Addressing Margins of Error in Small Areas of Data Delivered through the American Factfinder of the Census Transportation Planning Products Program

Final Report

NCHRP Project 8-36C, Task 135

Authors
Jane Li, Tom Krenzke

November 30, 2017

The information contained in this report was prepared as part of NCHRP Project 08-36(135), National Cooperative Highway Research Program.

Prepared for:
Larry Goldstein
Senior Program Officer
Cooperative Research Programs
Transportation Research Board
500 5th Street NW
Washington, D.C. 20001
(202) 334-1866

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

Table of Contents

Acknowledgements

The authors are grateful to Greg Erhardt, a Senior Analyst from RAND, whose travel modeling and transportation planning expertise was helpful in advising the plans for the tasks from the viewpoint of the end-user. Westat Data Scientist John Riddles provided invaluable skills to the project and the authors greatly appreciate his insights and efficient approach to data processing.

Executive Summary

The data users of the CTPP are provided the point estimates and the MOEs for each table cell; they do not have access to the microdata, that is, no data for individual households or persons can be accessed, and transportation planners use the CTPP to validate, or further calibrate their model results. The project team considered issues related to the aggregation of table cells and the computation of MOEs.

The Traffic Analysis Zones (TAZs) in the CTPP were made to be small so that transportation planners could piece them together in different ways for planning purposes. Combining cells of tables, either by combining TAZs, or combining categories of variables, may lead to greater precision. However, improvements can be made to the current practice of treating subgroups as independent when estimating the MOE for combined subgroups.
Generalized Variance Functions (GVFs) have been used to stabilize MOEs for surveys. For the purposes of stabilizing the MOEs and addressing the issue of estimating MOEs for aggregated estimates, the project team proposes and evaluates alternatives to the current MOE approach when aggregating subgroups. When the CTPP tables are aggregated to obtain estimates for larger geographies, we found that an adjusted GVF performed well in deriving MOEs. Given the natural desire to address high MOEs through aggregation, it is important that the aggregate MOEs be both reliable and easily calculable. Therefore, it is important that the research address both the mechanism to understand and communicate the MOEs as well as the stability of the MOEs themselves. The project team provides general guidance in Appendix A to help users understand and communicate the magnitude of the MOEs when combining subgroup estimates, or comparing subgroup estimates. In addition, the project team has created a simple CTPP MOE ToolKit to help compute the MOEs when aggregating table cells, and to conduct statistical testing when comparing two subgroups. The ToolKit was created to compute the estimated MOEs using four different procedures: the naïve approach, traditional GVF, weighted-adjusted GVF, and unweighted-adjusted GVF.

In the second main task of the project, an investigation was conducted for an approach that leads to a better understanding of the impact of MOEs when validating travel models. The objective was to facilitate a way to allow sampling error and perturbation error to propagate through to subsequent analysis and visually display results. The proposed solution is an approach that the project team refers to as "replicated tables". R code has been developed to allow users to prepare the replicated tables for a given CTPP table or aggregated table. The same function was also added to the CTPP

MOE ToolKit. The methodology was developed and evaluated to allow CTPP users to gauge the sensitivity of their results through the use of replicated tables that are generated from the original CTPP's estimates and published MOEs, or aggregated estimates and estimated MOEs. In practice, the transportation researchers can generate the replicated tables multiple times using the developed R function or the CTPP MOE ToolKit and fit the simulated values to their desired transportation models, one replicated table at a time. The replicated tables approach can be used as a diagnostic tool that provides a sensitivity assessment that takes into account the impact of the sampling and perturbation variance components in the CTPP tables. Various graphs (such as scaled bar chart, bar charts, and pie charts) can be produced to get a picture of the variation across the replicated tables, which helps the user visualize the precision of the CTPP estimates, given the MOE. The graphs allow the users to visualize the sampling and perturbation errors in the CTPP estimates

1 Introduction

The Census Transportation Planning Products (CTPP) comprise a set of special tabulations that are produced to meet the needs of transportation planners in understanding local journey-to-work patterns. The tables relate worker and household characteristics to travel mode based on the worker's residence, workplace, and travel from residence to workplace. The residence-based, workplace-based, and residence-to-workplace flow tables involve dozens of variables and provide cell aggregates, means, medians, and estimated margins of error (MOEs)[1] for small geographic units such as census tracts and Traffic Analysis Zones (TAZs) that are roughly the size of Census blocks or block groups. The 2006-2010 CTPP are based on five years of American Community Survey (ACS) data.

MOEs are commonly used in these transportation data to reflect the precision of the resulting estimates. Due to small sample sizes, the MOEs are important to take into account because the CTPP is generated from the five-year ACS sample, which replaced the Census long form as the data source. Figure 1-1 illustrates the large MOEs associated with mean travel time for a small number of TAZ flows in Fulton County, Atlanta, extracted from the CTPP data tool system. Such MOEs cannot be ignored in analysis.

Means of Transportation 4		Total, means of transportation		Car, truck, or van -- Drove alone		Car, truck, or van -- Carpooled		Public transportation, bicycle, walked, taxicab, m...	
Output		Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error	Estimate	Margin of Error
RESIDENCE	WORKPLACE								
TAZ 00000100, Fulton Coun...	TAZ 00000101, Fulton Coun...	43.9	35	2	1.5	-	**	60.4	40.5
TAZ 00000101, Fulton Coun...	TAZ 00000101, Fulton Coun...	7.2	4.6	5.5	2.4	-	**	8.7	7.7
TAZ 00000102, Fulton Coun...	TAZ 00000101, Fulton Coun...	9.5	6.7	9.5	6.7	-	**	-	**

Figure 1-1 Margins of error for mean travel time for select TAZ flows in Fulton County, Atlanta, Georgia

The reduced sample size of ACS compared to the Census long form sample is the main cause of inflated MOEs. The ACS sample sizes in the small geographic areas are generally small and, indeed, the smallest TAZs may have only 20 to 25 ACS sample workers (in the five-year ACS sample). In particular, the tables of flows for each TAZ consist of a large number of cells containing a very small

[1] In the CTPP tables, MOEs are calculated as standard errors multiplied by 1.645, where 1.645 is the z score associated with the standard normal distribution at the 90 percent confidence level.

number of workers. With millions of CTPP tables, including for small geographic units such as census tracts and TAZs, the ACS sample is spread too thinly. The American FactFinder also produces journey-to-work data from the ACS for larger geographic areas, however, they may also have small sample sizes since they may be based on a one-year sample.

Key Uses of ACS and CTPP Data

The ACS data, whether accessed through the American FactFinder or the CTPP, provide an important data source in travel demand modeling and transportation planning. To provide context for framing the challenges, common uses are described here, with the recognition that other applications abound.

Travel Demand Modeling Applications

The data provide an important source of information for calibrating travel demand models. Even with the reduced sample size, the sample size remains larger than most household travel surveys, providing an important means for observing the terms that are included. Most often, the data are used in modeling as a source of calibration targets. That is, data summaries are compared to equivalent summaries of model outputs, providing a basis for making model adjustments to better fit the observed data. In addition, the data are often used to develop socio-economic inputs to travel models, specifically the number of households in each TAZ for different classifications.

Several of the most important uses are:

- District-to-district worker flows for comparison to workplace location or home-based work trip distribution models.
- Commute mode shares for calibrating work mode choice models.
- Household vehicle ownership, by district, for calibrating auto ownership models.
- Small geographic area summaries of households by size, income, number of workers, household composition, and age of householder for the purpose of developing socio-economic data inputs by TAZ.

If the MOEs are high, it becomes difficult to distinguish between sampling error in the data and other errors in the models. Similarly, there remains a trade-off between the desires to analyze small geographic areas and aggregating those areas to achieve more certainty in the data estimates.

Transportation Planning Applications

In addition to their use in developing and calibrating travel models, the ACS data are frequently used directly in transportation planning applications. Several important examples include:

- Monitoring changes in commute mode shares, household composition or other measures.
- Assessing neighborhood population characteristics and the resulting transportation needs.
- Mapping low-income, minority and disadvantaged populations for the purpose of environmental justice analysis.

In each of these cases, there is a strong need to understand the reliability of the data, particularly as they are tabulated for smaller geographic areas.

The main challenge is to appropriately handle the levels of precision in the ACS/CTPP estimates when developing long range plans, validating and calibrating traffic models, and providing information to decision makers to deploy necessary funds. Accurate measures of precision are needed when conveying the information for decision making.

Addressing the Challenges of ACS and CTPP MOEs

The data users of the CTPP are provided the point estimates and the MOEs for each table cell; they do not have access to the microdata, that is, no data for individual households or persons can be accessed, and transportation planners use the CTPP to validate, or further calibrate their model results. The project team considers issues related to the aggregation of table cells and the computation of MOEs. The TAZs in the CTPP were made to be small so that transportation planners could piece them together in different ways for planning purposes. Much concern has been that the point estimates (weighted frequencies, means, medians) are unstable, however, the same concern exists for the measures of precision (e.g., MOEs). Combining cells of tables, either by

combining TAZs, or combining categories of variables, may lead to greater precision. However, the improvements can be made to the current practice of treating subgroups as independent when estimating the MOE for combined subgroups. GVFs have been used to stabilize MOEs for surveys. For the purposes of stabilizing the MOEs and addressing the issue of estimating MOEs for aggregated estimates, the project team proposes and evaluates alternatives to the current MOE aggregation approach, as presented in Section 2. When the CTPP tables are aggregated to obtain estimates for larger geographies, we found that an adjusted GVF performed well in deriving MOEs. Given the natural desire to address high MOEs through aggregation, it is important that the aggregate MOEs be both reliable and easily calculable. Therefore, it is important that the research address both the mechanism to understand and communicate the MOEs as well as the stability of the MOEs themselves. The project team provides general guidance in using MOEs in Appendix A when combining cell estimates, or comparing cell estimates. In addition, the project team has created a simple CTPP MOE ToolKit to help compute the MOEs when aggregating table cells, and to conduct statistical testing when comparing two subgroups.

In the second main task of the project, an investigation was conducted for an approach that leads to a better understanding of the impact of MOEs when validating travel models. The objective was to facilitate a way to allow sampling error and perturbation error to propagate through to subsequent analysis and visually display results. The proposed solution is an approach that the project teams refers to as "replicated tables", as described in Section 3. R code, provided in Appendix B has been developed to prepare the replicated tables for a given CTPP table or aggregated table. A series of worksheets in the CTPP MOE ToolKit can also be used to generate replicated tables.

