household income in the past 12 months (2008$) (1)	Households	1	√		√	√	
NEW	12	104	14104		Lifecycle of household (16)	Households	16	√		√	√	
NEW	12	105	14105		Household size (5)	Households	5	√		√	√	
NEW	12	NEW	14106A	a	Population in households (1)	Households	1	√		√	√	
NEW	12	106	14106		Mean number of Persons per household (1)	Households	1	√		√	√	
NEW	13	107	14107		Poverty status (4)	Households for which poverty status is determined	4	√		√	√	
NEW	11	108	14108		Telephone availability (3)	Occupied housing units	3	√		√	√	
1-061	11	109	14109		Tenure (5)	Occupied housing units	5	√		√	√	
NEW	11	110	14110		Vehicles available (6)	Occupied housing units	6	√		√	√	
1-109	11	111	14111		Aggregate Number of Vehicles Available in Households (1)	Occupied housing units	1	√		√	√	
NEW	12	112	14112		Number of Workers in household (6)	Households	6	√		√	√	
NEW	12	200	14200		Household income in the past 12 months (2008$) (5) by Lifecycle of household (16)	Households	80	√		√	√	
NEW	12	201	14201		Household income in the past 12 months (2008$) (9) by Number of Persons under 18 (4)	Households	36	√		√		√
1-114	12	202	14202		Aggregate Household income in the past 12 months (2008$) (1) by Number of workers in household (6)	Households	6	√		√	√	
1-115	12	203	14203		Aggregate Household income in the past 12 months (2008$) (1) by Vehicles available (6)	Households	6	√		√	√	
1-090	12	204	14204		Mean Household income in the past 12 months (2008$) (1) by Number of workers in household (6)	Households	6	√		√	√	
1-091	12	205	14205		Mean Household income in the past 12 months (2008$) (1) by Vehicles available (6)	Households	6	√		√	√	
1-088	12	206	14206		Median Household income in the past 12 months (2008$) (1) by Number of workers in household (6)	Households	6	√		√	√	

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables (Continued)

					Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables			3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
1-089	12	207	14207		Median Household income in the past 12 months (2008$) (1) by Vehicles available (6)	Households	6	√		√	√	
1-064	12	208	14208		Household size (5) by Household income in the past 12 months (2008$) (26)	Households	130	√		√		√
	12		14208C	c	Household size (5) by Household income in the past 12 months (2008$) (9)	Households	45	√		√	√	
1-062	12	209	14209		Household size (5) by Number of workers in household (6)	Households	30	√		√	√	
NEW	12	210	14210		Household size (5) by Units in Structure (9)	Households	45	√		√	√	
1-063	12	211	14211		Household size (5) by Vehicles available (6)	Households	30	√		√	√	
1-081	12	212	14212		Minority Status of the householder (3) by Telephone availability (3)	Households	9	√		√	√	
1-073	13	213	14213		Poverty status (4) by Telephone availability (3)	Households for which poverty status is determined	12	√		√	√	
1-070	12	214	14214		Vehicles available (6) by Age of householder (8)	Households	48	√		√	√	
1-067	12	215	14215		Vehicles available (6) by Household income in the past 12 months (2008$) (26)	Households	156	√		√		√
	12		14215C	c	Vehicles available (6) by Household income in the past 12 months (2008$) (9)	Households	54	√		√	√	
1-072	12	216	14216		Vehicles available (6) by Length of US residence of the householder (6)	Households	36	√		√	√	
1-080	12	217	14217		Vehicles available (6) by Minority Status of householder (3)	Households	18	√		√	√	
1-068	12	218	14218		Vehicles available (6) by Number of Persons 16 or over in household (6)	Households	36	√		√	√	
1-069	11	219	14219		Vehicles available (6) by Units in structure (7)	Occupied housing units	42	√		√		√
1-071	13	220	14220		Vehicles available (6) by Poverty status (4)	Households for which poverty status is determined	24	√		√	√	
1-066	12	221	14221		Number of Workers in household (6) by Household income in the past 12 months (2008$) (26)	Households	156	√		√		√
	12		14221C	c	Number of Workers in household (6) by Household income in the past 12 months (2008$) (9)	Households	54	√		√	√	
1-065	12	222	14222		Number of Workers in household (6) by Vehicles available (6)	Households	36	√		√	√	
NEW	12	300	14300		Lifecycle of household (16) by Number of Workers in Household (3) by Vehicles available (5)	Households	240	√		√	√	
NEW	12	301	14301		Household size (5) by Household income in the past 12 months (2008$) (5) by Number of Persons Under 18 (4)	Households	100	√		√		√
1-082	12	302	14302		Household size (5) by Household income in the past 12 months (2008$) (9) by Minority Status of householder (3)	Households	135	√		√	√	
1-075	12	303	14303		Household size (5) by Number of workers in household (6) by Household income in the past 12 months (2008$) (9)	Households	270	√		√		√
	12		14303C	c	Household size (5) by Number of workers in household (6) by Household income in the past 12 months (2008$) (5)	Households	150	√		√	√	
NEW	12	304	14304		Household size (5) by Number of workers in household (6) by Number of Persons Under 18 (4)	Households	120	√		√		√
1-074	12	305	14305		Household size (5) by Number of workers in household (6) by Vehicles available (5)	Households	150	√		√	√	
1-076	12	306	14306		Household size (5) by Vehicles available (5) by Household income in the past 12 months (2008$) (9)	Households	225	√		√		√

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 1, Residence Tables (Continued)

					Census Transportation Planning Products (CTPP) Tabulations -- Part 1, Residence-Based Tables			3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
								We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
2000 Table Number	Univ	Number	New Table Number	Collapse	Content	Universe	Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
	12		14306C	c	Household size (5) by Vehicles available (5) by Household income in the past 12 months (2008$) (5)	Households	125	√		√		√
NEW	12	307	14307		Household size (5) by Vehicles available (5) by Number of Persons Under 18 (4)	Households	100	√		√	√	
1-077	12	308	14308		Household size (5) by Vehicles available (5) by Units in structure (7)	Households	175	√		√		√
1-078	12	309	14309		Household size (5) by Vehicles available (5) by Tenure (5)	Households	125	√		√		√
1-079	12	310	14310		Number of Workers in household (6) by Vehicles available (5) by Household income in the past 12 months (2008$) (9)	Households	270	√		√		√
	12		14310C	c	Number of Workers in household (6) by Vehicles available (5) by Household income in the past 12 months (2008$) (5)	Households	150	√		√		√
NEW	12	311	14311		Number of Workers in households (6) by Household income in the past 12 months (2008$) (5) by Number of Persons Under 18 (4)	Households	120	√		√		√
NEW	12	312	14312		Number of Workers in households (6) by Vehicles available (5) by Number of Persons Under 18 (4)	Households	120	√		√		√
1-083	10	100	15100		Total Housing units(1)	All housing units	1	√		√	√	
1-084	10	101	15101		Housing units sampled [total] (1)	All housing units	1	√		√	√	
1-085	10	102	15102		Housing units sampled [percent] (1)	All housing units	1	√		√	√	
NEW	10	103	15103		Units in Structure (9)	All housing units	9	√		√	√	
1-087	10	104	15104		Vacancy status (6)	All housing units	6	√		√	√	
1-086	10	200	15200		Occupancy status (3) by Units in structure (7)	All housing units	21	√		√	√	

Census Transportation Planning Products (CTPP): Standard Tabulations - Part 2, Workplace Tables

Census Transportation Planning Products (CTPP)				3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
Tabulations -- Part 2, Workplace-Based Tables				We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
22100		Total Workers (1)	1	√		√	√	
22101		Age of Worker (10)	10	√		√	√	
22102		Class of worker (9)	9	√		√	√	
22103		Earnings in the past 12 months (2008$) (11)	11	√		√	√	
22104		Industry (15)	15	√		√	√	
22105		Industry (15)	15	√		√	√	
22106		Means of Transportation (18)	18	√		√	√	
22107		Occupation (25)	25	√		√	√	
22109		Workers per car, truck, or van (1)	1	√		√	√	
22110		Disability Status (7)	7		√			
22111		Hispanic Origin (3)	3	√		√	√	
22112		Usual Hours worked per week (7)	7	√		√	√	
22113		Length of US residence (6)	6	√		√	√	
22114		Race (5)	5	√		√	√	
22115		Sex (3)	3	√		√	√	
22116		Time arriving (41)	41	√		√	√	
22118		Travel time to work (12)	12	√		√	√	
22200		Age of Worker (8) by Earnings in the past 12 months (2008$) (11)	88	√		√		√
22201		Age of Worker (8) by Means of transportation (11)	88	√		√	√	
22201C	C	Age of Worker (8) by Means of transportation (7)	56	√		√	√	
22201C2	c2	Age of Worker (8) by Means of transportation (6)	48	√		√	√	
22201C3	c3	Age of Worker (8) by Means of transportation (4)	32	√		√	√	
22202		Aggregate Carpools (1) by Time leaving home (5)	5	√		√	√	
22203		Class of worker (9) by Means of transportation (11)	99		√			
22204		Disability status (7) by Means of transportation (11)	77		√			
22206		Earnings in the past 12 months (2008$) (11) by Travel time (12)	132	√		√		√
22206C	c	Earnings in the past 12 months (2008$) (6) by Travel time (12)	72	√		√	√	
22207		Earnings in 2007$ [aggregate] (1) by Means of transportation (11)	11		√			

Census Transportation Planning Products (CTPP): Standard Tabulations - Part 2, Workplace Tables (Continued)

		Census Transportation Planning Products (CTPP)		3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
		Tabulations -- Part 2, Workplace-Based Tables		We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
22208		Earnings [mean] (1) by Means of transportation (11)	11		√			
22209		Earnings [median] (1) by Means of transportation (11)	11		√			
22210		Hispanic origin (3) by Race (5)	15	√		√	√	
22211		Industry (15) by Earnings in the past 12 months (2008$) (11)	165	√		√		√
22211C	c	Industry (15) by Earnings in the past 12 months (2008$) (6)	90	√		√		√
22212		Industry (15) by Means of transportation to work (11)	165		√			
22214		Length of US residence (6) by Earnings in the past 12 months (2008$) (11)	66	√		√		√
22214C	c	Length of US residence (6) by Earnings in the past 12 months (2008$) (6)	36	√		√	√	
22215		Length of US residence (6) by Means of transportation (11)	66		√			
22216		Minority Status (3) by Class of worker (9)	27	√		√	√	
22217		Minority Status (3) by Earnings in the past 12 months (2008$) (11)	33	√		√	√	
22218		Minority Status (3) by Industry (15)	45	√		√	√	
22219		Minority Status (3) by Means of transportation (11)	33		√			
22220		Minority Status (3) by Occupation (25)	75	√		√	√	
22221		Minority Status (3) by Travel time (18)	54	√		√	√	
22222		Occupation (25) by Industry (15)	375	√		√	√	
22222C	c	Occupation (8) by Industry (15)	120	√		√	√	
22223		Occupation (25) by Means of transportation (11)	275		√			
22225		Sex (3) by Means of transportation (18)	54		√			
22227		Travel time (12) by Means of transportation (11)	132	√		√	√	
22227C	c	Travel time (12) by Means of transportation (7)	84	√		√	√	
22227C2	c2	Travel time (12) by Means of transportation (6)	72	√		√	√	
22227C3	c3	Travel time (12) by Means of transportation (4)	48	√		√	√	

Census Transportation Planning Products (CTPP): Standard Tabulations - Part 2, Workplace Tables (Continued)

		Census Transportation Planning Products (CTPP)		3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
		Tabulations -- Part 2, Workplace-Based Tables		We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
22228C0	c0	Aggregate travel time (1) by Means of transportation (18)	18	√		√	√	
22228		Aggregate travel time (1) by Means of transportation (11)	11	√		√	√	
22228C	c	Aggregate travel time (1) by Means of transportation (7)	7	√		√	√	
22228C2	c2	Aggregate travel time (1) by Means of transportation (6)	6	√		√	√	
22228C3	c3	Aggregate travel time (1) by Means of transportation (4)	4	√		√	√	
22229		Travel time [aggregate] (1) by Means of transportation (18)	18		√			
22230C0	c0	Mean travel time (1) by Means of transportation (18)	18	√		√	√	
22230		Mean travel time (1) by Means of transportation (11)	11	√		√	√	
22230C	c	Mean travel time (1) by Means of transportation (7)	7	√		√	√	
22230C2	c2	Mean travel time (1) by Means of transportation (6)	6	√		√	√	
22230C3	c3	Mean travel time (1) by Means of transportation (4)	4	√		√	√	
22231		Travel time [mean] (1) by Means of transportation (18)	18		√			
22232C0	c0	Median travel time (1) by Means of transportation (18)	18	√		√	√	
22232		Median travel time (1) by Means of transportation (11)	11	√		√	√	
22232C	c	Median travel time (1) by Means of transportation (7)	7	√		√	√	
22232C2	c2	Median travel time (1) by Means of transportation (6)	6	√		√	√	
22232C3	c3	Median travel time (1) by Means of transportation (4)	4	√		√	√	
22233		Travel time [median] (1) by Means of transportation (18)	18		√			
22234		Aggregate vehicles used (1) by Time leaving home (5)	5	√		√	√	
22235		Workers per carpool (1) by Time leaving home (5)	5	√		√	√	
22236		Workers per car, truck, or van (1) by Time leaving home (5)	5	√		√	√	

Census Transportation Planning Products (CTPP): Standard Tabulations - Part 2, Workplace Tables (Continued)

Census Transportation Planning Products (CTPP) Tabulations -- Part 2, Workplace-Based Tables				3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
				We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
22237		Industry (15) by Time arriving (17)	255	√		√		√
22237C	c	Industry (15) by Time arriving (10)	150	√		√	√	
22238		Time arriving (17) by Means of transportation (11)	187	√		√	√	
22238C	c	Time arriving (10) by Means of Transportation (7)	70	√		√	√	
22238C2	c2	Time arriving (10) by Means of transportation (6)	60	√		√	√	
22238C3	c3	Time arriving (10) by Means of Transportation (4)	40	√		√	√	
22239		Occupation (25) by Time arriving (17)	425	√		√		√
22239C	c	Occupation (8) by Time arriving (10)	80	√		√	√	
22240		Linguistic Isolation (3) by Language spoken at home(12)	36	√		√	√	
22303		Aggregate travel time (1) by Means of transportation (11) by Time arriving (17)	187	√		√	√	
22303C	c	Aggregate travel time (1) by Means of transportation (7) by Time arriving (10)	70	√		√		√
22303C2	c2	Aggregate travel time (1) by Means of transportation (6) by Time arriving (10)	60	√		√	√	
22303C3	c3	Aggregate travel time (1) by Means of transportation (4) by Time arriving (10)	40	√		√	√	
22304		Mean travel time (1) by Means of transportation (11) by Time arriving (17)	187	√		√	√	
22304C	c	Mean travel time (1) by Means of transportation (7) by Time arriving (10)	70	√		√		√
22304C2	c2	Mean travel time (1) by Means of transportation (6) by Time arriving (10)	60	√		√	√	
22304C3	c3	Mean travel time (1) by Means of transportation (4) by Time arriving (10)	40	√		√	√	
22305		Median travel time (1) by Means of transportation (11) by Time arriving (17)	187	√		√	√	

Census Transportation Planning Products (CTPP): Standard Tabulations - Part 2, Workplace Tables (Continued)

Census Transportation Planning Products (CTPP)				3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
Tabulations -- Part 2, Workplace-Based Tables				We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
22305C	c	Median travel time (1) by Means of transportation (7) by Time arriving (10)	70	√		√		√
22305C2	c2	Median travel time (1) by Means of transportation (6) by Time arriving (10)	60	√		√	√	
22305C3	c3	Median travel time (1) by Means of transportation (4) by Time arriving (10)	40	√		√	√	
23100		Total Workers in households (1)	1	√		√	√	
23101		Household income in the past 12 months (2008$) (26)	26	√		√	√	
23102		Vehicles Available (6)	6	√		√	√	
23103		Vehicles per household [average] (1)	1	√		√	√	
23200		Age group of youngest child in the household (5) by Means of transportation (11)	55		√			
23204		Household income in the past 12 months (2008$) (26) by Means of transportation (11)	286	√		√	√	
23204C	c	Household income in the past 12 months (2008$) (16) by Means of transportation (7)	112	√		√		√
23204C2	c2	Household income in the past 12 months (2008$) (16) by Means of transportation (6)	96	√		√	√	
23204C3	c3	Household income in the past 12 months (2008$) (16) by Means of transportation (4)	64	√		√	√	
23205		Household income in the past 12 months (2008$) (9) by Earnings in the past 12 months (2008$) (11)	99	√		√		√
23205C	c	Household income in the past 12 months (2008$) (9) by Earnings in the past 12 months (2008$) (6)	54	√		√		√
23206		Poverty status (4) by Means of transportation (11)	44		√			
23208		Vehicle availabilty [ratio] (4) by Length of US residence (6)	24		√			
23209		Vehicle availability [ratio] (4) by Means of transportation (11)	44		√			
23210		Vehicles available (6) by Length of US residence (6)	36	√		√	√	
23211		Vehicles available (6) by Means of transportation (11)	66	√		√	√	
23211C	c	Vehicles available (6) by Means of transportation (7)	42	√		√	√	
23211C2	c2	Vehicles available (6) by Means of transportation (6)	36	√		√	√	
23211C3	c3	Vehicles available (6) by Means of transportation (4)	24	√		√	√	
23212		Vehicles available (6) by Poverty status (4)	24	√		√	√	
23213		Vehicles available [ratio] (4) by Poverty status (4)	16		√			

Census Transportation Planning Products (CTPP): Standard Tabulations - Part 2, Workplace Tables (Continued)

		Census Transportation Planning Products (CTPP)		3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
		Tabulations -- Part 2, Workplace-Based Tables		We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5-year large area data	Proposed request for 5-year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
23214		Number of Workers in household (6) by Household size (5)	30	√		√	√	
23215		Workers in household (6) by Means of transportation (11)	66		√			
23216		Poverty status (4) by Time arriving (17)	68	√		√	√	
23300		Household income in 2007$ (5) by Minority Status (3) by Means of transportation (8)	120		√			
23301		Household income in the past 12 months (2008$) (5) by Number of Workers in household (6) by Age group of youngest child in the household (5)	150	√		√		√
23302		Household income in 2007$ (5) by Vehicles available (5) by Means of transportation (8)	200		√			
23303		Poverty status (4) by Minority Status(3) by Means of transportation (8)	96		√			
23304		Vehicles availability [ratio] (4) by Minority Status (3) by Means of transportation (8)	96		√			
23305		Vehicles available (3) by Minority Status (3) by Means of transportation (8)	72		√			

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 3, Worker Flow Table

Census Transportation Planning Products (CTPP)					3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
4/29/09	Tabulations -- Part 3, Worker Home-to-Work Flow Tables				We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	Universe	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5 year large area data	Proposed request for 5 year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
32100		Total Workers (1)	Workers 16 years and over	1	√		√	√	
32101		Age of Worker (8)	Workers 16 years and over	8	√		√	√	
32104		Industry (15)	Workers 16 years and over	15	√		√	√	
32105		Industry (15)	Workers 16 years and over who are not self-employed	15	√		√	√	
32106		Means of transportation (18)	Workers 16 years and over	18	√		√	√	
32108		Time leaving home to go to work (17)	Workers 16 years and over	17	√		√	√	
32117		Minority status (3)	Workers 16 years and over	3	√		√	√	
32118		Travel time to work (12)	Workers 16 years and over	12	√		√	√	
32201		Age of Worker (6) by Means of transportation (7)	Workers 16 years and over	42	√		√		√
32201	c	Age of Worker (6) by Means of transportation (4)	Workers 16 years and over	24	√		√		√
32202		Aggregate Carpools (1) by Time leaving home (5)	Workers 16 years and over who carpooled	5	√		√		√
32226		Time leaving home (5) by Means of transportation (7)	Workers 16 years and over	35	√		√	√	
32226	c	Time leaving home (5) by Means of transportation (4)	Workers 16 years and over	20	√		√		√
32227		Travel time (12) by Means of transportation (7)	Workers 16 years and over	84	√		√		√
32227	c	Travel time (12) by Means of transportation (4)	Workers 16 years and over	48	√		√	√	
32228		Aggregate Travel time (1) by Means of transportation (7)	Workers 16 years and over who did not work at home	7	√		√	√	
32228	c	Aggregate Travel time (1) by Means of transportation (4)	Workers 16 years and over who did not work at home	4	√		√	√	
32230		Mean Travel time (1) by Means of transportation (7)	Workers 16 years and over who did not work at home	7	√		√	√	
32230	c	Mean Travel time (1) by Means of transportation (4)	Workers 16 years and over who did not work at home	4	√		√	√	
32232		Median Travel time (1) by Means of transportation (7)	Workers 16 years and over who did not work at home	7	√		√	√	
32232	c	Median Travel time (1) by Means of transportation (4)	Workers 16 years and over who did not work at home	4	√		√	√	
32234		Aggregate Vehicles used (1) by Time leaving home (5)	Workers 16 years and over who used car, truck or van	5	√		√		√

Census Transportation Planning Products (CTPP): Standard Tabulations -- Part 3, Worker Flow Table (Continued)

Census Transportation Planning Products (CTPP)					3 Year Tables based on ACS 2006 - 2008 data		5 Year tables based on ACS 2006 - 2010 data		
4/29/09	Tabulations -- Part 3, Worker Home-to-Work Flow Tables				We are getting	We are not getting	We are asking for	We are asking for	We are not asking for
ACSO Table Number	Collapse	Content	Universe	# of Cells	requested and approved for 3-yr table	requested tables not obtainable due to 5 variable crosstab with Means of tansportation requirement (and disability question change)	Proposed request for 5 year large area data	Proposed request for 5 year small area data - (tract, TAZ, Block Group)	Proposed to not be requested for 5-year small area data - (tract, TAZ, Block Group)
32235		Workers per carpool (1) by Time leaving home (5)	Workers 16 years and over who carpooled	5	√		√		√
32236		Workers per car, truck, or van (1) by Time leaving home (5)	Workers 16 years and over who used car, truck or van	5	√		√		√
32300		Aggregate Travel time (1) by Means of transportation (7) and Time leaving home (5)	Workers 16 years and over who did not work at home	35	√		√		√
32300	c	Aggregate Travel time (1) by Means of transportation (4) and Time leaving home (5)	Workers 16 years and over who did not work at home	20	√		√		√
32301		Mean Travel time (1) by Means of transportation (7) and Time leaving home (5)	Workers 16 years and over who did not work at home	35	√		√		√
32301	c	Mean Travel time (1) by Means of transportation (4) and Time leaving home (5)	Workers 16 years and over who did not work at home	20	√		√		√
32302		Median Travel time (1) by Means of transportation (7) and Time leaving home (5)	Workers 16 years and over who did not work at home	35	√		√		√
32302	c	Median Travel Time (1) by Means of transportation (4) and Time leaving home (5)	Workers 16 years and over who did not work at home	20	√		√		√
32306		Travel time [mean] (1) by Means of transportation (5) and Time leaving home (4)	All workers	20		√			
32307		Travel time [median] (1) by Means of transportation (5) and Time leaving home (4)	All workers	20		√			
33100		Household income in past 12 months (2008$) (9)	Workers 16 years and over in households	9	√		√	√	
33103		Poverty status (4)	Workers 16 years and over for whom poverty status is determined	4	√		√	√	
33204		Household income in past 12 months (2008$) (5) by Means of transportation (7)	Workers 16 years and over in households	35	√		√	√	
33204	c	Household income in past 12 months (2008$) (5) by Means of transportation (4)	Workers 16 years and over in households	20	√		√	√	
33206		Poverty status (4) by Means of transportation (5)	Workers in households	20		√			
33208		Vehicle availability [ratio] (4) by Means of transportation (8)	Workers in households	32		√			
33211		Vehicles available (4) by Means of transportation (7)	Workers 16 years and over in households	28	√		√	√	
33211	c	Vehicles available (4) by Means of transportation (4)	Workers 16 years and over in households	16	√		√	√	

Appendix B

TRAFFIC ANALYSIS ZONE ESTIMATES FROM 2006-2008 ACS DATA, BY STATE

Traffic Analysis Zone Estimates from 2006-2008 ACS Data, By State

		Employment status; Population 16 years and over			www.census.gov/geo/www/tallies/vtdtally.html					
	Total Population Estimate	Estimate	In labor force; Estimate	Commuting to work; Workers 16 years and over Estimate	Census 2000 Number of Traffic Analysis Zones (TAZs)	Avg. Total Population per TAZ	Avg. Number of Workers 16+ per TAZ	Census 2000 Number of Block Groups	Avg. Total Population per Block Group	TAZ Count to Block Group Count Ratio (Higher Ratios Shaded More)
AL	4,625,354	3,635,377	2,204,041	1,999,506	3,416	1,354	585	3,329	1,389	1.03
AK	681,235	522,829	377,854	329,538	532	1,281	619	533	1,278	1.00
AS	2,830,047	2,211,056	1,350,667	1,225,650	1,732	1,634	708	2,135	1,326	0.81
AZ	6,343,952	4,852,103	3,050,473	2,803,030	2,195	2,890	1,277	3,570	1,777	0.61
CA	36,418,499	28,139,366	18,228,215	16,450,620	10,434	3,490	1,577	22,133	1,645	0.47
CO	4,844,568	3,784,783	2,659,963	2,457,317	3,929	1,233	625	3,278	1,478	1.20
CT	3,493,006	2,773,920	1,888,489	1,718,856	2,030	1,721	847	2,620	1,333	0.77
DE	861,804	680,599	444,132	406,645	MPO(s) did not participate			502	1,717	
DC	588,373	488,257	329,224	293,532	458	1,285	641	433	1,359	1.06
FL	18,182,321	14,638,681	8,969,628	8,171,223	12,738	1,427	641	9,112	1,995	1.40
GA	9,509,254	7,281,160	4,823,154	4,365,767	4,380	2,171	997	4,788	1,986	0.91
HI	1,280,273	1,027,471	682,982	637,084	MPO(s) did not participate			646	1,982	
ID	1,493,713	1,133,040	750,929	692,634	596	2,506	1,162	954	1,566	0.62
IL	12,829,014	10,014,013	6,704,699	6,056,254	9,432	1,360	642	9,843	1,303	0.96
IN	6,335,595	4,936,646	3,271,371	2,973,627	4,597	1,378	647	4,798	1,320	0.96
IA	2,984,391	2,358,232	1,637,615	1,523,848	2,201	1,356	692	2,634	1,133	0.84
KS	2,778,599	2,161,087	1,500,519	1,393,055	1,425	1,950	978	2,299	1,209	0.62
KY	4,234,999	3,347,494	2,047,529	1,857,122	5,871	721	316	3,157	1,341	1.86
LA	4,342,582	3,367,667	2,065,757	1,868,106	2,752	1,578	679	3,509	1,238	0.78
ME	1,315,069	1,072,844	705,001	644,540	1,504	874	429	1,143	1,151	1.32
MD	5,618,250	4,431,566	3,077,648	2,831,245	3,719	1,511	761	3,678	1,528	1.01
MA	6,469,770	5,209,389	3,528,670	3,222,621	558	11,595	5,775	5,053	1,280	0.11
MI	10,045,697	7,915,701	5,042,854	4,429,683	6,438	1,560	688	8,450	1,189	0.76
MN	5,181,962	4,077,063	2,916,785	2,702,896	2,317	2,236	1,167	4,082	1,269	0.57
MS	2,918,790	2,245,668	1,350,038	1,198,616	3,404	857	352	2,148	1,359	1.58
MO	5,874,327	4,621,185	3,034,581	2,773,582	2,409	2,438	1,151	4,540	1,294	0.53
MT	956,496	764,023	500,626	465,075	MPO(s) did not participate			874	1,094	
NE	1,770,896	1,377,678	984,520	919,180	1,675	1,057	549	1,591	1,113	1.05
NV	2,546,235	1,959,902	1,331,314	1,221,923	1,618	1,574	755	1,245	2,045	1.30

		Employment status; Population 16 years and over			www.census.gov/geo/www/tallies/vtdtally.html					
	Total Population Estimate	Estimate	In labor force; Estimate	Commuting to work; Workers 16 years and over Estimate	Census 2000 Number of Traffic Analysis Zones (TAZs)	Avg. Total Population per TAZ	Avg. Number of Workers 16+ per TAZ	Census 2000 Number of Block Groups	Avg. Total Population per Block Group	TAZ Count to Block Group Count Ratio (Higher Ratios Shaded More)
NH	1,312,298	1,053,594	740,747	686,366	1,132	1,159	606	874	1,501	1.30
NJ	8,658,668	6,841,756	4,561,929	4,159,120	1,411	6,137	2,948	6,510	1,330	0.22
NM	1,962,226	1,520,199	951,391	870,476	1,183	1,659	736	1,413	1,389	0.84
NY	19,428,881	15,525,409	9,858,485	8,958,424	4,126	4,709	2,171	15,079	1,288	0.27
NC	9,036,449	7,084,691	4,637,554	4,209,185	8,167	1,106	515	5,271	1,714	1.55
ND	638,613	513,248	358,541	338,845	549	1,163	617	630	1,014	0.87
OH	11,473,983	9,052,105	5,916,716	5,351,380	7,448	1,541	718	9,354	1,227	0.80
OK	3,606,200	2,811,023	1,780,714	1,644,051	1,620	2,226	1,015	2,901	1,243	0.56
OR	3,735,524	2,974,524	1,934,054	1,756,663	548	6,817	3,206	2,490	1,500	0.22
PA	12,418,756	9,985,133	6,315,780	5,782,102	4,895	2,537	1,181	10,387	1,196	0.47
RI	1,054,306	851,684	563,214	506,607	690	1,528	734	821	1,284	0.84
SC	4,403,175	3,472,645	2,185,385	1,973,349	5,175	851	381	2,859	1,540	1.81
SD	795,757	623,583	437,306	410,379	581	1,370	706	688	1,157	0.84
TN	6,144,104	4,845,792	3,079,701	2,780,135	1,692	3,631	1,643	4,014	1,531	0.42
TX	23,845,989	17,948,995	11,819,368	10,832,598	17,674	1,349	613	14,463	1,649	1.22
UT	2,663,500	1,921,988	1,336,789	1,252,752	1,358	1,961	922	1,481	1,798	0.92
VT	620,738	507,974	353,271	326,610	157	3,954	2,080	530	1,171	0.30
VA	7,698,738	6,089,016	4,117,421	3,822,594	4,688	1,642	815	4,749	1,621	0.99
WA	6,453,083	5,101,960	3,391,636	3,102,055	3,061	2,108	1,013	4,825	1,337	0.63
WV	1,810,358	1,470,138	819,510	746,145	1,969	919	379	1,588	1,140	1.24
WI	5,598,453	4,442,488	3,080,419	2,846,764	6,001	933	474	4,388	1,276	1.37
WY	522,833	412,009	290,593	272,209	262	1,996	1,039	398	1,314	0.66
PR	3,940,626	3,060,016	1,439,762	1,168,146						
Total w/o DE,HI, MT,PR	298,139,130	233,606,661	152,362,062	138,752,780	166,747	1,788	832	206,768	1,442	0.81

Appendix C

SET B TABLES

Part 1:	Residence
Part 2:	Workplace
Part 3:	Flows

PART 1: RESIDENCE

Table C-1. Part 1: Cell Means and Aggregates

Table #	Table Variable	Attribute	Development Phase	Validation Phase	Final Production
12202	TIME LEAVING HOME(5)	CARPOOLS	√	√	√
12228	MOT(18)	TRAVEL TIME	√	√	√
12234	TIME LEAVING HOME(5)	VEHICLES USED	√	√	√
12300	MOT(11)* TIME LEAVING HOME(17)	TRAVEL TIME	√	√	√
14102	--	HH INCOME	√	√	√
14106	--	PERSONS PER HH	√	√	√
14111	--	VEHICLES	√	√	√
14202	HH WORKERS(6)	HH INCOME	√	√	√
14203	VEHICLES(6)	HH INCOME	√	√	√

Note: only table #s for cell aggregates are shown.

Table C-2. Part 1: Cell Counts

Table #	Table Variable 1	Table Variable 2	Comment	Development Phase	Validation Phase	Final Production
12201	MOT(11)	AGE OF WORKER(8)		√	√	√
12226C	MOT(7)	TIME LEAVING HOME(10)		√	√	√
12226	MOT(11)	TIME LEAVING HOME(17)	Large geography only	√	√	√
12227	MOT(11)	TRAVEL TIME(12)		√	√	√
13204	MOT(11)	HH INCOME(26)		√	√	√
13211	MOT(11)	VEHICLES(6)		√	√	√
	MOT(11)	PRESENCE OF CHILDREN IN HH(3)			√	√
	MOT(11)	HH WORKERS(3)			√	√
	MOT(11)	MINORITY(3)			√	√

Note: large geography is county-level or higher

PART 2: WORKPLACE

Table C-3. Part 2: Cell Means and Aggregates

Table #	Table Variable	Attribute	Development Phase	Validation Phase	Final Production
22202	TIME LEAVING HOME(5)	CARPOOLS	√	√	√
22234	TIME LEAVING HOME(5)	VEHICLES USED	√	√	√
22228	MOT(18)	TRAVEL TIME	√	√	√
22303	MOT(11)* TIME ARRIVING(17)	TRAVEL TIME	√	√	√

Note: only table #s for cell aggregates are shown.

Table C-4. Part 2: Cell Counts

Table #	Table Variable 1	Table Variable 2	Development Phase	Validation Phase	Final Production
22201	MOT(11)	AGE OF WORKER(8)	√	√	√
22227	MOT(11)	TRAVEL TIME(12)	√	√	√
22238	MOT(11)	TIME ARRIVING(17)	√	√	√
23204	MOT(11)	HH INCOME(26)	√	√	√
23211	MOT(11)	VEHICLES(6)	√	√	√
	MOT(11)	PRESENCE OF CHILDREN IN HH (3)		√	√
	MOT(11)	HH WORKERS(3)		√	√
	MOT(11)	MINORITY(3)		√	√

PART 3: FLOWS

Table C-5. Part 3: Cell Means and Aggregates

Table #	Table Variable	Attribute	Comment	Development Phase	Validation Phase	Final Production
32202	TIME LEAVING HOME(5)	CARPOOLS	Large geography only	√	√	√
32228	MOT(7)	TRAVEL TIME		√	√	√
32234	TIME LEAVING HOME(5)	VEHICLES USED	Large geography only	√	√	√
32235	TIME LEAVING HOME(5)	WORKERS PER CARPOOL	Large geography only	√	√	√
32236	TIME LEAVING HOME(5)	WORKERS PER TRUCK, CAR, VAN	Large geography only	√	√	√
32300	MOT(7)* TIME LEAVING HOME(5)	TRAVEL TIME		√	√	√

Note: only table #s for cell aggregates are shown. Large geography is county-level or higher.

Table C-6. Part 3: Cell Counts

Table #	Table Variable 1	Table Variable 2	Comment	Development Phase	Validation Phase	Final Production
32101	AGE OF WORKER(8)			√	√	√
32104	INDUSTRY(8)			√	√	√
32105	INDUSTRY(8)		Workers 16 years and over who are not self-employed (Large geography only)	√	√	√
32108	TIME LEAVING HOME(17)			√	√	√
32117	MINORITY STATUS(3)			√	√	√
32118	TRAVEL TIME(12)			√	√	√
32201	MOT(7)	AGE OF WORKER(6)	Large geography only	√	√	√
32226	MOT(7)	TIME LEAVING HOME(5)		√	√	√
32227	MOT(4)	TRAVEL TIME(12)	MOT(7) for large geography	√	√	√
33100	HH INCOME(9)			√	√	√
33103	POVERTY STATUS(4)			√	√	√
33204	MOT(7)	HH INCOME(5)		√	√	√
33211	MOT(7)	VEHICLES(4)		√	√	√

Note: large geography is county-level or higher

Appendix D

SET OF PREDICTOR VARIABLES

Item	Variable Name	Variable Description	Level at Which Variable is Defined	Development Phase	Validation Phase	Final Production
1	AHINC	HH income	HH	X	X	X
2	APERN	Person earnings	Person	X	X	X
3	JWD_SHIFT	Work shift	Person	X	X	X
4	AGE9	Age	Person	X	X	X
5	MINORITY	Minority status	Person	X	X	X
6	SEX	Sex	Person	X	X	X
7	INDUSTRY_8	Industry	Person	X	X	X
8	OCCUPATION_7	Occupation	Person	X	X	X
9	YRS_US	Years of US residence	Person	X	X	X
10	YNGEST	Age of youngest child	HH	X	X	X
11	HH_WRK6	Number of HH workers	HH	X	X	X
12	VEHICLES6	Vehicles available	HH	X	X	X
13	HHLDRAGE	Householder age	HH	X	X	X
14	MINORITY_HH	Householder minority status	HH	X	X	X
15	YRSUS_HH	Householder years of US residence	HH	X	X	X
16	COW	Class of worker	Person	X	X	X
17	AVG_HHSIZE	Average HH size	CTAZ	X	X	X
18	AVG_WIH	Average number of workers in HH	CTAZ	X	X	X
19	AVG_VEH	Average vehicles available	CTAZ	X	X	X
20	C_PCTBLK	Percentage of population who are Black	Block	X	X	X
21	C_PCTHISP	Percentage of population who are Hispanic	Block	X	X	X
22	C_PCTOOCC	Percentage owner occupied	Block	X	X	X
23	MED_APERN	Median person earnings	CTAZ	X	X	X
24	MED_HHINC	Median HH income	CTAZ	X	X	X
25	PCT_BLK	Percentage of workers who are Black	CTAZ	X	X	X
26	PCT_COLL	Percentage of workers who are college graduates	CTAZ	X	X	X
27	PCT_HIGH	Percentage of workers with a high school diploma	CTAZ	X	X	X

Item	Variable Name	Variable Description	Level at Which Variable is Defined	Development Phase	Validation Phase	Final Production
28	PCT_HIS	Percentage of workers who are Hispanic	CTAZ	X	X	X
29	PCT_MAR	Percentage of workers who are married	CTAZ	X	X	X
30	PCT_POV	Percentage of workers in poverty	CTAZ	X	X	X
31	PCT_RENT	Percentage of workers who rent	CTAZ	X	X	X
32	PCT_PHONE	Percentage of workers with a phone line	CTAZ	X	X	X
33	PCT_UNDER 18	Percentage of workers under 18 years of age	CTAZ	X	X	X
34	PCT_WHT	Percentage of workers who are White	CTAZ	X	X	X
35	PCI	Principal city indicator	Person	X	X	X
36	UNDER18	Flag if under 18 years old	Person	X	X	X
37	UR	Urban rural indicator	Person	X	X	X
38	R18	Presence of persons < 18 years old	Person	X	X	X
39	NEW_HHT	Household family type recode	HH	X	X	X
40	MODE	Mode of data collection	HH	X	X	X
41	NOC	Number of own children	HH	X	X	X
42	NEW_NPF	Number of persons in family recode	HH	X	X	X
43	HH_OVER16 YRS	Number of persons 16 years old or older in HH	HH	X	X	X
44	HH_SIZE	HH size	HH	X	X	X
45	LNGI	Language indicator	Person	X	X	X
46	TEL	Telephone indicator	HH	X	X	X
47	BORNUSAH H	Householder country of birth	HH	X	X	X
48	NEW_HHLDS CHL	Householder education attainment recode	HH	X	X	X
49	HPOV	Household poverty index	HH	X	X	X

Item	Variable Name	Variable Description	Level at Which Variable is Defined	Development Phase	Validation Phase	Final Production
50	R60	Presence of persons > 60 years old	HH	X	X	X
51	RMS	Number of rooms	HH	X	X	X
52	HH_LIFE	Householder life cycle	HH	X	X	X
53	HHLDRMOT	Householder means of transportation	HH	X	X	X
54	STRUCTURE9	Household structure	HH	X	X	X
55	TEN	Tenure	HH	X	X	X
56	WKH	Hours worked per week	Person	X	X	X
57	WKW	Week worked in past 12 months	Person	X	X	X
58	ENROLLMENT	Enrollment status	Person	X	X	X
59	USUAL_HRS	Usual number of hours worked	Person	X	X	X
60	BORNUSA	Country of birth	Person	X	X	X
61	NEW_MIL	Military status recode	Person	X	X	X
62	MEANS11	Means of transportation	Person	X	X	X
63	ESR	Employment status	Person	X	X	X
64	NEW_MAR	Marital status recode	Person	X	X	X
65	NEW_MIG	Migration status recode	Person	X	X	X
66	NEW_POWPCI	Place of work principal city indicator recode	Person	X	X	X
67	NEW_VETSTAT	Veteran status recode	Person	X	X	X
68	NEW_JWMN	Travel time recode	Person	X	X	X
69	NEW_HHLDRHIS	Householder Ethnicity recode	Person	X	X	X
70	NEW_HHLDRACE	Householder race recode	Person	X	X	X
71	NEW_POVERTY	Poverty recode	Person			X
72	MOT*AHINC	MOT interaction with HH income	HH	X	X	X
73	MOT*APERN	MOT interaction with person earnings	Person	X	X	X
74	MOT*AGE9	MOT interaction with age categories	Person	X	X	X
75	MOT*MINORITY	MOT interaction with minority status	Person	X	X	X

Item	Variable Name	Variable Description	Level at Which Variable is Defined	Development Phase	Validation Phase	Final Production
76	MOT*SEX	MOT interaction with sex	Person	X	X	X
77	MOT*HH_WRK6	MOT interaction with number of workers in HH	HH	X	X	X
78	MOT*VEHICLES6	MOT interaction with vehicles available	HH	X	X	X
79	MOT*BORNUSA	MOT interaction with country of birth	Person	X	X	X
80	MOT*NEW_JWMN	MOT interaction with travel time	Person	X	X	X
81	MOT*NEW_POVERTY	MOT interaction with poverty status	Person	X	X	X
82	MOT*UNDER18	MOT interaction with under 18 years old	HH		X	X
83	MOT*HHLDRAGE	MOT interaction with householder's age	HH		X	X
84	MOT*MINORITY_HH	MOT interaction with householder's minority status	HH		X	X
85	MOT*R18	MOT interaction with Presence of persons < 18 years old	HH		X	X
86	MOT*BORNUSAHH	MOT interaction with householder's country of birth	HH		X	X
87	GQ	Group quarter indicator	Person	X		
88	DIR	Direction to work	Block	X		
89	FLOW_DIST	Derived distance of flow	Block	X		
90	DISTANCE2	Derived distance to work for MOT2	Block	X		
91	DISTANCE3	Derived distance to work for MOT3	Block	X		
92	DISTANCE4	Derived distance to work for MOT4	Block	X		
93	DISTANCE5	Derived distance to work for MOT5	Block	X		
94	DISTANCE6	Derived distance to work for MOT6	Block	X		
95	DISTANCE7	Derived distance to work for MOT7	Block	X		
96	DISTANCE8	Derived distance to work for MOT8	Block	X		
97	DISTANCE9	Derived distance to work for MOT9	Block	X		

Item	Variable Name	Variable Description	Level at Which Variable is Defined	Development Phase	Validation Phase	Final Production
98	DISTANCE10	Derived distance to work for MOT10	Block	X		
99	TRAVELER	Outlier commute flag	Block	X		
100	NEW_GQT	Group quarter status recode	Person	X		

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

Appendix E

DEVELOPMENT PHASE: TABULAR SUMMARY RESULTS FOR THE DATA UTILITY MEASURES

Table E-1. Median and Interquartile Range of Cell Mean Differences, TAZ Level, Full Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**Attribute**	**BYVAR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Parametric	AHINC	VEHICLES6_2	0.0	11263.6	0.0	11600.6	0.0	10685.2	0.0	11661.2
Parametric	AHINC	VEHICLES6_3	0.0	10882.0	0.0	11477.3	694.2	15399.3	342.3	9906.2
Parametric	AHINC	VEHICLES6_4	-1686.2	15237.5	-475.5	16858.9	-357.3	25041.1	-1269.3	14121.2
Parametric	AHINC	VEHICLES6_5	0.0	24985.0	0.0	24902.0	0.0	37522.1	0.0	24571.8
Parametric	AHINC	VEHICLES6_6	0.0	29809.0	378.1	28683.1	0.0	41303.3	45.0	34668.4
Parametric	JWMN	MEANS6_2	1.3	5.5	0.0	5.8	0.1	7.8	0.7	5.4
Parametric	JWMN	MEANS6_3	0.9	11.7	0.0	10.2	0.0	13.0	0.0	13.1
Parametric	JWMN	MEANS6_4	0.3	12.8	0.0	14.5	0.0	14.5	0.0	16.7
Parametric	JWMN	MEANS6_5	2.1	13.0	0.7	10.6	0.2	11.5	1.4	15.0
Parametric	JWMN	TM_LEAVE5_3	0.9	6.8	--0.0	6.8	-0.0	9.3	0.5	6.8
Parametric	JWMN	TM_LEAVE5_4	4.1	10.1	0.5	8.8	1.7	11.3	2.9	9.4
Semi-parametric	AHINC	VEHICLES6_2	0.0	7318.9	0.0	6595.3	0.0	11026.3	0.0	6888.1
Semi-parametric	AHINC	VEHICLES6_3	476.9	8596.6	1426.6	9149.1	908.5	12841.8	540.0	7986.5
Semi-parametric	AHINC	VEHICLES6_4	110.5	12917.2	-89.3	13654.6	0.0	24753.0	-68.1	12996.3
Semi-parametric	AHINC	VEHICLES6_5	0.0	20584.4	0.0	19800.7	-205.7	33911.9	0.0	19869.2
Semi-parametric	AHINC	VEHICLES6_6	0.0	23336.5	0.0	21327.7	0.0	40517.9	0.0	26129.4
Semi-parametric	JWMN	MEANS6_2	0.2	4.3	0.1	4.4	0.3	5.2	0.0	4.5
Semi-parametric	JWMN	MEANS6_3	0.0	9.3	0.3	7.5	0.0	8.9	0.5	8.1
Semi-parametric	JWMN	MEANS6_4	0.0	8.2	0.1	9.7	0.0	8.6	0.0	10.1
Semi-parametric	JWMN	MEANS6_5	0.6	10.0	0.0	6.2	0.3	9.2	0.1	9.9
Semi-parametric	JWMN	TM_LEAVE5_3	-0.2	5.7	0.0	5.0	0.0	6.5	-0.1	5.6
Semi-parametric	JWMN	TM_LEAVE5_4	2.4	9.8	0.6	7.6	1.4	8.6	1.5	9.2
Constrained hotdeck	AHINC	VEHICLES6_2	0.0	3770.0	0.0	4161.5	0.0	3113.3	0.0	4104.6
Constrained hotdeck	AHINC	VEHICLES6_3	62.9	3178.8	0.0	3501.9	0.0	4097.1	109.3	3042.2
Constrained hotdeck	AHINC	VEHICLES6_4	108.3	5115.1	0.0	5038.3	0.0	7592.5	0.8	4293.6
Constrained hotdeck	AHINC	VEHICLES6_5	0.0	7813.2	0.0	7628.4	0.0	10315.5	0.0	7572.8
Constrained hotdeck	AHINC	VEHICLES6_6	0.0	8529.0	0.0	8205.2	0.0	9367.0	0.0	10339.6
Constrained hotdeck	JWMN	MEANS6_2	0.0	2.1	-0.2	2.6	0.0	3.1	0.0	2.0
Constrained hotdeck	JWMN	MEANS6_3	0.0	4.9	0.0	5.6	0.0	5.9	0.0	5.4
Constrained hotdeck	JWMN	MEANS6_4	0.0	6.4	0.0	7.2	0.0	9.0	0.0	6.1
Constrained hotdeck	JWMN	MEANS6_5	0.0	6.4	0.0	5.4	0.0	5.2	0.0	5.6
Constrained hotdeck	JWMN	TM_LEAVE5_3	0.0	2.5	-0.3	2.8	0.0	3.4	0.0	2.3
Constrained hotdeck	JWMN	TM_LEAVE5_4	0.1	3.6	0.0	4.1	0.0	4.6	0.2	3.5

Table E-2. Median and Interquartile Range of Cell Mean Differences, TAZ Level, Partial Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Parametric	AHINC	VEHICLES6_2	0.0	6284.8	0.0	5391.0	0.0	9939.0	0.0	4332.2
Parametric	AHINC	VEHICLES6_3	0.0	7575.4	0.0	7399.4	0.0	12713.4	0.0	7089.8
Parametric	AHINC	VEHICLES6_4	-651.6	11891.6	-32.2	13209.5	101.4	23353.5	-746.9	12378.9
Parametric	AHINC	VEHICLES6_5	0.0	21036.2	0.0	20329.2	0.0	33329.2	0.0	22835.1
Parametric	AHINC	VEHICLES6_6	0.0	24785.8	52.7	24871.6	0.0	43190.7	0.0	30339.8
Parametric	JWMN	MEANS6_2	0.7	5.2	-0.2	5.5	0.0	7.6	0.3	5.3
Parametric	JWMN	MEANS6_3	0.7	10.3	0.0	9.5	0.5	13.4	0.0	12.2
Parametric	JWMN	MEANS6_4	0.9	12.8	0.0	12.6	0.0	13.6	0.5	16.3
Parametric	JWMN	MEANS6_5	1.0	12.5	0.5	10.9	0.0	12.9	1.0	13.4
Parametric	JWMN	TM_LEAVE5_3	0.5	5.9	0.0	5.8	-0.1	8.4	0.4	6.2
Parametric	JWMN	TM_LEAVE5_4	2.4	8.6	0.0	7.9	0.6	9.5	1.3	8.4
Semi-parametric	AHINC	VEHICLES6_2	0.0	3296.4	0.0	1979.8	0.0	6877.5	0.0	2292.0
Semi-parametric	AHINC	VEHICLES6_3	174.5	5697.6	100.2	5176.8	585.8	10968.0	262.3	4956.3
Semi-parametric	AHINC	VEHICLES6_4	69.7	10300.2	0.0	10574.9	140.3	20724.4	-103.5	9742.2
Semi-parametric	AHINC	VEHICLES6_5	-18.4	16232.4	0.0	15706.4	-525.5	29599.5	0.0	17904.9
Semi-parametric	AHINC	VEHICLES6_6	0.0	19457.7	0.0	16317.6	0.0	34864.5	0.0	21615.6
Semi-parametric	JWMN	MEANS6_2	0.1	4.5	0.0	3.9	0.0	5.6	-0.1	3.9
Semi-parametric	JWMN	MEANS6_3	0.0	7.4	0.0	7.0	0.0	8.5	0.3	8.1
Semi-parametric	JWMN	MEANS6_4	0.0	8.6	0.1	9.8	0.0	10.0	0.0	10.0
Semi-parametric	JWMN	MEANS6_5	0.1	9.4	0.0	5.9	0.0	9.4	0.0	10.5
Semi-parametric	JWMN	TM_LEAVE5_3	0.1	4.8	0.1	4.4	0.1	5.9	0.2	4.3
Semi-parametric	JWMN	TM_LEAVE5_4	0.3	6.7	0.0	5.5	0.0	7.0	0.0	6.1
Constrained hotdeck	AHINC	VEHICLES6_2	0.0	1733.9	0.0	1675.5	0.0	2288.3	0.0	1937.9
Constrained hotdeck	AHINC	VEHICLES6_3	0.0	2159.8	0.0	2146.3	0.0	3359.0	0.0	1773.3
Constrained hotdeck	AHINC	VEHICLES6_4	37.5	3998.9	28.6	3915.6	0.0	7040.9	0.9	3539.4
Constrained hotdeck	AHINC	VEHICLES6_5	0.0	6562.9	0.0	6435.0	0.0	9368.3	0.0	6308.6
Constrained hotdeck	AHINC	VEHICLES6_6	0.0	7559.9	0.0	7152.5	0.0	9359.0	0.0	9928.5
Constrained hotdeck	JWMN	MEANS6_2	0.1	1.8	-0.3	2.3	0.0	3.0	0.0	1.9
Constrained hotdeck	JWMN	MEANS6_3	0.0	3.8	0.0	4.8	0.0	5.2	0.0	4.8
Constrained hotdeck	JWMN	MEANS6_4	0.0	4.3	0.0	6.9	0.0	7.3	0.0	5.8
Constrained hotdeck	JWMN	MEANS6_5	0.0	6.3	0.0	5.1	0.0	6.0	0.0	5.5
Constrained hotdeck	JWMN	TM_LEAVE5_3	0.0	2.1	-0.3	2.4	0.0	3.2	-0.1	2.0
Constrained hotdeck	JWMN	TM_LEAVE5_4	0.1	3.1	-0.1	3.6	0.1	4.4	0.2	3.2

Table E-3. Median and Interquartile Range of Cell Mean Differences, County Level, Full Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Parametric	AHINC	VEHICLES6_2	-2341.2	4128.1	-2082.5	5427.8	-1854.5	0.0	-80.9	1056.9
Parametric	AHINC	VEHICLES6_3	413.1	2749.9	-420.6	3374.1	-68.4	0.0	32.2	1625.4
Parametric	AHINC	VEHICLES6_4	-1914.0	3645.6	-2853.2	4731.3	-1742.0	0.0	-1233.7	1967.4
Parametric	AHINC	VEHICLES6_5	-1345.9	4713.7	-399.9	6888.5	2996.5	0.0	647.6	7624.0
Parametric	AHINC	VEHICLES6_6	516.4	12209.1	118.1	8570.3	18038.1	0.0	6563.7	20731.0
Parametric	JWMN	MEANS6_2	1.5	0.6	0.8	1.6	-0.4	0.0	0.7	0.5
Parametric	JWMN	MEANS6_3	2.4	3.1	0.2	4.0	-1.1	0.0	1.0	2.1
Parametric	JWMN	MEANS6_4	4.0	5.7	0.4	7.2	0.0	0.0	1.2	3.5
Parametric	JWMN	MEANS6_5	6.7	6.9	2.5	3.7	-0.2	0.0	5.3	5.9
Parametric	JWMN	TM_LEAVE5_3	1.3	0.8	0.8	1.7	-0.6	0.0	0.6	1.4
Parametric	JWMN	TM_LEAVE5_4	5.0	2.6	1.0	3.0	1.1	0.0	2.5	2.2
Parametric	JWMN	Flow	0.0	6.0	-0.1	7.0	-0.4	5.5	0.0	5.1
Semi-parametric	AHINC	VEHICLES6_2	1536.3	3633.8	914.0	3715.1	-365.7	0.0	2186.8	1829.8
Semi-parametric	AHINC	VEHICLES6_3	962.6	1522.4	1465.5	3195.1	1929.8	0.0	1328.2	1583.8
Semi-parametric	AHINC	VEHICLES6_4	34.6	1838.6	-713.9	2493.5	1532.4	0.0	-987.4	2253.6
Semi-parametric	AHINC	VEHICLES6_5	-1992.6	3795.9	-1017.7	5855.8	-5956.9	0.0	-338.7	1930.0
Semi-parametric	AHINC	VEHICLES6_6	-3367.2	8901.4	-138.1	9521.4	3914.8	0.0	-4013.0	12022.2
Semi-parametric	JWMN	MEANS6_2	0.6	0.6	0.2	1.0	0.1	0.0	0.1	0.4
Semi-parametric	JWMN	MEANS6_3	0.7	1.8	1.1	2.9	0.1	0.0	1.3	1.8
Semi-parametric	JWMN	MEANS6_4	0.8	4.7	1.1	5.2	0.1	0.0	0.5	5.2
Semi-parametric	JWMN	MEANS6_5	2.7	4.5	0.4	1.9	0.5	0.0	1.8	2.5
Semi-parametric	JWMN	TM_LEAVE5_3	0.0	0.9	0.4	1.1	-0.4	0.0	-0.2	0.7
Semi-parametric	JWMN	TM_LEAVE5_4	2.5	2.0	0.1	2.4	1.7	0.0	1.8	2.1
Semi-parametric	JWMN	Flow	0.0	4.3	0.0	4.6	0.0	7.3	0.0	4.0
Constrained hotdeck	AHINC	VEHICLES6_2	219.9	1826.4	-70.2	1730.1	-129.7	0.0	-75.4	688.6
Constrained hotdeck	AHINC	VEHICLES6_3	23.2	830.7	-1.1	954.1	302.3	0.0	119.6	170.5
Constrained hotdeck	AHINC	VEHICLES6_4	318.1	973.3	62.0	1415.1	-286.6	0.0	57.6	615.6
Constrained hotdeck	AHINC	VEHICLES6_5	692.4	2087.2	195.6	2031.9	-370.5	0.0	102.3	1235.0
Constrained hotdeck	AHINC	VEHICLES6_6	-949.3	1819.3	-27.5	2312.3	115.0	0.0	1059.6	5992.5
Constrained hotdeck	JWMN	MEANS6_2	-0.1	0.4	0.8	1.2	0.0	0.0	0.1	0.2
Constrained hotdeck	JWMN	MEANS6_3	0.1	0.9	0.5	1.9	0.0	0.0	-0.4	1.1
Constrained hotdeck	JWMN	MEANS6_4	0.3	1.4	0.7	3.9	-0.3	0.0	0.7	1.0
Constrained hotdeck	JWMN	MEANS6_5	0.1	2.3	0.8	2.1	-0.5	0.0	0.3	1.1
Constrained hotdeck	JWMN	TM_LEAVE5_3	-0.1	0.6	0.8	1.2	-0.2	0.0	0.0	0.3
Constrained hotdeck	JWMN	TM_LEAVE5_4	0.2	0.8	0.7	1.5	0.2	0.0	0.4	0.6
Constrained hotdeck	JWMN	Flow	0.0	6.2	0.0	5.8	0.0	5.9	0.0	4.7

Table E-4. Median and Interquartile Range of Cell Mean Differences, County Level, Partial Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
Parametric	AHINC	VEHICLES6_2	576.4	2542.0	-51.3	781.5	-174.4	0.0	-78.6	884.7
Parametric	AHINC	VEHICLES6_3	141.1	911.0	87.6	666.7	246.5	0.0	544.0	1279.4
Parametric	AHINC	VEHICLES6_4	-257.9	1883.9	-315.2	1426.7	-745.4	0.0	-899.9	1346.5
Parametric	AHINC	VEHICLES6_5	-1020.2	4544.1	-69.3	2252.1	2523.5	0.0	2235.9	7241.8
Parametric	AHINC	VEHICLES6_6	-1609.4	10021.6	218.2	3329.1	13275.9	0.0	7895.4	21242.9
Parametric	JWMN	MEANS6_2	0.7	0.7	0.0	0.8	-0.6	0.0	0.3	0.4
Parametric	JWMN	MEANS6_3	1.9	2.2	0.1	2.8	-0.2	0.0	0.9	2.0
Parametric	JWMN	MEANS6_4	1.7	4.3	0.0	4.7	-1.0	0.0	3.6	7.2
Parametric	JWMN	MEANS6_5	6.4	8.8	2.1	3.1	0.0	0.0	4.4	5.0
Parametric	JWMN	TM_LEAVE5_3	0.5	0.5	0.1	1.0	-0.7	0.0	0.4	0.6
Parametric	JWMN	TM_LEAVE5_4	2.7	2.0	0.1	1.8	0.6	0.0	1.5	0.9
Parametric	JWMN	Flow	0.0	4.0	0.0	3.9	0.0	5.6	0.0	3.7
Semi-parametric	AHINC	VEHICLES6_2	-282.3	1318.7	49.7	409.7	1282.7	0.0	507.2	2586.5
Semi-parametric	AHINC	VEHICLES6_3	744.5	1211.3	199.7	568.3	2653.9	0.0	652.1	1203.0
Semi-parametric	AHINC	VEHICLES6_4	-508.2	1257.7	274.9	935.0	2730.6	0.0	-193.1	936.0
Semi-parametric	AHINC	VEHICLES6_5	-1115.9	2426.7	-308.4	2062.2	-8499.1	0.0	-2554.0	4581.7
Semi-parametric	AHINC	VEHICLES6_6	-2843.5	6732.0	-352.4	2367.8	-4442.8	0.0	-3428.5	6699.2
Semi-parametric	JWMN	MEANS6_2	0.1	0.4	-0.1	0.5	0.0	0.0	0.0	0.3
Semi-parametric	JWMN	MEANS6_3	0.3	2.6	0.2	1.7	0.3	0.0	0.9	2.5
Semi-parametric	JWMN	MEANS6_4	-0.5	3.3	1.1	3.9	-0.6	0.0	0.7	3.9
Semi-parametric	JWMN	MEANS6_5	2.0	4.3	0.3	1.6	0.1	0.0	0.7	2.0
Semi-parametric	JWMN	TM_LEAVE5_3	0.0	0.5	0.1	0.6	-0.2	0.0	0.2	0.5
Semi-parametric	JWMN	TM_LEAVE5_4	0.5	1.0	-0.1	1.2	0.3	0.0	-0.1	0.9
Semi-parametric	JWMN	Flow	0.0	3.0	0.0	2.0	0.0	4.5	0.0	2.9
Constrained hotdeck	AHINC	VEHICLES6_2	95.7	661.0	-0.2	204.1	50.4	0.0	-127.9	281.7
Constrained hotdeck	AHINC	VEHICLES6_3	91.6	403.2	-1.3	229.7	217.1	0.0	96.3	392.0
Constrained hotdeck	AHINC	VEHICLES6_4	5.0	772.4	108.0	496.6	-621.6	0.0	147.1	558.6
Constrained hotdeck	AHINC	VEHICLES6_5	154.0	1172.3	141.4	728.9	-139.9	0.0	228.5	1219.4
Constrained hotdeck	AHINC	VEHICLES6_6	-282.4	2231.1	98.1	878.3	32.4	0.0	-389.6	3648.1
Constrained hotdeck	JWMN	MEANS6_2	0.0	0.3	0.0	0.3	0.0	0.0	0.0	0.4
Constrained hotdeck	JWMN	MEANS6_3	-0.1	0.8	0.1	1.0	-0.1	0.0	0.1	0.5
Constrained hotdeck	JWMN	MEANS6_4	-0.1	2.1	0.0	2.5	-0.3	0.0	0.0	1.8
Constrained hotdeck	JWMN	MEANS6_5	0.1	2.5	0.4	1.2	-0.2	0.0	-0.5	1.0
Constrained hotdeck	JWMN	TM_LEAVE5_3	0.0	0.3	0.0	0.4	-0.1	0.0	-0.1	0.5
Constrained hotdeck	JWMN	TM_LEAVE5_4	0.1	0.3	0.0	0.7	0.2	0.0	0.1	0.3
Constrained hotdeck	JWMN	Flow	0.0	3.2	0.0	1.4	0.0	6.4	0.0	3.6

Table E-5. Median and Interquartile Range of Standard Error Differences, County Level, Full Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**Attribute**	**BYVAR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Parametric	AHINC	VEHICLES6_2	996.27	784.65	1378.14	2174.66	195.60	0.00	322.66	515.92
Parametric	AHINC	VEHICLES6_3	633.83	1117.87	529.56	1233.93	-99.10	0.00	481.09	709.21
Parametric	AHINC	VEHICLES6_4	1555.51	1925.58	1497.54	1824.06	1868.99	0.00	1155.80	1011.06
Parametric	AHINC	VEHICLES6_5	973.74	3105.10	1232.95	2617.82	4679.07	0.00	2333.41	4856.35
Parametric	AHINC	VEHICLES6_6	1942.72	5933.23	1640.41	3158.12	19172.14	0.00	4095.82	23835.75
Parametric	JWMN	MEANS6_2	1.12	0.62	0.56	0.76	0.25	0.00	0.49	0.46
Parametric	JWMN	MEANS6_3	1.43	2.07	0.75	1.89	0.74	0.00	0.63	0.89
Parametric	JWMN	MEANS6_4	2.05	3.68	1.17	3.48	0.07	0.00	1.57	1.69
Parametric	JWMN	MEANS6_5	3.30	5.98	1.55	2.63	0.00	0.00	3.00	4.96
Parametric	JWMN	TM_LEAVE5_3	0.96	0.72	0.63	0.85	0.42	0.00	0.40	1.04
Parametric	JWMN	TM_LEAVE5_4	3.68	2.24	0.56	1.45	0.79	0.00	1.70	2.06
Semi-parametric	AHINC	VEHICLES6_2	508.12	1798.44	441.10	1440.05	-619.37	0.00	852.34	2254.11
Semi-parametric	AHINC	VEHICLES6_3	541.93	670.71	915.47	2152.88	1262.52	0.00	817.52	1534.81
Semi-parametric	AHINC	VEHICLES6_4	256.25	344.64	476.01	1254.08	1069.06	0.00	297.17	985.86
Semi-parametric	AHINC	VEHICLES6_5	977.66	1850.71	848.20	2459.02	3558.69	0.00	174.15	724.40
Semi-parametric	AHINC	VEHICLES6_6	883.36	2654.61	1460.28	3967.18	631.08	0.00	2644.65	5261.50
Semi-parametric	JWMN	MEANS6_2	0.26	0.43	0.19	0.48	0.02	0.00	0.06	0.21
Semi-parametric	JWMN	MEANS6_3	0.34	0.73	0.67	1.36	0.03	0.00	0.80	0.90
Semi-parametric	JWMN	MEANS6_4	0.39	1.46	0.93	2.82	-0.06	0.00	0.95	2.71
Semi-parametric	JWMN	MEANS6_5	1.30	2.23	0.44	1.10	0.15	0.00	0.72	1.17
Semi-parametric	JWMN	TM_LEAVE5_3	0.21	0.43	0.23	0.43	0.25	0.00	0.23	0.25
Semi-parametric	JWMN	TM_LEAVE5_4	1.72	1.62	0.54	1.30	1.34	0.00	1.48	2.08
Constrained hotdeck	AHINC	VEHICLES6_2	178.47	359.60	120.84	644.27	58.94	0.00	20.17	64.39
Constrained hotdeck	AHINC	VEHICLES6_3	111.45	153.58	66.74	240.38	60.84	0.00	3.37	80.60
Constrained hotdeck	AHINC	VEHICLES6_4	258.45	458.79	146.08	451.22	-20.25	0.00	68.82	125.48
Constrained hotdeck	AHINC	VEHICLES6_5	286.80	801.68	160.89	621.43	73.08	0.00	220.89	555.14
Constrained hotdeck	AHINC	VEHICLES6_6	-86.56	742.78	195.97	813.61	178.52	0.00	698.56	1557.23
Constrained hotdeck	JWMN	MEANS6_2	0.07	0.18	0.34	0.62	0.02	0.00	0.06	0.06
Constrained hotdeck	JWMN	MEANS6_3	0.16	0.33	0.16	0.71	0.02	0.00	0.23	0.37
Constrained hotdeck	JWMN	MEANS6_4	-0.08	0.77	0.38	1.67	0.06	0.00	0.34	0.20
Constrained hotdeck	JWMN	MEANS6_5	0.14	0.64	0.47	1.07	0.17	0.00	0.06	0.57
Constrained hotdeck	JWMN	TM_LEAVE5_3	0.10	0.13	0.35	0.60	0.06	0.00	0.07	0.06
Constrained hotdeck	JWMN	TM_LEAVE5_4	0.06	0.22	0.21	0.59	0.08	0.00	0.15	0.26

Table E-6. Median and Interquartile Range of Standard Error Differences, County Level, Partial Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**Attribute**	**BYVAR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Parametric	AHINC	VEHICLES6_2	181.46	624.71	9.22	263.25	-102.66	0.00	73.94	271.08
Parametric	AHINC	VEHICLES6_3	113.04	382.01	50.24	231.36	-22.22	0.00	192.31	638.90
Parametric	AHINC	VEHICLES6_4	388.27	1156.77	95.64	367.69	1026.98	0.00	738.06	1110.37
Parametric	AHINC	VEHICLES6_5	760.90	1890.13	149.25	637.53	4161.82	0.00	2482.01	5506.16
Parametric	AHINC	VEHICLES6_6	1444.67	3618.57	239.75	838.40	13281.62	0.00	6671.74	23088.62
Parametric	JWMN	MEANS6_2	0.39	0.63	0.18	0.31	0.35	0.00	0.22	0.23
Parametric	JWMN	MEANS6_3	0.92	1.19	0.36	0.94	0.10	0.00	0.57	0.91
Parametric	JWMN	MEANS6_4	0.92	2.16	0.52	2.87	0.58	0.00	2.13	4.74
Parametric	JWMN	MEANS6_5	1.90	6.44	1.17	2.24	0.01	0.00	2.09	3.94
Parametric	JWMN	TM_LEAVE5_3	0.26	0.26	0.18	0.41	0.51	0.00	0.21	0.35
Parametric	JWMN	TM_LEAVE5_4	1.90	1.64	0.28	0.70	0.31	0.00	0.93	0.77
Semi-parametric	AHINC	VEHICLES6_2	65.60	432.37	62.92	300.00	471.83	0.00	176.82	719.79
Semi-parametric	AHINC	VEHICLES6_3	342.61	522.03	85.10	261.38	2013.01	0.00	219.78	778.43
Semi-parametric	AHINC	VEHICLES6_4	191.27	351.98	126.05	314.35	1687.03	0.00	142.64	101.96
Semi-parametric	AHINC	VEHICLES6_5	168.92	736.62	149.81	569.38	5566.19	0.00	983.57	630.51
Semi-parametric	AHINC	VEHICLES6_6	809.95	2660.57	187.62	719.74	666.85	0.00	1548.18	1521.02
Semi-parametric	JWMN	MEANS6_2	0.04	0.14	0.07	0.14	-0.02	0.00	0.09	0.10
Semi-parametric	JWMN	MEANS6_3	0.19	0.40	0.16	0.73	0.04	0.00	0.75	1.29
Semi-parametric	JWMN	MEANS6_4	0.14	1.34	0.60	1.83	0.22	0.00	0.63	1.59
Semi-parametric	JWMN	MEANS6_5	0.52	2.08	0.24	0.62	0.03	0.00	0.23	0.93
Semi-parametric	JWMN	TM_LEAVE5_3	0.08	0.14	0.10	0.16	0.06	0.00	0.14	0.20
Semi-parametric	JWMN	TM_LEAVE5_4	0.33	0.66	0.20	0.51	0.19	0.00	0.13	0.30
Constrained hotdeck	AHINC	VEHICLES6_2	62.23	156.86	7.26	66.97	36.22	0.00	25.99	126.43
Constrained hotdeck	AHINC	VEHICLES6_3	20.76	75.04	7.83	58.06	39.98	0.00	42.77	50.83
Constrained hotdeck	AHINC	VEHICLES6_4	99.95	154.01	27.53	109.73	70.82	0.00	104.04	161.78
Constrained hotdeck	AHINC	VEHICLES6_5	99.46	467.41	43.16	174.10	63.45	0.00	168.76	384.04
Constrained hotdeck	AHINC	VEHICLES6_6	163.36	506.26	21.98	202.87	66.73	0.00	145.53	910.06
Constrained hotdeck	JWMN	MEANS6_2	0.02	0.05	0.03	0.12	0.01	0.00	0.06	0.12
Constrained hotdeck	JWMN	MEANS6_3	0.05	0.26	0.04	0.31	0.03	0.00	0.01	0.16
Constrained hotdeck	JWMN	MEANS6_4	0.07	0.74	0.08	1.05	0.09	0.00	0.22	0.41
Constrained hotdeck	JWMN	MEANS6_5	0.05	0.83	0.20	0.60	0.14	0.00	-0.08	0.56
Constrained hotdeck	JWMN	TM_LEAVE5_3	0.02	0.08	0.03	0.15	0.05	0.00	0.09	0.15
Constrained hotdeck	JWMN	TM_LEAVE5_4	0.01	0.13	0.03	0.17	0.09	0.00	0.04	0.21

Table E-7. Median and Interquartile Range of Cramers V Differences, Full Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**GEOAREA**	**MOT**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Parametric	County	AGE9	0.00	0.02	-0.01	0.03	-0.02	0.00	-0.01	0.01
Parametric	County	HH_INC26	-0.01	0.03	-0.01	0.03	-0.02	0.00	-0.01	0.02
Parametric	County	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parametric	County	TRAVEL_TM12	0.00	0.01	-0.01	0.01	0.00	0.00	-0.01	0.01
Parametric	County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parametric	County flows	AGE9	0.00	0.14	0.00	0.13	0.00	0.09	0.00	0.14
Parametric	County flows	HH_INC26	0.00	0.06	0.00	0.11	0.00	0.10	0.00	0.06
Parametric	County flows	TM_LEAVE10	0.00	0.09	0.00	0.12	-0.04	0.18	0.00	0.08
Parametric	County flows	TRAVEL_TM12	0.00	0.08	0.00	0.15	0.00	0.14	0.00	0.09
Parametric	County flows	VEHICLES6	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
Parametric	TAZ	AGE9	0.00	0.14	0.00	0.15	0.00	0.18	0.00	0.14
Parametric	TAZ	HH_INC26	0.00	0.14	0.00	0.14	0.00	0.14	0.00	0.14
Parametric	TAZ	TM_LEAVE10	0.00	0.06	0.00	0.08	0.00	0.11	0.00	0.08
Parametric	TAZ	TRAVEL_TM12	-0.01	0.08	0.00	0.11	0.00	0.12	0.00	0.11
Parametric	TAZ	VEHICLES6	0.00	0.02	0.00	0.02	0.00	0.01	0.00	0.01
Semi-parametric	County	AGE9	0.00	0.01	-0.01	0.03	0.00	0.00	-0.01	0.01
Semi-parametric	County	HH_INC26	-0.01	0.02	-0.01	0.03	-0.03	0.00	-0.01	0.01
Semi-parametric	County	TM_LEAVE10	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Semi-parametric	County	TRAVEL_TM12	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.01
Semi-parametric	County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Semi-parametric	County flows	AGE9	0.00	0.14	0.00	0.13	-0.02	0.13	0.00	0.15
Semi-parametric	County flows	HH_INC26	0.00	0.04	0.00	0.10	0.00	0.29	0.00	0.06
Semi-parametric	County flows	TM_LEAVE10	0.00	0.08	0.00	0.13	0.00	0.13	0.00	0.06
Semi-parametric	County flows	TRAVEL_TM12	0.00	0.05	0.00	0.12	0.00	0.12	0.00	0.05
Semi-parametric	County flows	VEHICLES6	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00
Semi-parametric	TAZ	AGE9	0.00	0.13	0.00	0.14	0.00	0.18	0.00	0.13
Semi-parametric	TAZ	HH_INC26	0.00	0.11	0.00	0.11	0.00	0.13	0.00	0.13
Semi-parametric	TAZ	TM_LEAVE10	0.00	0.06	0.00	0.08	0.00	0.10	0.00	0.07
Semi-parametric	TAZ	TRAVEL_TM12	0.00	0.07	0.00	0.09	0.00	0.12	0.00	0.09
Semi-parametric	TAZ	VEHICLES6	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.01
Constrained hotdeck	County	AGE9	0.00	0.01	-0.01	0.04	0.00	0.00	0.00	0.01
Constrained hotdeck	County	HH_INC26	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.01
Constrained hotdeck	County	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table E-7. Median and Interquartile Range of Cramers V Differences, Full Replacement (Continued)

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**GEOAREA**	**MOT**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Constrained hotdeck	County	TRAVEL_TM12	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Constrained hotdeck	County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Constrained hotdeck	County flows	AGE9	0.00	0.14	0.00	0.12	-0.01	0.15	0.00	0.16
Constrained hotdeck	County flows	HH_INC26	0.00	0.02	0.00	0.07	0.00	0.03	0.00	0.02
Constrained hotdeck	County flows	TM_LEAVE10	0.00	0.01	0.00	0.01	0.00	0.01	0.00	0.00
Constrained hotdeck	County flows	TRAVEL_TM12	0.00	0.05	0.00	0.09	0.00	0.20	0.00	0.03
Constrained hotdeck	County flows	VEHICLES6	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
Constrained hotdeck	TAZ	AGE9	0.00	0.15	0.00	0.17	0.00	0.23	0.00	0.16
Constrained hotdeck	TAZ	HH_INC26	0.00	0.08	0.00	0.08	0.00	0.06	0.00	0.10
Constrained hotdeck	TAZ	TM_LEAVE10	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.02
Constrained hotdeck	TAZ	TRAVEL_TM12	0.00	0.05	0.00	0.07	0.00	0.08	0.00	0.06
Constrained hotdeck	TAZ	VEHICLES6	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.01

Table E-8. Median and Interquartile Range of Cramers V Differences, Partial Replacement

