Disclosure Avoidance Techniques to Improve ACS Data Availability for Transportation Planners

Requested by:

American Association of State Highway and Transportation Officials (AASHTO)

Standing Committee on Planning

Prepared by:

Cambridge Systematics, Inc.

Stephen E. Fienberg, PhD

Tanzy M.T.P. Love, PhD

May 2009

The information contained in this report was prepared as part of NCHRP Project 08-36, Task 71, National Cooperative Highway Research Program, Transportation Research Board.

Acknowledgments

This study was requested by the American Association of State Highway and Transportation Officials (AASHTO) and conducted as part of National Cooperative Highway Research Program (NCHRP) Project 08-36. The NCHRP is supported by annual voluntary contributions from the state departments of transportation. Project 08-36 is intended to fund quick response studies on behalf of the AASHTO Standing Committee on Planning. The report was prepared by Cambridge Systematics, Inc., Dr. Stephen Fienberg, and Dr. Tanzy M.T.P. Love. The project was managed by Lori Sundstrom and Nanda Srinivasan of NCHRP. The authors thank the technical advisory panel members, Elaine Murakami, Jonette Kreideweis, and Celia Boertlein, for their thoughtful and informed reviews.

Disclaimer

The opinions and conclusions expressed or implied are those of the researchers that performed the research, and are not necessarily those of the Transportation Research Board or its sponsors. The information contained in this document was taken directly from the submission of the authors. This document is not a report of the Transportation Research Board or of the National Research Council.

Technical Report Documentation Page

1. Report No. NCHRP 08-36, Task 71	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Disclosure Avoidance Techniques to Improve ACS Data Availability for Transportation Planners		5. Report Date May 2009 6. Performing Organization Code
7. Author(s) Kevin Tierney, Stephen E. Fienberg, and Tanzy M.T.P. Love		8. Performing Organization Report No. 7315-071
9. Performing Organization Name and Address Cambridge Systematics, Inc. 100 CambridgePark Drive, Suite 400 Cambridge, MA 02140		10. Work Unit No. (TRAIS) 11. Contract or Grant No.
12. Sponsoring Organization Name and Address National Cooperative Highway Research Program National Academy of Sciences 500 Fifth Street, NW Washington, DC 20001		13. Type of Report and Period Covered 14. Sponsoring Agency Code
15. Supplementary Notes		

16. Abstract

Data dissemination rules initiated by the DRB caused significant loss of data for CTPP 2000, and are expected to cause an even more serious loss when applied to products emanating from the American Community Survey. This report summarizes the development and testing of synthetic data techniques that the Census Bureau could employ to provide alternative data, particularly in regards to journey-to-work flow data at small geography from the Decennial Census (or from the 5-year accumulation of the American Community Survey). Multi-year ACS data tables are inherently protected from potential harmful disclosure of participants, and therefore do not require further disclosure restrictions. However, if disclosure-proofing is nevertheless deemed necessary, using a form of data synthesis is probably the best approach from the viewpoint of the transportation planning data user. This research examines alternative data synthesis approaches that could be employed for multi-year ACS journey-to-work tables, including iterative proportional fitting, combined Bayesian / iterative proportional fitting, and the Generalized Shuttle Algorithm.

17. Key Words ACS, Disclosure Avoidance, CTPP		18. Distribution Statement No restrictions.	
19. Security Classification (of this report) Unclassified.	20. Security Classification (of this page) Unclassified.	21. No. of Pages 60	22. Price NA

Table of Contents

List of Tables

List of Figures

1.0 Introduction and Summary

Census data are among the most often used public data sources in state and metropolitan transportation planning applications. Census Transportation Planning Product (CTPP) 2000, and the continuing suite of products from the American Community Survey (ACS) are vital for current and continuing transportation planning analyses at all states and MPOs. The applications and uses of these data for analyses of small geographic areas are many and varied[1].

Of all the tables in CTPP 2000, perhaps the most vital ones are the journey-to-work flow summaries by means of transportation cross-tabulated with other household characteristics, such as household income categories. These tables are provided at small level geography, and therefore often reflect very small numbers at the small area level of reporting. Due to changes in technology, and the continued concern for loss of privacy, the Census Bureau instituted a Disclosure Review Board (DRB) which reviews all tables before release to ensure confidentiality of responses. However, data dissemination rules initiated by the DRB caused significant loss of data for CTPP 2000, and are expected to cause an even more serious loss when applied to products emanating from the American Community Survey.

This report summarizes the development and testing of synthetic data techniques that the Census Bureau could employ to provide alternative data, particularly in regards to journey-to-work flow data at small geography from the Decennial Census (or from the 5-year accumulation of the American Community Survey).

In the remainder of this section, we provide an overview of a range of statistical data protection procedures that agencies employ to safeguard the confidentiality of individuals represented within their databases. We then discuss the specific procedures used by the Census Bureau for the 2000 CTPP data release and the anticipated procedures for the upcoming ACS 5-year data releases. This is followed by a discussion of how data users have been affected by the implementation of the disclosure avoidance procedures. Section 1 concludes with a brief summary of the research that was performed for this study.

Sections 2 through 4 then describe in more detail some potential techniques that could be employed to synthesize CTPP data, making it more useful to data users while maintaining data confidentiality. These sections describe the development and testing of the techniques.

---

[1] Please see a series of applications developed by local planners and models posted at http://www.fhwa.dot.gov/ctpp/srindex.htm, accessed on October 1, 2007.

1.1 STATISTICAL DATA PROTECTION

Many public agencies throughout the world, including the United States Census Bureau, collect and disseminate statistical data that could be used to expose, and ultimately harm, the individual people or entities that provide the data to the agencies. Agencies must balance the value of providing complete and unbiased databases to decision-makers and other data users against the very real potential that nefarious "intruders" could use the disseminated data to learn private information about specific respondents to the intruders' advantage and to the data respondents' disadvantage. The exposure of individuals or entities could harm those individuals and entities directly by providing dishonest or self-interested people and competitors with information that they should not have. In addition, the exposure could also weaken the overall usefulness of the data collection program, as a whole, because these exposures decrease the willingness of other individuals and entities to participate in the data collection effort in the future.

Statisticians interested in the fields of statistical disclosure control and protection have broadly classified disclosure limiting strategies into:

- Suppressions, in which some data that are more singular or identifiable are not provided to users;
- Recodings, in which cases and / or attributes are collapsed or swapped;
- Samplings, in which only a subset of the data of interest are provided to data users;
- Simulations, in which actual observed data are replaced by "pseudo-data."[2]

For tabular data releases, these disclosure limitation strategies are often accomplished through a few different techniques, including:

1. Rounding;
2. Cell suppression; and
3. Data swapping.[3]

The specific implementation of these different strategies limit disclosure to different degrees. In addition the strategies affect the usability of the data, and potentially bias the results of data analyses in different ways. Some common

---

[2] Stephen E. Fienberg. "Confidentiality and Disclosure Limitation," Encyclopedia of Social Measurement, Volume 1 (2005, Elsevier, Inc.).

[3] Stephen Fienberg and Leon C. R. J. Willenborg. ***Introduction to the Special Issue: Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data.*** Journal of Official Statistics, Vol. 14, No. 4, 1998, p.338.

disclosure limitation methods are simplistically illustrated in the following tables.

Table 1.1 is a representative two-dimensional three by four data table, with the marginal totals shown. This table could have been derived directly from any hypothetical data collection activity that enables the cross-tabulation of the variables of interest.

Because some of the cells have small counts, and therefore are more susceptible to allowing for the identification of individuals that participated in the data collection, the agency might want to employ a disclosure limiting strategy. In actual conditions, the agency would be less concerned about a single cross-tabulation like the example. They would more likely be concerned that intruders could use such a simple cross-tabulation along with other tabulations, microdata, and additional administrative data to learn details about specific individuals.

Table 1.1 An Example Data Table

	$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	
$t_{1,m}$	8	7	5	6	26
$t_{2,m}$	2	9	1	0	12
$t_{3,m}$	1	3	0	8	12
	11	19	6	14	50

Data disseminating agencies use one or more of a variety of different procedures to limit the exposure of individual data records.

Rounding

One option is for the agency to apply rounding to the tables that are provided to data users. In Table 1.2, all counts within Table 1.1 greater than 7 are rounded to the nearest 5, and counts between 1 and 7 are rounded to 4. The cells with zero remain zero. The resulting table makes it more difficult to determine which cells have a single entity or that have a very small number, because any such cells are shown with fours, along with other (presumably safer) cells that have counts up to seven. The rounding of the cells with larger counts to the nearest 5, including the marginal totals, makes it much more difficult for an intruder to isolate small count cells by subtracting the other cells in a row or column.

Table 1.2 Protection of the Example Table Records Through Rounding

	$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	
$t_{1,m}$	10	4	4	4	25
$t_{2,m}$	4	10	4	0	10
$t_{3,m}$	4	4	0	10	10
	10	20	4	15	50

Unfortunately for the data user, this particular rounding approach can change the distribution of the cells and the results of analyses of table cells measurably. In addition, the approach shown here leads to inconsistencies between cell values and marginal totals, so analyses that rely on the combination of cell values and marginal values (such as percentage calculations) are disrupted.

Data Suppression

A second disclosure limiting strategy is to suppress the cells with small count values. In Table 1.3, all the cells from the original table with counts of one to three are suppressed (labeled as "C"), while the other table values are retained.

Table 1.3 Protection of the Example Table Records Through Simple Data Suppression

	$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	
$t_{1,m}$	8	7	5	6	26
$t_{2,m}$	C	9	C	0	12
$t_{3,m}$	C	C	0	8	12
	11	19	6	14	50

Note that this table is still not disclosure-proof, because a user can discern suppressed values by comparing the marginal totals with reported cell values to determine what the suppressed cells would have to be. Consequently, agencies employ "secondary," or "complementary," suppression in addition to the initial suppression, such as that shown in Table 1.4, to make such calculations much less certain. For the example, the additional suppression of the cells labeled "CX" (which would not be distinguished from the cells labeled "C" in the table presentation) makes the identification of small counts much more difficult.

Table 1.4 Protection of the Example Table Records Through Primary and Complementary Data Suppression

	$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	
$t_{1,m}$	8	7	5	6	26
$t_{2,m}$	C	CX	C	0	12
$t_{3,m}$	C	C	CX	8	12
	11	19	6	14	50

Cell suppression, as implemented in the example, allows users to see actual marginal totals (provided that they have marginal total values of more than three), and to see accurate counts of most cells with values of more than three. Without secondary suppression, the users would know that the suppressed cell values are between one and three, but with the secondary suppression, a user could not conclude very much with any certainty about these suppressed cells.

Data Recoding

Another approach to disclosure limitation is to introduce perturbations into the dataset, either randomly or through decision rules. Table 1.5 shows how this approach could be implemented to help shield the more disclosable counts in the example table. In this approach, some cell values are changed before release. The perturbation details, including the decision rules used and the number of cells that are modified are not made known to data users, so it is difficult for them to make individual identifications.

Table 1.5 Protection of the Example Table Records Through Data Recoding (Perturbation)

	$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	
$t_{1,m}$	8	7+1=8	5-1=4	6	26
$t_{2,m}$	2-1=1	9	1+1=2	0	12
$t_{3,m}$	1+1=2	3-1=2	0	8	12
	11	19	6	14	50

This strategy maintains consistency between individual cells and marginal totals, and as long as the perturbation rules (which could be much more sophisticated than in this simple example) are not known by users, the method provides an excellent means to make it difficult for intruders to identify individual data collection participants.

On the other hand, there is no way for data users to assess whether the perturbations have a significant effect on the specific analyses they are conducting, and the accuracy of the analyses will vary depending on which specific cells are subject to perturbation.

Data Synthesis

The final general approach that we will illustrate is data synthesis. In some ways, this approach is an extension of the data recoding approach, but in this approach, no specific cell level counts are reported to users. Instead, a mathematical model is developed based on the actual data, and the model is then applied to all table cells. The users are then given the model outputs, rather than the initial data. Table 1.6 shows a very simple data synthesis approach.

Table 1.6 Protection of the Example Table Records Through Data Synthesis

	$t_{n,1}$	$t_{n,2}$	$t_{n,3}$	$t_{n,4}$	
$t_{1,m}$	50*(11/50)*(26/50)	50*(19/50)*(26/50)	50*(6/50)*(26/50)	50*(14/50)*(26/50)	26
$t_{2,m}$	50*(11/50)*(12/50)	50*(19/50)*(12/50)	50*(6/50)*(12/50)	50*(14/50)*(12/50)	12
$t_{3,m}$	50*(11/50)*(12/50)	50*(19/50)*(12/50)	50*(6/50)*(12/50)	50*(14/50)*(12/50)	12
	11	19	6	14	50

From the data user perspective, the method provides an internally consistent tabulation. It is difficult for the user to ascertain how well the synthesized data fit the actual (but unreleasable) data--the data user is reliant on the agency to maintain the integrity of the underlying data with its synthesis approach--but the accuracy of the synthesized data is not biased by subjective perturbation.

1.2 DISCLOSURE LIMITATION FOR CTPP 2000

For the 2000 Decennial Census, AASHTO worked with the Census Bureau to develop the CTPP 2000, a data product consisting of more than 200 tabulations and cross-tabulations of variables from the Census long form data. The product is designed to provide tables of particular utility for transportation planners at geographic levels of specificity defined by state and local transportation planners to support their specific analyses.

CTPP 2000 consists of 121 tables/cross-tables of variables summarized at the residence of the Census participants (so-called Part 1 tables); 68 tables/cross-tables of variables summarized at the place-of-work of the Census participants (Part 2 tables); and 14 tables/cross-tables of variables summarized for residence-workplace flows. Like other Census Bureau special tabulations, the special tabulations of the Census data in CTPP Part 1 provide transportation planners

greater insights than the standard Summary File 3 (SF3) compilation of long form data because the tables anticipate common transportation planning analyses. In addition, the provision of data summarized at the workplace geography and for worker flows make the CTPP data product a valuable asset for the transportation community and a relatively unique product for the Census Bureau, which tends to provide tabulations primarily by residence geography.