Generalized Variance Functions for Computing MOEs for Aggregate Estimates

2

Margins of error (MOEs) are commonly used in these transportation data to reflect the precision of the resulting estimates. Due to small sample sizes, the MOEs of the CTPP tables are unstable. Data users may want to combine areas or categories of variables to arrive at more precise estimates. A challenge is to estimate the MOE when aggregating geographic areas that result in an area that is not published, or more generally, when aggregating any table cell estimates that result in an estimate that is not published in the set of tables.

There are different options to compute the MOE for aggregated estimates. Asiala (2012) evaluated the resulting MOEs associated with aggregated estimates under the current practice. Suppose the estimates X_1 and X_2 are to be combined. In general, when aggregation is done, the resulting MOE is equal to $MOE = \sqrt{MOE_{X_1}^2 + MOE_{X_2}^2 + Covariance\ term}$. However, in the context of static tables provided in the American FactFinder (ACS) and CTPP, the covariance between estimates X_1 and X_2 is not provided. Therefore, the common practice is to treat estimates X_1 and X_2 as independent (i.e., $MOE = \sqrt{MOE_{X_1}^2 + MOE_{X_2}^2}$). This method is referred to the "Naïve" estimator hereafter. Asiala explained that the resulting MOE is an overestimate (when aggregating) and may seriously break down when aggregating more than four estimates. Asiala suggested to aggregate the fewest number of estimates as possible, to try aggregating in different ways and see how sensitive the calculated MOE is to the method, to calculate the estimates using the Public Use Microdata Sample (PUMS), and to request a special tabulation (fee based and certain criteria apply).

Development of Approach to Address the Challenge

One potential solution is to consider generalized variance functions (GVFs), which have been used in large national surveys mainly as a way to easily estimate the variance associated with a resulting point estimate (count, rate, proportion, means, etc). GVFs were used for the ACS PUMS from 2000 to 2004. The GVF approaches were re-evaluated using a design effect approach for counts, means and medians in Fuller (2010) for computing standard errors at the state and public use microdata area levels.

Using a GVF to compute the MOE is simple once the model has been established. For example, a subset of estimated weighted frequencies for each subgroup or table cell c (X_c), and the associated directly estimated relative variances ($V_{X_c}^2$) are selected[2]. The GVF is a curve of the form $V_{X_c}^2 = a + \frac{b}{X_c}$, where a and b are parameters to be estimated by an iterative weighted least squares process. Diagnostics such as residual analyses, residual plots with influential observations marked, and Cook's D influence statistics based on deletion of influential observations can be used to check the model fitting. The parameters (a and b) can be published such that the user can use the aggregated estimate for the combined group or set of cells g, X_g, in the formula to estimate its relative variance, which results in an estimated MOE through algebra ($MOE = 1.645 \times \sqrt{aX_g^2 + bX_g}$). The linear regression provides a smoothed set of variances, which is largely considered as an approach to stabilize the variance estimates.

For a proportion, $p_g = \frac{X_g}{X}$, as given by Wolter (1985), where X represents an estimate for a certain subpopulation, the estimated relative variance is approximated as $V_{p_g}^2 \approx V_{X_g}^2 - V_X^2$. As shown in Appendix A, this leads to $Var_{p_g} = p_g^2(\frac{SE_{X_g}^2}{X_g^2} - \frac{SE_X^2}{X^2})$. A special case results when using the same GVF function for both X_g and X, which result in $Var_{p_g} = \frac{b}{X} p_g(1 - p_g)$. Note that the use of this formula assumes independence between the numerator X_g and the denominator X. The GVF functions for statistics like means, ratios, etc., may take different forms and are not in the scope of this project.

We have developed and evaluated a new GVF approach using a heuristic adjustment for the purpose of improving the precision of variance estimates when combining estimates. The adjustment comes from the likelihood that a GVF could not do well for the numerous possibilities of combining subgroups. For example, a set of TAZs being combined may each be impacted by a design effect[3]. Because variances are impacted by the sample design and weighting adjustments, conceptually a GVF model would be created for each subgroup associated with the same magnitude of design effect. Furthermore, because the CTPP contains many tables and estimates, it is impractical to develop many GVF functions to appropriately account for impact from stratification, differential sampling rates, clustering, weighting adjustment for nonresponse, etc. The research team recognized this issue and have investigated a heuristic solution to adjust the GVF variance estimates for the given subgroups (table cells) that are combined, in order to improve the precision of the GVFs. If a

[2] Relative variances are equal to the variance divided by the estimate squared.

[3] Design effect is ratio of the variance of a complex sample over the variance of a simple random sample.

set of *C* cells are combined to obtain a combined cell estimate X_g, the actual MOEs and the GVF-based MOEs of each cell estimate can be used to adjust the GVF-based MOE of the aggregated estimate, X_g. For example, suppose the user wants to combine 10 TAZs within a Traffic Analysis District (TAD). A TAD is a combination of TAZs and has over 20,000 residents[4]. There can be multiple TADs within a state. If the GVF-based MOEs overestimate the actual MOEs for a majority of the 10 TAZs, the GVF based MOE of X_g should be adjusted downward to avoid overestimation. The adjusted GVF variance estimate is defined as

$$var_{gvf-adj}(X_g) = var_{gvf}(X_g)/f, \quad (2.1)$$

where *f* is the adjustment factor. Two options were considered to compute the factor *f* -- unweighted and weighted. The unweighted-adjusted GVF method computes *f* as the unweighted mean of the ratio of the GVF-based variance to the actual variance across all the cells to be combined.

$$f_{uw} = \frac{1}{C}\sum_{c=1}^{C} f_c, \quad (2.2)$$

Where, *C* = number of cells to combine, and $f_c = var_{gvf}(X_c)/var_{actual}(X_c)$.

The weighted-adjusted GVF method computes the mean of the ratio of the GVF-based variance to the actual variance, weighted by the cell's point estimate, X_c.

$$f_w = \frac{\sum_{c=1}^{C} f_c * X_c}{\sum_{c=1}^{C} X_c}. \quad (2.3)$$

For both adjustment methods, in the CTPP MOE ToolKit, cell estimates with counts less than 20 are excluded from calculating the factors since their GVF-based variances may be unstable.

Model Estimation

The initial data set includes all tables from CTPP. The following were then excluded:

- Geography levels other than TAZ and TAD

[4] More information on the formation of TAZs and TADs is available at: http://ctpp.transportation.org/Documents/TAZ_MTPS_Participant_Guidelines_Final.pdf

- The three evaluation tables: one from residence tables (B102203), one workplace (B203206C3), and one flows (B302105)
- Any data points with missing MOEs

The data were then divided into 15 (3x5) strata. The strata were divided by type and size:

- 3 Types: Residence, Workplace, Flow
- 5 Sizes: 1-1000, 1001-2000, 2001-5000, 5001-10,000, 10,001-100,000

Data points with a count of 0 or a count greater than 100,000 were excluded. A new sample was drawn from the stratified data as follows:

- The Flow/10,001+ stratum contained only a few hundred data points, so all data in the stratum were included.
- For each 1001+ stratum, excluding the Flow/10,001+ stratum, a systematic sample of size 1500 was selected.
- Due to the size of the three 1-1000 strata and the computer intensity of the selection process (especially the sorting of records), a random sample of size 1 million was initially selected. From this initial sample, a systematic sample of size 1500 was drawn.
- A new set of parameter values for a and b was estimated for the full sample. Parameters were estimated for each of the residence, workplace and flow tables separately, however, initial evaluation results did not show enough improvement to justify having separate parameter estimates.

A weighted least squares regression $V_{X_c}^2 = a + \frac{b}{X_c}$ was fit, where the weights were $\frac{1}{(\hat{V}_{X_c}^2)^2}$, where $\hat{V}_{X_c}^2$ is based on the values from the last iteration. The a and b parameters for the GVF were estimated for relative variances related to the estimated number of workers. The resulting estimated parameters were $a = -0.00023$ and $b = 24.8988$. We emphasize that the approach should be used only when the combined cell estimate X_g is between 0 and 100,000, which is within the range of the modeled X_c's.

Evaluation

An evaluation was conducted to compare different methods for computing the precision (in terms of MOEs) of the aggregated estimates. We used three evaluation tables (one from residence tables

(B102203), one workplace (B203206C3), and one flows (B302105)) which were not included in the estimation of the GVF parameters a and b.

A nationally representative subset of 200 TADs was selected with probability proportionate to the number of workers from a list sorted by state and Metropolitan Planning Organization (MPO). All TAZs from within the selected TADs were selected. For the evaluation set of tables, the chosen residence table was Means of Transportation (MOT) by age group, which provides a large range of totals. The age group has seven categories (16-17; 18-24; 25-44; 45-59; 60-64; 65-74; 75 and over) and MOT has 3 categories (Car, truck, or van -- Drove alone; Car, truck, or van – Carpooled; Public transportation, others or worked at home). The chosen workplace table was presence of children by MOT. The presence of children has two categories (with children; no children), and MOT is defined as in the chosen residence table. The chosen flow table was minority status, which consists of two categories (Non-Hispanic White; Other). The weighted frequencies (estimated number of workers) in the CTPP tables had been rounded to multiples of 5 (some cells have counts of 0 or 4) by the Census Bureau.

In the evaluation, we viewed the TAD-level estimates as the aggregates of the corresponding TAZ-level estimates and computed the MOEs using different approaches illustrated above (Naive estimator, GVF estimator, GVF with unweighted adjustment, and GVF with weighted adjustment). After that, we compared those MOEs against the actual MOEs published in the CTPP tables for TADs and TAZs, which serve as the gold standard in the comparison.

Ratios of MOEs were computed using the MOEs published in the CTPP tables as the base for each method. Two sets of results are provided for each residence, workplace and flows table. The first result table shows select percentiles and interquartile range (IQR) of the ratios by the number of TAZs to combine in a TAD. The second result table shows the percentage of MOE ratios that are not too far from 1 for each method, by the number of TAZs to combine. We computed the percentage of MOE ratios that are within 20% (or 10%) of 1, i.e., the ratios fall into the interval (0.8, 1.2) (or (0.9, 1.1)).

After review of the initial results, it was decided to use the one set of GVF parameters that was derived from the full sample. There is only marginal improvement seen, if any, by fitting to residence, workplace and flows, separately. The evaluation criteria were based on the following:

- Extreme MOE. This is measured by the difference between the 95th percentile and the 5th percentile.

- Stable MOE. This is measured by the interquartile range (IQR).
- Median ratio closest to 1.00
- Interval, that is, the percentage of MOE ratios within the range of 0.8 to 1.2.