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**GEOAREA**	**MOT**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Parametric	County	AGE9	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.00
Parametric	County	HH_INC26	-0.01	0.01	0.00	0.02	-0.02	0.00	-0.01	0.02
Parametric	County	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parametric	County	TRAVEL_TM12	0.00	0.01	0.00	0.01	0.00	0.00	-0.01	0.01
Parametric	County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Parametric	County flows	AGE9	0.00	0.02	0.00	0.01	0.00	0.02	0.00	0.02
Parametric	County flows	HH_INC26	0.00	0.04	0.00	0.03	0.00	0.12	0.00	0.05
Parametric	County flows	TM_LEAVE10	0.00	0.01	0.00	0.01	0.00	0.10	0.00	0.02
Parametric	County flows	TRAVEL_TM12	0.00	0.05	0.00	0.07	-0.01	0.10	0.00	0.04
Parametric	County flows	VEHICLES6	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.01
Parametric	TAZ	AGE9	0.00	0.04	0.00	0.04	0.00	0.05	0.00	0.04
Parametric	TAZ	HH_INC26	0.00	0.11	0.00	0.12	0.00	0.14	0.00	0.13
Parametric	TAZ	TM_LEAVE10	0.00	0.03	0.00	0.03	0.00	0.04	0.00	0.03
Parametric	TAZ	TRAVEL_TM12	0.00	0.08	0.00	0.10	0.00	0.12	0.00	0.09
Parametric	TAZ	VEHICLES6	0.00	0.01	0.00	0.02	0.00	0.01	0.00	0.01
Semi-parametric	County	AGE9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Semi-parametric	County	HH_INC26	0.00	0.01	0.00	0.01	-0.03	0.00	0.00	0.05
Semi-parametric	County	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Semi-parametric	County	TRAVEL_TM12	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01
Semi-parametric	County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Semi-parametric	County flows	AGE9	0.00	0.01	0.00	0.01	0.00	0.02	0.00	0.03
Semi-parametric	County flows	HH_INC26	0.00	0.03	0.00	0.02	0.02	0.32	0.00	0.14
Semi-parametric	County flows	TM_LEAVE10	0.00	0.01	0.00	0.01	0.00	0.04	0.00	0.03
Semi-parametric	County flows	TRAVEL_TM12	0.00	0.03	0.00	0.04	0.00	0.06	0.00	0.14
Semi-parametric	County flows	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Semi-parametric	TAZ	AGE9	0.00	0.03	0.00	0.02	0.00	0.04	0.00	0.00
Semi-parametric	TAZ	HH_INC26	0.00	0.09	0.00	0.10	0.00	0.10	-0.01	0.01
Semi-parametric	TAZ	TM_LEAVE10	0.00	0.02	0.00	0.02	0.00	0.03	0.00	0.00
Semi-parametric	TAZ	TRAVEL_TM12	0.00	0.07	0.00	0.09	0.00	0.11	0.00	0.00
Semi-parametric	TAZ	VEHICLES6	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00

Table E-8. Median and Interquartile Range of Cramers V Differences, Partial Replacement (Continued)

			ATLANTA		IOWA		MADISON		ST.LOUIS	
Approach	**GEOAREA**	**MOT**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**	**Median**	**IQR**
Constrained hotdeck	County	AGE9	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
Constrained hotdeck	County	HH_INC26	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.03
Constrained hotdeck	County	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Constrained hotdeck	County	TRAVEL_TM12	0.00	0.01	0.00	0.01	0.00	0.00	0.00	0.05
Constrained hotdeck	County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Constrained hotdeck	County flows	AGE9	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.03
Constrained hotdeck	County flows	HH_INC26	0.00	0.01	0.00	0.01	0.00	0.03	0.00	0.12
Constrained hotdeck	County flows	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.03
Constrained hotdeck	County flows	TRAVEL_TM12	0.00	0.02	0.00	0.02	0.00	0.08	0.00	0.08
Constrained hotdeck	County flows	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
Constrained hotdeck	TAZ	AGE9	0.00	0.03	0.00	0.03	0.00	0.04	0.00	0.01
Constrained hotdeck	TAZ	HH_INC26	0.00	0.07	0.00	0.07	0.00	0.05	0.00	0.01
Constrained hotdeck	TAZ	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Constrained hotdeck	TAZ	TRAVEL_TM12	0.00	0.05	0.00	0.06	0.00	0.06	0.00	0.01
Constrained hotdeck	TAZ	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table E- 9. Pairwise Correlations Between Key Ordinal Variables, Full Replacement, by Perturbation Run, Atlanta

				Run				
Approach	**Var1**	**Var2**	**Actual**	**1**	**2**	**3**	**4**	**5**
Parametric	AHINC	AGE9	0.0949	0.0389	0.0416	0.0423	0.0422	0.0427
	AHINC	POVERTY	0.2510	0.1270	0.1277	0.1297	0.1269	0.1305
	AGE9	POVERTY	0.1308	0.1466	0.1504	0.1518	0.1481	0.1438
	POVERTY	JWMN	0.0061	-0.0065	-0.0087	-0.0103	-0.0006	-0.0129
	JWMN	SynJWD	-0.1296	-0.1112	-0.1072	-0.1113	-0.1170	-0.1136
	JWMN	FLOWDIST	0.1081	0.1362	0.1366	0.1372	0.1390	0.1359
Semi-parametric	AHINC	AGE9	0.0949	0.0720	0.0759	0.0744	0.0787	0.0721
	AHINC	POVERTY	0.2510	0.1828	0.1918	0.1849	0.1873	0.1847
	AGE9	POVERTY	0.1308	0.1177	0.1269	0.1180	0.1230	0.1248
	POVERTY	JWMN	0.0061	0.0061	0.0017	0.0041	0.0036	0.0042
	JWMN	SynJWD	-0.1296	-0.0895	-0.0876	-0.0825	-0.0836	-0.0805
	JWMN	FLOWDIST	0.1081	0.0945	0.0961	0.0959	0.0968	0.0959
Constrained hotdeck	AHINC	AGE9	0.0949	0.0860	0.0879	0.0875	0.0852	0.0882
	AHINC	POVERTY	0.2510	0.2480	0.2509	0.2497	0.2496	0.2466
	AGE9	POVERTY	0.1308	0.0773	0.0791	0.0809	0.0797	0.0789
	POVERTY	JWMN	0.0061	0.0095	0.0052	0.0077	0.0037	0.0032
	JWMN	SynJWD	-0.1296	-0.1277	-0.1275	-0.1267	-0.1255	-0.1237
	JWMN	FLOWDIST	0.1081	0.0980	0.0941	0.0925	0.0933	0.0964

Table E-10. Pairwise Correlations Between Key Ordinal Variables, Partial Replacement, by Perturbation Run, Atlanta

Approach	Var1	Var2	Actual	Run 1	2	3	4	5
Parametric	AHINC	AGE9	0.0949	0.0522	0.0533	0.0528	0.0543	0.0557
	AHINC	POVERTY	0.2510	0.1474	0.1453	0.1433	0.1451	0.1480
	AGE9	POVERTY	0.1308	0.1375	0.1383	0.1367	0.1382	0.1375
	POVERTY	JWMN	0.0061	0.0017	0.0000	-0.0033	-0.0009	-0.0014
	JWMN	SynJWD	-0.1296	-0.1238	-0.1161	-0.1197	-0.1169	-0.1183
	JWMN	FLOWDIST	0.1081	0.1299	0.1319	0.1317	0.1278	0.1323
Semi-parametric	AHINC	AGE9	0.0949	0.0825	0.0832	0.0798	0.0809	0.0832
	AHINC	POVERTY	0.2510	0.2263	0.2286	0.2260	0.2256	0.2251
	AGE9	POVERTY	0.1308	0.1298	0.1317	0.1309	0.1303	0.1322
	POVERTY	JWMN	0.0061	0.0040	0.0077	0.0060	0.0038	0.0066
	JWMN	SynJWD	-0.1296	-0.1340	-0.1351	-0.1347	-0.1301	-0.1328
	JWMN	FLOWDIST	0.1081	0.0936	0.0972	0.0971	0.0962	0.0969
Constrained hotdeck	AHINC	AGE9	0.0949	0.0929	0.0926	0.0952	0.0950	0.0904
	AHINC	POVERTY	0.2510	0.2504	0.2507	0.2542	0.2523	0.2520
	AGE9	POVERTY	0.1308	0.1235	0.1238	0.1259	0.1246	0.1242
	POVERTY	JWMN	0.0061	0.0059	0.0065	0.0028	0.0025	0.0047
	JWMN	SynJWD	-0.1296	-0.1255	-0.1273	-0.1271	-0.1267	-0.1269
	JWMN	FLOWDIST	0.1081	0.0995	0.0961	0.0953	0.0945	0.0955

Table E-11. Multivariate Association Measure U, by Perturbation Run

Test site	Approach	Replacement amount	Run 1	2	3	4	5
ATLANTA	Parametric	Full	0.000547	0.000589	0.000533	0.000581	0.000515
IOWA	Parametric	Full	0.000202	0.000241	0.000220	0.000232	0.000214
MADISON	Parametric	Full	0.000353	0.000362	0.000347	0.000275	0.000476
ST.LOUIS	Parametric	Full	0.000312	0.000221	0.000257	0.000227	0.000233
ATLANTA	Semi-parametric	Full	0.000138	0.000144	0.000124	0.000127	0.000142
IOWA	Semi-parametric	Full	0.000135	0.000134	0.000142	0.000128	0.000153
MADISON	Semi-parametric	Full	0.000126	0.000110	0.000076	0.000116	0.000104
ST.LOUIS	Semi-parametric	Full	0.000075	0.000081	0.000074	0.000071	0.000076
ATLANTA	Constrained hotdeck	Full	0.000012	0.000011	0.000008	0.000007	0.000008
IOWA	Constrained hotdeck	Full	0.000025	0.000021	0.000020	0.000023	0.000024
MADISON	Constrained hotdeck	Full	0.000011	0.000008	0.000004	0.000006	0.000005
ST.LOUIS	Constrained hotdeck	Full	0.000001	0.000001	0.000002	0.000001	0.000003
ATLANTA	Parametric	Partial	0.000115	0.000110	0.000122	0.000108	0.000118
IOWA	Parametric	Partial	0.000025	0.000023	0.000025	0.000021	0.000025
MADISON	Parametric	Partial	0.000091	0.000093	0.000142	0.000105	0.000126
ST.LOUIS	Parametric	Partial	0.000109	0.000113	0.000101	0.000106	0.000104
ATLANTA	Semi-parametric	Partial	0.000022	0.000018	0.000019	0.000017	0.000019
IOWA	Semi-parametric	Partial	0.000008	0.000008	0.000013	0.000008	0.000012
MADISON	Semi-parametric	Partial	0.000049	0.000037	0.000046	0.000045	0.000073
ST.LOUIS	Semi-parametric	Partial	0.000035	0.000027	0.000023	0.000039	0.000026
ATLANTA	Constrained hotdeck	Partial	0.000002	0.000001	0.000000	0.000001	0.000001
IOWA	Constrained hotdeck	Partial	0.000014	0.000013	0.000016	0.000014	0.000014
MADISON	Constrained hotdeck	Partial	0.000004	0.000003	0.000004	0.000006	0.000002
ST.LOUIS	Constrained hotdeck	Partial	0.000003	0.000001	0.000001	0.000002	0.000002

Table E-12. Multivariate Association Measure U, by Perturbation Run, Includes Interaction Effects

Test site	Approach	Replacement amount	Run 1	2	3	4	5
ATLANTA	Parametric	Full	0.001530	0.001559	0.001477	0.001534	0.001490
IOWA	Parametric	Full	0.001227	0.001272	0.001246	0.001200	0.001206
MADISON	Parametric	Full	0.001094	0.001192	0.001168	0.001222	0.001275
ST.LOUIS	Parametric	Full	0.001354	0.001204	0.001325	0.001196	0.001176
ATLANTA	Semi-parametric	Full	0.002008	0.001931	0.002092	0.001903	0.002090
IOWA	Semi-parametric	Full	0.002305	0.002380	0.002425	0.002443	0.002354
MADISON	Semi-parametric	Full	0.002034	0.002753	0.002219	0.002367	0.002392
ST.LOUIS	Semi-parametric	Full	0.002974	0.002848	0.003204	0.002894	0.003162
ATLANTA	Constrained hotdeck	Full	0.000026	0.000027	0.000022	0.000019	0.000017
IOWA	Constrained hotdeck	Full	0.000078	0.000077	0.000074	0.000077	0.000075
MADISON	Constrained hotdeck	Full	0.000052	0.000066	0.000040	0.000047	0.000063
ST.LOUIS	Constrained hotdeck	Full	0.000036	0.000040	0.000038	0.000043	0.000043
ATLANTA	Parametric	Partial	0.000180	0.000183	0.000183	0.000164	0.000201
IOWA	Parametric	Partial	0.000094	0.000095	0.000099	0.000082	0.000096
MADISON	Parametric	Partial	0.000277	0.000253	0.000369	0.000259	0.000259
ST.LOUIS	Parametric	Partial	0.000217	0.000243	0.000209	0.000238	0.000212
ATLANTA	Semi-parametric	Partial	0.000159	0.000149	0.000148	0.000149	0.000145
IOWA	Semi-parametric	Partial	0.000148	0.000150	0.000142	0.000140	0.000142
MADISON	Semi-parametric	Partial	0.000433	0.000283	0.000320	0.000411	0.000342
ST.LOUIS	Semi-parametric	Partial	0.000280	0.000332	0.000278	0.000270	0.000293
ATLANTA	Constrained hotdeck	Partial	0.000005	0.000005	0.000003	0.000005	0.000005
IOWA	Constrained hotdeck	Partial	0.000023	0.000024	0.000024	0.000022	0.000024
MADISON	Constrained hotdeck	Partial	0.000026	0.000026	0.000036	0.000032	0.000023
ST.LOUIS	Constrained hotdeck	Partial	0.000031	0.000020	0.000022	0.000021	0.000027

Appendix F

DEVELOPMENT PHASE: PLOTS OF ACS AND PERTURBED DATA FOR MEAN TRAVEL TIME

Note: The dots in Figures F-1 through F-8 are shown only for localities with 30 or more ACS sample cases. The dots in Figures F-9 through F12 are shown only for flows with at least 10 ACS cases.

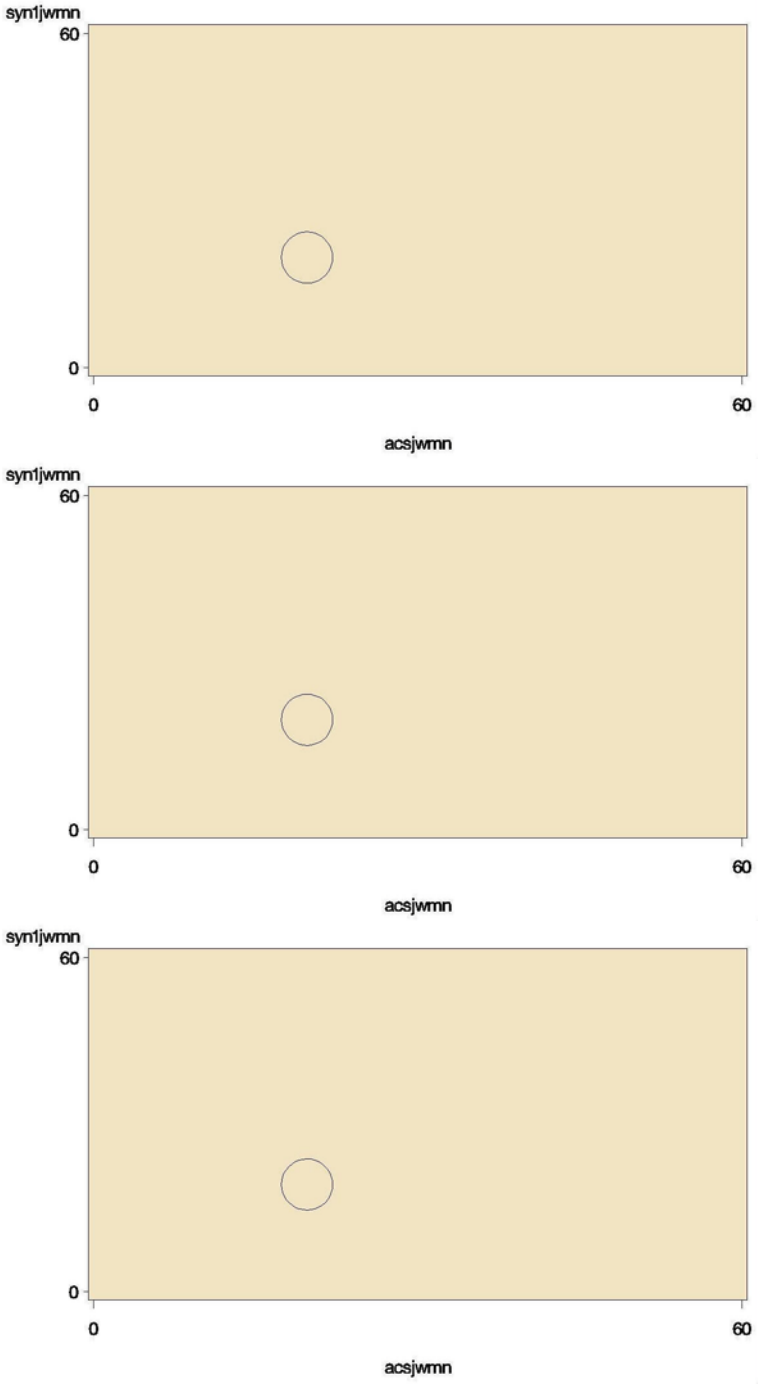


Figure F-1. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Madison's County: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

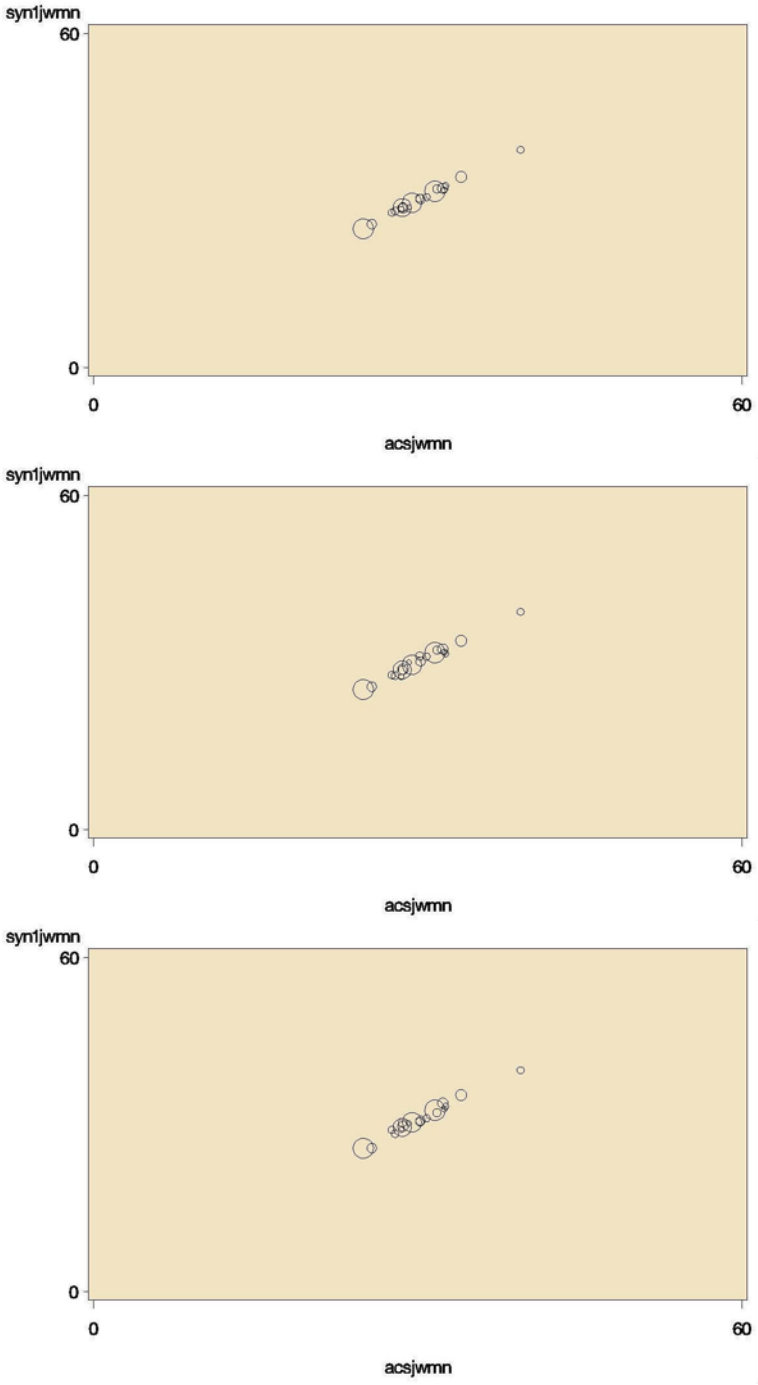


Figure F-2. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

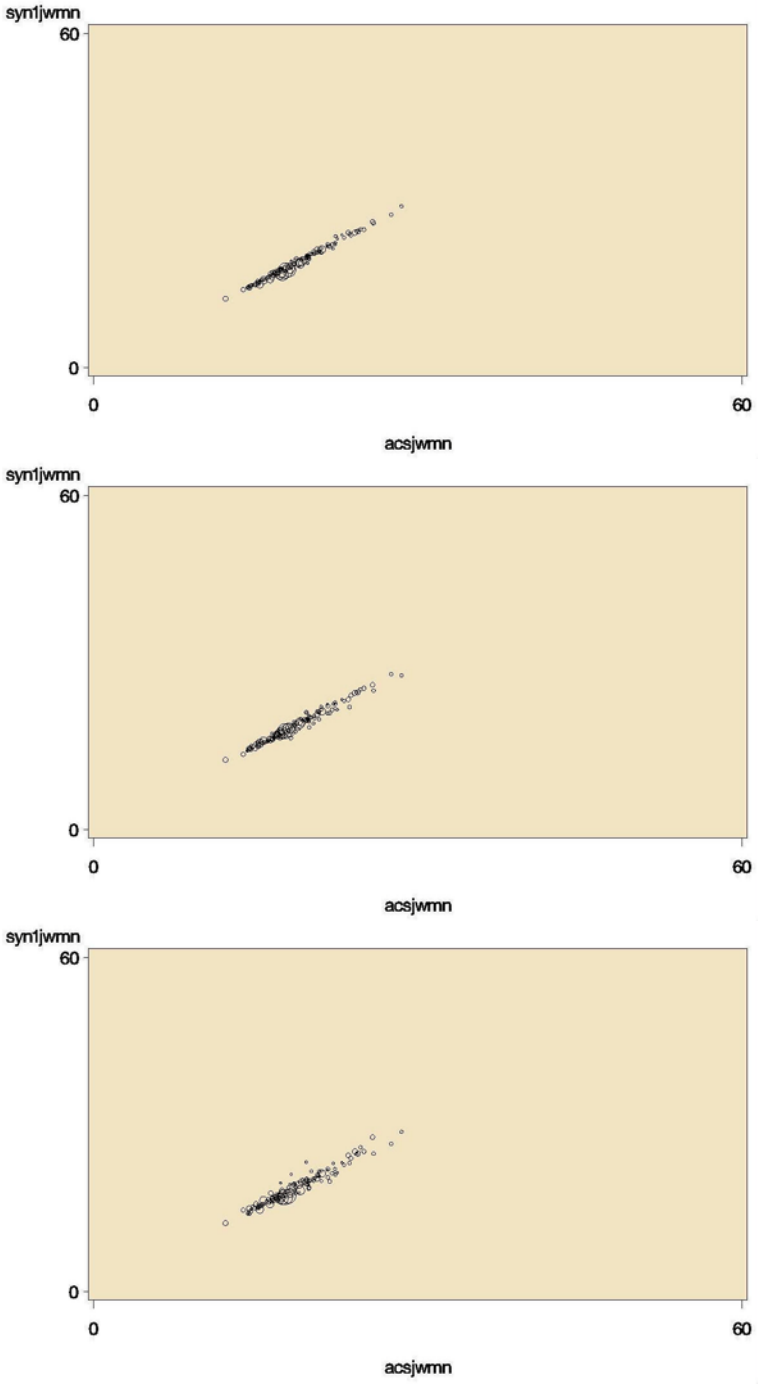


Figure F-3. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Iowa's Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

Figure F-4. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for St. Louis' Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

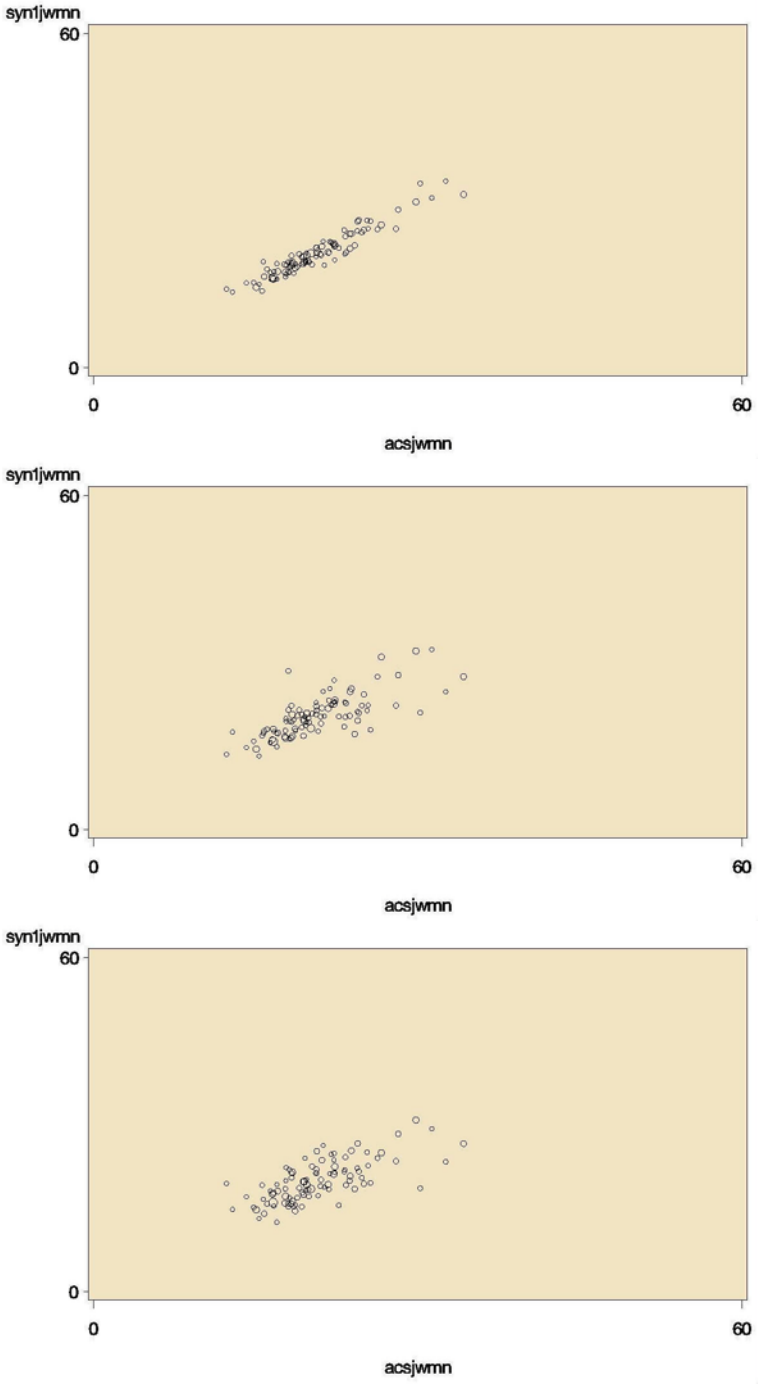


Figure F-5. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Madison's TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

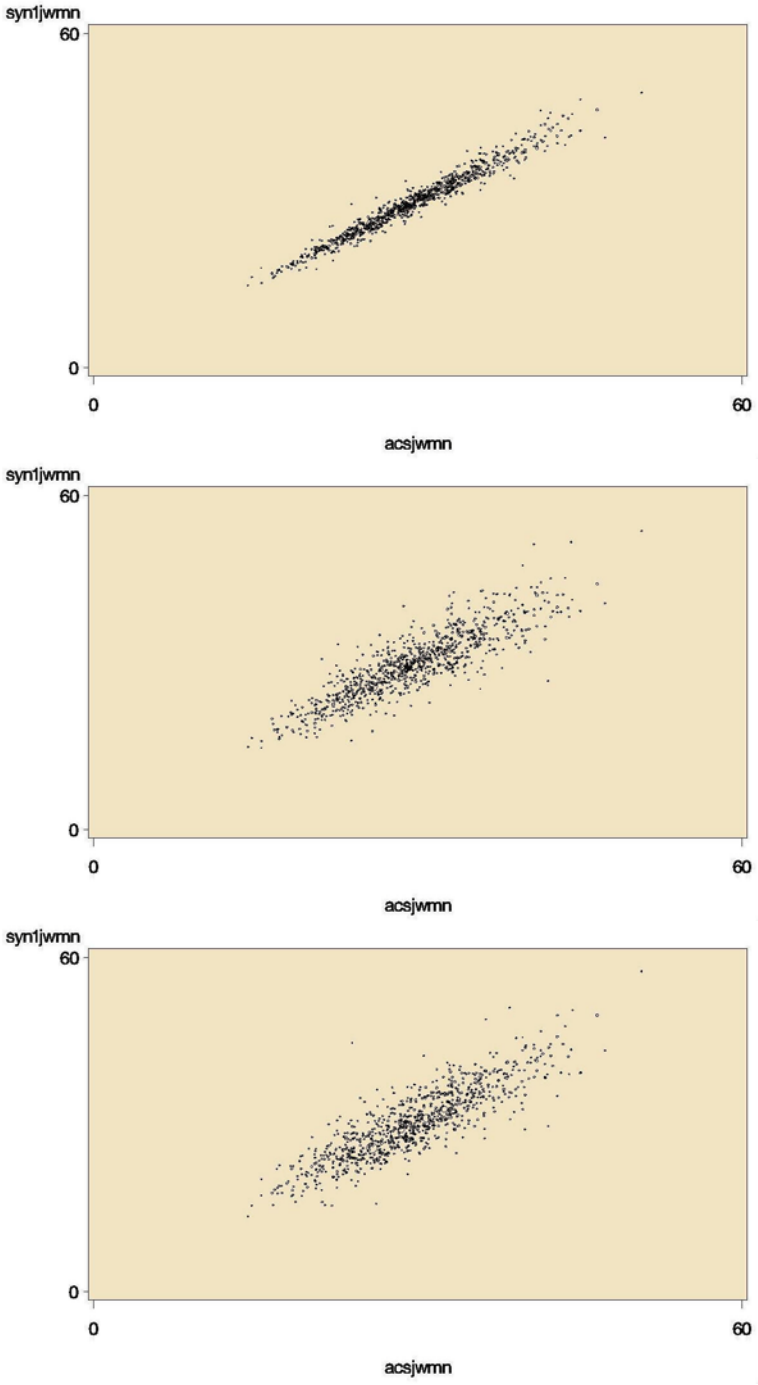


Figure F-6. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

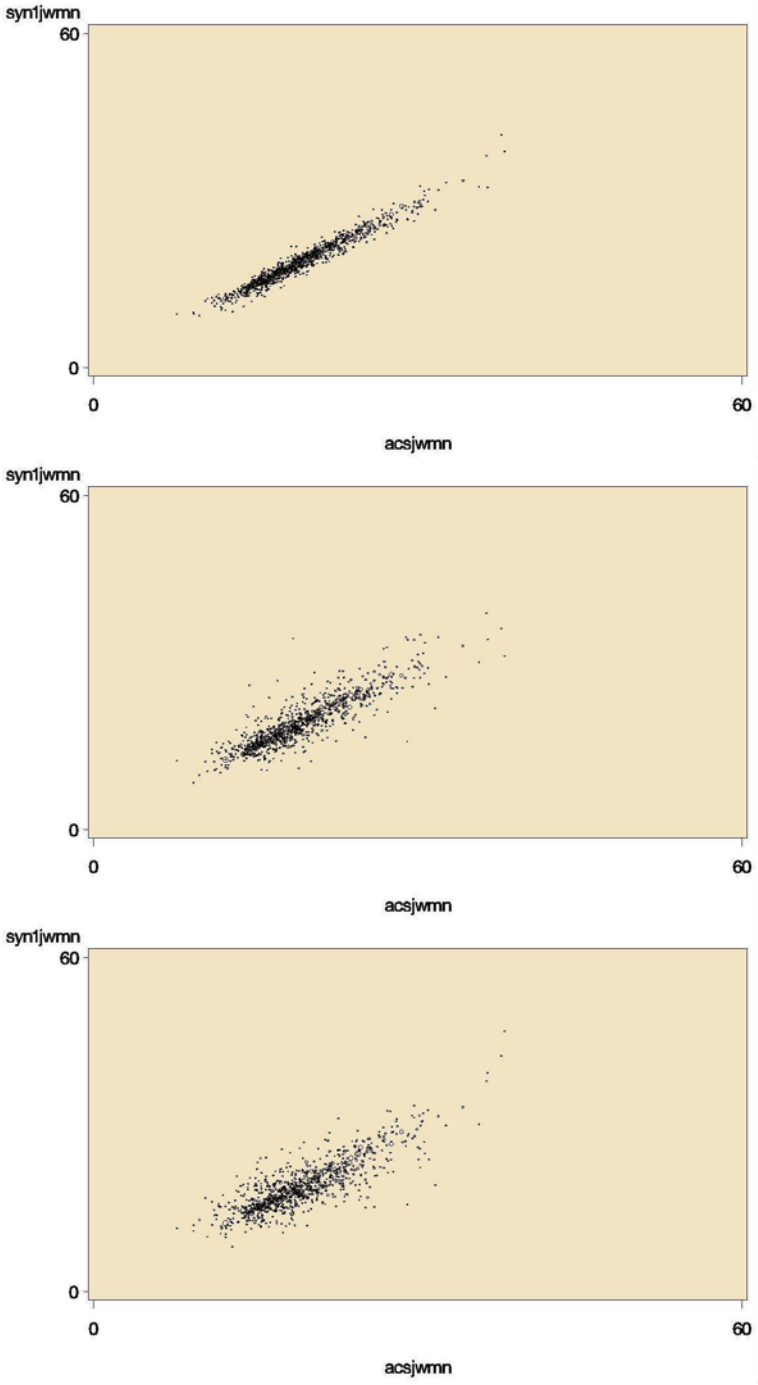


Figure F-7. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Iowa's TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

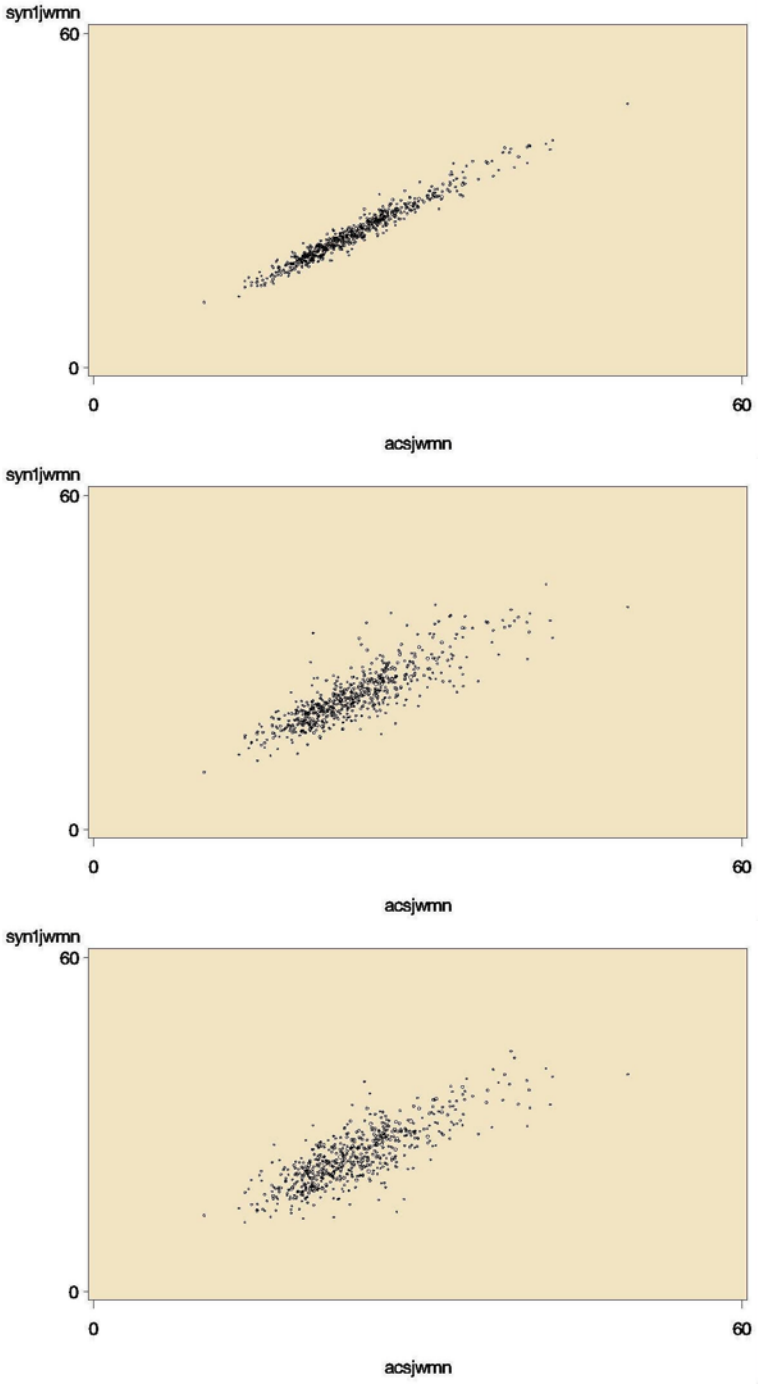


Figure F-8. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for St. Louis' TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

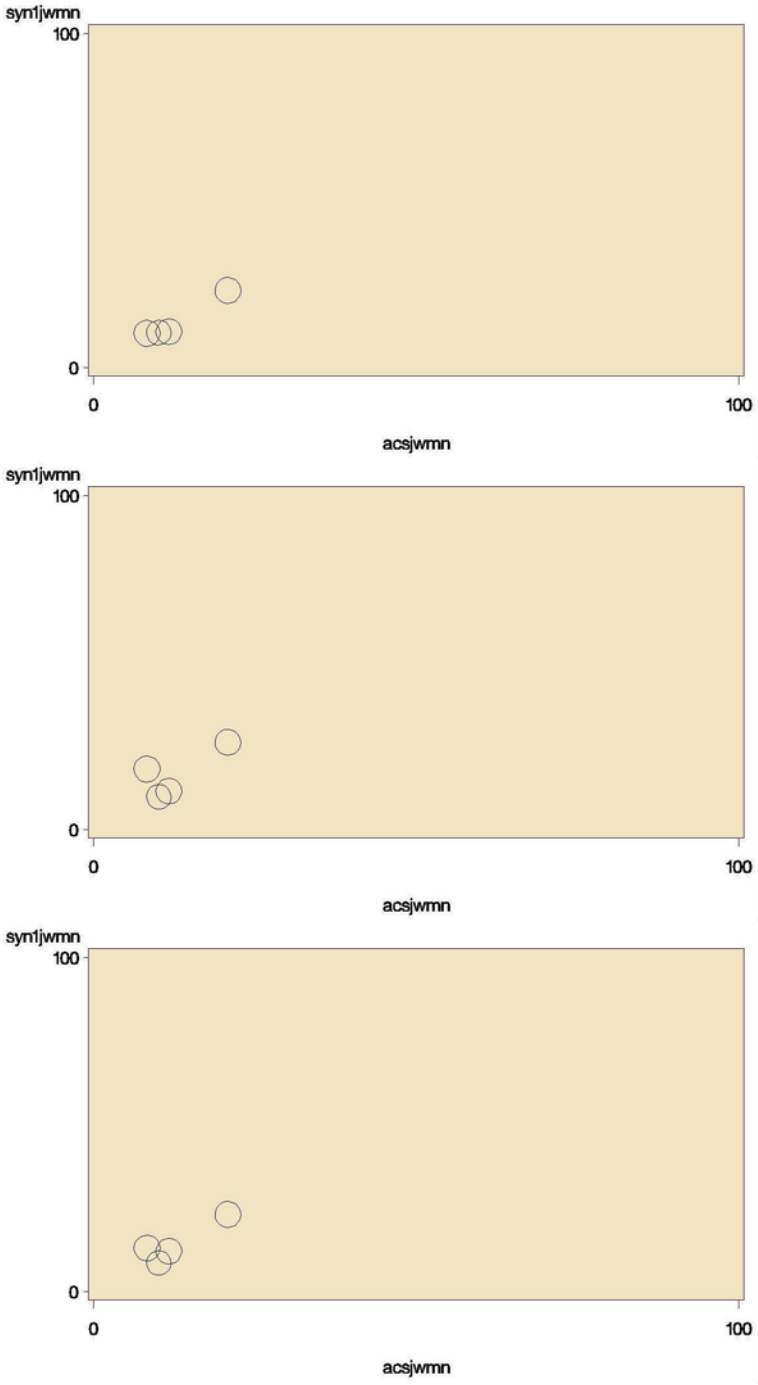


Figure F-9. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Madison's TAZ Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥10)

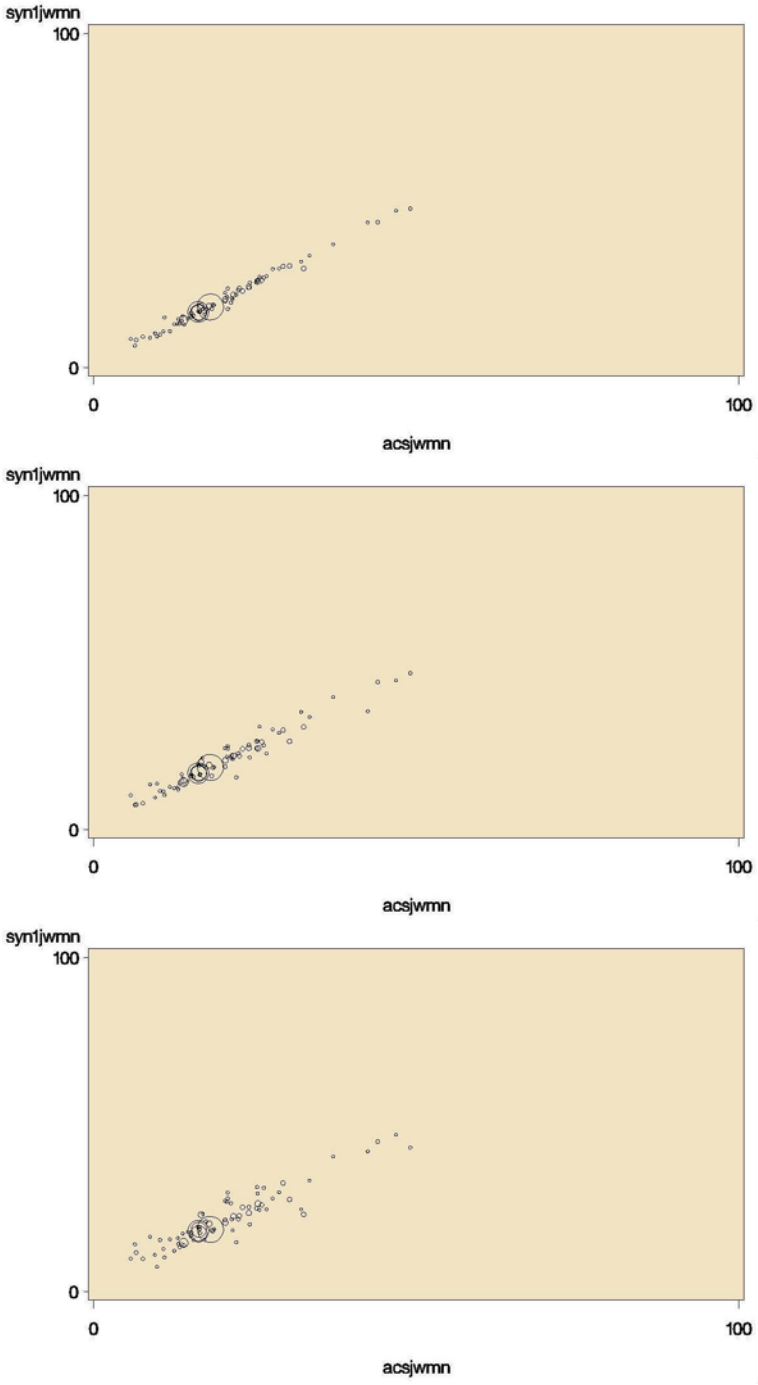


Figure F-10. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's TAZ Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥10)

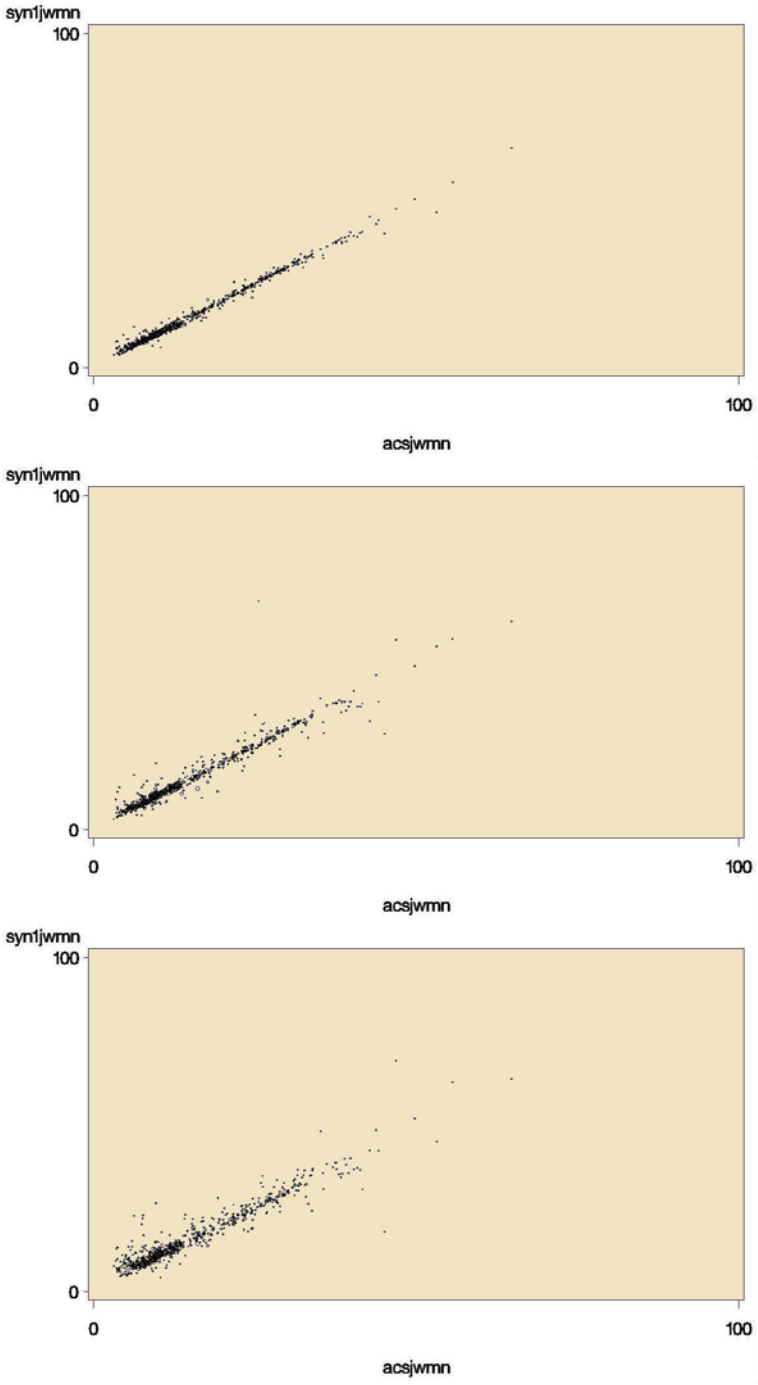


Figure F-11. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Iowa's TAZ Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥10)

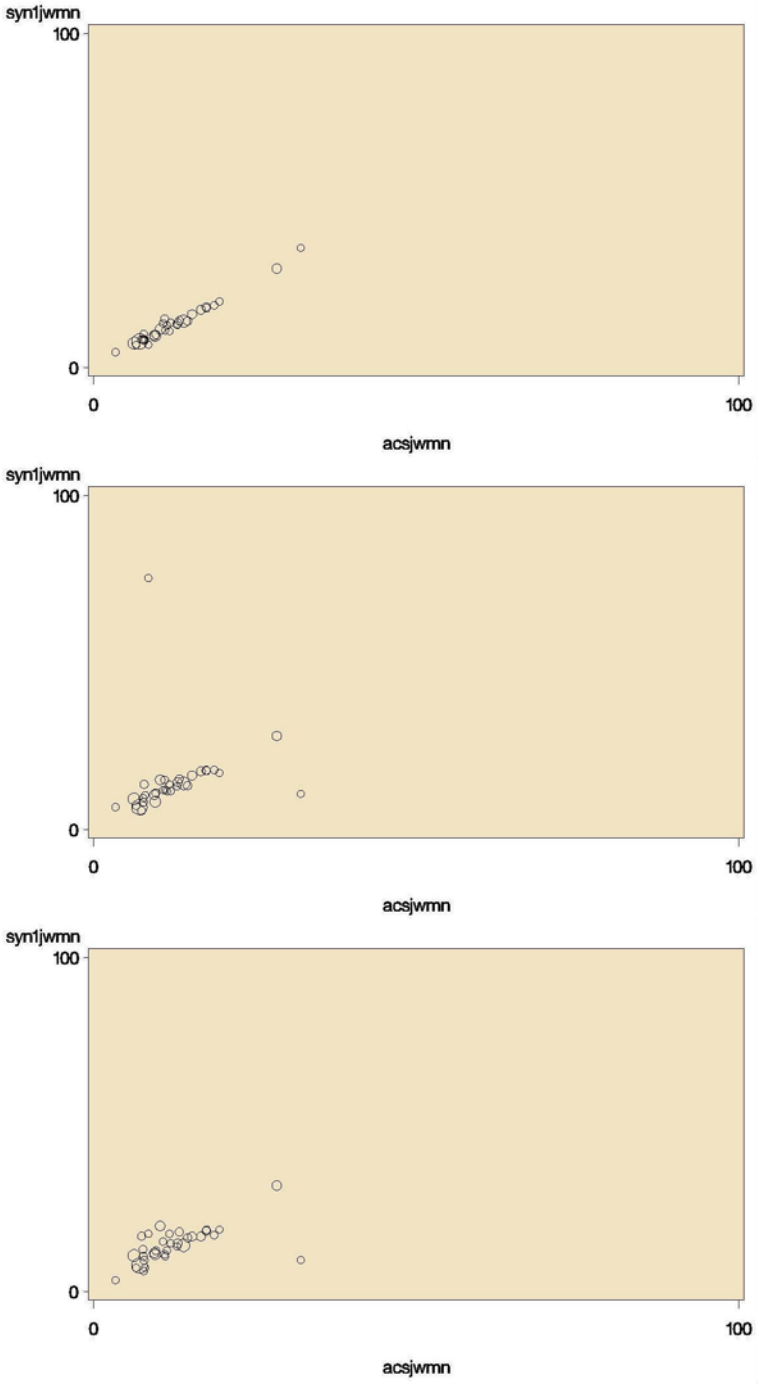


Figure F-12. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for St. Louis' TAZ Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥10)

Appendix G

DEVELOPMENT PHASE: PLOTS OF ACS AND PERTURBED DATA FOR MEAN HOUSEHOLD (HH) INCOME

Note: The dots in Figures G-1 through G-8 are shown only for localities with 30 or more ACS cases.

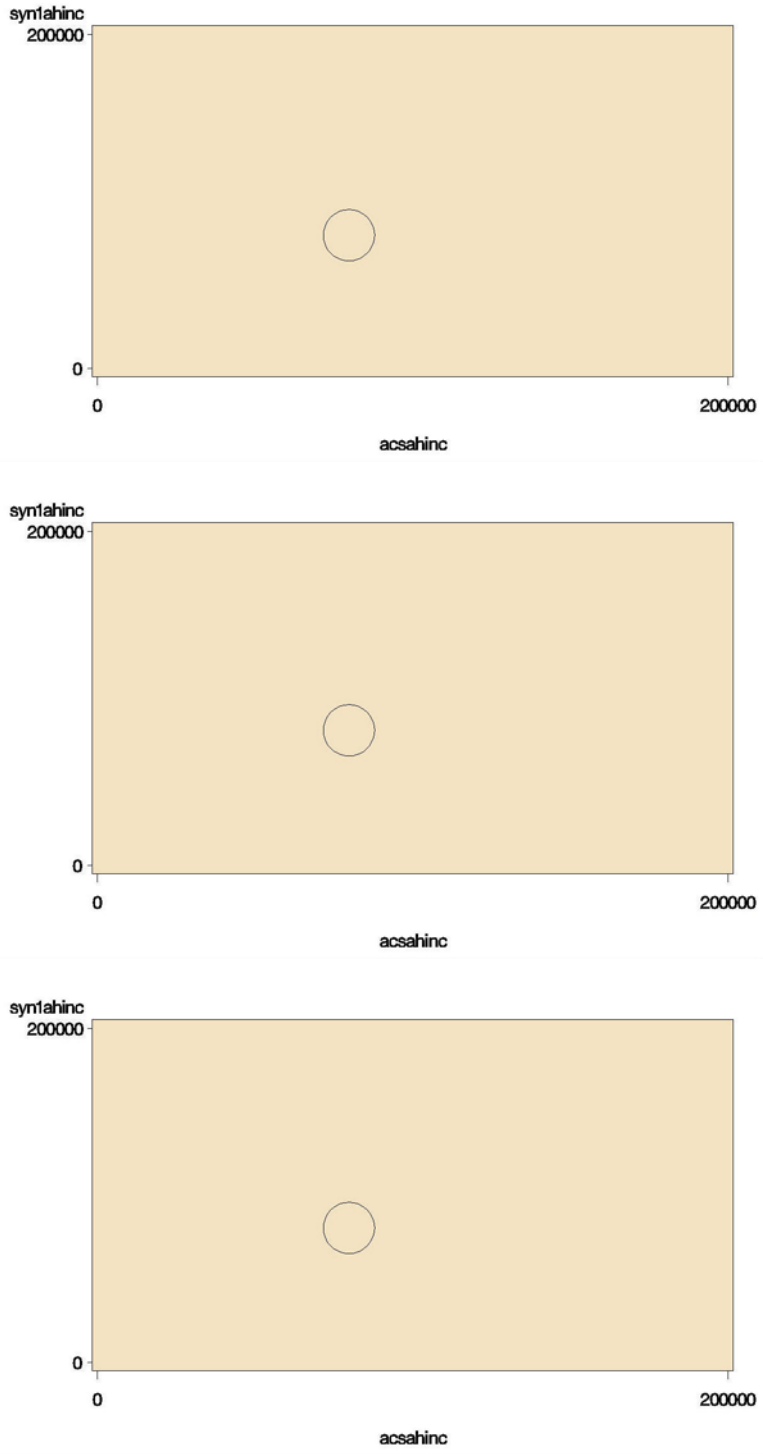


Figure G-1. Bubble Plots of ACS and Partial Perturbed Mean HH Income for Madison's County: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

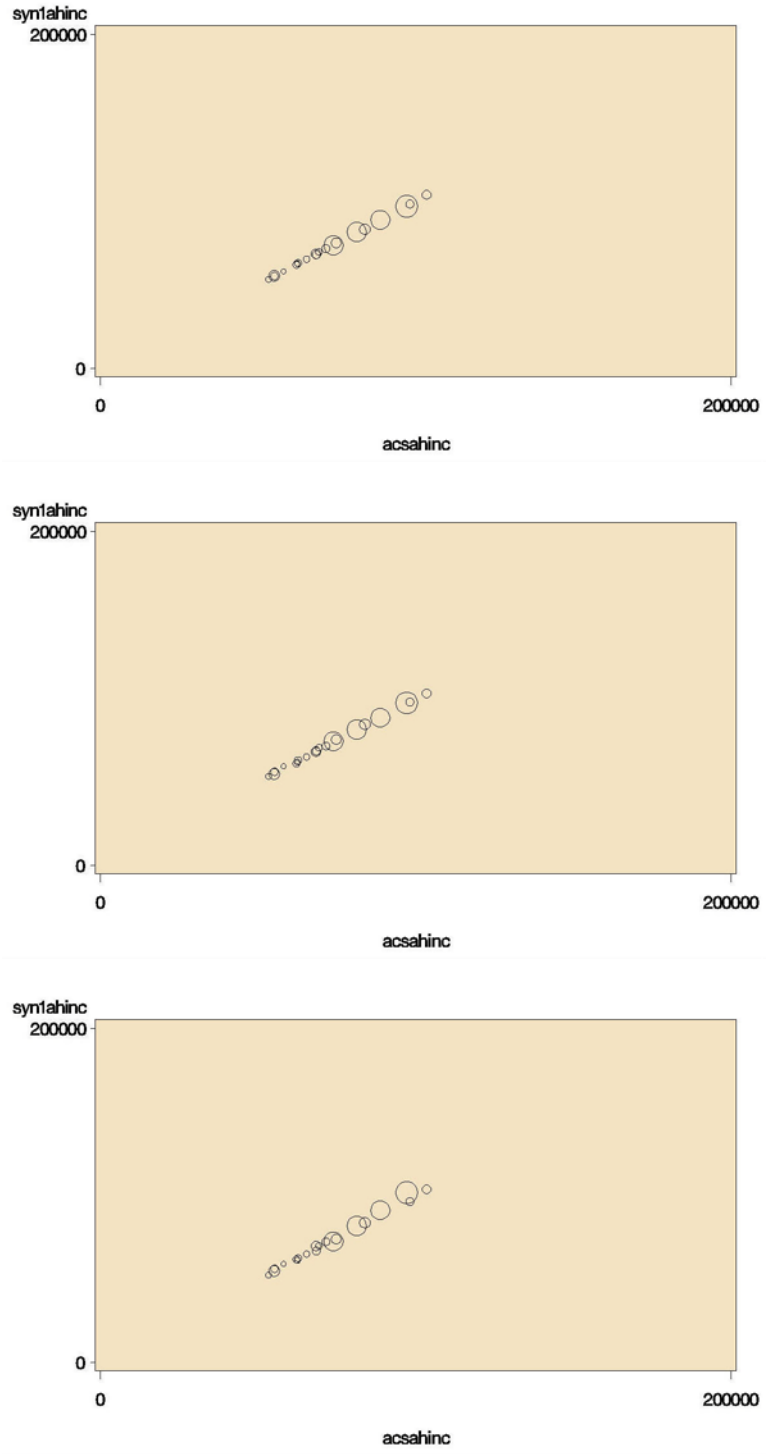


Figure G-2. Bubble Plots of ACS and Partial Perturbed Mean HH Income for Atlanta's Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

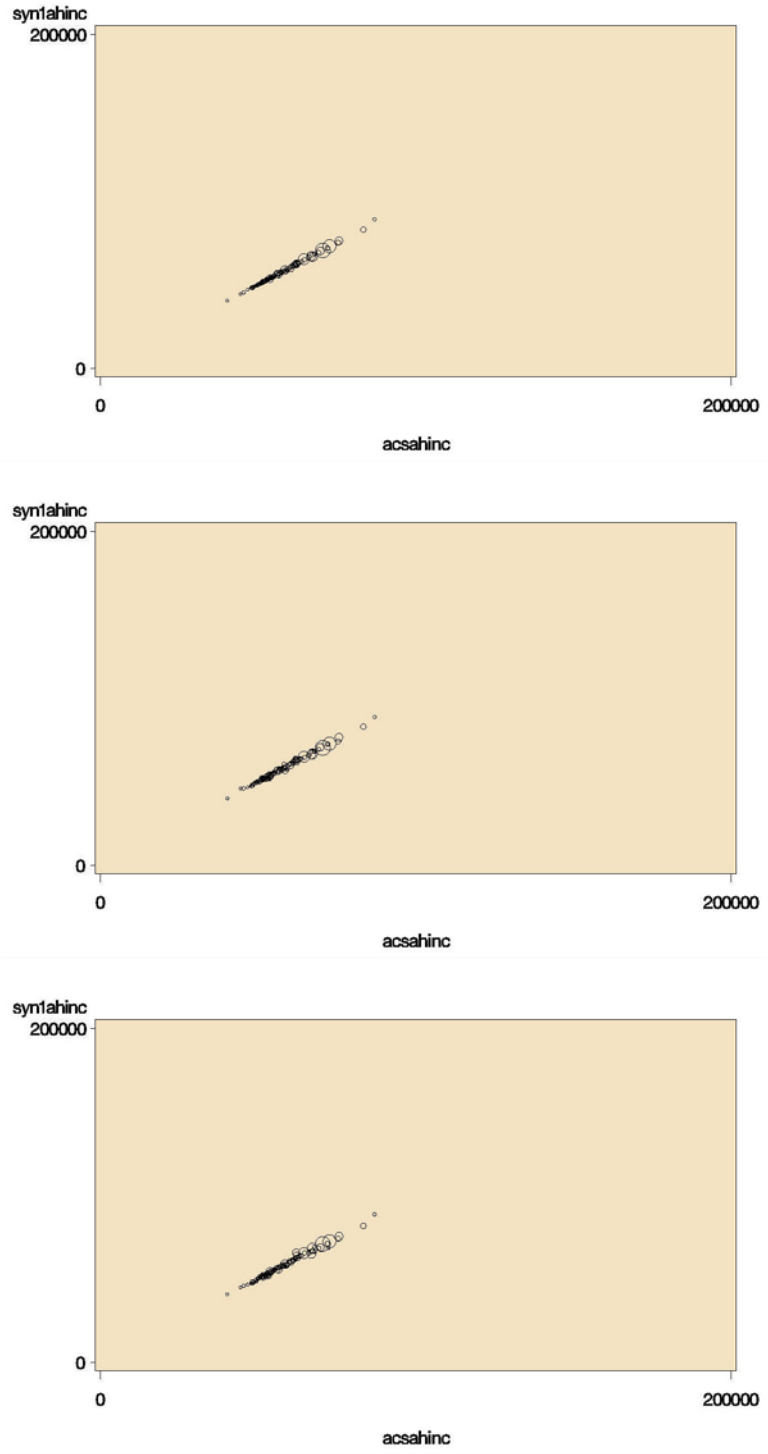


Figure G-3. Bubble Plots of ACS and Partial Perturbed Mean HH Income for Iowa's Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

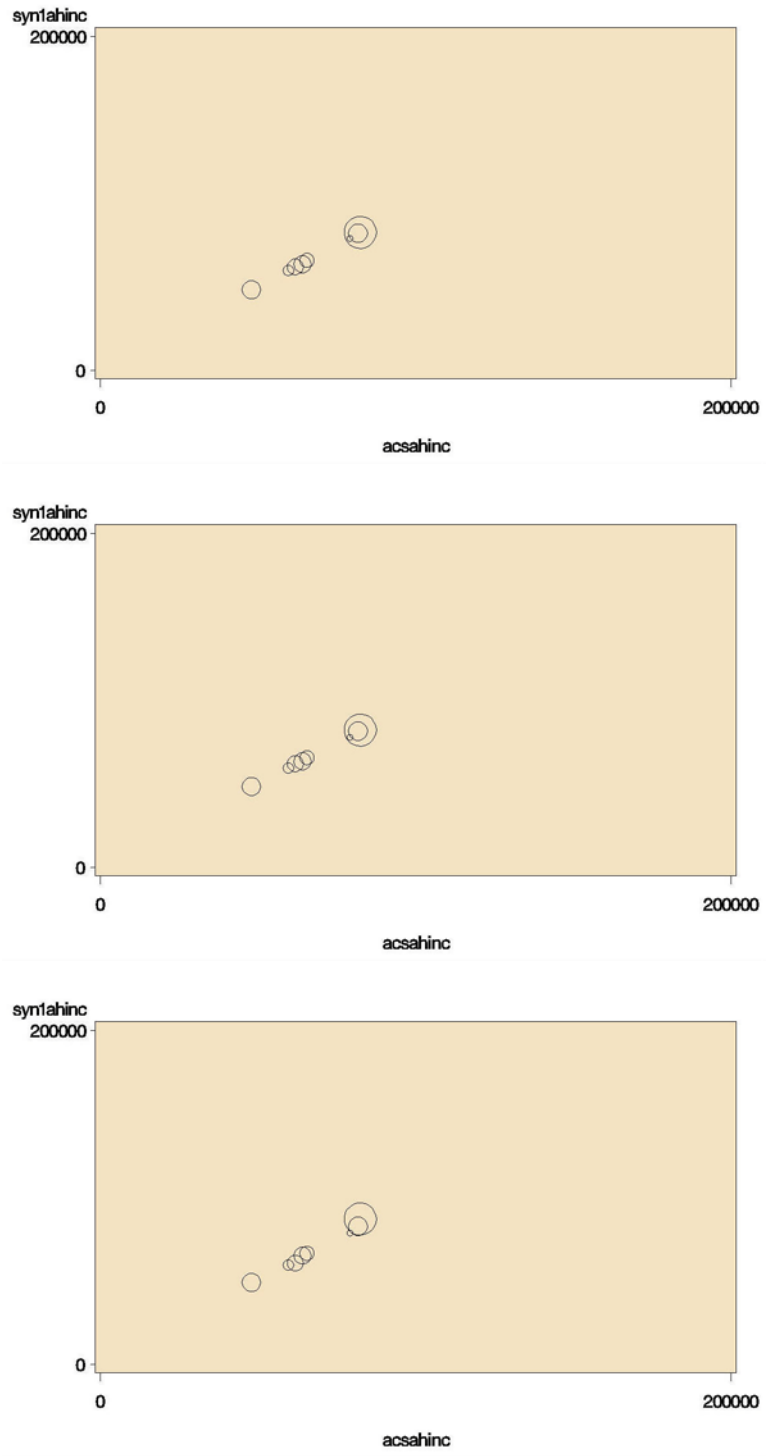


Figure G-4. Bubble Plots of ACS and Partial Perturbed Mean HH Income for St. Louis' Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

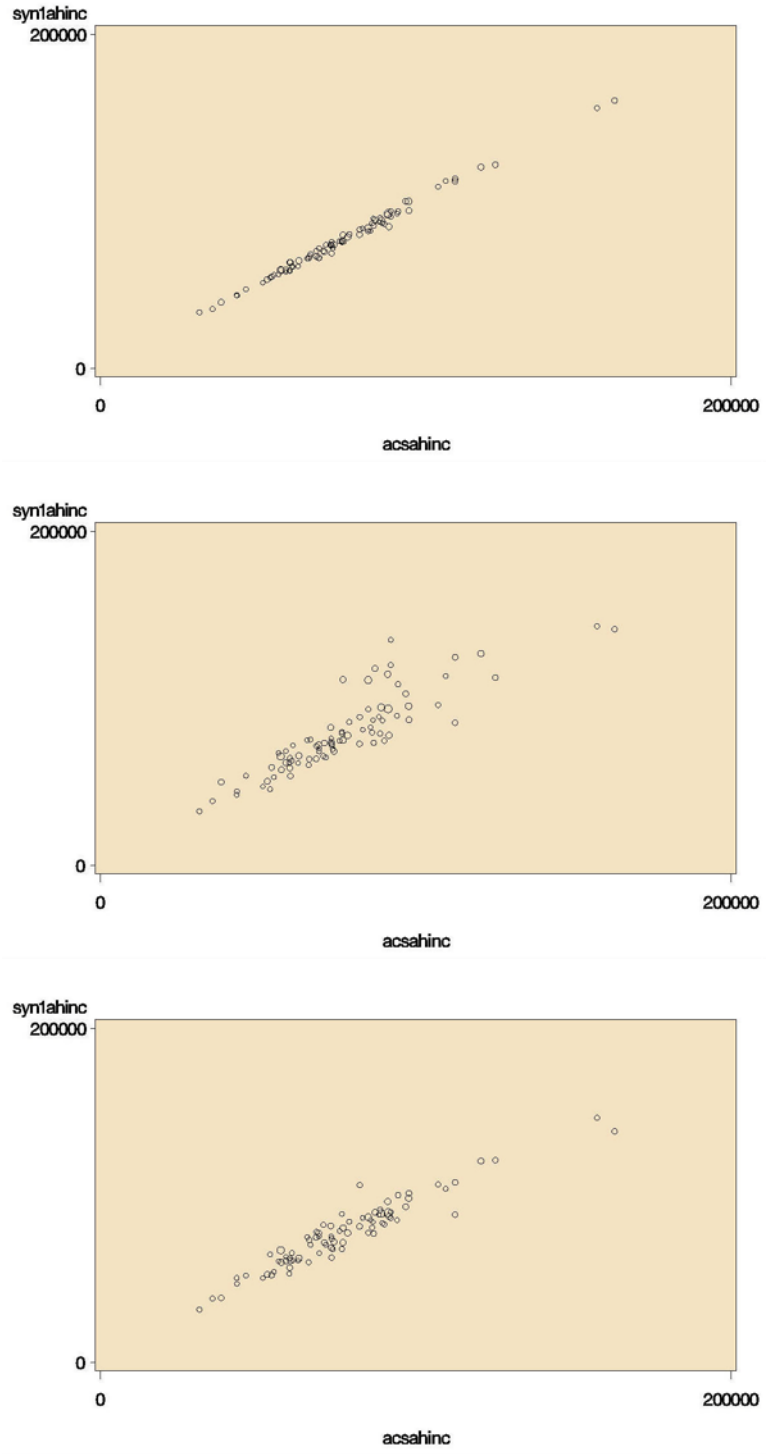


Figure G-5. Bubble Plots of ACS and Partial Perturbed Mean HH Income for Madison's TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

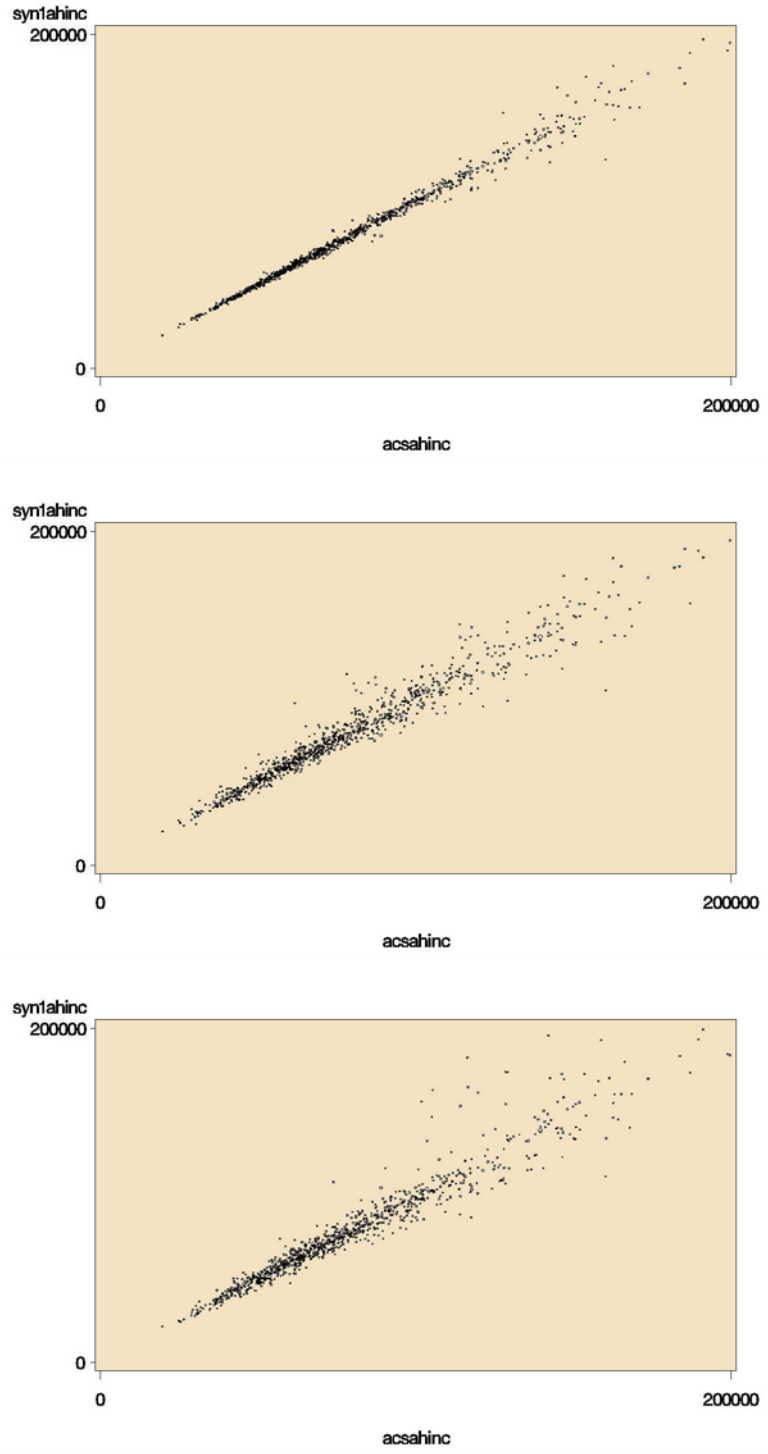


Figure G-6. Bubble Plots of ACS and Partial Perturbed Mean HH Income for Atlanta's TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

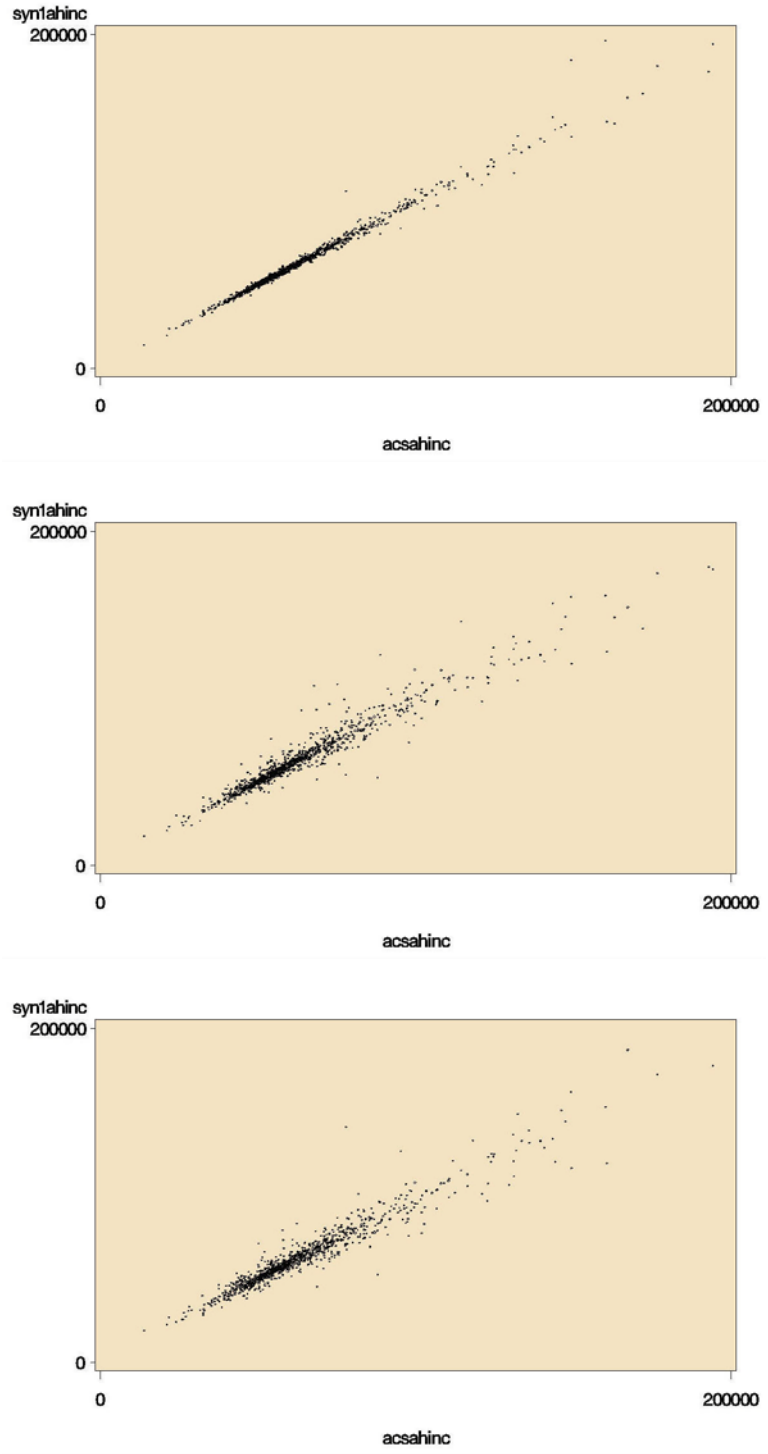


Figure G-7. Bubble Plots of ACS and Partial Perturbed Mean HH Income for Iowa's TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

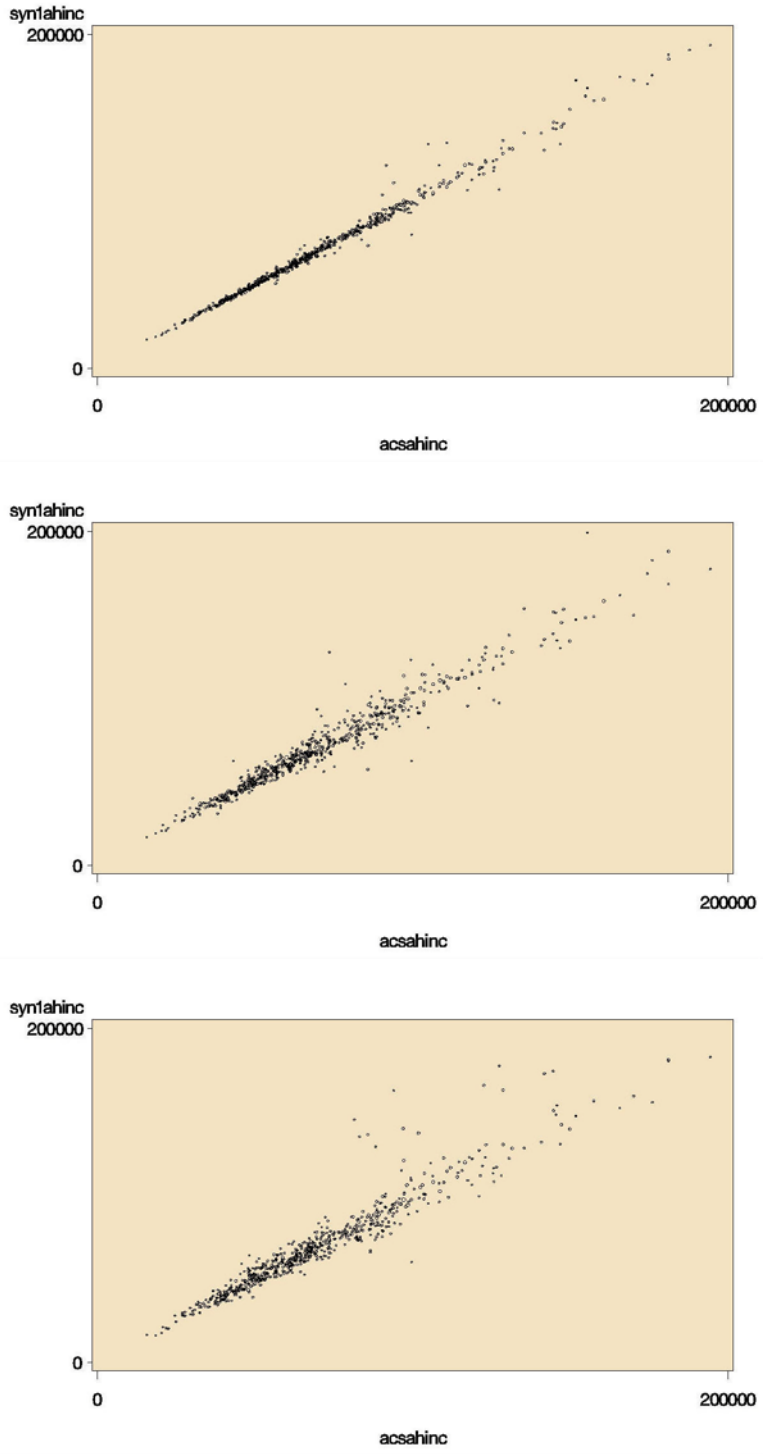


Figure G-8. Bubble Plots of ACS and Partial Perturbed Mean HH Income for St.Louis' TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric (n≥30)

Appendix H

DEVELOPMENT PHASE: PLOTS OF ACS AND PERTURBED DATA FOR WEIGHTED COUNTS – FULL REPLACEMENT

Note: For the county and TAZ level plots, a dot is shown only if there are at least 10 sample cases and 10 perturbed cases in the cell. For county flow plots, a dot is shown only for flows with at least 5 sample cases and 5 perturbed cases in the cell.