The Census Bureau's Disclosure Review Board applied specific rounding disclosure rules to all specially tabulated Census 2000 products, including CTPP 2000. In addition, the flow tabulations were further suppressed by adding a rule that there needed to be three or more completed Census forms pertaining to any geography definition before data release. Because the long form of the Census is a sample survey, the requirement that only three or more completed Census forms can be released meant that a weighted total of at least 30 to 40 workers had to be present in any origin-destination pair before data release. Table 1.7 summarizes the Census Bureau disclosure rules for CTPP 2000.

Table 1.7 CTPP 2000 Disclosure Review Board Data Dissemination Rules

CTPP Part	Rounding Requirements	Data Suppression/Thresholds
Part 1: At Residence (121 Tables)	All tables rounded: - Zero = 0; - 1 – 7 = 4; - 8 or more=Nearest multiple of 5	No thresholds
Part 2: At Workplace (68 Tables)	All tables rounded: - Zero = 0; - 1 – 7 = 4; - 8 or more=Nearest multiple of 5	No thresholds
Part 3: Worker Flows (14 Tables)	Most, but not all tables rounded: - Zero = 0; - 1 – 7 = 4; - 8 or more=Nearest multiple of 5	Some tables subject to minimum threshold of 3 unweighted records

The rounding and threshold rules set by the Census Bureau severely impacted the utility of several journey-to-work tables for CTPP 2000. In fact, several tables had to be eliminated from the original data request because of the level of restriction introduced by the disclosure limitation. A working paper published by U.S. DOT examined the effect of the rounding rules and thresholds to show the severity of the impact.[4]

As described in the simple example earlier, the imposition of rounding introduced inconsistencies between table cell values and marginal totals. In

[4] Ed Christopher and Nanda Srinivasan, "Disclosure and Utility of Census Journey-to-Work Flow Data from the American Community Survey - Finding the Right Balance." Posted at http://www.fhwa.dot.gov/ctpp/balance.htm.

addition, the rounding caused inconsistencies between different geographic levels of detail, because the combination of rounded data for adjacent small geographic areas within a larger area did not match the larger area's data (which has less rounding).

More than the rounding rule, however, the more problematic aspect of the DRB requirements was the application of thresholds set for the Part 3 Worker Flow tables. As shown in Table 1.8, several key tables were released with the requirement of three unweighted observations for any origin-destination pair.

Table 1.8 CTPP Part 3 Worker Flow Tables

Table	Content	Disclosure Proofing
1	Total Workers (1)	No threshold; rounding only
2	Vehicles available (3 – zero, one, or two+) by Means of Transportation (7 modes)	No threshold; rounding only
3	Poverty Status (3 categories)	3 unweighted records and rounding
4	Minority Status (2 – white non Hispanic and all others)	3 unweighted records and rounding
5	Household Income (8 classifications)	3 unweighted records and rounding
6	Means of Transportation (17 modes)	3 unweighted records and rounding
7	Household Income (4 classifications) by Means of Transportation (4 modes)	3 unweighted records and rounding
8	Mean Travel Time by Means of Transportation to Work (7 modes) and Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding
9	Median Travel Time by Means of Transportation to Work (7 modes) and Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding
10	Aggregate Number of Vehicles by Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding
11	Number of Workers per Vehicle by Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding
12	Aggregate Number of Carpools by Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding
13	Number of Workers per Carpool by Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding
14	Aggregate Travel Time by Means of Transportation to Work (7 modes) and Time Leaving Home for Work (2 – AM peak and all other times)	No threshold; No rounding

CTPP Table 3-01, Total Worker Flows, and CTPP Table 3-02, Vehicles Available per Household (3 vehicle availability levels) by Means of Transportation to Work (7 modes) were released without any record thresholds. CTPP Tables 3-03 to 3-07 were subjected to the rule-of-three threshold. Tables 3-08 to 3-14 were exempt from both rounding and thresholds since they fell under the Census Bureau's "normal" process for reporting aggregates, means, medians and standard deviations.

Applying threshold rules to CTPP 2000 resulted in elimination of at least 80 percent of the data at small geography (Census Tracts or lower) for tables 3-003 to 3-007. As detailed below in the discussion of ACS disclosure, analyses of specific locations performed by FHWA staff and by Cambridge Systematics for the development of the NCHRP ACS Guidebook found significant loss of data with the introduction of thresholds both for CTPP and for ACS test data.

These results were consistently observed across the nation, and planners have regularly complained about the loss of these valuable data at regional workshop settings as well as national conferences[5]. Table 1.9 shows the results of a typical Part 3 table subjected to CTPP threshold rules. The majority of the table was found to be suppressed, leading to complete inability to use the tabulation for any planning or modeling purpose.

[5] See results from a transportation peer exchange held on ACS posted at http://trbcensus.com/SCOP/docs/acs_peer_exchange_may2007.pdf, accessed on October 1, 2007.

Table 1.9 CTPP 2000 Table 3-06 After Suppression (3 Record Threshold) for Downtown Miami Census Tract Flows

MEANS OF TRANSPORTATION (ALL WORKERS)	Total	Drove Alone	2 Person Carpool	3+ Person Carpool	Transit	Other means	Worked at home
FL, 120110101.01-FL, 120860030.01	0	0	0	0	0	0	0
FL, 120110101.02-FL, 120860048.00	25	15	10	0	0	0	0
FL, 120110101.02-FL, 120860067.01	0	0	0	0	0	0	0
FL, 120110102.00-FL, 120860030.04	0	0	0	0	0	0	0
FL, 120110102.00-FL, 120860037.01	0	0	0	0	0	0	0
FL, 120110102.00-FL, 120860048.00	0	0	0	0	0	0	0
FL, 120110103.01-FL, 120860029.00	0	0	0	0	0	0	0
FL, 120110103.01-FL, 120860037.01	0	0	0	0	0	0	0
FL, 120110103.01-FL, 120860048.00	0	0	0	0	0	0	0
FL, 120110103.03-FL, 120860037.01	0	0	0	0	0	0	0
FL, 120110103.03-FL, 120860067.01	0	0	0	0	0	0	0
FL, 120110104.01-FL, 120860024.01	0	0	0	0	0	0	0
FL, 120110104.01-FL, 120860036.01	0	0	0	0	0	0	0
FL, 120110104.01-FL, 120860048.00	0	0	0	0	0	0	0
FL, 120110104.01-FL, 120860049.01	0	0	0	0	0	0	0
FL, 120110104.04-FL, 120860037.01	0	0	0	0	0	0	0
FL, 120110104.04-FL, 120860037.02	0	0	0	0	0	0	0
FL, 120110104.04-FL, 120860048.00	0	0	0	0	0	0	0
FL, 120110104.04-FL, 120860066.01	0	0	0	0	0	0	0
FL, 120110104.04-FL, 120860067.02	0	0	0	0	0	0	0
FL, 120110104.05-FL, 120860017.03	0	0	0	0	0	0	0
FL, 120110104.05-FL, 120860037.01	0	0	0	0	0	0	0
FL, 120110104.05-FL, 120860048.00	0	0	0	0	0	0	0
FL, 120110104.05-FL, 120860050.01	0	0	0	0	0	0	0
FL, 120110105.01-FL, 120860024.01	0	0	0	0	0	0	0
FL, 120110105.01-FL, 120860029.00	0	0	0	0	0	0	0
FL, 120110105.01-FL, 120860030.04	35	35	0	0	0	0	0
FL, 120110105.01-FL, 120860037.01	35	35	0	0	0	0	0
FL, 120110105.01-FL, 120860048.00	30	30	0	0	0	0	0
FL, 120110105.01-FL, 120860062.00	0	0	0	0	0	0	0

1.3 ACS DISCLOSURE LIMITATION

The American Community Survey samples about three million households on an annual basis. Data are collected by mail, and Census Bureau staff follow-up with a sample of those who do not respond according to a regular schedule. The continuous data collection offers the appeal of more current data and better data collection efficiency for the Census Bureau. However, the survey's continuous

methodology "spreads" out the amount of data collection equivalent to the long form data collection to at least 8 years.

On an annual basis, the ACS provides estimates of demographic, housing, social, and economic characteristics for all states, as well as for all cities, counties, metropolitan areas, and population groups of 65,000 people or more. However, for smaller areas such as Census tracts and most places, it will take three to five years to accumulate sufficient sample to produce publishable data. Beginning this year, three-year average ACS data are being published for areas of 20,000 to 65,000. For smaller rural areas and city neighborhoods or population groups of less than 20,000 people, it will take five years to accumulate publishable data, and this accumulated sample will still be substantially smaller than the long form sample of the decennial census.

Likely Effects of ACS Data Disclosure Limitation

Because even the five-year accumulated ACS sample sizes are considerably smaller than the Census long form data, the application of rounding and thresholds will be much larger. Cambridge Systematics investigated the effects of rounding for NCHRP 08-48 "ACS Guidebook" and found a more pronounced effect on the ACS test data for Hampden County, MA (Table 1.10).[6]

Table 1.10 Comparison of the Effects of Disclosure Limitation on Census 2000 and ACS test data – Hampden County

	Part 3: Without Thresholds		Part 3: With Thresholds		Part 1
Data	Total Records	Total Workers	Total Records	Total Workers	Total Workers
Census 2000	8,228	207,120	2,644	147,080	199,220
ACS	6,368	181,563	1,673	118,234	202,024

Further, Christopher and Srinivasan (2005), using Franklin County, OH ACS Test site data as an example (Table 1.11 and Table 1.12) showed the progressive loss of flow interaction data for smaller geographic areas in their U.S. DOT sponsored work[7]. For the CTPP data, the rounding performed with CTPP Table 3-01 led to small differences (0 to 4 percent) in the estimates of workers living and working

[6] As part of work on NCHRP 08-48 "ACS Guidebook", Cambridge Systematics analysis of Part 3 data showed severe loss of data for ACS Special Tabulations for the nine-test sites.

[7] Christopher, Ed, and Srinivasan, Nanda, 2005, "Disclosure and Utility of Census Journey-to-Work Flow Data from the American Community Survey: Is There a Right Balance?" Accessed from http://www.fhwa.dot.gov/ctpp/balance.htm on October 3, 2007.

in Franklin County as geographic detail is increased. The rounding and thresholds applied to CTPP Table 3-06 (cross-tabulated means of transportation and income) had a more significant effect. A little less than a third of the worker counts for CTPP 2000 were suppressed at the Census tract level, and almost two-thirds of the worker flows were suppressed at the traffic analysis zone flow level.

Table 1.11 Effects of Data Suppression on Data Availability for the Franklin County ACS Test Site

Data Product		County-County	Place-Place	Tract-Tract	Zone-Zone
CTPP2000 (Total Workers Living and Working in the County [Census 2000] = 508,393)	Table 3-01 (No Thresholds)	508,395	508,361	500,426	487,979
	Percent Loss	0.00%	0.01%	1.57%	4.02%
	Table 3-06 (Thresholds)	508,395	507,604	358,170	177,643
	Percent Loss	0.00%	0.16%	29.55%	65.06%
ACS (1999, 2000 and 2001) (Total Workers Living and Working in the County [ACS, 3-yr] = 498,220)	Table 3-01 (No Thresholds)	498,220	498,168	447,446	N/A
	Percent Loss	0.00%	0.01%	10.19%	
	Table 3-03 (Thresholds)	498,220	495,840	233,920	N/A
	Percent Loss	0.00%	0.48%	53.05%	

Based on the ACS test data, the imposition of the same disclosure avoidance rules for 5-year accumulated data will lead to significantly less available data. CTPP Table 3-01 would be affected by the rounding rules to a greater degree than the Census 2000 data (10.19 percent loss vs. 1.57 percent loss at the tract-to-tract level), and more importantly, the thresholds on CTPP Table 3-06 would result in 53 percent data loss at the tract-to-tract geographic level, compared to a 29 percent loss for 2000 Census data. Data losses for the TAZ-to-TAZ level were not estimated for the ACS test data, but the loss would be almost total.

Table 1.12 shows the number of origin-destination pairs for which the data were reported in CTPP 2000 and the number of origin-destination pairs that would be reported for the 5-year ACS, assuming the rule-of-three threshold. For the 2000 Census, 20 percent of the place-to-place flows, 71 percent of the tract-to-tract flows, and 89 percent of the block group-to-block group flows were suppressed. Christopher and Srinivasan estimate that for the 5-year ACS 31 percent of place-to-place flows and 82 percent of tract-to-tract flows would be suppressed.

Table 1.12 Summary of OD Pairs Lost Due to Thresholds for Franklin County ACS Test Site Data

	CTPP OD Pairs w/Trips			ACS OD Pairs w/Trips		
Geography	Without Thresholds	With Thresholds	Percent Lost	Without Thresholds	With Thresholds	Percent Lost
Place-Place	384	306	20%	334	229	31%
Tract-Tract	23,289	6,794	71%	13,380	2,459	82%
Block Group – Block Group	44,266	5,045	89%	--	--	--

Inherent ACS Data Disclosure Limitation

Although not yet known, the specific disclosure rules for the American Community Survey (ACS) five-year data releases may be similar to those used for CTPP 2000, and could even be stricter. However, there is reason to believe that this level of disclosure avoidance may not be warranted for the ACS data release.

Because of the ACS survey methodology, sample size, and population dynamics during the period of data collection, the risk of disclosure of individuals by ACS should be far less than for the decennial Census long form data collection effort. Some of the inherent disclosure avoidance features of the ACS data collection effort include:

- Survey Differences:
  - Data Accumulation: ACS data for small geographic areas are accumulated over five years before they are released. The time-period of the data collection for specific records is not available to users in the final data tabulations, so singular or outlying data records in the tabulations could correspond to any of the 60 months in the five year time period. This accumulation period makes the acquisition and detection of specific respondents' data by intruders both more difficult and less useful to intruders.
  - Imputation: Some of the household variables reported in the CTPP tables, such as income, have high levels of non-response, and thus are imputed for the tabulations. In addition, about 25 percent of workplace locations in the Census Long Form were allocated either by a standard allocation process, or by an extended place of work allocation process. These are "educated guesses" about a worker's location based on other observed data, including travel time to work, means of transportation to work, industry, and occupation. In other words, about one-quarter of the ACS workplace and worker flow data already are synthetic. The

imputation makes the comparison of CTPP data to other Census data and third-party databases more difficult and speculative.