The following notation was used for the approaches:

N = Naïve
G = Traditional GVF
U = Unweighted adjusted GVF
W = Weighted adjusted GVF

From the review of Tables 2-1 through 2-6, the following rankings were observed for each type of table, for each criterion. For example, (WU)GN means the top two (the two GVF adjusted methods) were indistinguable, followed by the traditional GVF and the Naïve approach.

- Residence
  - Extreme… (WU)GN
  - Stable… (WU)NG
  - Median… WUNG
  - Interval… UWNG
- Workplace
  - Extreme… (WU)NG
  - Stable… (WU)NG
  - Median… NWUG
  - Interval… (WUN)G
- Flows
  - Extreme… (WU)GN
  - Stable… (WUN)G
  - Median… (WN)UG

- Interval… WNUG

Of the 12 scenarios, the weighted-adjusted GVF is in top 2 in all 12, unweighted-adjusted is in the top 2 in 9 scenarios, Naïve is in the top 2 in 5 scenarios, and the traditional GVF is never in the top 2. For this evaluation, based on the selected tables, the weighted adjustment works well to correct GVF estimated MOEs (the median of the MOE ratios is very close to 1), and tends to outperform the other approaches.

Figure 2-1 shows the fit of the traditional GVF curve through the data points, for the residence, workplace and flows evaluation tables, respectively. The line is a nonlinear fitted line that reflects the traditional GVF model $V_{X_g}^2 = a + \frac{b}{X_g}$. The naive, unweighted, and unadjusted estimates are not included in the plots.

Table 2-1 Distribution of MOE ratios for select residence table for evaluation, by number of TAZs combined and method

# of TAZs combined	# of TAD estimates	MOE methods[1]	5th percentile	25th percentile	Median	75th percentile	95th percentile	IQR
<10	2,956	Naïve	0.91	0.99	1.04	1.12	2.97	0.13
		Trad. GVF	0.85	1.09	1.25	1.43	1.76	0.35
		U-adjusted GVF	0.76	0.90	0.99	1.08	1.24	0.17
		W-adjusted GVF	0.82	0.93	1.00	1.08	1.25	0.15
10-49	2,480	Naive	0.94	1.04	1.15	1.31	1.95	0.28
		Trad. GVF	0.85	1.07	1.24	1.44	1.85	0.37
		U-adjusted GVF	0.76	0.89	0.98	1.08	1.28	0.20
		W-adjusted GVF	0.82	0.92	1.00	1.10	1.29	0.17
50-99	191	Naive	0.97	1.11	1.19	1.34	1.61	0.23
		Trad. GVF	0.87	1.01	1.16	1.32	1.65	0.32
		U-adjusted GVF	0.76	0.89	0.96	1.06	1.23	0.17
		W-adjusted GVF	0.79	0.92	0.99	1.08	1.23	0.16
>=100	87	Naive	1.00	1.18	1.29	1.48	1.83	0.31
		Trad. GVF	0.83	1.12	1.28	1.41	1.85	0.29
		U-adjusted GVF	0.70	0.87	0.95	1.07	1.22	0.20
		W-adjusted GVF	0.73	0.90	0.98	1.09	1.28	0.20

[1] U is abbreviation for "Unweighted" and W is an abbreviation for "Weighted"

Table 2-2 Proportion of MOE ratios within specified range for select residence table for evaluation, by number of TAZs combined and method

# of TAZs combined	# of TAD estimates	MOE methods[1]	MOE Ratios $\in(0.8, 1.2)$	MOE Ratios $\in(0.9, 1.1)$
<10	2,956	Naive	0.82	0.67
		Trad. GVF	0.39	0.20
		U-adjusted GVF	0.84	0.55
		W-adjusted GVF	0.89	0.61
10-49	2,480	Naive	0.59	0.38
		Trad. GVF	0.42	0.22
		U-adjusted GVF	0.81	0.50
		W-adjusted GVF	0.86	0.56
50-99	191	Naive	0.55	0.22
		Trad. GVF	0.55	0.31
		U-adjusted GVF	0.82	0.55
		W-adjusted GVF	0.86	0.56
>=100	87	Naïve	0.30	0.15
		Trad. GVF	0.33	0.14
		U-adjusted GVF	0.80	0.47
		W-adjusted GVF	0.78	0.52

[1] U is abbreviation for "Unweighted" and W is an abbreviation for "Weighted"

Table 2-3 Distribution of MOE ratios for select workplace table for evaluation, by number of TAZs combined and method

# of TAZs combined	# of TAD estimates	MOE methods[1]	5th percentile	25th percentile	Median	75th percentile	95th percentile	IQR
<10	679	Naïve	0.32	0.74	0.98	1.04	1.19	0.31
		Trad. GVF	0.87	1.01	1.10	1.22	1.53	0.21
		U-adjusted GVF	0.77	0.87	0.93	1.00	1.12	0.13
		W-adjusted GVF	0.80	0.90	0.95	1.02	1.12	0.12
10-49	1,067	Naïve	0.60	0.86	0.98	1.07	1.26	0.21
		Trad. GVF	0.88	1.01	1.11	1.25	1.54	0.24
		U-adjusted GVF	0.74	0.84	0.90	0.97	1.06	0.12
		W-adjusted GVF	0.79	0.89	0.94	1.01	1.09	0.12
50-99	99	Naïve	0.65	0.86	0.95	1.03	1.22	0.17
		Trad. GVF	0.78	0.93	1.05	1.16	1.30	0.23
		U-adjusted GVF	0.66	0.78	0.85	0.91	1.01	0.13
		W-adjusted GVF	0.68	0.82	0.90	0.96	1.02	0.13
>=100	51	Naïve	0.76	0.93	0.99	1.05	1.11	0.12
		Trad. GVF	0.55	0.93	1.08	1.17	1.47	0.24
		U-adjusted GVF	0.43	0.76	0.82	0.88	0.93	0.12
		W-adjusted GVF	0.45	0.79	0.89	0.91	0.97	0.12

[1] U is abbreviation for "Unweighted" and W is an abbreviation for "Weighted"

Table 2-4 Proportion of MOE ratios within specified range for select workplace table for evaluation, by number of TAZs combined and method

# of TAZs combined	# of TAD estimates	MOE methods[1]	MOE Ratios $\in(0.8, 1.2)$	MOE Ratios $\in(0.9, 1.1)$
<10	679	Naïve	0.69	0.57
		Trad. GVF	0.70	0.43
		U-adjusted GVF	0.89	0.55
		W-adjusted GVF	0.95	0.68
10-49	1,067	Naïve	0.73	0.52
		Trad. GVF	0.66	0.41
		U-adjusted GVF	0.87	0.50
		W-adjusted GVF	0.94	0.65
50-99	99	Naïve	0.82	0.57
		Trad. GVF	0.69	0.51
		U-adjusted GVF	0.67	0.33
		W-adjusted GVF	0.85	0.48
>=100	51	Naïve	0.93	0.79
		Trad. GVF	0.70	0.33
		U-adjusted GVF	0.60	0.14
		W-adjusted GVF	0.72	0.30

[1] U is abbreviation for "Unweighted" and W is an abbreviation for "Weighted"

Table 2-5 Distribution of MOE ratios for select flow table for evaluation, by number of TAZs combined and method

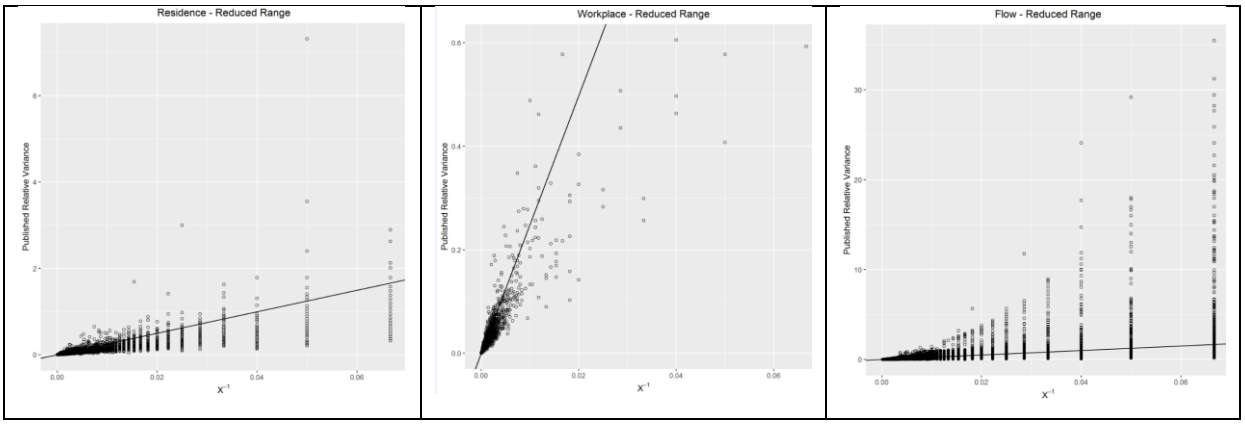
# of TAZs combined	# of TAD estimates	MOE methods[1]	5th percentile	25th percentile	Median	75th percentile	95th percentile	IQR
<10	54,215	Naïve	0.90	0.99	1.03	1.08	2.98	0.09
		Trad. GVF	0.85	1.08	1.27	1.45	1.74	0.37
		U-adjusted GVF	0.84	0.98	1.06	1.13	1.28	0.15
		W-adjusted GVF	0.88	0.99	1.06	1.13	1.28	0.14
10-49	4,713	Naïve	0.92	1.00	1.06	1.16	1.92	0.16
		Trad. GVF	0.91	1.08	1.21	1.35	1.62	0.27
		U-adjusted GVF	0.80	0.91	1.00	1.09	1.27	0.18
		W-adjusted GVF	0.85	0.95	1.02	1.11	1.28	0.16
50-99	117	Naïve	0.93	1.00	1.07	1.18	1.36	0.17
		Trad. GVF	0.97	1.06	1.16	1.22	1.33	0.16
		U-adjusted GVF	0.79	0.86	0.91	0.98	1.08	0.12
		W-adjusted GVF	0.82	0.90	0.94	1.01	1.09	0.11
>=100	16	Naïve	0.98	1.03	1.05	1.26	1.26	0.09
		Trad. GVF	1.02	1.09	1.14	1.36	1.36	0.09
		U-adjusted GVF	0.82	0.88	0.89	1.03	1.03	0.07
		W-adjusted GVF	0.86	0.91	0.94	1.09	1.09	0.06

[1] U is abbreviation for "Unweighted" and W is an abbreviation for "Weighted"