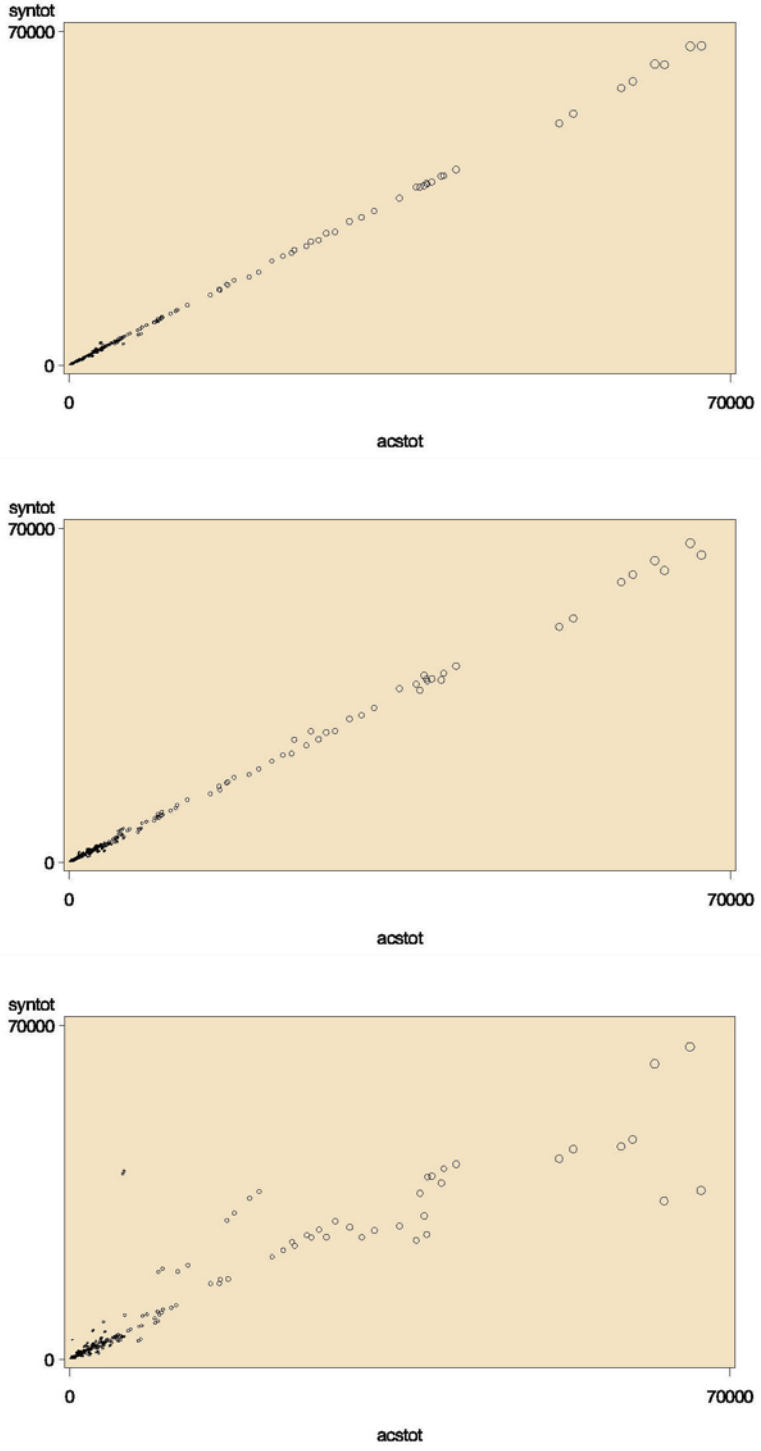


Figure H-1. Plot of ACS and Fully Perturbed Weighted Counts for Madison's County: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

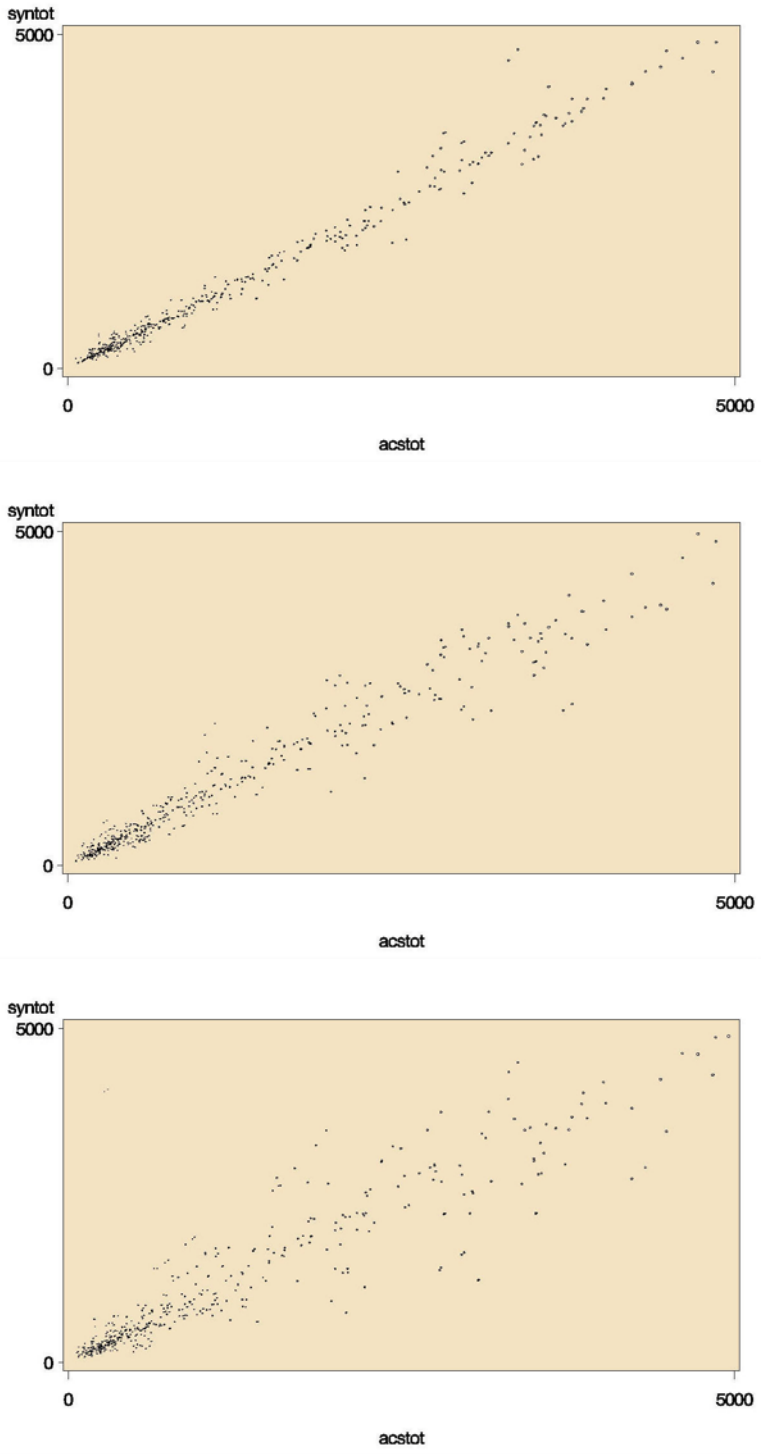


Figure H-2. Plot of ACS and Fully Perturbed Weighted Counts for Madison TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

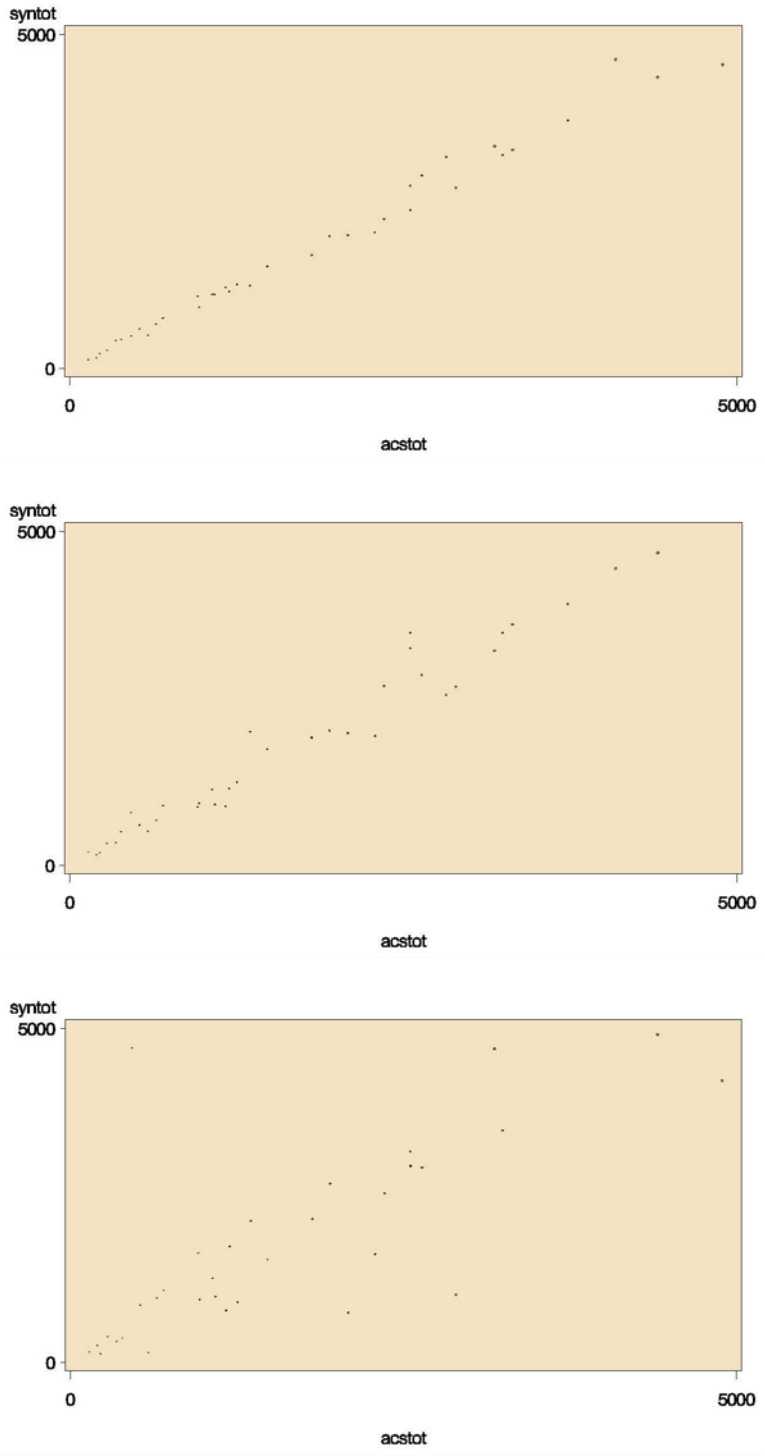


Figure H-3. Plot of ACS and Fully Perturbed Weighted Counts for Madison County Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

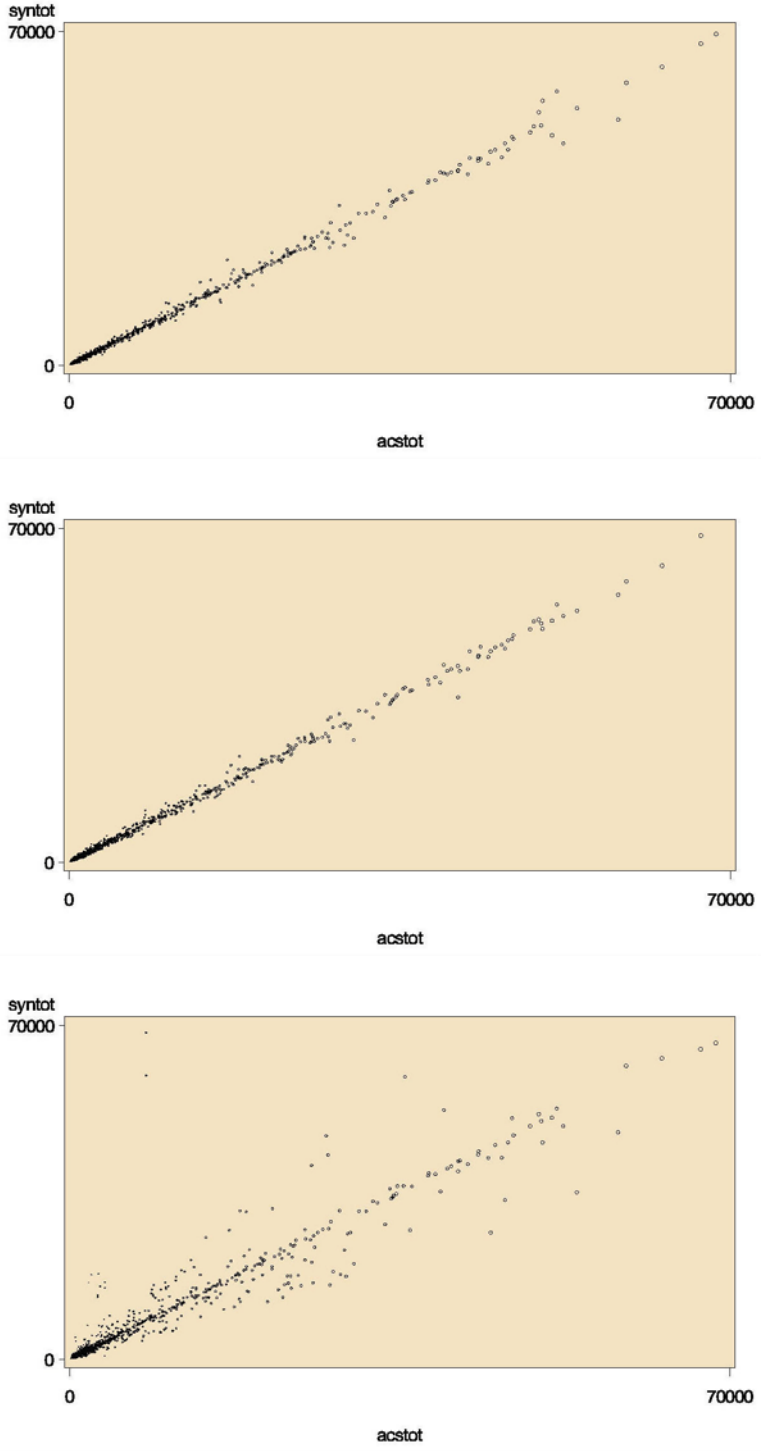


Figure H-4. Plot of ACS and Fully Perturbed Weighted Counts for St. Louis Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

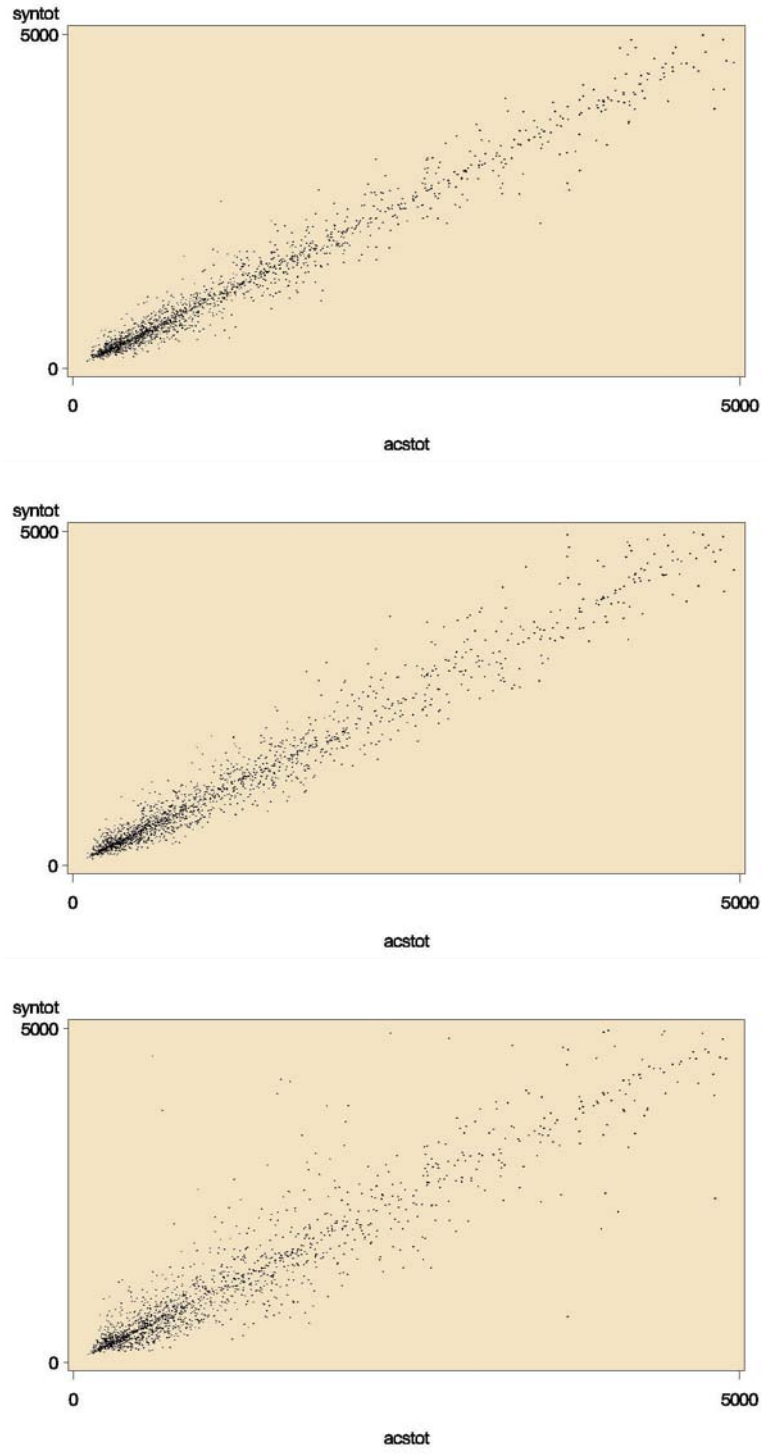


Figure H-5. Plot of ACS and Fully Perturbed Weighted Counts for St. Louis TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

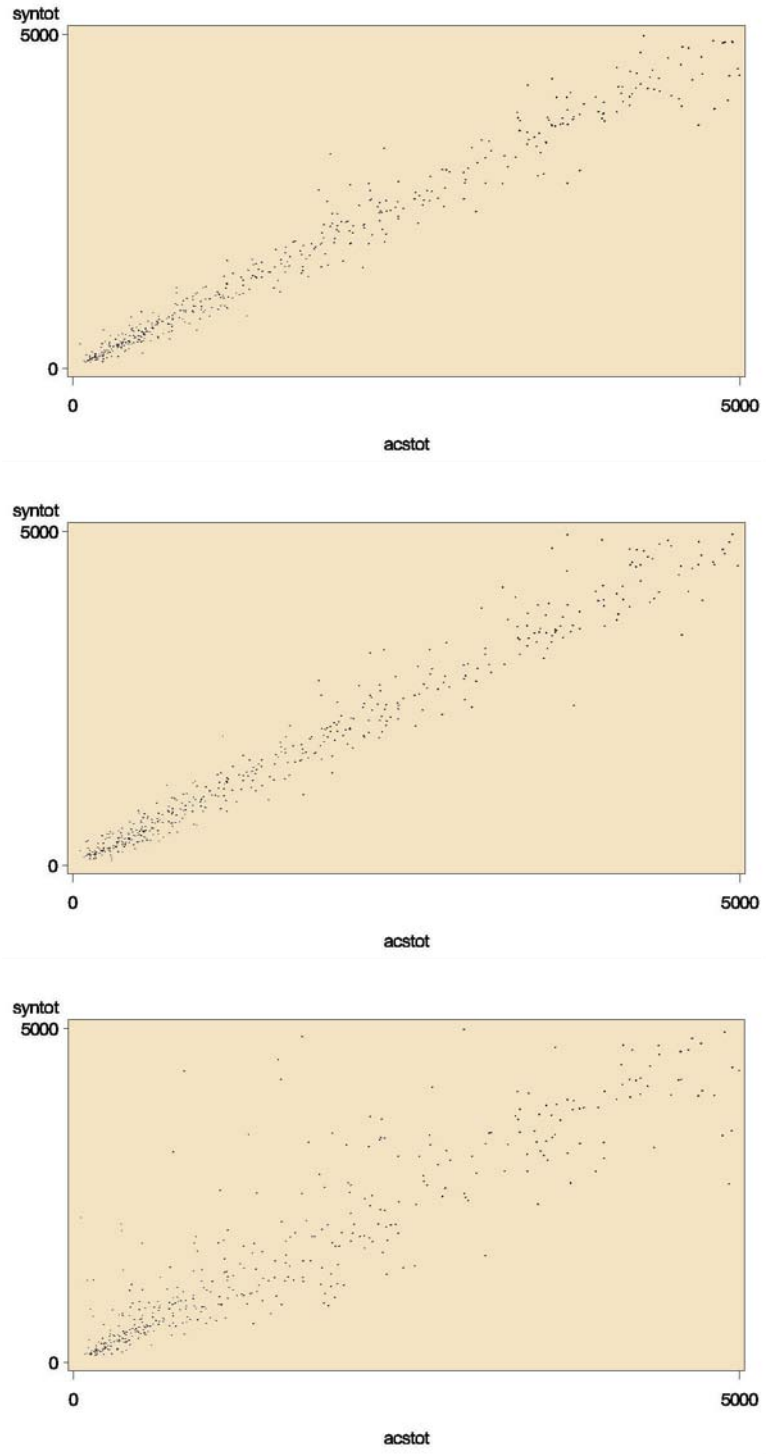


Figure H-6. Plot of ACS and Fully Perturbed Weighted Counts for St. Louis County Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

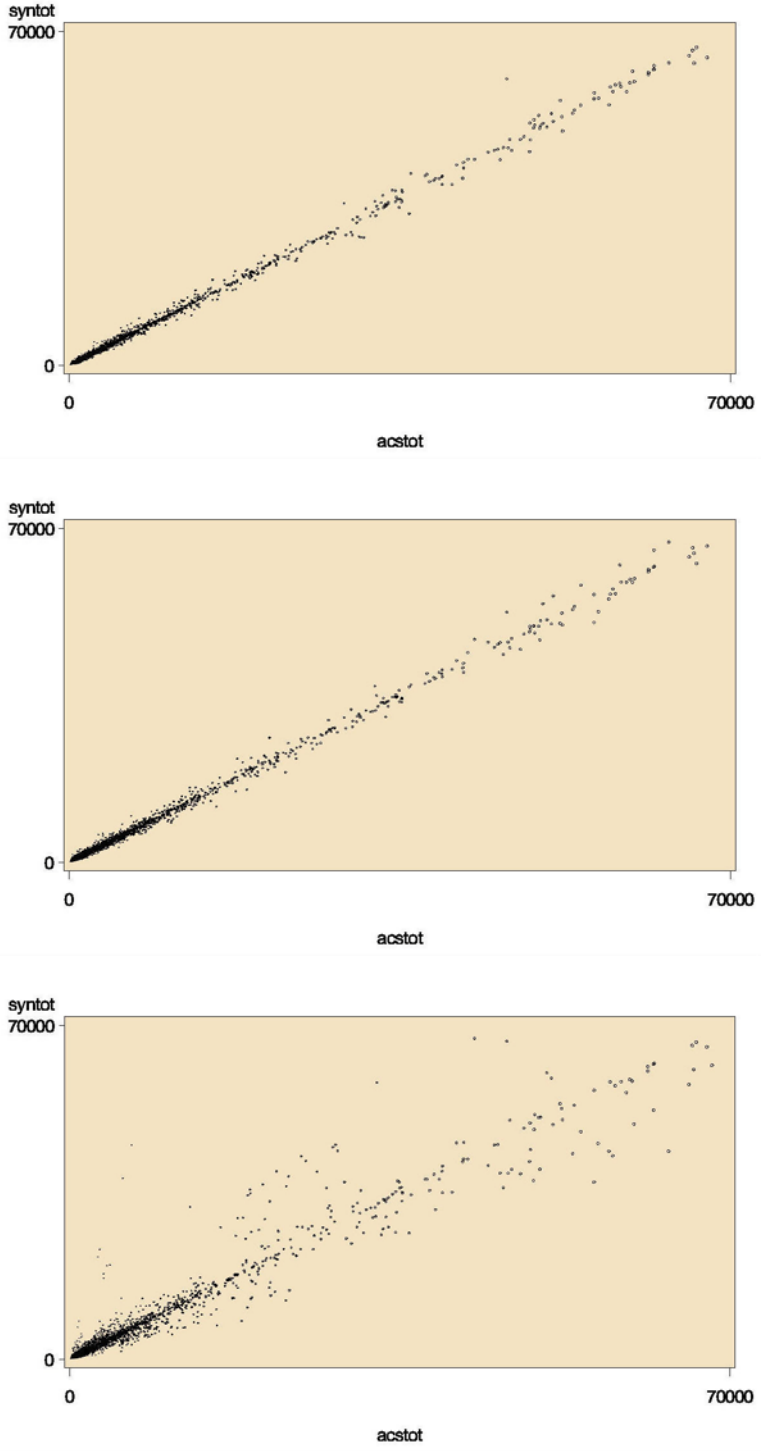


Figure H-7. Plot of ACS and Fully Perturbed Weighted Counts for Atlanta Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

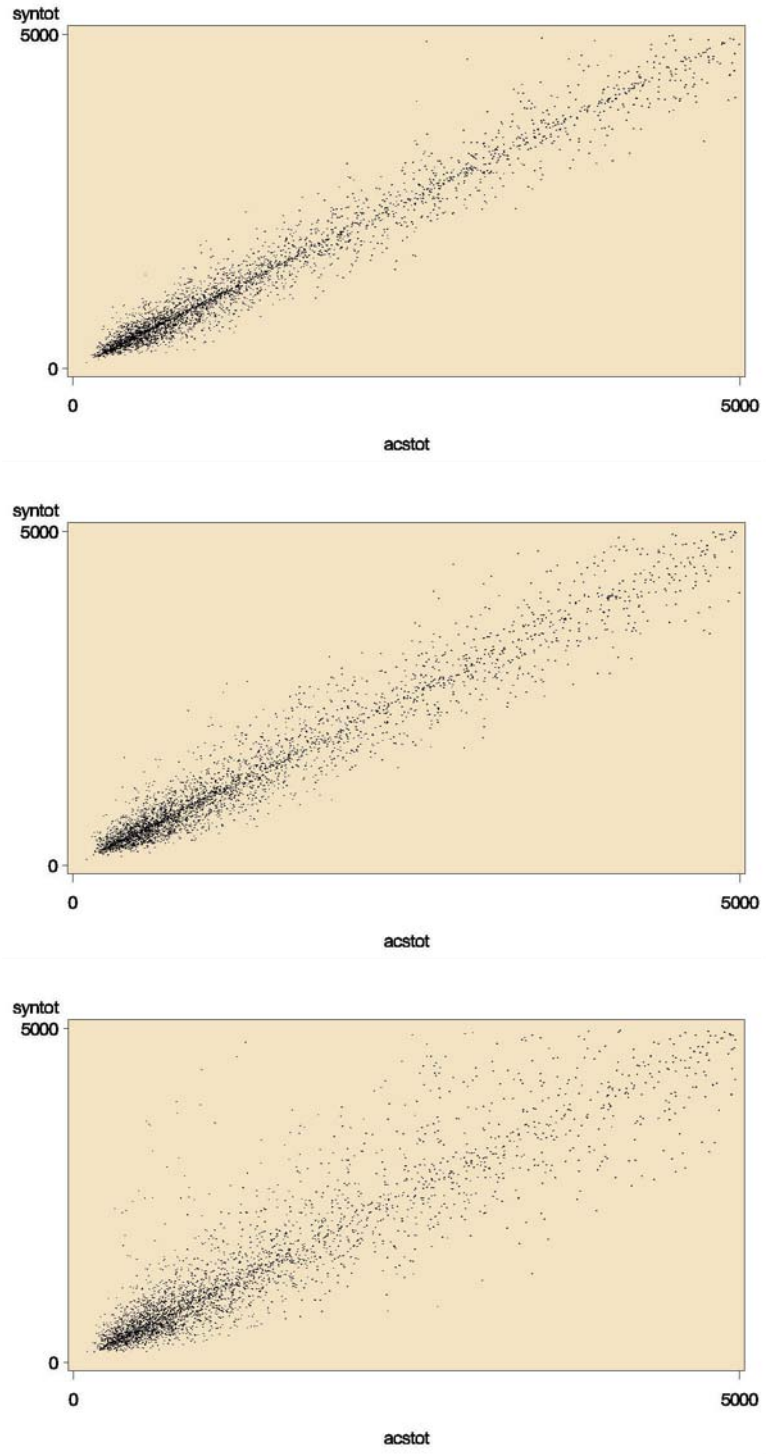


Figure H-8. Plot of ACS and Fully Perturbed Weighted Counts for Atlanta TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

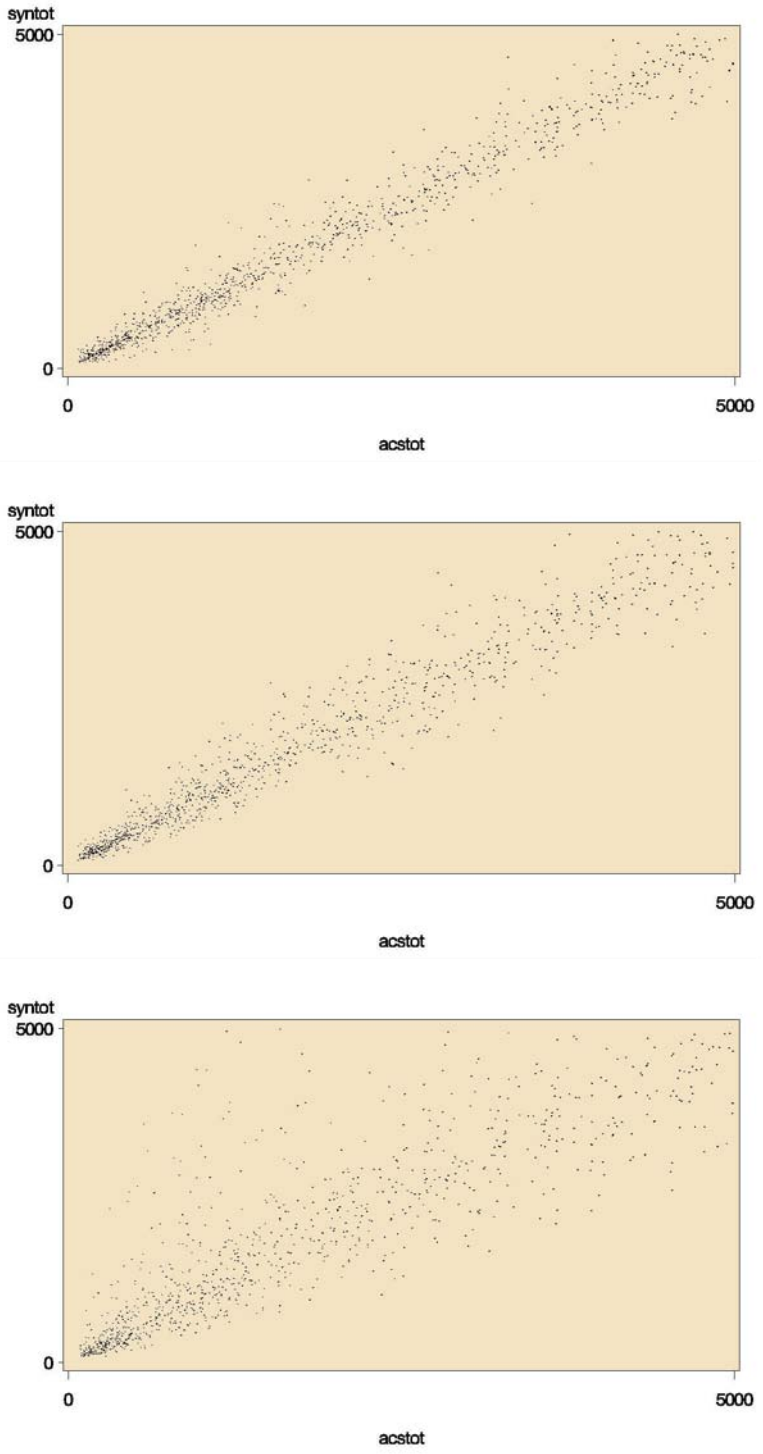


Figure H-9. Plot of ACS and Fully Perturbed Weighted Counts for Atlanta County Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

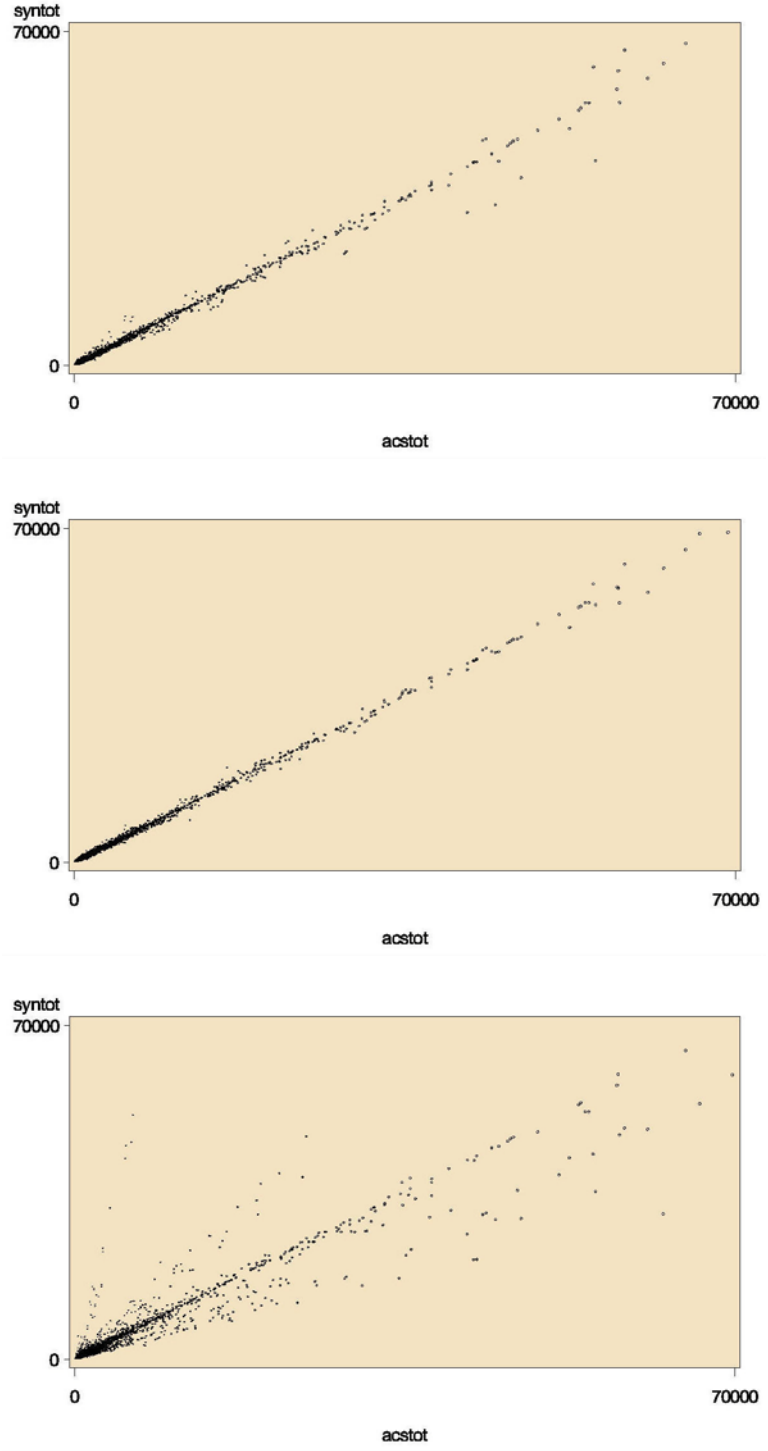


Figure H-10. Plot of ACS and Fully Perturbed Weighted Counts for Iowa Counties: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

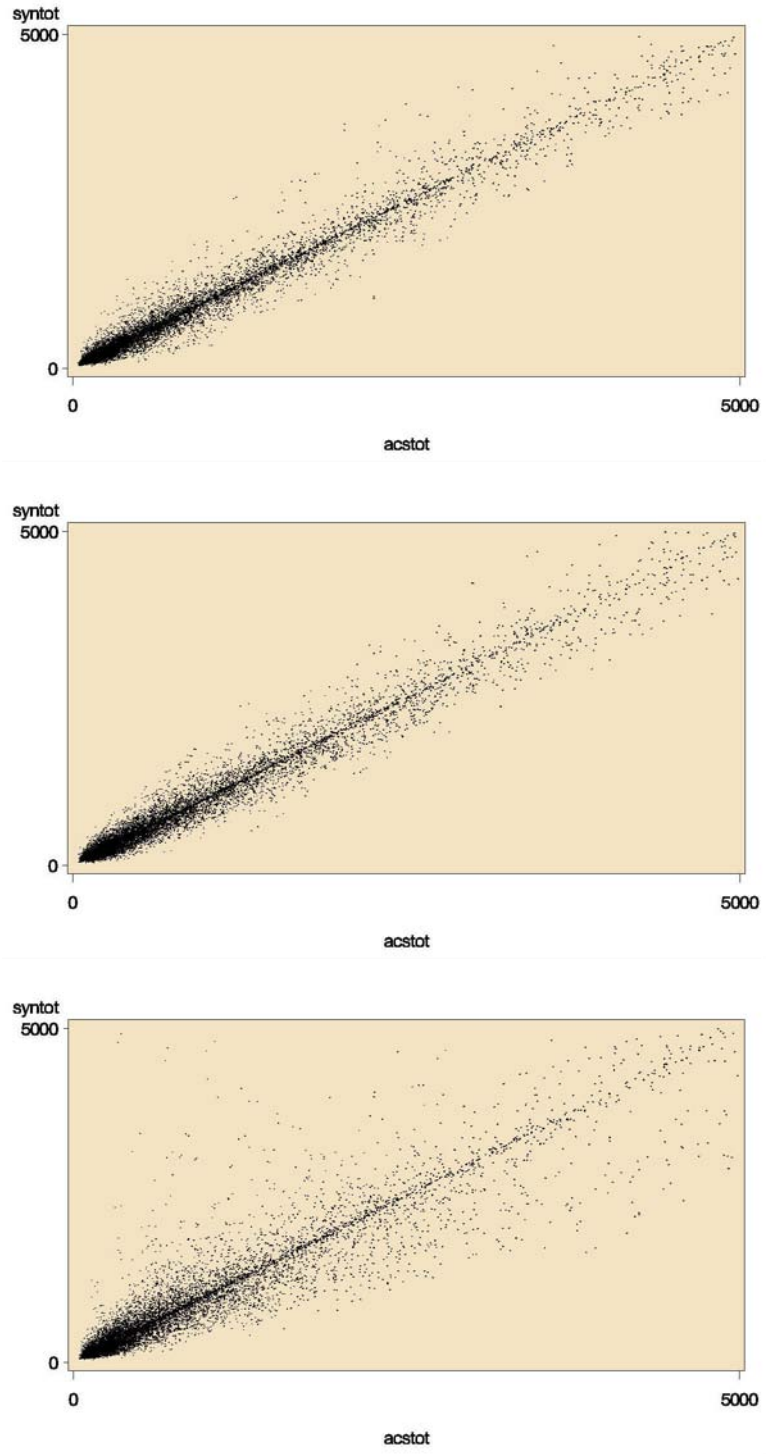


Figure H-11. Plot of ACS and Fully Perturbed Weighted Counts for Iowa TAZs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

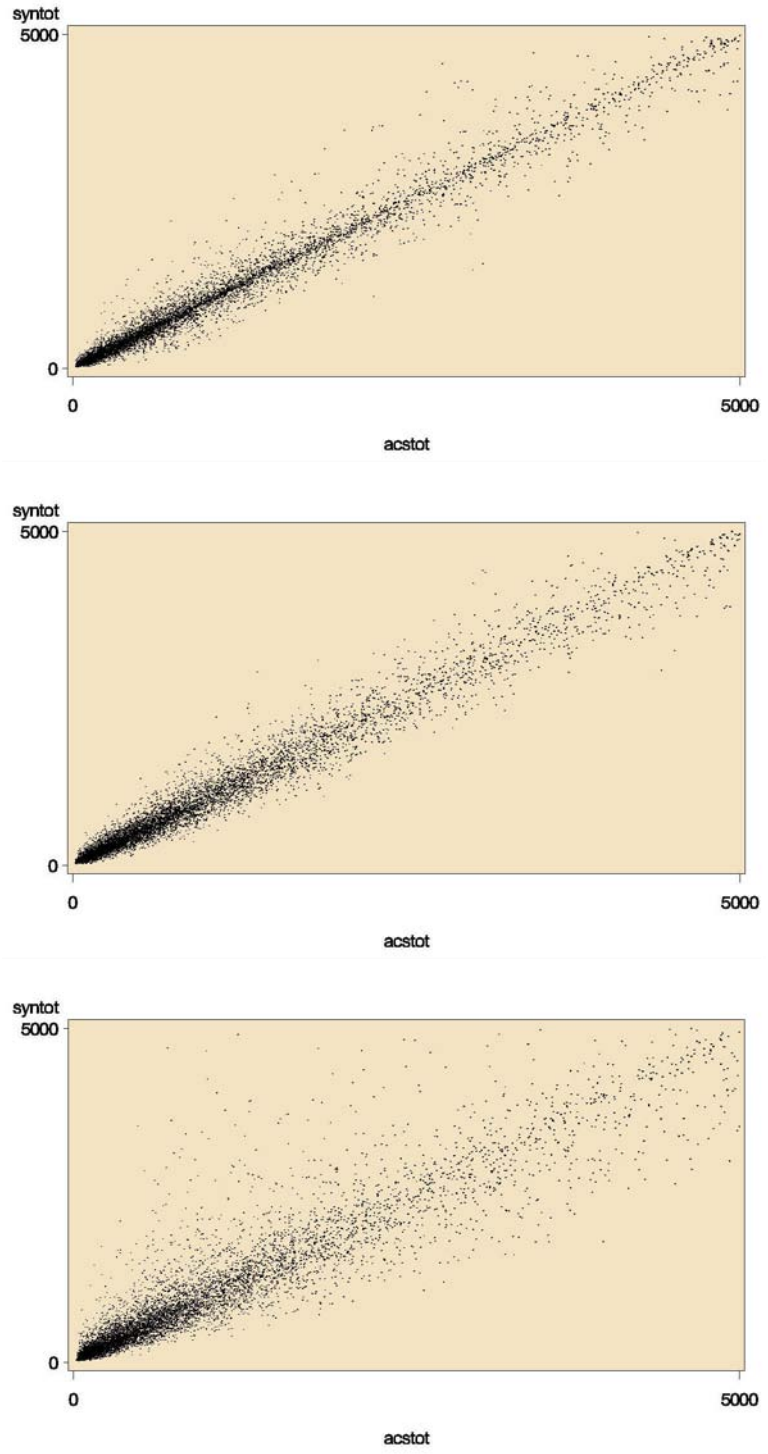


Figure H-12. Plot of ACS and Fully Perturbed Weighted Counts for Iowa County Flows: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

Appendix I

DEVELOPMENT PHASE: PLOTS OF ACS AND CONSTRAINED HOTDECK DATA FOR WEIGHTED COUNTS – PARTIAL REPLACEMENT

Note: For the county and TAZ level plots, a dot is shown only if there are at least 10 sample cases and 10 perturbed cases in the cell. For county flow plots, a dot is shown only for flows with at least 5 sample cases and 5 perturbed cases in the cell.

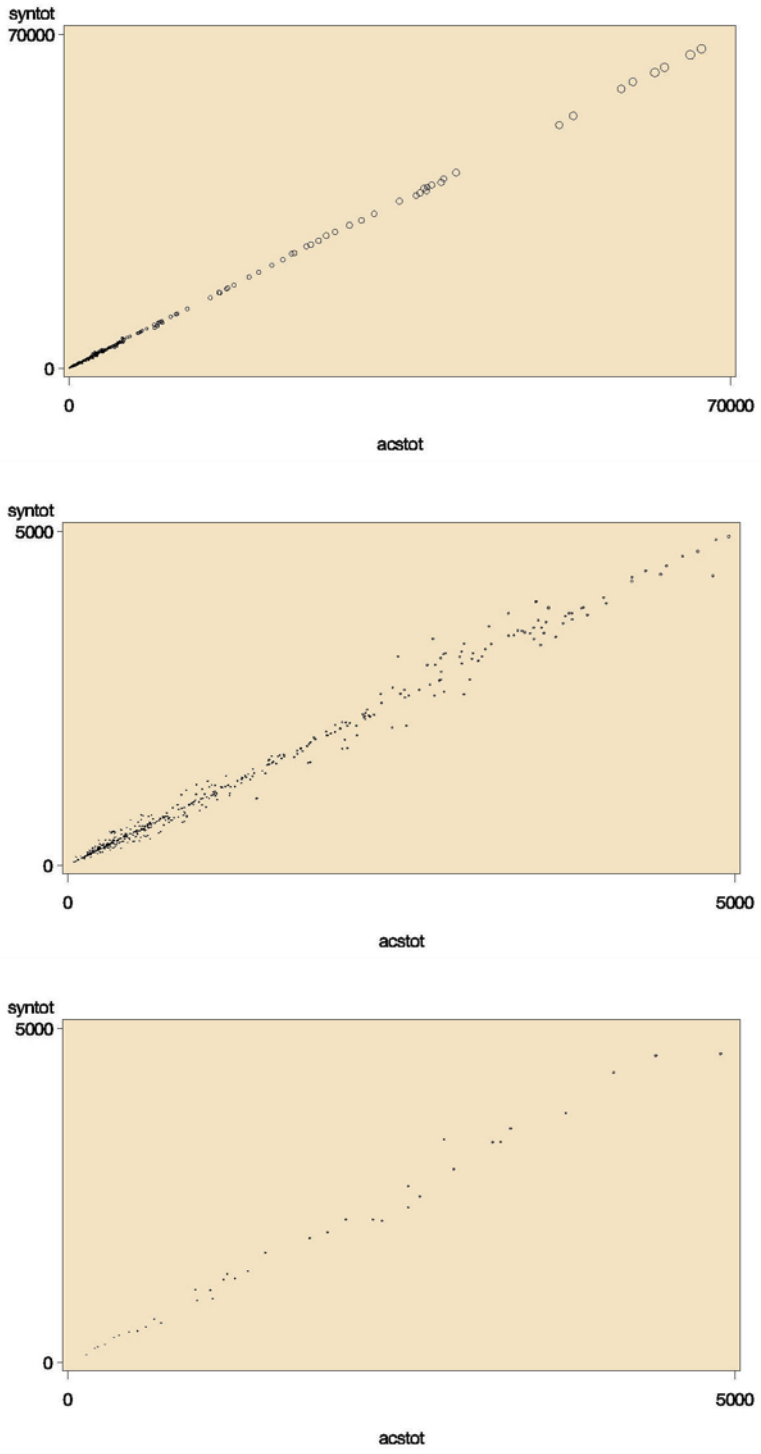


Figure I-1. Plot of ACS and Partial Constrained Hotdeck Weighted Counts for Madison: Top: Counties, Middle: TAZs, Bottom: County Flows

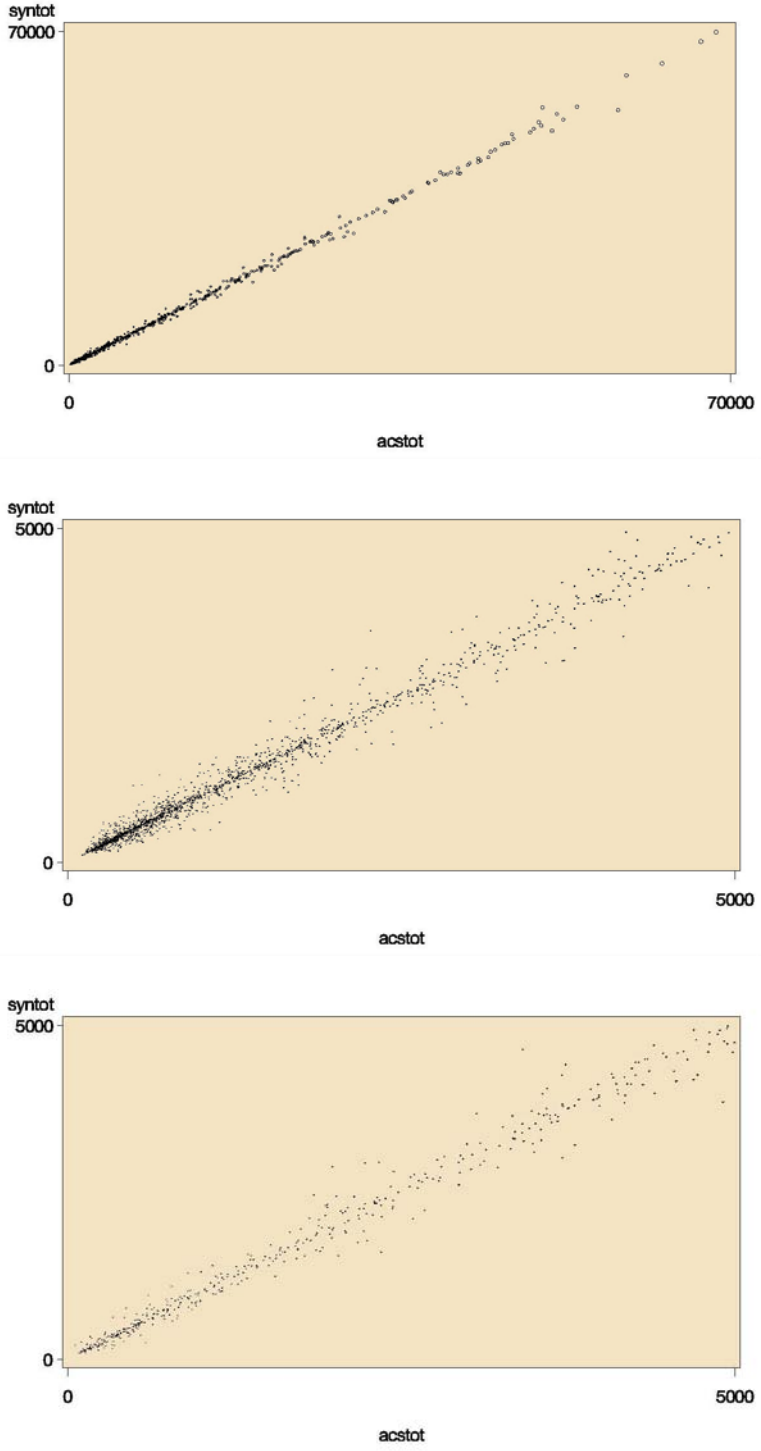


Figure I-2. Plot of ACS and Partial Constrained Hotdeck Weighted Counts for St. Louis: Top: Counties, Middle: TAZs, Bottom: County Flows

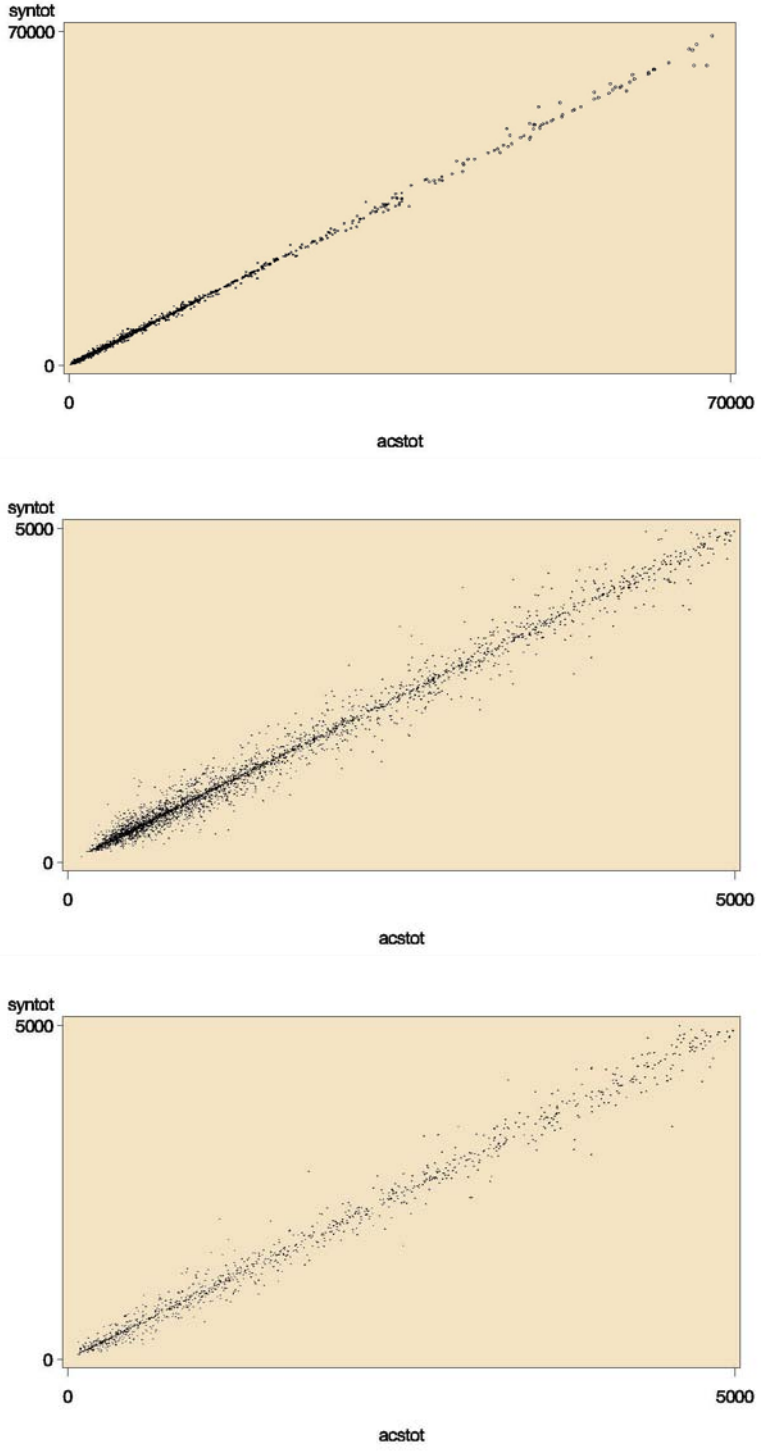


Figure I-3. Plot of ACS and Partial Constrained Hotdeck Weighted Counts for Atlanta: Top: Counties, Middle: TAZs, Bottom: County Flows

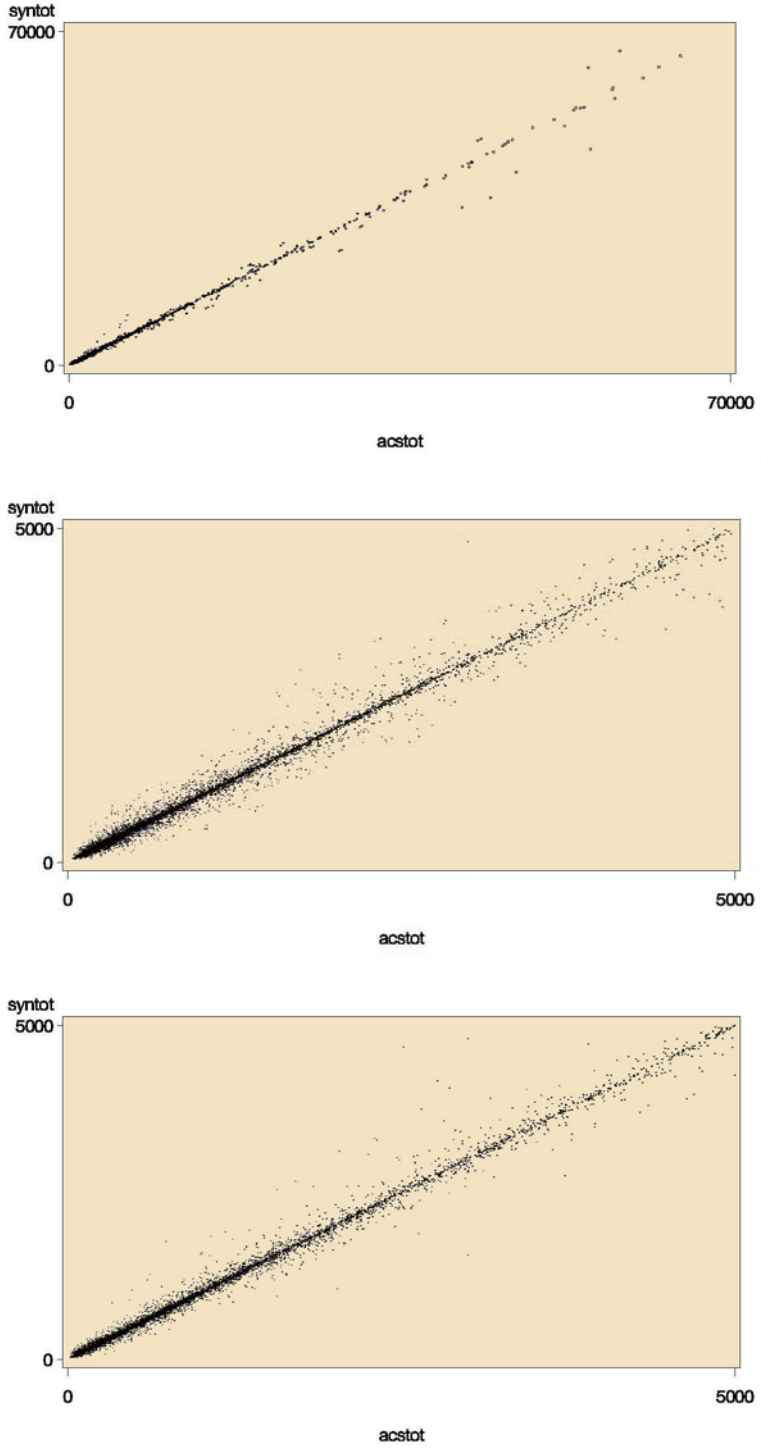


Figure I-4. Plot of ACS and Partial Constrained Hotdeck Weighted Counts for Iowa: Top: Counties, Middle: TAZs, Bottom: County Flows

Appendix J

DEVELOPMENT PHASE: PLOTS OF ACS AND PERTURBED DATA FOR PAIRWISE CORRELATIONS

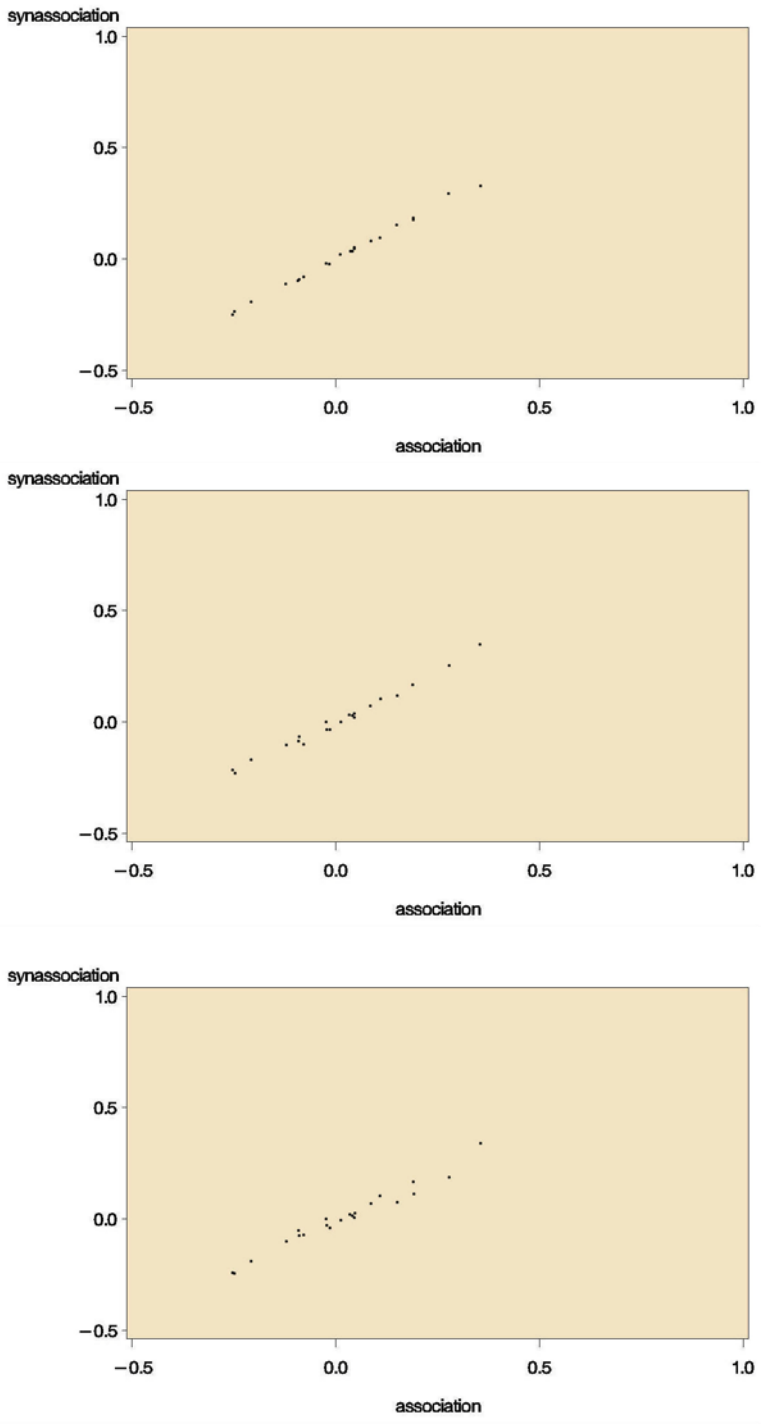


Figure J-1. Plots of ACS and Partial Perturbed Pairwise Correlations for Madison's PUMAs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

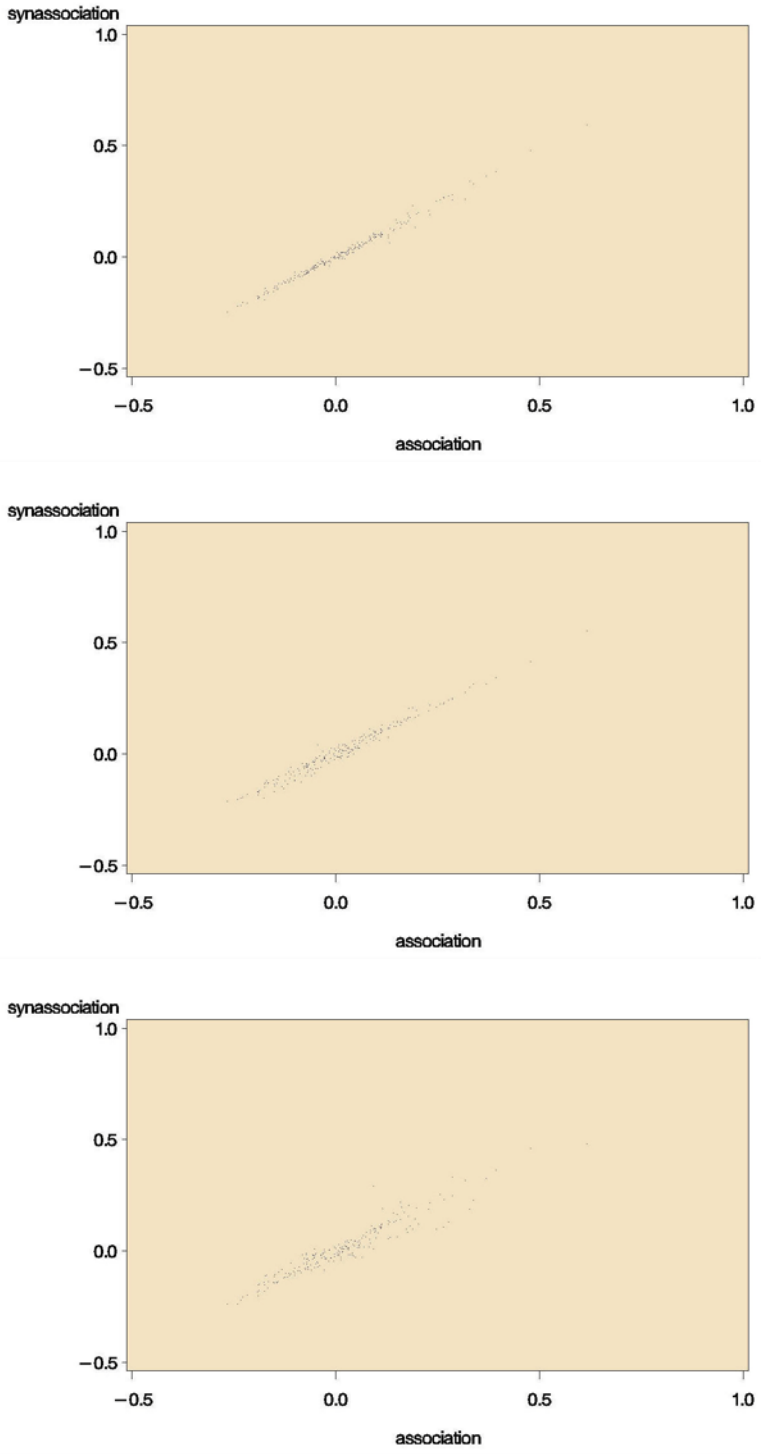


Figure J-2. Plots of ACS and Partial Perturbed Pairwise Correlations for Atlanta's PUMAs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

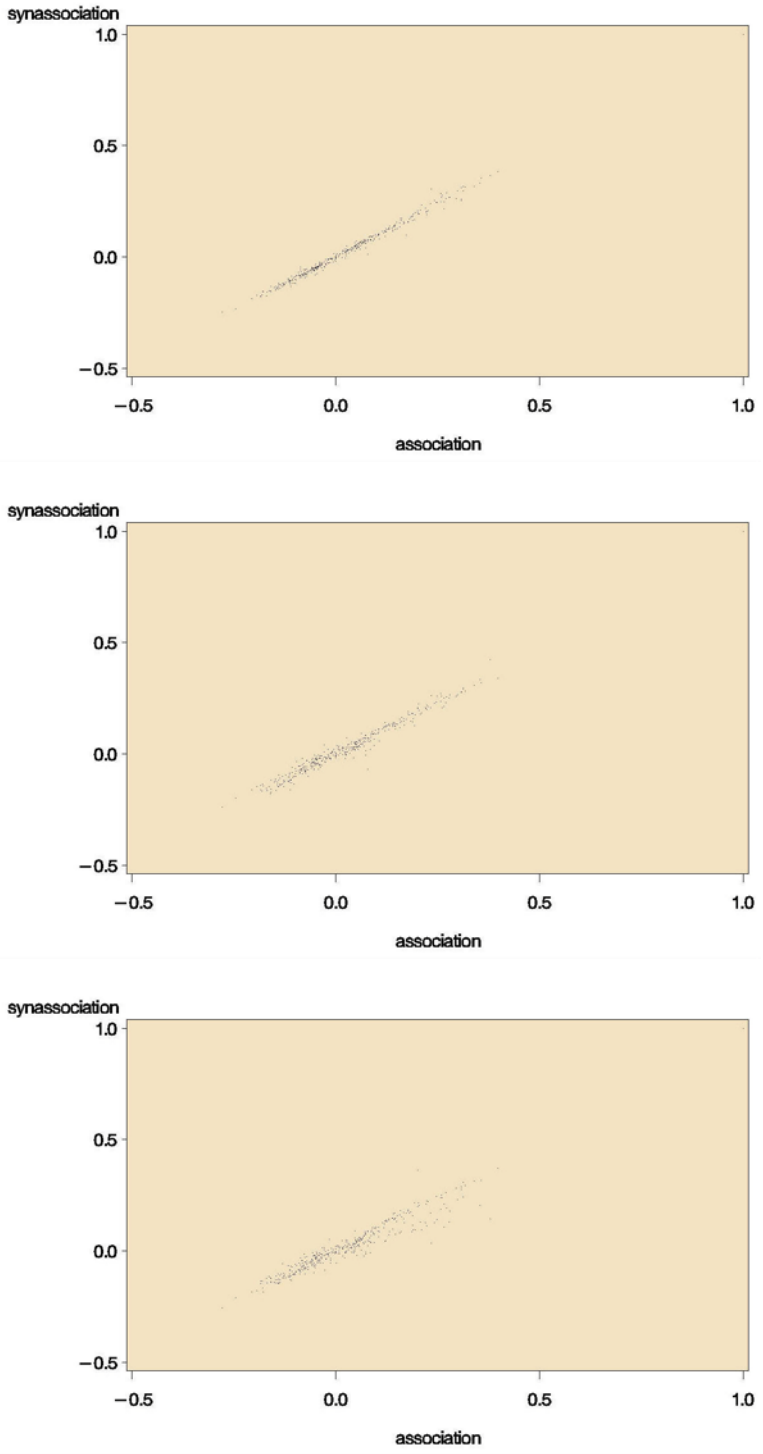


Figure J-3. Plots of ACS and Partial Perturbed Pairwise Correlations for Atlanta's PUMAs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

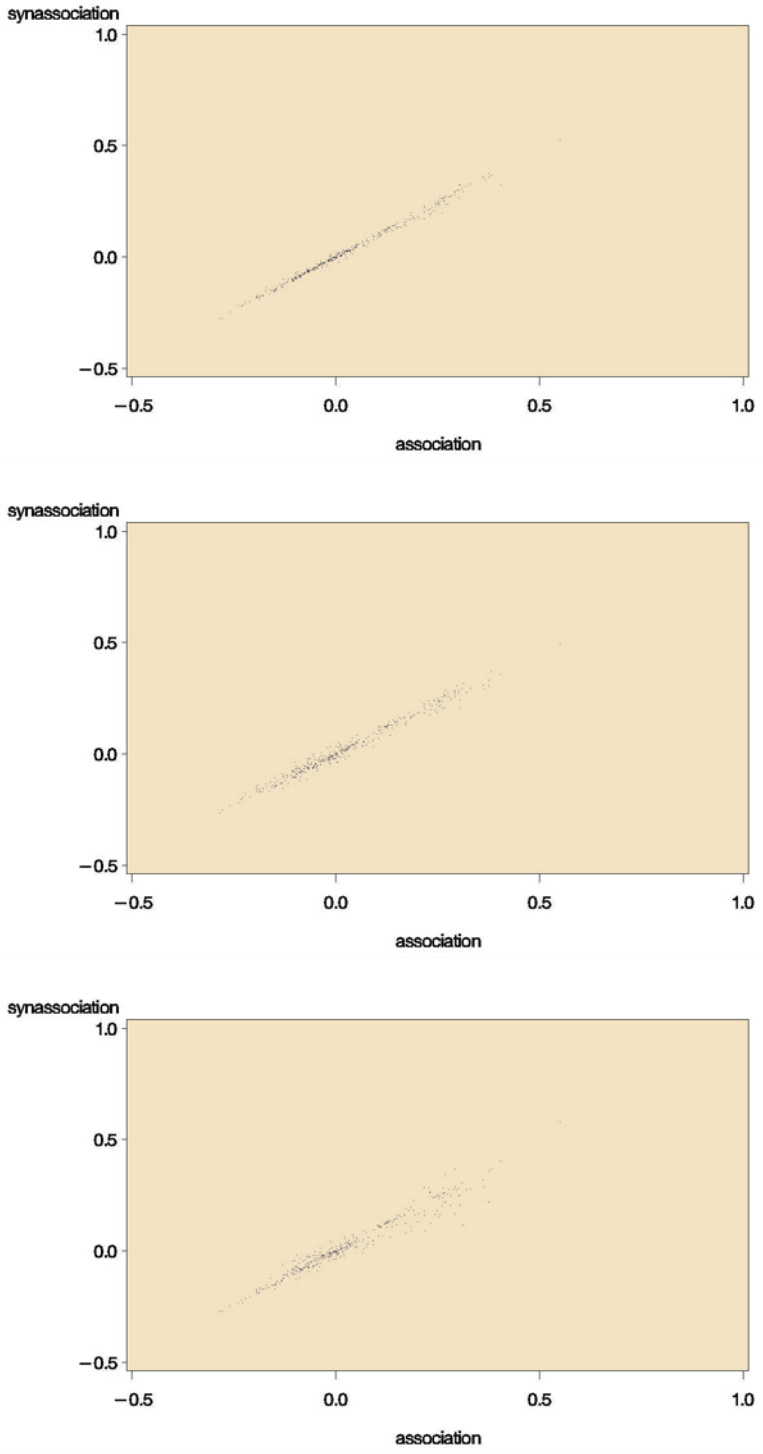


Figure J-4. Plots of ACS and Partial Perturbed Pairwise Correlations for Iowa's PUMAs: Top: Constrained Hotdeck, Middle: Semi-Parametric, Bottom: Parametric

Appendix K

DEVELOPMENT PHASE: TABLES FOR TRAVEL MODEL OUTPUT COMPARISONS

Appendix K presents the results of the comparison tests for the four Development Phase test sites: Atlanta, GA; Madison, WI; St. Louis, MO, and Iowa statewide. As noted earlier, the results for Atlanta are not yet complete and will be the subject of a supplemental report. Table 2-12 in the main body of the report shows which comparison tests were made for which test site and at which level of geography.