  - Income adjustment using CPI: Accumulated ACS data on income and other dollar value variables are adjusted for CPI to the end of the data period. Without access to the time period of data collection for specific household records, these adjustments, along with the fairly dynamic nature of annual incomes, help to mask the actual reported data.
  - Sampling and sample size: Section 4.6 describes the inherent data disclosure protection of tables based on sampling. Small area ACS data accumulated and tabulated over a five year period are likely to be based on substantially fewer raw observations (perhaps as low as one-half) than the traditional Census long form. This means the likelihood that the tabulations will include specific individuals of interest for potential intruders is smaller. As discussed below, it also means that the application of data suppression routines that were previously applied to the larger sample will have a larger effect on the usability of data.
  - Weighting: Because of the multimode data collection effort and the variable subsampling rates used for ACS, the tabulations of ACS results rely on complex weighing schemes that will make it more difficult for intruders to recreate raw responses.

- Population dynamics over five years
  - Changing residence location: Because ACS data for small areas are accumulated over five years, a substantial portion of the individual data records will not reflect current conditions. In the year 2000 Census, about one-half of household residents reported moving into the housing unit for which they were reporting Census data within the previous five years.
  - Changing workplace location: The percentage of workers changing work locations over a five year period is also thought to be approaching one-half of workers. So, by the time small area ACS data are released, the chances of finding specific workers who continue to commute the same way is significantly reduced.
  - Changing means of transportation to work: In addition to varying travel patterns because of changes in origins and destinations, commuting characteristics such as mode of travel, departure and arrival times, and travel times change over time for a large proportion of workers. Workers who take transit, bike, or carpool are likely to change their mode often.[8] In general, users of these traditionally lower market share modes would

[8] Nancy McGuckin and Nanda Srinivasan, Journey to Work Trends Report, Exhibit 1.22. Posted at http://www.fhwa.dot.gov/ctpp/jtw/jtw1.htm, Accessed on October 13, 2006.

be more susceptible to being identified by database intruders, but the impermanence of their mode decisions would impede any identification effort.

- Data Interpretation issues: The ACS user community is yet to understand and employ "period" estimates as opposed to point-in-time estimates. The variability in individual household characteristics measures as obtained from ACS over time are not fully understood, so it will not likely to be possible to accurately model changes, such as those noted above, within a five year period to be able to identify specific survey respondents.

Sections 4.5 through 4.7 further discuss the issues of disclosure protection for sample-derived data tables from a statistical viewpoint.

Given the inherent disclosure protection features in the ACS data, and the sparse sample data in the ACS, it would not be unreasonable for the DRB to consider dropping the disclosure rules altogether for the multiyear data releases or to reduce the threshold. The threshold of at least three records was derived from economic censuses where one firm could benefit from a date release that allowed them to discern information on a competitor. With a minimum threshold of three firms, the firms can not identify another firm singularly. ACS respondents and others who complete demographic surveys in a large area generally cannot benefit from such information, and in any case, are likely to never be able to identify the specific similar people, so the need for thresholding is less. If the geography were increased to tract level or super-tract level, the probability of interaction between individuals is much smaller, so it is hoped that bivariate tables would be available without thresholds or rounding, at least at a larger geography.

1.4 SUMMARY OF THIS RESEARCH

Based on the discussions above, it can be argued that:

- Multi-year ACS data tables are inherently protected from potential harmful disclosure of participants, and therefore do not require further disclosure restrictions; and
- If disclosure-proofing is nevertheless deemed necessary, using a form of data synthesis is probably the best approach from the viewpoint of the transportation planning data user.

The remaining sections of this report summarize research performed to test alternative data synthesis approaches for ACS disclosure avoidance. Section 2 describes the use of iterative proportional fitting methods to synthesize the journey-to-work cross-tabulations. Section 3 describes an extension to the iterative proportional fitting method that incorporates Bayesian methods similar to the ones used by the Census Bureau for the Longitudinal Employer

Household Dynamics (LEHD) program data dissemination. Section 4 describes the use of the Generalized Shuttle Algorithm to synthesize the journey-to-work cross-tabulations.

Iterative Proportional Fitting (IPF) Data Synthesis

Iterative Proportional Fitting (IPF), also known as raking in market research circles or "Frataring" by transportation planners and modelers, is a commonly used mathematical technique to fill matrices based on the availability of marginals. An IPF based approach for synthesizing ACS Journey-to-Work flows was examined because:

- There is a high level of familiarity among the transportation modeling community with this method;
- There is a high level of familiarity among Census Bureau staff with the IPF method; and
- There is an assurance of convergence in IPF given non-zero marginals.

As discussed below, data from Montgomery County, Maryland were used to test the validity of the IPF approach for synthesizing CTPP crosstabulations. The inputs for the test were (at Census tract level):

- Part 1 (residence) Tables by Income (4 categories – less than $25,000, $25,000-$44,999, $45,000-$74,999, more than $75,0000) by Mode (4 categories – Drove alone, Carpooled, Transit, Others);
- Part 2 (workplace) Tables by Income (4 categories – less than $25,000, $25,000-$44,999, $45,000-$74,999, more than $75,0000) by Mode (4 categories – Drove alone, Carpooled, Transit, Others); and
- Part 3 (flow) Tables with worker flows by Mode (4 categories – Drove alone, Carpooled, Transit, Others).

We began by defining "super-tracts" by combining adjacent tracts, and then applied the IPF steps listed in section 2 to the larger geography marginal totals in order to synthesize the more detailed geographic data that would not be available to data users without the synthesis. We compared the results of the synthesis effort to the actual known (but unreportable) table values.

The Montgomery County test confirmed the overall validity of the IPF approach, so the next step was to apply it to another geography to test whether the approach is transferable. For this purpose, Cook County, IL was selected.

These initial tests gave results that were close to the real data, but the model correlations were relatively weak and the Cook County procedure did not converge after multiple iterations. So, further fine tuning was required to the procedure.

In new tests, the super-tract univariate tables for mode (4 categories – Drove alone, Carpooled Transit, All others) and Income (4 categories – Less than

$25,000, $25,000-$44,999, $45,000-$74,999, $75,000 or more) were used as inputs, along with the tract level mode by income tables for residence end (part 1), workplace end (part 2), and tract level worker flow by mode. The tract level inputs were the same as what was used for the initial tests in Montgomery County, Maryland and Cook County, Illinois. The difference between this and the previous tests is that a two phased IPF implementation was used initially to develop a synthetic IPF distribution of mode by income for super-tract-to-super-tract flows.

The procedure first used the super-tract mode and super-tract income univariate tables along with the super-tract level distribution of mode by income to produce a synthesized IPF distribution of mode (4 categories) by income (4 categories) for super-tract-to-super-tract flows. This bivariate table was then used as an input along with the other input tables as in the previous tests. Then the same procedure as used in Montgomery County, Maryland and Cook County, Illinois previously was used to generate tract-to-tract synthetic flows by mode and income.

Further tests were performed with various IPF data synthesis approaches, and it was concluded that the IPF methods were adequate and feasible, but that certain biases in the synthesized data seemed to be present. Consequently, a combined IPF/Bayesian synthesis method was considered.

IPF/Bayesian Data Synthesis

The IPF procedures explored Section 2 relied on using univariate Journey-to-Work Flow tables at super-tract geography. In Section 3, we explored the possibility of combining a Bayesian method along with the IPF. The method generates super-tract level bivariate tables using super-tract level univariate tables where a disclosure threshold of two or more records is met, and using just one implicate where the disclosure threshold is not met.

There is precedence for using Bayesian methods for data synthesis in the Census Bureau, specifically in the Longitudinal Employer Household Dynamics (LEHD) data. Bayesian techniques are used to synthesize workers' place of residence conditional on disclosable counts of workers by place of work, industry, age, and earnings categories.

The combined method was first applied on a representative data set with four residence zones, four workplace zones, and 16 income categories. The results of this test were promising, because the synthetic data were found to be correlated very well with the real data, and because the modeled fit between the real and synthetic data was also very close. Because of these results, the procedure was applied to test cases in Cook County, Illinois and to the Seattle, Washington MSA. Based on the results of the comparison between synthesized data and real flow data from Cook County, IL and Seattle MSA at the Super-tract, tract and TAZ levels, the Bayesian (implicate) plus IPF method performed quite well.

Generalized Shuttle Algorithm

The generalized shuttle algorithm (GSA) is a method for determining bounds for table cell and super-cell counts. There are ***fixed*** super-cells, whose counts are fixed by either empirical observation or assumption, and ***free*** super-cells, whose counts are unobserved or simply unconstrained. In practice, certain marginals are known and used as the fixed super-cells. In general terms, the GSA uses the fixed super-cell counts and the intrinsic dependencies between counts of super-cells to iteratively refine estimates of counts for free super-cells. The algorithm establishes an order for evaluating cell values, and then goes from cell to cell establishing feasible upper and lower bounds. The algorithm will either be halted by encountering an inconsistency between consecutive calculations of bounds for super-cell counts, or will reach a point where two consecutive iterations do not change any of the bounds for the super-cell counts. The synthesized data table can then be formed from these otpimal bounds.

The GSA was applied to the same representative data set with four residence zones, four workplace zones, and 16 income categories as was used in the combined IPF/Bayesian evaluation. The GSA successfully developed a synthetic data set, but the computational requirements of completing the synthesis were very large. In the short run, it is probably infeasible to apply such an approach for ACS Journey-to-Work analyses because of the computing resource requirements. However, over time, the feasibility of the approach should improve.

Evaluation Findings and Further Research

The IPF and Bayesian methods, taken individually, are feasible approaches to data sysnthesis, but they also have some problems which make it difficult to use them for ACS disclosure avoidance. For instance, one of the disadvantages of IPF is that too much noise is introduced. Implicates are much cleaner, but the resulting tables can vary significantly from the true data tables.

However, a method combining the implicate and IPF methods reduces the some of the undesirable properties. The one potential data synthesis issue that appears to remain is that the distributions between real and synthetic data tend to be off for lower value estimates. Therefore, it is recommended that the implicate plus IPF method be tested further for an ACS test site to determine the validity of the method.

We also recommend further analyses of how a data synthesis approach such as the Bayesian/IPF methods could be integrated into the data dissemination protocols. The Bayesian/IPF approach would require some ad hoc steps, such as the definition of super-tracts and the selection of implicates. Super-tracts could be defined separately from the ACS data development effort, as they would be based on tract population and employment estimates, and could be developed using cluster analyses or other similar methods. Alternatively, secrecy could be

maintained in the selection of super-tracts, as they would for the selection of Bayesian implicates, to improve disclosure-proofing.

The GSA approach offers a more statistically grounded method for disclosure-proofing, but the computational requirements almost certainly make it unreasonable in the near term. Research could assess the available computer resources into the future for the possible adoption of GSA.

When we compare GSA to the other methods that were evaluated for the ACS data, we find that applying the GSA approach for the universe of ACS five-year data would be very difficult, because of the significant computational requirements. However, in the future, the GSA may provide a better and more reasonable way to avoid data disclosure. GSA's sharp bounds for each cell entry produce clear measures of the risk of data disclosure. The algorithm can be used both to develop the synthesized data and to identify those cells that are at the most risk of harmful disclosure. In contrast, while the tables generated with the earlier methods appear to protect confidentiality, it is unclear whether or not an intruder could learn the identity or additional information about individuals in the database. Further research on applying the GSA for future Census data releases could be worthwhile.

2.0 Iterative Proportional Fitting

Iterative Proportional Fitting (IPF), also known as raking in market research circles or "Frataring" by transportation planners and modelers, is a commonly used mathematical technique to fill matrices based on the availability of marginals.[9] The IPF procedure was introduced in 1940 by Deming and Stephan[10] at the U.S. Census Bureau to estimate cell probabilities subjected to marginal constraints. The method's convergence and statistical properties have been investigated since then by several authors and by several different methods. The method is widely in use in travel demand modeling applications for trip distribution analyses that require the development of joint distributions given one- or two-way marginal totals. For example, using one-way distributions of income or mode-to-work to develop a two-way joint matrix of mode-to-work by income.

Table 2.1 shows a simple form of IPF for two tracts (A and B), and for two categories of income (1 and 2). The total workers who lived and worked in these two tracts was 400. The first step of the process is to assemble the marginal totals. For the example, 300 workers fall in income category 1, and 100 workers fall in income category 2. One-quarter (100) of the workers travel between Tract A residences and Tract B; one-quarter (100) travel between Tract B residences and Tract B workplaces; 150 workers travel between Tract A residences and Tract A workplaces; and 50 workers travel between Tract B residences and Tract A workplaces.

The next step in the IPF routine is to initialize all missing cells either to a unitary matrix populated with ones in each cell or to another seed matrix. Initial ratios for multiplying the individual cells in the matrix are obtained by dividing the target totals by the new row totals.

In step 3, the matrix cell values are factored by the calculated row ratios. This step fills the cells of the matrix with new values that are consistent with the origin-destination marginals, but that are not necessarily consistent with the income marginal totals. Ratios of the cell column totals and the marginal totals are calculated, and in step 4, the cell values are factored by these ratios. Steps 3

---

[9] For detailed descriptions of the methodology and information on its historical development linked to the 1940 decennial census see Bishop, Y.M. M., Fienberg, S.E. and Holland, P.W. (1975). ***Discrete Multivariate Analysis: Theory and Practice.*** M.I.T. Press, Cambridge, Massachusetts.

[10] W.E. Deming and F.F. Stephan, On a least square adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math.Statist. 11 (1940), pp. 427–444.

and 4 can be repeated iteratively until convergence is achieved, and the cell values are consistent with both sets of marginal totals.