Table 2-6 Proportion of MOE ratios within specified range for select flow table for evaluation, by number of TAZs combined and method

# of TAZs combined	# of TAD estimates	MOE methods[1]	MOE Ratios $\in(0.8, 1.2)$	MOE Ratios $\in(0.9, 1.1)$
<10	54,215	Naïve	0.86	0.74
		Trad. GVF	0.37	0.20
		U-adjusted GVF	0.85	0.56
		W-adjusted GVF	0.87	0.59
10-49	4,713	Naïve	0.79	0.62
		Trad. GVF	0.48	0.24
		U-adjusted GVF	0.86	0.56
		W-adjusted GVF	0.88	0.61
50-99	117	Naïve	0.79	0.56
		Trad. GVF	0.68	0.32
		U-adjusted GVF	0.92	0.55
		W-adjusted GVF	0.99	0.72
>=100	16	Naïve	0.94	0.69
		Trad. GVF	0.81	0.25
		U-adjusted GVF	1.00	0.48
		W-adjusted GVF	1.00	0.81

[1] U is abbreviation for "Unweighted" and W is an abbreviation for "Weighted"



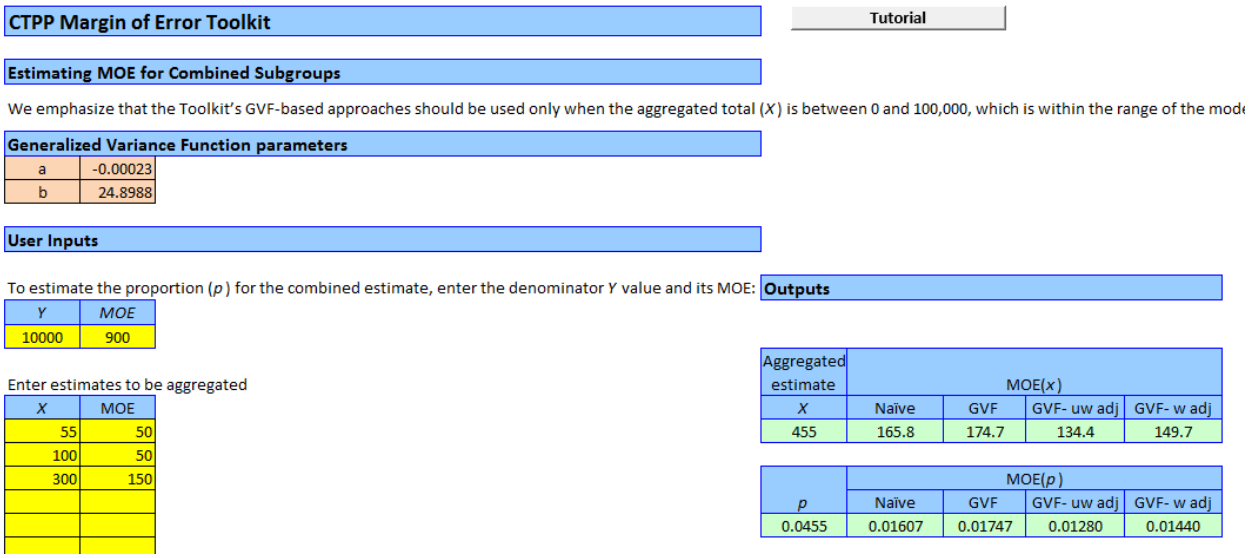
Note: The line is a nonlinear fitted line reflects the traditional GVF model $V_{X_g}^2 = a + \frac{b}{X_g}$.

Figure 2-1 Fit of the traditional GVF curve through the data points, for the residence, workplace and flows evaluation tables, respectively

As mentioned above, a simple toolkit (CTPP MOE ToolKit) has been developed to help with the computations. The CTPP MOE ToolKit is an Excel file that serves three purposes:

- To estimate MOE for totals and proportions for combined subgroups
- To compare proportions between two subgroups
- To replicate tables to reflect the published MOEs, for use in subsequent sensitivity travel demand analysis results (will discuss later)

The sheet named "CombineSubgroups" serves the purpose of estimating the MOE for a combination of subgroups (levels) from the CTPP tables. The MOE is estimated for the aggregated total X from the combined subgroups. Combined subgroups can be areas (e.g., groups of TAZs) or levels of a variable (e.g., categories of Means of Transportation). It also produces the Naïve, the traditional GVF and the adjusted GVFs. Figure 2-2 provides a screenshot of the interface that provides the user the ability to estimate the MOEs when combining subgroups.



CTPP Margin of Error Toolkit

Tutorial

Estimating MOE for Combined Subgroups

We emphasize that the Toolkit's GVF-based approaches should be used only when the aggregated total (*X*) is between 0 and 100,000, which is within the range of the mod

Generalized Variance Function parameters

a	-0.00023
b	24.8988

User Inputs

To estimate the proportion (*p*) for the combined estimate, enter the denominator *Y* value and its MOE:

Y	*MOE*
10000	900

Enter estimates to be aggregated

X	MOE
55	50
100	50
300	150	

Outputs

Aggregated estimate	MOE(*x*)			
X	Naïve	GVF	GVF- uw adj	GVF- w adj
455	165.8	174.7	134.4	149.7

	MOE(*p*)			
p	Naïve	GVF	GVF- uw adj	GVF- w adj
0.0455	0.01607	0.01747	0.01280	0.01440

Figure 2-2 CTPP MOE ToolKit Screenshot of Combining Subgroups Sheet

The sheet named "CompareSubgroups" has the purpose of comparing proportions between two subgroups. To statistically compare two subgroups, the MOE of the estimated difference in proportions needs to be computed. There are two scenarios that impact the MOE computations:

- The two subgroups that are being compared are independent from each other.
- The two subgroups that are being compared are dependent on each other.

Figure 2-3 provides the user with the ability to compare two subgroups. Appendix A provides some information about the computations in the ToolKit. We acknowledge that other organizations have developed toolkits for producing MOEs when combining estimates from the CTPP. From our review, the available toolkits use the Naïve approach. Further work on the CTPP MOE Toolkit can be done to expand, or the organizations with existing toolkits can consider the methods used in this section. User guidance for both sheets in the CTPP MOE ToolKit is provided.

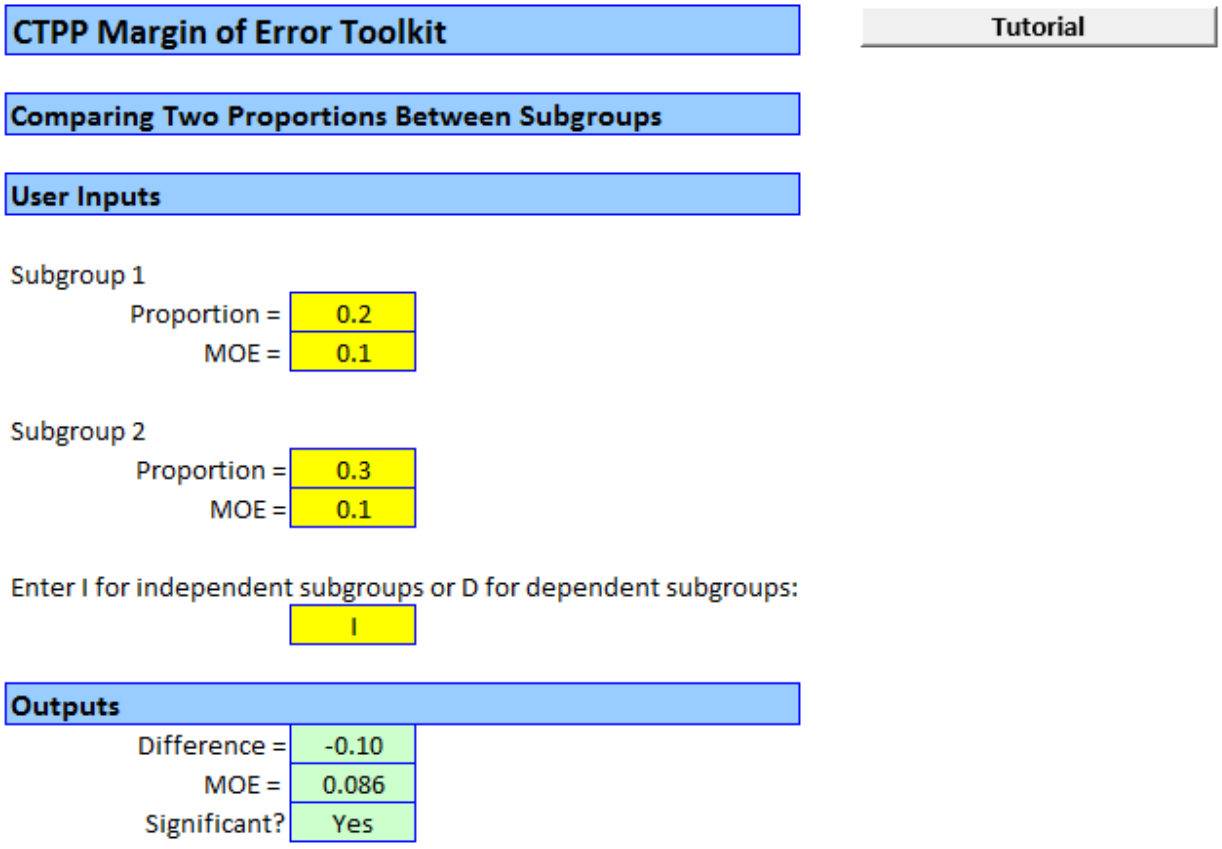


Figure 2-3 CTPP MOE ToolKit Screenshot of Comparing Subgroups Sheet

Handling MOEs through Replicated Tables 3

The main challenge is to appropriately handle the levels of precision in the CTPP estimates when developing long range plans, validating and calibrating traffic models, and providing information to decision makers to deploy necessary funds. Incorporating the sampling error associated with CTPP estimates in travel demand modeling is difficult to do. At the very least, accurate measures of precision are needed when conveying the information for decision making.

Development of Approach to Address the Challenge

A solution we developed and evaluated under this contract is referred to the "replicated tables approach". It allows sampling and perturbation error from the ACS to propagate through subsequent usage of the CTPP tables. Given a table with its point estimates and MOEs, it is possible to generate replicated (or simulated) copies of those tables. Each replicated table is a representation of the original table. To illustrate, suppose the original CTPP table of residence-to-workplace flows A and B crossed with Means of Transportation (MOT) 1 and 2 is as shown in Table 3-1.