Table K-1. Results for Test 1 (Population by Age Category), Atlanta, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	29,000	252%	29,000	249%	335	3%
Median	-800	-12%	-900	-13%	24	<1%

Matrix Size: 112 estimates
ACS Sample Size: 88,722 respondents

Table K-2. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), Atlanta, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	6,000	40%	6,000	40%	1	0%
Median	-1,500	-11.50%	-1,600	-11.62%	0.0	0.00%

Matrix Size: 98 estimates
ACS Sample Size: 70,488 respondents

Table K-3. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), St. Louis, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif)	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	51,000	296%				
Median	12,000	64%	Same as raw ACS	Same as raw ACS	Within +/- 20 for all cells	Within +/- 1% for all cells

Matrix Size: 24 estimates
ACS Sample Size: 45,895 respondents

Table K-4. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), St. Louis, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	11,000	250%				
Median	2,000	70%	Same as raw ACS	Same as raw ACS	Within +/- 100 for all cells with two outliers	Within +/-12% for all cells with two outliers

Matrix Size: 105 estimates
ACS Sample Size:45,895 respondents

Table K-5. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), Madison, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	5,919	15%				
Median	-568	-1%	Same as raw ACS	Same as raw ACS	Zero	Zero

Matrix Size: 4 estimates
ACS Sample Size: 10,799 respondents

Table K-6. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), Madison, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	379	15%	389	34%	<10	1%
Median	-8	0%	-10	-1%	Zero	Zero

Matrix Size: 100 estimates
ACS Sample Size: 10,799 respondents

Table K-7. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), Atlanta County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	3,854	101%	3,810	103%	17	2%
Median	-394.0	-26.32%	-391.1	-26.01%	1.7	0.42%

Matrix Size: 196 estimates
ACS Sample Size: 88,722 respondents

Table K-8. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), Atlanta, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	716	95%	713	96%	6	2%
Median	-126	-33%	-126	-33%	0.6	0%

Matrix Size: 6,084 estimates
ACS Sample Size: 88,722 respondents

Table K-9. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), St. Louis, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	4,000	49%			42	3%
Median	-800	-19%	Same as raw ACS	Same as raw ACS	1	<1%

Matrix Size: 64 estimates
ACS Sample Size: 54,759 respondents

Table K-10. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), St. Louis, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif (000s)	PctDif	AbsDif	PctDif
IQR	660	52%			9	3%
Median	-174	-29%	Same as raw ACS	Same as raw ACS	Zero	Zero

Matrix Size: 1,225 estimates
ACS Sample Size: 54,759 respondents

Table K-11. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), Madison, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif (000s)	PctDif	AbsDif	PctDif
IQR	464	118%	470	118%	4	2%
Median	-55	-23%	-57	-22%	0	0

Matrix Size: 125 estimates
ACS Sample Size: 14,645 respondents

Table K-12. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), Iowa Statewide, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	1,366	48%			395	112%
Median	-504	-30%	Same as raw ACS	Same as raw ACS	-11	-0.1%

Matrix Size: 324 estimates
ACS Sample Size: 139,088 respondents

Table K-13. Results for Test 4(Estimated Average Home-to-Work Travel Time), St. Louis, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
Auto*	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	16 minutes	31%	16 minutes	33%	2 minutes	5%
Median	-6 minutes	-16%	-7 minutes	-14%	-1 minute	-2%
Transit	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	26 minutes	28%	26 minutes	28%	3 minutes	10%
Median	-20 minutes	-28%	-23 minutes	-32%	-1 minute	-2%

Matrix Size: 64 estimates
ACS Sample Size: 54,759 respondents
*ACS average of drive alone and carpool

Table K-14. Results for Test 4 (Estimated Average Home-to-Work Travel Time), St. Louis, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
Auto*	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	13 minutes	37%	14 minutes	38%	4 minutes	11%
Median	<1 minute	<1%	<1 minute	-1%	<1 minute	-1%
Transit	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	45 minutes	45%	46 minutes	44%	5 minutes	11%
Median	-30 minutes	-40%	-31 minutes	-42%	Zero	Zero

Matrix Size: 1225 estimates
ACS Sample Size: 54,759 respondents
*ACS average of drive alone and carpool

Table K-15. Results for Test 4 (Estimated Average Home-to-Work Travel Time), Madison, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
Auto*	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	9	30%	10	30%	4	17%
Median	-4	-16%	-5	-18%	-1	-3%
Transit	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	25	36%	29	41%	9	27%
Median	-22	-41%	-25	-43%	-2	-10%

Matrix Size: 125 estimates
ACS Sample Size: 14,645 respondents
*ACS average of drive alone and carpool

Appendix L

DEVELOPMENT PHASE: FIGURES FOR TRAVEL MODEL OUTPUT COMPARISONS

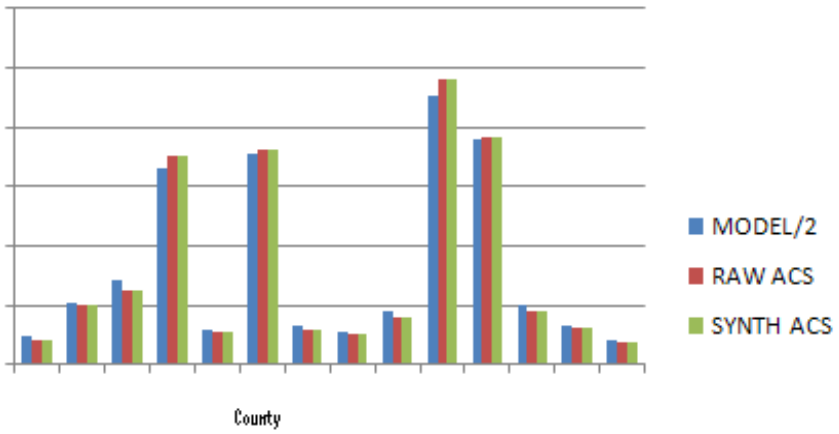


Figure L-1. Test 1: Distribution of Total Workers, Atlanta, County Level

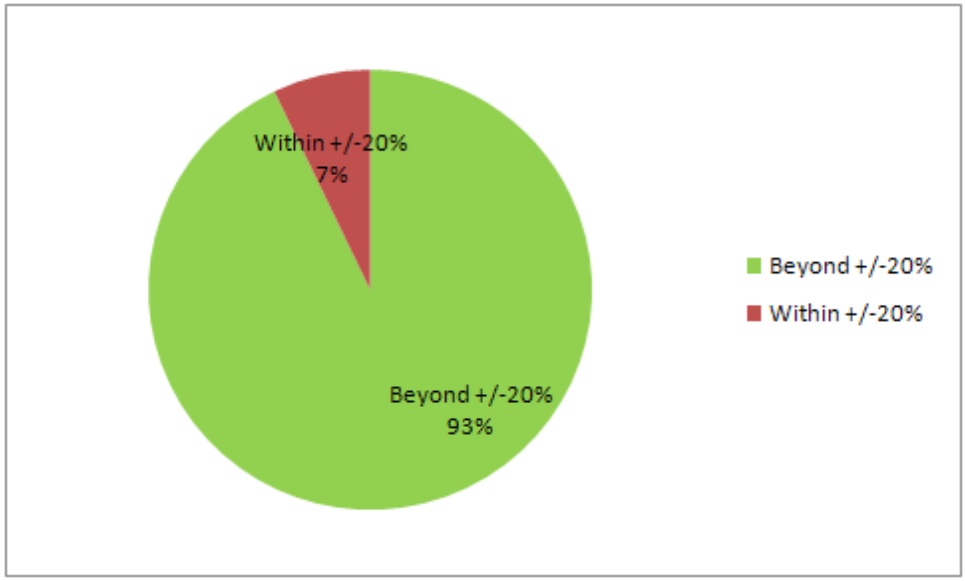


Figure L-2. Test 1 Results within +/- 20%, Atlanta, County Level

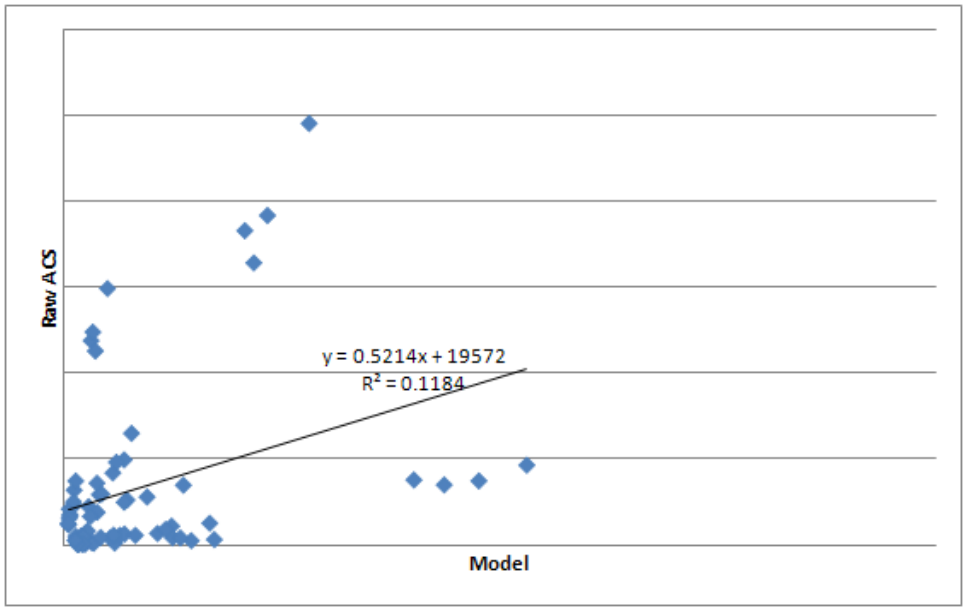


Figure L-3. Test 1: Scatterplot of Model vs. Raw ACS, Atlanta, County Level (all categories)

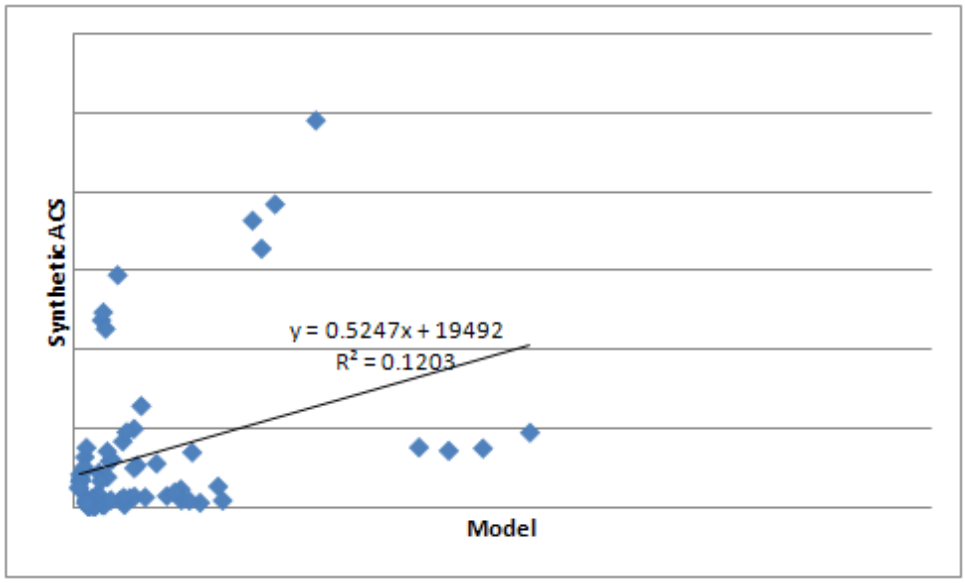


Figure L-4. Test 1: Scatterplot of Model vs. Perturbed ACS, Atlanta, County Level

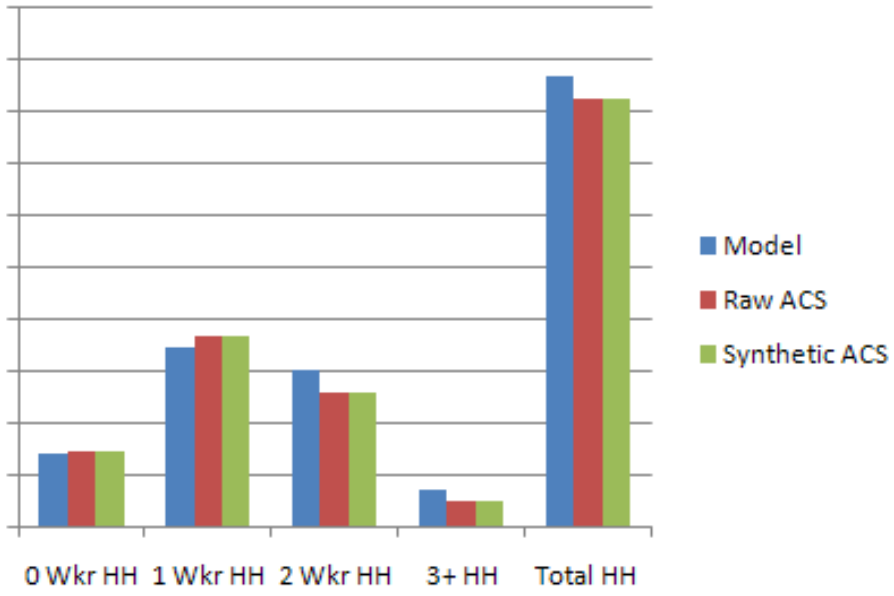


Figure L-5. Test 2: Distribution of Households by Number of Workers, Atlanta

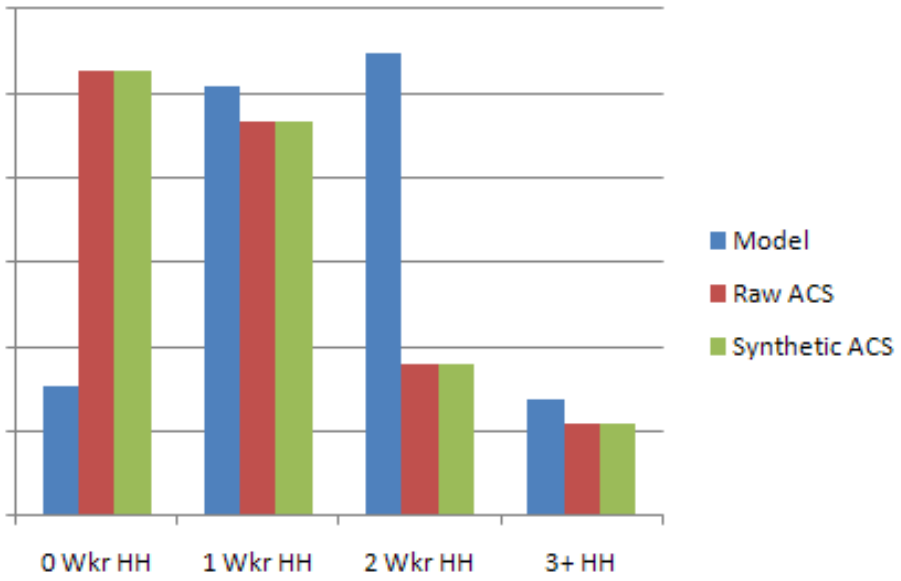


Figure L-6. Test 2: Distribution of Households by Number of Workers, a Select County, Georgia

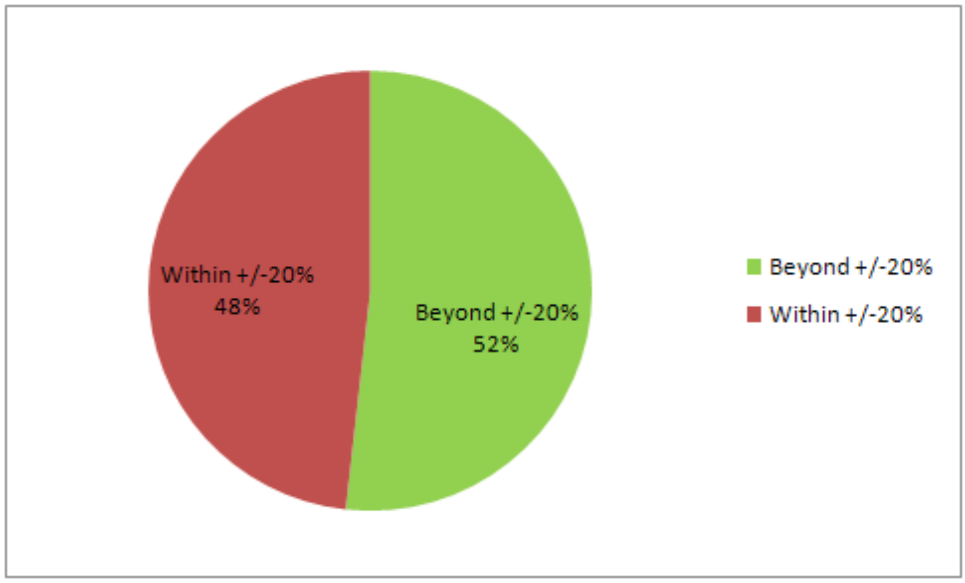


Figure L-7. Test 2: Results within +/- 20%, Atlanta, County Level, Model vs. Raw ACS (all categories, Perturbed results identical)

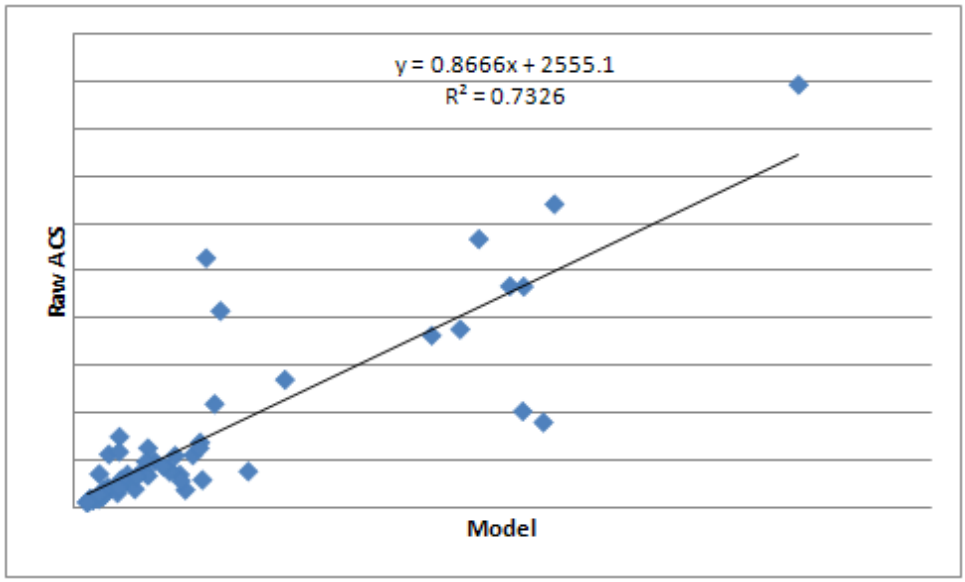


Figure L-8. Test 2: Scatterplot of Model vs Raw ACS, Atlanta, County Level, all categories

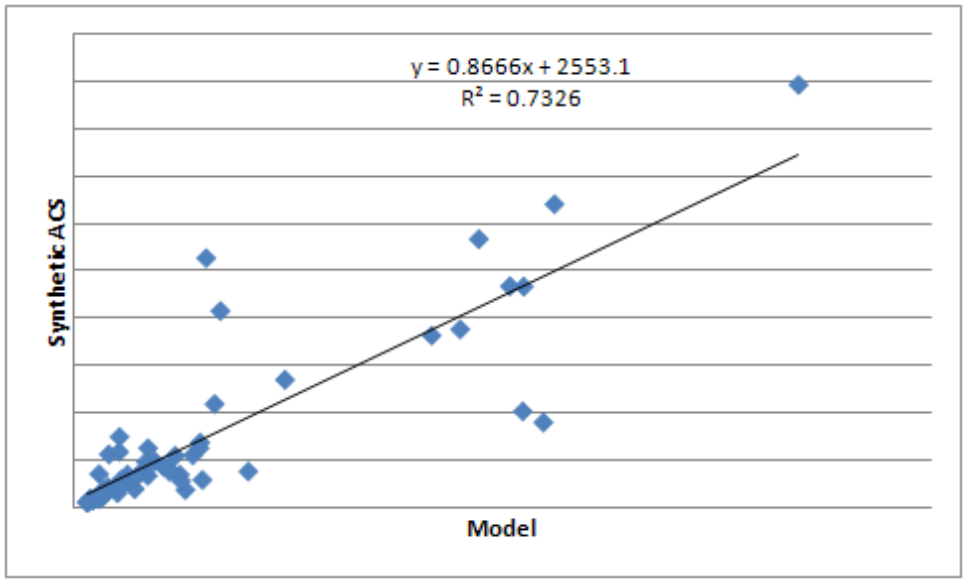


Figure L-9. Test 2: Scatterplot of Model vs. Perturbed ACS, Atlanta, All Categories

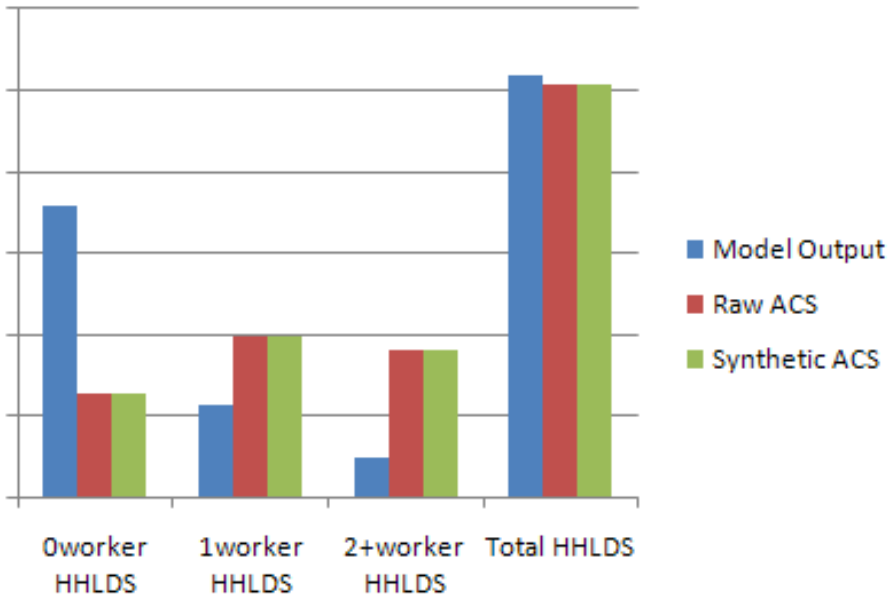


Figure L-10. Test 2: Distribution of Households by Number of Workers, St. Louis MPO Area

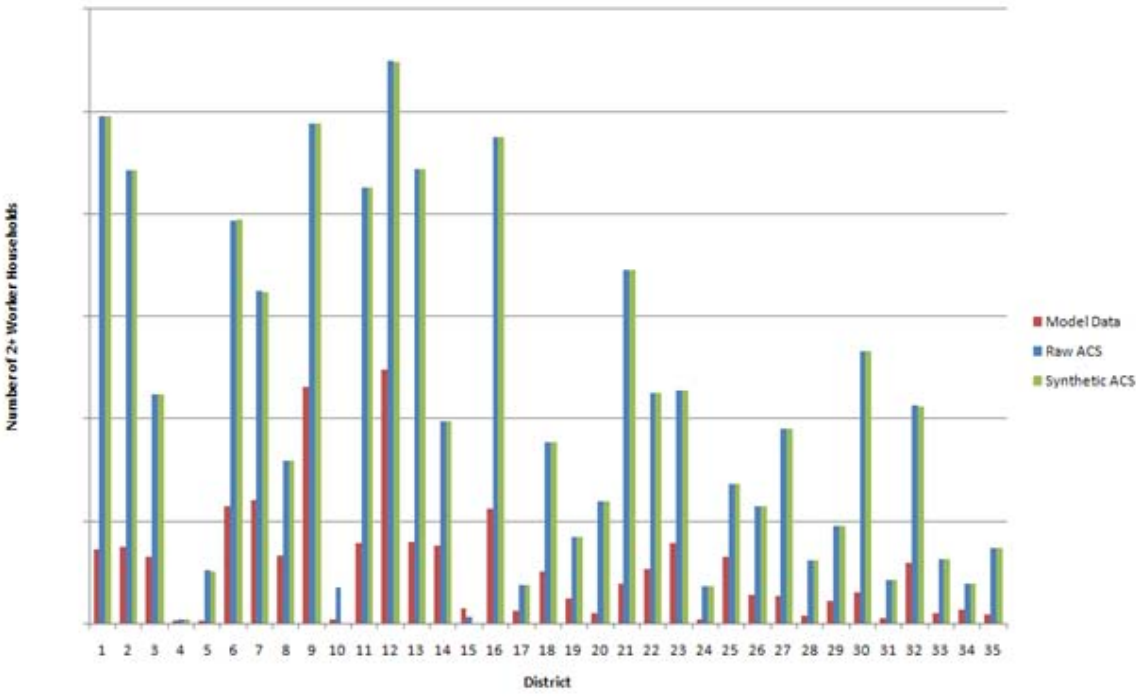


Figure L-11. Test 2: Number of 2+ Worker Households by District, St. Louis MPO Area

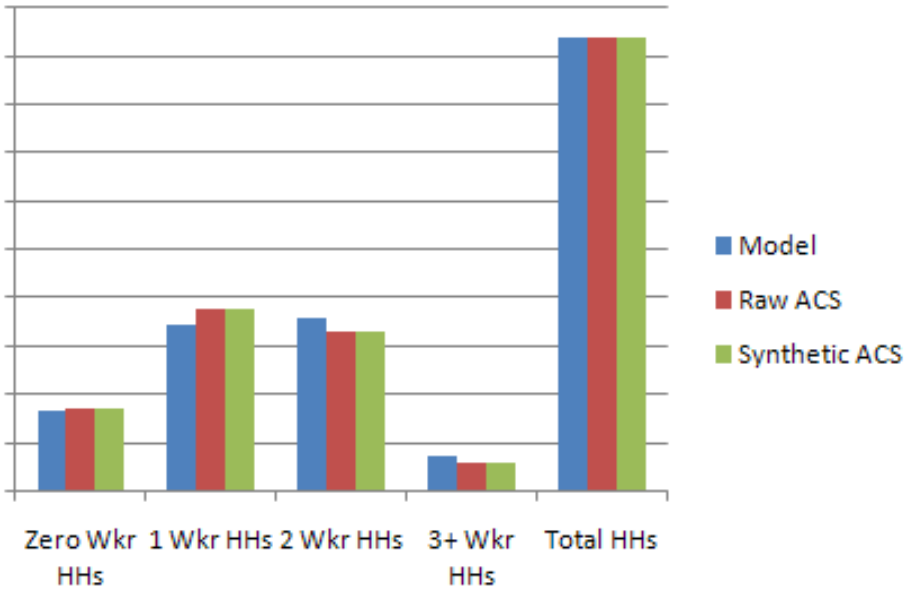


Figure L-12. Test 2: Distribution of Households by Number of Workers, Dane County, Wisconsin

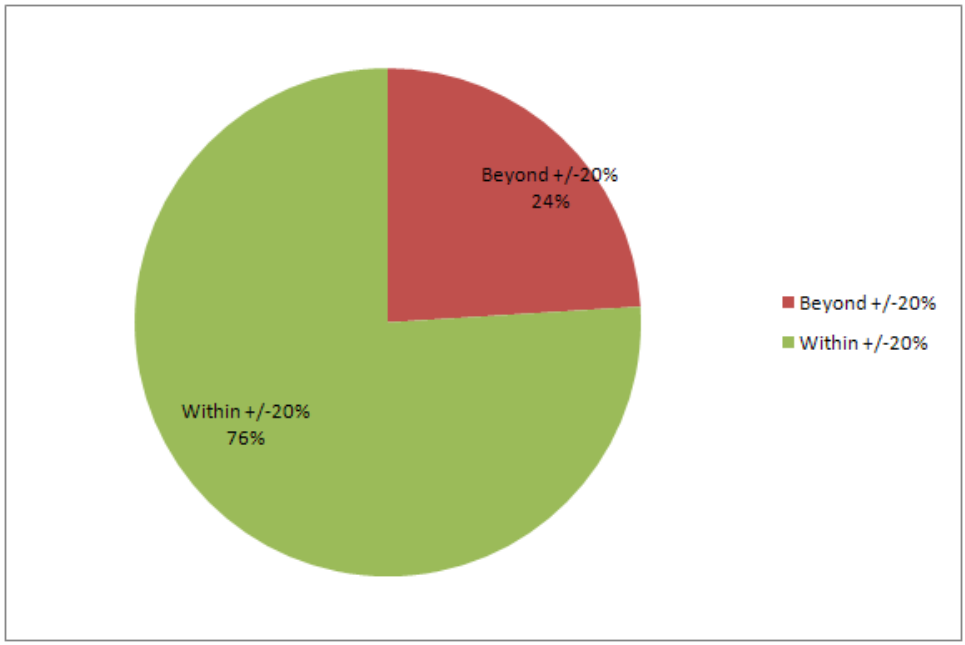


Figure L-13. Test 2: Estimates Within +/-20% -- Raw ACS vs. Model, Madison, District Level

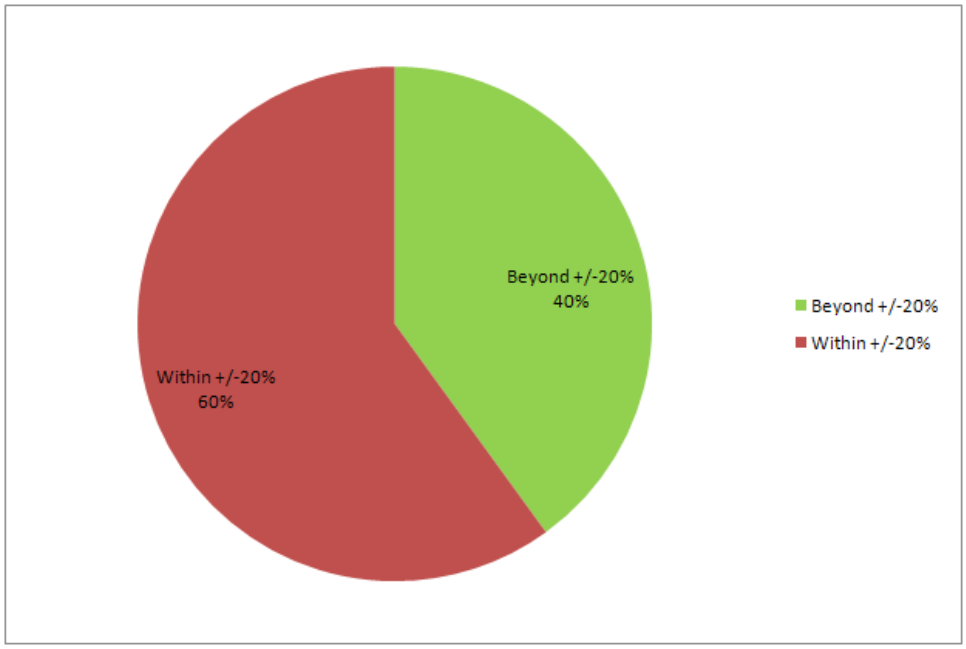


Figure L-14. Test 2: Estimates Within +/-20% -- Perturbed ACS vs. Model, Madison, District Level

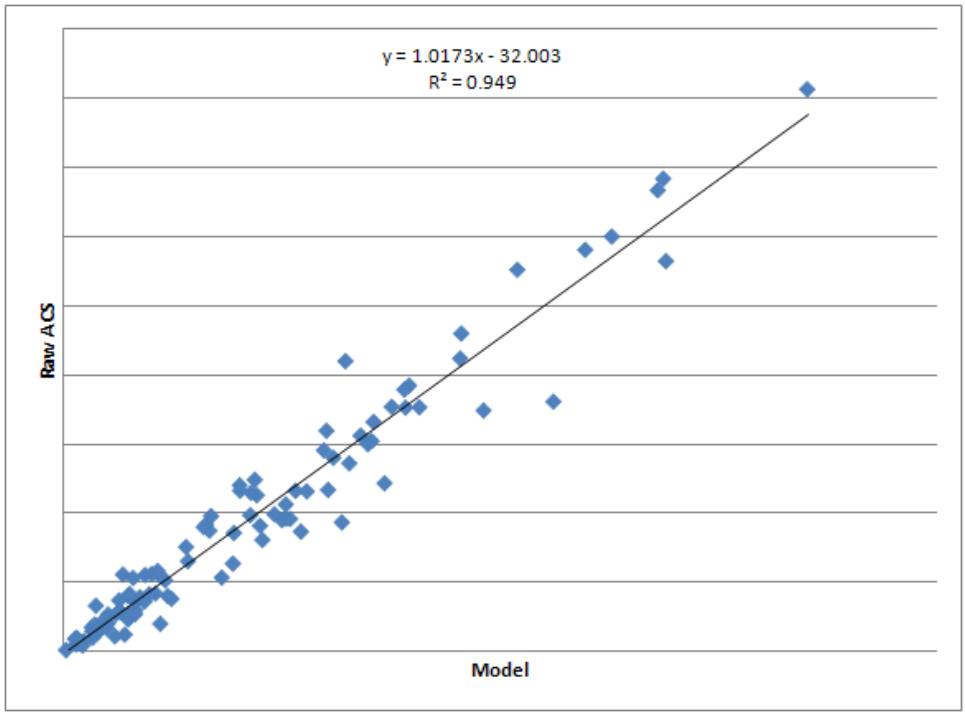


Figure L-15. Test 2: Scatterplot for Madison, Model vs. Raw ACS, District Level (all categories)

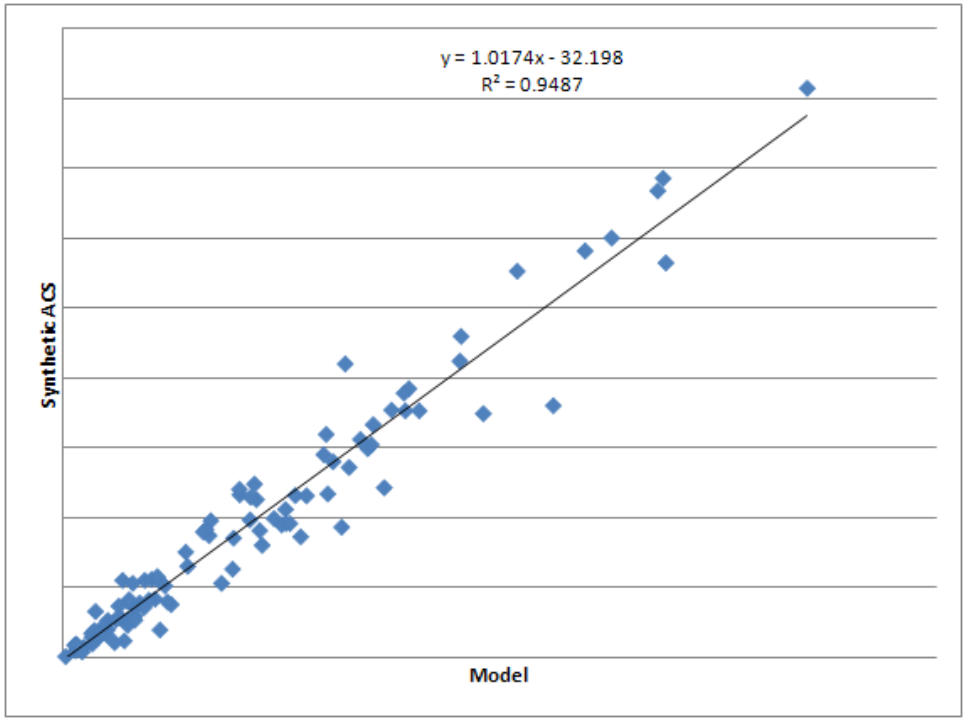


Figure L-16. Test 2: Scatterplot for Madison, Model vs. Perturbed ACS, District Level (all categories)

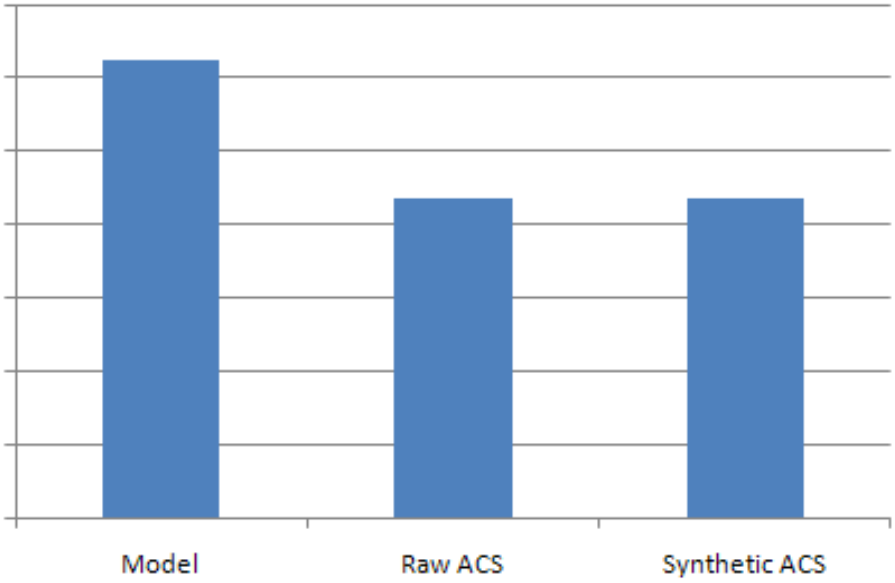


Figure L-17. Test 3: Total Home-Work Flow Estimates, Atlanta 13 County MPO Region

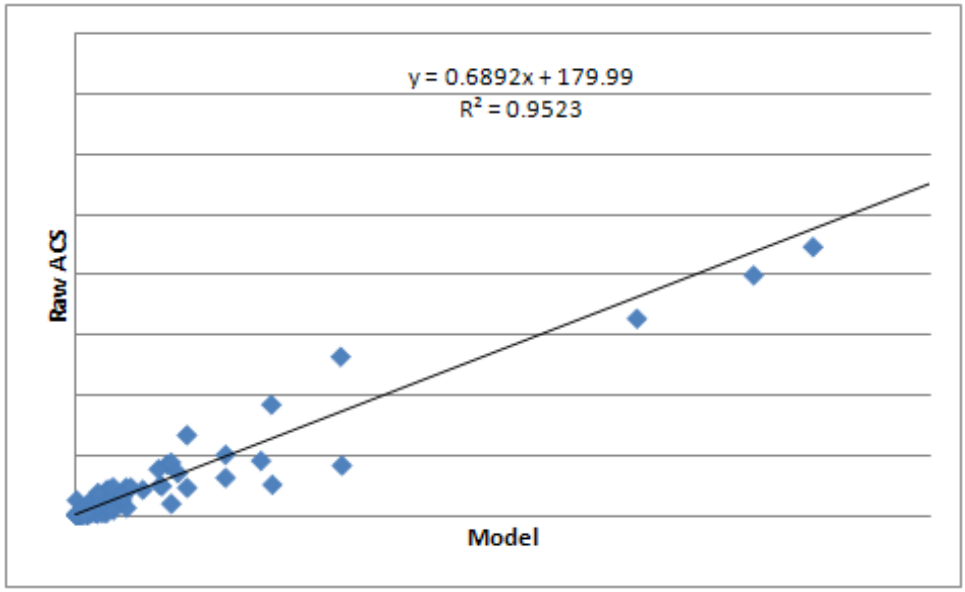


Figure L-18. Test 3: Scatterplot for Atlanta, County Level, Model vs. Raw ACS

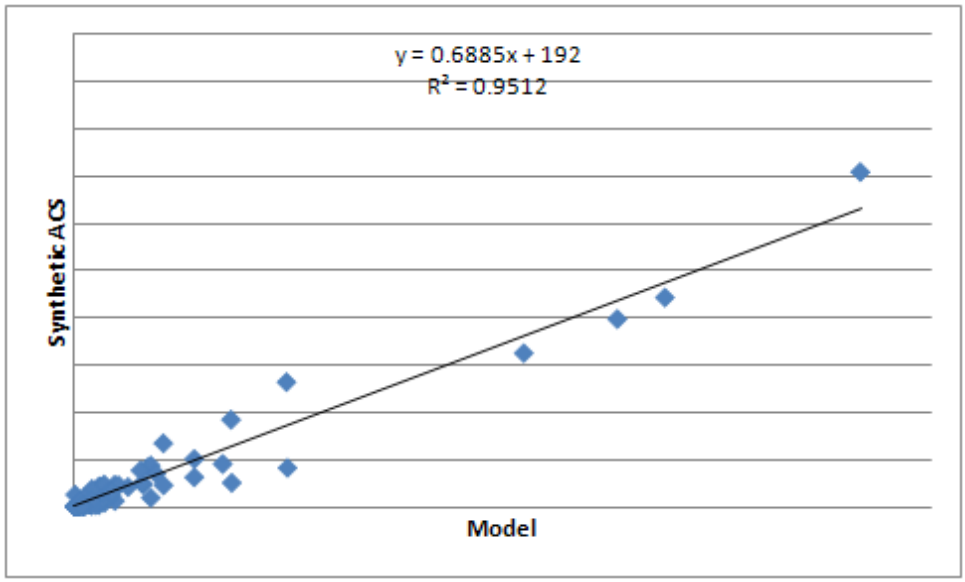


Figure L-19. Test 3: Scatterplot for Atlanta, County Level, Model vs. Perturbed ACS

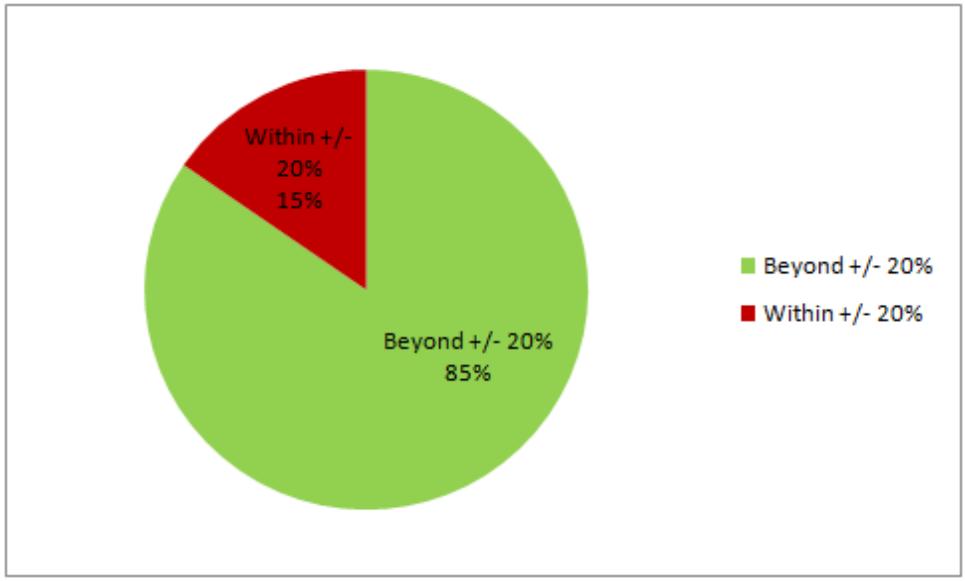


Figure L-20. Test 3: Results within +/- 20% Relative Difference, Atlanta, Model vs. Raw ACS, District Level

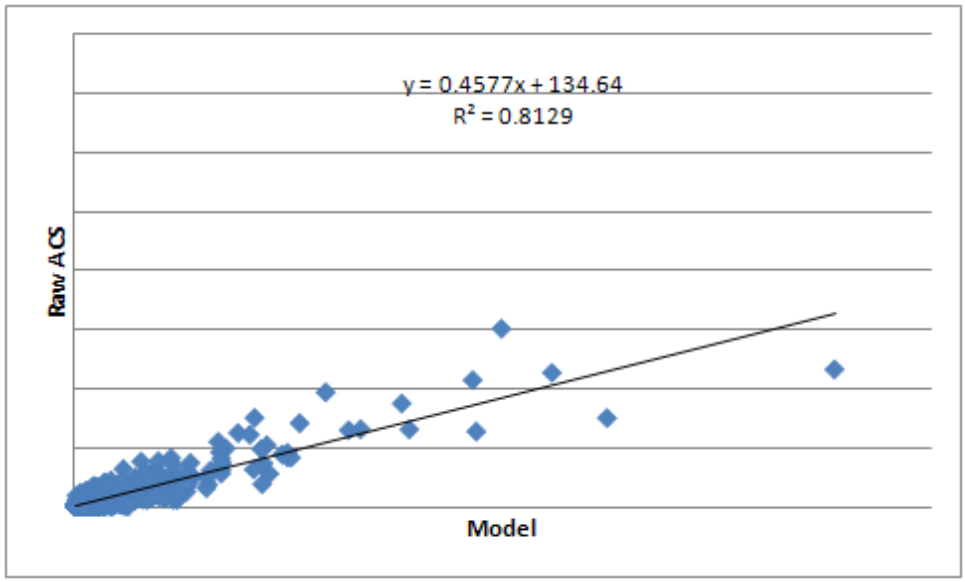


Figure L-21. Test 3: Scatterplot of Model vs. Raw ACS, Atlanta, District Level

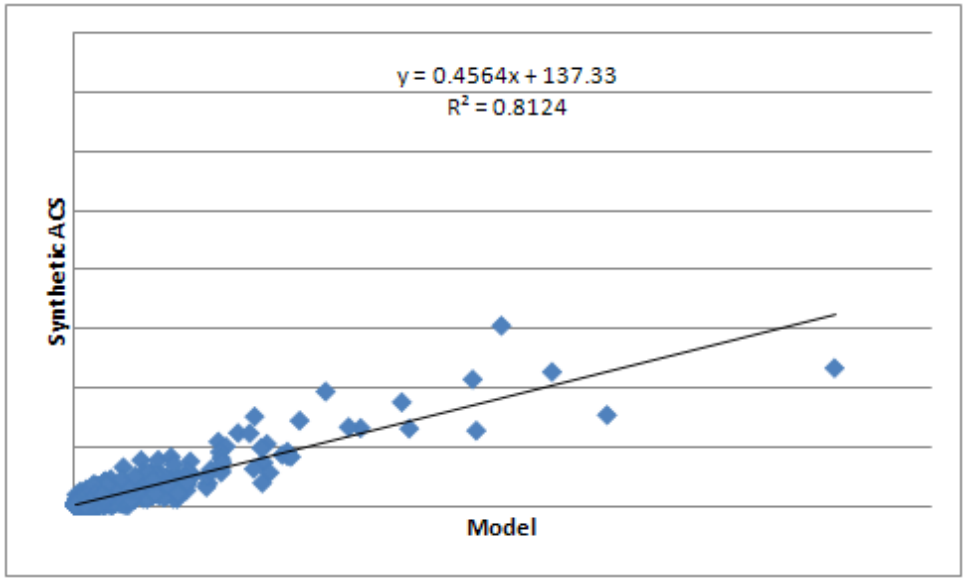


Figure L-22. Test 3: Scatterplot of Model vs. Perturbed ACS, Atlanta, District Level

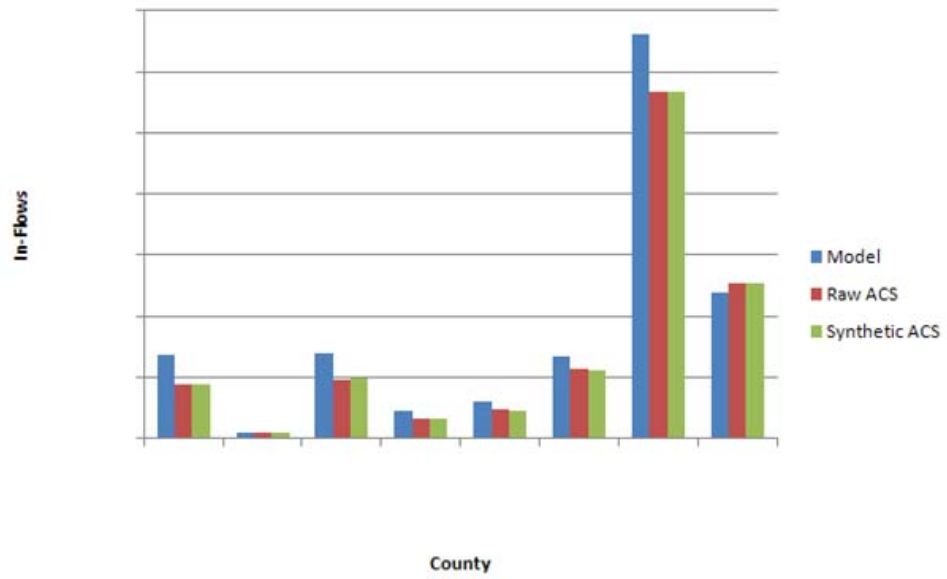


Figure L-23. Test 3: Total In-Flows by County, St. Louis

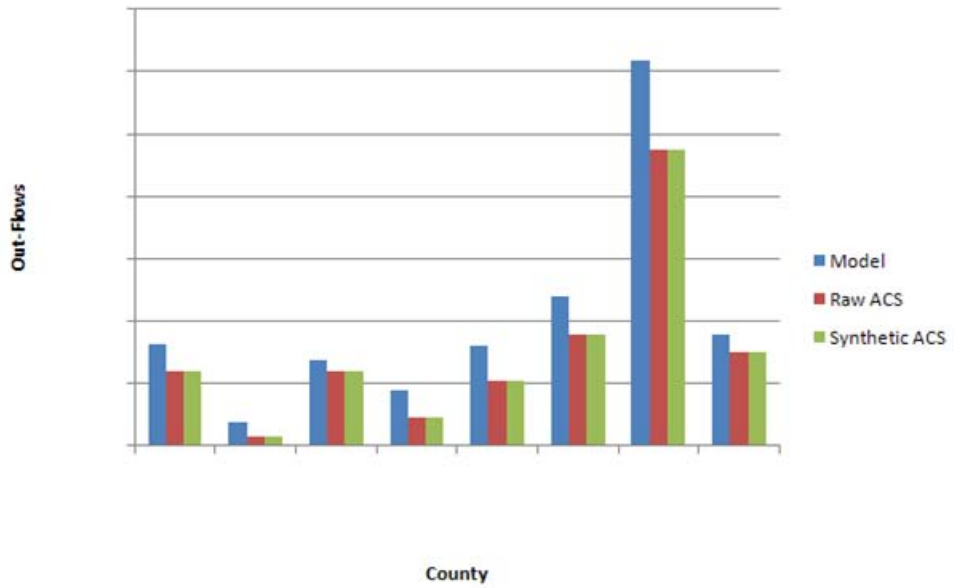


Figure L-24. Test 3: Total Out-Flows by County, St. Louis

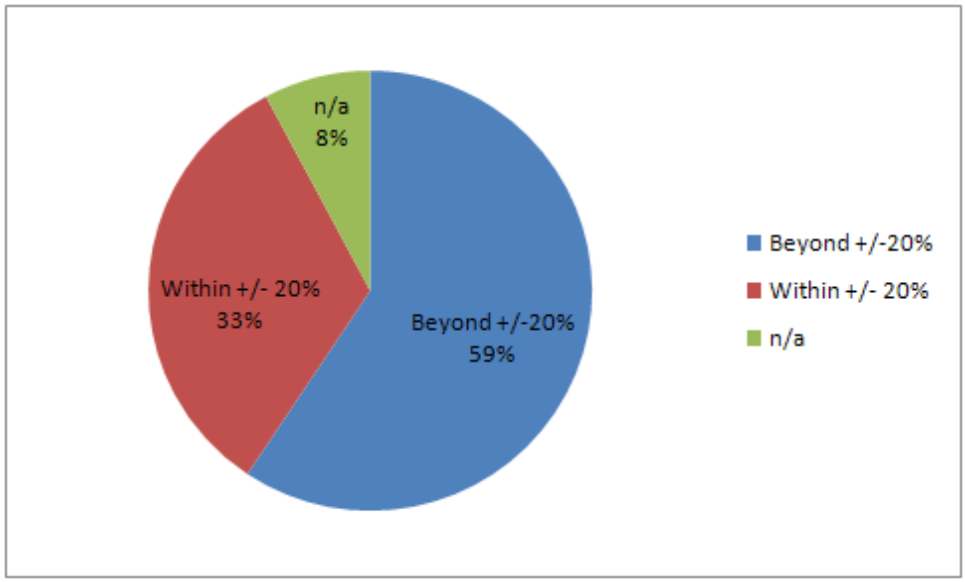


Figure L-25. Test 3: Results within +/- 20% for St.Louis, County-Level, Raw ACS vs. Model

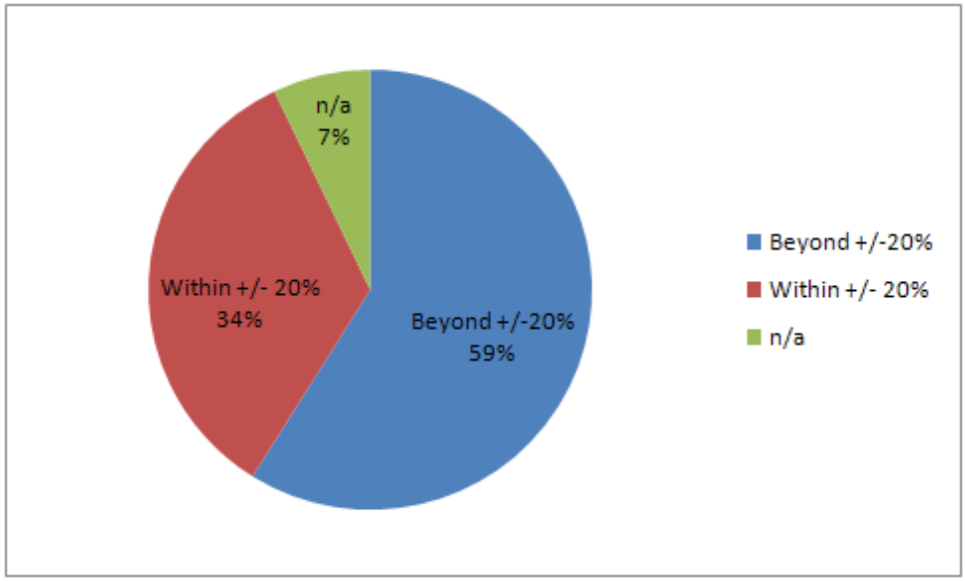


Figure L-26. Test 3: Results within +/- 20% for St.Louis, County-Level, Perturbed ACS vs. Model

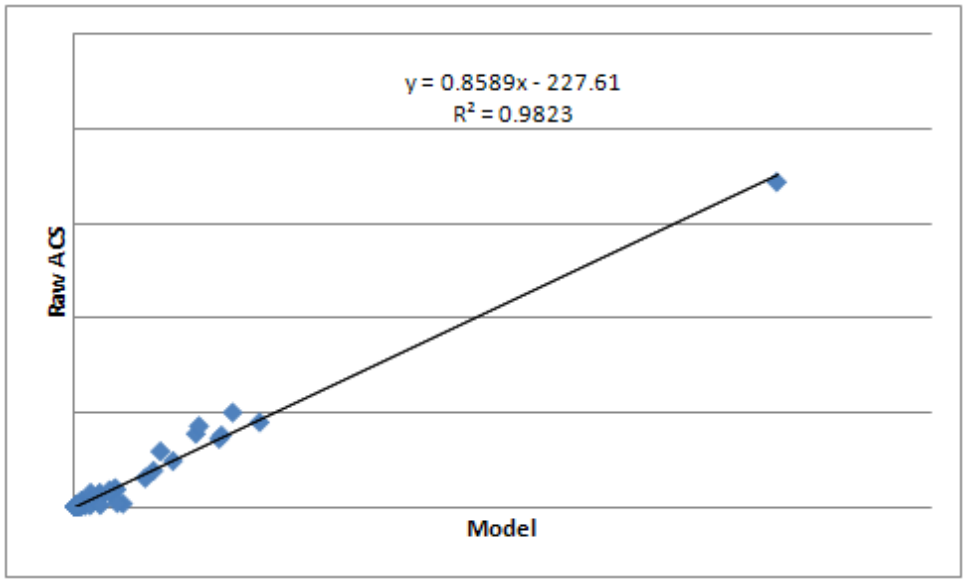


Figure L-27. Test 3: Scatterplot of Raw ACS vs. Model – St. Louis, County Level

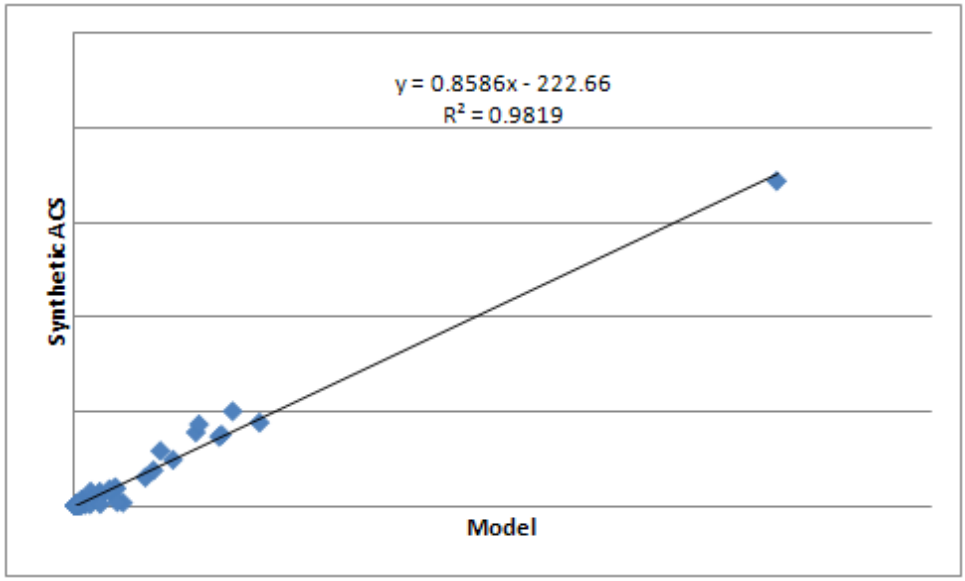


Figure L-28. Test 3: Scatterplot of Perturbed ACS vs. Model – St. Louis, County Level

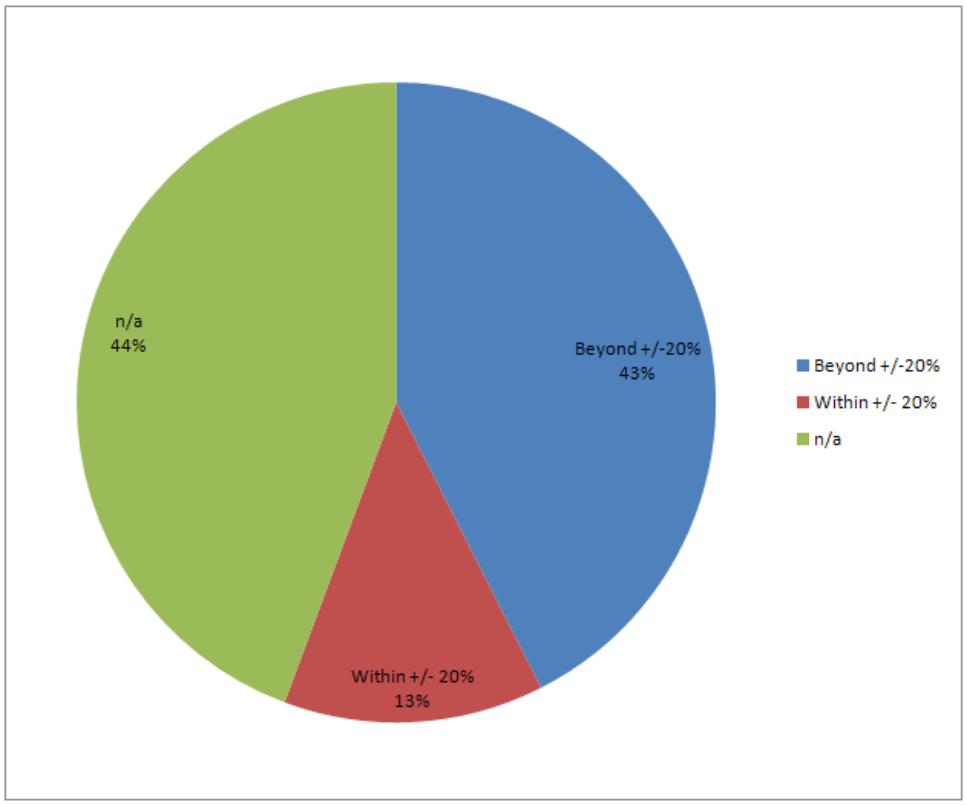


Figure L-29. Test 3: Results Within +/- 20% for St. Louis, District, Raw ACS vs. Model

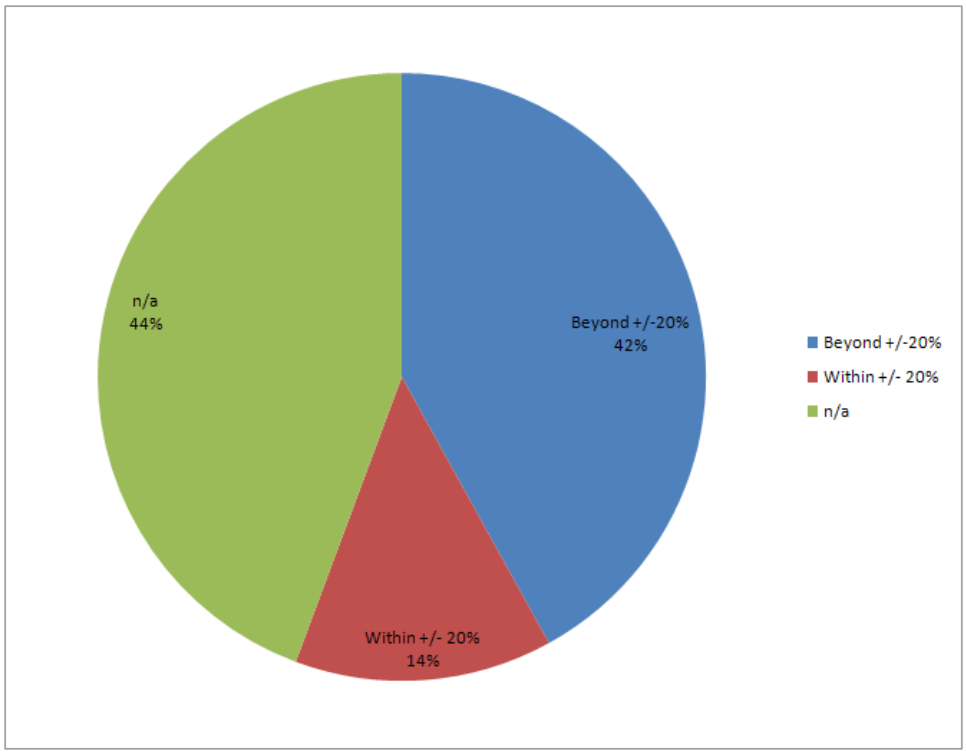


Figure L-30. Test 3: Results Within +/-20% for St. Louis, District, Perturbed ACS vs. Model

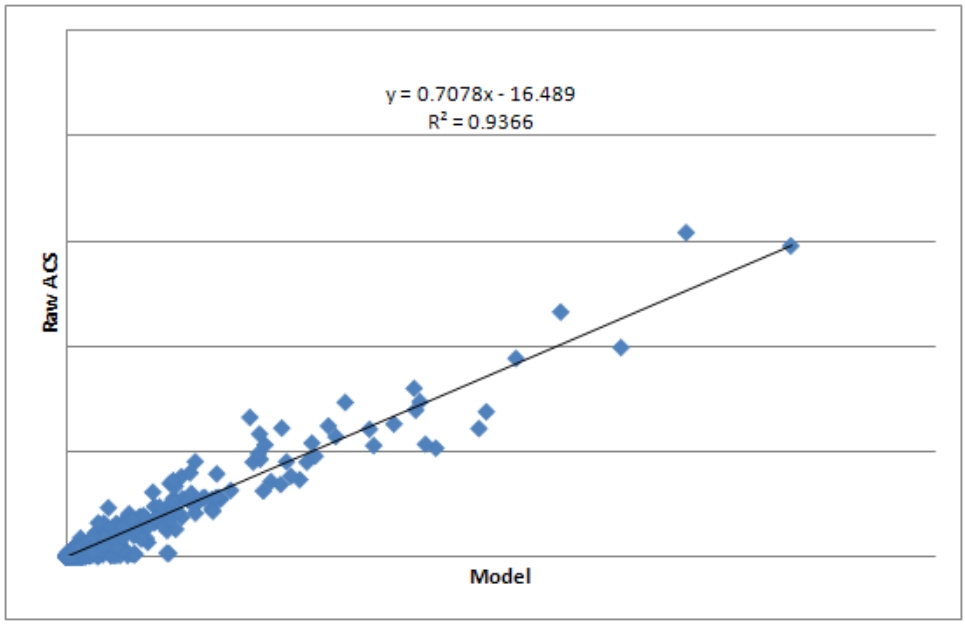


Figure L-31. Test 3: Scatterplot of Model vs. Raw ACS – St. Louis, District Level

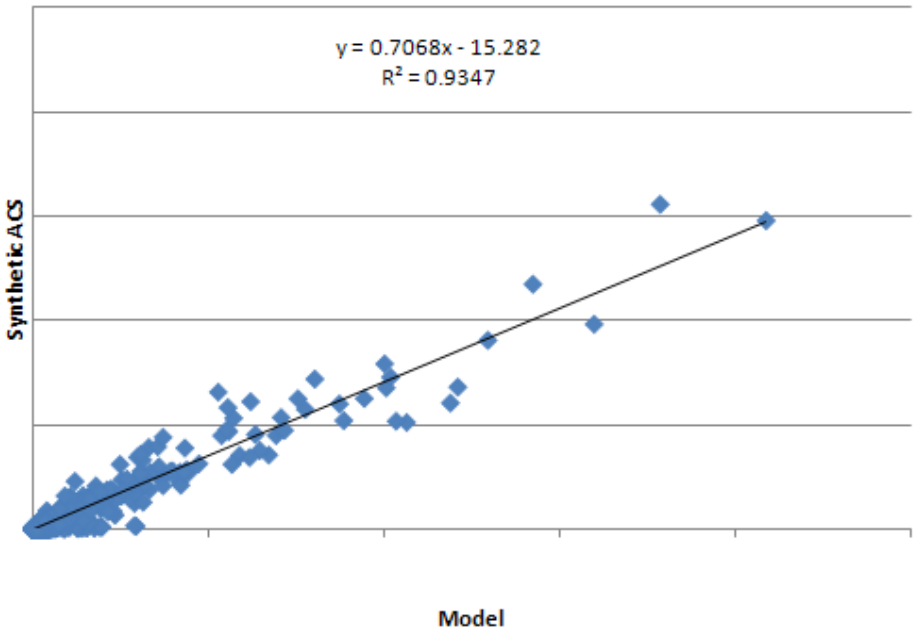


Figure L-32. Test 3: Scatterplot of Model vs. Perturbed ACS, St. Louis, District Level

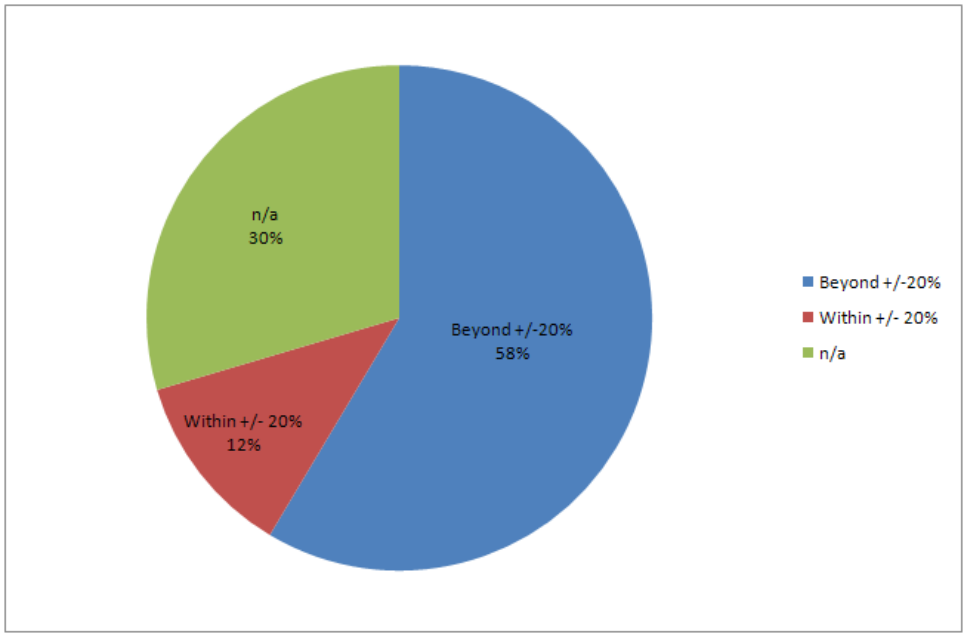


Figure L-33. Test 3: Results within +/-20%, Madison, District Level, Raw ACS vs. Model

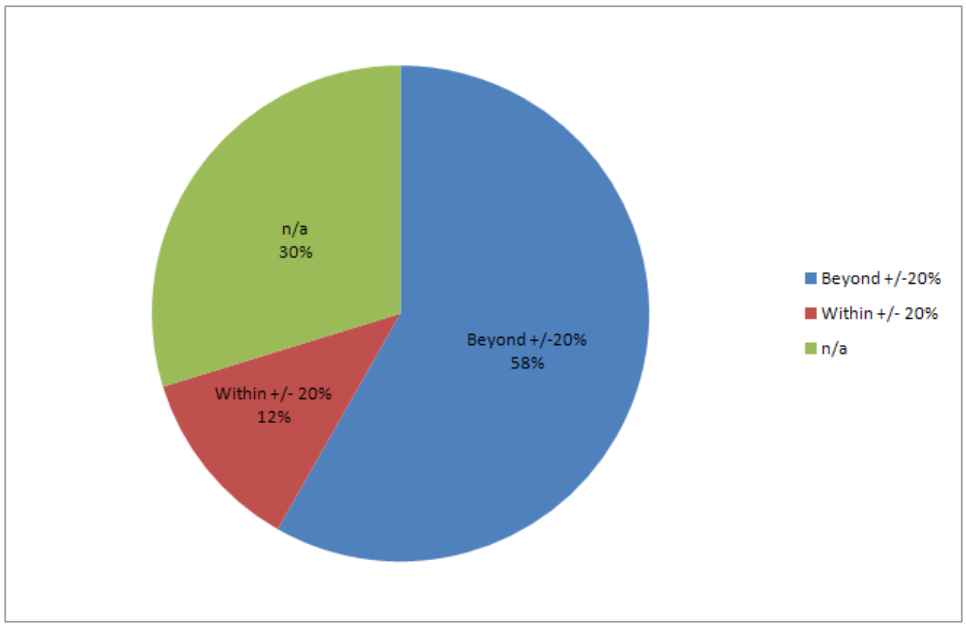


Figure L-34. Test 3: Results within +/-20%, Madison, District Level, Perturbed ACS vs. Model

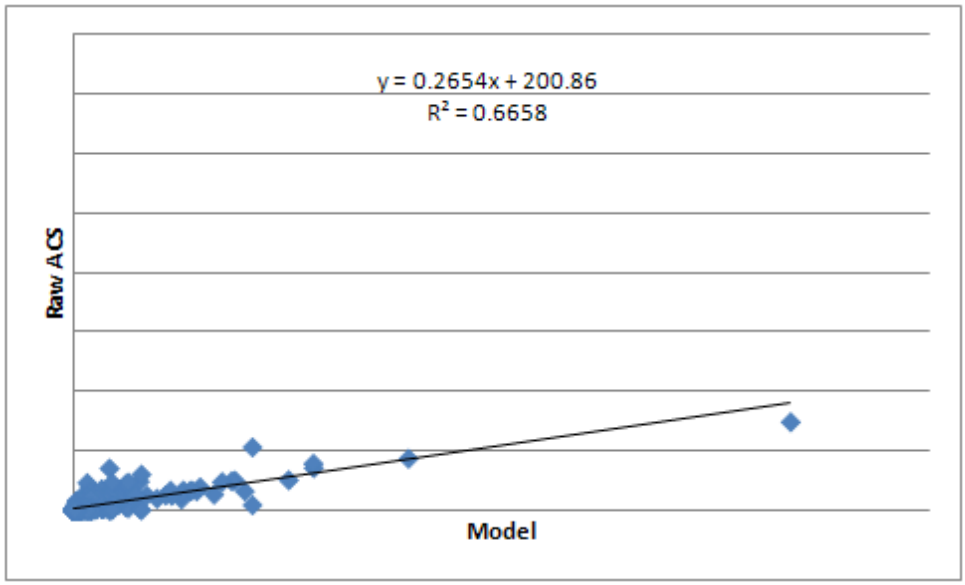


Figure L-35. Test 3: Scatterplot of Model vs. Raw ACS, Madison, District Level (Perturbed results are similar)

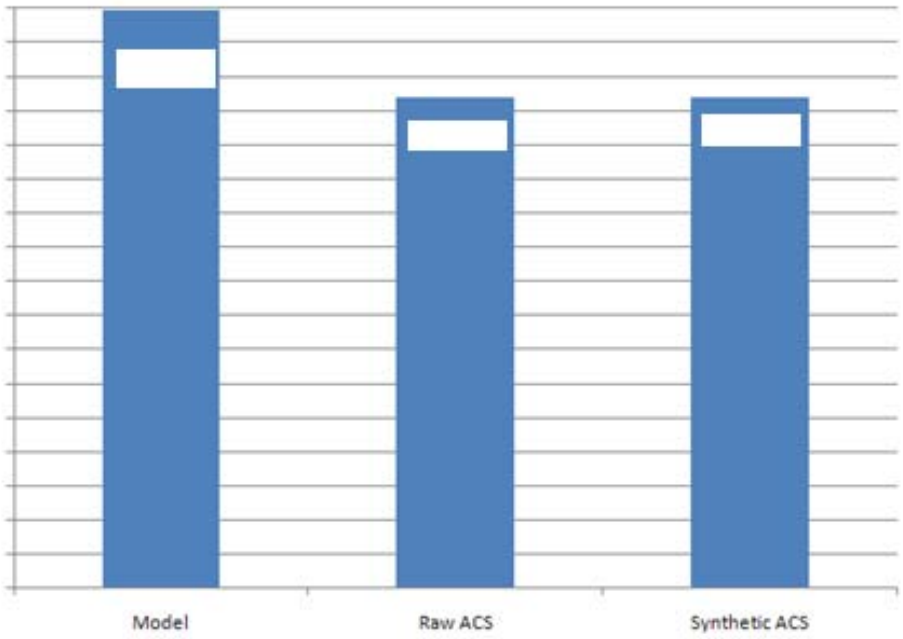


Figure L-36. Test 3: Total Home-Work Flows in Iowa

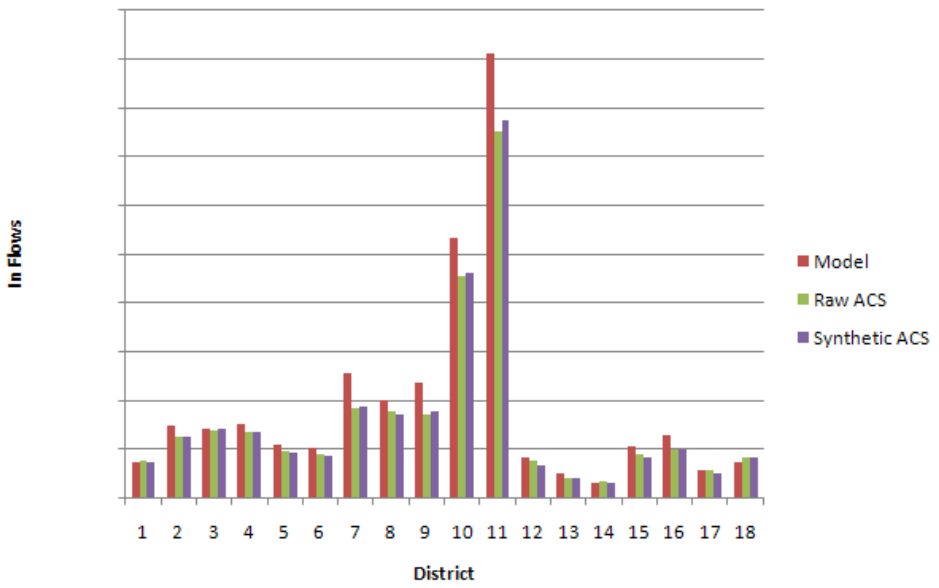


Figure L-37. Test 3: Total Inflows by District, Iowa

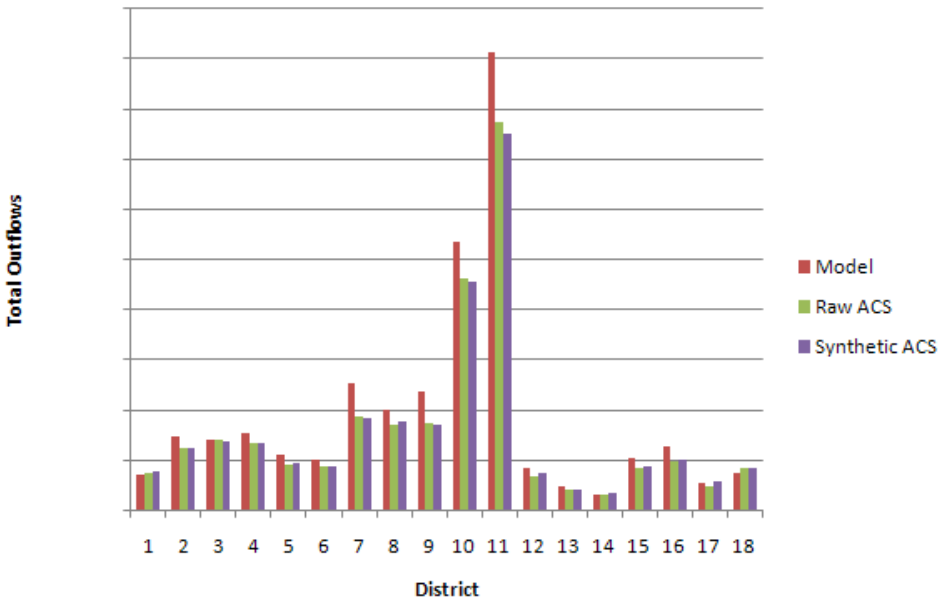


Figure L-38. Test 3: Total Outflows by District, Iowa

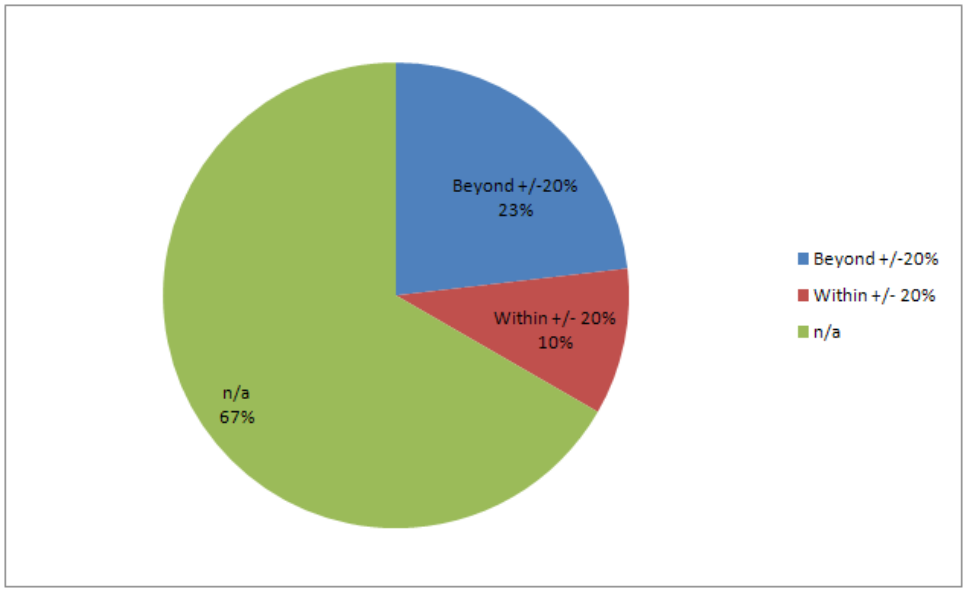


Figure L-39. Test 3: Results within +/- 20% for Iowa, Model vs. Raw ACS, District Level