Table 2.1 Illustration of the Iterative Proportional Fitting Procedure

Step 1: Establish Marginal Totals						
Origin	Destination	Income1	Income2	Total		
A	B			100		
A	A			150		
B	A			50		
B	B			100		
	Total	300	100			
Step 2: Initialize Table Cells						
Origin	Destination	Income1	Income2	New Total	Target Total	Row ratio
A	B	1	1	2	100	50
A	A	1	1	2	150	75
B	A	1	1	2	50	25
B	B	1	1	2	100	50
	Total	4	4			
Step 3: Row Factoring						
Origin	Destination	Income1	Income2	New Total	Target Total	
A	B	50	50	100	100	`
A	A	75	75	150	150	
B	A	25	25	50	50	
B	B	50	50	100	100	
	New Total	200	200			
	Target Total	300	100			
	Column Ratio	1.5	0.5			
Step 4: Column Factoring						
Origin	Destination	Income1	Income2	New Total	Target Total	Row ratio
A	B	75	25	100	100	1
A	A	112.5	37.5	150	150	1
B	A	37.5	12.5	50	50	1
B	B	75	25	100	100	1
	New Total	300	100			
	Target Total	300	100			

An IPF based approach for synthesizing ACS Journey-to-Work flows was examined because:

- There is a high level of familiarity among the transportation modeling community with this method;
- There is a high level of familiarity among Census Bureau staff with the IPF method; and
- There is an assurance of convergence in IPF given non-zero marginals.

U.S. DOT[11] has already worked on some IPF data synthesis strategies at the Census Bureau, and has produced results indicating that a rudimentary IPF with some higher-way marginals can produce 60-80 percent of the original flows. The U.S. DOT research assumed:

- Residence end joint distribution of mode and income (or other variables);
- Work end joint distribution of mode and income (or other variables);
- Residence-Work flow by mode; and
- A larger geography (Super-tract/Super-TAZ) distribution of mode and income.

This research extended the techniques investigated for U.S. DOT.

2.1 IPF SYNTHESIS USING DATA FROM MONTGOMERY COUNTY, MARYLAND

In order to make sure that the IPF procedure was a sensible approach, a test with Montgomery County, MD was undertaken. Data on workers who lived and worked in Montgomery County were obtained from the Census long form. Seventy-one super-tracts were created, each with a population of at least 10,000. The data from the super-tracts are used to develop an IPF routine to synthetically generate data at the Tract or TAZ level. Table 2.2 shows some statistics about the CTPP data for the county for tracts and super-tracts, and Figure 2.1 shows the county tracts and super-tracts. The inputs for the test were (at Census tract level):

- Part 1 (residence) Tables by Income (4 categories – less than $25,000, $25,000-$44,999, $45,000-$74,999, more than $75,0000) by Mode (4 categories – Drove alone, Carpooled, Transit, Others);

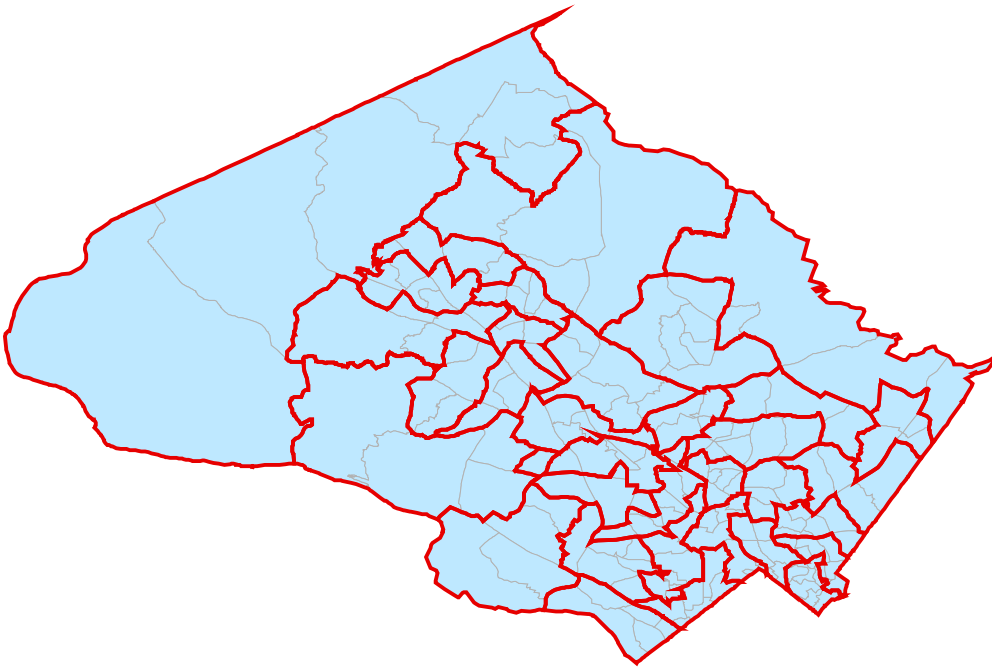
[11] Performed by Nanda Srinivasan, while at Cambridge Systematics, and posted at ftp://ftp.camsys.com/Clientsupport/CTPPdata/NCHRP/srinivasan_report3.ppt.

- Part 2 (workplace) Tables by Income (4 categories – less than $25,000, $25,000-$44,999, $45,000-$74,999, more than $75,0000) by Mode (4 categories – Drove alone, Carpooled, Transit, Others); and
- Part 3 (flow) Tables with worker flows by Mode (4 categories – Drove alone, Carpooled, Transit, Others).

Table 2.2 Montgomery County CTPP Data

Description	Tracts	Super-tracts
Number of Areas	177	71
Average Population	~4,000	~13,000
Number of Possible Combinations	31,329	5,041
Number of Combinations that Occur in the Dataset	10,549	3,886
Percent of Possible Combinations that Occurred	33.67%	77.09%
Percent of Cases based on less than 3 observations	78.50%	39.80%
Number of Possible Cells in the Combinations that Occurred	611842	225388
Number of Non-Zero Cells	20,195	12,300

Figure 2.1 Montgomery County Tracts and Super-tracts



The IPF procedure for creating the synthetic data is as follows:

- Step 0: Initialize Tract-Tract flows by mode and income to 1 in the result matrix;
- Step 1: Distribute income by mode using Super-tract income by mode marginals;
- Step 2: Use real data from Part 1 to factor summed flows on the residence end as follows:
  - Sum flows by mode and income by residence Tract for result matrix;
  - Factor summed flows with Part 1 data and get a factor matrix; and
  - Multiply all values in result matrix using Residence tract factor in factor matrix.
- Step 3: Repeat Step 2 on work geography using Part 2 data;
- Step 4: Use total workers at tract to tract to develop a third factor, and factor every cell using the flow factor; and
- Step 5: Repeat steps 1 through 4 multiple times to achieve closure.

The closure in total workers from tract–to-tract was observed to be:

Mean = 0.999991, STDEV = 0.003506, Max = 1.095617 and Min = 0.932641

Table 2.3 shows the percent of cells that have differences in the number of workers between real and synthetic data for the number of workers in the real data. As can be seen from the table, while the data seem to correlate pretty well, there are number of cells with high differences (around six percent have differences of more than three).

Table 2.3 Distribution Between Real and IPF Synthesized Data

	Number of Workers in Real Data					
Difference in Number of Workers between Real and Synthetic Data	0	1-10	11-20	20-30	>30	Total
Total Pairs	100%	100%	100%	100%	100%	100%
Difference = 0	93%	40%	32%	23%	21%	86%
Difference = 1	2%	13%	8%	8%	8%	3%
Difference = 2	2%	12%	10%	9%	8%	3%
Difference = 3	1%	11%	9%	9%	9%	2%
Difference = 4	1%	9%	9%	7%	7%	2%
Difference = 5	0%	6%	7%	6%	6%	1%
Difference > 5	1%	8%	24%	38%	40%	3%

Table 2.4 shows the distribution between real and synthetic data for non-zero cells. As can be seen from Table 2.4 for non-zero values, the results seem close (within 3 or less) to original values 70 percent of the time. An advantage of IPF is that all the marginals are kept. The correlation between synthetic and real data is around 0.89, and the regression relationship between the synthesized and real data is:

Real value = 0.95 x IPF Data + 2.3.

Table 2.4 Distribution between Real and IPF Synthesized Data for Non-Zero Cells

	Number of Workers in Real Data				
Difference in Number of Workers between Real and Synthetic Data	1-10	11-20	20-30	>30	Total
Total Pairs	100%	100%	100%	100%	100%
Difference = 0	45%	36%	25%	22%	40%
Difference = 1	12%	7%	7%	8%	10%
Difference = 2	11%	9%	9%	7%	10%
Difference = 3	10%	9%	9%	9%	10%
Difference = 4	8%	8%	7%	7%	8%
Difference = 5	6%	7%	6%	6%	6%
Difference > 5	7%	23%	37%	40%	15%

So far, the focus has been on comparing the IPF matrix with the original origin-destination matrix by income and mode. Based on the test results in Tables 2.3 and 2.4, the IPF approach seems to perform pretty well with all the marginals retained. Another measure of the adequacy of the approach is the determination of how consistent the IPF matrix is with CTPP Part 1 and Part 2 tables. Table 2.5 shows the results of the comparison, and indicates that for the most part it is mostly consistent. In this table, the mode and income are kept at their most disaggregate categories of seven mode and eight income categories respectively and hence the number of cells for origin and destination tracts are multiplied by 56.

Table 2.5 Differences between the IPF Synthesized Matrix and CTPP Part 1 and Part 2 Mode and Income Tables

PART 1 COMPARISON		PART 2 COMPARISON	
Difference in Workers for Origin Tracts	Number of Cells	Difference in Workers for Destination Tracts	Number of Cells
Total Cells (=177*56)	9,912	Total Cells (=177*56)	9,912
0	8,989	0	9,215
1	675	1	504
2	196	2	118
3	40	3	42
4	11	4	19
5	1	5	9
6 through 9		6 through 9	5

Having confirmed the overall validity of the IPF approach, the next step was to apply it to another geography to test whether the approach is transferable. For this purpose, Cook County, IL was selected.

2.2 IPF SYNTHESIS USING DATA FROM COOK COUNTY, ILLINOIS

Table 2.6 shows the data for the Cook County, IL at the Tract and Super-tract levels. The super-tracts had a greater population threshold (25,000) compared to 10,000 for Montgomery County, MD. As can be seen from the data, even at the super-tract level, about half of the flow pairs are based on less than 3 observations. However, the thinking is that at such large geography the probability of two respondents interacting with each other in a demographic survey is pretty low.

Similar to the previous exercise an IPF routine was run for Cook County using the super-tract marginals by mode and income and the tract level information at the residence (part 1) and workplace (part 2) ends. The IPF produced 373,388 non-zero cells. This means a significant portion of the zero cells were not estimated correctly. However, the IPF results are only off by 1 or 2. The relationship between the real and IPF data is:

Real Value = 0.92 x IPF Data + 3.26

with a R(Squared) of 0.84 indicating a reasonable fit.

Table 2.6 Cook County CTPP Data

Description	Tracts	Super-tracts
Number of Areas	1344	217
Average Population	~4,000	~25,000
Number of Possible Combinations	1,806,336	47,089
Number of Combinations that Occur in the Dataset	128,857	33,028
Percent of Possible Combinations that Occurred	7.10%	70.10%
Percent of Cases based on less than 3 observations	88.90%	49.90%
Number of Possible Cells in the Combinations that Occurred	2,061,712	528,448
Number of Non-Zero Cells	194,311	96,333

In addition, univariate statistics were also obtained to determine how well the distribution between the real and IPF data compared. Table 2.7 shows the quantile distribution between Real and IPF data, and shows a wide disparity between the Real and IPF synthesized data estimates. Around 25 percent of the IPF data have estimates of 2 or 1 compared to around one percent for the Real data. This indicates that further data modifications are required for the analysis to be used reliably.

Table 2.7 Comparison of Quantile Distributions of Real Data and IPF Synthesized Data

Quantile	Real Data	IPF Data
100%Max	1,081	1,114
99%	48	33
95%	24	16
90%	18	11
75%	12	7
50%	8	4
25%	6	2
10%	5	1
5%	4	1
1%	3	1
0%	1	1

Since the analysis whose results are reported in Table 2.7 did not try to use the fact that a majority of CTPP numbers are greater than three, the next test was to

eliminate the numbers less than or equal to three from the original IPF run and rerun the IPF procedure multiple times. The relationship between the real and IPF synthesized data from this test is:

Real Value = 0.93 x IPF Data + 2.

The R(Squared) dropped from 0.84 to 0.76 indicating some problems with the fit. However, as seen from Table 2.8 the quantile distribution between Real and IPF data are much closer. However, it should be noted that the IPF did not converge after 32 iterations.

Table 2.8 Comparison of Quantile Distributions of Real Data and IPF Synthesized Data With the Elimination of Small Cell Values

Quantile	Real Data	IPF Data	IPF (Run2)
100%Max	1,081	1,114	1,105
99%	48	33	44
95%	24	16	22
90%	18	11	17
75%	12	7	12
50%	8	4	8
25%	6	2	6
10%	5	1	5
5%	4	1	3
1%	3	1	1
0%	1	1	1

2.3 IPF SYNTHESIS WITH REDUCED SUPER-TRACT DATA

While the previous test reducing the IPF data to include only cells with counts of greater than three gave results that were close to the real data, the fact that the model correlation was weak and did not converge after multiple iterations indicated that further fine tuning was required to the procedure.

In this test, the super-tract univariate tables for mode (4 categories – Drove alone, Carpooled Transit, All others) and Income (4 categories – Less than $25,000, $25,000-$44,999, $45,000-$74,999, $75,000 or more) are used as inputs, along with the tract level mode by income tables for residence end (part 1), workplace end (part 2), and tract level worker flow by mode. The tract level inputs are the same

as what was used for the initial tests in Montgomery County, Maryland and Cook County, Illinois. The difference between this and the previous tests is that a two phased IPF implementation is used initially to develop a synthetic IPF distribution of mode by income for super-tract-to-super-tract flows.

The procedure first uses the super-tract mode and super-tract income univariate tables along with the super-tract level distribution of mode by income to produce a synthesized IPF distribution of mode (4 categories) by income (4 categories) for super-tract-to-super-tract flows. This bivariate table is now used as an input along with the other input tables as in the previous tests. Then the same procedure as used in Montgomery County, Maryland and Cook County, Illinois previously is used to generate tract-to-tract synthetic flows by mode and income.