Table 3-1 Original CTPP table

		MOT	
		1	2
Flow	A	100	200
	B	300	400

The result is a set of *M* tables, illustrated below in Table 3-2. Once the tables are replicated, each table can be used in the analysis, and the variation across the results can be computed. In another context, graphs can be generated for each simulated table. An advantage of the simulated tables approach is that it is robust to the type of graph. Whether the graphs of interest are heat maps, pie charts, or histograms, simply show the same graph multiple times, side-by-side, to display the variation.

Table 3-2 Replicated tables

1	2	3	...	M
	1 2			
A	110 230			
B	305 370			
	1 2			
A	70 220			
B	230 410			
	1 2			
A	95 150		...	
B	320 370			
	1 2			
A	90 300			
B	320 380			

A goal of the replicated tables is to arrive at a set of *M* tables such that the variation across the tables is approximately the same as the MOE in the single published CTPP table or in an aggregated table which combines multiple published CTPP tables. The challenge is to incorporate the correlation between the cell estimates, that is, the cell estimates are not independent of one another. The Dirichlet distribution (Connor et al, 1969) is useful for simulating contingency tables since it takes into account the relationship between the table cells. For example, in one application the Dirichlet distribution was used to synthesize results for the Longitudinal Employment Household Dynamics (LEHD) OnTheMap system (Machanavajjhala et al (2008)). A given CTPP table has *K* cells. Assume the overall weighted frequency X for the table follows a normal distribution, $N(\mu, \sigma^2)$. Suppose X_k is the weighted frequency in cell *k*. The cell proportion in cell *k* can be expressed as $p_k = \frac{X_k}{X}$. Assume that the cell proportions $(p_1, p_2, \dots p_k, \dots, p_K)$ are the realizations of a set of random variables that follow a Dirichlet distribution model with parameters $(\alpha_1, \alpha_2, \dots \alpha_k, \dots, \alpha_K)$. The means and variances of the cells proportions are $E(p_k) = \frac{\alpha_k}{\alpha_0}$ and $Var(p_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0{}^2(\alpha_0+1)}$, where $\alpha_0 = \alpha_1 + \alpha_2 + \cdots + \alpha_K$.

A basic algorithm for the replicated tables approach is as follows. First, randomly draw the overall weighted frequency, X, from a normal distribution.. Next, randomly draw a set of cell proportions, p_k's , from the Dirichlet distribution. Then derive the cell counts of a replicated table as $X_k = Xp_k$. Repeat the above steps *M* times to generate *M* replicated tables. After producing *M* replicated tables, the resulting variation among the replicated tables is checked.

We need to set up the parameters of the normal distribution and the Dirichlet distribution to generate the replicated tables. The parameters of the normal distribution μ and σ^2 can be estimated by the overall weighted total and its associated variance (MOE^2 divided by 1.645^2) for a published CTPP table. For an aggregated table which combines multiple published CTPP tables, the adjusted GVF variance from (2.1) may be used. Two methods are proposed to estimate the parameters $(\alpha_1, \alpha_2, \dots \alpha_k, \dots, \alpha_K)$ of the Dirichlet distribution.

GVF Method

The parameters of the Dirichlet distribution can be estimated from the observed cell proportions $(p_1, p_2, \ldots p_k, \ldots, p_K)$. Set

$$\hat{\alpha}_k = \left(\frac{X}{b} - 1\right) p_k \tag{3.1}$$

where $b = 24.8988$ is estimated from the GVF model in Section 2. As a result,

$$Var(p_k) = \frac{\hat{\alpha}_k(\hat{\alpha}_0 - \hat{\alpha}_k)}{\hat{\alpha}_0{}^2(\hat{\alpha}_0 + 1)} = \frac{b}{X} p_k(1 - p_k), \tag{3.2}$$

where $\hat{\alpha}_0 = \hat{\alpha}_1 + \cdots + \hat{\alpha}_K$. Note that (3.2) is identical to the formula which uses the GVF functions to estimate the variance of a cell proportion in Section 2. If the GVF based variances are not good approximations of the observed CTPP variances, this method will not perform well since the variation among the cell estimates in the replicated tables will reflect the variances computed from the GVF method, not the variances in the published CTPP table (or the weighted GVF adjusted variances in the aggregated table).

Distance Function Method

As an alternative to the GVF method, the parameters of the Dirichlet distribution can be estimated through a distance function method so that the variation among the replicated tables may be a better approximation of the published CTPP variances or the GVF adjusted variances. The goal is to find a factor f which minimizes the distance function

$$\sum_k \left(\frac{\hat{\alpha}_k(\hat{\alpha}_0 - \hat{\alpha}_k)}{\hat{\alpha}_0{}^2(\hat{\alpha}_0 + 1)} - v_k\right)^2, \tag{3.3}$$

where $\hat{\alpha}_k = f p_k$ and v_k is the variance of a cell proportion for cell k computed from the method given by Wolter (1985) (See Appendix A for more details):

$$var(p_k) = p_k{}^2 \left(\frac{var(X_k)}{X_k{}^2} - \frac{var(X)}{X^2}\right) = v_k$$

where $var(X_k)$ and $var(X)$ are either the CTPP published variances (MOE/1.645) or the weighted-adjusted GVF variances of X_k and X, respectively.

The GVF method is a special case of the distance function method where $f = \frac{X}{b} - 1$. The distance function method may improve the GVF method in the sense that it finds an f that minimizes the distance between the variances from the Dirichlet distribution, $\frac{\hat{\alpha}_k(\hat{\alpha}_0 - \hat{\alpha}_k)}{\hat{\alpha}_0{}^2(\hat{\alpha}_0+1)}$, and the variances of the cell proportions derived from the observed CTPP variances (or the weighted-adjusted GVF variances). In the case that the GVF based variances are very close to the observed CTPP variances or the weighted-adjusted GVF variances, the Dirichlet parameters estimated from the GVF method and the distance function method would be very similar.

The Census Bureau published the "Variance Replicate Tables" for select tables from the American Communities Survey (ACS) 5-year (2011-2015) (https://www.census.gov/programs-surveys/acs/data/variance-tables.html). For this select set of tables, in addition to the original published estimates and MOEs, the standard errors and 80 variance replicate estimates (*Var_Rep*) are provided. The ACS uses a successive differences replication (SDR) variance estimation methodology to derive the MOEs in tables (https://www2.census.gov/programs-surveys/acs/replicate_estimates/2015/documentation/5-year/2011_2015_Variance_Replicate_Tables_Documentation.pdf). The SDR variance is calculated using the ACS estimate and the 80 variance replicate estimates (*Var_Rep1* to *Var_Rep80*). The variance is the sum of the squared differences between the ACS estimate and each of the 80 variance replicate estimates, multiplied by 4/80.

$$variance = \frac{4}{80}\sum_{i=1}^{80}(Var\_Rep_i - ACS\ estimate)^2$$

The MOE is calculated by multiplying the standard error (the square root of the variance) by the factor 1.645, which is associated with a 90 percent confidence level. Although this select set of tables include a wide range of topics, they do not cover all the CTPP tables. The Census Bureau has no plan to add more tables for the special tabulations of CTPP.

The variance replicate estimates (with additional information – a factor as discussed below) make it possible to calculate MOEs for the user-defined aggregated tables without using approximation formulas. However, it should be noted that measuring the variation among the Census Bureau's variance replicate estimates directly does not give correct variance estimates. That is, by using the

replicate tables directly, we mean conducting a sensitivity analysis by using a number of the replicated tables directly in travel demand models to see the variation in the travel demand model results. The variance replicate estimate tables from the Census Bureau would yield results with not enough variation. Technically, this can be seen by the factor used in the SDR variance estimator of 4/80 (shown in equation 1 on page 4 of the Census Bureau documentation cited above). That is, using the Census Bureau's replicated tables directly in the sensitivity assessment, the user is conceptually applying a factor of 1/80, not 4/80, when reviewing the results. Therefore, the variation that is seen in the results for this purpose, is not enough. There should be four times more variation.

The approaches (GVF and distance function) designed in this section serve the purpose of generating the right amount (or close to it)) of variation among the replicated tables for the purpose of seeing how sensitive the transportation model results are to the sampling error and perturbation error in the CTPP estimates.

Evaluation

We demonstrate the use of replicated tables through an example. Table 3-3 shows the weighted frequencies and the original MOEs, along with three other columns

- Original standard error (SE), computed by original MOE divided by 1.645
- GVF based standard error (SE), computed from the traditional GVF method
- Cell proportions

for a published CTPP table "Age by Means of Transportation (MOT)" in a TAD (MPO ID is 34198200 and TAD ID is 00000063). This table has 18 internal cells. The parameters for Dirichlet distribution $\hat{\alpha}_k$ were estimated using both methods described above. An R function *rep.tab* is provided to generate replicated tables based on the algorithm described above (see Appendix B for more details about the R function).

Tables 3-4 and 3-5 show the five replicated tables in columns Rep1 through Rep5 which were generated from the GVF method and the distance function method, respectively. The parameters of the Dirichlet distribution, Alphas, are also listed in a column for each table. The Alphas are slightly different for the two methods. The standard deviation (SD) among the five replicated tables for each

cell estimate was computed. As the number of replicated tables increases, the standard deviation among the replicate estimates gets closer to the traditional GVF-based standard errors or the CTPP standard errors (or GVF weighted adjusted standard error depending upon how v_k was computed), as illustrated in formulae (3.2) and (3.3). Therefore, in both tables we included the standard deviation among ten thousand randomly drawn replicated tables.

In Table 3-6, four columns were taken from Tables 3-3, 3-4, and 3-5 for comparison purposes.

- Original standard error (SE), computed by original MOE divided by 1.645
- GVF based standard error (SE), computed from the traditional GVF method
- GVF-based standard deviation (SD) among 10,000 replicated tables
- Difference Function standard deviation (SD) among 10,000 replicated tables

As expected, the GVF-based SDs among 10,000 replicated tables are almost identical to the GVF SEs. The difference function SDs among 10,000 replicated tables are very close to the original SEs but they are not the same. The distance function method intends to find a solution which minimizes the distance between the variances of the Dirichlet distribution and the CTPP variances (or the GVF weighted adjusted variances), but the solution may not be able to equalize the two sets of variances. However, Table 3-6 does show that the difference function SDs among 10,000 replicated tables are closer to the original SEs than the GVF-based SDs among 10,000 replicated tables for most of the table cells. In the case that the traditional GVF variances do not approximate the true CTPP variances well, the distance function method would be able to generate a set of replicated tables which works better to reflect the CTPP variances.