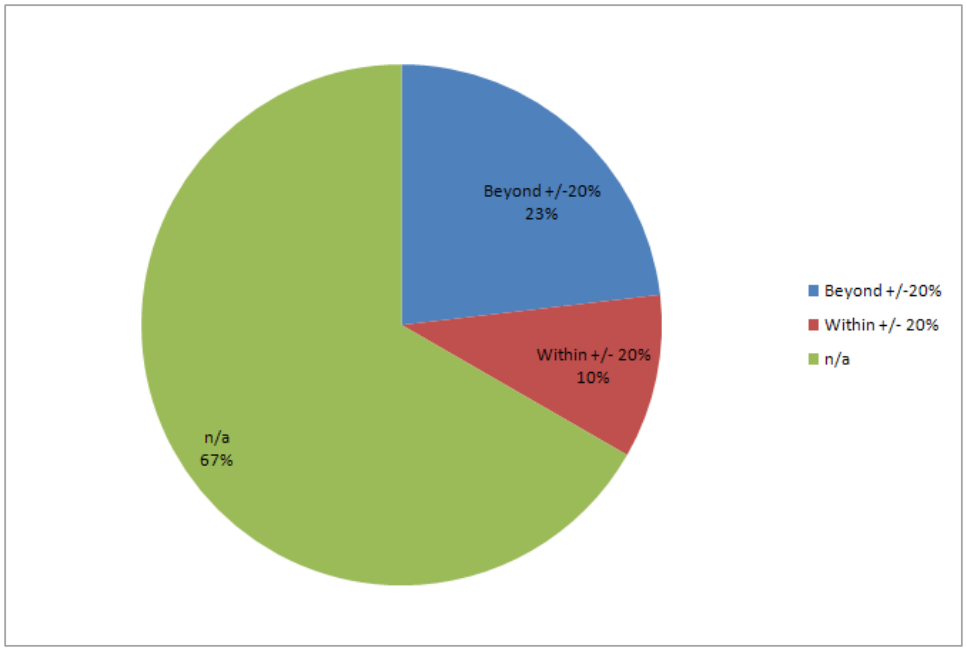


Figure L-40. Test 3: Results within +/- 20% for Iowa, Model vs. Perturbed ACS, District Level

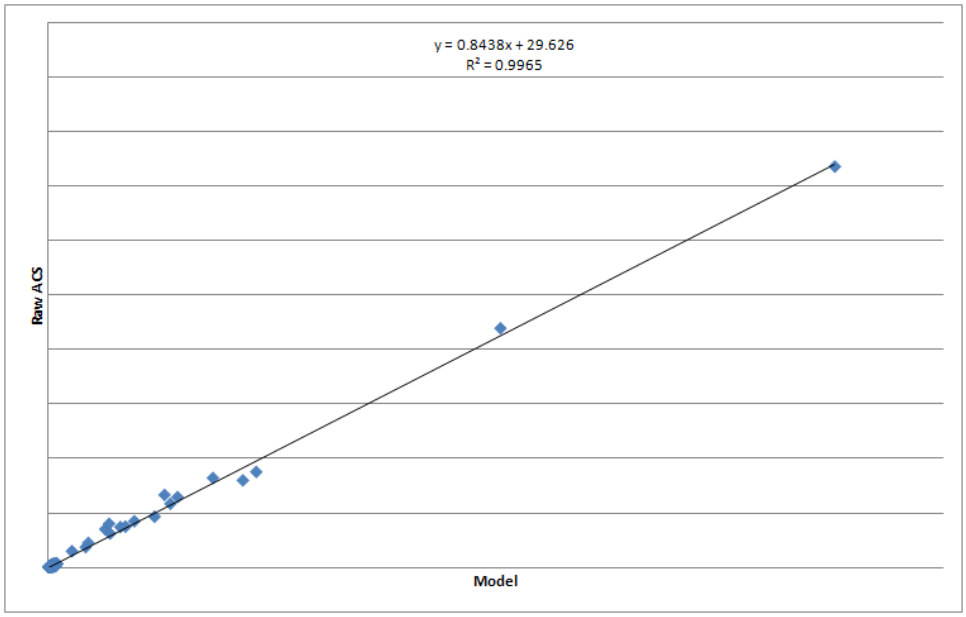


Figure L-41. Test 3: Scatterplot of Model vs. Raw ACS Iowa District

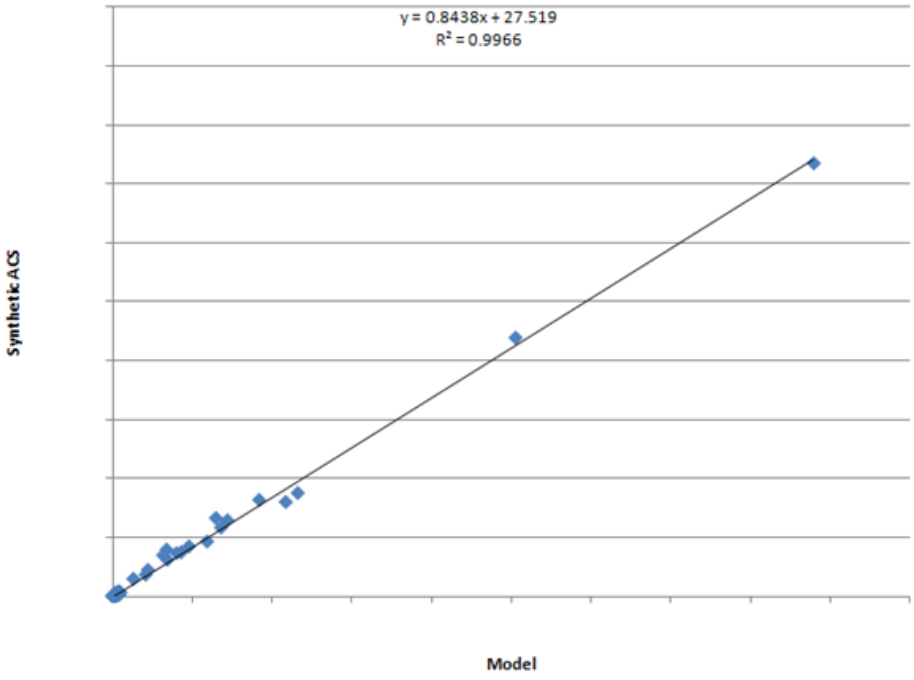


Figure L-42. Test 3: Scatterplot of Model vs Perturbed ACS, Iowa, District Level

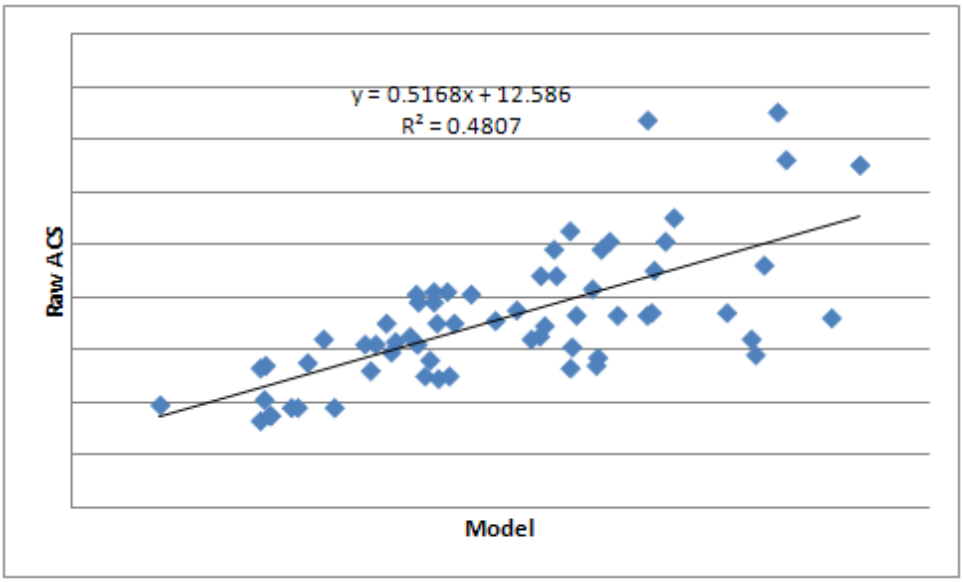


Figure L-43. Test 4: Scatterplot of Model (Auto) vs. Raw ACS, St. Louis, County Level

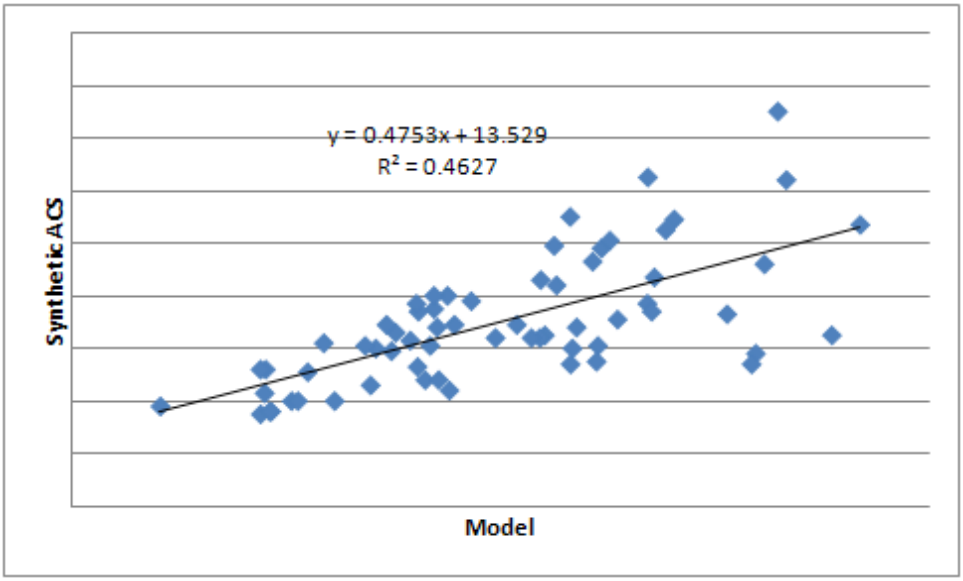


Figure L-44. Test 4: Scatterplot of Model vs. Perturbed ACS, St. Louis, County Level, Auto

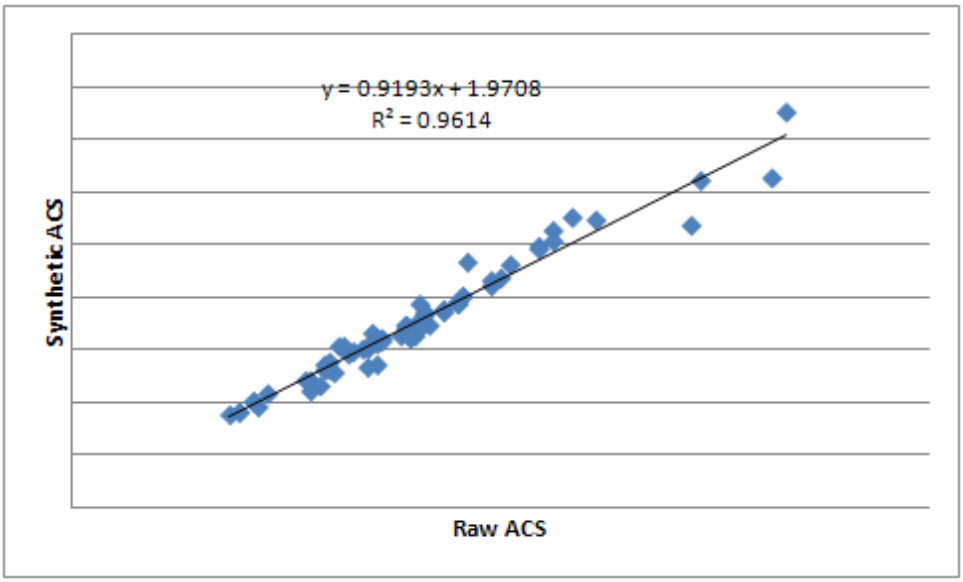


Figure L-45. Test 4: Scatterplot of Raw ACS vs. Perturbed ACS, St. Louis, County Level, Auto

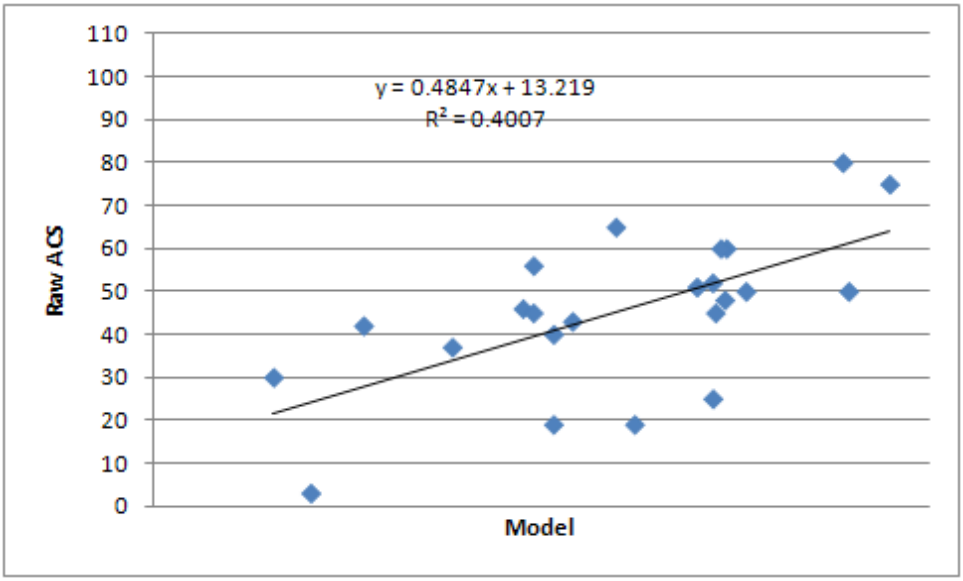


Figure L-46. Test 4: Scatterplot of Model vs. Raw ACS, St. Louis, County Level (Transit)

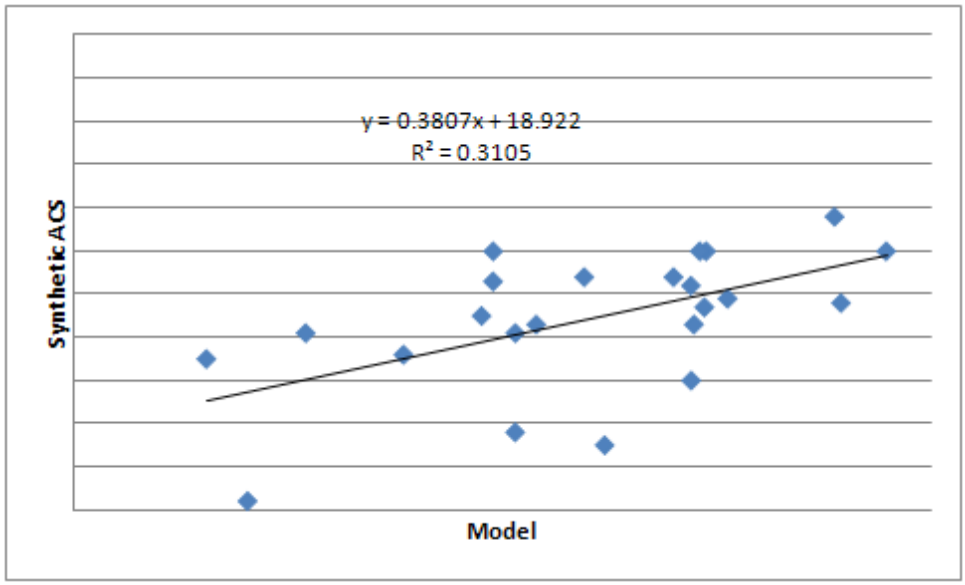


Figure L-47. Test 4: Scatterplot of Model vs. Perturbed ACS, St. Louis, County Level (Transit)

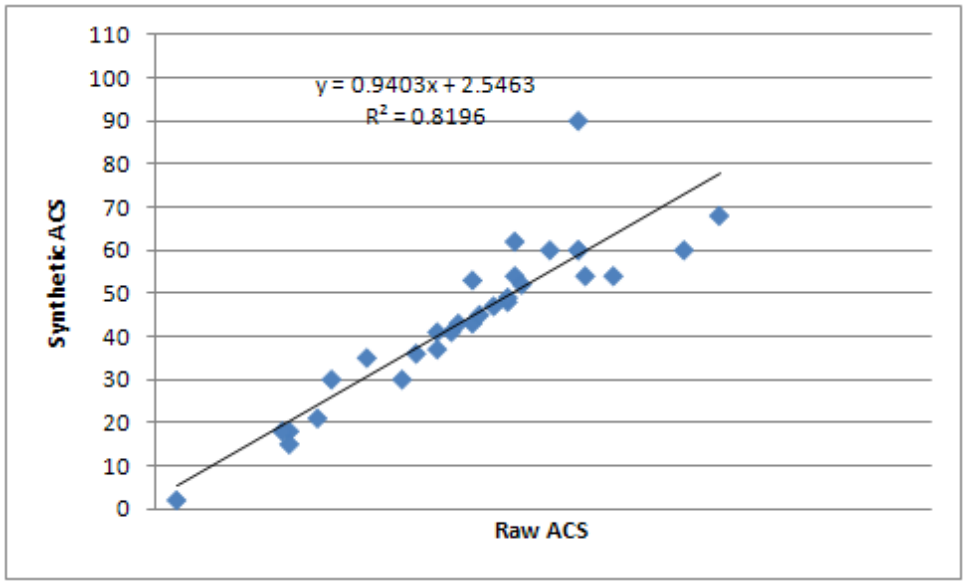


Figure L-48. Test 4: Scatterplot of Raw ACS vs. Perturbed ACS, St. Louis, County Level (Transit)

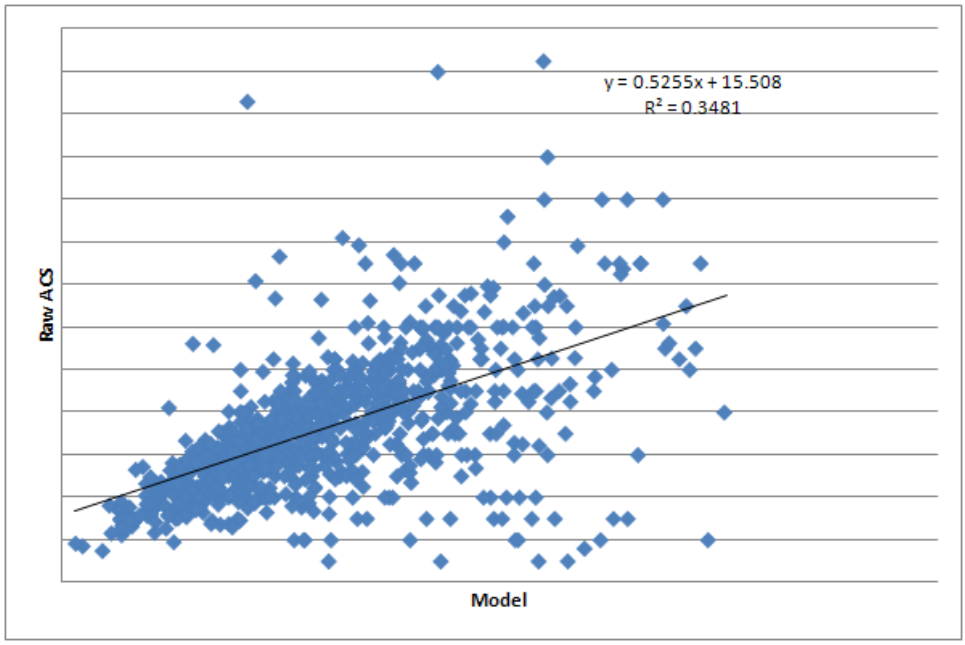


Figure L-49. Test 4: Scatterplot of Model vs Raw ACS, St. Louis, Auto, District Level

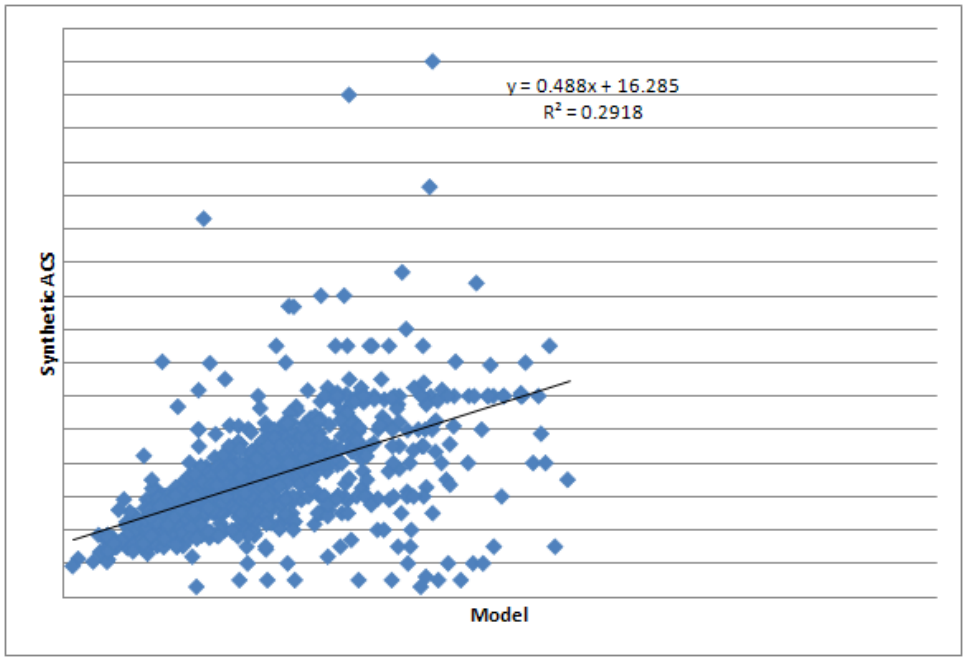


Figure L-50. Test 4: Scatterplot of St. Louis, Model vs Perturbed ACS, District Level (Auto)

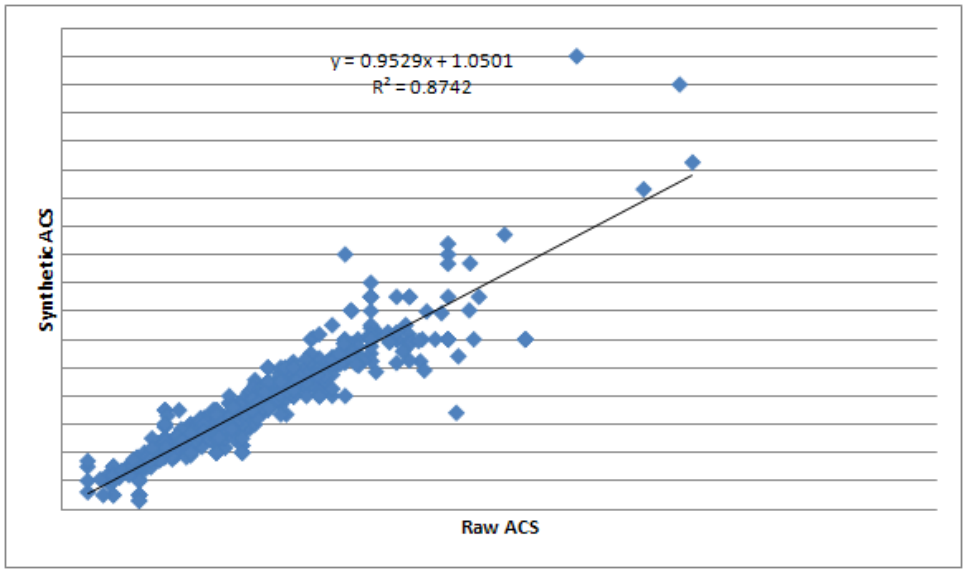


Figure L-51. Test 4: Scatterplot of Raw vs Perturbed ACS, St. Louis, District Level Auto

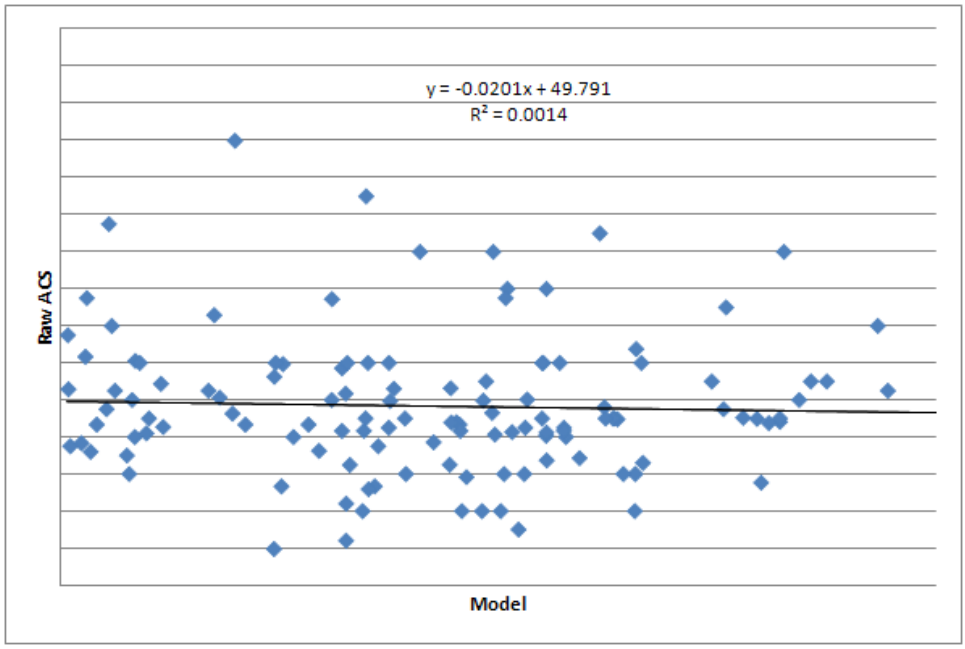


Figure L-52. Test 4: Scatterplot of Model vs Raw ACS, St. Louis, District, Transit

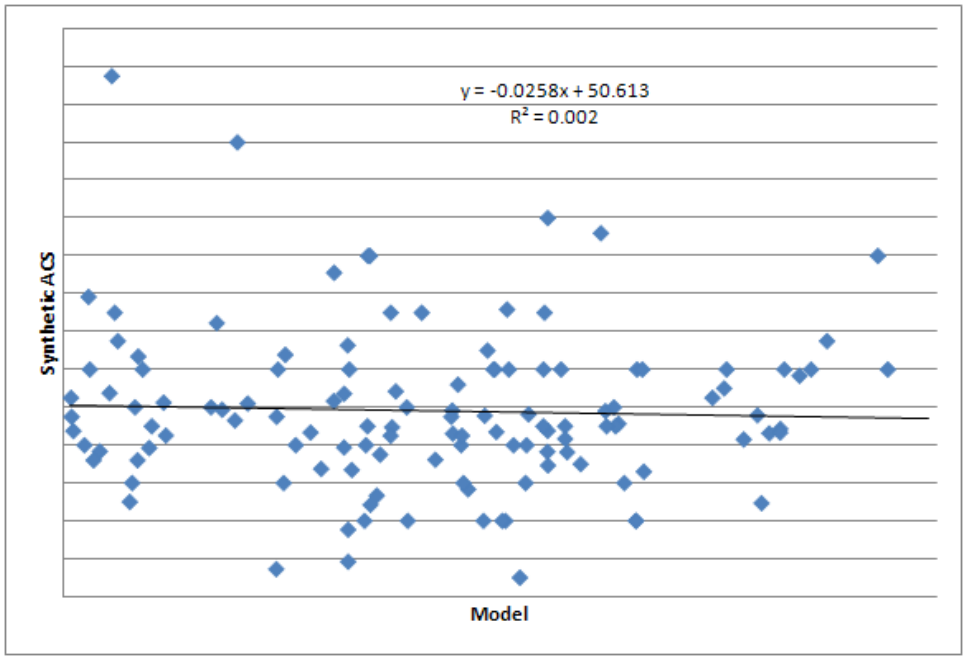


Figure L-53. Test 4: Scatterplot of Model vs. Perturbed ACS, St. Louis, District, Transit

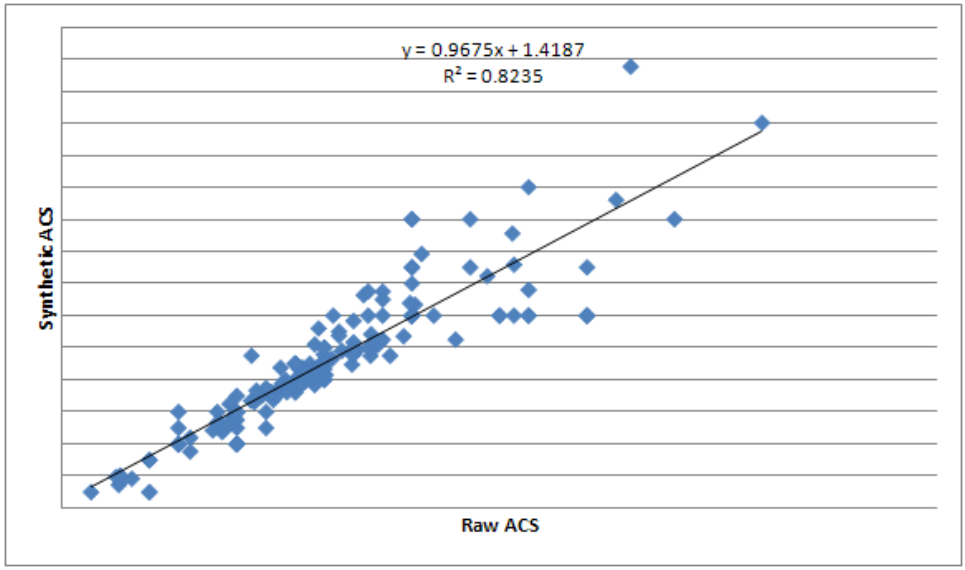


Figure L-54. Test 4: Scatterplot of Raw ACS vs. Perturbed ACS, St. Louis, District, Transit

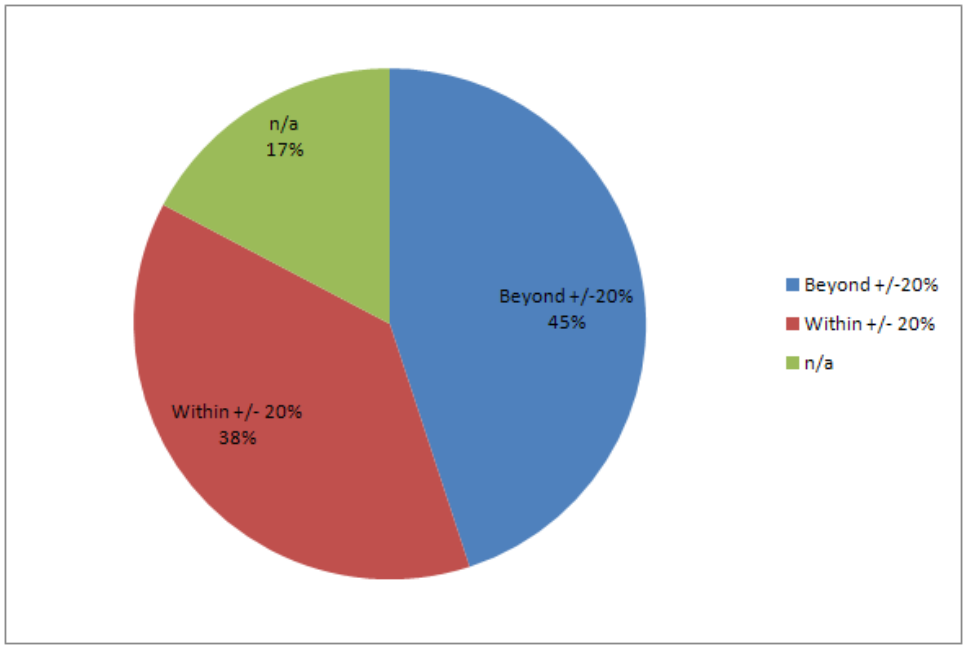


Figure L-55. Test 4: Results within +/-20%, Madison, Raw vs. Model Auto

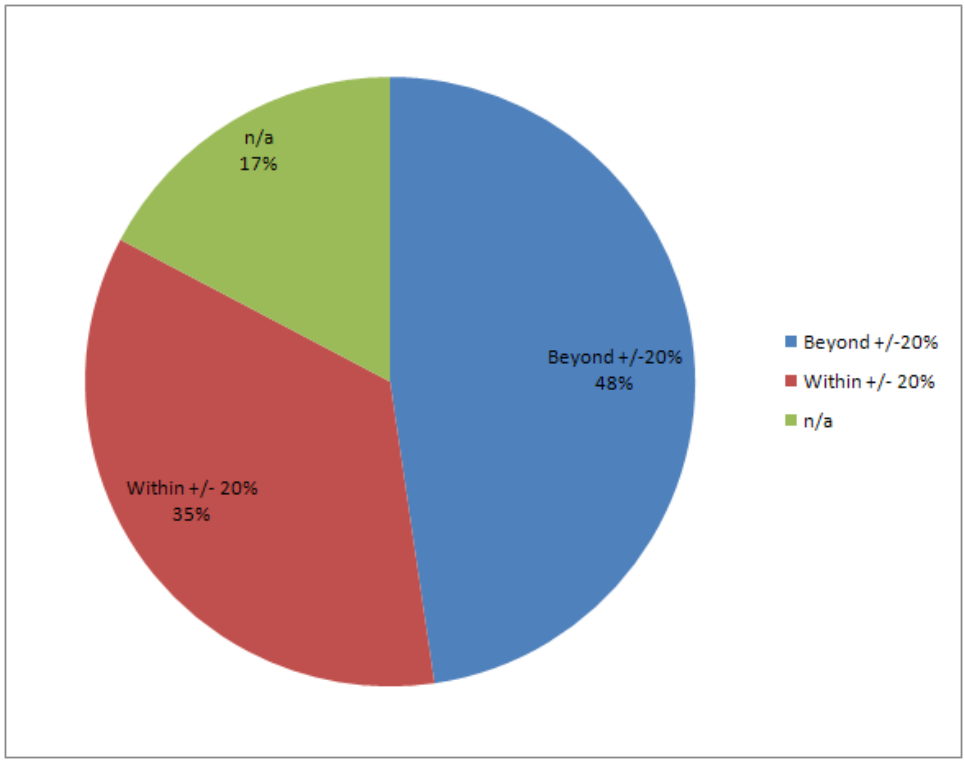


Figure L-56. Test 4: Results within +/- 20% Madison Perturbed vs Model, District, Auto

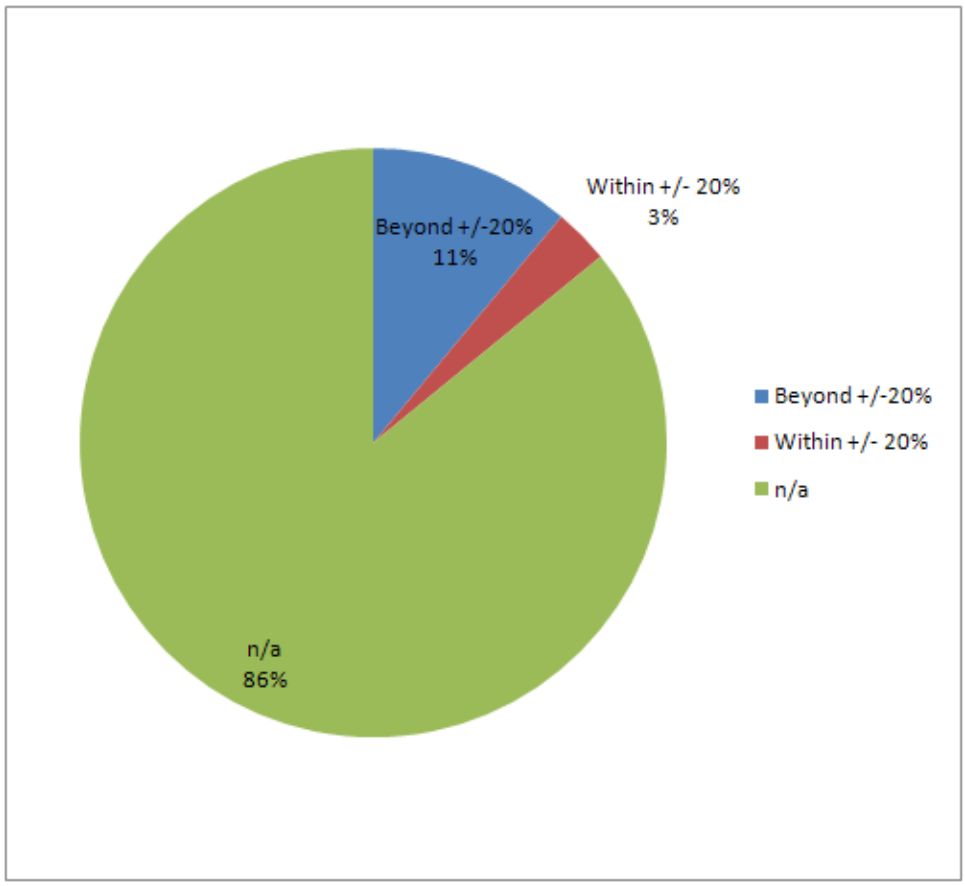


Figure L-57. Test 4: Results within +/- 20% Raw ACS vs Model, Transit, Madison

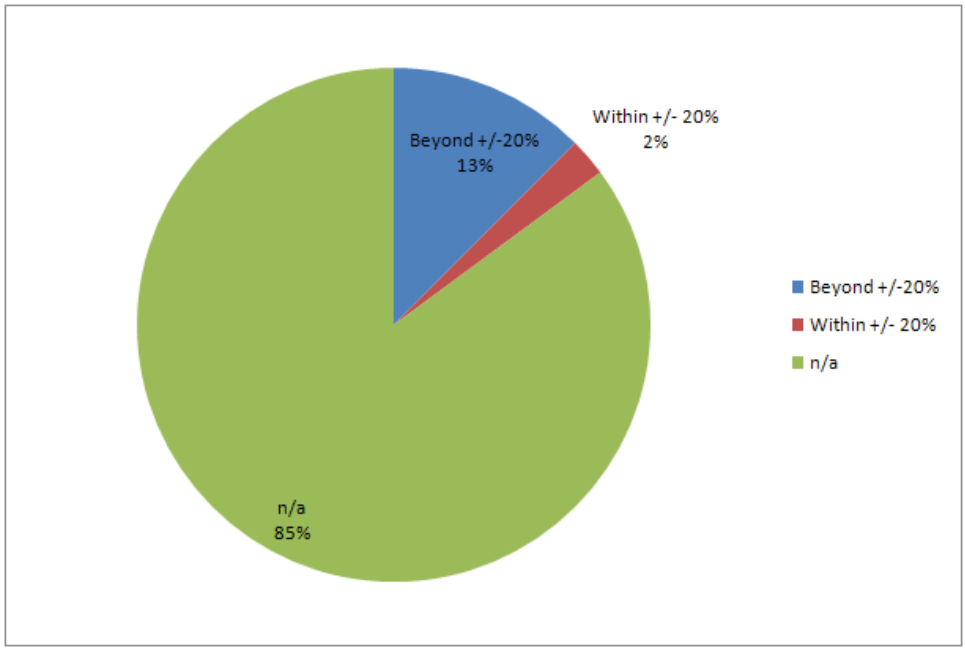


Figure L-58. Test 4: Results within +/-20% Madison, Perturbed vs Model Transit

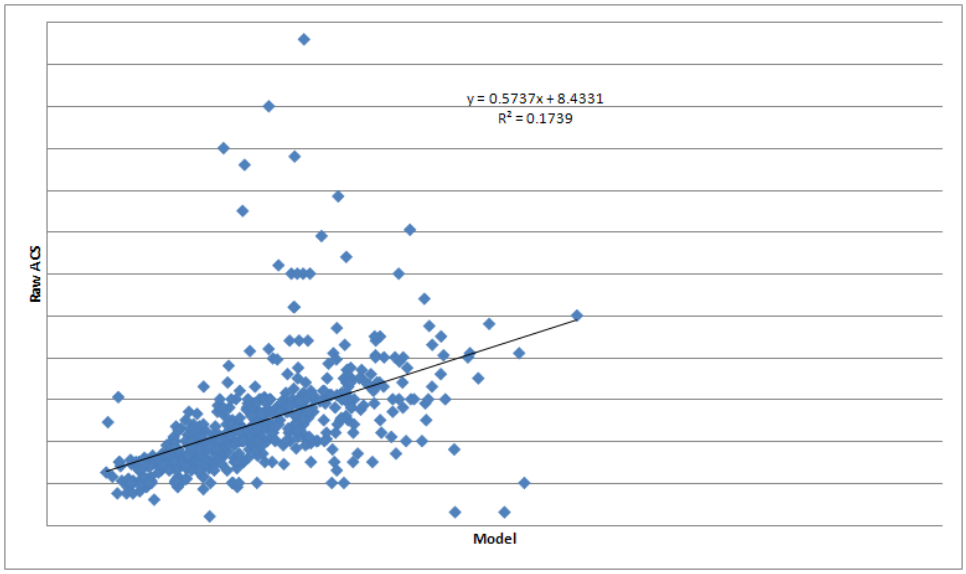


Figure L-59. Test 4: Scatterplot of Model vs Raw ACS, Madison, District, Auto

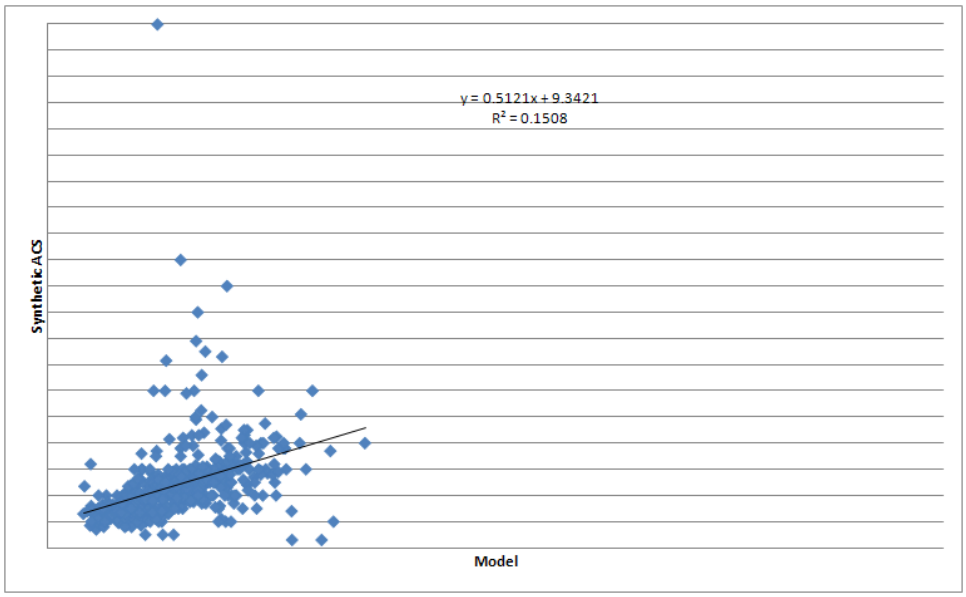


Figure L-60. Test 4: Scatterplot of Model vs Perturbed ACS, Madison, District, Auto

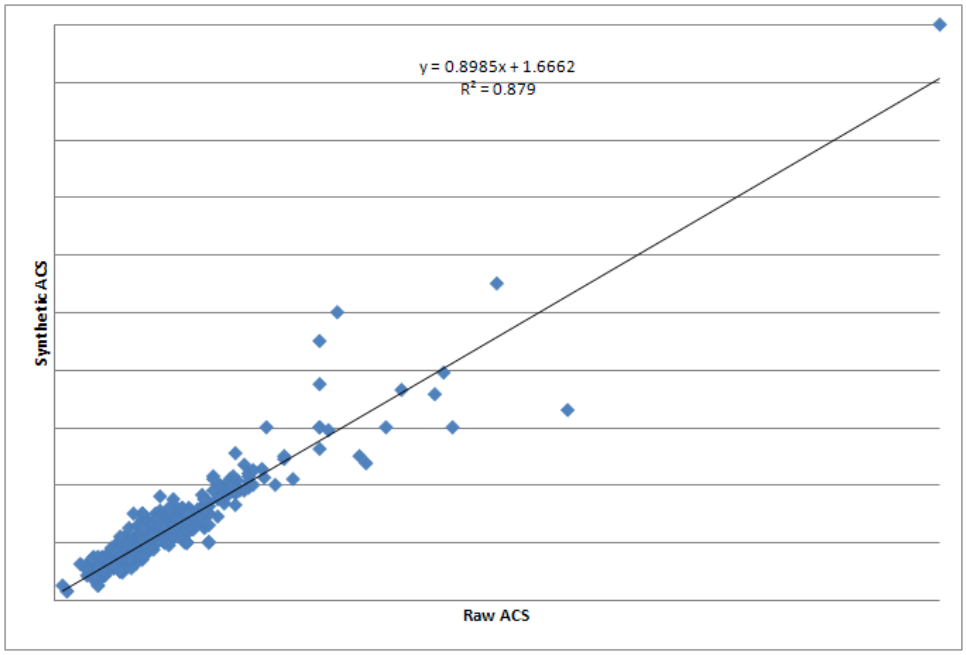


Figure L-61. Test 4: Scatterplot of Raw ACS vs Perturbed ACS, Madison, District, Auto

Appendix M

VALIDATION PHASE: TABULAR SUMMARY RESULTS FOR THE DATA UTILITY MEASURES

Table M-1. Median and Interquartile Range of Cell Mean Differences, TAZ Level, Partial Replacement

		Development (3-year ACS)				Validation (5-year ACS)			
		ATLANTA		MADISON		ATLANTA		OLYMPIA	
Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
AHINC	VEHICLES6_2	0.0	1733.9	0.0	2288.3	0.0	637.0	0.0	1085.9
AHINC	VEHICLES6_3	0.0	2159.8	0.0	3359.0	0.0	1183.1	-12.9	1014.3
AHINC	VEHICLES6_4	37.5	3998.9	0.0	7040.9	32.9	3199.3	-4.3	3277.2
AHINC	VEHICLES6_5	0.0	6562.9	0.0	9368.3	10.1	4986.0	0.0	5510.0
AHINC	VEHICLES6_6	0.0	7559.9	0.0	9359.0	0.0	4839.0	0.0	6534.5
JWMN	MEANS6_2	0.1	1.8	0.0	3.0	0.0	1.3	0.0	2.2
JWMN	MEANS6_3	0.0	3.8	0.0	5.2	0.0	2.3	0.0	3.0
JWMN	MEANS6_4	0.0	4.3	0.0	7.3	0.0	0.4	0.0	0.3
JWMN	MEANS6_5	0.0	6.3	0.0	6.0	0.0	2.3	0.0	3.6
JWMN	TM_LEAVE5_3	0.0	2.1	0.0	3.2	0.0	1.7	0.1	2.2
JWMN	TM_LEAVE5_4	0.1	3.1	0.1	4.4	0.0	2.8	0.0	4.1

Table M-2. Median and Interquartile Range of Cell Mean Differences, County Level, Partial Replacement

		Development (3-year ACS)				Validation (5-year ACS)			
		ATLANTA		MADISON		ATLANTA		OLYMPIA	
Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
AHINC	VEHICLES6_2	95.7	661.0	50.4	0.0	-0.9	86.0	-178.5	0.0
AHINC	VEHICLES6_3	91.6	403.2	217.1	0.0	-4.9	79.2	-29.8	0.0
AHINC	VEHICLES6_4	5.0	772.4	-621.6	0.0	-8.7	220.8	-144.4	0.0
AHINC	VEHICLES6_5	154.0	1172.3	-139.9	0.0	92.1	418.5	-346.4	0.0
AHINC	VEHICLES6_6	-282.4	2231.1	32.4	0.0	-44.9	941.5	-298.0	0.0
JWMN	MEANS6_2	0.0	0.3	0.0	0.0	0.0	0.1	0.0	0.0
JWMN	MEANS6_3	-0.1	0.8	-0.1	0.0	0.0	0.3	0.0	0.0
JWMN	MEANS6_4	-0.1	2.1	-0.3	0.0	0.0	0.2	-0.1	0.0
JWMN	MEANS6_5	0.1	2.5	-0.2	0.0	0.1	0.4	0.1	0.0
JWMN	TM_LEAVE5_3	0.0	0.3	-0.1	0.0	0.0	0.1	0.0	0.0
JWMN	TM_LEAVE5_4	0.1	0.3	0.2	0.0	0.0	0.2	0.0	0.0
JWMN	Flow	0.0	3.2	0.0	6.4	0.0	1.8	0.0	2.2

Table M-3. Median and Interquartile Range of Cell Quantile Differences, TAZ Level, Partial Replacement

		Cell Median				Cell 75th Percentiles			
		ATLANTA		OLYMPIA		ATLANTA		OLYMPIA	
Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
AHINC	VEHICLES6_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	371.8
AHINC	VEHICLES6_3	0.0	0.0	0.0	0.0	0.0	148.5	0.0	0.0
AHINC	VEHICLES6_4	0.0	1991.0	0.0	358.0	0.0	4745.0	0.0	3619.0
AHINC	VEHICLES6_5	0.0	2999.4	0.0	3507.3	0.0	5432.0	0.0	4013.0
AHINC	VEHICLES6_6	0.0	912.5	0.0	3261.0	0.0	2575.6	0.0	8973.0
JWMN	MEANS6_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
JWMN	MEANS6_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	MEANS6_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	MEANS6_5	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0
JWMN	TM_LEAVE5_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0
JWMN	TM_LEAVE5_4	0.0	0.0	0.0	4.0	0.0	0.0	0.0	5.0

Table M-4. Median and Interquartile Range of Cell Quantile Differences, County Level, Partial Replacement

		Cell Median				Cell 75th Percentiles			
		ATLANTA		OLYMPIA		ATLANTA		OLYMPIA	
Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
AHINC	VEHICLES6_2	0.0	28.5	-33.0	0.0	0.0	0.0	0.0	0.0
AHINC	VEHICLES6_3	0.0	96.5	7.0	0.0	0.0	193.5	-271.0	0.0
AHINC	VEHICLES6_4	0.0	222.3	-157.0	0.0	-31.0	500.0	-162.0	0.0
AHINC	VEHICLES6_5	12.5	841.3	-384.0	0.0	-0.5	944.6	-544.0	0.0
AHINC	VEHICLES6_6	0.0	421.5	-10.0	0.0	0.0	1226.5	-198.0	0.0
JWMN	MEANS6_2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	MEANS6_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	MEANS6_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	MEANS6_5	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0
JWMN	TM_LEAVE5_3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	TM_LEAVE5_4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
JWMN	Flow	0.0	0.0	0.0	5.0	0.0	0.0	0.0	3.0

Table M-5. Median and Interquartile Range of Standard Error Differences, County Level, Partial Replacement

		Development (3-year ACS)				Validation (5-year ACS)			
		ATLANTA		MADISON		ATLANTA		OLYMPIA	
Attribute	BYVAR	Median	IQR	Median	IQR	Median	IQR	Median	IQR
AHINC	VEHICLES6_2	62.23	156.86	36.22	0.00	2.29	19.70	-81.68	0.00
AHINC	VEHICLES6_3	20.76	75.04	39.98	0.00	6.07	27.79	2.56	0.00
AHINC	VEHICLES6_4	99.95	154.01	70.82	0.00	21.31	84.80	67.26	0.00
AHINC	VEHICLES6_5	99.46	467.41	63.45	0.00	41.14	190.77	-109.98	0.00
AHINC	VEHICLES6_6	163.36	506.26	66.73	0.00	20.36	238.24	73.67	0.00
JWMN	MEANS6_2	0.02	0.05	0.01	0.00	0.02	0.04	0.02	0.00
JWMN	MEANS6_3	0.05	0.26	0.03	0.00	0.03	0.06	0.00	0.00
JWMN	MEANS6_4	0.07	0.74	0.09	0.00	0.01	0.06	-0.19	0.00
JWMN	MEANS6_5	0.05	0.83	0.14	0.00	0.03	0.09	0.06	0.00
JWMN	TM_LEAVE5_3	0.02	0.08	0.05	0.00	0.01	0.08	0.05	0.00
JWMN	TM_LEAVE5_4	0.01	0.13	0.09	0.00	0.02	0.05	0.01	0.00

Table M-6. Median and Interquartile Range of Cramers V Differences, Partial Replacement

		Development (3-year ACS)				Validation (5-year ACS)			
		ATLANTA		MADISON		ATLANTA		OLYMPIA	
GEOAREA	MOT	Median	IQR	Median	IQR	Median	IQR	Median	IQR
County	AGE9	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00
County	HH_INC26	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00
County	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
County	TRAVEL_TM12	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
County	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
County flows	AGE9	0.00	0.01	0.00	0.01	0.00	0.06	0.00	0.11
County flows	HH_INC26	0.00	0.01	0.00	0.03	0.00	0.01	0.00	0.01
County flows	TM_LEAVE10	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
County flows	TRAVEL_TM12	0.00	0.02	0.00	0.08	0.00	0.01	0.00	0.05
County flows	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TAZ	AGE9	0.00	0.03	0.00	0.04	0.00	0.06	0.00	0.09
TAZ	HH_INC26	0.00	0.07	0.00	0.05	0.00	0.06	0.00	0.05
TAZ	TM_LEAVE10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.03
TAZ	TRAVEL_TM12	0.00	0.05	0.00	0.06	0.00	0.04	0.00	0.06
TAZ	VEHICLES6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table M-7. Pairwise Correlations Between Key Ordinal Variables, Partial Replacement, by Perturbation Run

				Perturbed Run				
Phase	Var1	Var2	Actual	1	2	3	4	5
Development (Atlanta)	AHINC	AGE9	0.0949	0.0929	0.0926	0.0952	0.0950	0.0904
	AHINC	POVERTY	0.2510	0.2504	0.2507	0.2542	0.2523	0.2520
	AGE9	POVERTY	0.1308	0.1235	0.1238	0.1259	0.1246	0.1242
	POVERTY	JWMN	0.0061	0.0059	0.0065	0.0028	0.0025	0.0047
Validation (Atlanta)	AHINC	AGE9	0.0933	0.0840	0.0878	0.0858	0.0852	0.0875
	AHINC	POVERTY	0.2562	0.2511	0.2516	0.2550	0.2506	0.2530
	AGE9	POVERTY	0.1312	0.1107	0.1150	0.1108	0.1131	0.1119
	POVERTY	JWMN	0.0050	0.0057	0.0062	0.0054	0.0045	0.0072
	JWMN	JWD	-0.1195	-0.1186	-0.1184	-0.1175	-0.1178	-0.1169
	AHINC	HH WORKERS	0.1771	0.1747	0.1767	0.1765	0.1771	0.1755
	AGE9	HH WORKERS	-0.1949	-0.1742	-0.1741	-0.1743	-0.1724	-0.1743
	POVERTY	HH WORKERS	0.0624	0.0618	0.0612	0.0612	0.0610	0.0604
Validation (Olympia)	AHINC	AGE9	0.1123	0.1023	0.0954	0.1007	0.1020	0.1096
	AHINC	POVERTY	0.2845	0.2949	0.2846	0.2941	0.2940	0.2927
	AGE9	POVERTY	0.1700	0.1314	0.1366	0.1326	0.1358	0.1540
	POVERTY	JWMN	0.0512	0.0400	0.0367	0.0323	0.0330	0.0295
	JWMN	JWD	-0.1245	-0.1290	-0.1301	-0.1281	-0.1315	-0.1276
	AHINC	HH WORKERS	0.2976	0.2948	0.2970	0.3032	0.2925	0.3009
	AGE9	HH WORKERS	-0.2210	-0.1894	-0.1890	-0.1935	-0.1903	-0.1981
	POVERTY	HH WORKERS	0.0827	0.0956	0.0913	0.0976	0.0958	0.0926

Table M-8. Multivariate Association Measure U, by Perturbation Run

			Run				
Phase	**Amount**	**Test site**	**1**	**2**	**3**	**4**	**5**
Development	Full	ATLANTA	0.000026	0.000027	0.000022	0.000019	0.000017
		IOWA	0.000078	0.000077	0.000074	0.000077	0.000075
		MADISON	0.000052	0.000066	0.000040	0.000047	0.000063
		ST.LOUIS	0.000036	0.000040	0.000038	0.000043	0.000043
	Partial	ATLANTA	0.000005	0.000005	0.000003	0.000005	0.000005
		IOWA	0.000023	0.000024	0.000024	0.000022	0.000024
		MADISON	0.000026	0.000026	0.000036	0.000032	0.000023
		ST.LOUIS	0.000031	0.000020	0.000022	0.000021	0.000027
Validation	Partial	ATLANTA	0.000010	0.000009	0.000009	0.000001	0.000008
		OLYMPIA	0.000060	0.000054	0.000051	0.000044	0.000057

Appendix N

VALIDATION PHASE: PLOTS OF ACS AND PERTURBED DATA FOR MEAN TRAVEL TIME

Note: The dots in Figures E-1 through E-4 are shown only for localities with 30 or more ACS sample cases. The dots in Figures E-5 through E-6 are shown only for flows with at least 10 ACS cases. ACS values are on horizontal axis. Perturbed values are on vertical axis. The line has a slope of 1. Bubble radii are proportional to the estimated worker population.

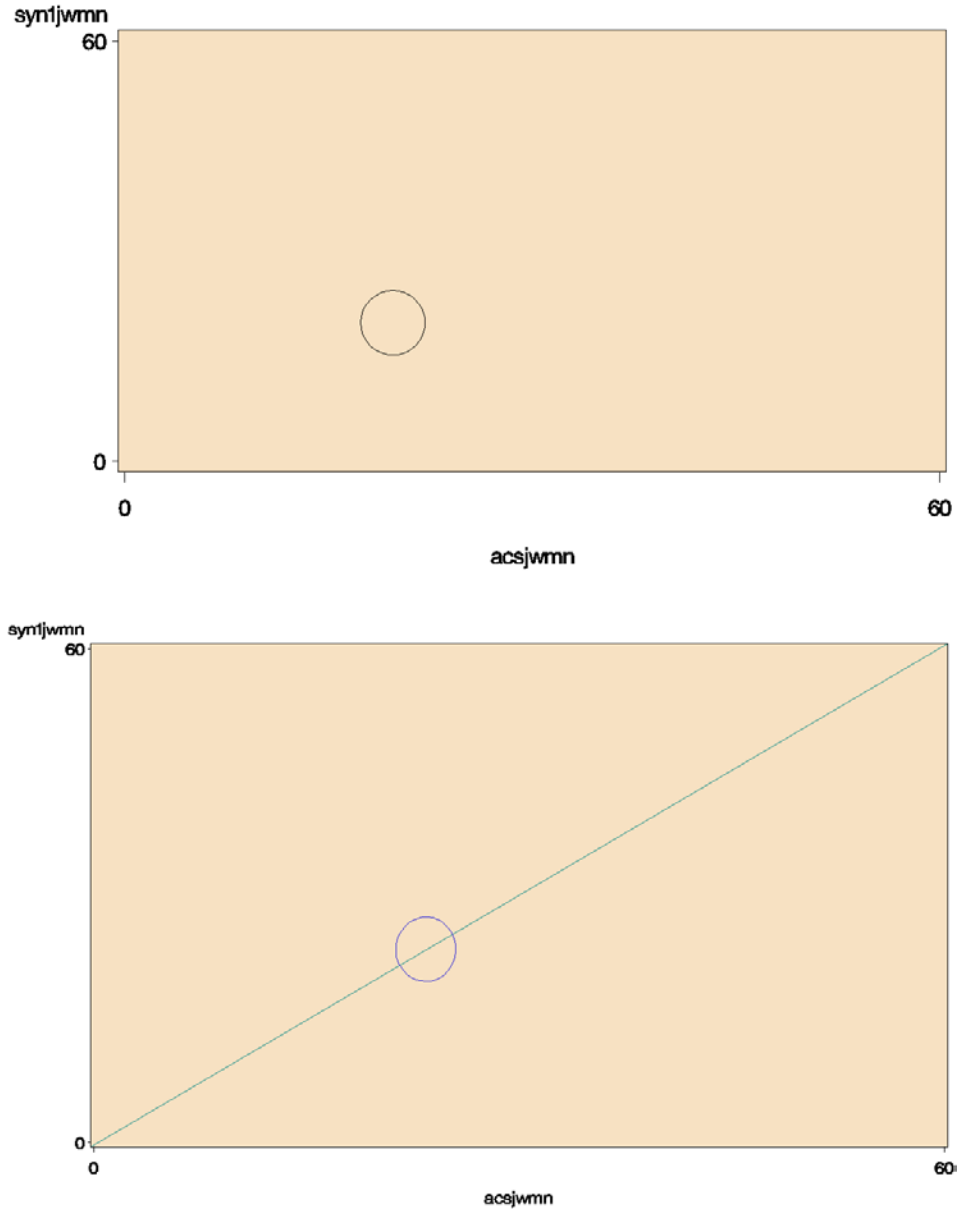


Figure N-1. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Madison's and Olympia's County: Top: Development Phase -- Madison, Bottom: Validation Phase – Olympia (n≥30)

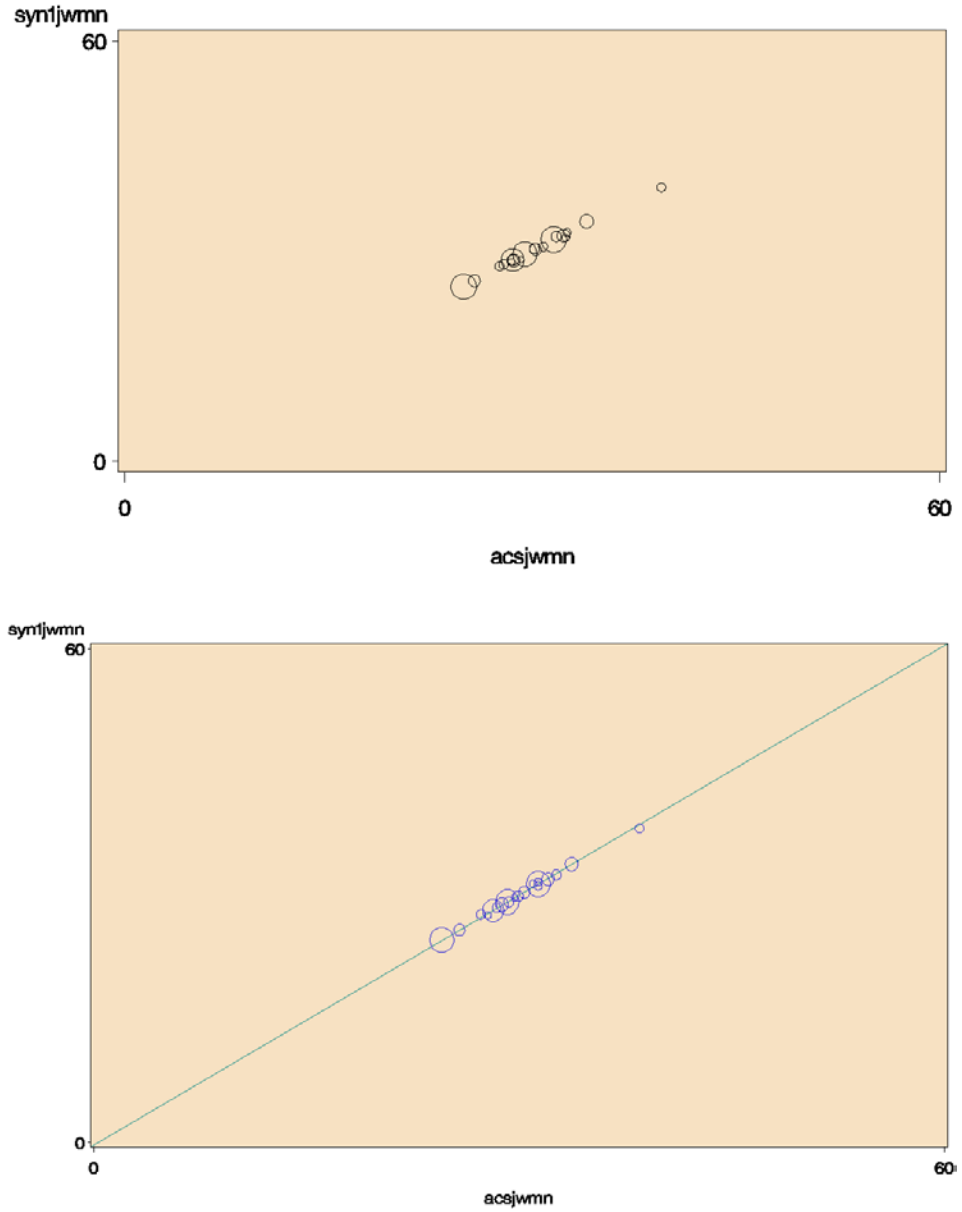


Figure N-2. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's Counties: Top: Development Phase, Bottom: Validation Phase (n≥30)

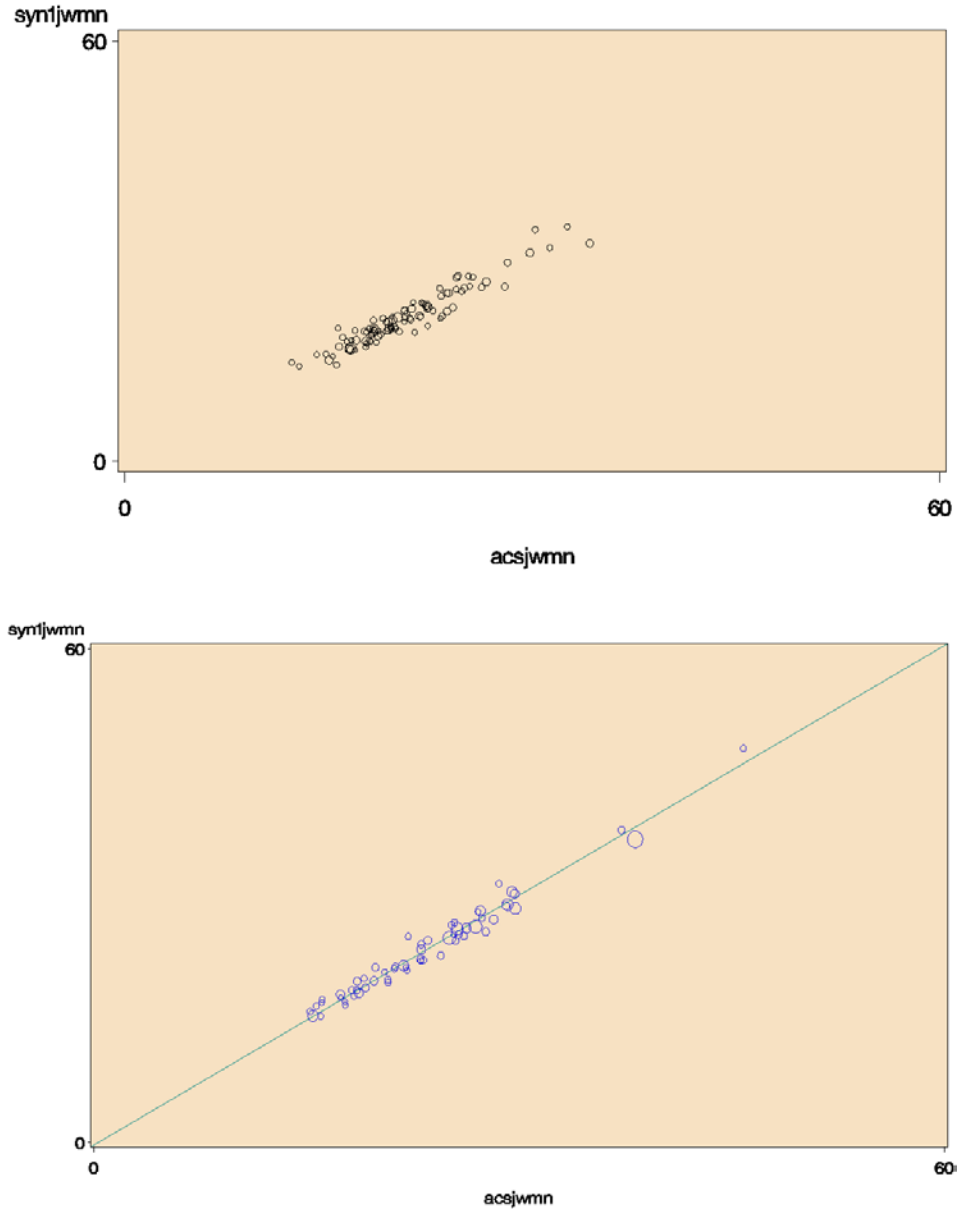


Figure N-3. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Madison's and Olympia's TAZs: Top: Development Phase -- Madison, Bottom: Validation Phase -- Olympia (n≥30)

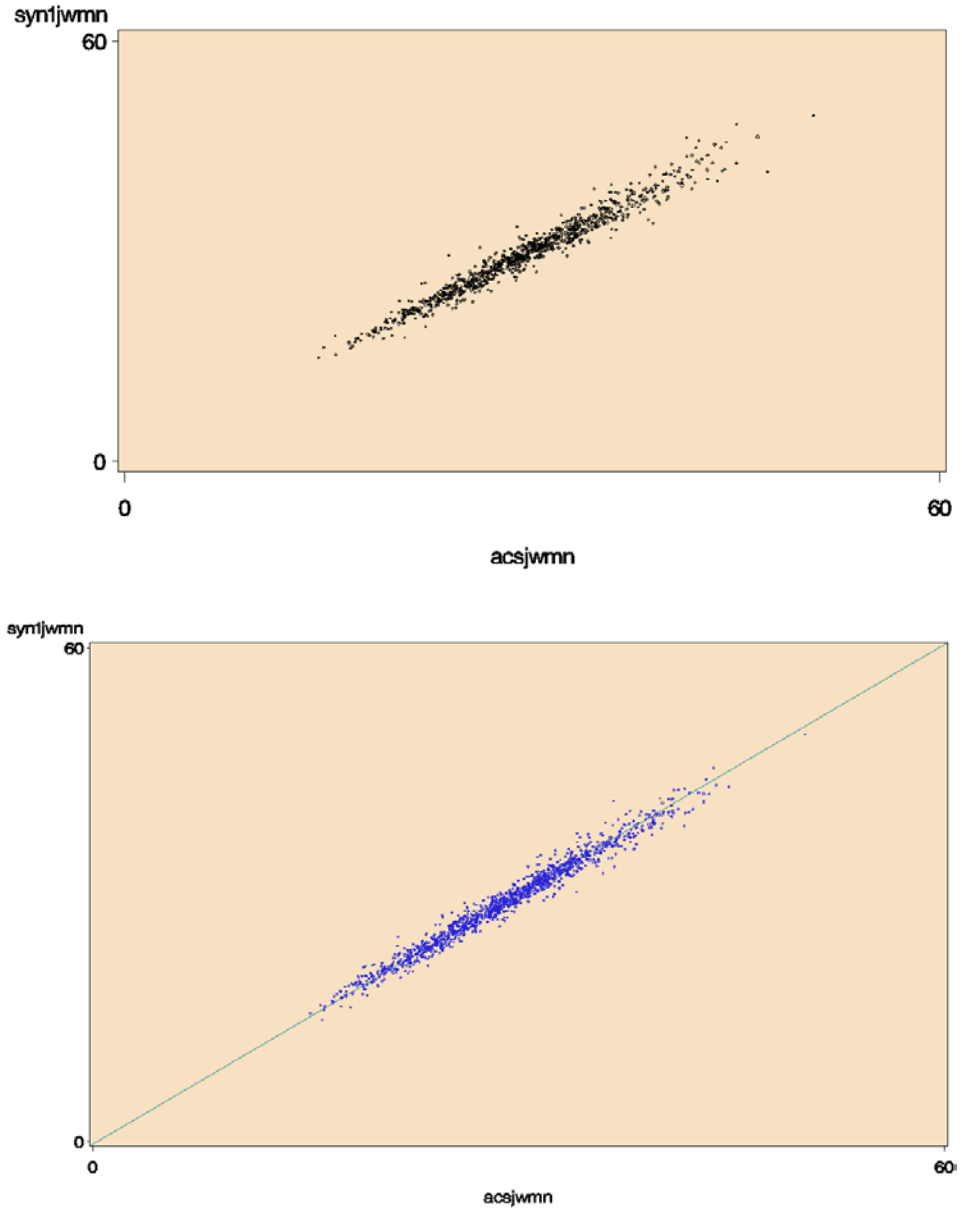


Figure N-4. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's TAZs: Top: Development Phase, Bottom: Validation Phase (n≥30)

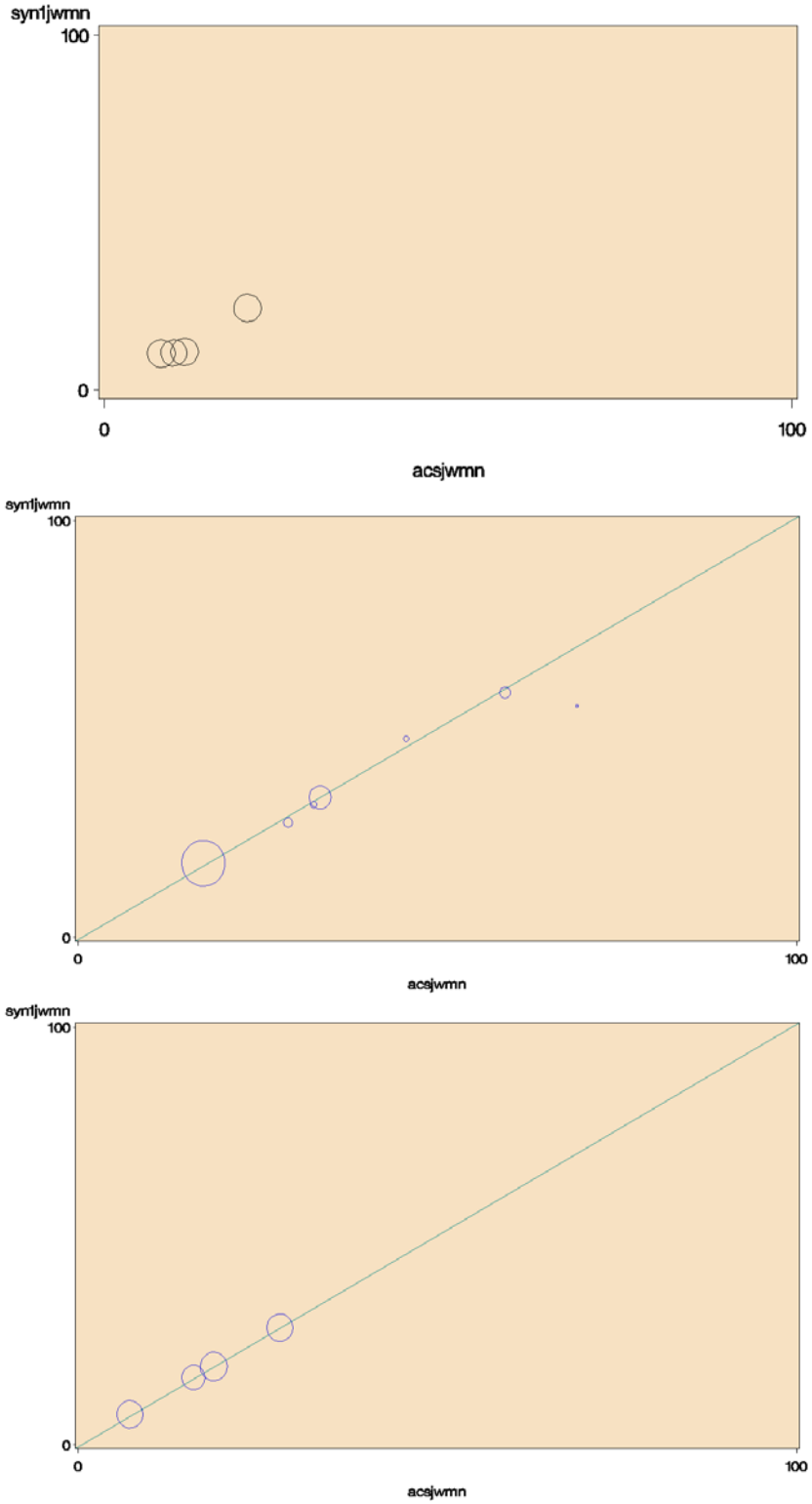


Figure N-5. Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Madison's and Olympia's Flows: Top: Development phase TAZ flows for Madison, Middle: Validation phase county flows for Olympia, Bottom: Validation phase TAZ flows for Olympia (n≥10)

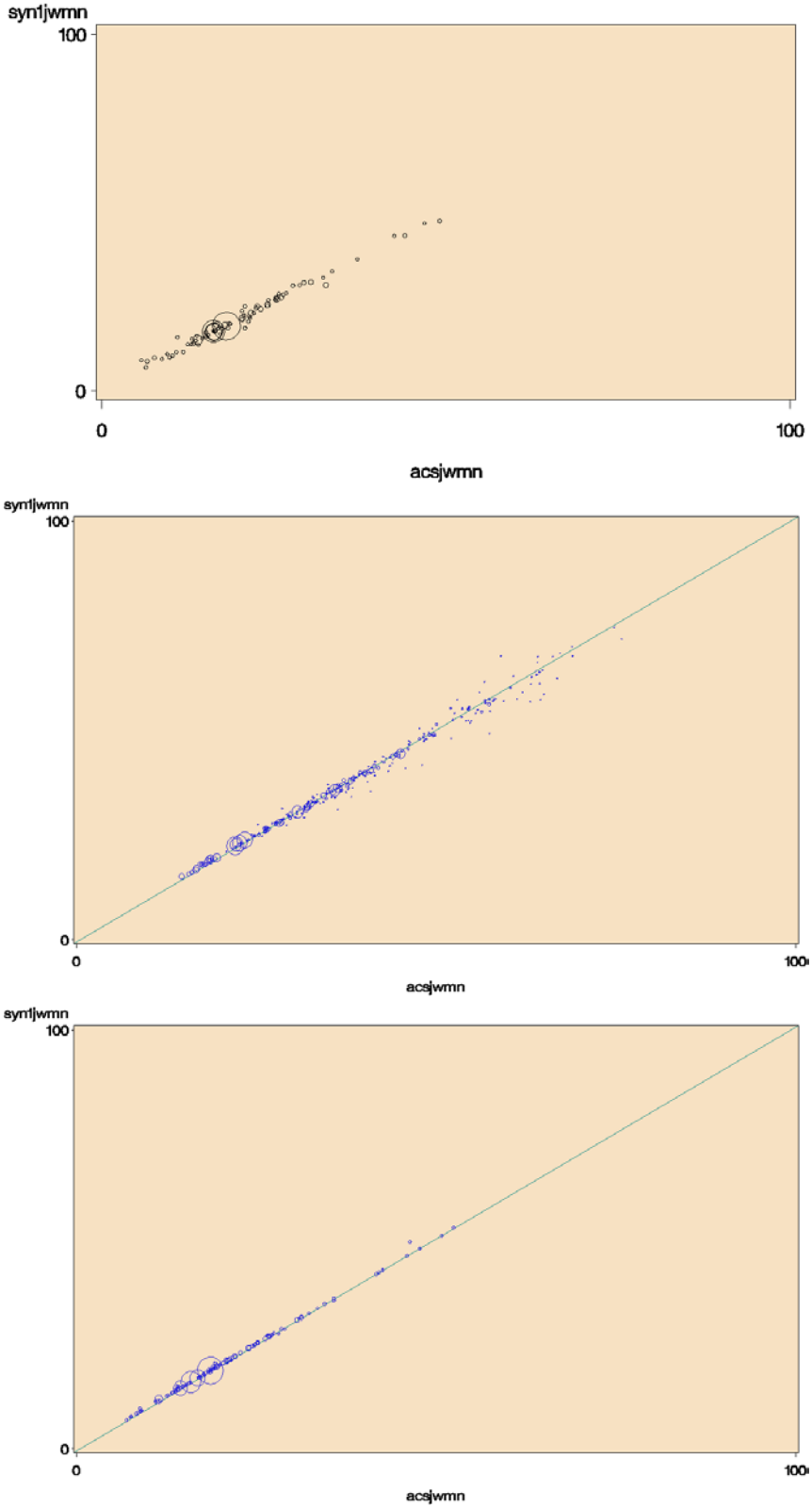


Figure N-6. *Bubble Plots of ACS and Partial Perturbed Mean Travel Time for Atlanta's Flows: Top: Development phase TAZ flows, Middle: Validation phase county flows, Bottom: Validation phase TAZ flows (n≥10*

Appendix O

VALIDATION PHASE: PLOTS OF ACS AND PERTURBED DATA FOR MEAN HOUSEHOLD (HH) INCOME

Note: The dots in Figures F-1 through F-4 are shown only for localities with 30 or more ACS cases. ACS values are on horizontal axis. Perturbed values are on vertical axis. The line has a slope of 1. Bubble radii are proportional to the estimate worker population.