Using this method, the IPF data produced more than 500,000 non-zero cells. This means a significant portion of the zero cells were not estimated correctly. IPF results are only off by 1 –4 for the lower values. The real value is related to the IPF cell value as:

Real Value = 0.90 x IPF Cell value + 5

with a R(Squared) of 0.8 and correlation of 0.9.

This indicates that IPF values are off by five to eight for almost all values, pointing to the fact that the results are very unimpressive.

For values greater than five, the results still remained very similar:

Real Value = 0.89 x IPF Cell value + 5.7

with R(Squared) of 0.8 and correlation of 0.9.

However, the sum total of the IPF dataset for values over 5 was 1,133,182, while for the real data set was 1,944,895. This means that a significant number of smaller values are present in the IPF dataset.

Because the first IPF result showed many cells being equated to 1, 2, and 3, and because these values occur very less frequently in the original distribution, the IPF routines were run after setting any number less than three to be equal to zero. The model did not achieve convergence even after 100 iterations and the resulting dataset lost about 182,534 workers out of 2,077,798 workers, a loss of 8.7 percent. The correlation came down to 0.85 and the regression came down to

Real Value = 0.9 x IPF Cell value + 2.49

with a R(Squared) of 0.72.

The model matched much better at higher values as shown by the differences in the quantile distributions of the real and IPF synthesized data, as shown in Table 2.9. As can be seen from Table 2.9, five percent of the IPF data had values of 3 or less whereas only one percent of the real data had values of 3 or less.

To see if any improvements could be made on the original two phased IPF implementation, the IPF routines were run after making any number less than or

equal to two as zero. The model did not achieve convergence even after 50 iterations and the resulting dataset lost about 54,060 workers out of 2,077,798 workers, a loss of 2.6 percent. The correlation performed slightly better (0.88) and the regression came down to:

Real Value = 0.93 x IPF Cell value + 3

with a R(Squared) of 0.77.

The model did not match much better between the real and IPF data compared to making values less than three equal to zero, and are shown in Table 2.10.

Table 2.9 Comparison of Quantile Distributions of Real Data and IPF Synthesized Data With the Elimination of Cell Values of Less than Three

Quantile	Real Data	IPF Data
100%Max	1,081	1,116
99%	48	44
95%	24	22
90%	18	17
75%	12	12
50%	8	8
25%	6	6
10%	5	4
5%	4	3
1%	3	1
0%	1	1

Table 2.10 Comparison of Quantile Distributions of Real Data and IPF Synthesized Data With the Elimination of Cell Values of Less than Two

Quantile	Real Data	IPF Data
100%Max	1,081	1,114
99%	48	39
95%	24	18
90%	18	14
75%	12	9
50%	8	7
25%	6	5
10%	5	3
5%	4	3
1%	3	1
0%	1	1

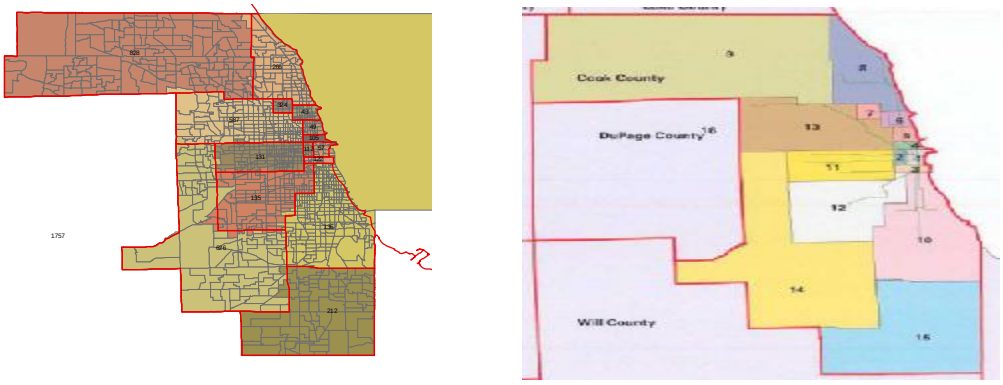
While the regression results and Tables 2.9 and 2.10 show that IPF gives spurious distributional differences and does not converge when changes are made to the results in terms of reducing small cell counts to zeros, the question that needs to be asked is whether the results of the IPF methodology hold up under a typical transportation planning scenario such as adding up TAZs to super-districts.

2.4 APPLICATION OF THE IPF DATA SYNTHESIS METHODOLOGY TO A REAL WORLD APPLICATION

In travel demand forecasting, it is a common practice to roll up TAZs to super-districts so that CTPP data can be compared to model results. The reason for doing this is to present results of analyses in a manner that is easily understood by decision-makers.

A super-district file for Cook County was obtained from the Federal Transit Administration (FTA) and a near equivalency file to the super-districts with Census tracts was established. Figure 2.2 shows the super-districts and the census tract equivalents.

Figure 2.2 FTA Super-Districts in Cook County, Illinois



The first step was to obtain bivariate tract data tables of Mode (4 categories) by Income (4 categories) from the original unrounded data, and then to aggregate it to the super-district level. Similarly, the bivariate synthetic IPF tract data tables of Mode (4 categories) by Income (4 categories) obtained by the two phased IPF procedure were aggregated to the super-district level.

The results of this test indicated a remarkably good correlation (0.99) and the

Real Value (Super-district Level) = 1.00017x IPF Cell Value +1.14.

Further tests were done by forcing the IPF data with values less than or equal to 3 or 2 to zero, but they did not perform that well indicating that it might not be a good idea to force the IPF to ignore the lower values. Even though the data seem skewed towards lower values in the original IPF, they behave better when grouped to larger geography.

The final test was to use the reduced super-tract data but with Univariate Tables of Income and Mode at super-tract level to create Mode by Income summaries at the Tract level. Here it was assumed that Mode by Income information is available for those rows based on 3 or more observations at the tract level. This test also used lesser information than in any of the runs at the super-tract level, and assumed a disclosure threshold of 3 (just as in CTPP 2000) at tract level. The original result assuming super-tract bivariate tables was:

Real Value = 0.90 x IPF Cell value + 5

with a correlation of 0.90.

In this test we created a Mode (4 categories) x Income (4 categories) table at the tract level and suppressed all rows that had less than three observations. This was followed by creating super-tract univariate tables by Mode (4 categories) and Income (4 categories). In addition, residence end (Part 1), and workplace end (Part 2) like tables for Mode (4 categories) by Income (4 categories) were created. The IPF procedure was revised so that it uses univariate super-tract tables to create a bivariate distribution of Mode (4 categories) by Income (4

categories) for tract flows only for those tract pairs where the observations are greater than three.

The synthetic data obtained from above is now fed into a second IPF routine for generating flows by Mode (4 categories) x Income (4 categories) at the tract level. The second IPF routine is run by assuming that tract flow pairs would show these data for observations greater than two and only these flows are initialized in the IPF routine. The resulting regression shows that the:

Real Value = 0.92 x IPF Cell Value + 2.83

with a correlation of 0.96.

This is an excellent result because it gives a much better correlation, even though we are using a two-pronged IPF with lesser information than before for sparse cells. In addition, it considers disclosure issues much better than the other tests.

As seen from Table 2.11 the distribution still showed the median to be off, but previous tests have shown that changing values of 3s, and 2s in the synthetic data to zero does not help provide a better fit between the real and synthetic data.

Table 2.11 Comparison of Quantile Distributions of Real Data and IPF Synthesized Data for the Final IPF Test

Quantile	Real Data	IPF Data
100%Max	1,081	1,081
99%	48	37
95%	24	17
90%	18	12
75%	12	7
50%	8	4
25%	6	2
10%	5	1
5%	4	1
1%	3	1
0%	1	1

3.0 Joint Bayesian/IPF Data Synthesis

The IPF procedures explored in the previous section relied on using univariate Journey-to-Work Flow tables at super-tract geography. In this section the possibility of combining a Bayesian method along with the IPF is explored. The method generates super-tract level bivariate tables using super-tract level univariate tables where a disclosure threshold of two or more records is met, and using just one implicate where the disclosure threshold is not met.

There is precedence for using Bayesian methods for data synthesis in the Census Bureau, specifically in the Longitudinal Employer Household Dynamics (LEHD) data. In the first version of the LEHD Origin-Destination database, produced in 2003, cell suppression was used to protect confidentiality. If the number of workers residing in a block was less than 5, or if the number of different blocks in which they were employed was less than 3, then the observation was suppressed. The cell suppression rules seriously limited the resulting data. Therefore, instead of relying on cell suppression, subsequent versions of the LEHD data set have been disclosure-proofed by "synthesizing" them from the true data. The key statistical property to preserve in the synthetic data is the joint distribution of workers across home and work areas. Bayesian techniques are used to synthesize workers' place of residence conditional on disclosable counts of workers by place of work, industry, age, and earnings categories.

The basic idea is to replace the true home blocks by drawing synthetic home blocks for the workers with each set of characteristics on each work block. The draws are from the distribution of actual home blocks among workers with similar age, earnings, and industry characteristics, combined in a certain measure with another distribution of home blocks known as the "prior" distribution. The actual home blocks of workers employed on a certain block are defined according to Quarterly Workforce Indicators (QWI) disclosure rules governing counts, with one modification. In the QWIs, there must be at least 3 entities on which to base a releasable statistic; here, a draw of 1 or 2 substitutes for any true count of workers less than 3. In this application, the prior is simply the distribution of home blocks among a larger worker group encompassing the target group. First the prior for each set of worker characteristics in the Census tract containing the target work block is constructed. In cases where this does not provide population over a sufficient number of residence blocks, worker types are aggregated. If this is not enough, the population is aggregated to larger geographies with or without distinctions between worker types. The prior assures that some trips in the synthetic data are absent from the true data, by allowing draws of home blocks from which none of the workers on this block

actually originate. It also assures that the home block of a unique worker can be synthesized.[12]

While the Bayesian method adopted in this task is similar to the LEHD method, because most of the flow data were already available, the Bayesian method is applied only for univariate income tables and hence there are some differences that are highlighted below:

- Residence end marginals are used to constrain the flows (for sampling for priors);
- In the posterior distribution stage, any cell that had more than 50 percent probability had about 20 percent donated to the immediate adjoining cell;
- The result provides for "methodological" protection, i.e., no single cell observation will really be shown on the final univariate table;
- Only one implicate which while crude is sufficient for the purposes of this task; and
- While the final univariate table is pretty asymmetric, however, the IPF in the next stage will create an income by mode table and remove the asymmetry introduced.

The joint Bayesian and IPF procedure is as follows:

- Use Implicate method to develop a disclosure proofed one-way income table for flows with 2 or less observations;
- Use Super-Tract to Super-Tract Flow univariate Income (4 categories) data:
  - If the number of observations for each flow is greater than 2, then keep the data,
  - If the number of observations is less than or equal to 2, then apply implicate method to generate one implicate,
  - Final output will be an univariate income file at super-tract geography;
- Apply IPF as before to generate a bivariate income x mode to work table; and
- Test final bivariate table against original table to see correlations.

[12] Marc Roemer, An Origin-Destination Matrix, Area Characteristics Files and Quarterly Workforce Indicators for the Employment and Training Administration, LEHD, Census Bureau, July 15, 2005.

3.1 TEST OF THE JOINT BAYESIAN / IPF DATA SYNTHESIS USING AN EXAMPLE DATA SET

Table 3.1 shows an example origin-destination univariate table with 16 income categories. The rows highlighted in yellow need to be disclosure proofed because they have two or fewer observations.

The implicates are developed using the following procedure:

- Step 1. Use total residence workers as the basis of the "prior";
- Step 2. Put a weight on "prior" (of 10, for example) to get samples. While the weight can be any value, it has to be developed with some thought;
- Step 3. Develop a posterior distribution using the current super-tract to super-tract flow by income (4 categories) distribution perturbated by the addition of the sample; and
- Step 4. Using total worker counts as the basis, develop a uniform random distribution of workers and populate the cells based on the posterior distribution.

The "prior" is developed by merging the residence end super-tract worker data with the super-tract origin-destination table and then dividing each total flow by total residence workers. The "prior" is then multiplied by the weight to get a sample for each super-tract pair.

The posterior matrix is developed by first adding the prior matrix sample to each of the income total cells, and then by taking the total proportion of the samples. An additional step is added here since the implicate method preserves "single cell" values. Cell values are distributed around to neighboring cells using the following rule:

- If a cell had more than 50 percent probability of occurrence, 20 percent is transferred to a neighboring cell.

The thinking behind this is that single observations are now "made to feel" like two observations and two observations with more weights on one of them are made to feel like three observations. Following this step, we develop a uniform random distribution of workers to create a new distribution for the output. This step ensures that any concerns about privacy and confidentiality are addressed. Including this step totally distorts the distribution of the numbers in the cells. Hence, the IPF is introduced to correct the loss of symmetry. The IPF balances the productions (Place-of-Residence) and attractions (Place-of-Work).

Table 3.1 Sample Data to be Disclosure Proofed

Ozone	Dzone	Inc1	Inc2	Inc3	Inc4	Inc5	Inc6	Inc7	Inc8	Inc9	Inc10	Inc11	Inc12	Inc13	Inc14	Inc15	Inc16	Freq
a	A	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
a	B	46	34	0	23	0	0	0	0	0	0	0	0	0	0	0	0	5
a	C	243	200	0	0	45	0	0	0	70	0	0	80	0	0	0	0	5
a	D	0	0	0	0	0	0	0	45	60	0	0	0	0	0	0	0	2
b	A	4	9	15	14	18	17	0	0	17	18	22	44	33	0	16	16	8
b	B	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0	0	1
c	A	14	24	36	34	14	16	17	18	0	18	12	0	44	34	33	33	12
c	B	0	0	14	0	16	18	18	34	12	16	44	22	16	18	12	14	14
c	C	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	1
d	A	12	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
d	B	14	12	67	9	22	66	14	14	34	37	38	12	24	22	16	18	15
d	C	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	1
d	D	0	0	0	0	0	0	18	0	0	22	0	0	0	0	0	0	2

Table 3.2 shows the real and synthesized data after using the implicate method for two or less observations.