An R function *rep.tab* was developed by the Westat team to fulfill the task of generating replicated tables. Appendix B provides relevant guidance on the use of this R function. Also, a series of worksheets are added to the CTPP MOE ToolKit for creating replicated tables using the two methods illustrated above. A tutorial is prepared to guide the use of the ToolKit including the function of generating replicated tables.

Table 3-3 Original Age by Means of Transportation (MOT) table (in MPO 34198200 and TAD 00000063)

Cell	AGE	MOT	Weighted Frequency	Original MOE	Original SE =MOE/1.645	GVF SE	Cell Percent p_k
1	18-24	CAR,TRK,VAN,DROVE_ALONE	350	126	77	93	0.024
2	25-44	CAR,TRK,VAN,DROVE_ALONE	2755	444	270	259	0.193
3	45-59	CAR,TRK,VAN,DROVE_ALONE	1585	258	157	197	0.111
4	60-64	CAR,TRK,VAN,DROVE_ALONE	290	125	76	85	0.020
5	65-74	CAR,TRK,VAN,DROVE_ALONE	160	80	49	63	0.011
6	75+	CAR,TRK,VAN,DROVE_ALONE	25	38	23	25	0.002
7	18-24	CARPOOL	175	88	53	66	0.012
8	25-44	CARPOOL	705	221	134	132	0.049
9	45-59	CARPOOL	475	164	100	109	0.033
10	60-64	CARPOOL	70	62	38	42	0.005
11	65-74	CARPOOL	15	21	13	19	0.001
12	16-17	OTHER	40	44	27	32	0.003
13	18-24	OTHER	1115	223	136	166	0.078
14	25-44	OTHER	4180	453	275	316	0.292
15	45-59	OTHER	1730	288	175	206	0.121
16	60-64	OTHER	210	100	61	72	0.015
17	65-74	OTHER	365	111	67	95	0.026
18	75+	OTHER	55	53	32	37	0.004
Total			14300	820	498	556	

Table 3-4 Five replicated tables generated from alphas based on the GVF method (method 1)

Cell	AGE	MOT	Alpha (Method 1)	Weighted Frequencies					SD among 5 replicated tables	SD among 10,000 replicated tables
				Rep1	Rep2	Rep3	Rep4	Rep5		
1	18-24	CAR,TRK,VAN,DROVE_ALONE	14.03	559	482	299	393	396	99	93
2	25-44	CAR,TRK,VAN,DROVE_ALONE	110.46	2750	2873	2646	2705	2835	93	257
3	45-59	CAR,TRK,VAN,DROVE_ALONE	63.55	1776	1653	1569	1658	1570	85	195
4	60-64	CAR,TRK,VAN,DROVE_ALONE	11.63	279	242	298	261	415	68	84
5	65-74	CAR,TRK,VAN,DROVE_ALONE	6.41	194	205	95	147	201	47	63
6	75+	CAR,TRK,VAN,DROVE_ALONE	1.00	16	33	24	96	11	35	25
7	18-24	CARPOOL	7.02	184	188	105	237	192	48	66
8	25-44	CARPOOL	28.27	572	556	651	815	763	114	132
9	45-59	CARPOOL	19.04	381	457	407	500	684	120	108
10	60-64	CARPOOL	2.81	10	72	35	4	37	27	43
11	65-74	CARPOOL	0.60	35	22	48	1	5	20	19
12	16-17	OTHER	1.60	114	35	16	18	13	43	32
13	18-24	OTHER	44.70	973	1031	873	1114	1187	122	165
14	25-44	OTHER	167.59	4444	4147	4708	4495	3569	441	307
15	45-59	OTHER	69.36	1440	1807	1807	1878	1623	178	203
16	60-64	OTHER	8.42	258	130	272	215	141	66	72
17	65-74	OTHER	14.63	451	358	454	418	280	74	96
18	75+	OTHER	2.21	124	122	48	25	50	46	37
Total				14559	14410	14353	14981	13970		

Table 3-5 Five replicated tables generated from alphas based on the distance method (method 2)

Cell	AGE	MOT	Alpha (Method 2)	Weighted Frequencies					SD among 5 replicated tables	SD among 10,000 replicated tables
				Rep1	Rep2	Rep3	Rep4	Rep5		
1	18-24	CAR,TRK,VAN,DROVE_ALONE	16.97	533	489	289	449	367	98	85
2	25-44	CAR,TRK,VAN,DROVE_ALONE	133.58	2716	2975	2523	3189	2621	273	236
3	45-59	CAR,TRK,VAN,DROVE_ALONE	76.85	1737	1713	1492	1720	1545	114	177
4	60-64	CAR,TRK,VAN,DROVE_ALONE	14.06	277	257	283	339	247	36	77
5	65-74	CAR,TRK,VAN,DROVE_ALONE	7.76	189	209	96	132	200	49	58
6	75+	CAR,TRK,VAN,DROVE_ALONE	1.21	17	35	24	18	8	10	23
7	18-24	CARPOOL	8.48	182	195	106	158	160	34	60
8	25-44	CARPOOL	34.18	577	593	624	736	848	115	119
9	45-59	CARPOOL	23.03	385	477	393	327	526	79	98
10	60-64	CARPOOL	3.39	13	76	37	47	54	23	38
11	65-74	CARPOOL	0.72	31	20	52	102	1	39	18
12	16-17	OTHER	1.94	107	38	87	31	43	39	29
13	18-24	OTHER	54.06	973	1080	1281	1173	1310	140	152
14	25-44	OTHER	202.67	4363	4314	4490	3822	3743	339	286
15	45-59	OTHER	83.88	1447	1871	1584	1925	1502	219	189
16	60-64	OTHER	10.18	252	143	106	147	204	58	65
17	65-74	OTHER	17.70	438	374	455	359	382	42	87
18	75+	OTHER	2.67	117	121	49	13	49	47	33
Total				14353	14980	13970	14687	13810		

Table 3-6 Comparison summary table for replicated tables evaluation

Cell	AGE	MOT	Original SE =MOE/1.645	GVF SE	GVF-based SD among 10,000 replicated tables	Difference Function SD among 10,000 replicated tables
1	18-24	CAR,TRK,VAN,DROVE_ALONE	77	93	93	85
2	25-44	CAR,TRK,VAN,DROVE_ALONE	270	259	257	236
3	45-59	CAR,TRK,VAN,DROVE_ALONE	157	197	195	177
4	60-64	CAR,TRK,VAN,DROVE_ALONE	76	85	84	77
5	65-74	CAR,TRK,VAN,DROVE_ALONE	49	63	63	58
6	75+	CAR,TRK,VAN,DROVE_ALONE	23	25	25	23
7	18-24	CARPOOL	53	66	66	60
8	25-44	CARPOOL	134	132	132	119
9	45-59	CARPOOL	100	109	108	98
10	60-64	CARPOOL	38	42	43	38
11	65-74	CARPOOL	13	19	19	18
12	16-17	OTHER	27	32	32	29
13	18-24	OTHER	136	166	165	152
14	25-44	OTHER	275	316	307	286
15	45-59	OTHER	175	206	203	189
16	60-64	OTHER	61	72	72	65
17	65-74	OTHER	67	95	96	87
18	75+	OTHER	32	37	37	33

4 Summary

For the NCHRP Project 8-36C, Task 135, methodology has been developed and evaluated to estimate MOEs based on a heuristic weighted-adjusted GVF approach. A CTPP MOE ToolKit was created to compute the estimated MOEs using the weighted-adjusted GVF approach and three other procedures. The ToolKit can also be used to compare subgroup estimates. Guidance has been written and provided in Appendix A on the use of MOEs.

In addition, methodology was developed and evaluated to allow CTPP users to gauge the sensitivity of their transportation model results by using several replicated tables that are generated from the original CTPP's estimates and published MOEs, or aggregated estimates and estimated MOEs. In practice, the transportation researchers can generate the replicated tables multiple times using the developed R function or the CTPP MOE ToolKit and fit the simulated values to their desired transportation models, one replicated table at a time. The replicated tables approach can be used as a diagnostic tool that provides a sensitivity assessment that takes into account the impact of the sampling and perturbation variance components in the CTPP tables. Various graphs can be produced using the R function to get a picture of the variation across the replicated tables, which helps the user visualize the precision of the CTPP estimates, given the MOE.

References R

Asiala, M. (2012). Topics on American Community Survey. Presented at the California Regional / Affiliate Data Center meeting, June 1, 2012. http://www.dof.ca.gov/research/demographic/state_census_data_center/meetings/documents/CASDC_AnnualMtg2012_Asiala-ACSUpdate.pdf (accessed 12/26/2014)

Connor, R. J., Mosimann, J. E (1969). Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association*. American Statistical Association. 64 (325): 194–206

Fuller, S. (2010). Analyzing Generalized Variances for the American Community Survey 2005 Public Use Microdata Sample. Final Report, April 20, 2010. Decenniel Statistical Studies Division. U.S. Bureau of the Census. (Accessed January 6, 2015, http://www.census.gov/acs/www/Downloads/library/2010/2010_Fuller_01.pdf).

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. ICDE 2008. Proceedings of the 2008 IEEE 24th International Conference on Data Engineering; April 7-12; Cancun, Mexico. Washington, DC: IEEE Computer Society; 2008: 227–286.

Appendix A

CTPP Margin of Error (MOE) User Guidance

This appendix provides relevant guidance on the use of the Census Transportation Planning Products (CTPP) margins of error (MOE).

What is the purpose of the MOE?

The MOE is a measure of the precision of an estimate.

What does the MOE represent?

The CTPP MOE represents the half-width of a 90% confidence interval. Suppose the estimated proportion (p) of workers in a specific traffic analysis zone (TAZ) who drive alone, as opposed to other means of transportation, is 0.70, and the MOE is 0.20. Then we are 90 percent confident the true proportion is between 0.50 and 0.90. The interval tells you how precise the estimate is (or sometimes, how imprecise).

What can you conclude about the precision of estimates?

The chance that the 90 percent confidence interval (derived from the MOE) contains the true parameter value (e.g., proportion, total) is not uniform across the interval. That is, assuming the estimated parameter value is an unbiased estimate, there is a relatively higher chance that the true value is close to the estimated parameter value. Table A-1 below provides the probability that the interval contains the true value as a function of the MOE. For example, there is a 32 percent chance the true value is contained within +/- 0.25 of the MOE, and a 50 percent chance within +/- 0.41 of the MOE.