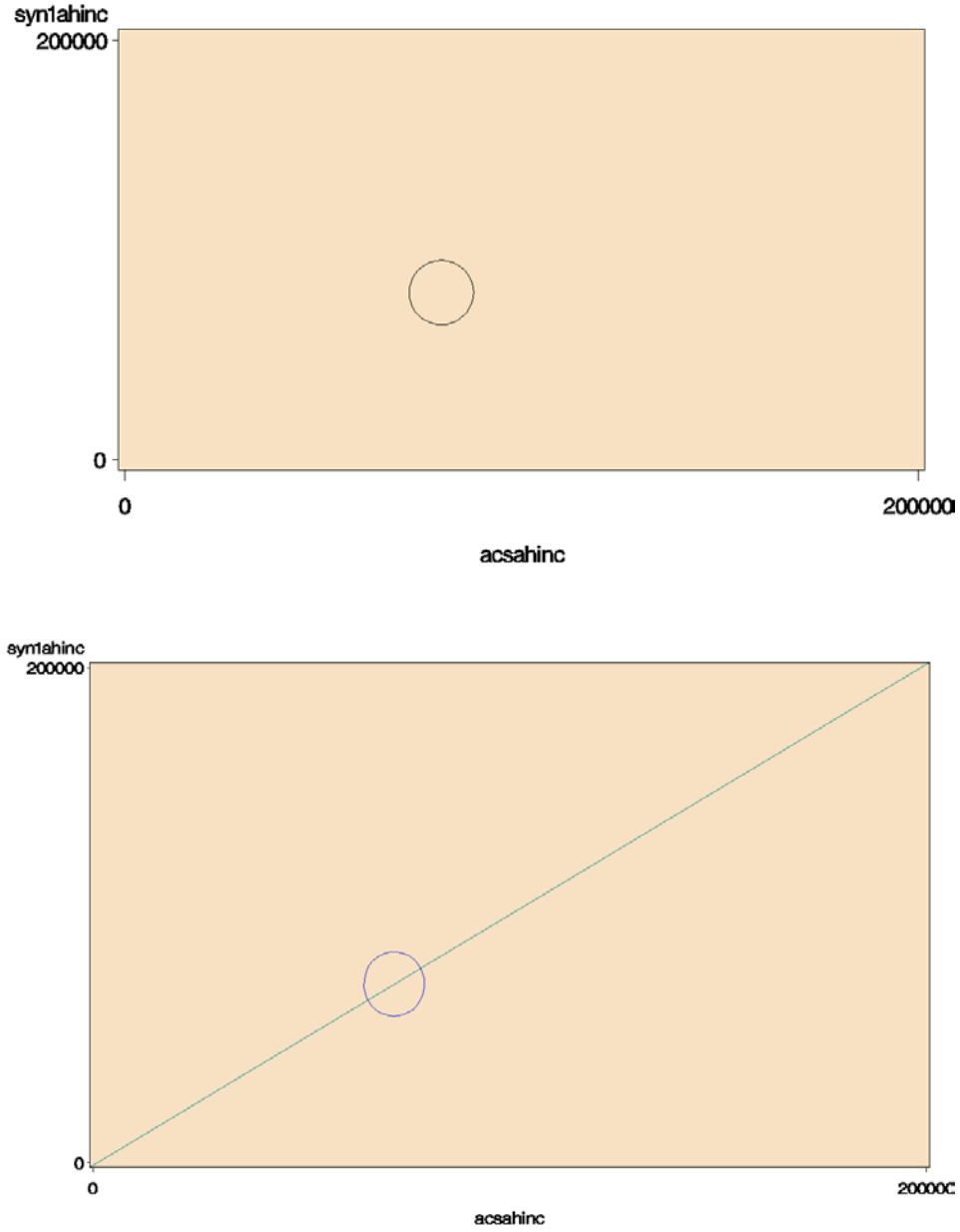


Figure O-1. Bubble Plots of ACS and Partial Perturbed Mean Household Income for Madison's and Olympia's County: Top: Development Phase -- Madison, Bottom: Validation Phase – Olympia (n≥30)

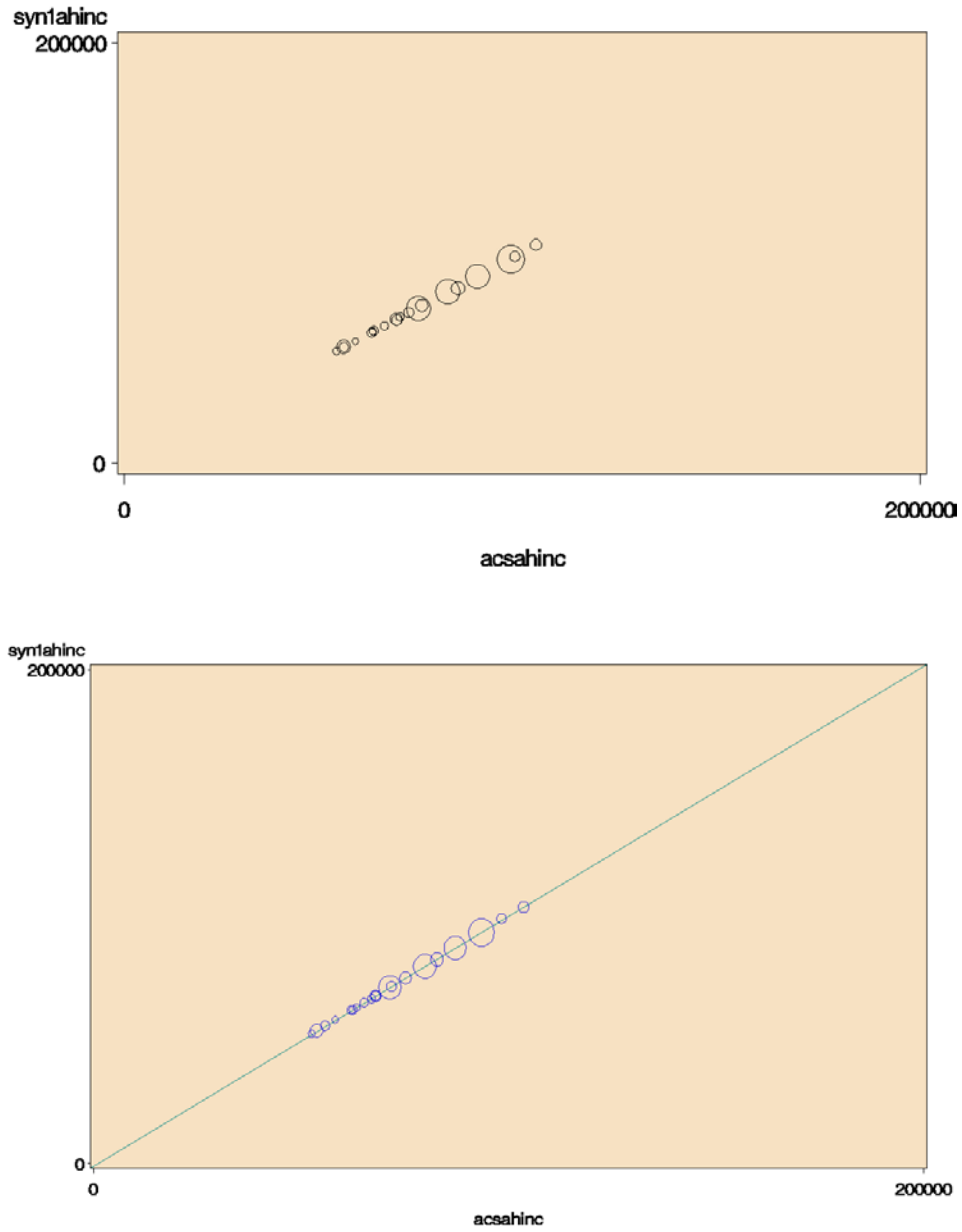


Figure O-2. Bubble Plots of ACS and Partial Perturbed Mean Household Income for Atlanta's Counties: Top: Development Phase, Bottom: Validation Phase (n≥30)

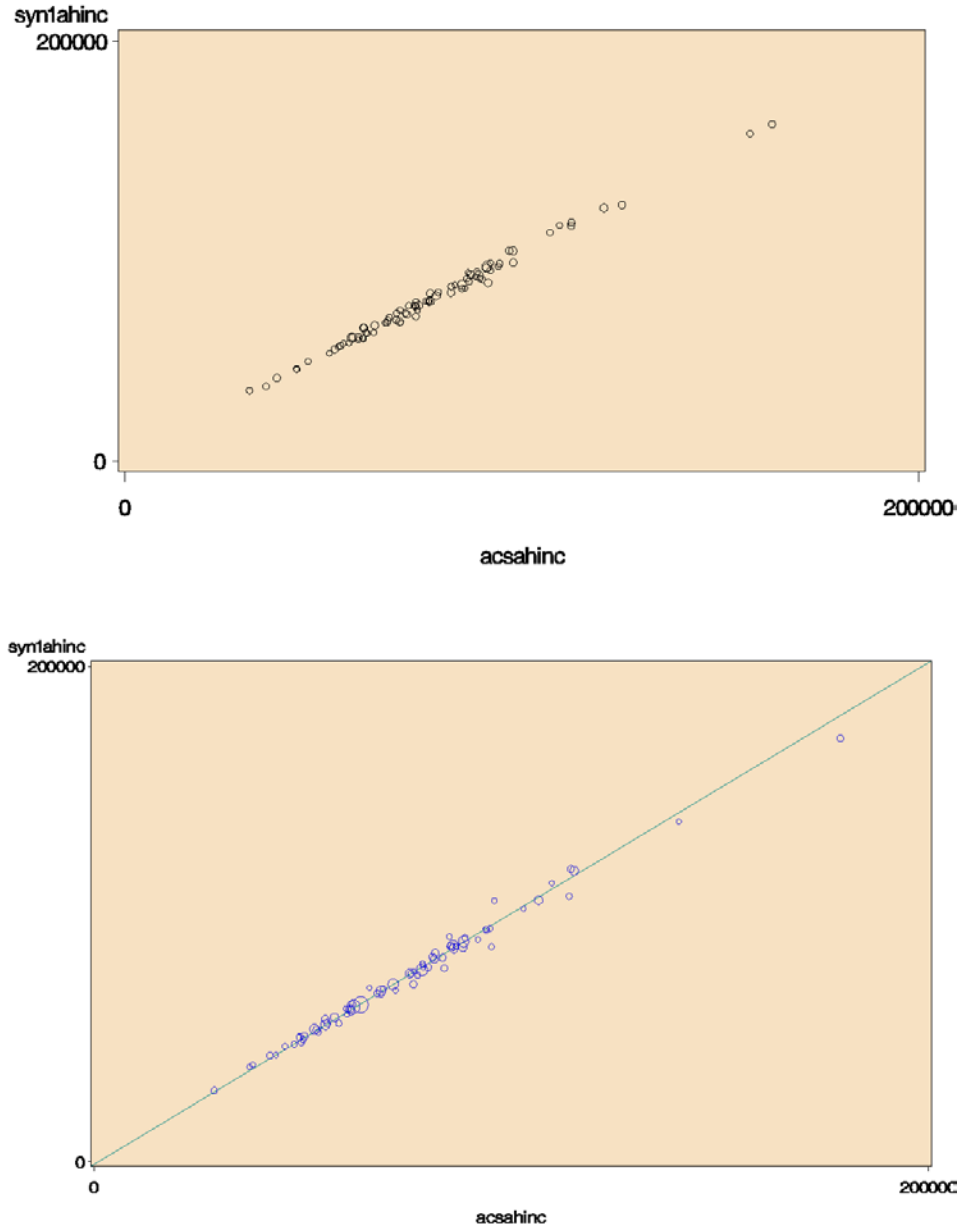


Figure O-3. Bubble Plots of ACS and Partial Perturbed Mean Household Income for Madison's and Olympia's TAZs: Top: Development Phase -- Madison, Bottom: Validation Phase -- Olympia (n≥30)

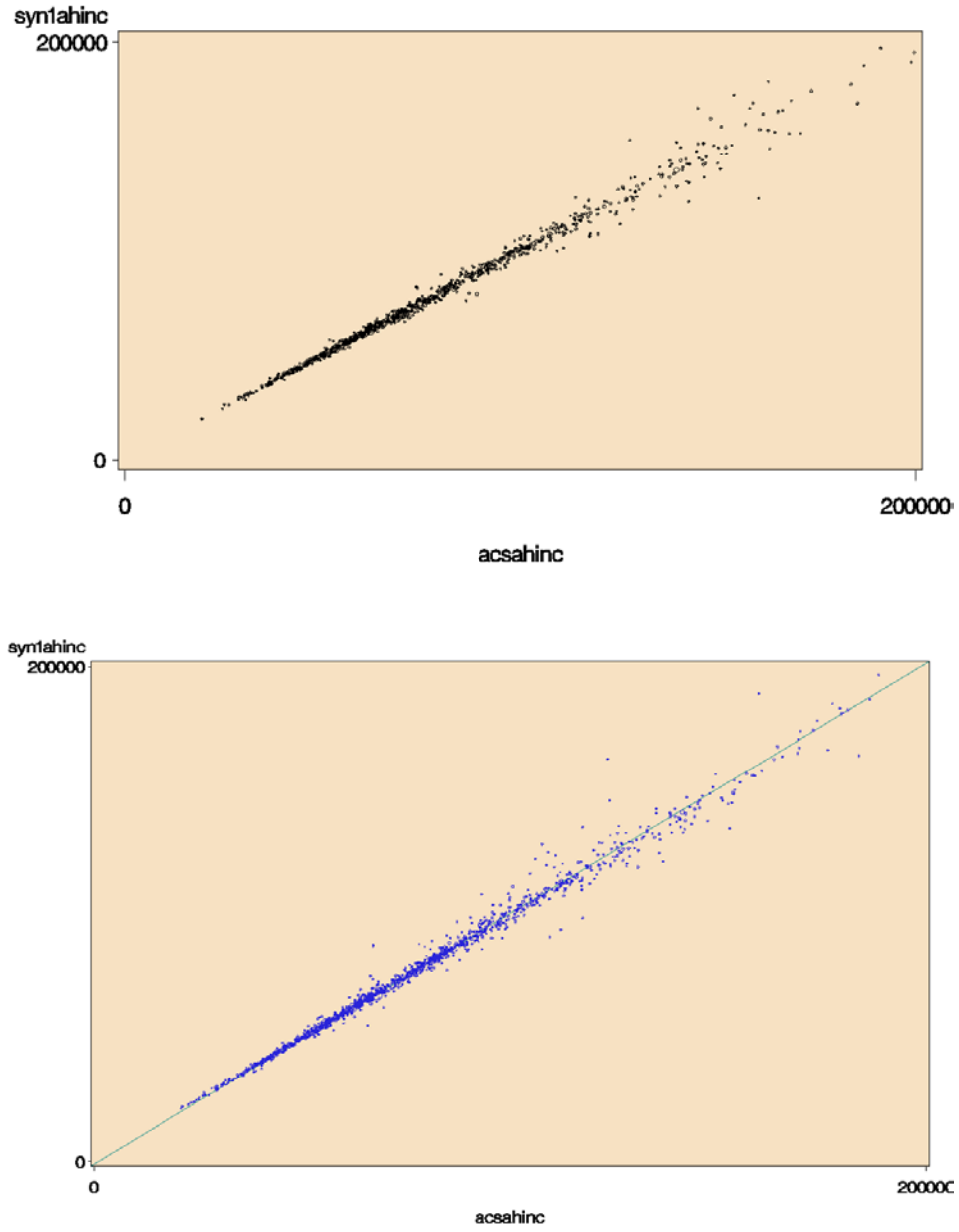


Figure O-4. Bubble Plots of ACS and Partial Perturbed Mean Household Income for Atlanta's TAZs: Top: Development Phase, Bottom: Validation Phase (n≥30)

Appendix P

VALIDATION PHASE: PLOTS OF ACS PARTIALLY PERTURBED DATA FOR WEIGHTED COUNTS

Note: For the county and TAZ level plots, a dot is shown only if there are at least 10 sample cases and 10 perturbed cases in the cell. For county flow plots, a dot is shown only for flows with at least 5 sample cases and 5 perturbed cases in the cell. ACS values are on horizontal axis. Perturbed values are on the vertical axis. The line has a slope of 1. Bubble radii are proportional to the estimated population off workers. CTAZ1000 denotes TAZs that were combined until there were at least 1000 ACS sample cases, which represents approximately 25,000 in population.

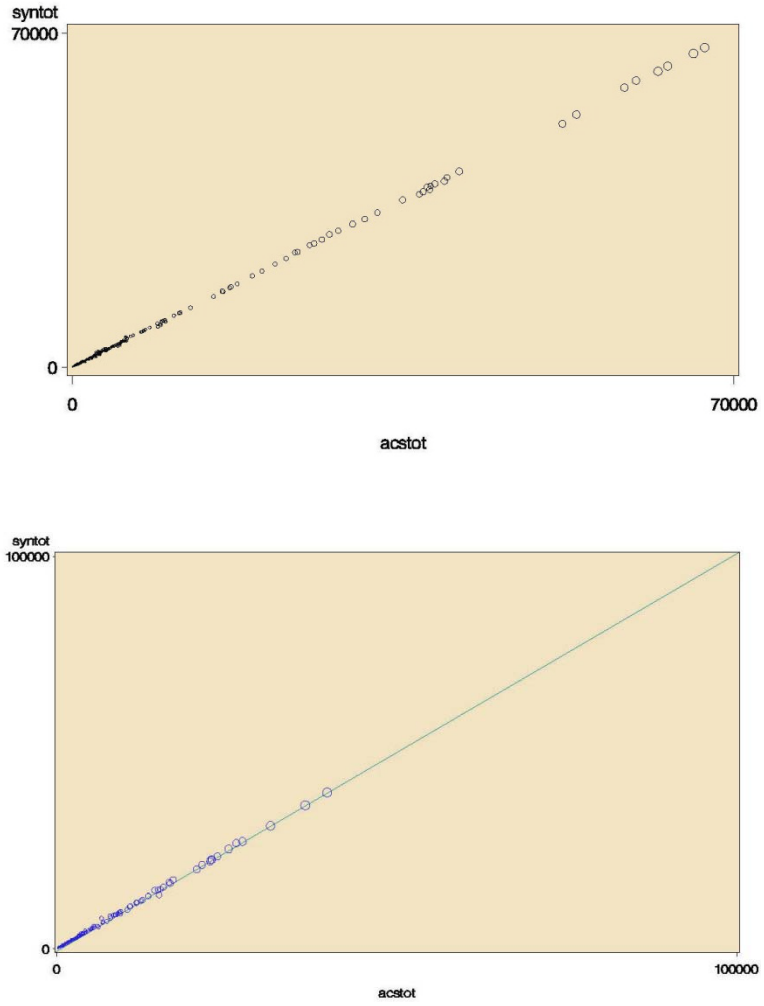


Figure P-1. Plot of ACS Partially Perturbed Weighted Counts for Madison's and Olympia's County: Top: Development Phase, Bottom: Validation Phase (Different Scale)

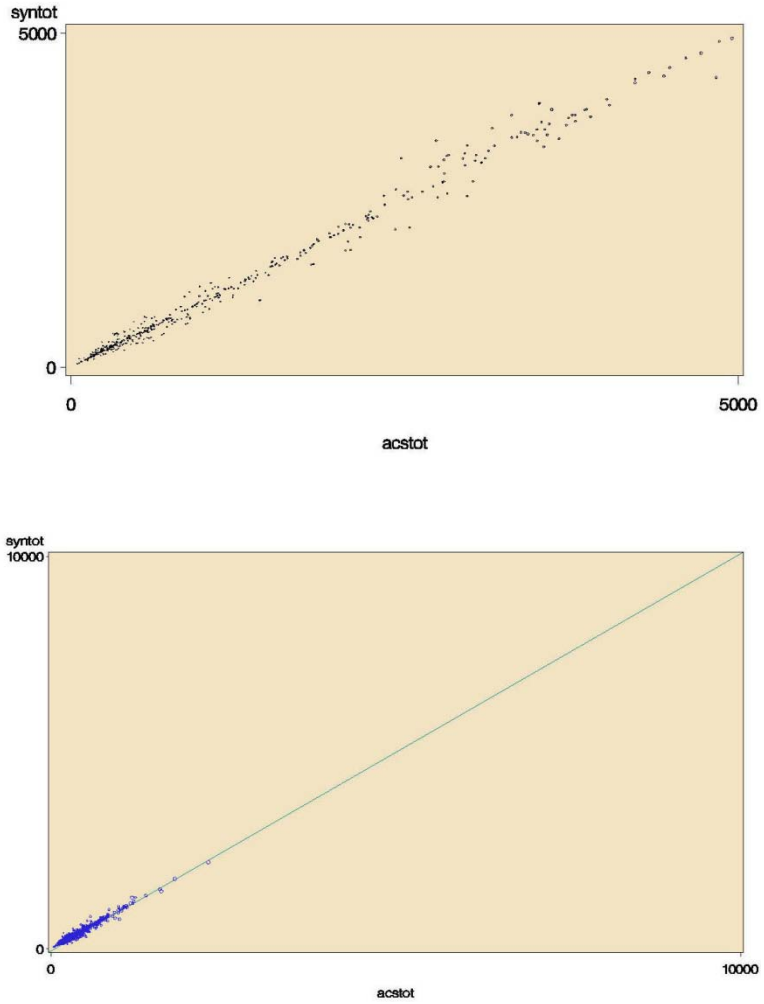


Figure P-2. Plot of ACS Partially Perturbed Weighted Counts for Madison's and Olympia's TAZs: Top: Development Phase, Bottom: Validation Phase (Different Scale)

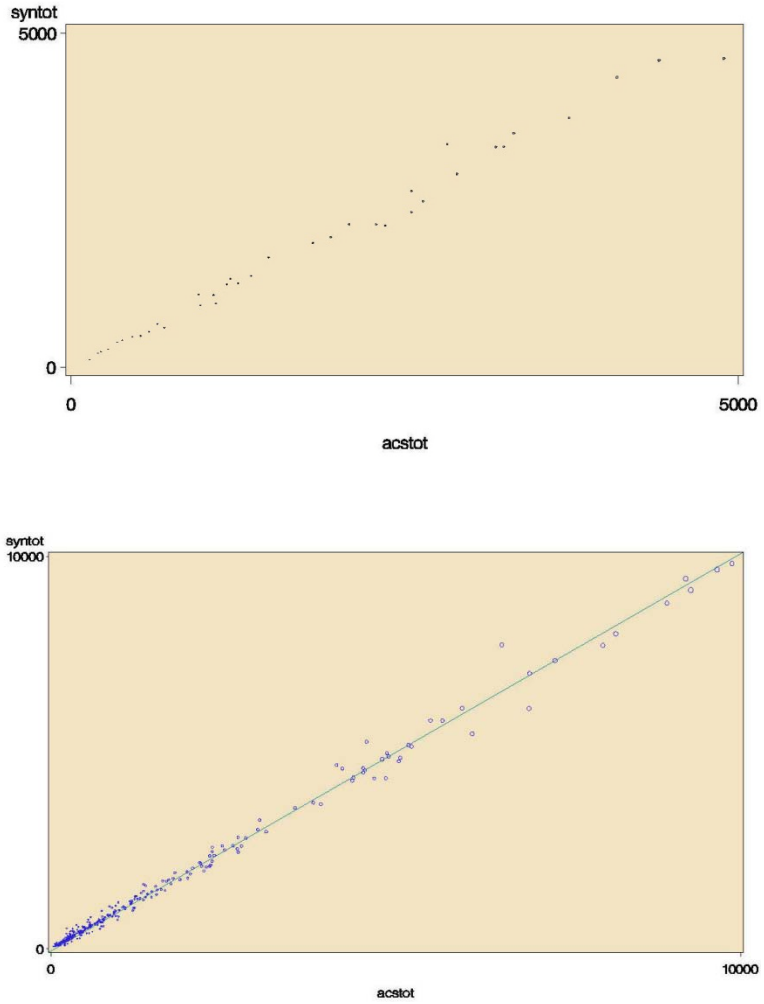


Figure P-3. Plot of ACS Partially Perturbed Weighted Counts for Madison's and Olympia's County Flows: Top: Development Phase, Bottom: Validation Phase (Different Scale)

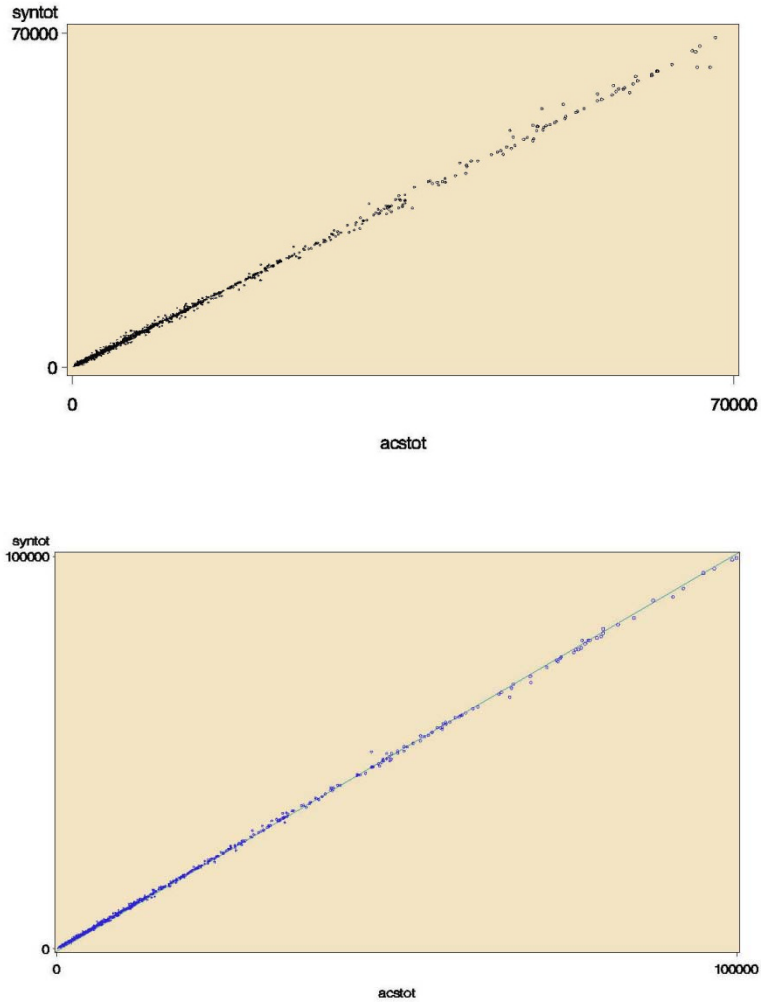


Figure P-4. Plot of ACS Partially Perturbed Weighted Counts for Atlanta's Counties: Top: Development Phase, Bottom: Validation Phase (Different Scale)

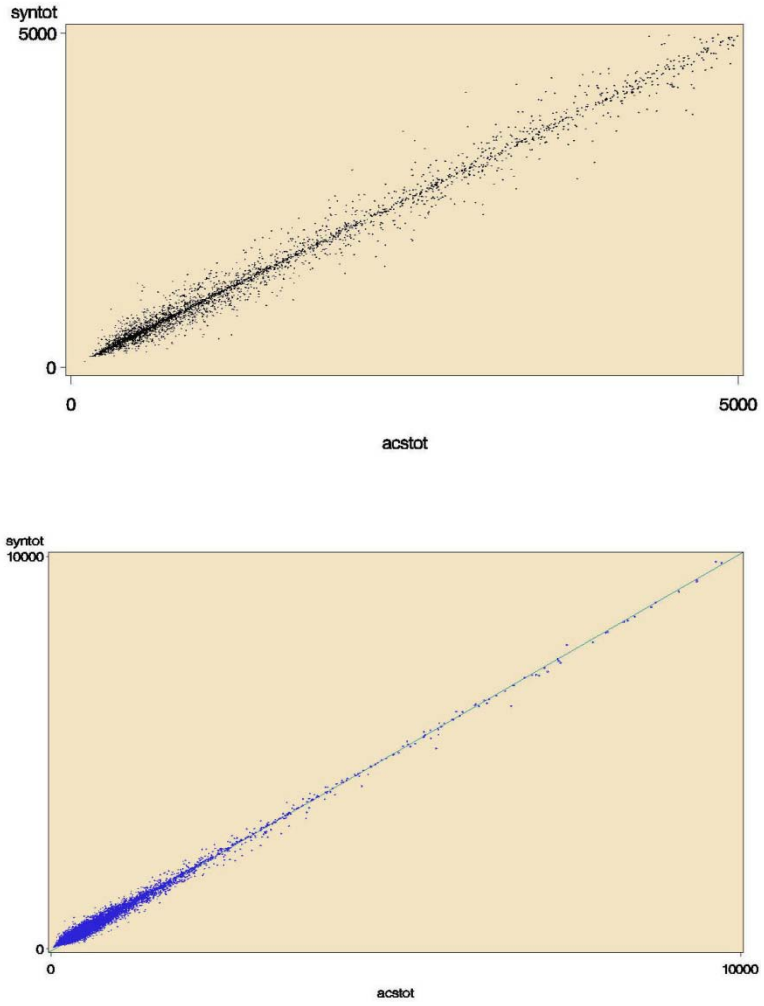


Figure P-5. Plot of ACS Partially Perturbed Weighted Counts for Atlanta's TAZs: Top: Development Phase Bottom: Validation Phase (Different Scale)

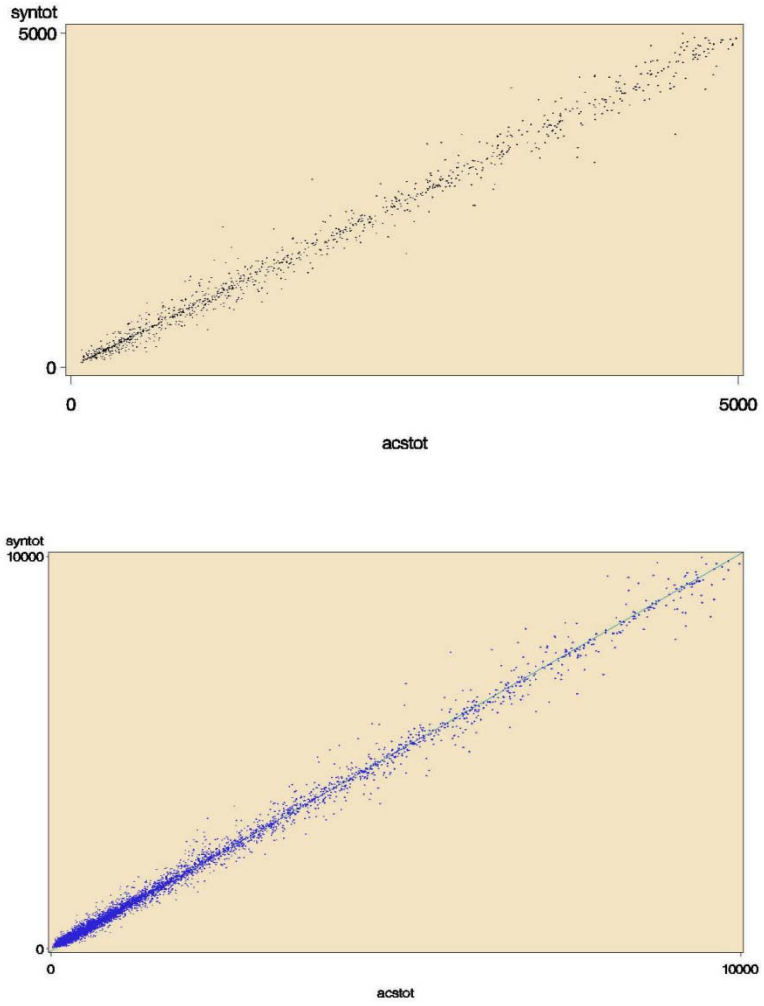


Figure P-6. Plot of ACS Partially Perturbed Weighted Counts for Atlanta's County Flows: Top: Development Phase, Bottom: Validation Phase (Different Scale)

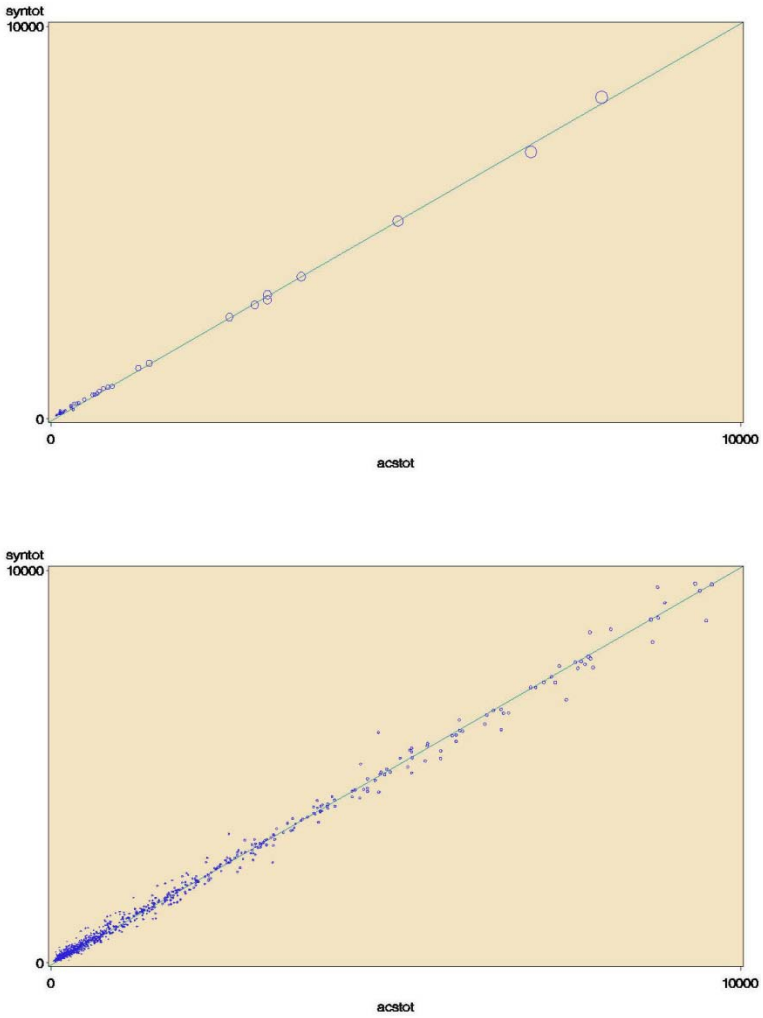


Figure P-7. Plot of ACS Partially Perturbed Weighted Counts for County Flow Tabulations Involving Industry: Top: Validation Phase – Olympia, Bottom: Validation Phase – Atlanta

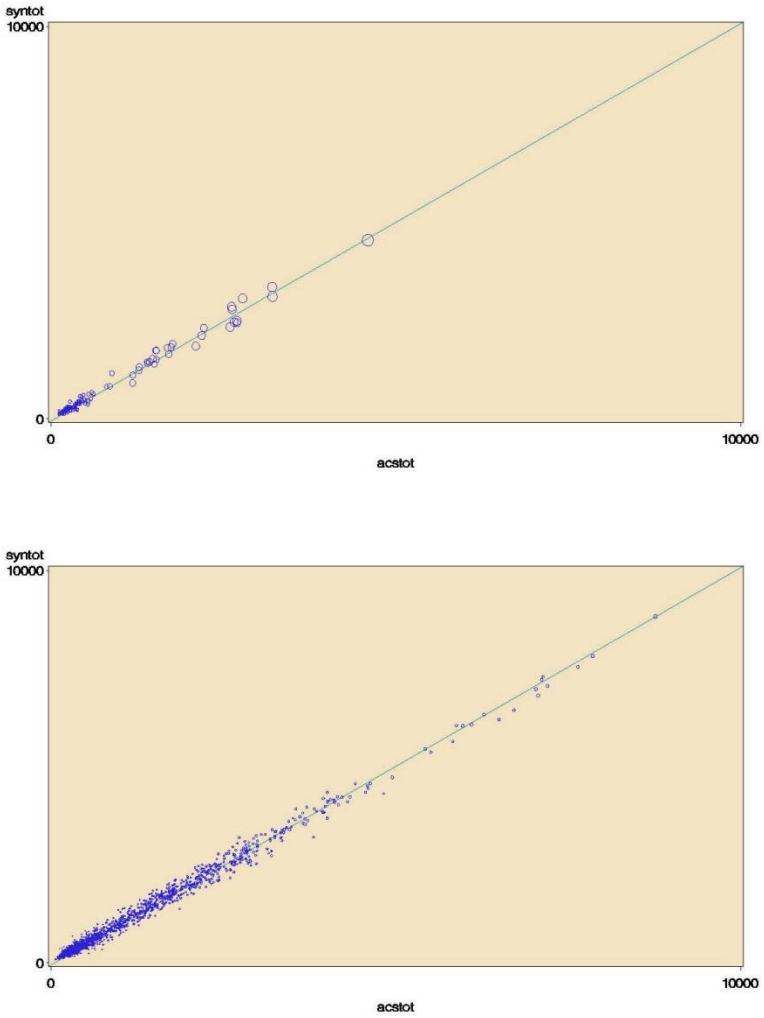


Figure P-8. Plot of ACS Partially Perturbed Weighted Counts for CTAZ1000 Tabulations Involving MOT11 by AHINC26: Top: Validation Phase – Olympia, Bottom: Validation Phase – Atlanta

Appendix Q

VALIDATION PHASE: PLOTS OF ACS AND PERTURBED DATA FOR PAIRWISE CORRELATIONS

Figure Q-1. Plots of ACS and Partial Perturbed Pairwise Correlations for Madison's and Olympia's PUMAs: Top: Development Phase -- Madison, Bottom: Validation Phase -- Olympia

Figure Q-2. Plots of ACS and Partial Perturbed Pairwise Correlations for Atlanta's PUMAs: Top: Top: Development Phase, Bottom: Validation Phase

Note: One point represents one PUMA. ACS values are on the horizontal axis. Perturbed values are on the vertical axis. The line has a slope of 1.

The scatterplots for the development phase were generated for 11 select pairwise correlations computed for each PUMA. The 11 pairs were:

- Travel time with each of the following: time leaving home, HH income, derived flow distance, poverty status, and age
- Time leaving home with: HH income, poverty status, and age
- HH income with: age, and poverty status
- Poverty status with age

Scatterplots for the validation phase were generated for 11 select pairwise correlations computed for each PUMA. The 11 pairs were:

- Travel time with each of the following: time leaving home and age
- Time leaving home with age
- HH income with each of the following: age, and poverty status, number of workers in HH, number of vehicles in HH
- Poverty status with each of the following age, number of vehicles in HH
- Number of workers in HH with number of vehicles in HH, and with Age

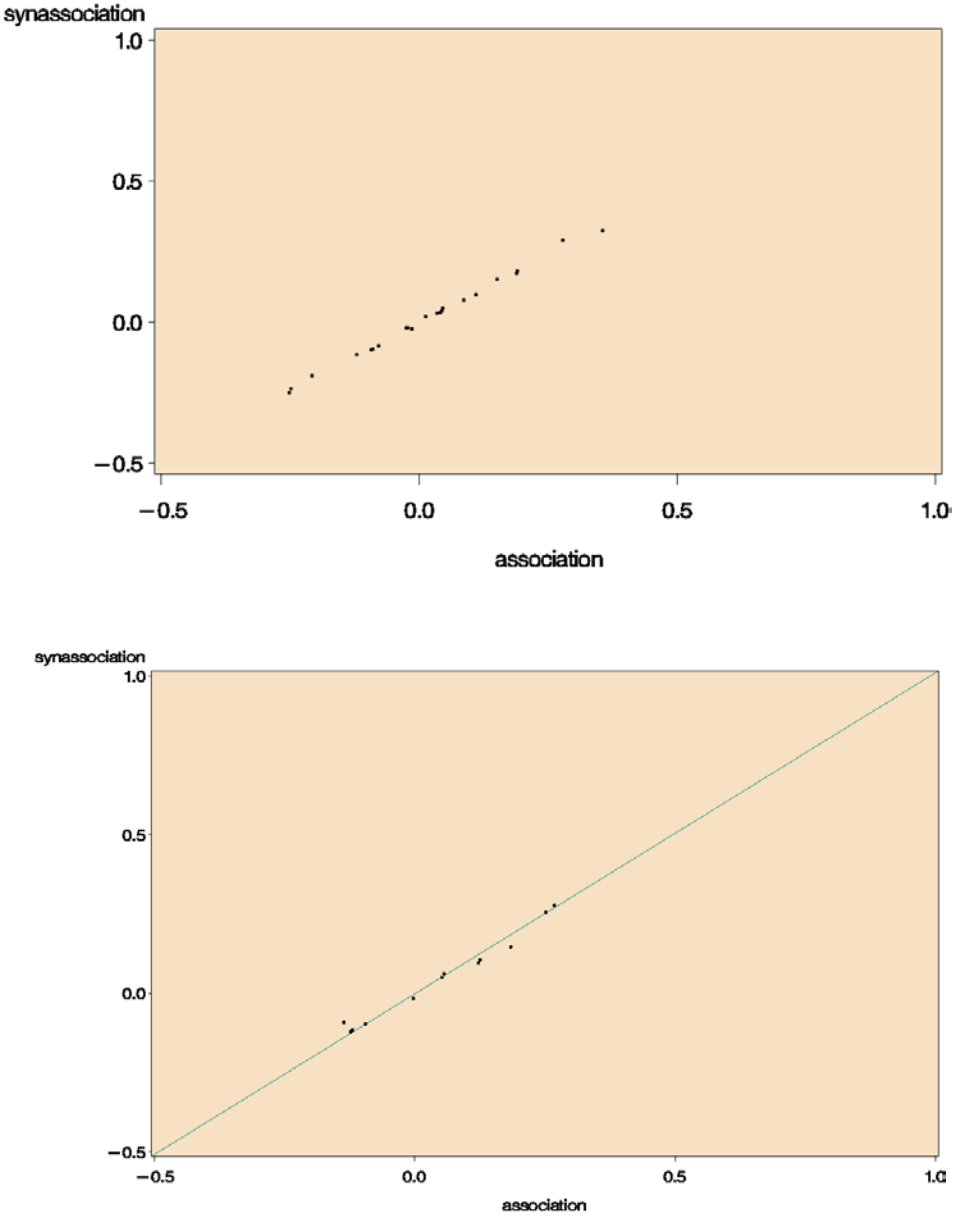


Figure Q-1. Plots of ACS and Partial Perturbed Pairwise Correlations for Madison's and Olympia's PUMAs: Top: Development Phase -- Madison, Bottom: Validation Phase -- Olympia

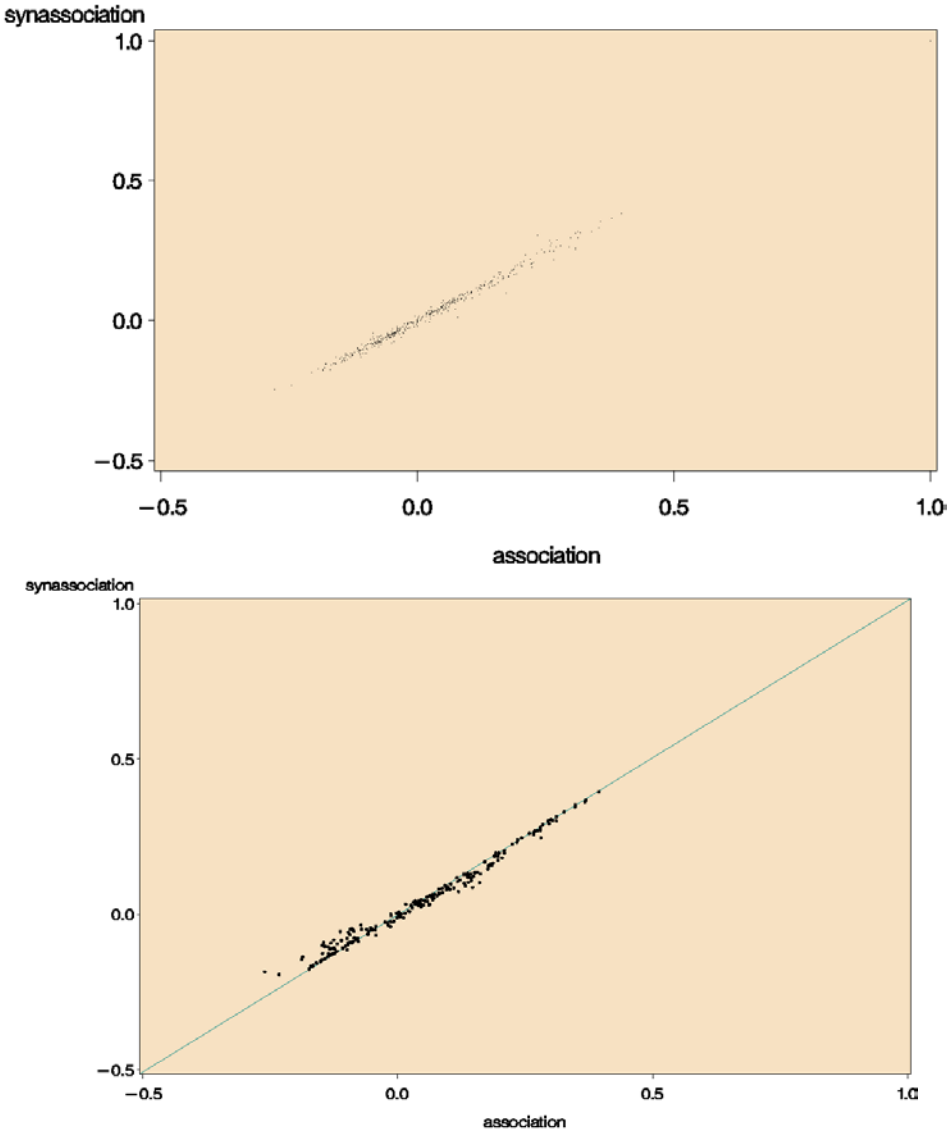


Figure Q-2. Plots of ACS and Partial Perturbed Pairwise Correlations for Atlanta's PUMAs: Top: Top: Development Phase, Bottom: Validation Phase

Appendix R

VALIDATION PHASE: TABLES FOR TRAVEL MODEL OUTPUT COMPARISONS

This appendix presents the results of the comparison tests for the two Validation Phase test sites: Atlanta, GA; and Olympia, WA. Table 3-15 in the main body of the report shows which comparison tests were made for which test site and at which level of geography.

Table R-1. Results for Test 1 (Estimates Population by Age Category), Atlanta, TAZ Level

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	700	201%	718	203%	151	42%
Median	70	33%	77	36%	0.1	0.1%

Matrix Size: 7,163 estimates
ACS Sample Size: 117,545 respondents

Table R-2. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), Olympia, County Level*

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	7,703	40%	7,703	40%	1	0.01%
Median	733	-6%	731	-6%	-2	-0.01%

Matrix Size: 5 estimates
ACS Sample Size: 6,439 respondents
*ACS categories collapsed to match MPO categories

Table R-3. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), Olympia, TAZ Level

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
Zero Worker	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	65	200%	65	201%	0	0%
Median	2	0.4%	2	0.4%	0	0.0%
One Worker	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	73	105%	76	111%	1	0.9%
Median	8	7%	8	9%	0	0.0%
Two Worker	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	82	95%	83	96%	1	0.7%
Median	-20	-25%	-21	-24%	0	0.0%
Total HH	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	153	80%	149	79%	1	0.3%
Median	-31	-13%	-28	-12%	0	0.0%

Matrix Size: 420 estimates
ACS Sample Size: 6,439 respondents

Table R-4. Results for Test 2 (Estimated Number of Households by Categorized Number of Workers), Atlanta, TAZ Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
All	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	766	180%	766	180%	0.91	0.33%
Median	-69	-21%	-68	-21%	0.03	0.01%
Zero Worker	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	831	13%	831	13%	0.33	0.16%
Median	-933	-81%	-933	-81%	0.03	0.01%
One-Worker	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	334	248%	340	250%	1.7	0.4%
Median	237	182%	238	183%	0.2	0.1%
Two Worker	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	268	59%	271	59%	1.8	0.5%
Median	-56	-20%	-56	-19%	-0.2	-0.1%
Three Plus Workers	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	410	11%	410	11%	0.3	0.20%
Median	-586	-82%	-586	-82%	0.1	0.04%

Matrix Size: All: 4,178 estimates; Zero-Worker: 1,181 estimates; One-Worker: 1,448 estimates; Two-Worker: 1,305 estimates; Three-Plus Workers: 244 estimates
ACS Sample Size: All: 96,636 respondents; Zero-Worker: 18,934 respondents; One-Worker: 43,869 respondents; Two-Worker: 31,708 respondents; Three-Plus Workers: 2,125 respondents

Table R-5. Results for Test 3 (Estimated Number of Home-to-Work Person Trips), Olympia, TAZ Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	410	1,596%	409	1,575%	0.5	0.8%
Median	-301	-468%	-300	-466%	0.1	0.1%

Matrix Size: 18 estimates
ACS Sample Size: 284 respondents

Table R-6. Results for Test 4(Estimated Average Home-to-Work Travel Time), Olympia, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
All Modes	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**	**AbsDif**	**PctDif**
IQR	156	59%	157	63%	1	0.00%
Median	-75.5	-59%	-75	-57%	0.5	0.00%

Matrix Size: 4 estimates
ACS Sample Size: 4,963 respondents

Table R-7. Results for Test 4 (Estimated Average Home-to-Work Travel Time), Olympia, TAZ Level*

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	**PctDif**	**AbsDif (000s)**	**PctDif**	**AbsDif**	**PctDif**
IQR						
Median						

*Could not be meaningfully computed due to sparseness of data

Table R-8. Results for Test 4 (Estimated Average Home-to-Work Travel Time), Atlanta, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	16	24%	17	25%	2	5%
Median	-10	-21%	-10	-21%	0	0%

Matrix Size: 301 estimates
ACS Sample Size: Drive Alone: 96,890 respondents ; Carpool: 12,061 respondents; Public Transportation: 3,852 respondents; Bike/Walk: 1,570 respondents

Table R-9. Results for Test 4 (Estimated Average Home-to-Work Travel Time), Atlanta, District Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
All Modes	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	10	36%	11	38%	0.2	1%
Median	-4	-19%	-4	-18%	0.0	0%

Matrix Size: 118 estimates
ACS Sample Size: Drive Alone: 3,403 respondents; Carpool: 23 respondents; Bike/Walk: 103 respondents

Table R-10. Results for Test 5 (Estimated Person Trips by Household Income Category by Means of Transportation [7]), Olympia, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
Total, MOT, all income categories*	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	21,383	35%	21,152	35%	235	1%
Median	-24,441	-36%	-24,526	-37%	-78	-0.2%

Matrix Size: 5 estimates
ACS Sample Size: 5,320 respondents
*The MPO could not provide trips by income by travel mode but provided trip productions (all travel modes) by income. ACS income categories were collapsed to match the MPO's income categories.

Table R-11. Results for Test 5 (Estimated Person Trips by Household Income Category by Means of Transportation [7]), Olympia, TAZ Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
Total, MOT, all income categories*	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	547	96%	963	10%	26	0.21%
Median	-124	-85%	-183	-97%	0.00	0.00%

Matrix Size: 33 estimates
ACS Sample Size: 746 respondents
*The MPO could not provide trips by income by travel mode but provided trip productions (all travel modes) by income. ACS income categories were collapsed to match the MPO's income categories.

Table R-12. Results for Test 5 (Estimated Person Trips by Household Income Category by Means of Transportation [7]), Atlanta, County Level

Raw ACS Minus Model			Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	1,763	71%	1,596	69%	524	240%
Median	-411	-52%	-381	-52%	-0.03	-0.01%

Matrix Size: 752 estimates
ACS Sample Size: 73,670 respondents

Table R-13. Results for Test 5 (Estimated Person Trips by Household Income Category by Means of Transportation [7]), Atlanta, District Level

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	2,446	22%	2,485	19%	20	26%
Median	-1,212	-94%	-1,212	-92%	-0.3	-0.4%

Matrix Size: 127 estimates
ACS Sample Size: 2,647 respondents

Table R-14. Results for Test 6 (Estimated Person Trips by Age of Worker Category by Means of Transportation [7]), Atlanta, County Level

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	1,866	58%	1,798	63%	106	11%
Median	-321	-37%	-311	-37%	0.04	0.05%

Matrix Size: 283 estimates
ACS Sample Size: 8,250 respondents

Table R-15. Results for Test 7 (Estimated Person Trips by Age of Worker Category by Means of Transportation [4]), Atlanta, County Level

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	1,085	98%	1,133	97%	74	10%
Median	-96	-26%	-76	-27%	0.3	0.3%

Matrix Size: 221 estimates
ACS Sample Size: 9,253 respondents

Table R-16. Test 7 (Estimated Person Trips by Age of Worker Category by Means of Transportation [4]), Atlanta, TAZ Level

	Raw ACS Minus Model		Perturbed ACS Minus Model		Perturbed ACS Minus Raw ACS	
	AbsDif	PctDif	AbsDif	PctDif	AbsDif	PctDif
IQR	86	643%	86	645%	1	1%
Median	65	193%	64	190%	0.1	0.05%

Matrix Size: 290 estimates
ACS Sample Size: 5,588 respondents

Appendix S

VALIDATION PHASE: FIGURES FOR TRAVEL MODEL OUTPUT COMPARISONS

The scatterplots are subset to flows with more than five sample cases due to disclosure concerns. With the perturbation targeted to tables and flows with a small number of cases, and most, if not all, such flows removed from the scatterplot, and the scatterplot may show perfect or near-perfect correlation between the raw and perturbed data for a very small subset of all total flows.

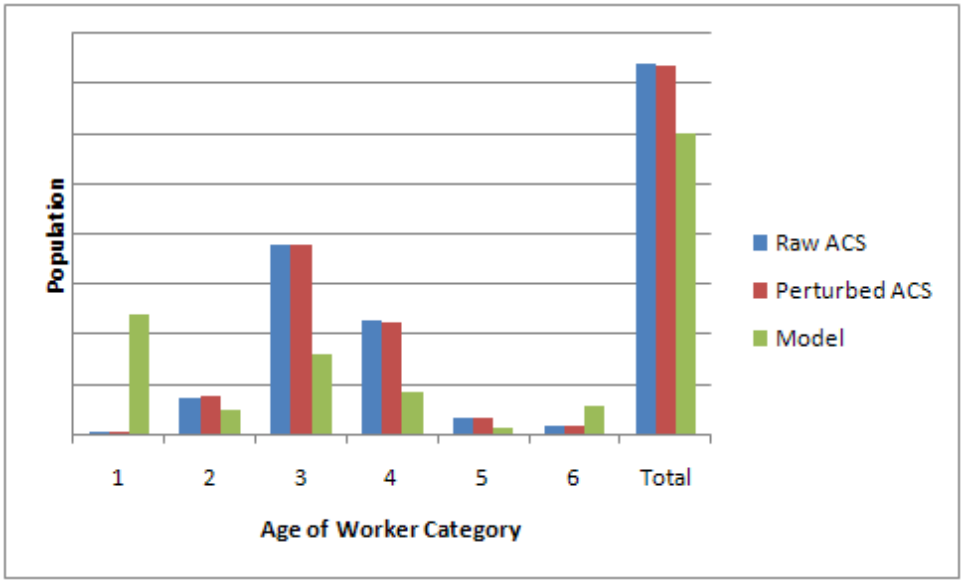


Figure S-1. Test 1: Population by Age of Worker, Atlanta

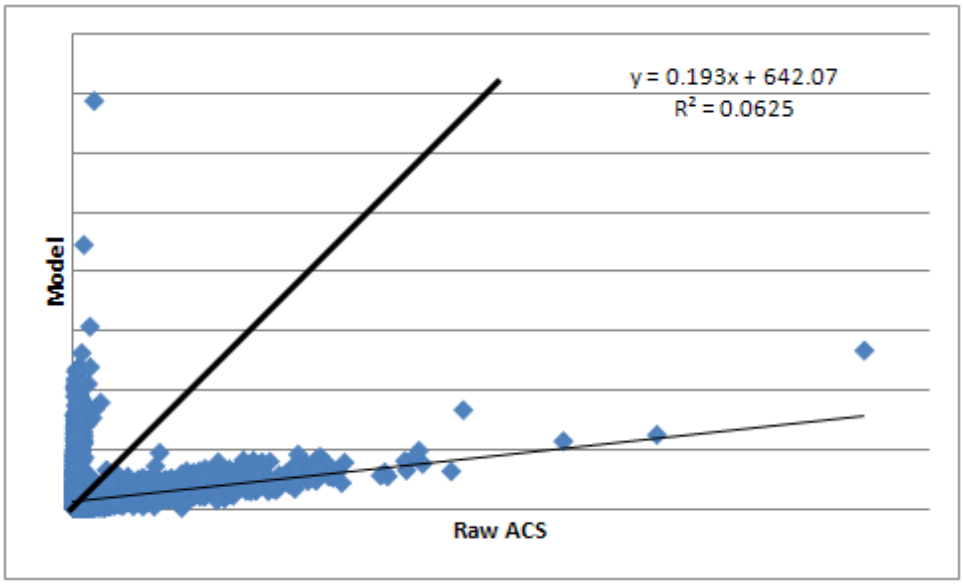


Figure S-2. Test 1: Scatterplot of Raw ACS vs. Model, All Age of Worker Categories, Atlanta, TAZ Level

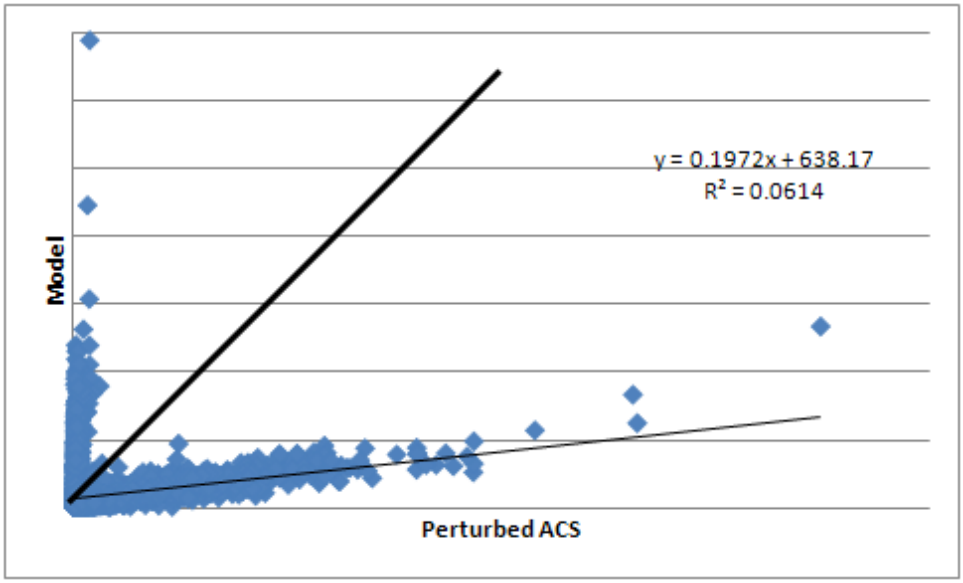


Figure S-3. Test 1: Scatterplot of Perturbed ACS vs. Model, All Age of Worker Categories, Atlanta, TAZ Level

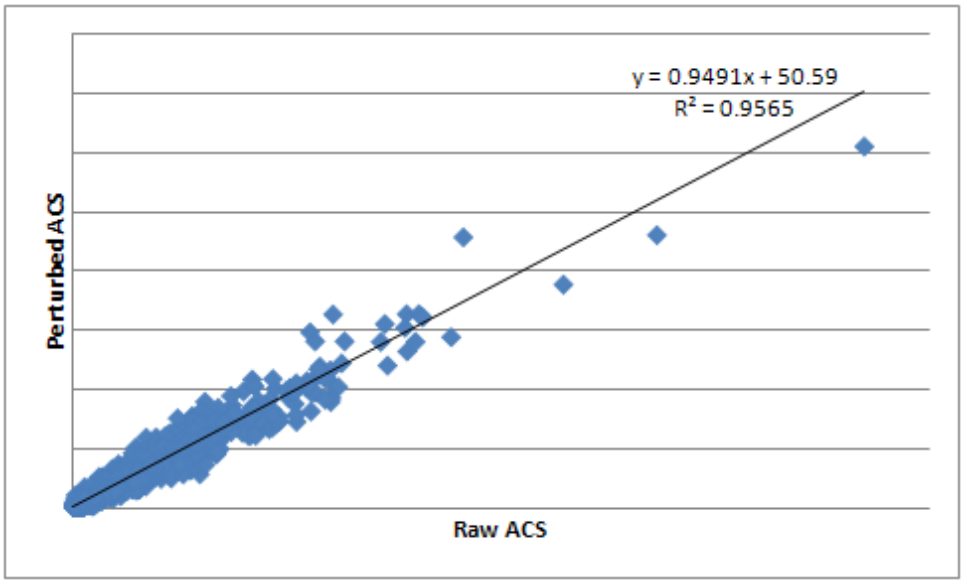


Figure S-4. Test 1: Scatterplot of Raw ACS vs. Perturbed ACS, All Age of Worker Categories, Atlanta, TAZ Level

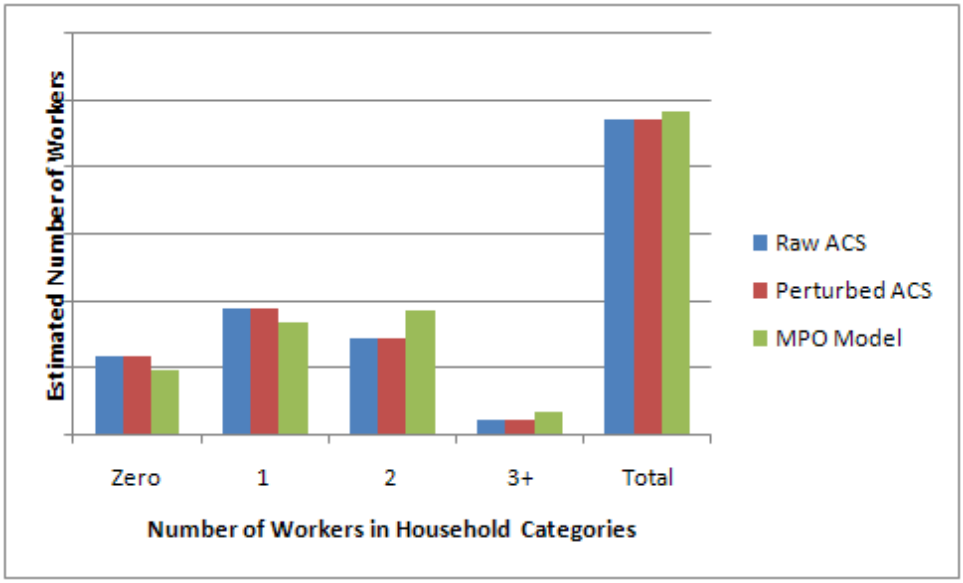


Figure S-5. Test 2: Distribution of Households by Number of Workers, Olympia, County Level

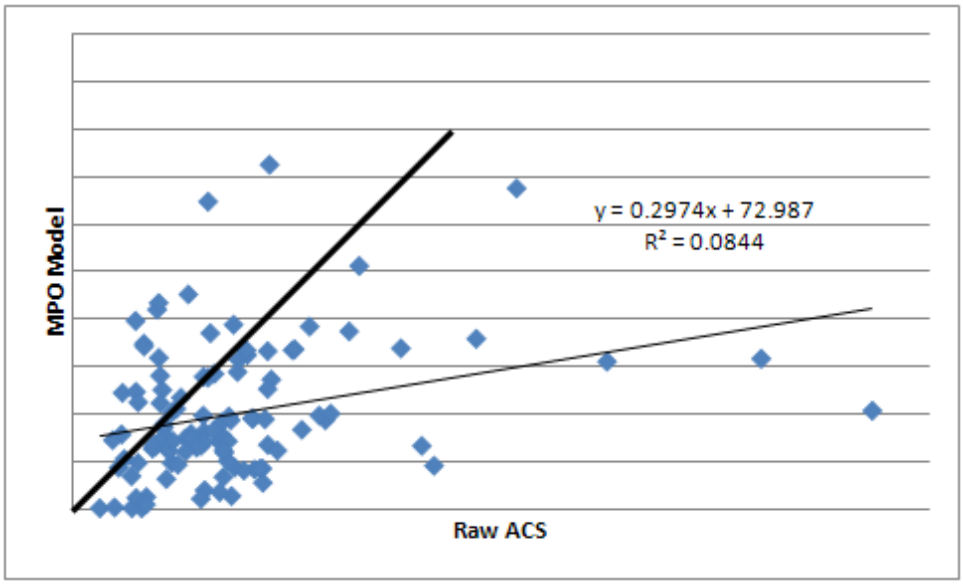


Figure S-6. Test 2: Scatterplot of Model vs. Raw ACS, Olympia, TAZ Level, Two-Worker Households

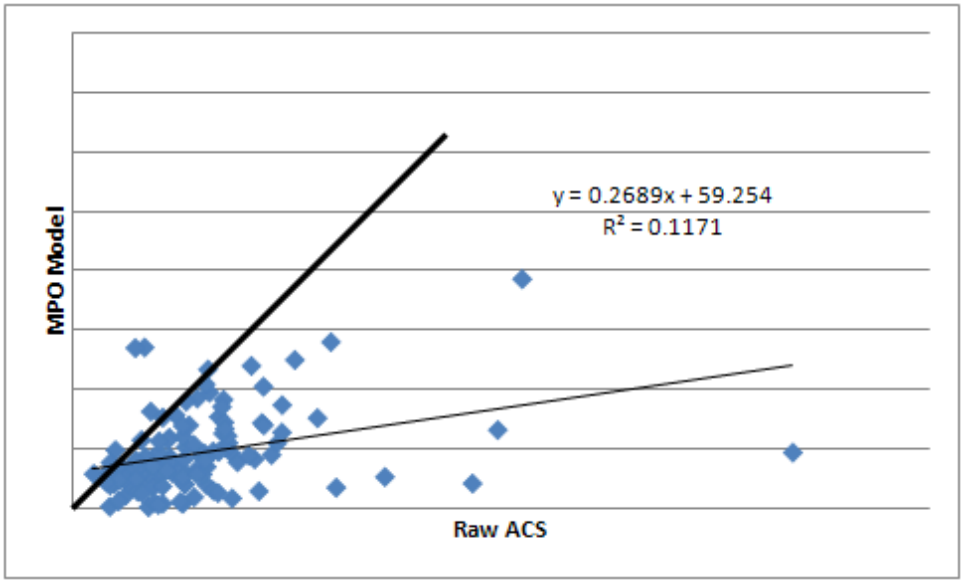


Figure S-7. Test 2: Scatterplot of Model vs. Raw ACS, Olympia, TAZ Level, One-Worker Households

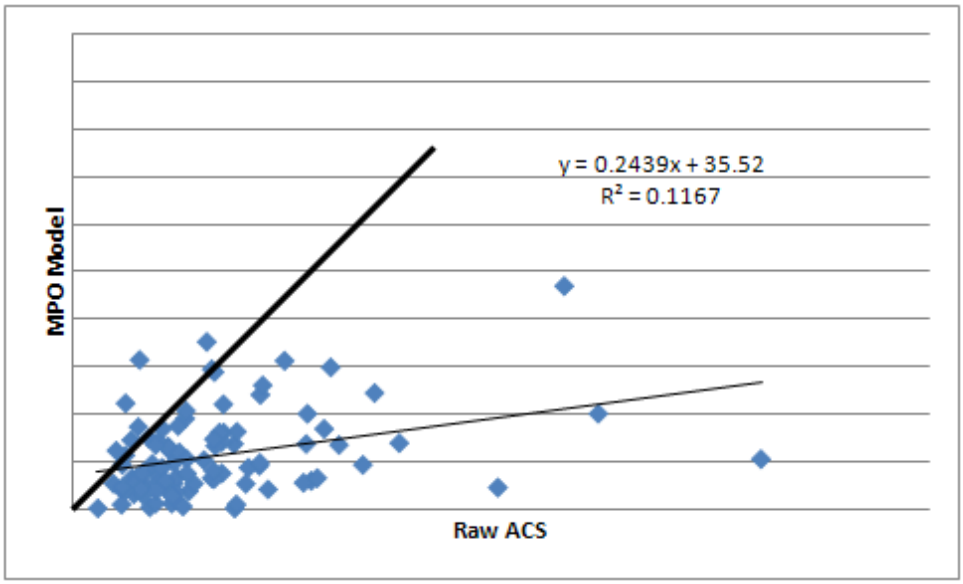


Figure S-8. Test 2: Scatterplot of Model vs. Raw ACS, Olympia, TAZ level, Zero-Worker Households

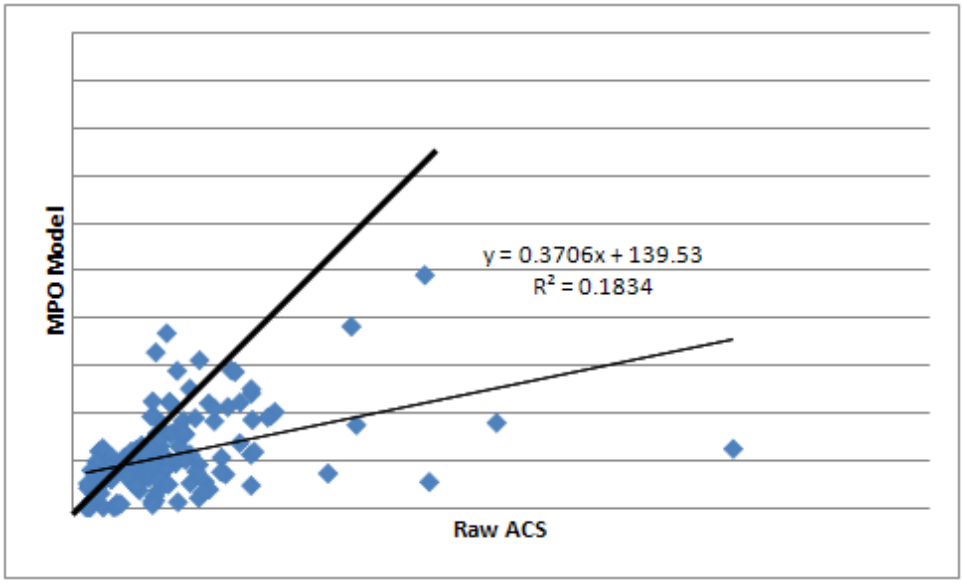


Figure S-9. Test 2: Scatterplot of Model vs. Raw ACS, Olympia, TAZ Level, Total Households