Table 3.2 Real and Synthesized Data After Using Implicate Method

		Real Data					Synthesized Data			
Ozone	Dzone	FREQ	Inctot1	Inctot2	Inctot3	Inctot4	Inctot1	Inctot2	Inctot3	Inctot4
a	A	1	9	0	0	0	8	1	0	0
a	B	5	103	0	0	0	103	0	0	0
a	C	5	443	45	150	0	443	45	150	0
a	D	2	0	45	60	0	1	36	54	14
b	A	8	42	35	101	65	42	35	101	65
b	B	1	0	0	0	78	4	1	21	52
c	A	12	108	65	30	144	108	65	30	144
c	B	14	14	86	94	60	14	86	94	60
c	C	1	0	7	0	0	0	7	0	0
d	A	2	30	0	0	0	24	6	0	0
d	B	15	102	116	121	80	102	116	121	80
d	C	1	0	18	0	0	0	16	1	1
d	D	2	0	18	22	0	2	19	15	4

Finally, the IPF routine is applied to obtain a bivariate Income (4 categories) by Mode (4 categories) table with the following tables as inputs:

- One way flow mode (4 categories) table (original)
- One way flow income (4 categories) table (modified)
- Place of Residence Two way mode (4 categories) x income (4 categories)
- Place of work Two way mode (4 categories) x income (4 categories)

Table 3.3 shows the final synthetic data after applying the IPF and mimics the original data (Table 3.1).

Table 3.3 Synthetic Data after Disclosure Proofing

Ozone	Dzone	Inc1	Inc2	Inc3	Inc4	Inc5	Inc6	Inc7	Inc8	Inc9	Inc10	Inc11	Inc12	Inc13	Inc14	Inc15	Inc16
A	A	9	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
A	B	46	34	0	23	0	0	0	0	0	0	0	0	0	0	0	0
A	C	243	200	0	0	45	0	0	0	70	0	0	80	0	0	0	0
A	D	0	0	0	0	0	0	0	45	60	0	0	0	0	0	0	0
B	A	4	9	15	14	18	17	0	0	17	18	22	44	33	0	16	16
B	B	0	3	0	0	0	1	0	0	0	8	0	0	0	78	0	0
C	A	14	24	36	34	14	16	17	18	0	18	12	0	44	34	33	33
C	B	0	0	14	0	16	18	18	34	12	16	44	22	16	18	12	14
C	C	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0
D	A	11	17	0	0	2	3	0	0	0	0	0	0	0	0	0	0
D	B	14	12	67	9	22	66	14	14	34	37	38	12	24	22	16	18
D	C	0	0	0	0	16	0	0	0	1	0	0	0	0	0	0	0
D	D	0	0	0	0	0	0	18	0	0	22	0	0	0	0	0	0

The synthetic data are correlated very well with the real data (0.99997). Also, the fit between the real and synthetic data is also very close with

Real Value = 0.99934 x Synthetic Value + 0.07

with a R(squared) of 0.9999.

Table 3.4 shows the distribution between the real and synthetic data. As in previous tests, the distribution is off for large estimates despite the excellent overall fit.

Table 3.4 Comparison of Quantile Distributions of Real Data and Bayesian/IPF Synthesized Data

Quantile	Real Data	Synthetic Data
100% Max	243	243
95%	78	70
90%	60	46
75%	34	34
50%	18	18
25%	14	14
10%	12	8
5%	9	3
1%	4	1
0%	4	1

3.2 APPLICATION OF THE JOINT BAYESIAN / IPF APPROACH TO REAL-WORLD DATA

While the method combining Bayesian and IPF gave synthetic data that were close to the real data, we needed to address the question of how effective the method is at analyzing both large (Super-tract) and small (TAZ) geographies.

Cook County Super-tract Geography

To test this, the Bayesian / IPF method was applied to super-tract geography for Cook County, IL; then to tract geography for Seattle MSA with enhancements and to TAZ geography for Seattle MSA.

Applying the combined Bayesian and IPF method to super-tracts in Cook County, IL gives a very high correlation (0.99346) and close match between the real value and synthetic value as:

Real Value = 0.99 x Synthetic Value + 2

with a R (squared) of 0.987. While the fit is good, the distribution between the real and synthetic data is the familiar tale of being slightly off (Table 3.5).

Table 3.5 Comparison of Quantile Distributions of Real Data and Bayesian/IPF Synthesized Data for Cook County Super-tracts

Quantile	Real Data	Synthetic Data
100% Max	3,501	3,435
95%	68	47
90%	39	26
75%	19	34
50%	10	12
25%	7	6
10%	5	3
5%	5	1
1%	3	1
0%	1	1

Seattle MSA Tract Geography

The next test was to apply the Bayesian plus IPF method to the Seattle MSA Tract geography. While the method is similar to that done with the Cook County, IL super-tract data, the fact that about 33 percent of flow observations are already allocated in some way or form at tract level needs to be considered.

The five-county Seattle MSA had about 735 Tracts. Of these, 69,916 tract-flow-pairs occurred in the flow dataset (representing 1,642,680 workers). A total of 25,532 of these pairs were based on 3 or more observations or were entirely imputed observations (representing 1,100,962 workers). About a third of the workers and 63 percent of flow pairs would have been suppressed had the CTPP 2000 rules been used at the tract level.

Using the Bayesian and IPF methods together on the Seattle MSA Tract Geography gave us a very highly correlated result between the real and synthetic data (0.99474). Further, regressions indicate that the

Real Value = 0.992 x Synthetic Value + 0.7

with a R (squared) of 0.9895. The result is better than those obtained at the Super-tract geography. But the income univariate table is adjusted as before. In addition, the result is symmetric and complete, while being synthetic, and disclosure-proofed, where it needs to be.

Seattle MSA TAZ Geography

In this instance, the combined Bayesian and IPF methods are applied to the Seattle MSA TAZ geography. There are 936 TAZs in the region, compared to 177 tracts. There were a number of "zzzzzz" polygons, so there was incomplete TAZ coverage. About 92,115 combinations were observed. For the smaller geography, there are bound to be many more cells with counts of one or two introduced because of the IPF portion of this method.

The results show a high correlation between real and synthetic data (0.99731) and the regressions indicate that there is a close match between the real and synthetic data:

Real Value = 0.996 x Synthetic Value + 0.87

with a R (squared) of 0.9946. The result is better than those obtained at the Tract geography, but the income univariate table is adjusted as before.

3.3 SUMMARY

Based on the results of the comparison between synthesized data and real flow data from Cook County, IL and Seattle MSA at the Super-tract, tract and TAZ levels, the Bayesian (implicate) plus IPF method performed quite well. While the IPF and Bayesian methods, taken individually, have advantages, they also have some problems which make it difficult to use them for ACS disclosure avoidance. One of the disadvantages of IPF is that too much noise is introduced. Implicates are much cleaner, but they can vary significantly from the true data.

Nevertheless, a method combining the implicate and IPF methods reduces the probability of populating "zero cells," while achieving the same amount of disclosure protection. The one caveat for all three methods is that the distributions between real and synthetic data tend to be off for lower value estimates. Therefore, it is recommended that the implicate plus IPF method be tested for an ACS test site to determine the validity of the method.

4.0 Generalized Shuttle Algorithm

4.1 OVERVIEW

The Census Transportation Planning Product (CTPP) and the suite of products from the American Community Survey (ACS) are used for current and continuing planning by state and metropolitan planning organizations. These include journey-to-work tables at small-level geography and often include very small numbers of responses for a particular location. The Census Bureau created a Disclosure Review Board (DRB) to review tables before release and ensure confidentiality of responses. However, the rules implemented by the DRB are expected to cause severe loss of information when applied to the ACS tables because of the large proportion of small values in these tables. Such loss of information reduces the usefulness of the journey-to-work tables in metropolitan planning.

The objective of both the methods described above and the method that is outlined here is to produce synthetic data that the Census Bureau could release which protect against the disclosure of confidential information and which provide a richer array of information than current Census Bureau procedures.

This chapter is comprised of the following components:

- Description of the generalized shuttle algorithm (GSA) for computing bounds on table values given a set of marginal values;
- Implementation of the GSA on the example journey-to-work tables that were used to test the other table synthesis method;
- Description of the problem of disclosure and measures of the risk of disclosure in journey-to-work tables; and
- A recommendation for implementing disclosure limitation in the release of journey-to-work tables.

4.2 SOME TECHNICAL SPECIFICATIONS REGARDING CONTINGENCY TABLE DATA

A contingency table is an array of non-negative integers that arises from the cross-classification of N objects based on their observed value for each of k categorical variables of interest, $X = (X_1, \ldots, X_k)$ (see [4] and [12]). An object's observed value on the variable X_r can fall into one of the l_r

possible categories $\{q_1^r,....,q_{l_r}^r\}$; q_i^r is the label for the ***i***th category of the ***r***th categorical variable. By defining an index set for each variable, I_r = $\{1,....,l_r\}$, and the index set for the table, $I = I_1 \times ... \times I_k$, any coordinate $\langle i_1,....,i_k \rangle \in I$ is referred to as a cell. For each cell, $n(i_1,....,i_k)$ is the count of cell $\langle i_1,....,i_k \rangle$, where $n(i_1,....,i_l)$ is the number of objects whose observed value of X is $\langle q_{i_1}^1,....,q_{i_k}^k \rangle$. In other words, $n(i_1,....,i_l)$ is the number of objects whose observed value of X_r is $q_{i_r}^r$ for r = 1,..., ***k.*** We represent the contingency table as a vector of nonnegative integers $n = \{n(i)\}_{i \in I}$.

If $J_r \subset I_r$ for r = 1,..., ***k,*** then we call

$$\langle J_1,....,J_k \rangle = J_1 \times ... \times J_k = \{\langle i_1,...,i_k \rangle \in I : i_r \in J_r, r = 1,...,k\}$$

a "super-cell", and denote

$$n(J_1,...,J_k) = \sum_{i_1 \in J_1} ... \sum_{i_k \in J_k} n(i_1,...,i_k)$$

as the count of super-cell $J_1 \times ... \times J_k$. Note that a cell is also a super-cell, and that $n(J_1,...,J_k)$ is the number of objects whose observed value of X is a member of the set

$$\bigcup_{i_1 \in J_i,...,i_k \in J_k} \langle q_{i_k}^1,...,q_{i_k}^k \rangle.$$

Equivalently, $n(J_1,...,J_k)$ is the number of objects whose observed value of X_r is a member of $\{q_{i_r}^r\}_{i_r \in J_r}, r = 1,...,k.$

Following Dobra and Fienberg, we denote ***T*** as the set of all super-cells, $T = \{J \in 2^I : J = \langle J_1,...,J_k \rangle, J_r \subset I_r, r = 1,...,k\}$, and ***Q(T)*** as the set off all "dependencies" in T,

$$Q(T) = \{\langle J_1, J_2, J_3 \rangle \in T \times T \times T : J_1 \cup J_2 = J_3\}.$$

It is worth noting that not all subsets of ***I*** are super-cells. If the union of two super-cells is a super-cell, then their constituent indices differ for only one variable. That is, if $J_s = \langle J_1^s,...,J_k^s \rangle \in T$ for ***s*** = 1, 2, 3, and $J_1 \cup J_2 = J_3$, there exists at most one 1 ≤ ***i*** ≤ ***k*** such that $J_i^1 \neq J_i^2$. Also, if there is a dependency between super-cells, $\langle J_1, J_2, J_3 \rangle \in Q(T)$, then their cell counts sum,

$$n(J_1) + n(J_2) = n(J_3). \tag{1}$$

4.3 THE GENERALIZED SHUTTLE ALGORITHM

The generalized shuttle algorithm (GSA), as proposed by Dobra and Fienberg (see [6, 7, 8, 9]) and implemented in [11], is a method for determining bounds for cell and super-cell counts. There are ***fixed*** super-cells, whose counts are fixed by either empirical observation or assumption, and ***free*** super-cells, whose counts are unobserved or simply unconstrained. In practice, certain marginals are known and used as the fixed super-cells.

In general terms, the GSA uses the fixed super-cell counts and the intrinsic dependencies between counts of super-cells $\langle J_1, J_2, J_3 \rangle \in Q(T)$ as defined in equation (1) to iteratively refine estimates of counts for free super-cells.

We set $|Q(T)| = q$, and let $T_{0,q} \subset T$ be the set of all fixed super cells, i.e. the count of super-cell J is fixed if and only if J ∈ $\boldsymbol{T_{0,q}}$. When J ∈ $T_{0,q}$, we denote its fixed count as c(J). The algorithm begins with a set of initial lower bounds for super-cell counts, $L_{0,q}(T) = \{L_{0,q}(J) : J \in T\}$, and a set of initial upper bounds for super-cell counts, $\boldsymbol{U_{0,q}(T) = \{U_{0,q}(J)}$: J ∈ $\boldsymbol{T\}}$. These initial bounds are set such that for J ∈ $\boldsymbol{T_{0,q}}$, $L_{0,q}(J) = \boldsymbol{U_{0,q}(J)} = c\boldsymbol{(J)}$, while for J ∈ $\boldsymbol{T \setminus T_{0,q}}$, $L_{0,q}(J) = 0$, $U_{0,q}(J) = \boldsymbol{N}$.

The GSA algorithm begins by selecting a visiting schedule $\{\langle J_1, J_2, J_3 \rangle_i\}_{i=1}^{q}$, which is nothing more than an enumeration of ***Q(T)***. Following this schedule, each member of ***Q(T)*** will be visited during each iteration of the GSA. For ***i*** = 1,2,...,***j*** = 1,..., q, let $L_{i,j}(T) = L_{i,j}(J) : J \in T$ be the lower bounds for the super-cell counts during iteration ***i*** after having visited dependency $j, \langle J_1, J_2, J_3 \rangle_j, U_{i,j}(T) = \{U_{i,j}(J) : J \in T\}$ be the upper bounds for the super-cell counts during iteration ***i*** after having visited dependency j, and $T_{i,j}$ be the collection of all super-cells J ∈ ***T*** such that $L_{i,j}(J) = U_{i,j}(J)$.