Table A-1 Relationship between portions of the MOE and the probability of containing the true parameter value

Portion of MOE	Probability of containing the true estimated proportion
1 MOE	0.90
0.5 MOE	0.59
0.41 MOE	0.50
0.25 MOE	0.32
0.1 MOE	0.13

Sometimes a rule of thumb is used based on the relative standard error (sometimes referred to as the coefficient of variation, or CV). The relative standard error is the ratio of the standard error to the estimated value (of the proportion, total, etc). For example, if the CV is no more than 0.1, the point estimate may be considered as having good precision; if the CV is between 0.1 and 0.3, some caution is needed; if the CV is between 0.3 and 0.5, a warning is given to the point estimate; and if the CV is greater than 0.5, the point estimate should not be published. Table A-2 presents the rule of thumb based on the ratio of the MOE to the estimated value. This rule of thumb has limitations, such as the relative standard error is not as useful for small (or large) estimated proportions. Another limitation is that p could easily be considered 1-p, however, the relative standard error for p is not the same as the relative standard error for 1-p. When estimates are used under caution, warning or not published, one solution may be to combine areas or categories to arrive at estimates at higher precision.

Table A-2 Rule of thumb intervals for the precision of estimated proportions

$\frac{MOE\ of\ X}{X}$	Rule of Thumb
<= 0.1*1.645	Good
>0.1*1.645 and <=0.3*1.645	Caution
>0.3*1.645, and <=0.5*1.645	Warning
>0.5*1.645	No publish

How does one compute the MOE when combining subgroups?

Subgroups can be defined in different ways, such as combining TAZs or combining means of transportation categories for a TAZ. In any case, the MOE will need to be estimated for the resulting combined subgroup.

MOEs for totals

The following four MOE estimates are available in the CTPP MOE ToolKit for the total X_g from combining C subgroups:

1. Naïve: The estimated MOE is derived by treating each subgroup as independent. This is known to be unstable, however it may be a reasonable estimate when combining a small number of subgroups, such as four or less.

$$MOE = \sqrt{\sum MOE_{X_c}^2}$$

2. GVF: The estimated MOE is derived from a generalized variance function.

$$MOE = 1.645 \times \sqrt{aX_g^2 + bX_g}$$

where, a = -0.00023 and b = 24.8988.

3. GVF unweighted adjustment

The adjusted GVF variance estimate is defined as

$$var_{gvf-adj}(X_g) = var_{gvf}(X_g)/f,$$

Where, *f* is the adjustment factor. Two options were considered to compute the factor *f* -- unweighted and weighted. The unweighted method computes *f* as the unweighted mean of the ratio of the GVF-based variance to the actual variance across all the cells (*c*) to be combined.

$$f_{uw} = \frac{1}{C}\sum_{c=1}^{C} f_c,$$

Where, *C* = number of subgroups, and $f_c = var_{gvf}(X_c)/var_{actual}(X_c)$.

4. GVF weighted adjustment

The weighted method computes the mean of the ratio of the GVF-based variance to the actual variance, weighted by each cell's point estimate, X_c.

$$f_w = \frac{\sum_{c=1}^{C} f_c * X_c}{\sum_{c=1}^{C} X_c}.$$

For both adjustment methods, small TAZs are excluded from calculating the factors since their GVF-based variances may be unstable.

The naïve estimate has typically been used in CTPP, however, it tends to be unstable as more subgroups are combined. The weighted adjusted GVF has performed the best under evaluations that have been conducted. We emphasize that the approach should be used only when the combined cell estimate X_g is between 0 and 100,000, which is within the range of the set of modeled X_c's.

MOEs for proportions

The following describes how to estimate the MOE of a proportion p after combining categories. The computation is provided through the CTPP MOE ToolKit. Let the proportion $p_g = X_g/X$, where X_g is the estimated total for the combined cells, and X is the estimated total in a certain subpopulation. Then, the relative variance of p_g can be expressed as follows:

$$V_{p_g}^2 = \frac{MOE_{p_g}^2}{1.645^2 {p_g}^2}$$

It follows that the MOE is equal to:

$$MOE_{p_g} = \sqrt{1.645^2 p_g^2 V_{p_g}^2}$$

Then, from Wolter (1985), we estimate $V_{p_g}^2 \approx V_{X_g}^2 - V_X^2$, and substitute for $V_{p_g}^2$ as shown here:

$$MOE_{p_g} = \sqrt{1.645^2 p_g^2 (V_{X_g}^2 - V_X^2)}$$

$$MOE_{p_g} = \sqrt{1.645^2 p_g^2 (\frac{MOE_{X_g}^2}{1.645^2 X_g^2} - \frac{MOE_X^2}{1.645^2 X^2})}$$

$$MOE_{p_g} = \sqrt{p_g^2 (\frac{MOE_{X_g}^2}{X_g^2} - \frac{MOE_X^2}{X^2})}$$

With some algebra, this becomes:

$$Var_{p_g} = p_g^2 (\frac{SE_X^2}{X_g^2} - \frac{SE_X^2}{X^2})$$

When combining subgroups and the term $V_{X_g}^2$ is based on the naïve, GVF unweighted adjustment, or GVF weighted adjustment, the term V_X^2 is estimated from the direct estimate from the CTPP

table. If the term $V_{X_g}^2$ is based on the unadjusted GVF, then the term V_X^2 is also estimated from the GVF. If no result is given for the naïve, GVF unweighted adjustment, or GVF weighted adjustment approaches, it is likely due to relative variance of the base estimate causing negative variance, in which case the variance of the proportion is not estimable.

How can the MOE for an estimate from one subgroup be used to compare to another subgroup?

To test for the difference between two subgroups proportions, the MOE of the difference needs to be computed. The CTPP MOE ToolKit provides the computations. There are two scenarios that impact the MOE computations:

1) The two subgroups that are being compared are independent from each other.
2) The two subgroups that are being compared are dependent on each other.

Table A-3 provides an example of each scenario from a two-way table.

Table A-3 Example for comparing two subgroups

		Means of Transportation		Total
		Drive alone	Other	
TAZ	1	10	20	30
	2	30	40	70
Total		40	60	100

When the two subgroups are independent

An example of an independent test is comparing two column proportions, that is, the proportion among those who drive alone in TAZ 1 compared to the proportion among those who use other Means of Transportation in TAZ 1. The computation of the difference is (10/40 – 20/60).
Let the difference between two subgroup proportions (p_1 and p_2) be represented by $d = p_1 - p_2$.
The MOE of the difference *d* is expressed as:

$$MOE_d = 1.645\sqrt{Var_d} \quad \text{(A1)}$$

where Var_d is the variance of d, which is computed as:

$$Var_d = Var_{p_1} + Var_{p_2} \quad \text{(A2)}$$

When the two subgroups are dependent

An example of comparing two dependent subgroups is comparing two row percentages in the same row, such that they have the same denominator. That is, the proportion in TAZ 1 that drive alone compared to the proportion in TAZ 1 that use other Means of Transportation. The computation of the difference is (10/30 – 20/30).

From Scheaffer, Mendenhall III, and Ott (1995),

$$Var_d = Var_{p_1} + Var_{p_2} - 2\,cov(p_1, p_2)$$
$$Var_d = Var_{p_1} + Var_{p_2} + 2\frac{p_1 p_2}{n}$$

For the first two terms under the square root symbol, the Var_{p_1} and Var_{p_2} can be derived from the estimated MOE as follows (shown for p_1 only):

$$MOE_{p_1} = 1.645\sqrt{Var_{p_1}}$$
$$\left(\frac{MOE_{p_1}}{1.645}\right)^2 = Var_{p_1}$$

For the third term, the sample size *n* needs to be estimated, if not known. To do so, we make use of the following formula for Var_{p_1} (and likewise for Var_{p_2}) :

$Var_{p_1} = \frac{p_1(1-p_1)}{n_1}$, which implies that, $n_1 = \frac{p_1(1-p_1)}{Var_{p_1}}$. Then, for the CTPP MOE ToolKit, we set *n* equal to the minimum of n_1 and n_2. Then Var_d is computed from equation (A2), and the MOE_d is computed from equation (A1).

References

Scheaffer, R., Mendenhall III, W., and Ott, R.L. (1995). *Elementary Survey Sampling*. Duxbury Press. Fifth Edition.

Wolter, K. (1985). *Introduction to Variance Estimation*. Springer-Verlag, New York, Inc.

Appendix B

R function *rep.tab* User Guidance

Replicate tables get at the essence of the MOE – that a specific ACS estimate is a function of one small sample, and that alternative small samples could yield very different results. Taking the ACS multiple times is not feasible, but these simulations are what the results might look like if we could, and in the process, it shows how uncertain small area ACS estimates can be. Users can determine if they would be comfortable with any of the alternatives, or whether they would draw different conclusions depending on which one they had to use.

This appendix provides relevant guidance on the use of the R function *rep.tab* developed by the Westat team under Task 2 of NCHRP135. The code is as follows.

```
rep.tab <- function(Xs, MOE.X, MOE.Xs, rep=5, gvf.method=FALSE, graph=NULL) {
   X <- sum(Xs)
   p.k <- Xs/X
   var.Xs <- (MOE.Xs/1.645)^2
   var.X <- (MOE.X/1.645)^2
   v.k <- p.k^2 * (var.Xs/Xs^2 - var.X/X^2)

   if (gvf.method)
      alpha.k <- (X/b-1)*p.k
   else
      alpha.k <- (sum(p.k^2 * (1 - p.k)^2)/sum(p.k * (1 - p.k) * v.k) - 1)*p.k

   rep.tables <- gtools::rdirichlet(rep, alpha.k)*(rnorm(rep, X, MOE.X/1.645))
   tab.sd <- apply(rep.tables, 2, function(x) sqrt(var(x)))
   rep.tables10000 <- gtools::rdirichlet(10000, alpha.k)*(rnorm(10000, X, MOE.X/1.645))
   tab.sd10000 <- apply(rep.tables10000, 2, function(x) sqrt(var(x)))
   ctpp.sd <- MOE.Xs/1.645
   tabs <- t(rep.tables)

   if (!is.null(graph)) {
      if (graph == 1)
         barplot(height = t(tabs/rowSums(tabs)), xlab = paste("Cell"))
      else if (graph == 2) {
         par(mfrow=c(2, ceiling(rep/2)))
         for (i in 1:rep) {
```

```
            barplot(t(tabs[, i]), xlab = paste("Rep", i))
        }
    } else if (graph == 3) {
        par(mfrow=c(2, ceiling(rep/2)))
        for (i in 1:rep) {
            pie(t(tabs[, i]/sum(tabs[, i])), xlab = paste("Rep", i))
        }
    }
    par(mfrow=c(1, 1))
  }
  return(data.frame(tabs, tab.sd, tab.sd10000, ctpp.sd))
}
```

What is the purpose of the function?