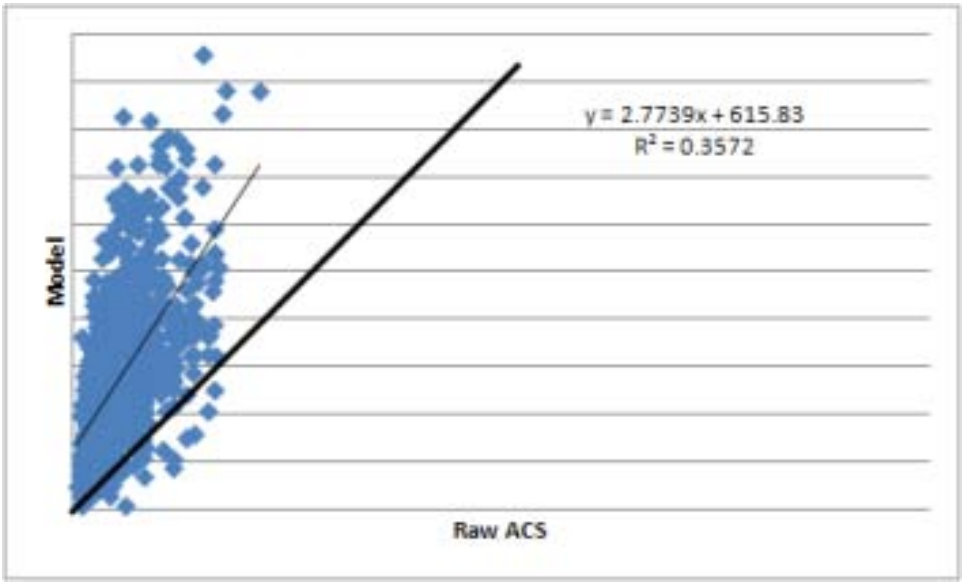


Figure S-10. Test 2: Scatterplot of Raw ACS vs. Model, Atlanta, TAZ Level, Zero-Worker Households

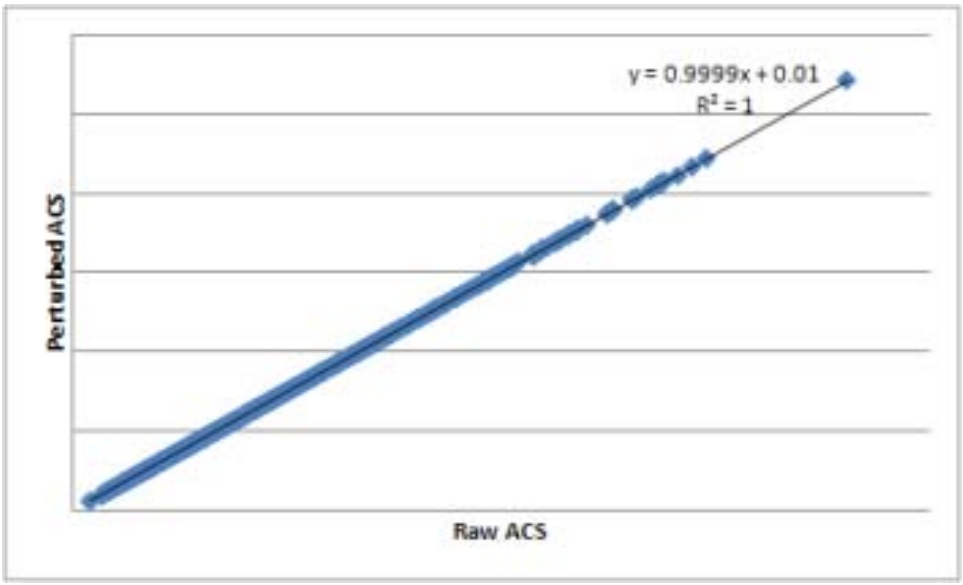


Figure S-11. Test 2: Scatterplot of Perturbed ACS vs. Model, Atlanta, TAZ Level, Zero-Worker Households

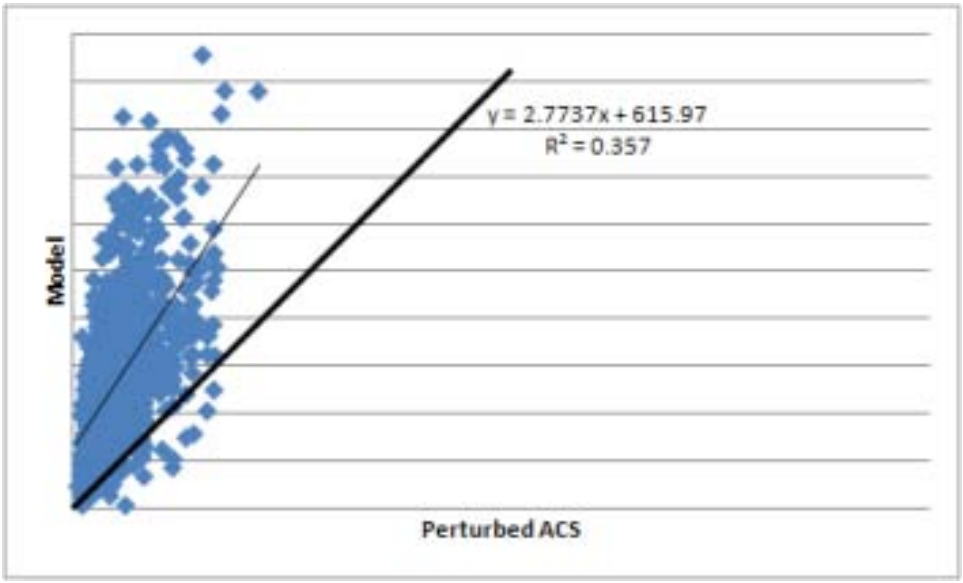


Figure S-12. Test 2: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, TAZ Level, Zero-Worker Households

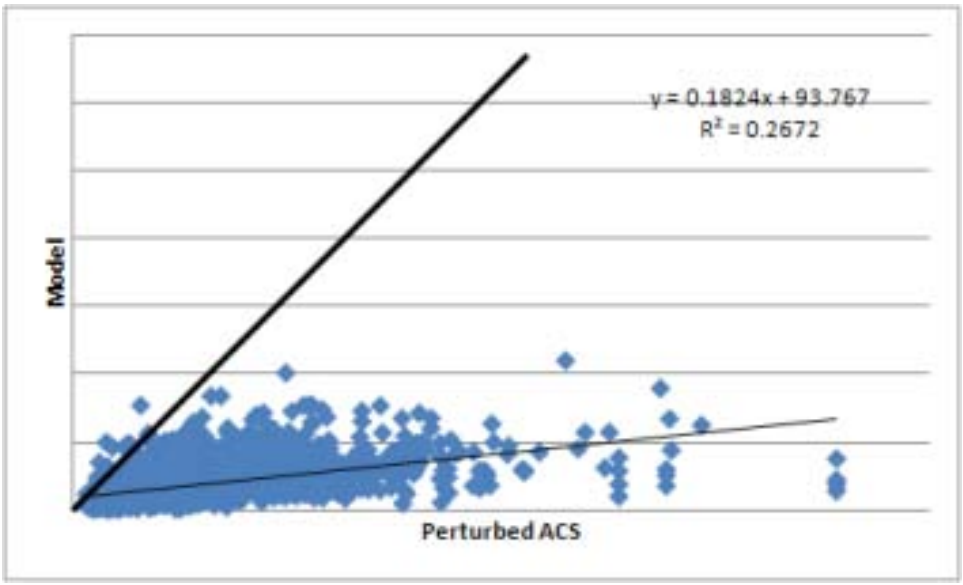


Figure S-13. Test 2: Scatterplot of Raw ACS vs. Model, Atlanta, TAZ Level, One-Worker Households

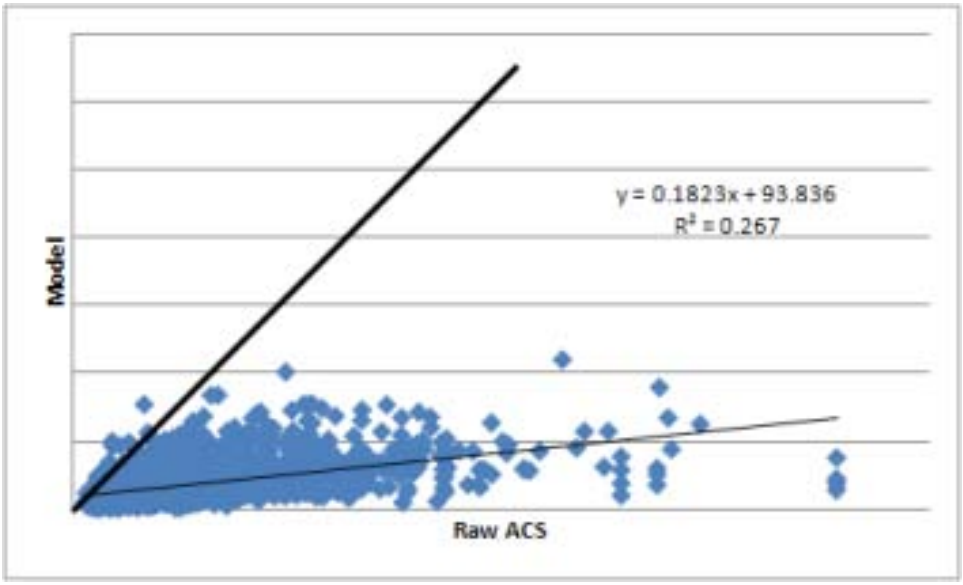


Figure S-14. Test 2: Scatterplot of Perturbed ACS vs. Model, Atlanta, TAZ Level, One-Worker Households

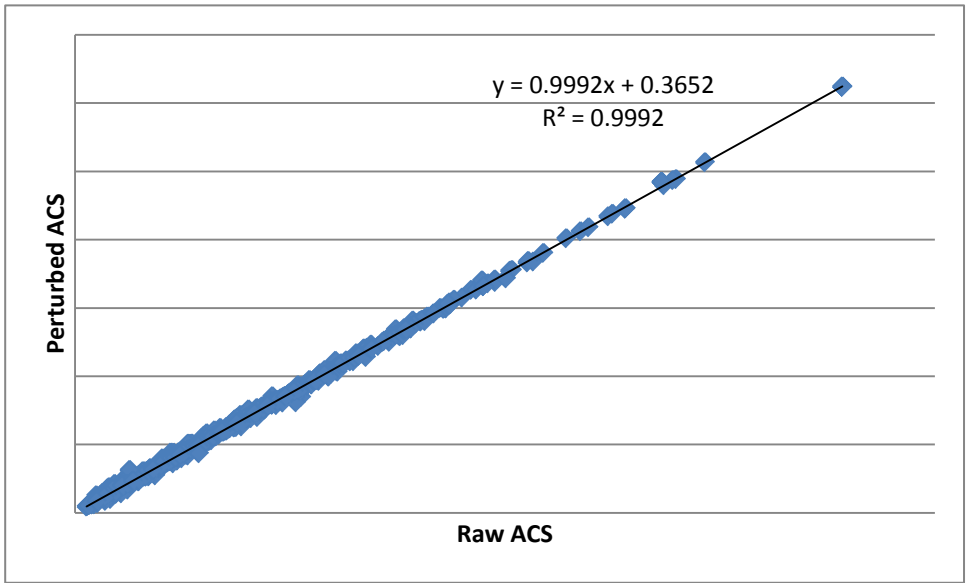


Figure S-15. Test 2: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, TAZ Level, One-Worker Households

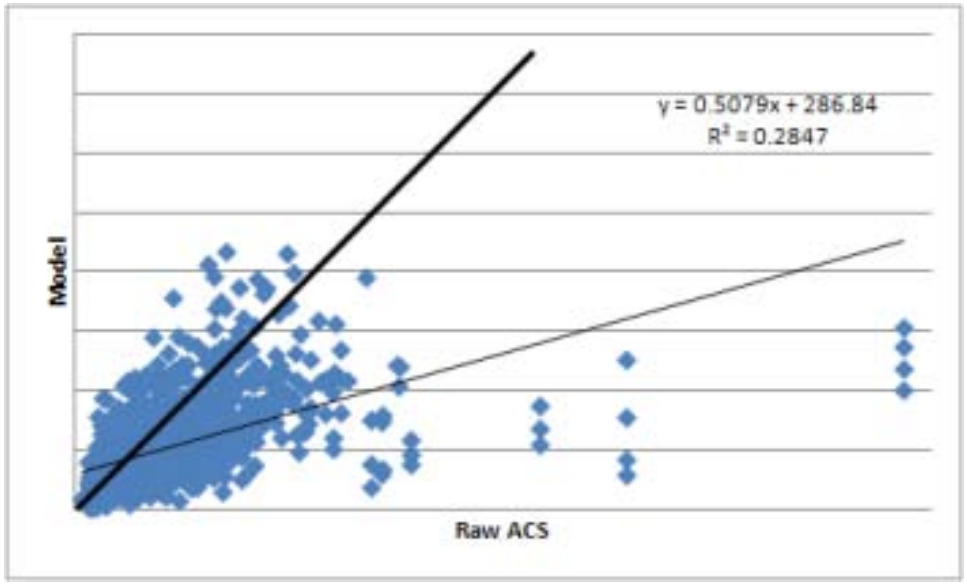


Figure S-16. Test 2: Scatterplot of Raw ACS vs. Model, Atlanta, TAZ Level, Two-Worker Households

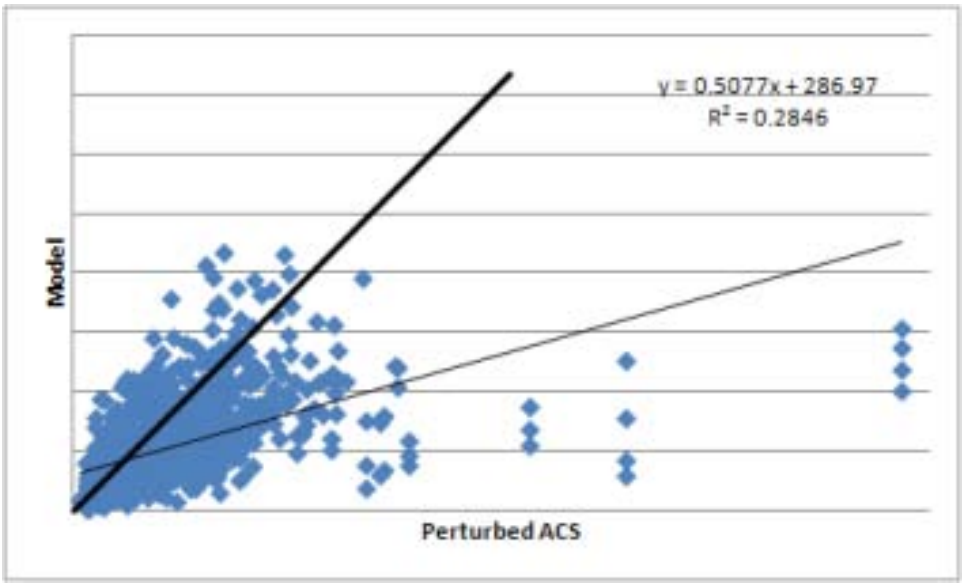


Figure S-17. Test 2: Scatterplot of Perturbed ACS vs. Model, Atlanta, TAZ Level, Two-Worker Households

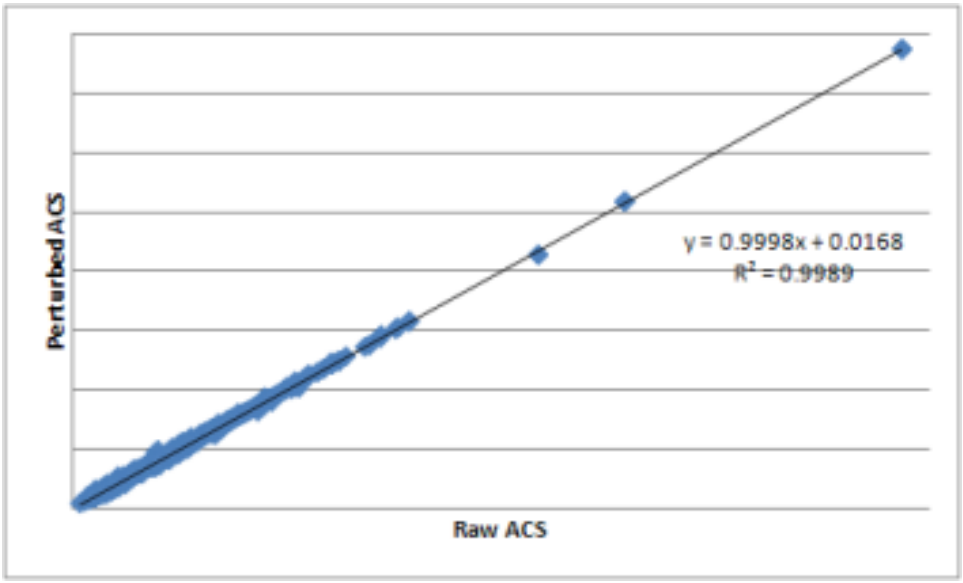


Figure S-18. Test 2: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, TAZ Level, Two-Worker Households

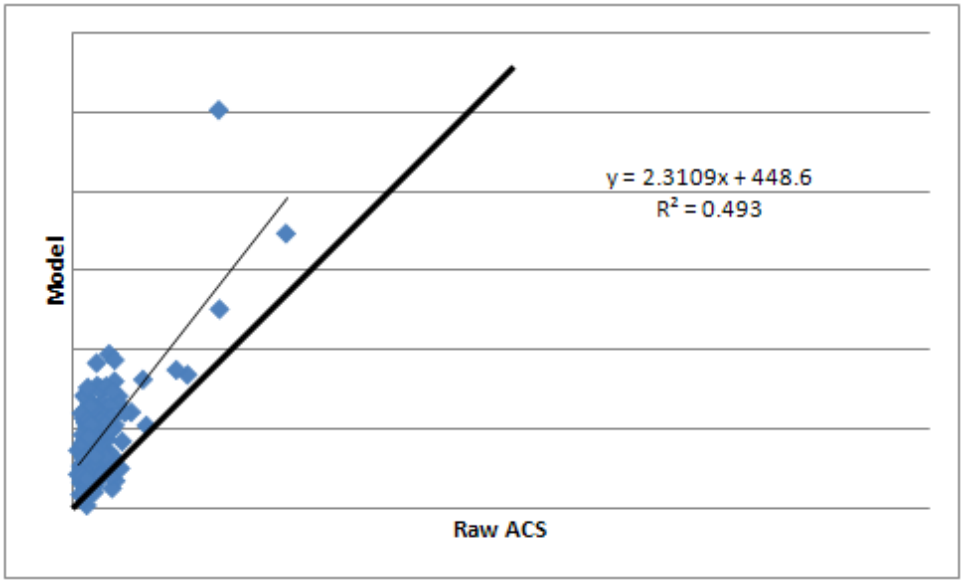


Figure S-19. Test 2: Scatterplot of Raw ACS vs. Model, Atlanta, TAZ Level, Three-Plus Worker Households

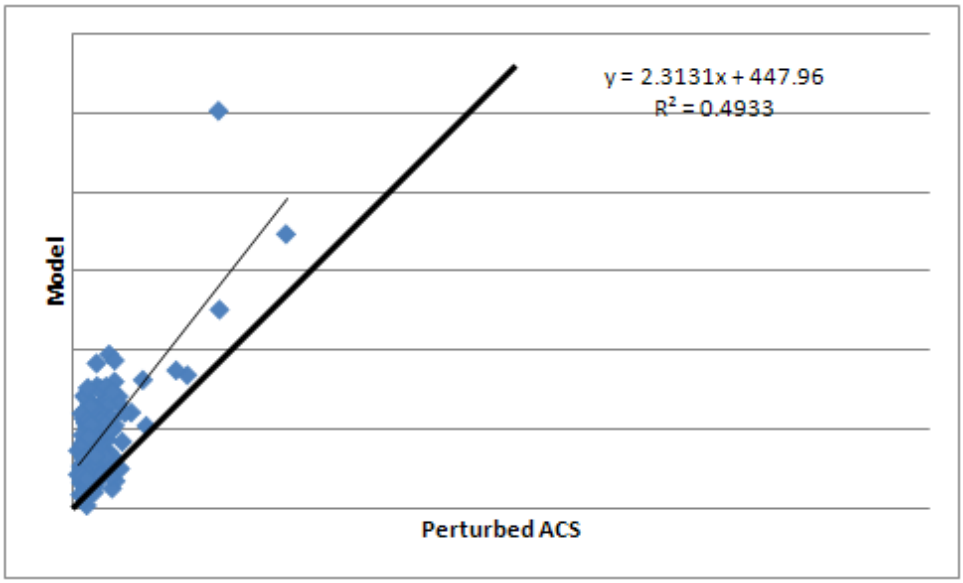


Figure S-20. Test 2: Scatterplot of Perturbed ACS vs. Model, Atlanta, TAZ Level, Three-Plus Worker Households

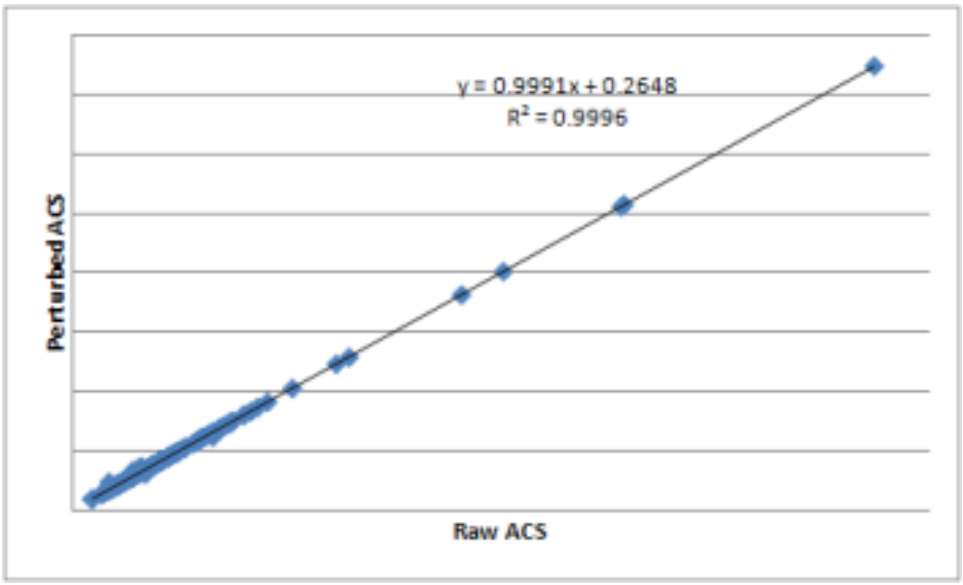


Figure S-21. Test 2: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, TAZ Level, Three-Plus Worker Households

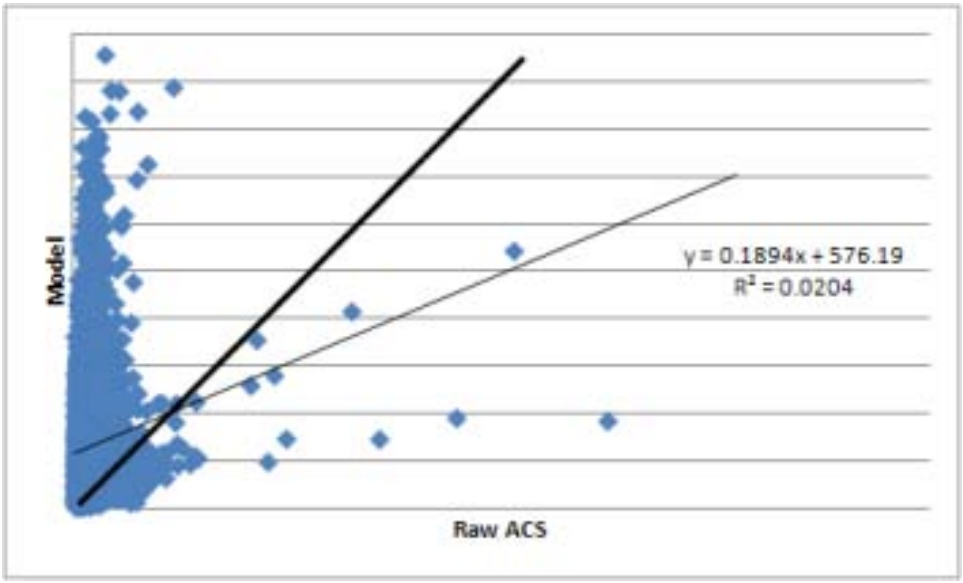


Figure S-22. Test 2: Scatterplot of Raw ACS vs. Model, Atlanta, TAZ Level, All Number of Worker Categories

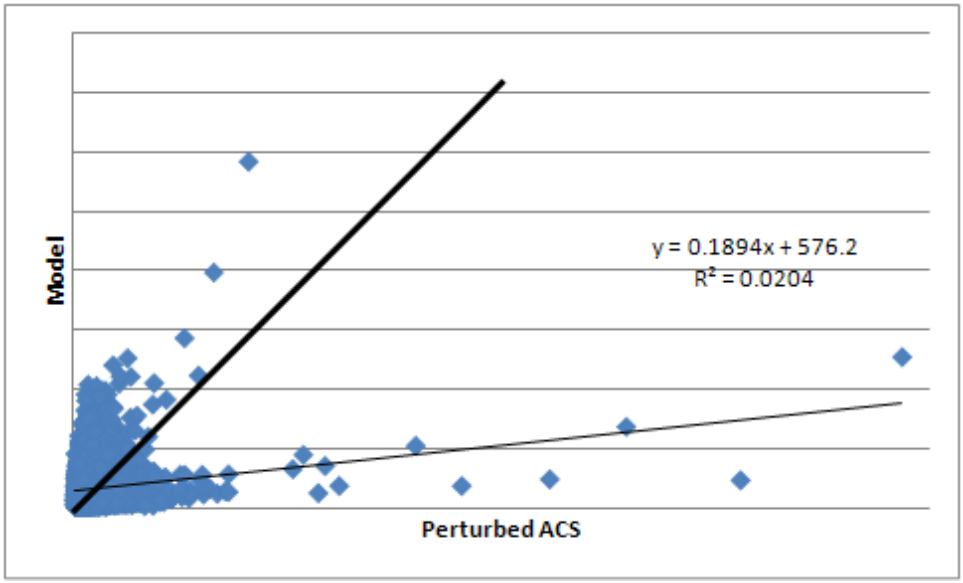


Figure S-23. Test 2: Scatterplot of Perturbed ACS vs. Model, Atlanta, TAZ Level, All Number of Workers Categories

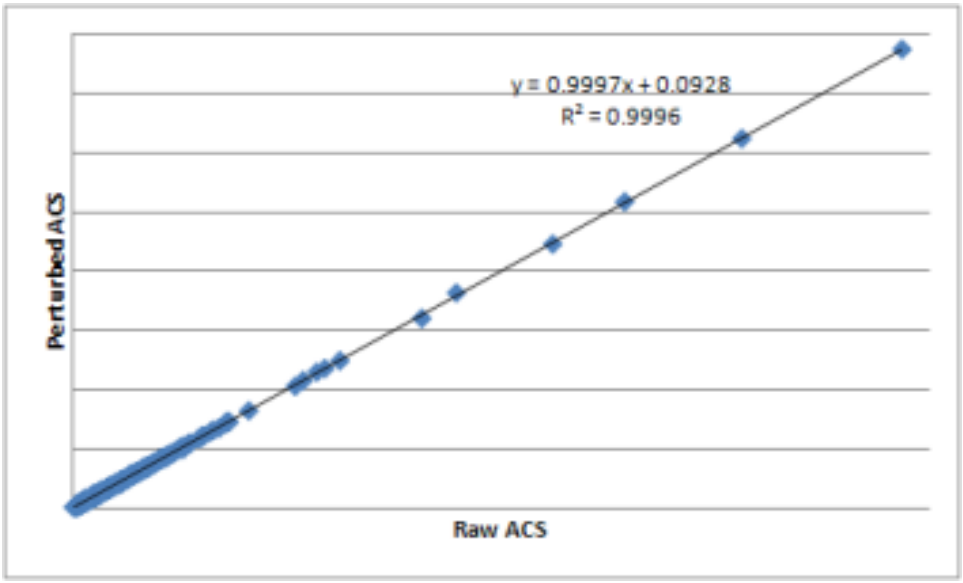


Figure S-24. Test 2: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, TAZ Level, All Number of Workers Categories

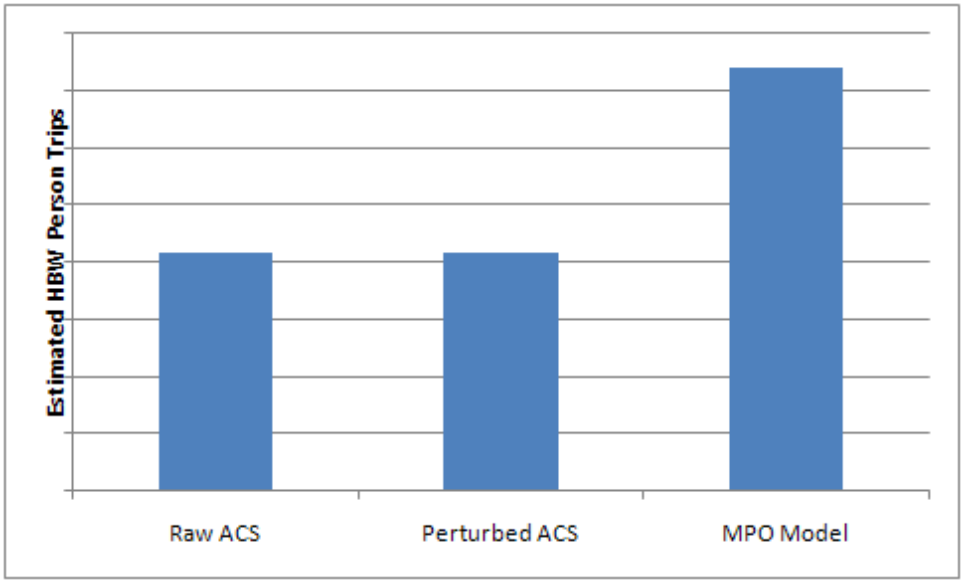


Figure S-25. Test 3: Total Estimated HBW Person Trips, Olympia, County Level

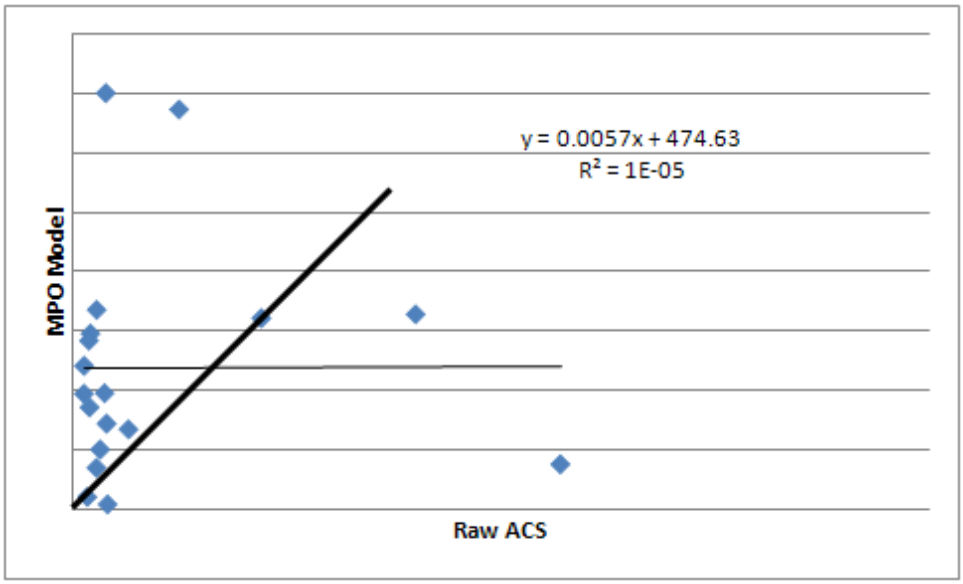


Figure S-26. Test 3: Scatterplot of Model vs. Raw ACS, Olympia, TAZ Level

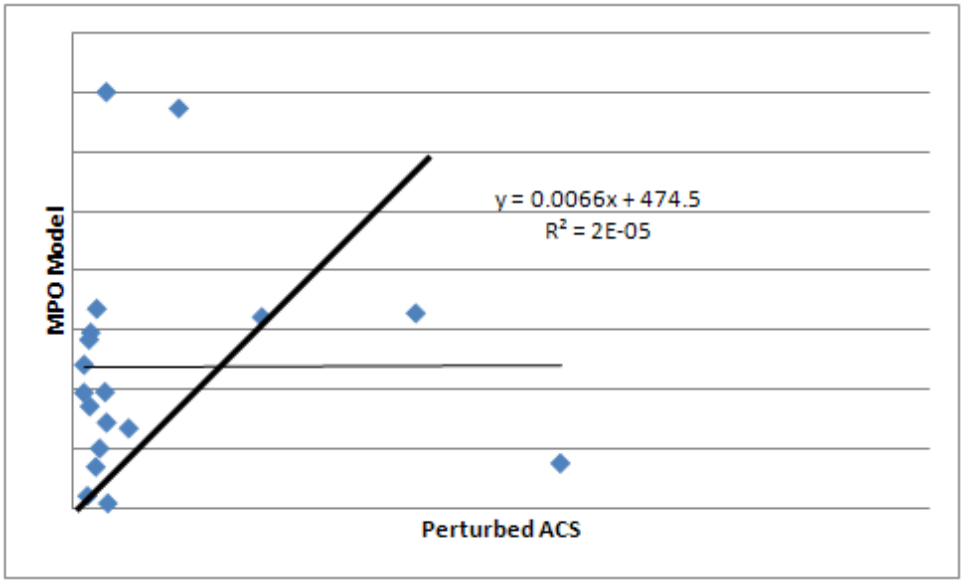


Figure S-27. Test 3: Scatterplot of Model vs. Perturbed ACS, Olympia, TAZ Level

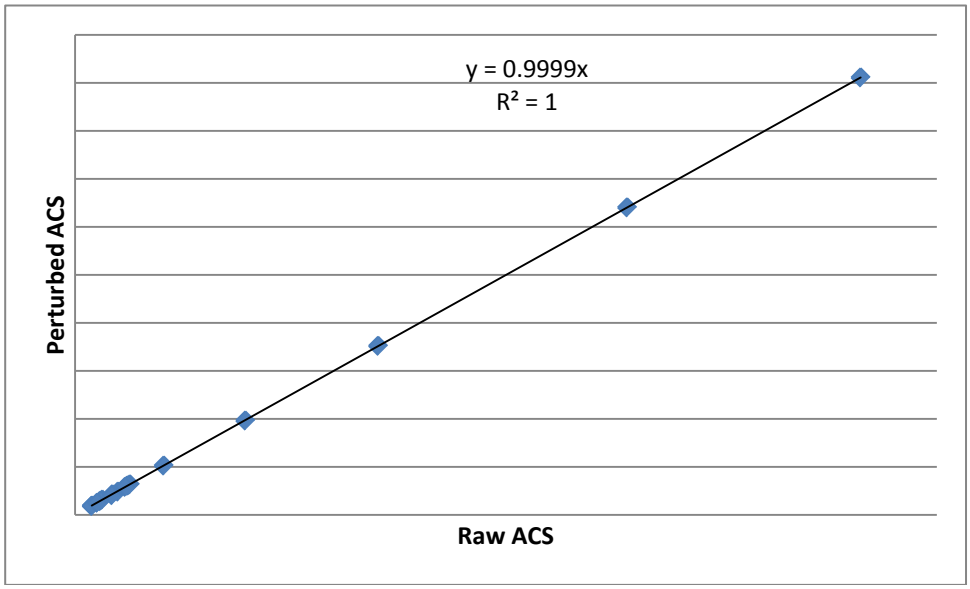


Figure S-28. Test 3: Scatterplot of Raw ACS vs. Perturbed ACS, Olympia, TAZ Level

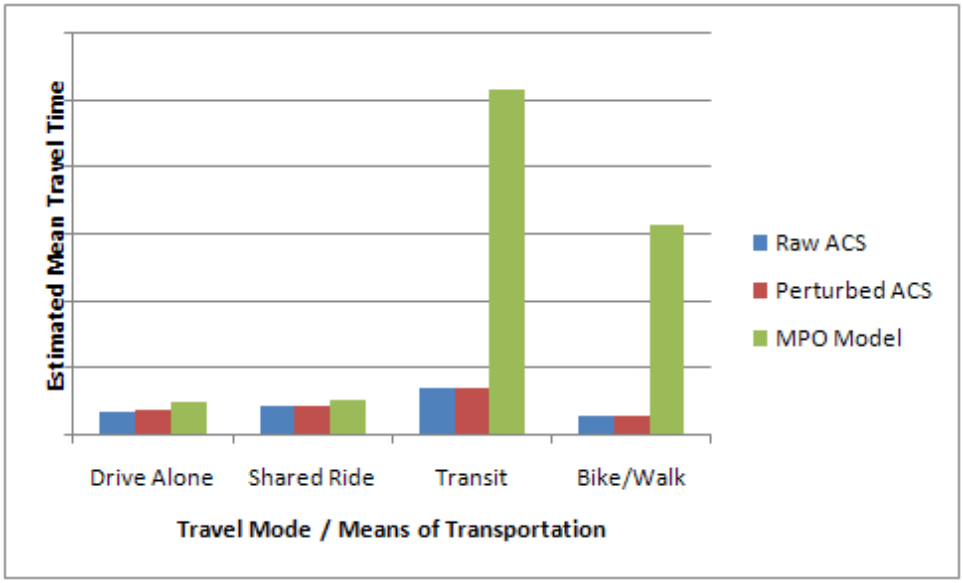


Figure S-29. Test 4: Estimated HBW Mean Travel Time by Mode / Means of Transportation, Olympia, County Level

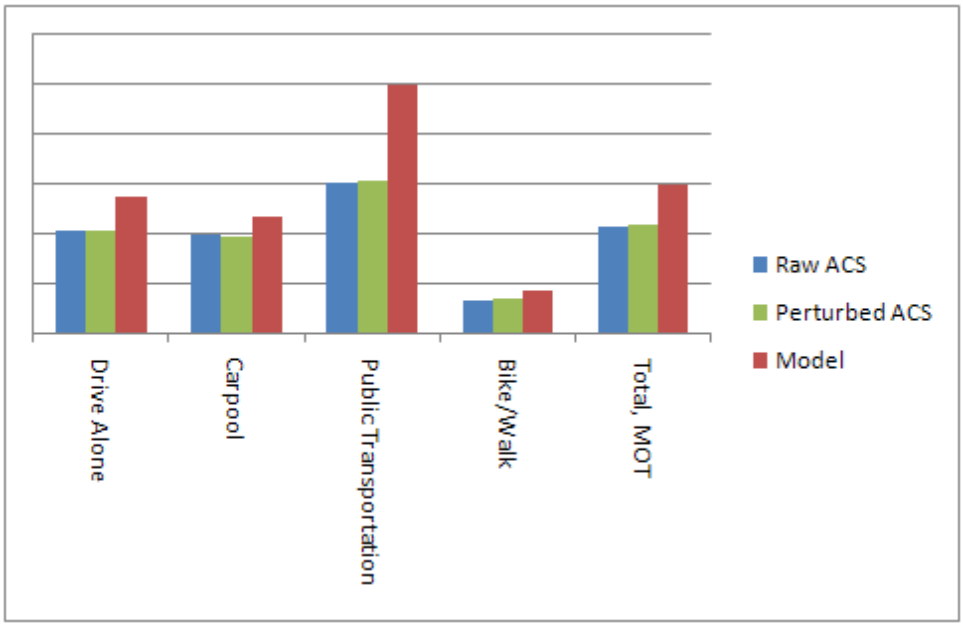


Figure S-30. Test 4: Average Travel Time by Means of Transportation, Atlanta, All Counties

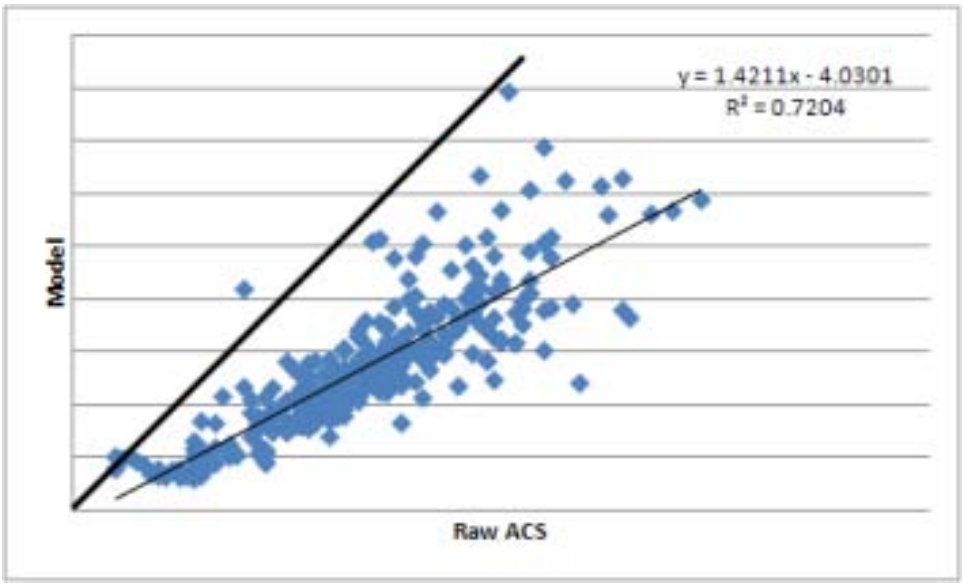


Figure S-31. Test 4: Scatterplot of Raw ACS vs. Model, Atlanta, County Level, All Modes

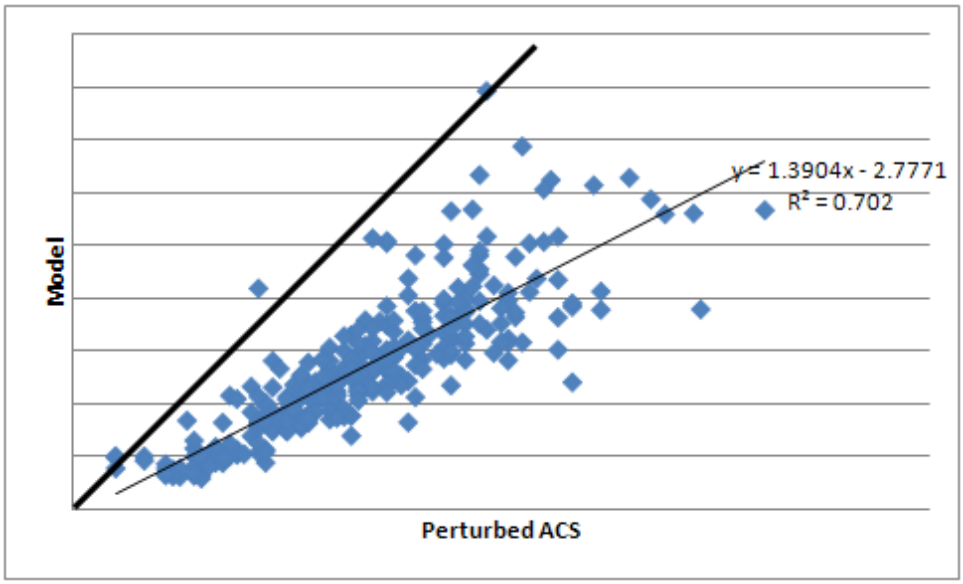


Figure S-32. Test 4: Scatterplot of Perturbed ACS vs. Model, Atlanta, County Level, All Modes

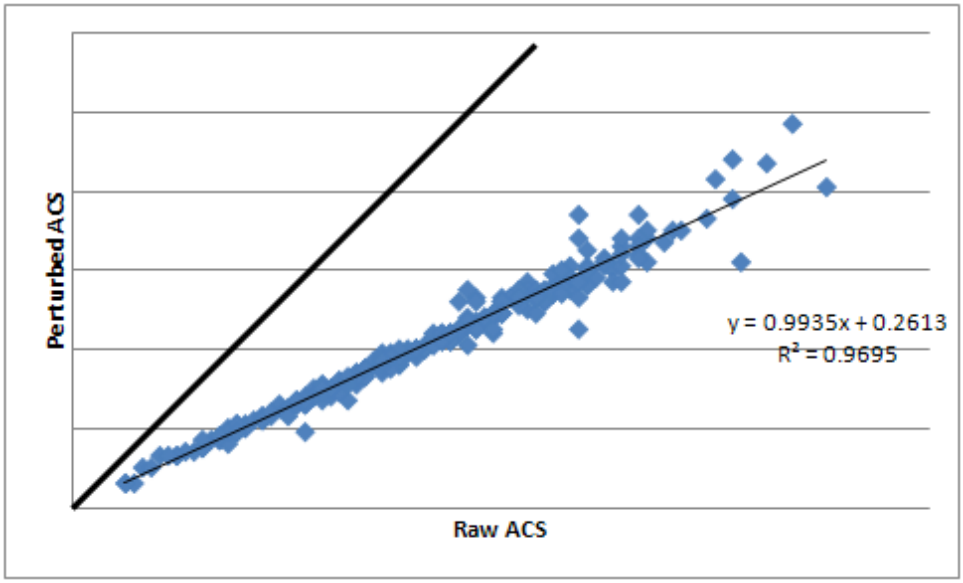


Figure S-33. Test 4 Scatterplot of Raw ACS vs Perturbed ACS, Atlanta, County Level, All Modes

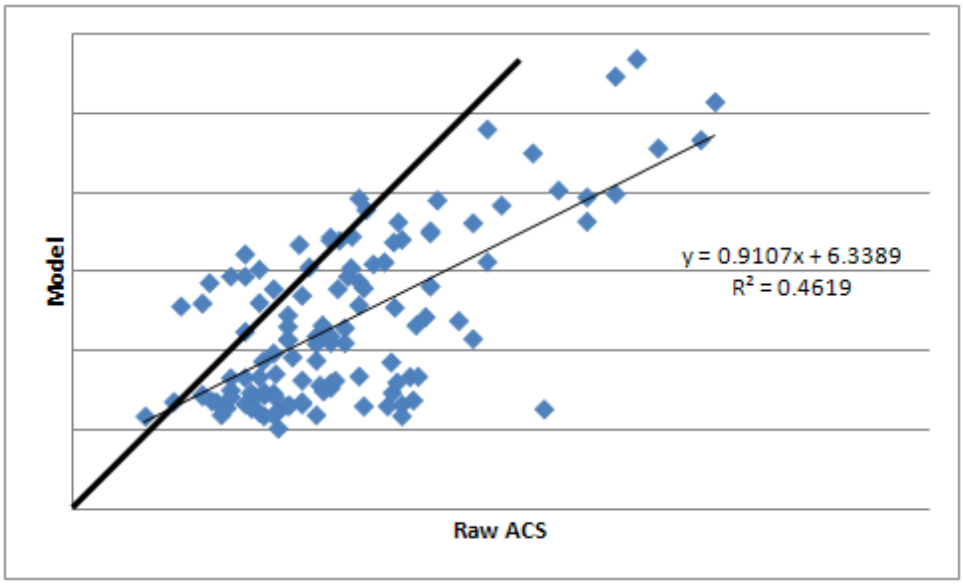


Figure S-34. Test 4: Scatterplot of Raw ACS vs. Model, Atlanta, District Level, All Modes

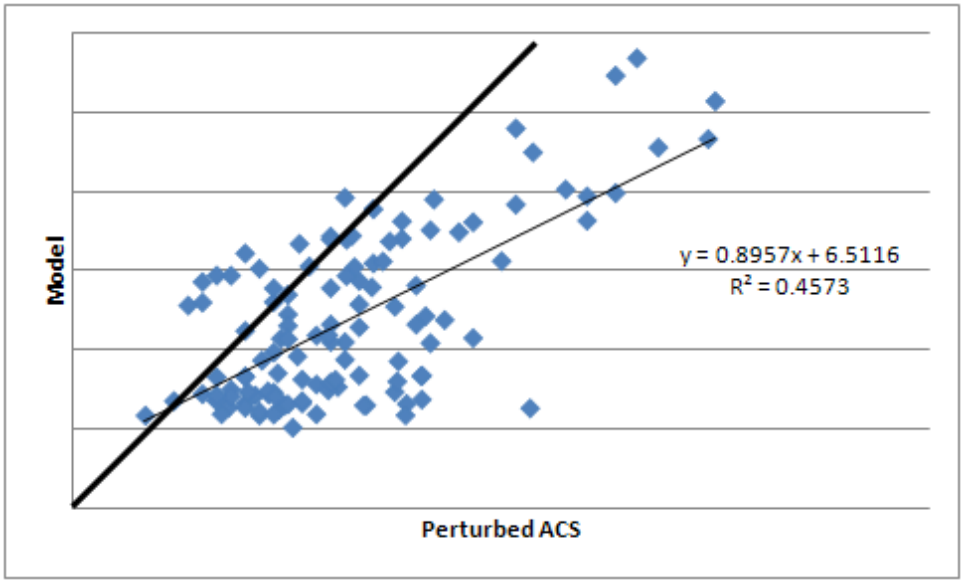


Figure S-35. Test 4: Scatterplot of Perturbed ACS vs. Model, Atlanta, District Level, All Modes

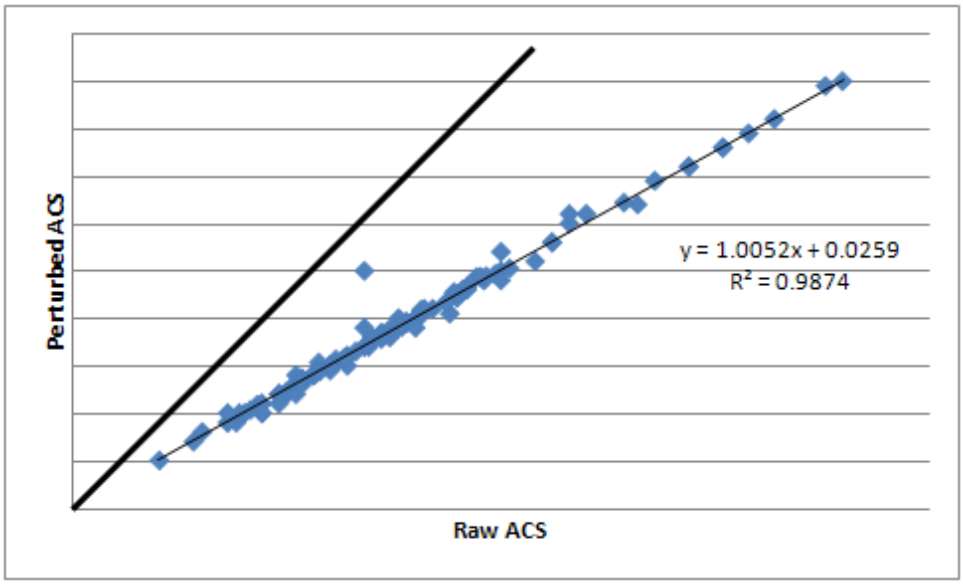


Figure S-36. Test 4: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, District Level, All Modes

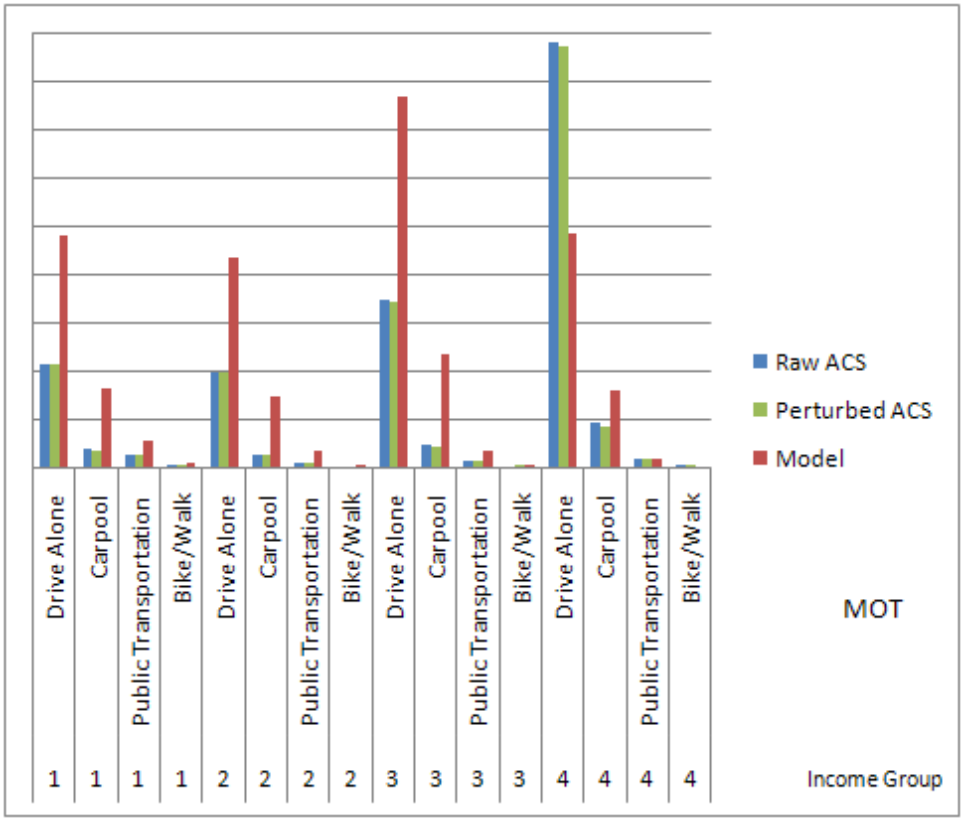


Figure S-37. Test 5: Total HBW Flows by Income by Means of Transportation, Atlanta, County Level

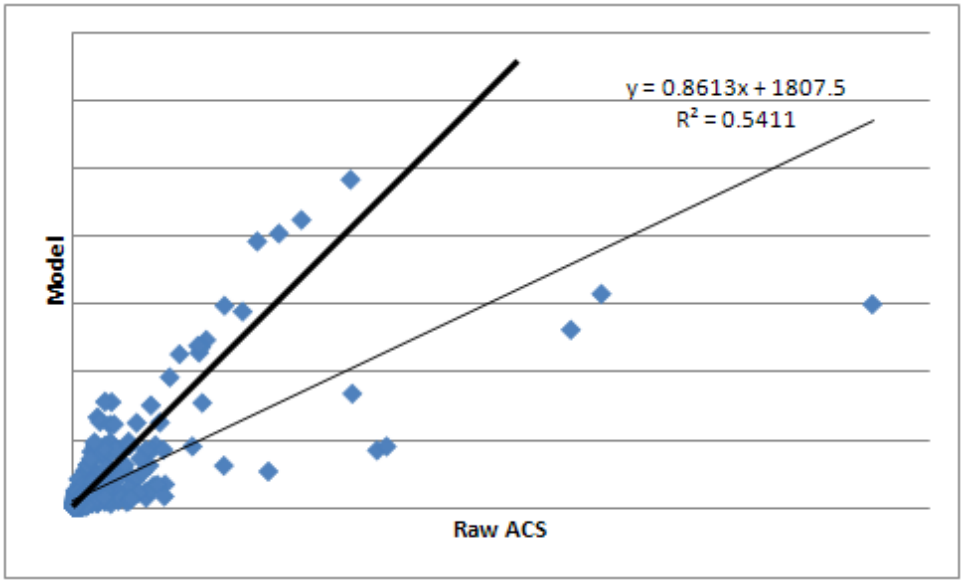


Figure S-38. Test 5: Scatterplot of Raw ACS vs. Model, Atlanta, County Level, All Modes and Income Groups

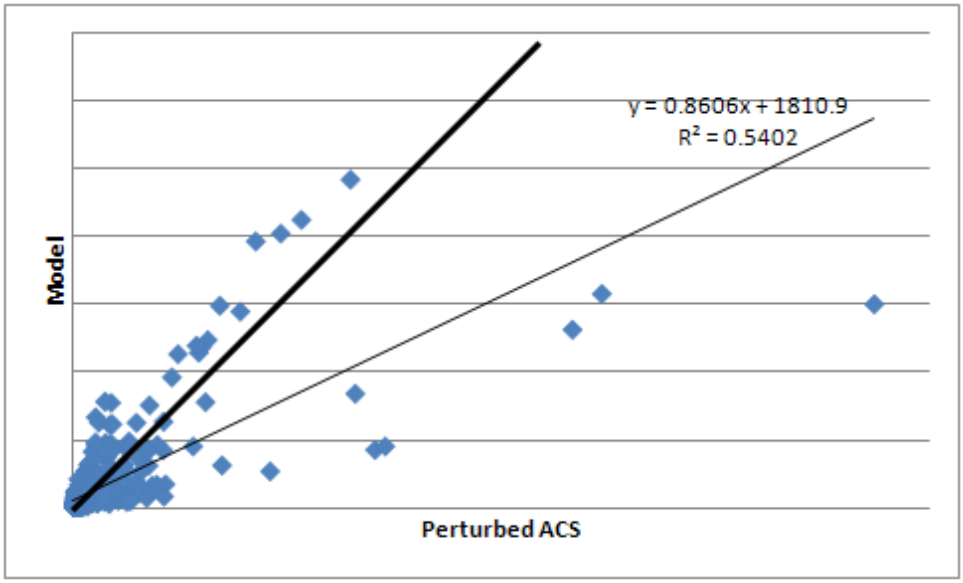


Figure S-39. Test 5: Scatterplot of Perturbed ACS vs. Model, Atlanta, County Level, All Modes and Income Groups

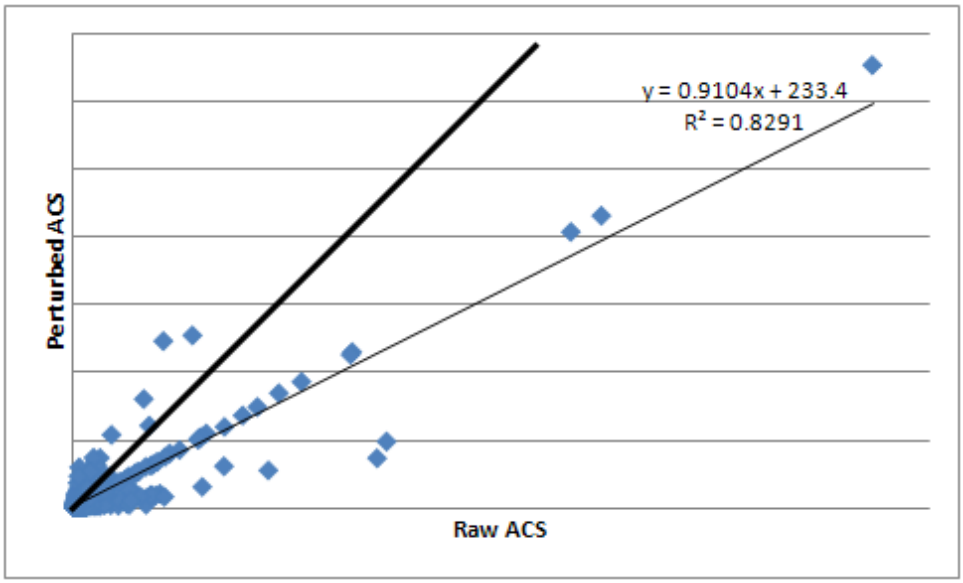


Figure S-40. Test 5: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, County Level, All Modes and Income Groups

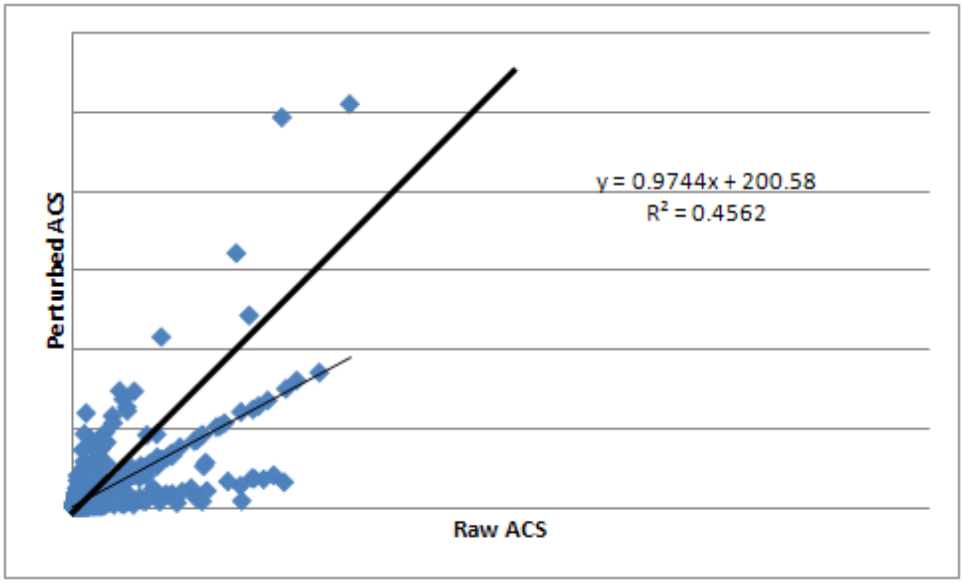


Figure S-41. Test 5: Scatterplot of Raw ACS vs. Perturbed ACS, Atlanta, County Level, All Modes and Income Groups (Flows Under 20,000)

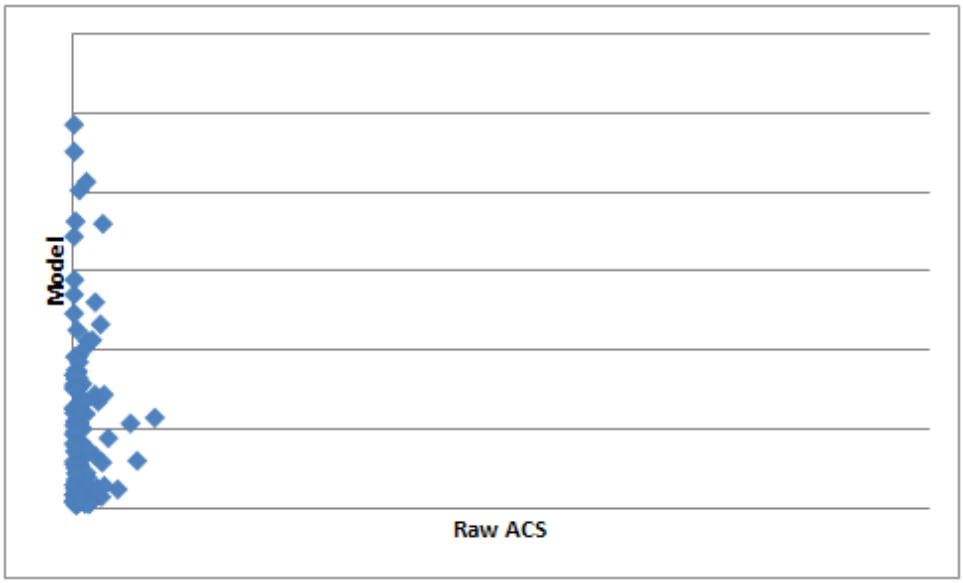


Figure S-42. Test 5: Scatterplot of Raw ACS vs. Model, Atlanta, District Level, All Modes and Income Groups

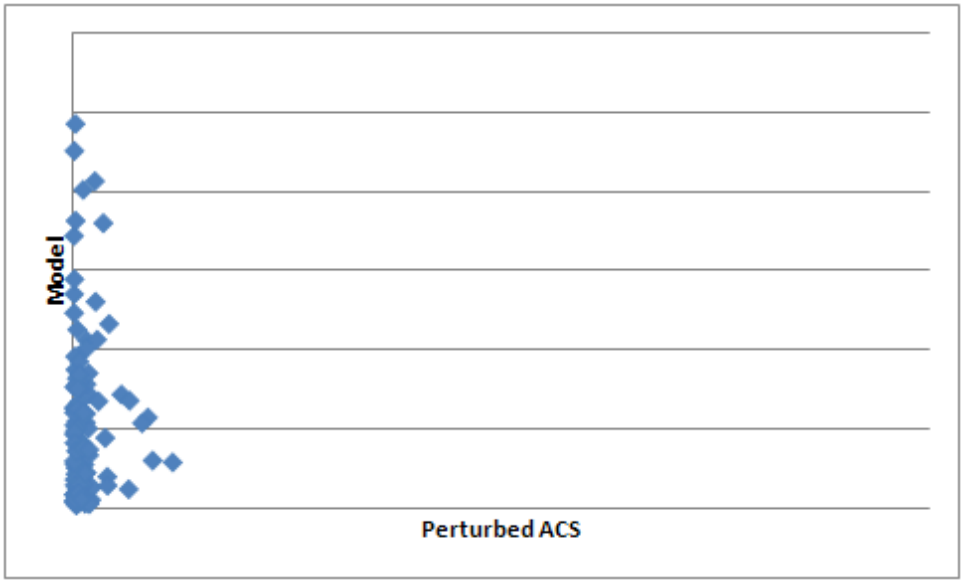


Figure S-43. Test 5: Scatterplot of Perturbed ACS vs. Model, Atlanta, District Level, All Modes and Income Groups

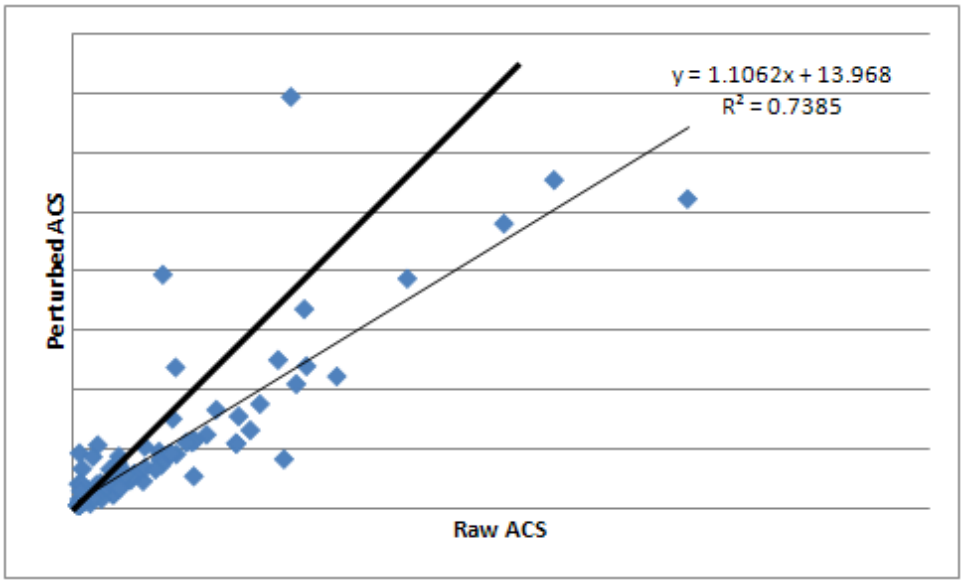


Figure S-44. Test 5: Scatterplot of Raw vs. Perturbed ACS, Atlanta, District Level, All Modes and Income Groups

Figure S-45. Test 5: Number of Trips by Income Level, All Modes, Olympia, County Level

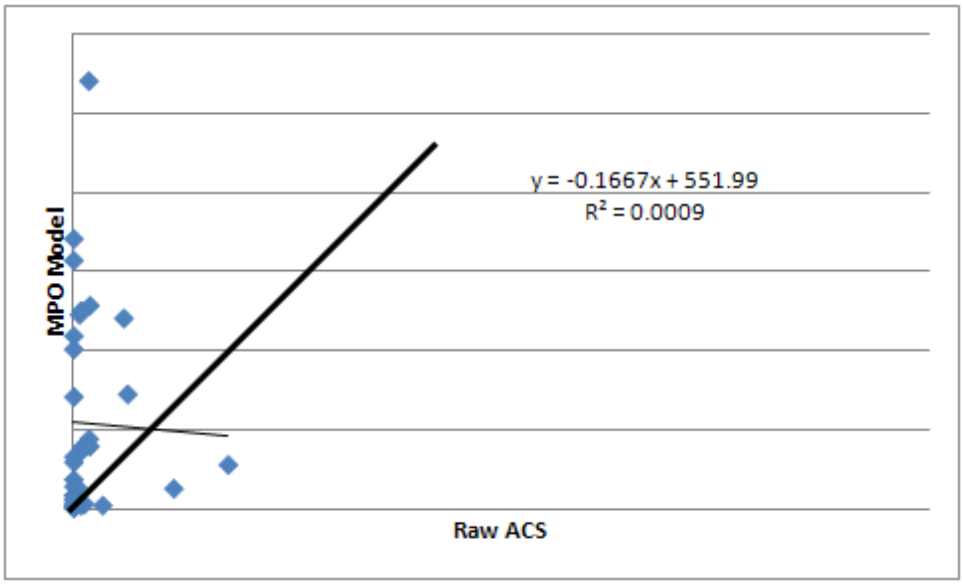


Figure S-46. Test 5: Scatterplot of Raw ACS vs. Model, All Modes, All Income Groups, Olympia, TAZ Level

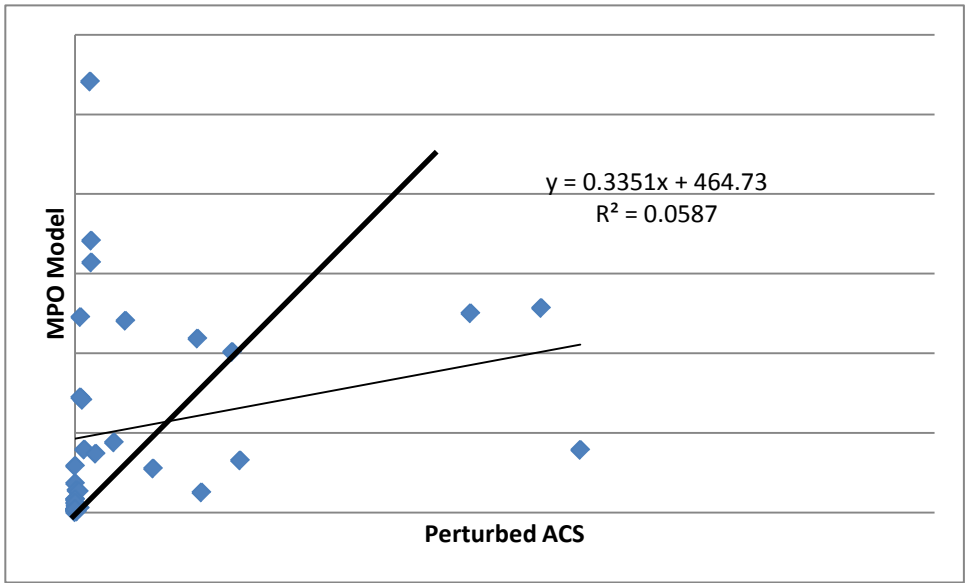


Figure S-47. Test 5: Scatterplot of Perturbed ACS vs. Model, All Modes, All Income Groups, Olympia, TAZ Level

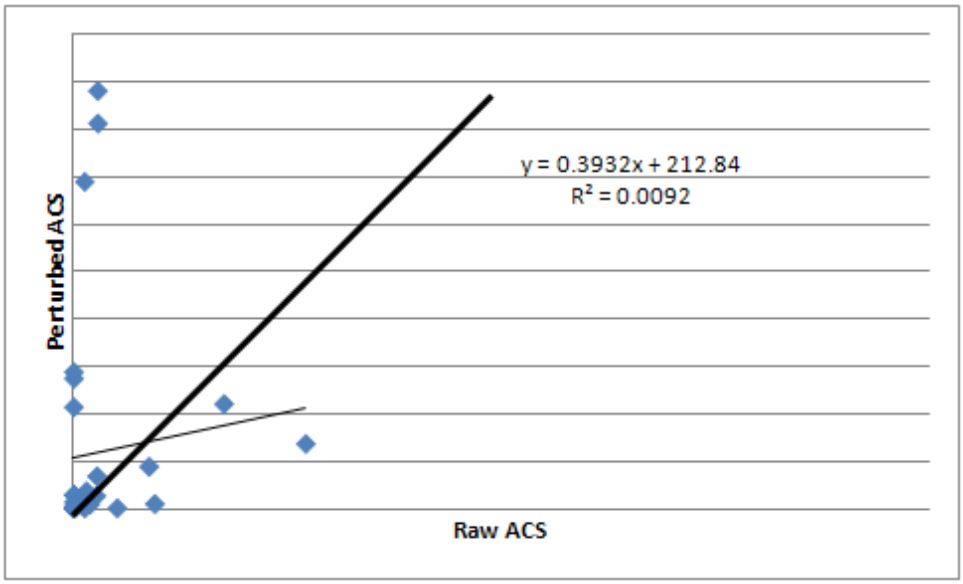


Figure S-48. Test 5: Scatterplot of Raw ACS vs. Perturbed ACS, All Modes, All Income Groups, Olympia, TAZ Level

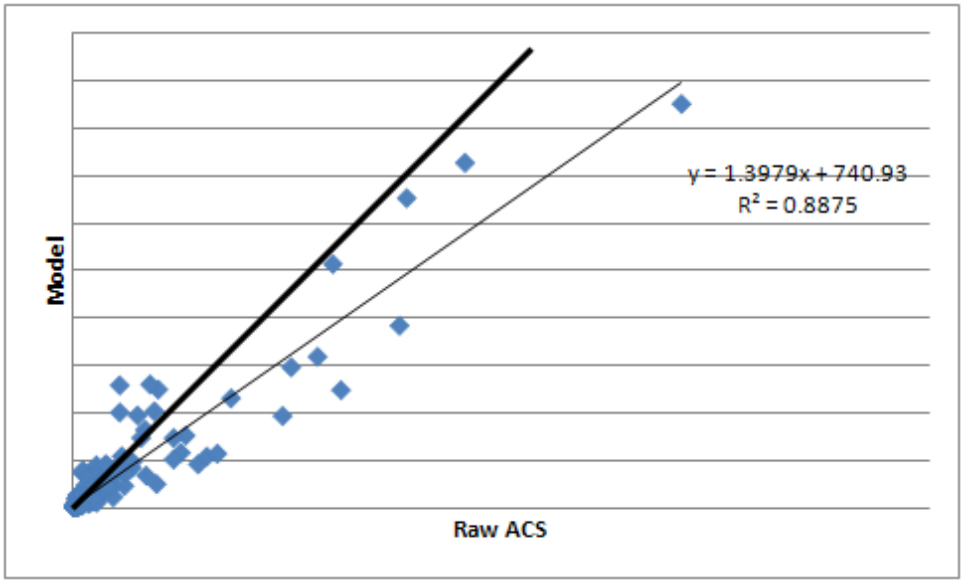


Figure S-49. Test 6: Scatterplot of Raw ACS vs. Model, All Modes and Age of Worker Categories, Atlanta, County Level

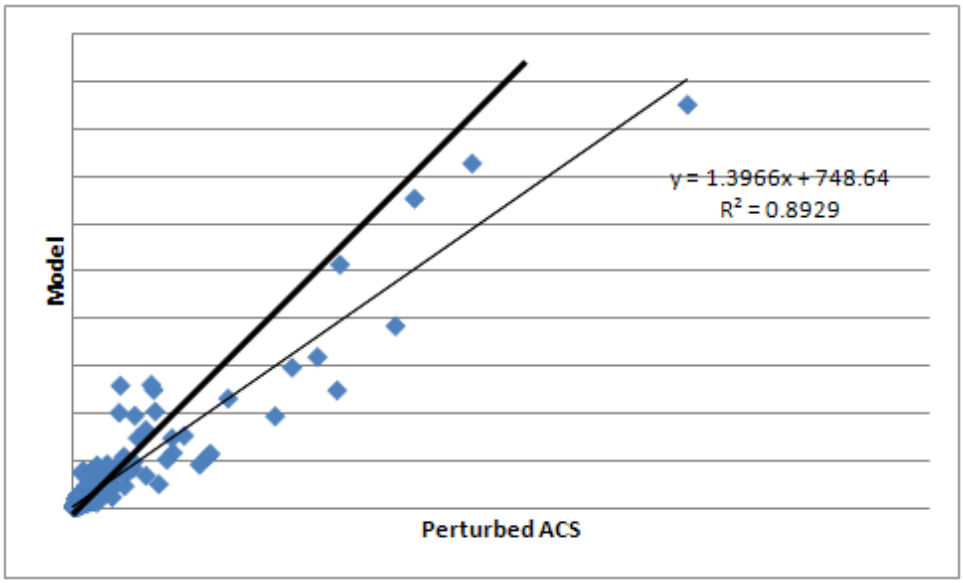


Figure S-50. Test 6: Scatterplot of Perturbed ACS vs. Model, All Modes and Age of Worker Categories, Atlanta, County Level