Suppose that during the ***ith*** iteration, we are at dependency $\langle J_1, J_2, J_3 \rangle_{j+1} \in Q(T), 0 \le j \le q-1$. If we set $U_{i,0} = U_{i-1,q}, L_{i,0} = L_{i-1,q}$, and $T_{i,0} = T_{i-1,q}$, we then update as follows: for super-cells not in the current dependency, $J \in T \setminus \{J_s\}_{s=1}^{3}$, the bounds do not change

$$U_{i,j+1}(J) = U_{i,j}(J)$$
$$L_{i,j+1}(J) = L_{i,j}(J),$$

otherwise,

$$U_{i,j+1}(J_1) = \begin{cases} U_{i,j}(J_3) - U_{i,j}(J_2) & : \langle J_2, J_3 \rangle \in T^2_{i,j} \\ U_{i,j}(J_1) & : \langle J_2, J_3 \rangle \notin T^2_{i,j}, J_1 \in T_{i,j} \\ \min\{U_{i,j}(J_1), U_{i,j}(J_3) - L_{i,j}(J_2)\} & : else \end{cases}$$

$$L_{i,j+1}(J_1) = \begin{cases} L_{i,j}(J_3) - L_{i,j}(J_2) & : \langle J_2, J_3 \rangle \in T^2_{i,j} \\ L_{i,j}(J_1) & : \langle J_2, J_3 \rangle \notin T^2_{i,j}, J_1 \in T_{i,j} \\ \max\{L_{i,j}(J_1), L_{i,j}(J_3) - U_{i,j}(J_2)\} & : else \end{cases}$$

$$U_{i,j+1}(J_2) = \begin{cases} U_{i,j}(J_3) - U_{i,j}(J_1) & : \langle J_1, J_3 \rangle \in T^2_{i,j} \\ U_{i,j}(J_2) & : \langle J_1, J_3 \rangle \notin T^2_{i,j}, J_2 \in T_{i,j} \\ \min\{U_{i,j}(J_2), U_{i,j}(J_3) - L_{i,j+1}(J_1)\} & : else \end{cases}$$

$$L_{i,j+1}(J_2) = \begin{cases} L_{i,j}(J_3) - L_{i,j}(J_1) & : \langle J_1, J_3 \rangle \in T^2_{i,j} \\ L_{i,j}(J_2) & : \langle J_1, J_3 \rangle \notin T^2_{i,j}, J_2 \in T_{i,j} \\ \max\{L_{i,j}(J_2), L_{i,j}(J_3) - U_{i,j+1}(J_1)\} & : else \end{cases}$$

$$U_{i,j+1}(J_3) = \begin{cases} U_{i,j}(J_1) + U_{i,j}(J_2) & : \langle J_1, J_2 \rangle \in T^2_{i,j} \\ U_{i,j}(J_3) & : \langle J_1, J_2 \rangle \notin T^2_{i,j}, J_3 \in T_{i,j} \\ \min\{U_{i,j}(J_3), U_{i,j+1}(J_1) + U_{i,j+1}(J_2)\} & : else \end{cases}$$

$$L_{i,j+1}(J_3) = \begin{cases} L_{i,j}(J_1) + L_{i,j}(J_2) & : \langle J_1, J_2 \rangle \in T^2_{i,j} \\ L_{i,j}(J_3) & : \langle J_1, J_2 \rangle \notin T^2_{i,j}, J_3 \in T_{i,j} \\ \max\{L_{i,j}(J_3), L_{i,j+1}(J_1) + L_{i,j+1}(J_2)\} & : else \end{cases}$$

$$A = \{J \in \{J_1, J_2, J_3\} : L_{i,j+1}(J) = U_{i,j+1}(J)\}$$

$$T_{i,j} \cup A$$

If for some $1 \le s \le 3$,

$$U_{i,j+1}(J_s) < L_{i,j}(J_s) - or - L_{i,j+1}(J_s) > U_{i,j}(J_s),$$

then no contingency table exists with fixed super-cells $T_{0,q}$ and concomitant fixed counts c(J), $J \in T_{0,q}$, so the algorithm stops. Otherwise, when j + 1 < q, the algorithm iterates to $\langle J_1, J_2, J_3 \rangle_{j+2} \in Q(T)$, or, when j + 1 = q, the algorithm either halts if $U_{i,q}(T) = U_{i+1,q}(T)$ and $L_{i,q}(T) = L_{i+1,q}(T)$, or iterates to i + 1 and $\langle J_1, J_2, J_3 \rangle_1 \in Q(T)$ otherwise.

If the GSA is not halted by encountering any inconsistencies between consecutive calculations of bounds for super-cell counts, then eventually there will be two consecutive iterations which do not change any of the bounds for the super-cell counts. In particular, an ***i*** will be reached such that $U_{i,q}(T) = U_{i+1,q}(T)$ and $L_{i,q}(T) = L_{i+1,q}(T)$. Of course, these bounds are not necessarily sharp; however, they are the optimal bounds possible utilizing the dependencies of Q(T).

The fact that the GSA will eventually stabilize in the absence of inconsistencies may be seen in various ways. One direct route is to observe that the GSA produces a decreasing sequence of integer-valued upper bounds, and an increasing sequence of integer-valued lower bounds, both of which therefore must stabilize. A small example will hopefully prove illuminating.

Consider the 2 × 2 contingency table (k=2, l_1=2, l_2 = 2) in Table 4.1, which we represent by the vector $\boldsymbol{n}=(\boldsymbol{n}(1,1),\boldsymbol{n}(1,2),\boldsymbol{n}(2,1),\boldsymbol{n}(2,2))'$. The marginals of that table are $(n_{1+}, n_{2+}, n_{+1}, n_{+2})'$, which represent the counts for the super-cells. There are a total of $(2^2 - 1)^2 = 9$ super-cells,

$$T = \{\langle 1,1\rangle, \langle 1,2\rangle, \langle 2,1\rangle, \langle 2,2\rangle, \langle 1,\{1,2\}\rangle, \langle 2,\{1,2\}\rangle, \langle \{1,2\},1\rangle, \langle \{1,2\},2\rangle, \langle \{1,2\},\{1,2\}\rangle\}.$$

Table 4.1 A Contingency Table With Marginals

n_{11}	n_{12}	n_{1+}
n_{21}	n_{22}	n_{2+}
n_{+1}	n_{+2}	n_{++}

For $1 \le i, \boldsymbol{j} \le 2$, cell $\langle i, j \rangle$ has cell count $\boldsymbol{n}(\boldsymbol{i},\boldsymbol{j})$, super-cell ($\boldsymbol{i}$, {1,2}) has count n_{i+}, super-cell ({1,2},$\boldsymbol{j}$) has count $\boldsymbol{n}_{+\boldsymbol{j}}$, and super-cell $\langle \{1,2\},\{1,2\} \rangle$ has count $\boldsymbol{n}_{++} = \boldsymbol{N}$. There are $\boldsymbol{q}$ = 6 members of ***Q(T)***, enumerated in the following visiting schedule:

$$1.\langle\langle 1,1\rangle,\langle 1,2\rangle,\langle 1,\{1,2\}\rangle\rangle$$
$$2.\langle\langle 1,1\rangle,\langle 2,1\rangle,\langle\{1,2\},1\rangle\rangle$$
$$3.\langle\langle 2,1\rangle,\langle 2,2\rangle,\langle 2,\{1,2\}\rangle\rangle$$
$$4.\langle\langle 1,2\rangle,\langle 2,2\rangle,\langle\{1,2\},2\rangle\rangle$$
$$5.\langle\langle 1,\{1,2\}\rangle,\langle 2,\{1,2\}\rangle,\langle\{1,2\},\{1,2\}\rangle\rangle$$
$$6.\langle\langle\{1,2\},1\rangle,\langle\{1,2\},2\rangle,\langle\{1,2\},\{1,2\}\rangle\rangle.$$

If $\boldsymbol{n}(2,2)$, $\boldsymbol{n}_{2+} = \boldsymbol{n}(2,\{1,2\})$, and $\boldsymbol{n}_{+2} = \boldsymbol{n}(\{1,2\},2)$ are initially fixed, and our visiting schedule corresponds to the enumeration of $\boldsymbol{Q(T)}$ above, then upon first iteration of the GSA, $\boldsymbol{n}(2,1)$, $\boldsymbol{n}(1,2)$, $\boldsymbol{n}_{1+} = \boldsymbol{n}(1,\{1,2\})$ and $\boldsymbol{n}_{1+} = \boldsymbol{n}(\{1,2\},1)$ are fixed, and after the second iteration, the remaining super-cell count $\boldsymbol{n}(1,1)$ is fixed as well. In this example, there is only one table with the super-cell counts which were initially fixed, so all of the cells will have upper and lower bounds which are the same.

In order to refine these bounds to sharp bounds, GSA is again implemented to discover the values of $\boldsymbol{t} \in \boldsymbol{T}$ between the bounds that correspond to feasible tables. These marginals can then be used to generate alternative synthetic tables that will be useful by transportation analysts while at the same time protecting the confidentiality of the individuals whose data is embedded in the tables. For examples, see Duncan et al. [10].

4.4 JOURNEY-TO-WORK TABLES

The synthetic table produced for the evaluation of the joint IPF/Bayesian approach (Table 3.1) has four zones of origin (home) and four zones of destination (work) along with sixteen income categories, and is repeated as Table 4.2. We implemented the generalized shuttle algorithm on this table with the three pairwise marginals (Home by Work, Home by Income, and Work by Income) initially fixed. This generated the bounds for each cell count in Table 4.3.

Despite the size of this table, there are quite a few marginal counts which are zero. This is caused by expected sparsity in the data. We expect some neighborhoods to originate no low income workers because they could not afford to live there; additionally, some destination blocks will not include any highly paid positions. These zero marginals lead many cell counts to be fixed based on the marginals. In particular, from the three sets of pairwise marginals in the example data, we know that there are definitely four individuals living in zone b and working in zone a in the smallest income category. Since this small cell count can be exactly determined from the marginals, there is a possibility of disclosure of these individuals.

If these were actual ACS survey data we had analyzed, instead of releasing the original ACS table, we could release a synthetic table. In order for the synthetic table to be a useful tool for transportation planning applications, we generate a table which agrees with the fixed marginal counts. This synthetic table will have cell counts within the bounds computed by the GSA. Therefore, fixing the three pairwise marginals in the example data set means that any synthetic table we create would have a count of four individuals living in zone b and working in zone a in the smallest income category. In this way, we could create and release synthetic data tables that will provide assessable privacy protection for the ACS survey respondents determined by the bounds computed by the GSA.

Table 4.2 Sample Data to be Disclosure Proofed

Ozone	Dzone	Inc1	Inc2	Inc3	Inc4	Inc5	Inc6	Inc7	Inc8	Inc9	Inc10	Inc11	Inc12	Inc13	Inc14	Inc15	Inc16	Freq
A	A	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
A	B	46	34	0	23	0	0	0	0	0	0	0	0	0	0	0	0	5
A	C	243	200	0	0	45	0	0	0	70	0	0	80	0	0	0	0	5
A	D	0	0	0	0	0	0	0	45	60	0	0	0	0	0	0	0	2
B	A	4	9	15	14	18	17	0	0	17	18	22	44	33	0	16	16	8
B	B	0	0	0	0	0	0	0	0	0	0	0	0	0	78	0	0	1
C	A	14	24	36	34	14	16	17	18	0	18	12	0	44	34	33	33	12
C	B	0	0	14	0	16	18	18	34	12	16	44	22	16	18	12	14	14
C	C	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	1
D	A	12	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
D	B	14	12	67	9	22	66	14	14	34	37	38	12	24	22	16	18	15
D	C	0	0	0	0	18	0	0	0	0	0	0	0	0	0	0	0	1
D	D	0	0	0	0	0	0	18	0	0	22	0	0	0	0	0	0	2

4.5 DISCLOSURE RISK

The GSA's utility in establishing cell-count bounds is a fundamental component of disclosure limitation, permitting an individual, to identify cells for which public release of data would present unacceptable disclosure risk. However, it bears remarking that within the context of the ACS-based journey-to-work tabulations, implementation of the GSA occurs at a second-stage of disclosure limitation, and is likely a measure of ancillary value in this regard. The primary means of disclosure limitation is implicit in the tables' construction, for the cross-classified data in these journey-to-work tables are acquired through the ACS, and are therefore an aggregate combination of sample tables collected over time from the U.S. population.

Sampling itself often provides effective protect against disclosure limitation since a collection of categorical values unique in a sample table are not necessarily unique in the population, where the latter is the focus of possible disclosure. The likelihood of non-uniqueness in demographic attributes within the population then attenuates the risk of identity disclosure when publicly releasing ACS-based journey-to-work tables, even when the tables themselves are completely uncensored. We proceed to discuss more formally the issue of sampling and disclosure risk attenuation, following the work of Skinner and Elliot [13].

Additionally, we comment on methods for producing synthetic data, especially as implemented in Abowd [1, 2, 3], which aim to limit disclosure risk by replacing information provided by actual respondents with that for "synthetic individual information" in the released data.