The R function *rep.tab* is written for generating a set of replicated tables for a given CTPP table or an aggregated table. The methodology of generating replicated tables is designed for reflecting the sampling error and perturbation error in the transportation models and evaluating the sensitivity of models to data quality.

What R package is required to use the function?

The R package ***gtools*** needs to be installed and loaded before running *rep.tab* because within *rep.tab* it calls ***rdirichlet***. For details about ***rdirichlet***, see http://svitsrv25.epfl.ch/R-doc/library/gtools/html/dirichlet.html.

What are the parameters for the function?

- Xs – A vector of weighted frequencies in the internal cells of a table.
- MOE.X – MOE of the overall total of a table.
- MOE.Xs – A vector of MOEs for Xs.

- gvf.method – If TRUE, the parameters that are calculated within *rep.tab* are based on the GVF method. If FALSE, the distance method is used.
- rep – Number of replicated tables. Default value is 5.
- graph – Set to 1 for a scaled bar chart; 2 for a set of regular bar charts; and 3 for a set of pie charts.

How to call the function?

Suppose one wants to generate 5 replicated tables based on a CTPP table with 6 inner cells, as shown in Table B-1. CTPP estimates (weighted frequencies and MOEs) are shown in the table below. The overall total of the table is 11,365 and its associated MOE is 800.

Table B-1. Example of a CTPP table with six inner cells

Table cell	Estimate/Weighted frequency	MOE
1	2755	150
2	1585	100
3	1115	85
4	4180	525
5	1730	110
6	635	70
Total	12000	800

Step 1: Set up the seed. This ensures that consistent results will be generated in case of reruns.

set.seed(7169) ←7169 is an example of a random number used to set up the seed.

Step 2: Define Xs as an R vector containing the 6 cell estimates.

Xs <- c(2755,1585, 1115, 4180,1730, 635)

Step 3: Define MOE.Xs as an R vector containing the MOEs for the 6 cells.

MOE.Xs <- c(150, 100, 85, 525, 110, 70)

Step 4: Call *rep.tab*

If alphas are to be computed based on the GVF method

rep.tab(Xs=Xs, MOE.X = 800, MOE.Xs = MOE.Xs, rep = 5, graph = 1, gvf.method = TRUE)

If alphas are to be computed based on the distance method

rep.tab(Xs=Xs, MOE.X = 800, MOE.Xs = MOE.Xs, rep = 5, graph = 1, gvf.method = FALSE)

What is the output from calling the function?

The output shows a matrix of numbers in Table B-2. Columns *X1* through *X5* are the cell estimates in the five replicated tables. The column *tab.sd* shows the standard deviation among the five replicates for each table cell. The column *tab.sd10000* shows the standard deviation among the replicated tables if 10,000 of them are generated for each table cell. The column *ctpp.sd* is the standard errors calculated from the provided margin of errors.

Table B-2. Output matrix from calling R function

	X1	X2	X3	X4	X5	tab.sd	tab.sd 10000	ctpp.sd
1	2896.5061	2836.6244	2441.8605	2923.3754	2666.9241	200.51912	190.35917	91.18541
2	1499.5951	1485.5212	1623.4476	1566.5837	1351.6484	102.16050	139.10399	60.79027
3	1154.1924	1043.7335	912.2104	1102.4740	1035.8030	90.60335	113.44763	51.67173
4	3851.2737	3822.5210	4029.0215	3984.3165	4041.2885	102.05425	241.83410	319.14894
5	1683.9835	1901.1452	1736.5143	1798.1235	1730.3982	83.78308	145.66241	66.86930
6	591.5961	835.2359	736.5924	688.6089	616.7395	97.78319	84.98242	42.55319

What are the graphs in the output?

The function call can generate three types of graphs by setting different values to the parameter *graph*. If the parameter *graph* is left blank, no graphs will be produced. If *graph* is set to 1, a scaled bar chart as in Figure B-1 is produced.

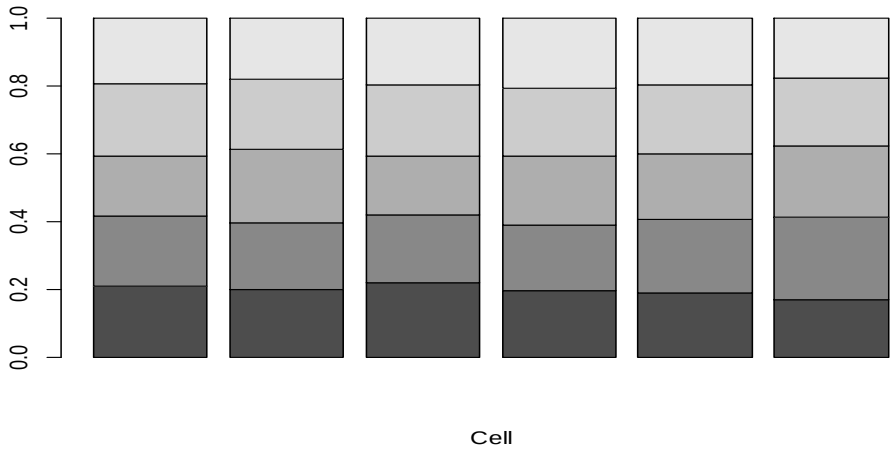


Figure B-1 Scaled bar chart for five replicated tables

In the scaled bar chart, each bar shows the relative magnitude of the five replicated estimates for each table cell. If the variation among the five replicates is very small, each replicate accounts for approximately 20% of the bar. The rightmost bar in Figure 1 shows that the estimate from replicate 1 (591.5961) is relatively small compared to the estimate from replicate 2 (835.2359).

If *graph* is set to 2, a set of regular bar charts are produced as in Figure B-2. The bars in each chart represent the estimates from each of the five replicated tables.

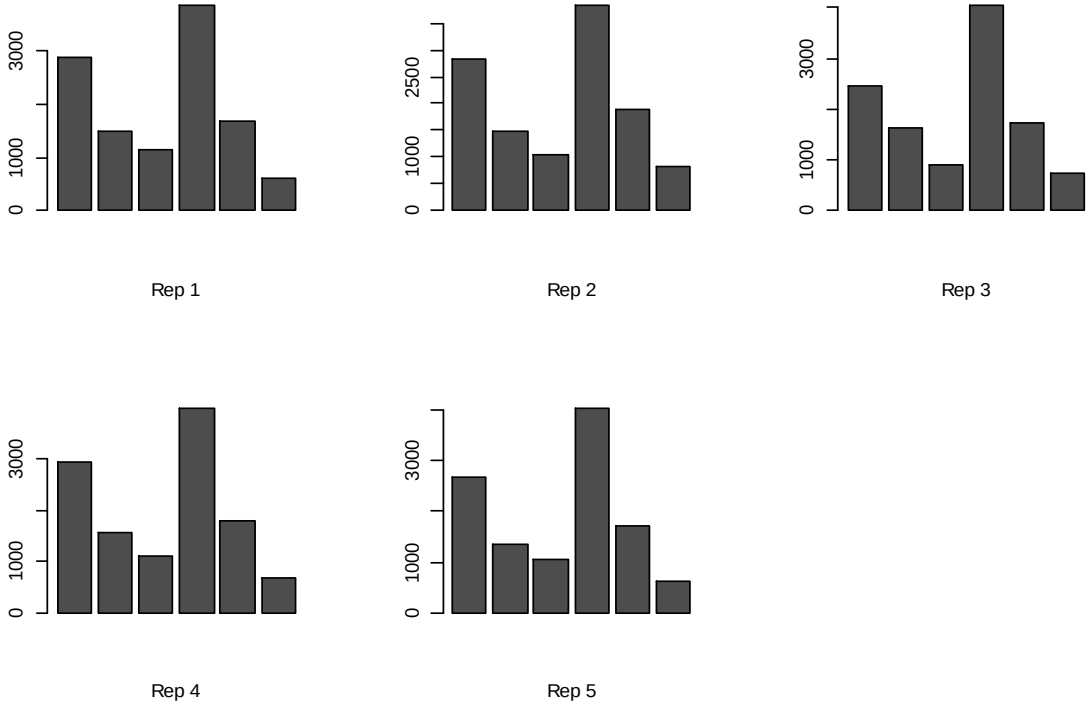


Figure B-2 **A set of bar charts for five replicated tables**
Note: **Each of the five bar charts represents a replicated table and the areas of the bars represent the magnitude of the replicated estimates.**

If *graph* is set to 3, a set of pie charts are produced as in Figure B-3. Rep1 through Rep5 show five replicated tables generated for the CTPP table in Table B-1, where the areas of the pie slices numbered 1 through 6 represent the magnitude of the replicated estimates in six inner cells.

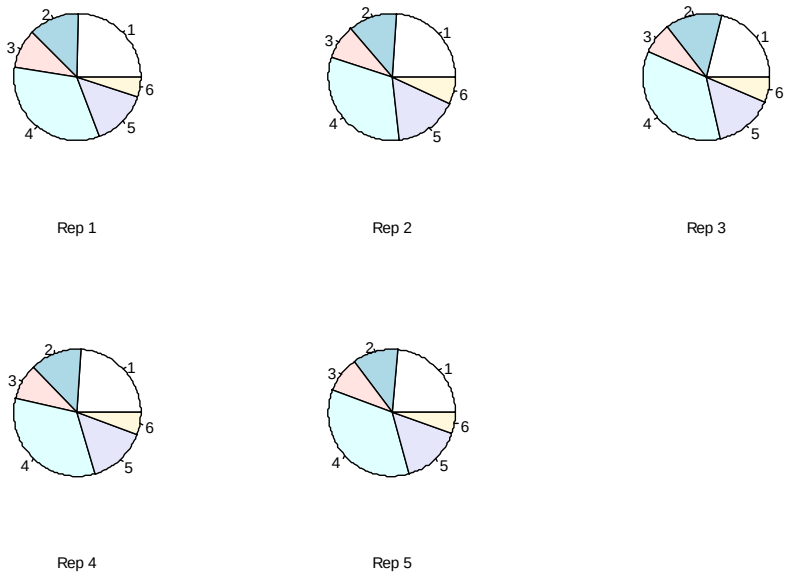


Figure B-3. A set of pie charts for five replicated tables
Note: Each of the five pie charts represents a replicated table and the areas of the pie slices represent the magnitude of the replicated estimates.

The graphs allow the users to visualize the sampling and perturbation errors in the CTPP estimates.