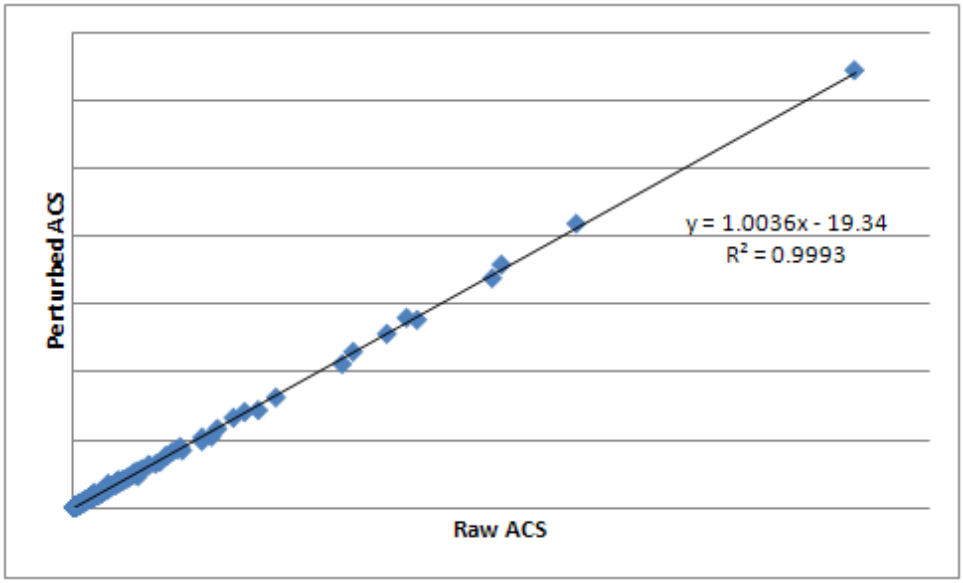


Figure S-51 *Test 6: Scatterplot of Raw ACS vs. Perturbed ACS, All Modes, All Age of Worker Groups, Atlanta, County Level*

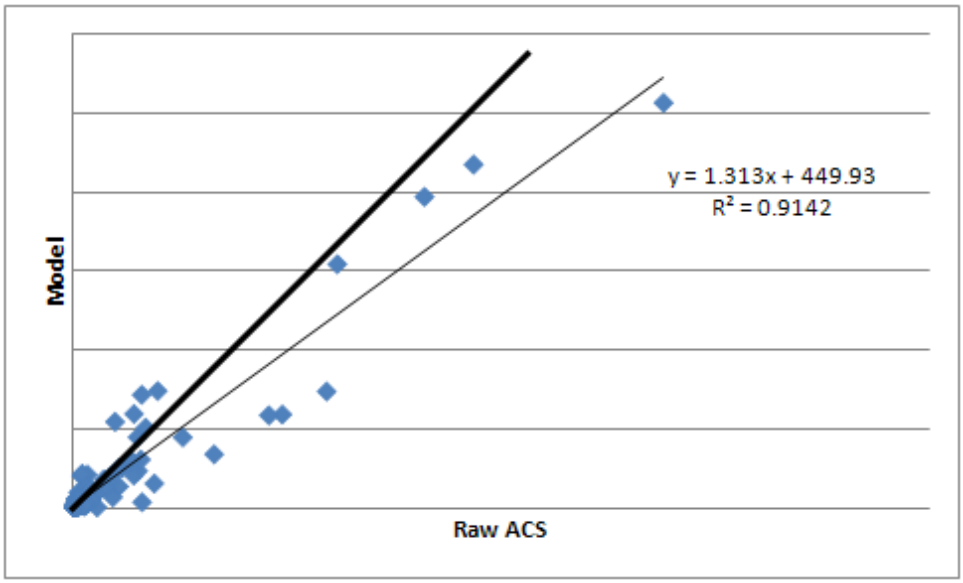


Figure S-52. *Test 7: Scatterplot of Raw ACS vs. Model, All Modes and Age of Worker Groups, Atlanta, County Level*

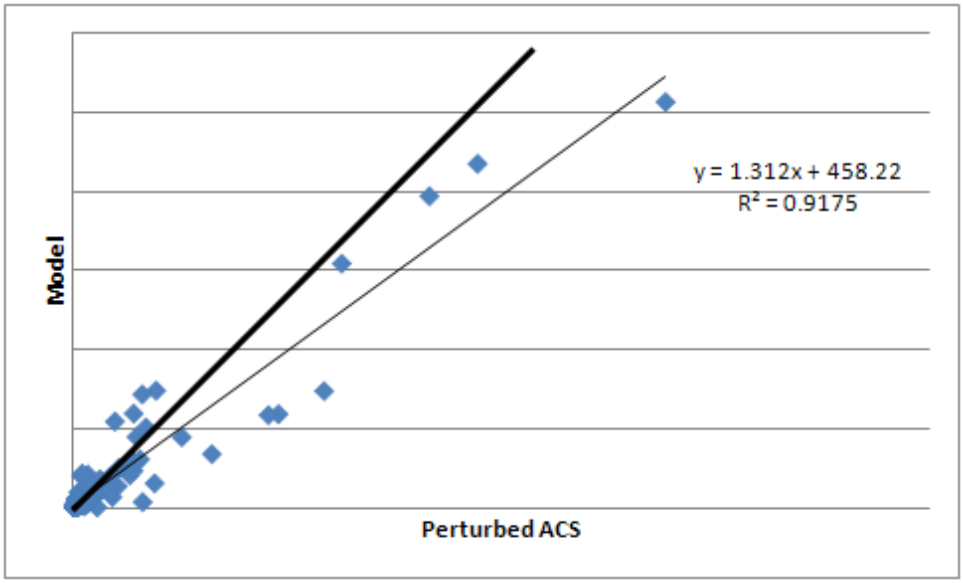


Figure S-53. Test 7: Scatterplot of Perturbed ACS vs. Model, All Modes and Age of Worker Groups, Atlanta, County Level

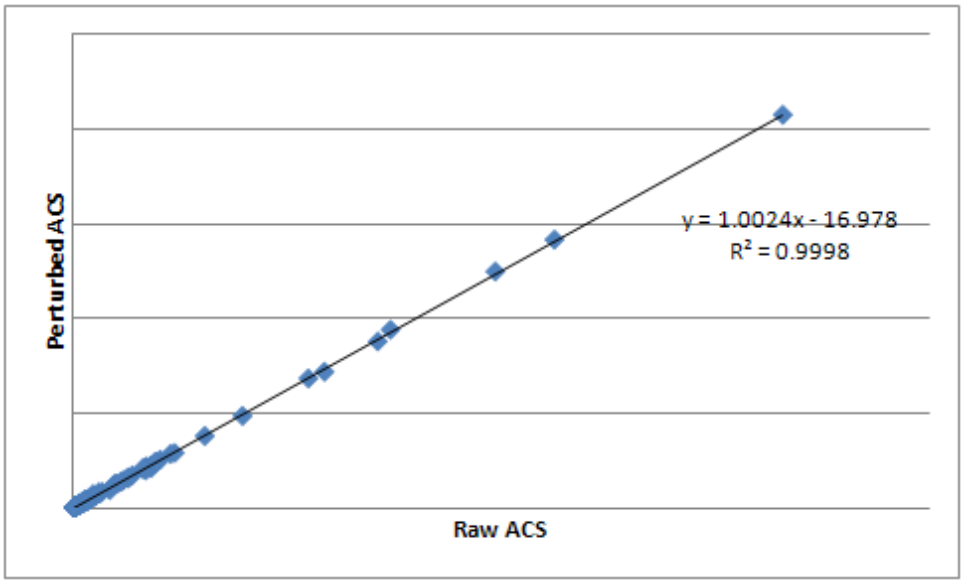


Figure S-54. Test 7: Scatterplot of Raw ACS vs. Perturbed ACS, All Modes and Age of Worker Groups, Atlanta, County Level

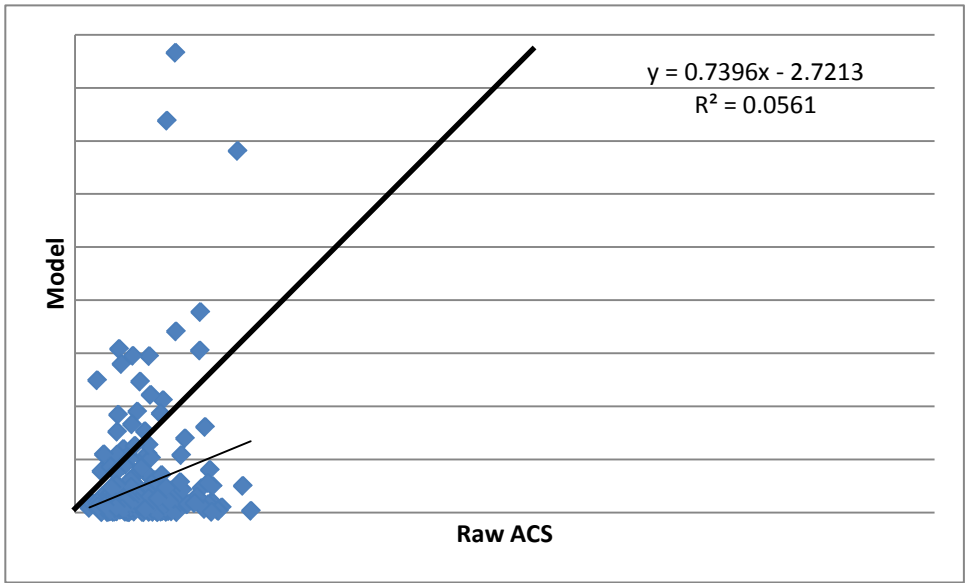


Figure S-55. Test 7: Scatterplot of Raw ACS vs. Model, All Modes and Age of Worker Categories, Atlanta, TAZ Level

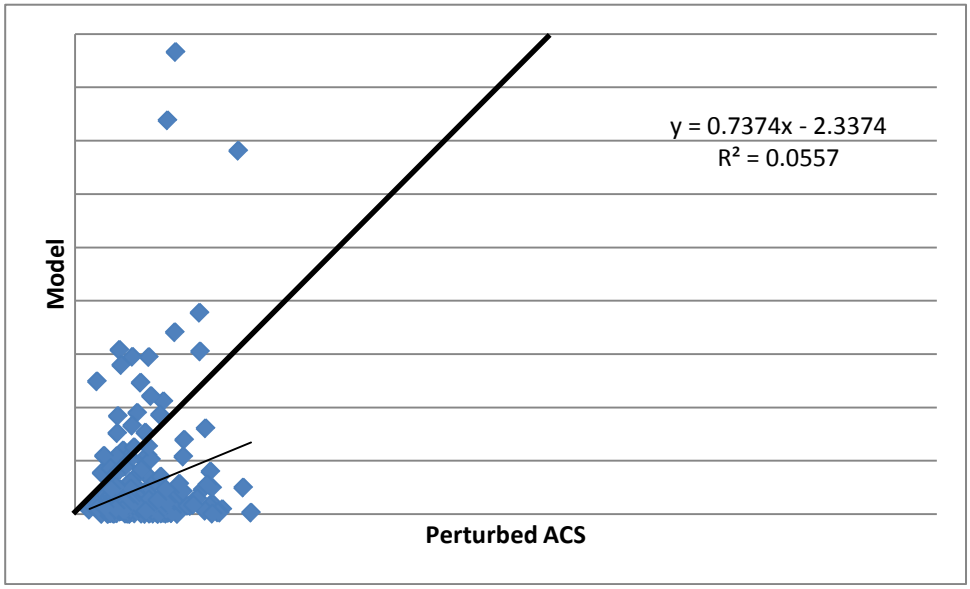


Figure S-56. Test 7: Scatterplot of Perturbed ACS vs. Model, All Modes and Age of Worker Categories, Atlanta, TAZ Level

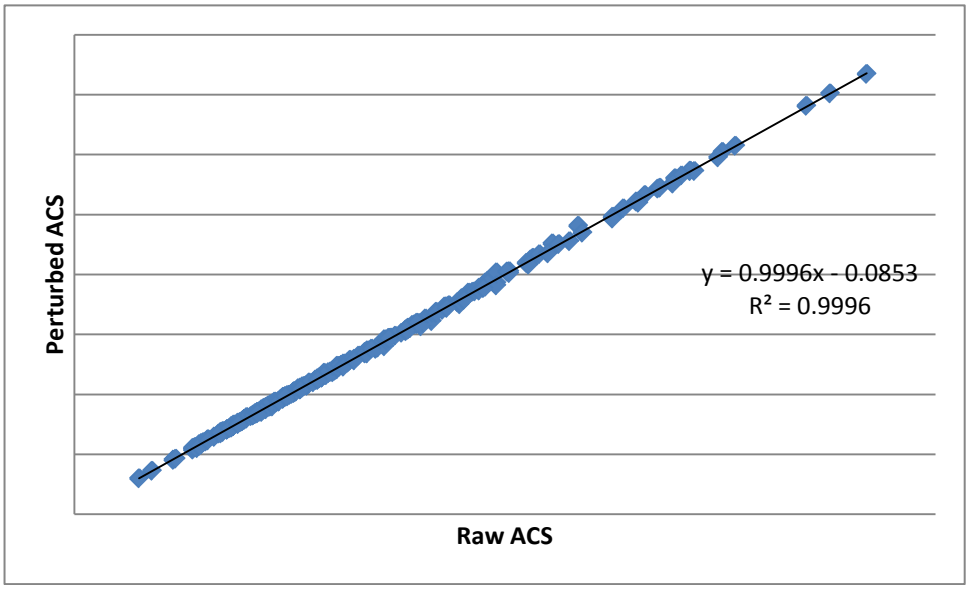


Figure S-57. Test 7: Scatterplot of Raw ACS vs. Perturbed ACS, All Modes and Age of Worker Categories, Atlanta, TAZ Level

Appendix T

SAS CODE FOR DRIVER PROGRAMS

ctpp_main_driver.sas
ira_main_driver.sas
data_replacement.sas
raking_driver.sas
utility.sas
risk.sas
cleanup.sas

CTPP_MAIN_DRIVER.SAS

```
*************************************************************************************************;
**                                                                                             **;
**  Program Name: CTPP_MAIN_DRIVER.SAS                                                         **;
**                                                                                             **;
**  Purpose: This program runs the entire NCHRP system. All phases of NCHRP is submitted by this  **;
**           program.  Initial Risk Analysis, Data Replacement, Raking with Control totals,    **;
**           Utility Analysis, Risk Analysis, and Cleanup programs are all submitted by this   **;
**           program.                                                                          **;
**                                                                                             **;
**  Called Programs: IRA_MAIN_DIRVER.SAS  => The driver program for Initial Risk Analysis      **;
**                   DATA_REPLACEMENT.SAS => The driver program for Data Replacement           **;
**                   RAKING_DRIVER.SAS    => The driver program for Raking with control totals  **;
**                   UTILITY.SAS          => The driver program for Utility Analysis           **;
**                   RISK.SAS             => The Risk Analysis program                         **;
**                   CLEANUP.SAS          => Performs cleanup and creates the delivery files   **;
**                                                                                             **;
*************************************************************************************************;
options ls=130 ps=50 nocenter validvarname=upcase mprint symbolgen;
%LET MACDIR=/cenwork/ctpp/Task7/programs/all_macros;
OPTIONS SET = SRCLIB "&MACDIR" SASAUTOS=('!SRCLIB' SASAUTOS) MAUTOSOURCE;

%include "/cenwork/ctpp/Task7/programs/all_macros/ira_main_driver.sas" ;

%include "/cenwork/ctpp/Task7/programs/all_macros/data_replacement.sas";

%include "/cenwork/ctpp/Task7/programs/all_macros/raking_driver.sas";

%include "/cenwork/ctpp/Task7/programs/all_macros/utility.sas";

%include "/cenwork/ctpp/Task7/programs/all_macros/risk.sas";

%include "/cenwork/ctpp/Task7/programs/all_macros/cleanup.sas";
```

IRA_MAIN_DRIVER.SAS

```
/*************************************************************************************************************;
**                                                                                                            **;
**  Macro Name: IRA_MAIN_DRIVER                                                                               **;
**                                                                                                            **;
**  Purpose: Main Driver Program for the Initial Risk Analysis                                                **;
**                                                                                                            **;
**  Initial Risk Analysis Description:                                                                        **;
**                                                                                                            **;
**        The set of initial risk analysis modules were processed to generate tables. The tables              **;
**        were generated to flag data values that violate the DRB rules and therefore are at the              **;
**        highest risk of disclosure. It is important to note that the table generator needs to               **;
**        incorporate any additional tables requested. Several steps are necessary within the                 **;
**        initial risk analysis component to prepare for the application of the perturbation                  **;
**        approach, including: Creation of Combined TAZs, ACS area-level covariates, and input                **;
**        data preparation.  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, or swapped through the ACS disclosure process, will not be                      **;
**        replaced, that is, they will be considered as to have already been perturbed.                       **;
**        This approach is acceptable to the DRB. As part of the initial risk analysis, data                  **;
**        values were classified according to risk strata. The following flags were created to                **;
**        assist in the perturbation process as well as in the disclosure risk measures:                      **;
**        VarName_FLG, VarName_FLG2,VarName_FLG3, VarName_FLG4, VarName_RPL, VarName_FULL,                    **;
**        VarName_SING, VarName_STRT.                                                                         **;
**                                                                                                            **;
**  Function:  To run this program, make sure the following parameters are current:                           **;
**             housfile = (the most current household file)                                                   **;
**             persfile = (the most current Person file)                                                      **;
**             datadir  = (the directory for all input and output files)                                      **;
**             progdir  = (the directory where the programs are located)                                      **;
**                                                                                                            **;
**  Called by: CTPP_MAIN_DRIVER.SAS                                                                           **;
**                                                                                                            **;
**  Called Macros: MAIN_PROG1    => creates Housing and Person CTAZ files                                     **;
**                 M30_STEP1     => creates subset files for both Person and Household files                  **;
**                 M30           => calls Initial Risk Analysis programs for Person                           **;
**                 M30_MAIN_HOUS => calls Initial Risk Analysis programs for Household                        **;
**                                                                                                            **;
**  Input Files:                                                                                              **;
**                                                                                                            **;
**        VPERS5REC.SAS7BDAT                                                                                  **;
**        VHOUS5REC.SAS7BDAT                                                                                  **;
**        HU_SWAPOUTUS_2005.SAS7BDAT                                                                          **;
**        HU_SWAPOUTUS_2006.SAS7BDAT                                                                          **;
**        HU_SWAPOUTUS_2007.SAS7BDAT                                                                          **;
**        HU_SWAPOUTUS_2008.SAS7BDAT                                                                          **;
**        HU_SWAPOUTUS_2009.SAS7BDAT                                                                          **;
**        GQ_CHG_2006.SAS7BDAT                                                                                **;
**        GQ_CHG_2007.SAS7BDAT                                                                                **;
**        GQ_CHG_2008.SAS7BDAT                                                                                **;
**        GQ_CHG_2009.SAS7BDAT                                                                                **;
**        COUNTYFILE.SAS7BDAT                                                                                 **;
**        SF1BLK2K.SAS7BDAT                                                                                   **;
**                                                                                                            **;
**  Output Files:  VPERS5REC_SUB.SAS7BDAT                                                                     **;
**                 VHOUS5REC_SUB.SAS7BDAT                                                                     **;
**                 VPERS5REC_CTAZ.SAS7BDAT                                                                    **;
```

```
**                VHOUS5REC_CTAZ.SAS7BDAT                                                **;
**                VPERS5REC_IRA.SAS7BDAT                                                 **;
**                VHOUS5REC_IRA.SAS7BDAT                                                 **;
**                                                                                       **;
*******************************************************************************************/;

%macro ira_main_driver;

  title "CTPP - Initial Risk Analysis";

  %main_prog1

  %m30_step1

  %m30

  %m30_main_hous

  /* Special CTAZ Reporting programs */

  /*
  %main_prog2_ctaz

  %main_prog2_ctaz_hous

  %main_prog2_ctaz_hous_300

  %main_prog2_ctaz_300
  */

%mend ira_main_driver;
```

DATA_REPLACEMENT.SAS

```
/*****************************************************************************************************;
**                                                                                                    **;
**  Macro Name: DATA_REPLACEMENT       -- EXAMPLE                                                     **;
**                                                                                                    **;
**  Purpose: Main program for Data Replacement using SP and CH/RL approaches.                         **;
**                                                                                                    **;
**  Data Replacement Description:                                                                     **;
**                                                                                                    **;
**    This program combines the constrained hotdeck and semi-parametric approaches into one program.**;
**    The initial steps prior to 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      **;
**    perturbed and the rates were used to select and flag (VarName_PARTIAL) a sample of data       **;
**    values for replacement for each of the test sites. The Variable Prep step is processed in     **;
**    order to prepare recodes and prepare variables as predictors for the semi-parametric approach.**;
**    The master index file (MIF) drives the process and identifies the variables to be perturbed   **;
**    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 and unordered      **;
**    categorical. For the unordered categorical variables, indicator variables were created.       **;
**    Select interaction terms to be added to the pool of candidate predictor variables are         **;
**    identified as well.                                                                           **;
**                                                                                                  **;
**    Once the variable prep processing was completed, then the model selection approach is         **;
**    processed for all variables identified in the MIF that undergo the semi-parametric approach.  **;
**    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 perturbation step.                               **;
**                                                                                                  **;
**    One by one, the target variables are processed through the Main Loop. Either the constrained  **;
**    hotdeck or the semi-parametric approach is processed, depending on the variable type of the   **;
**    target variable. First, household-level variables are perturbed, then the perturbed household **;
**    variables are transferred to the person level, where the process continues with the           **;
**    perturbations on person-level variables.                                                      **;
**                                                                                                  **;
**    After processing, pre-post checks are conducted in order to have an initial look at the       **;
**    impact of the perturbations. Frequencies, means and correlations are generated before and     **;
**    after perturbation. Lastly, recodes are processed in order to prepare for the raking step.    **;
**                                                                                                  **;
**  Called by:  CTPP_MAIN_DRIVER.SAS                                                                **;
**                                                                                                  **;
**  Called Macros:  PARTIAL_FLAGS_NEW                                                               **;
**                  VARIABLE PREP                                                                   **;
**                  MODELING_STEPS                                                                  **;
**                  MAIN_LOOP                                                                       **;
**                  PRE_POST_CHECKS                                                                 **;
**                  CHANGE_SUMMARY                                                                  **;
**                  RECODES                                                                         **;
**                                                                                                  **;
**  Input Files:  VPERS5REC_IRA.SAS7BDAT                                                            **;
**                VHOUS5REC_IRA.SAS7BDAT                                                            **;
**                ALL_MIF_V13.SAS7BDAT                                                              **;
**                                                                                                  **;
**  Output Files:  V1HNAT_PARTIAL.SAS7BDAT                                                          **;
**                 V1PNAT_PARTIAL.SAS7BDAT                                                          **;
**                                                                                                  **;
*****************************************************************************************************/;
```

```
%MACRO DATA_REPLACEMENT;

  %GLOBAL MIF HH_DSN PER_DSN SEED MERGEVAR MERGEVAR2 X AMOUNT DEBUG AMOUNT1 SITE
        RAREA WAREA HH_OUT_DSN PER_OUT_DSN HHMOT_MAX PERMOT_MAX
        DEPVAR TYPE PREDPOOL START END HDCELLVAR TRANS_VARLIST ACSVARS_H ACSVARS_P
  ;

  %LET MIF=ALL_MIF_V13;
  %LET HH_DSN=VHOUS5REC_IRA;
  %LET PER_DSN=VPERS5REC_IRA;
  %LET RUN_NUM=1;
  %LET SEED=54651;
  %LET MERGEVAR=cmid;
  %LET MERGEVAR2=pnum;
  %LET X=M;
  %LET AMOUNT=PARTIAL;
  %LET AMOUNT1=%SUBSTR(&AMOUNT,1,4);
  %LET DEBUG=N;
  %LET RAREA=RArea;
  %LET WAREA=wArea;
  %LET FSWGT=REPW0;
  %LET HH_OUT_DSN=V&RUN_NUM.HNAT_&AMOUNT;
  %LET PER_OUT_DSN=V&RUN_NUM.PNAT_&AMOUNT;
  %LET SITE=NAT;

  LIBNAME HERE "&DATADIR";

  TITLE "CTPP - DATA REPLACEMENT ON &HH_DSN AND &PER_DSN";

  DATA &HH_DSN;
    SET HERE.&HH_DSN;
     /* WHERE ST = '11'; */
  RUN;

  DATA &PER_DSN;
    SET HERE.&PER_DSN;
     /* WHERE ST = '11'; */
  RUN;


  /*********************************************/
  /* Creation of partial flags -- EXAMPLE      */
  /*********************************************/

  %Partial_Flags_New(filename=&PER_DSN,
                     varname=JWD,
                     samprate=(1 1 1 1),
                     seed=28461557,
                     flag= PARTIAL)

  %Partial_Flags_New(filename=&PER_DSN,
                     varname=JWMN,
                     samprate=(1 1 1 1),
                     seed=32898476,
                     flag= PARTIAL)
```

...

```
   DATA &HH_DSN(drop = rc_h yngest ur r18 lngi tel hpov r60 rms under18
                rename = (yngest_n  = yngest
                          ur_n      = ur
                          r18_n     = r18
                          lngi_n    = lngi
                          tel_n     = tel
                          hpov_n    = hpov
                          r60_n     = r60
                          rms_n     = rms
                          under18_n = under18));
    SET &HH_DSN;
    sum_&AMOUNT = sum();
  length yngest_n  8;
  length ur_n      8;
  length r18_n     8;
  length lngi_n    8;
  length tel_n     8;
  length hpov_n    8;
  length r60_n     8;
  length rms_n     8;
  length under18_n 8;
  yngest_n  = put(yngest,$1.);
  ur_n      = put(ur,$1.);
  r18_n     = put(r18,$1.);
  lngi_n    = put(lngi,$1.);
  tel_n     = put(tel,$1.);
  hpov_n    = put(hpov,$1.);
  r60_n     = put(r60,$1.);
  rms_n     = put(rms,$1.);
  under18_n = put(under18,$1.);
  RUN;

  DATA &PER_DSN(drop = rc_h yngest ur under18 sex wkw new_mil esr
                rename=(yngest_n=yngest
                        ur_n=ur
                        under18_n=under18
                        sex_n=sex
                        wkw_n=wkw
                        new_mil_n=new_mil
                        esr_n=esr));
    SET &PER_DSN;
  sum_&AMOUNT = sum();
  length yngest_n   8;
  length ur_n       8;
  length under18_n  8;
  length sex_n      8;
  length wkw_n      8;
  length new_mil_n  8;
  length esr_n      8;
  yngest_n   = put(yngest,$1.);
  ur_n       = put(ur,$1.);
  under18_n  = put(under18,$1.);
  sex_n      = put(sex,$1.);
  wkw_n      = put(wkw,$1.);
  new_mil_n  = put(new_mil,$1.);
```

```
esr_n      = put(esr,$1.);

if jwmn_partial=0 and jwd_partial=1 then jwmn_partial=1;
if jwmn_partial=1 and jwd_partial=0 then jwd_partial=1;

RUN;
/*
proc freq data = &HH_DSN;
  tables mos_hous * mos_pers * sum_&AMOUNT /list missing;
  title2'CrossTab of MOS_HOUS, MOS_PERS, and SUM_&AMOUNT';
run;

proc freq data = &PER_DSN;
  tables mos_pers * sum_&AMOUNT /list missing;
  title2 'CrossTab on MOS_PERS and SUM_&AMOUNT';

run;
*/

proc means data = &HH_DSN n min median mean max;
class sum_&amount;
var mos_hous;
title2 "Household Level mos_hous by sum_&amount";
run;

proc means data = &PER_DSN n min median mean max;
class sum_&amount;
var mos_pers;
title2 "Person Level mos_pers by sum_&amount";
run;


/************************************/

DATA &MIF;
  SET HERE.&MIF;
RUN;

/* VARIABLE PREPARATION */

%VARIABLE_PREP

/* MODELING STEPS */

%MODELING_STEPS

/* DETERMINE MINIMUM AND MAXIMUM PROCESS NUMBERS FOR HOUSEHOLD DATA REPLACEMENT */

/*
PROC MEANS NOPRINT DATA=&MIF.2 (WHERE=(VHOUS=1 AND Approach NE ' '));
  VAR ProcessNumber;
  OUTPUT OUT=HHPROC (DROP=_FREQ_ _TYPE_) MIN=PN_MIN_HH MAX=PN_MAX_HH;
RUN;

DATA _NULL_;
  SET HHPROC;
  CALL SYMPUT('START',LEFT(PUT(PN_MIN_HH,8.)));
```

```
  CALL SYMPUT('END',LEFT(PUT(PN_MAX_HH,8.)));
RUN;
*/

proc sort data = &mif.2 out = subset1 nodupkey;
  WHERE VHOUS=1 AND
      Approach NE ' ';
  by processnumber;

run;

proc print data = subset1;
  var varname processnumber;
  title2 'housing varname and processnumber';

run;

title2 ' ';

data _null_;
set subset1 end=eof;

cnt_h + 1;

start = 1;

call symputx('Pronum'||trim(left(put(cnt_h,2.))),ProcessNumber);

if eof then do;

  CALL SYMPUT('START',LEFT(PUT(start,8.)));
  CALL SYMPUT('END',LEFT(PUT(cnt_h,8.)));

end;

run;

/* CALL MAIN LOOP FOR HOUSEHOLD DATA REPLACEMENT */

%MAIN_LOOP (INDSN=&HH_DSN,filelevel=HH)

/* RECOPY HOUSEHOLD VARIABLES ONTO PERSON FILE AND RECREATE PERSON INTERACTION VARIABLES */

DATA TOMERGE;
  SET &HH_DSN (KEEP=&MERGEVAR &TRANS_VARLIST);
RUN;

PROC SORT DATA=&PER_DSN;
  BY &MERGEVAR;
RUN;

DATA &PER_DSN;
  MERGE &PER_DSN (in=aa DROP=&TRANS_VARLIST)
        TOMERGE;
  BY &MERGEVAR;
  /* set any missings on trans_varlist to 0 (for GQ) */
  %setzero
```

```
  if aa then output;

if HH_WRK6 in (4,5,6) then D_WRK = 2;
else D_WRK = 1;

if ahinc<50000 then D_AHINC = 1;
else D_AHINC = 2;


RUN;

/* DETERMINE MINIMUM AND MAXIMUM PROCESS NUMBERS FOR HOUSEHOLD DATA REPLACEMENT */

proc sort data = &mif.2 out = subset1 nodupkey;
  WHERE VHOUS=0 AND
        VPERS=1 AND
      Approach NE ' ';
  by processnumber;

run;
/*
proc print data = subset1;
  var varname processnumber;
  title2 'Person varname and processnumber';

run;

title2 ' ';
*/
data _null_;
set subset1 end=eof;

cnt_p + 1;

start = 1;

call symputx('Pronum'||trim(left(put(cnt_p,2.))),ProcessNumber);

if eof then do;

  CALL SYMPUT('START',LEFT(PUT(start,8.)));
  CALL SYMPUT('END',LEFT(PUT(cnt_p,8.)));

end;

run;

/* CALL MAIN LOOP FOR PERSON DATA REPLACEMENT */

%MAIN_LOOP (INDSN=&PER_DSN,filelevel=PP)

/* RUN PRE AND POST DATA CHECKS */

%pre_post_checks

/* RUN SUMMARY OF CHANGES */
```

```
  %change_summary

  %recodes

%MEND DATA_REPLACEMENT;
```

RAKING_DRIVER.SAS

```
/*************************************************************************************************************;
**                                                                                                            **;
**  Macro Name: RAKING_DRIVER                                                                                 **;
**                                                                                                            **;
**  Purpose: Main Driver program for Control Totals and Raking.                                               **;
**                                                                                                            **;
**  Raking Description:                                                                                       **;
**                                                                                                            **;
**       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 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. In addition, a dimension              **;
**       was added to calibrate to the estimated total number of workers at the CTAZ300 level,               **;
**       which are areas of about 8,000 in population. The weight calibration process employed               **;
**       sample-based raking, meaning that the estimates for the modified estimates reflected the            **;
**       sampling error of the five-year ACS control totals, rather than considering these totals            **;
**       to be error-free, as is often the case with calibration methods. For sample-based raking,           **;
**       each replicate weight for the modified file was raked to its corresponding replicate                **;
**       weight estimated totals from the five-year ACS.                                                     **;
**                                                                                                            **;
**  Called by: CTPP_MAIN_DRIVER.SAS                                                                           **;
**                                                                                                            **;
**  Called Macro:  CONTROLTOTALS_HH   => Generates Household Level Control total file                        **;
**                 CONTROLTOTALS_PERS => Generates Person Level Control total file                           **;
**                 RAKING_HH          => Performs Household Level Raking                                     **;
**                 RAKING_PERS        => Performs Person Level Raking                                        **;
**                                                                                                            **;
**  Input Files:  VHOUS5REC_IRA.SAS7BDAT                                                                      **;
**                VPERS5REC_IRA.SAS7BDAT                                                                      **;
**                V1HNAT_PARTIAL.SAS7BDAT                                                                     **;
**                V1PNAT_PARTIAL.SAS7BDAT                                                                     **;
**                TABLE1.DAT                                                                                  **;
**                TABLE3.DAT                                                                                  **;
**                                                                                                            **;
**  Output Files:  CVHOUS5REC_IRA.SAS7BDAT                                                                    **;
**                 CVPRES5REC_IRA.SAS7BDAT                                                                    **;
**                 RV1HNAT_PARTIAL.SAS7BDAT                                                                   **;
**                 RV1PNAT_PARTIAL.SAS7BDAT                                                                   **;
**                                                                                                            **;
*************************************************************************************************************/;
* RAKING_HH.SAS ;
* kt 5/2010 ;

* autocall macros:  summ,  rake, conv_report;

%MACRO RAKING_DRIVER;

  %LET HH_DSN=VHOUS5REC_IRA;
  %LET PER_DSN=VPERS5REC_IRA;
  %LET RUN_NUM=1;
  %LET HH_OUT_DSN=V&RUN_NUM.HNAT_&AMOUNT;
  %LET PER_OUT_DSN=V&RUN_NUM.PNAT_&AMOUNT;
  %LET SUFFIX=NAT;
  %LET AMOUNT=PARTIAL;

  %controltotals_hh
```

```
  %controltotals_pers

  %raking_hh

  %raking_pers

%MEND RAKING_DRIVER;
```

UTILITY.SAS

```
/*************************************************************************************************************;
**                                                                                                             **;
**  Program Name:  UTILITY.SAS           -- EXAMPLE                                                            **;
**                                                                                                             **;
**  Purpose:  Performs Data Utility measures.                                                                  **;
**                                                                                                             **;
**  Utility Description:                                                                                       **;
**                                                                                                             **;
**        The data perturbation 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 perturbed ACS data. The                  **;
**        comparisons check cell means, 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                    **;
**        perturbed estimates (across estimates for geographic areas) were computed where                      **;
**        appropriate in order to give indications of potential bias introduced by the                         **;
**        perturbation.  The interquartile range for the differences provided an indication of the             **;
**        variation caused by the perturbations. Lastly, there is a check on the differences for               **;
**        medians and 75th percentiles of travel time for table cells across estimates for                     **;
**        geographic areas.                                                                                    **;
**                                                                                                             **;
**  Called by: CTPP_MAIN_DRIVER.SAS                                                                            **;
**                                                                                                             **;
**  Called Programs: ACS_CMR.SAS                                                                               **;
**                   ACS_CQR.SAS                                                                               **;
**                   ACS_CMSE.SAS                                                                              **;
**                   CRV.SAS                                                                                   **;
**                   ACS_ASSOC.SAS                                                                             **;
**                   CELL_MEAN_RATIOS.SAS                                                                      **;
**                   CELL_QUANTILE_RATIOS.SAS                                                                  **;
**                   PERT_CMSE.SAS                                                                             **;
**                   PAIR_ASSOC.SAS                                                                            **;
**                   MULT_ASSOC.SAS                                                                            **;
**                                                                                                             **;
**  Input Files:  CVPERS5REC_IRA.SAS7BDAT                                                                      **;
**                CVHOUS5REC_IRA.SAS7BDAT                                                                      **;
**                RV1PNAT_PARTIAL.SAS7BDAT                                                                     **;
**                RV1HNAT_PARTIAL.SAS7BDAT                                                                     **;
**                CMR_TABLE.DAT                                                                                **;
**                CQR_TABLE.DAT                                                                                **;
**                CMSE_TABLE.DAT                                                                               **;
**                                                                                                             **;
**                                                                                                             **;
*************************************************************************************************************/;
* Utility.SAS ;
* kt 5/2010 ;

* autocall macros:  cell_mean_ratios cmr_compute ;

title "CTPP - Data Utility Measures" ;

%MACRO UTILITY;
```

```
  %let debug = N ;

  %let type1 = H ;
  %let type2 = P ;
  %let amount1 = PARTIAL ;

  %global geo g;

  Proc datasets library = here nolist ;
     delete cellMedianDiffs cellp75Diffs cellMeanDiffs SERDiffs cramerVDiffs Associations Ustat Ustat2
;
  run ;

  %macro loop;

  %do t = 1 %to 2 ;

     %if &&type&t = H %then %let byvar = VEHICLES6;
     %else
     %if &&type&t = P %then %let byvar = MEANS6;

     %do s = 1 %to 1 ;

        data CV&&type&t.&&site&s.;

        set CV&&type&t.&&site&s.;

       %if &&type&t = P %then %do;

        If 5 <= MEANS11 <= 10 then MEANS6 = 5;
        else if means11 = 11 then means6 = 6;
        Else
          MEANS6 = Means11;

        DUMMY = 1;

        If MINORITY = 2 then MINORITY_1 = 1;
        Else
          MINORITY_1 = 0;

        If POVERTY = 2 then POVERTY_1 = 1;
        Else
          POVERTY_1 = 0;

        If POVERTY = 3 then POVERTY_2 = 1;
        Else
          POVERTY_2 = 0;

        If INDUSTRY8 = 2 then INDUSTRY_1 = 1;
        Else
          INDUSTRY_1 = 0;

        If INDUSTRY8 = 3 then INDUSTRY_2 = 1;
        Else
          INDUSTRY_2 = 0;

        If INDUSTRY8 = 4 then INDUSTRY_3 = 1;
```

```
  Else
    INDUSTRY_3 = 0;

  If INDUSTRY8 = 5 then INDUSTRY_4 = 1;
  Else
    INDUSTRY_4 = 0;

  If INDUSTRY8 = 6 then INDUSTRY_5 = 1;
  Else
    INDUSTRY_5 = 0;

  If INDUSTRY8 = 7 then INDUSTRY_6 = 1;
  Else
    INDUSTRY_6 = 0;

  If INDUSTRY8 = 8 then INDUSTRY_7 = 1;
  Else
    INDUSTRY_7 = 0;

%end;

run;

%acs_cmr (
   Type     = &&type&t ,
   Site     = &&site&s
   )

%acs_cqr (
   Type     = &&type&t ,
   Site     = &&site&s
   )

%acs_cmse (
   Type     = &&type&t ,
   Site     = &&site&s
   )
%crv (
   Approach = ACS ,
   Type     = &&type&t ,
   Site     = &&site&s
   )
%acs_assoc (
   Type     = &&type&t ,
   Site     = &&site&s
   )

proc datasets lib = work ; run ;
   %do run = 1 %to 1 ;

             %cell_mean_ratios (
            Run      = &run ,
            Type     = &&type&t ,
            Site     = &&site&s ,
            Amount   = &&amount1
            )
```

```
                    %cell_quantile_ratios (
                   Run      = &run ,
                   Type     = &&type&t ,
                   Site     = &&site&s ,
                   Amount   = &&amount1
                   )

                %pert_cmse (
                   Run      = &run ,
                   Type     = &&type&t ,
                   Site     = &&site&s ,
                   Amount   = &&amount1
                   )
                %crv (
                   Run      = &run ,
                   Type     = &&type&t ,
                   Site     = &&site&s ,
                   Amount   = &&amount1
                   )
                %pair_assoc (
                   Run      = &run ,
                   Type     = &&type&t ,
                   Site     = &&site&s ,
                   Amount   = &&amount1
                   )

                %mult_assoc (
                   Run      = &run ,
                   Type     = &&type&t ,
                   Site     = &&site&s ,
                   Amount   = &&amount1
                   )

  %end ;
/*      proc datasets lib=work kill memtype=data nolist ; run;*/
   %end ;
  %end ;

  %mend loop ;


  %let site1 = NAT ;

  %macro utility2;

    %do ii = 1 %to 1;

    data CVp&site1;
    set here.cvpers5rec_ira;
    run;

    data CVh&site1.;
    set here.cvhous5rec_ira;
    run;

    data rv1p&site1._&amount1;
    set here.rv1pnat_&amount1;
```

```
    run;

    data rv1h&site1._&amount1;
    set here.rv1hnat_&amount1;
    run;

    %loop;
    %end;

  %mend;

  %utility2;

  title4 Tables with cell quantiles - Cell Quantile Diffs ;

  Proc summary data = here.CellMedianDiffs nway ;
    Where ProcRun = '1';
    Class TestSite  Amt GeoArea Attr byvar;
    Var p1 p25 p50 p75 p99;
    Output out = here.Tab1CQR50 ( drop = _: ) median =;
  Run;

  data here.Tab1CQR50;
  set here.Tab1CQR50;
  iqr = p75-p25;
  run;

  proc print data = here.Tab1CQR50;
  var TESTSITE AMT GEOAREA ATTR BYVAR p50 iqr;
  title "Tab1CQR50";
  run;

  Proc summary data = here.CellP75Diffs nway ;
   Where ProcRun = '1';
   Class TestSite  Amt GeoArea Attr byvar;
   Var p1 p25 p50 p75 p99;
   Output out = here.Tab1CQR75 ( drop = _: ) median =;
  Run;

  data here.Tab1CQR75;
  set here.Tab1CQR75;
  iqr = p75-p25;
  run;

  proc print data = here.Tab1CQR75;
  var TESTSITE AMT GEOAREA ATTR BYVAR p50 iqr;
  title "Tab1CQR75";
  run;

  title4 Tables with cell means - Cell mean Diffs ;

  Proc summary data = here.CellMeanDiffs nway ;
    Where ProcRun = '1';
    Class TestSite  Amt GeoArea Attr byvar;
    Var p1 p25 p50 p75 p99;
    Output out = here.Tab1CMR ( drop = _: ) median =;
  Run;
```

```
data here.Tab1CMR;
set here.Tab1CMR;
iqr = p75-p25;
run;

proc print data = here.Tab1CMR;
var TESTSITE AMT GEOAREA ATTR BYVAR p50 iqr;
title "Tab1CMR";
run;

title4 Tables with cell means - Standard error Diffs ;

Proc summary data = here.SERDiffs nway ;
  Where ProcRun = '1';
  Class TestSite  Amt GeoArea Attr byvar;
  Var p1 p25 p50 p75 p99;
  Output out = here.Tab1SER ( drop = _: ) median =;
Run;

data here.Tab1SER;
set here.Tab1SER;
IQR = p75-p25;
run;

proc print data = here.Tab1SER;
var TESTSITE AMT GEOAREA ATTR BYVAR p50 IQR;
title "Tab1SER";
run;

title4 Tables with weighted counts - Cramer's V Diffs ;

Proc summary data = here.CramerVDiffs nway;
  Where ProcRun = '1';
  Class TestSite  Amt GeoArea TableVar ;
  Var p10 p25 p50 p75 p90;
  Output out = here.Tab1CrV ( drop = _: ) median =;
Run;

data here.Tab1CrV;
set here.Tab1CrV;
IQR = p75-p25;
run;

proc print data = here.Tab1CrV;
var TESTSITE AMT GEOAREA TABLEVAR p50 iqr;
title "Tab1CrV";
run;

title4 Association between variables - Pairwise associations ;

proc print data = here.Associations;
title "Associations";
run;


title4 Association between variables - Multivariate associations ;
```

```
  Proc summary data = here.UStat nway ;
    Class TestSite  Amt ProcRun;
    Var U;
    Output out = Tab1U ( drop = _: ) median =;
  Run;

  proc transpose data = Tab1U out = here.Tab1U ( drop = _: ) prefix=U_Run_ ;
    by TestSite  Amt ;
    id ProcRun ;
    var U;
  run ;

  proc print data = here.Tab1U;
  title "Tab1U";
  run;

%MEND UTILITY;
```

RISK.SAS

```
/*************************************************************************************************************;
**                                                                                                            **;
**  Program Name: RISK.SAS              --  EXAMPLE                                                          **;
**                                                                                                            **;
**  Purpose: Performs Disclosure Risk analysis.                                                               **;
**                                                                                                            **;
**  Risk Description:                                                                                         **;
**                                                                                                            **;
**   Risk measures have been 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. The research team and the           **;
**   Census DRB recognize that combinations of just a few variables can lead to a single sample              **;
**   unit (sometimes referred to as a sample unique or singleton). The DRB has set up rules to               **;
**   reduce the risks associated with small cells. We consider additional 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 our perturbed          **;
**   CTPP data.                                                                                               **;
**                                                                                                            **;
**   The general approach is to bring together measures of various risk elements, including a                **;
**   measure of the amount of changed information. The measures were found acceptable by the DRB.            **;
**   While these risk components can be looked at separately, we build up a series of factors and             **;
**   therefore consider the product of the following risk components to quantify the overall risk             **;
**   as a score.                                                                                              **;
**                                                                                                            **;
**  Called by:  CTPP_MAIN_DRIVER.SAS                                                                          **;
**                                                                                                            **;
**  Called Programs: None                                                                                     **;
**                                                                                                            **;
**  Input Files:  RISKTAB.DAT                                                                                 **;
**                V1HNAT_PARTIAL.SAS7BDAT                                                                     **;
**                CVPERS5REC_IRA.SAS7BDAT                                                                     **;
**                RV1HNAT_PARTIAL.SAS7BDAT                                                                    **;
**                RV1PNAT_PARTIAL.SAS7BDAT                                                                    **;
**                                                                                                            **;
*************************************************************************************************************/;
* Risk.SAS ;
* kt 6/2010 ;

* autocall macros:  cell_mean_ratios cmr_compute ;

%MACRO RISK;

  title CTPP - Disclosure Risk Measures ;

  %let type1 = H ;
  %let type2 = P ;
  %let site1 = NAT ;
  %let amount1 = PARTIAL ;

  filename riskin "&Datadir./risktab.dat" lrecl=80;

  data table1;
  infile riskin truncover;
  input
    @1      attribute  $char14.
    @15   level1         $char1.
```

```
  @18   level2              $char1.
  @21   fullflg             $char14.
  @35   strtflg        $char20.;

if attribute = " " then delete;

run;

data table1;
set table1 end=eof;

cnt + 1;

call symputx('attribute'||trim(left(put(cnt,2.))),attribute);
call symputx('level1'||trim(left(put(cnt,2.))),level1);
call symputx('level2'||trim(left(put(cnt,2.))),level2);
call symputx('fullflg'||trim(left(put(cnt,2.))),fullflg);
call symputx('strtflg'||trim(left(put(cnt,2.))),strtflg);

if eof then call symput('numrecs', cnt);

run;

proc print data = table1; run;

Proc datasets library = here nolist ;
  delete RiskChg RiskChg_flg RiskChg_vio RiskSummary1 RiskSummary2 ;
run ;


%macro loop ;

  %let pnumrecs = &numrecs;

   %let numrecs = &pnumrecs;

 %do s = 1 %to 1 ; /* Nation */

    proc sort data = here.V1h&&site&s.._&amount1
     (keep = cmid st ahinc_strt ahinc_full ahinc_&amount1 ...)
      out = cvh&&site&s;
      by cmid;

    run;

    proc sort data = here.cvpers5rec_ira out=cvp&&site&s;
      by cmid;

    run;

    data cvp&&site&s;
    merge cvp&&site&s (in=aa) cvh&&site&s;
      by cmid;

    if aa then output;

    run;
```

```
   proc sort data = cvp&&site&s
      ( keep = CMID PNUM st RepW0 FLOW_FLG ahinc_full ahinc_strt …
      %do i = 1 %to &numrecs;

        &&attribute&i

      %end;
      rename = (

        %do i = 1 %to &numrecs ;

              &&attribute&i = ACS&&attribute&i

         %end ;
        ) )
      out = cvp&&site&s ;
             by  CMID PNUM ;
           run ;

 /*
   proc sort data = here.vhous5rec_ira (rename = (ahinc=acsahinc)) out = cvh&&site&s;
     by cmid;

   run;
*/


      %do run = 1 %to 1 ; /* 1 run for nation */

                  proc sort data = here.RV&run.h&&site&s.._&amount1
                                  (keep = cmid st ahinc_strt ahinc_full ahinc_&amount1 ahinc
                                          …)
                                out = RV&run.h&&site&s.._&amount1;

                    by cmid;

                  run;

                    proc sort data = here.RV&run.p&&site&s.._&amount1
                                  out = RV&run.p&&site&s.._&amount1;

                    by cmid;

                  run;

                  data RV&run.p&&site&s.._&amount1;
                  merge RV&run.p&&site&s.._&amount1 (in=aa) RV&run.h&&site&s.._&amount1;
                    by cmid;

                  if aa then output;
                  run;

       /*
                 data BASE&run.h&&site&s.._&amount1 ;
                 merge RV&run.h&&site&s.._&amount1 cvh&&site&s;
                 by cmid;
```

```
                    If ACSahinc ne ahinc then ahinc_CHG = 1 ;
                    Else ahinc_CHG = 0 ;
                       If ahinc_STRT in ( 1 , 2 ) then Iahinc_STRT = 1 ;
                       Else Iahinc_STRT = 0 ;
                    run;
          */

                    data BASE&run.h&&site&s.._&amount1 ;
                    set RV&run.h&&site&s.._&amount1;
                    If ACSahinc ne ahinc then ahinc_CHG = 1 ;
                    Else ahinc_CHG = 0 ;
                       If ahinc_STRT in ( 1 , 2 ) then Iahinc_STRT = 1 ;
                       Else Iahinc_STRT = 0 ;
                    run;

                     proc sort data = RV&run.p&&site&s.._&amount1
                                ( keep = CMID pnum
                               %do i = 1 %to &numrecs ;

                                 &&attribute&i &&fullflg&i.._&amount1 &&fullflg&i.._STRT

                               %end ;
                                )
                        out = BASE&run.p&&site&s.._&amount1 ;
                        by  CMID PNUM ;

                    run ;

                        data BASE&run.p&&site&s.._&amount1 ;
                              merge BASE&run.p&&site&s.._&amount1 ( in = inr )
                                     cvp&&site&s ;

                                by  CMID PNUM ;
                              if inr ;
                              RiskStrat = max ( of _numeric_ ) ;

                              flagcount = 0;
                              chgcount1 = 0;
                              chgcount2 = 0;
                              viocount  = 0;
                              chgcount3 = 0;
                              iscount   = 0;

                              %do i = 1 %to &numrecs ; /* for each attribute in table 1 */

                               /*      %let fullflag = &&fullflg&i.._&amount1;*/

                                     If ACS&&attribute&i ne &&attribute&i then &&attribute&i.._CHG =
1 ;
                                     Else &&attribute&i.._CHG = 0 ;

                                     RiskStrat = min ( RiskStrat , &&fullflg&i.._STRT ) ;

                                     If &&fullflg&i.._STRT in ( 1 , 2 ) then I&&fullflg&i.._STRT = 1
;
                                     Else I&&fullflg&i.._STRT = 0 ;
```

```
                                    If &&strtflg&i =4 then SI&&strtflg&i = 1 ;
                                    Else SI&&strtflg&i = 0 ;

                                    If &&attribute&i.._CHG = 1 or  SI&&strtflg&i = 1 then
&&attribute&i.._CHG2 = 1 ;
                                    Else &&attribute&i.._CHG2 = 0 ;
                            %end ;

                FlagCount = sum(… jwmn_&amount1, jwd_&amount1, jwmn_&amount1,
                                                                    …);

                        ChgCount1 = sum(… travel_tm12_CHG2, tm_leave5_CHG2, jwmn_CHG2,
                                                                    …);

   ChgCount2 = sum(… travel_tm12_CHG,
                                                tm_leave5_CHG, jwmn_CHG, …
                                                                                    );

              VioCount = sum(… tm_leave5_flg,
                                jwmn_flg, travel_tm12_flg, … );

                             ChgCount3 = sum(hh_inc26_flg*hh_inc26_chg, …
                                             tm_leave5_flg*tm_leave5_chg,
                                            jwmn_flg*jwmn_chg, travel_tm12_flg*travel_tm12_chg,
                                            …);

                              ISCount = sum(hh_inc26_flg*SIahinc_strt, …,
                                             tm_leave5_flg*tm_leave5_chg,
                                            jwmn_flg*sijwd_strt, travel_tm12_flg*sijwmn_strt,
                                            …
                                            );

                           F = 1 / RepW0;
                           if f = 1 then do;

                              r3 = 1;

                              if RiskStrat = 1 then r1 = 1;
                              else
                                if RiskStrat = 2 then r1 = 0.5;
                                else
                                  r1 = 0;

                           end;
                           else do;

                             If RiskStrat = 1 then do ; /*singles*/
                                    R1= 1;
                                    if F ^= 1 then R3 = -LOG(F)*(F/(1-F)) ;

                             end ;
                             Else if RiskStrat = 2 then do ; /*doubles */
                                   R1 = 0.5;
                                   if F ^= 1 then R3 = (F/((1-F)**2))*(F*LOG(F)+(1-F)) ;

                             end ;
                             Else do ;
```

```
                    R1 = 0;
                    R3 = F/2;
                end ;

            end;

            R2 = 0.30 ;
            R4_1 = 1-0.34;
            R4_2 = 1-0.42;
            R4_3 = 1 - (0.34 + 0.42 - 0.34*0.42);
            R5A = 1 - ChgCount1/10;
            if VioCount ^= 0 then R5B = 1 - (ChgCount3 + ISCount) / VioCount;
            else
              R5B = 0;

            P1 = /*R1*/ R2* R3* R4_3* R5A;
           /* P2 = R1* R2* R3* R4_3* R5B; */
run ;


proc freq data = BASE&run.p&&site&s.._&amount1 ;
  tables FlagCount*ChgCount1 / missing nocol nopercent;
  title2 'CrossTab of FlagCount by ChgCount1 where Flow_flg = 1';
  where flow_flg = 1;

run;

proc summary data = BASE&run.p&&site&s.._&amount1 missing nway;
  class FlagCount … travel_tm12 tm_leave5 jwmn hh_inc9 …
  ;
  output out = cross_sum (rename = (_freq_=count) drop=_type_);
  where flow_flg = 1 and FlagCount =10 and ChgCount1 = 0;

run;
/*
proc print data = cross_sum ;
  title2 'CrossTab of cases with FlagCount =10 and ChgCount1 = 0 and Flow_flg =
1';


run;
*/
proc freq data = BASE&run.p&&site&s.._&amount1 ;
  tables FlagCount*ChgCount2 / missing nocol nopercent;
  title2 'CrossTab of FlagCount by ChgCount2 where Flow_flg = 1';
  where flow_flg = 1;

run;

proc means data = BASE&run.p&&site&s.._&amount1 min p25 median p75 max;
  where flow_flg = 1;
  var FlagCount ChgCount1 ChgCount2;
  title2 'Proc Means on FlagCount, ChgCount1, and ChgCount2 where Flow_flg =
1';


run;

proc means data = BASE&run.p&&site&s.._&amount1 min p25 median p75 max;
```

```
                        where flow_flg = 1;
                        class st;
                        var FlagCount ChgCount1 ChgCount2;
                        title2 'Proc Means on FlagCount, ChgCount1, and ChgCount2 where Flow_flg = 1,
by state';

                      run;

                      proc freq data = BASE&run.p&&site&s.._&amount1 ;
                        tables FlagCount*ChgCount1 / missing nocol nopercent;
                        title2 'CrossTab of FlagCount by ChgCount1 where Flow_flg = 2';
                        where flow_flg = 2;

                      run;

                      proc freq data = BASE&run.p&&site&s.._&amount1 ;
                        tables FlagCount*ChgCount2 / missing nocol nopercent;
                        title2 'CrossTab of FlagCount by ChgCount2 where Flow_flg = 2';
                        where flow_flg = 2;

                      run;

                      proc means data = BASE&run.p&&site&s.._&amount1 min p25 median p75 max;
                        where flow_flg = 2;
                        var FlagCount ChgCount1 ChgCount2;
                        title2 'Proc Means on FlagCount, ChgCount1, and ChgCount2 where Flow_flg =
2';

                      run;


          %do i = 1 %to &numrecs ; /* for each attribute in table 1 */

          %let fullflag = &&fullflg&i.._&amount1;

          %if &&level1&i =1 %then %do;
          %if &i = 1 %then %let type = H;
          %else %let type = P;


                            proc summary data = BASE&run.p&&site&s.._&amount1 nway;
                              var &fullflag i&&strtflg&i;
                              output out=pflg&run.p&&site&s.._&amount1 sum=;
                              where &&strtflg&i in (1,2);

                            run;

                            proc summary data = BASE&run.p&&site&s.._&amount1 nway;
                                   var &&attribute&i.._CHG &&fullflg&i.._&amount1 ;
                                   output out=Pchg&run.p&&site&s.._&amount1 sum=;
                                   where &fullflag = 1 and &&strtflg&i in (1,2);
                            run;

                            Data Pflg&run.p&&site&s.._&amount1.&&attribute&i;
                              Set Pflg&run.p&&site&s.._&amount1;
                              length
                                 ProcRun   $ 2
```

```
                              TestSite  $ 3
                              Amt       $ 8
                        Attr      $ 12
                              ;
                              Pflg = &fullflag / i&&strtflg&i; /* percent flagged for replacement
in stratum 1 and 2 */

                              Attr = "&&attribute&i" ;
                              ProcRun = "&run" ;
                              TestSite = "&&site&s" ;
                              Amt = "&amount1" ;
                              Keep Pflg Attr ProcRun TestSite Amt;
                             Run;

                        proc append base = here.RiskChg
                           data = Pflg&run.p&&site&s.._&amount1.&&attribute&i;
                        run ;


                        Data Pchg&run.p&&site&s.._&amount1.&&attribute&i;
                        Set Pchg&run.p&&site&s.._&amount1;
                        length
                          ProcRun   $ 2
                          TestSite  $ 3
                          Amt       $ 8
                          Attr      $ 12
                        ;
                        Pchg_flg = &&attribute&i.._chg / &fullflag; /* percent flagged for replacement
in stratum 1 and 2 */

                        Attr = "&&attribute&i" ;
                        ProcRun = "&run" ;
                        TestSite = "&&site&s" ;
                        Amt = "&amount1" ;
                        Keep Pchg_flg Attr ProcRun TestSite Amt;

                        Run;

                        proc append base = here.RiskChg_flg
                           data = Pchg&run.p&&site&s.._&amount1.&&attribute&i;
                        run ;
                    %end;

                    %if &&attribute&i ne AHINC and &&attribute&i ne JWD %then %do;
                        proc summary data = BASE&run.p&&site&s.._&amount1 nway;
                          var &&attribute&i.._chg &&attribute&i.._chg2 %if &&attribute&i = industry8
%then %do;
                         industry8_flg
                         %end;
                         %else %do;
                        &&attribute&i.._flg
                         %end;
                         si&&strtflg&i;
                          output out=Pvio&run.p&&site&s.._&amount1 sum=;
                         %if &&attribute&i = … %then %do;
                         where …_flg = 1;
                         %end;
```

```
 %else %do;
 where &&attribute&i.._flg = 1;
 %end;
run;

Data Pvio&run.p&&site&s.._&amount1.&&attribute&i;
Set Pvio&run.p&&site&s.._&amount1;
length
  ProcRun   $ 2
  TestSite  $ 3
  Amt       $ 8
  Attr      $ 12
;

Attr = "&&Attribute&i";
ProcRun = "&run" ;
TestSite = "&&site&s" ;
Amt = "&amount1" ;
Pchg_vio1 = &&attribute&i.._chg /  &&attribute&i.._flg ;
Pchg_vio2 = &&attribute&i.._chg / (&&attribute&i.._flg
  - (&&attribute&i.._chg2-&&attribute&i.._chg) );
Keep ProcRun TestSite Amt Pchg_vio1 Pchg_vio2 Attr;
Run;

proc append base = here.RiskChg_vio
           data = Pvio&run.p&&site&s.._&amount1.&&attribute&i ;
run ;

  %end ;
%end;
%if &&site&s = OLY and &amount1 = PART2 %then %do ;

       proc print data = BASE&run.p&&site&s.._&amount1 ( obs = 20) ;
              var r: f i: val: p: ;
       run ;

%end ;

Data Summary2&run.p&&site&s.._&amount1;
Set Base&run.p&&site&s.._&amount1;
  length

     ProcRun   $ 2
     TestSite  $ 3
     Amt       $ 8
  ;

  ProcRun = "&run" ;
  TestSite = "&&site&s" ;
  Amt = "&amount1" ;
  Keep ProcRun TestSite Amt R5B RiskStrat;
Run;

proc append base = here.RiskSummary2
  data = Summary2&run.p&&site&s.._&amount1;

run ;
```

```
                    Data Summary1&run.p&&site&s.._&amount1;
                    Set Base&run.p&&site&s.._&amount1;
                    length

                      ProcRun   $ 2
                      TestSite  $ 3
                      Amt       $ 8
                    ;
                    ProcRun = "&run" ;
                    TestSite = "&&site&s" ;
                    Amt = "&amount1" ;
                    Keep ProcRun TestSite Amt p1 R5A RiskStrat;
                    where flow_flg in (1,2);
                    Run;

                    proc append base = here.RiskSummary1
                   data = Summary1&run.p&&site&s.._&amount1;

                    run ;


            %end ;

   proc datasets lib=work kill memtype=data nolist ; run;
%end ;


%mend loop ;

%loop

Proc summary data = here.RiskSummary1 nway;
 Where ProcRun = '1';
 Class TestSite Amt RiskStrat;
 Var P1;
 Output out = here.Tab1CMR ( drop = _type_ _freq_) median  = P1_MED
                                                   max     = P1_MAX;
Run;

Proc summary data = here.RiskSummary1 nway;
 Where ProcRun = '1';
 Class TestSite Amt RiskStrat;
 Var R5A;
 Output out = here.Tab3CMR ( drop = _type_ _freq_) median  = R5A_MED
                                                   max     = R5A_MAX;
Run;

proc print data = here.Tab1CMR;
title2 "Tab1CMR";
run;

proc print data = here.Tab3CMR;
title2 "Tab3CMR";
run;

/*
```

```
  proc export data=here.Tab3CMR
     outfile='r:\data\Risk\risk.xls'
     dbms=excel replace;
     sheet=Tab3CMR ;
  run;


  Proc GLM data = here.RiskSummary1;
   Where RiskStrat in ( 1 , 2 ) ;
   Class TestSite  Amt;
   Model P1 = TestSite  Amt;
  Run;

  Proc summary data = here.RiskSummary2 nway;
   Where ProcRun = '1';
   Class TestSite  Amt RiskStrat;
   Var R5B;
   Output out = here.Tab2CMR ( drop = _type_ _freq_) median  = R5B_MED
                                                     max     = R5B_MAX;
  Run;

  proc print data = here.Tab2CMR;
  title "Tab2CMR";
  run;
  */
  proc print data = here.riskchg_flg;
  title2 "riskchg_flg";
  run;

  proc print data = here.riskchg_vio;
  title2 "riskchg_vio";
  run;

  proc print data = here.riskchg;
  title2 "riskchg";
  run;

%MEND RISK;
```

CLEANUP.SAS

```
/*************************************************************************************************************;
**                                                                                                            **;
**  Program Name:  CLEANUP.SAS                          --EXAMPLE                                             **;
**                                                                                                            **;
**  Purpose:  Performs Cleanup and creates delivery files                                                     **;
**                                                                                                            **;
**  Cleanup Description:                                                                                      **;
**                                                                                                            **;
**  This version of the cleanup program creates the delivery files after the processes of                     **;
**  initial risk analysis, perturbation, raking, risk, and utility are finished. The final files              **;
**  at the household and person levels will include the ID variables, the perturbed variables                 **;
**  and their recodes which will be used in Set B tables.                                                     **;
**                                                                                                            **;
**                                                                                                            **;
**  Called by:  CTPP_MAIN_DRIVER.SAS                                                                          **;
**                                                                                                            **;
**  Called Programs: None                                                                                     **;
**                                                                                                            **;
**  Input Files:  RV1HNAT_PARTIAL.SAS7BDAT                                                                    **;
**                VHOUS5REC.SAS7BDAT                                                                          **;
**                RV1PNAT_PARTIAL.SAS7BDAT                                                                    **;
**                CVPERS5REC_IRA.SAS7BDAT                                                                     **;
**                                                                                                            **;
**  Output Files:  PERT_VHOUS5.SAS7BDAT                                                                       **;
**                 PERT_VPERS5.SAS7BDAT                                                                       **;
**                                                                                                            **;
*************************************************************************************************************/;
%MACRO CLEANUP;

title "CTPP - Clean up";

data here.pert_vhous5;
set here.rv1hnat_partial (keep = cmid …);

run;

proc freq data = here.pert_vhous5;
  tables … /list missing;
  title2 'Data Check 1: Frequency on … using file pert_vhous5';

run;

proc freq data = here.vhous5rec_ira;
  tables … /list missing;
  title2 'Data Check 2: Frequency on … using file vpers5rec_ira';

run;

proc means data = here.pert_vhous5 n nmiss min p25 p50 p75 max mean;
  var …;
  title2 'Data Check 3: Univariate on … using file PERT_VHOUS5';

run;

proc means data = here.vhous5rec_ira n nmiss min p25 p50 p75 max mean;
  var …;
```

```
  title2 'Data Check 4: Univariate on … using file vhous5rec_ira';

run;

data here.pert_vpers5;
set here.rv1pnat_partial (keep = CMID PNUM JWMN …);

run;

proc freq data = here.pert_vpers5;
  tables TM_LEAVE5*TM_LEAVE10*TM_LEAVE17
         … /list missing;
  title2 'Data Check 5: Proc Frequency on variables from file PERT_VPERS5';

run;

proc means data = here.pert_vpers5 min max mean median;
  class travel_tm12;
  var jwmn;
  title2 'Data Check 6: Proc Means on JWMN by TRAVEL_TM12 using file PERT_VPERS5';

run;

proc freq data = here.cvpers5rec_ira;
  tables TM_LEAVE5*TM_LEAVE10*TM_LEAVE17
         … /list missing;
  title2 'Data Check 7: Proc Frequency on variables from file cvpers5rec_ira';

run;

proc means data = here.cvpers5rec_ira min max mean median;
  class travel_tm12;
  var jwmn;
  title2 'Data Check 8: Proc Means on JWMN by TRAVEL_TM12 using file cvpers5rec_ira';

run;

proc contents data = here.cvhous5rec_ira;
  title2 'Contents on file cvhous5rec_ira';

run;

proc contents data = here.cvpers5rec_ira;
  title2 'Contents on file cvpers5rec_ira';

run;

proc contents data = here.pert_vhous5;
  title2 'Contents on file PERT_VHOUS5';

run;

proc contents data = here.pert_vpers5;
  title2 'Contents on file PERT_VPERS5';

run;
```

```
%MEND CLEANUP;
```

Appendix U

LIST OF SAS PROGRAMS

HIERARCHY OF PROGRAMS:

CTPP_DRIVER.SAS
- IRA_MAIN_DRIVER.SAS
  - MAIN_PROG1.SAS
    - CREATE_INPUTS.SAS
    - GENTAZ.SAS
    - ADJUST_POW.SAS
    - CTAZ_HOUS_LEVEL.SAS
    - CTAZ_PERS_LEVEL.SAS
    - SUMCHECK.SAS
  - M30_STEP1.SAS
  - M30.SAS
    - SWAP_FLAG.SAS
    - NEW_CEN_MERGE.SAS
    - M30_STEP2_1.SAS
    - M30_STEP3.SAS
      - FLGCELL3.SAS
        - GETCELL2.SAS
    - M30_VIOLATION_PER.SAS
    - M30_SINGLETON.SAS
      - FLGCELL4.SAS -
        - GETCELL2.SAS
  - M30_MAIN_HOUS.SAS
    - M30_STEP2_2.SAS
    - SWAP_FLAG_HOUS.SAS
    - NEW_CEN_MERGE_HH2.SAS
    - M30_VHOUS_PREP_1A.SAS
    - M30_VHOUS_PREP_3.SAS
      - GETCELL2.SAS
    - M30_VHOUS_PREP_4.SAS
    - M30_VHOUS_PREP_5.SAS
      - GETCELL2.SAS
- DATA_REPLACEMENT.SAS
  - PARTIAL_FLAGS_NEW.SAS
  - VARIABLE_PREP.SAS
    - SETVARS.SAS
    - SETZERO.SAS
    - INDICATOR.SAS
    - HHINTERACTION.SAS
    - PERSONINTERACTION.SAS
    - FINALIZE_PREDPOOL.SAS
  - MODELING_STEPS
    - MODEL_SELECTION.SAS
      - PREDPOOL.SAS
  - MAIN_LOOP.SAS
    - CONSTRAINEDHOTDECK.SAS
      - PROCBIN.SAS
      - CHDLLOOP.SAS
    - INDICATOR.SAS
    - PERSONINTERACTION.SAS
    - HHINTERACTION.SAS

- SEMI_PARA.SAS
    - PREDICTION.SAS
    - HOTDECK.SAS
    - FASTCLUS.SAS
    - SYNTHESIZE_OCUC.SAS
- RECODE.SAS
    - RECODEP.SAS
    - RECODEH.SAS
- PRE_POST_CHECKS.SAS
    - POST_FREQ.SAS
    - POST_MEAN.SAS
- CHANGE_SUMMARY.SAS
- RECODES.SAS
    - RECODEP.SAS
    - RECODEH.SAS

- RAKING_DRIVER.SAS
    - CONTROLTOTALS_HH.SAS
        - RECODEH.SAS
        - SUMM.SAS
            - RAKE.SAS
    - CONTROLTOTALS_PERS.SAS
        - RECODEP.SAS
        - SUMM.SAS
            - RAKE.SAS
    - RAKING_HH.SAS
        - SUMM.SAS
            - RAKE.SAS
        - CONV_REPORT.SAS
        - COMPTOTALS.SAS
    - RAKING_PERS.SAS
        - RECODE_JWMN4.SAS
        - SUMM.SAS
            - RAKE.SAS
        - CONV_REPORT.SAS
        - COMPTOTALS.SAS

- UTILITY.SAS
    - ACS_CMR.SAS
    - ACS_CQR.SAS
    - ACS_CMSE.SAS
        - CMSE_COMPUTE.SAS
    - CRV.SAS
        - CRV_COMPUTE.SAS
    - ACS_ASSOC.SAS
    - CELL_MEAN_RATIOS.SAS
        - CMR_COMPUTE.SAS
    - CELL_QUANTILE_RATIOS.SAS
        - CQR_COMPUTE.SAS
    - PERT_CMSE.SAS
        - CMSE_COMPUTE.SAS
    - PAIR_ASSOC.SAS
    - MULT_ASSOC.SAS

- RISK.SAS
- CLEANUP.SAS

Figure U-1. Hierarchical List of Programs by Major Component

Macro Name	Sub System	Function
acs_assoc.sas	Utility	Creates multivariate associations
acs_cmr.sas	Utility	Creates ACS cell means
acs_cmse.sas	Utility	Creates ACS standard errors
acs_cqr.sas	Utility	Creates ACS cell quantiles
adjust_pow.sas	Initial Risk Analysis	Merge place of work PUMA code onto person file
cell_mean_ratios.sas	Utility	Creates cell mean differences
cell_quantile_ratios.sas	Utility	Creates cell quantile differences
change_summary.sas	Data Replacement	Provides a summary of changed values in the household and person level files
chdlloop.sas	Data Replacement	Performs main steps of the constrained hotdeck approach
cleanup.sas	CTPP Main	Cleanup and creation of delivery files
cmr_compute.sas	Utility	Computes cell means
cmse_compute.sas	Utility	Computes cell standard errors
comptotals.sas	Raking	Checks the difference between control totals and sample totals after raking
constrainedhotdeck.sas	Data Replacement	Performs data replacement using the constrained hotdeck approach
controltotals_hh.sas	Raking	Creates control total files at household level
controltotals_pers.sas	Raking	Creates control total files at person level
conv_report.sas	Raking	Creates convergence report in the raking process
cqr_compute.sas	Utility	Computes cell quantiles
create_inputs.sas	Initial Risk Analysis	Create input files for CTAZ creation
crv.sas	Utility	Creates Cramer's V differences
crv_compute.sas	Utility	Computes Cramer's V
ctaz_hous_level.sas	Initial Risk Analysis	Creates CTAZs in household level file
ctaz_pers_level.sas	Initial Risk Analysis	Creates CTAZs in person level file
data_replacement.sas	CTPP Main	Driver program for data replacement
fastclus.sas	Data Replacement	Creates clusters for UC variables
finalize_predpool.sas	Data Replacement	Finalizes the pool of predictors for modeling
flgcell3.sas	Initial Risk Analysis	Calls Macro GETCELL2 to create violation flags (threshold=3)
flgcell4.sas	Initial Risk Analysis	Calls Macro GETCELL2 to create violation flags (threshold=2)
gentaz.sas	Initial Risk Analysis	Macro that assigns CTAZs
getcell2.sas	Initial Risk Analysis	Marco that generates violation flags
hhinteraction.sas	Data Replacement	Creates household-level interaction terms
hotdeck.sas	Data Replacement	Creates hot deck cells
indicator.sas	Data Replacement	Creates indicator variables for UC variables
ira_main_driver.sas	CTPP Main	Driver program for Initial Risk Analysis
m30.sas	Initial Risk Analysis	Sub-Driver program for person level Initial Risk Analysis
m30_main_hous.sas	Initial Risk Analysis	Sub-Driver program for household level Initial Risk Analysis
m30_singleton.sas	Initial Risk Analysis	Creates person level singleton and stratum flags
m30_step1.sas	Initial Risk Analysis	Creates household and person subset files
m30_step2_1.sas	Initial Risk Analysis	Creates person level recode variables
m30_step2_2.sas	Initial Risk Analysis	Create household level recode variables
m30_vhous_prep_1a.sas	Initial Risk Analysis	Merges person level flag variables ontp household level file
m30_vhous_prep_3.sas	Initial Risk Analysis	Creates household level violation flags
m30_vhous_prep_4.sas	Initial Risk Analysis	Creates household level full and replacemant flags
m30_vhous_prep_5.sas	Initial Risk Analysis	Creates household level singleton and stratum flags
m30_violation_pers.sas	Initial Risk Analysis	Creates person level full and replacement flags
main_loop.sas	Data Replacement	Main loop to perform data replacement
main_prog1.sas	Initial Risk Analysis	Creates CTAZs and CTAZ level covariates at person level
model_selection.sas	Data Replacement	Performs stepwise model selection for the semi-parametric method

Macro Name	Sub System	Function
modeling_steps.sas	Data Replacement	Runs all of the modeling steps needed for the semi-parametric approach
mult_assoc.sas	Utility	Creates multivariate associations
new_cen_merge.sas	Initial Risk Analysis	Merges census block level predictos to the person file
new_cen_merge_hh2.sas	Initial Risk Analysis	Merges census block level predictos to the household file
pair_assoc.sas	Utility	Creates pairwise associations
partial_flags_new.sas	Data Replacement	Creates partial flags
personinteraction.sas	Data Replacement	Creates person-level interaction terms
pert_cmse.sas	Utility	Creates standard error differences
post_freq.sas	Data Replacement	Checks the frequencies of the perturbed variables
post_mean.sas	Data Replacement	Checks the distributions of the perturbed variables
pre_post_checks.sas	Data Replacement	Performs pre and post replacement data checks
prediction.sas	Data Replacement	Computes the predicted values of the dependent variables in the models
predpool.sas	Data Replacement	Creates a pool of variables as the predictors
procbin.sas	Data Replacement	Creates bin variables for constrained hotdeck approach
rake.sas	Raking	Performs raking
raking_driver.sas	CTPP Main	Driver program for raking
raking_hh.sas	Raking	Performs household level raking
raking_pers.sas	Raking	Performs person level raking
recode.sas	Data Replacement	Creates recode variables NEW_JWMN, JWD_SHIFT, and NEW_POVERTY
recode_jwmn4.sas	Raking	Creates recode variable JWMN4
recodeh.sas	Data Replacement & Raking	Creates household level recodes
recodep.sas	Data Replacement & Raking	Creates person level recodes
recodes.sas	Data Replacement	Creates recodes
risk.sas	CTPP Main	Risk analysis program
semi_para.sas	Data Replacement	Performs data replacement using the semi-parametric approach
setbins.sas	Data Replacement	Creates bin variables for constrained hotdeck approach
setvars.sas	Data Replacement	Creates ACS versions of variables
setzero.sas	Data Replacement	Sets missing values to zero
sumcheck.sas	Initial Risk Analysis	Performs checks after CTAZs are created
summ.sas	Raking	Computes control totals or sample totals
swap_flag.sas	Initial Risk Analysis	Merges swap flags and GQ change flags to person file
swap_flag_hous.sas	Initial Risk Analysis	Merges swap flags and GQ change flags to household file
synthesize_ocuc.sas	Data Replacement	Synthesizes OC or UC variables
utility.sas	CTPP Main	Driver program for utility analysis
variable_prep.sas	Data Replacement	Prepares the list of predictors for semi-paramtric approach

Figure U-2. List of Programs in Alphabetical Order