Table 4.3 Synthetic Data after Disclosure Proofing using the GSA Method

O	D	Inc1	Inc2	Inc3	Inc4	Inc5	Inc6	Inc7	Inc8	Inc9	Inc10	Inc11	Inc12	Inc13	Inc14	Inc15	Inc16
A	A	[0,9]	[0,9]	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
A	B	[46,55]	[25,34]	{0}	{23}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
A	C	{243}	{200}	{0}	{0}	{45}	{0}	{0}	{0}	{70}	{0}	{0}	{80}	{0}	{0}	{0}	{0}
A	D	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{45}	{60}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
B	A	{4}	{9}	[1,15]	{14}	[2,18]	[0,17]	{0}	{0}	[5,17]	[2,18]	[0,22]	[22,44]	[17,33]	[0,34]	[4,16]	[2,16]
B	B	{0}	{0}	[0,14]	{0}	[0,16]	[0,17]	{0}	{0}	[0,12]	[0,16]	[0,22]	[0,22]	[0,16]	[44,78]	[0,12]	[0,14]
B	C	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
B	D	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
C	A	{14}	{24}	[36,50]	{34}	[14,30]	[16,33]	{17}	{18}	[0,12]	[18,34]	[12,34]	[0,22]	[44,60]	[0,34]	[33,45]	[33,47]
C	B	{0}	{0}	[0,14]	{0}	[0,16]	[1,25]	{18}	{34}	[0,12]	[0,16]	[22,44]	[0,22]	[0,16]	[18,52]	[0,12]	[0,14]
C	C	{0}	{0}	{0}	{0}	{0}	[0,7]	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
C	D	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
D	A	[12,21]	[9,18]	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
D	B	[5,14]	[12,21]	{67}	{9}	[22,29]	[59,66]	{14}	{14}	{34}	{37}	{38}	{12}	{24}	{22}	{16}	{18}
D	C	{0}	{0}	{0}	{0}	[11,18]	[0,7]	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}	{0}
D	D	{0}	{0}	{0}	{0}	{0}	{0}	{18}	{0}	{0}	{22}	{0}	{0}	{0}	{0}	{0}	{0}

4.6 DISCLOSURE AND SAMPLE DATA

Our discussion is based upon a general framework originally established by Bethlehem et al. [5], and subsequently used by Skinner and Elliot. Suppose a sample ***s*** is chosen from a population, and identity information is collected for each individual in ***s*** via measurement on a set of ***identifying variables.*** In addition, sensitive information is recorded on each member of s, income being one example. The sensitive information may be a subset of the identity information. A contingency table is made for the sample from a cross-classification of ***s*** using the identifying variables. The possible values of the identifying variables define the cells of the contingency table and we say that the subpopulation of all individuals whose values equal a particular set of values corresponds to that cell. The public release of a contingency table with accompanying sensitive information might risk disclosure of this sensitive information if a cell count is nonzero, and in addition, only a few individuals comprise the subpopulation corresponding to the cell itself.

The nature of disclosure risk is made particularly transparent in the case that a cell with cell-count one corresponds to a set of values for the identifying variables that are possessed by only one member of the overall population. In this case, if an ***intruder*** matches a known individual's identifying variable values to this cell, the intruder immediately gains access to the sensitive information of his victim. Another scenario illustrating the risk of disclosure when releasing cross-tabulations occurs when an intruder is cognizant that all cells with count one correspond to unique individuals in the population. In this instance, an intruder chooses a target individual from the population and simply matches the individual's values of the identifying variables against those of all the cells with single counts. If the individual's values are found to agree with that of any of these cells, the intruder then immediately gains access to the individual's sensitive information. The same general technique may be used by an intruder who can verify only that each cell with count one correspond to a small number of individual's in the population. In this case, a correct match between a target's values for the identifying variables and a cell with count one only permits an intruder to gain sensitive information on the individual with some probability. Of course the danger lies in the fact that matching may not be a prohibitively labor-intensive task, so an intruder may test many different individuals against the contingency table, and identify members of the subpopulation corresponding to the cell with count one.

In contrast, suppose that we are able to determine that a particular cell with a cell-count of one corresponds to values of the identifying variables that are shared by a large number of individuals in the overall population, say ν individuals. In this case, the risk of disclosure even if cell-count information on this cell were to be made publicly available would be quite minimal. The

best that an intruder can hope for is to be able to isolate all members of the population that share the cell's values for the identifying variables, in which case, absent any further information, the probability that the intruder correctly correlates an individual from the population with the sensitive information provided with the contingency table is $1/\nu$. Therefore, disclosure risk is quite small even when ν can only be bounded below by 100. Returning to the scenario in which an intruder matches a randomly chosen individual against all cells with single counts, then again, even if a positive match is found, the risk of disclosure is small if we know that all the sub-populations corresponding to cells of single counts are quite large, say, greater than or equal to 100. In this case, the probability the intruder correctly correlates an individual from the population with the sensitive information provided with the contingency table is no greater than $1/\nu$.

The preceding is meant to illustrate that the release of uncensored contingency tables with small counts does not necessarily pose a risk of disclosure when the contingency table is based on a sample from an underlying population. Indeed, in such case, an integral component of disclosure risk assessment is the ability to either estimate or ascertain the number of individuals in the overall population corresponding to cells with small counts. If we are able to satisfy ourselves that the magnitude of any subpopulation corresponding to a low-count cell is fairly large, than our risk disclosing sensitive information even upon the release of a completely uncensored contingency table is acceptably small. On the other hand, if there exists a cell with a low count, and the subpopulation corresponding to it is of limited membership, then a release of uncensored contingency table information may pose a serious risk of identity disclosure.

We need to stress here that subpopulation sizes corresponding to cells of low count is absolutely vital information in assessing disclosure risk for sample tables. In very large populations, one expects most of these subpopulations to be, in relative terms, non-trivial. Although proportionally these subpopulations may represent only a tiny fraction of the overall population, they are large enough to be sampled, and so the sheer size of the overall population demands that the subpopulations must not have an insubstantial number of members. Suppose we were to base a contingency table on a simple random sample of size 100 from a population of 10,000 people, then a cell with count one would imply that approximately 100 people in the overall population correspond to the cell.

The risk of disclosure for cross-classified data from a sample can be assessed with one of three measures. To describe these measures, we adopt the terminology of Skinner and Elliot [13]. An individual is called ***population unique*** if no other members of the population have the same values of the identifying variables as the individual. An individual in the sample is called ***sample unique*** if no other members of the sample have the same values of the identifying variables as the individual.

The first measure of risk, denoted as ***Pr(PU),*** is the proportion of the overall population that is population unique. If an intruder randomly samples from the population, the probability he finds a population unique individual is ***Pr(PU).*** If the intruder is capable of linking a population unique individual to any cell with a single count, the intruder immediately acquires the individual's sensitive information. Therefore, whether a population unique individual is at risk of having sensitive information disclosed depends solely upon whether the individual is additionally a member of the sample.

The second measure of risk, denoted as ***P(Pr|SU),*** is simply the proportion of the sample unique individuals that are simultaneously population unique. If an intruder randomly chooses a single-count cell, and is able to match the identifying variables of every individual in the population against this cell, then ***P(Pr|SU)*** is the probability the intruder will acquire sensitive information on some individual.

The third and final measure was introduced by Skinner and Elliot [13]. Let ***V*** be the union of all subpopulations that correspond to some sample unique individual. Skinner and Elliot denote their measure as θ***,*** and calculate it as the number of sample unique individuals over the number of individuals in ***V.*** Suppose that an intruder randomly chooses an individual from the overall population. θ is then just the probability that the individual is a sample unique given that they are a member of ***V,*** i.e., given that the value of the identifying variables for this individual agrees with those of one of the sample unique individuals. This is the probability that an intruder will find the individual whose sensitive information is actually presented.

These measures can be adapted to accommodate intruders who use different attack methods than those here discussed; however, Skinner and Elliot [13] argue that under the most likely attack scenarios, θ provides the most appropriate, or at least most useful, measure of disclosure risk. Furthermore, Skinner and Elliot show that consistent inference for θ can be achieved simply and without the reliance upon strong modeling assumptions, unlike those required for consistent estimation of ***Pr(U)*** or ***P(Pr|SU).***

The subpopulations corresponding to each cell in the ACS tables include every individual who matched the identifying variables in any of the three years over which the survey was aggregated. This will make θ slightly more difficult to estimate and smaller than if ACS data were collected at one fixed time.

4.7 SYNTHETIC DATA

The group of Abowd et al. [1, 2, 3] used journey to work data of a similar type to the ACS data. They performed a privacy analysis on their method of creating synthetic data for origin and destination block pairs. In that application, the desired output were individual records to be used in mapping software, which correspond to a two variable contingency table.

The goal was to examine journey-to-work paths and no data other than origin (home) and destination (work) were to be released.

In the creation of synthetic data sets for journey to work tables, Abowd et al. [1] sought to protect against an intruder who had obtained the identifying and sensitive information for every individual in the sample except one. They used a probabilistic criteria, probabilistic differential privacy (pdp), to weight (across possible synthetic datasets and across individuals) the risks of disclosing the sensitive information of the last individual. This weighted total risk is limited instead of an unweighted total of risks across all possibilities.

The criteria of protecting against an intruder who knows information on all but one sample individual is stronger than that imposed on the CTPP tables. In any table of real data, the final individual could be disclosed with probability one if all other individuals were known. This criteria would require that synthetic data be produced in lieu of real data for every table released, even those with no small counts.

To provide protection at the level of these stringent criteria, synthetic data sets are created to be released in the place of the real data. It is possible that using the synthetic data, the intruder's information about the first $\boldsymbol{n} - 1$ individuals would not match the data released. Even if there were matches to the information about the real people, there is some probability that the sensitive data value for the last synthetic data point does not match the sensitive data value for the real person they are trying to discover. Therefore, the best that an intruder with near perfect information can do is get a possible value for the sensitive information of the last individual. The probability that this value is correct depends on the algorithm for creating the synthetic data.

Using entirely synthetic data to achieve this extremely high level of disclosure prevention creates the added burden of determining whether the synthetic data are representative of the original data. If the synthetic data do not represent the true underlying behaviors, then it is possibly damaging to release this data to policy makers and analysts. Abowd et al. [1, 2, 3] propose a criteria for evaluating representativeness of synthetic datasets which preserves the probabilistic disclosure limitation of the synthetic data.

The Generalized Shuttle Algorithm (GSA) derives from formal contingency table methodology and we can use it to compute sharp bounds for a table of counts given any set of marginals. The algorithm is computationally intensive[13]; however, it is only necessary to implement the GSA once for each set of tables from which data are to be released. In the past, this type of disclosure limitation strategy was thought by many to be far too complicated to implement in statistical practice. As computational power continues to increase,

[13] The actual computational task is a variant of one that is known to be NP-hard.

we believe that utilizing procedures such as GSA will not only be feasible in practice but could easily become routine.

When we compare our approach with the other methods presented above for the ACS data, these sharp bounds for each cell entry produce clear measures of the risk of data disclosure. In contrast, while the tables generated with the earlier methods appear to protect confidentiality, it is unclear whether or not an intruder could learn the identity or additional information about individuals in the database, c.f., [11]. The sharp bounds produced by GSA show the types of tables that are feasible for a given set of marginals, and we believe that these will prove useful for generating alternative synthetic tables that will be useful for local planning by transportation analysts.

The disclosure protection of examining θ**,** the chance of finding a sample unique individual from among the subpopulation which corresponds to the identifying variables, is a useful criteria for examining disclosure risk in a table generated from a sample. In combination with the generalized shuttle algorithm, it can inform the Census Bureau of the chance of disclosure for sample data if the original table is released and the cells which will be de facto released by releasing certain marginal totals.

4.8 REFERENCES

[1] Abowd, J.M., Machanavajjhala, A., Kifer, D., Gehrke, J. and Vilhuber, L. (2008). "Privacy: Theory Meets Practice On the Map." ***International Conference on Data Engineering (ICDE).***

[2] Wu, J.S. and Graham, M.R. (2008) "OnTheMap: An Innovative Mapping and Reporting Tool." U.S. Census Bureau. Available at:

<http://unstats.un.org/unsd/statcom/statcom_09/seminars/innovation/ Innovation%20Seminar/USA-OntheMap.pdf >

[3] Wu, J. and Abowd, JM. (2007). "Synthetic Data for Administrative Record Applications at LEHD. Joint Statistical Meetings Invited Session 82 presentation. U.S. Census Bureau, PN-2007-05. Available at:

<http://lehd.did.census.gov/led/library/presentations/Wu-Abowd-20070831.pdf>

[4] Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). ***Discrete Multivariate Analysis: Theory and Practice.*** MIT Press, Cambridge, MA. Reprinted (2007), Springer-Verlag, New York.

[5] Bethlehem, J.G., Keller, W.J., and Pannekoek, J. (1990). "Disclosure control of micro-data." ***Journal Of The American Statistical Association,*** 85, 38-45.

[6] Dobra, A. and Fienberg, S.E. (2000). "Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. ***Proceedings of the National Academy of Sciences,*** 97, No. 22, 11885-11892.

[7] Dobra, A. and Fienberg, S.E. (2003). "Bounding entries in multi-way contingency tables given a set of marginal totals. In Y. Haitovsky, H.R. Lerche, and Y. Ritov, eds., ***Foundations of Statistical Inference: Proceedings of the Shoresh Conference 2000.*** PhysicaVerlag, Heidelberg, 3-16.

[8] Dobra, A. and Fienberg, S.E. (2009) "The generalized shuttle algorithm," In P. Gibilisco, E. Riccomagno, M.-P. Rogantin, eds., ***Algebraic and Geometric Methods in Probability and Statistics.*** Cambridge University Press, New York, in press.

[9] Dobra, A., Fienberg, S.E., Rinaldo, A., Slavkovic, A.B., and Zhou, Y. (2008). "Algebraic statistics and contingency table problems: Log-linear models, likelihood estimation, and disclosure limitation." In M. Putinar and S. Sullivan, eds., ***Emerging Applications of Algebraic Geometry,*** IMA Series in Applied Mathematics. Springer-Verlag, New York, 63-88.

[10] Duncan, G.T., Fienberg, S.E., Krishnan, R., Padman, R., and Roehrig, S.F. (2001), "Disclosure limitation methods and information loss for tabular data." In P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, (eds.), ***Confidentiality, Disclosure and Data Access.*** North-Holland, Amsterdam, 135-166.

[11] Fienberg, S.E. and Slavkovic, A.B. (2008). "A survey of statistical approaches to preserving confidentiality of contingency table entries." In C. Aggarwal and P.S. Yu, eds., ***Privacy Preserving Data Mining: Models and Algorithms.*** Springer-Verlag, New York, 289-310.

[12] Lauritzen, S.L. (1996). ***Graphical Models.*** Oxford University Press, New York.

[13] Skinner, C.J. and Elliot, M.J. (2002). "Measure of disclosure risk for microdata." ***Journal of the Royal Statistical Society, Series B,*** 855-867.