Statistical Journal of the IAOS 33 (2017) 469–485
DOI 10.3233/SJI-160259
IOS Press

Producing multiple tables for small areas with confidentiality protection

Tom Krenzke[a,*], Jianzhu Li[a] and Laura McKenna[b]
[a]*Westat, 1600 Research Boulevard, Rockville, MD, USA*
[b]*U.S. Bureau of the Census, 4600 Silver Hill Road, Washington, DC, USA*

Abstract. One approach for protecting data confidentiality with tabular data is to apply statistical disclosure control (SDC) methods that randomly perturb selected variables in the underlying microdata set. A pre-tabular SDC approach has the attraction of retaining both consistency and additivity across tables. However, the effect of the perturbation needs to be reflected in the variance estimates for the estimates presented in the table cells. This paper describes the pre-tabular SDC method used for the 2006–2010 Census Transportation Planning Products and describes a method for estimating the variances of the cell estimates that account for the perturbation variance.

Keywords: Disclosure limitation, model-assisted constrained hot deck, random perturbation, variance estimation

1. Introduction

Survey results are often published in tabular form, with the table cells defined by cross-classifications of categorical variables. Tables can be susceptible to disclosure attacks. When the estimate in a table cell is based on data from a single respondent (often known as a "sample unique"), it may be possible to identify a series of characteristics for that respondent from a combination of related tables. If an intruder knows that a person with the characteristics defining a sample unique is a survey respondent, the intruder can learn a set of the respondent's other characteristics. Even if the intruder does not know about survey participation, a subset of the characteristics derived for a sample unique may identify a "population unique", that is, the only individual in the whole population with that subset of characteristics. The intruder may then be able to find about other characteristics of that individual.

Further concerns when generating multiple tables for small areas can arise from table differencing and possible resultant table-linking attacks [1]. Sometimes a set of tables is produced that contains tables that are based on two marginally different universes of interest (e.g., workers aged 16 years and older in one table and workers aged 17 years and older in another) or two different overlapping classifications (e.g., one age classification of 16–24 and 25–34, and another of 16–23 and 24–34). Even if two tables individually pass disclosure rules –such as the requirement that each cell contains at least three respondents– the difference between the tables can reveal a "sliver", that is, an implicit table with only one sampled record in one or more table cells. Table differencing is illustrated in Table 1 for two tables for Means of Transportation (MOT) in a small area that pass the disclosure rules. By taking the difference between the two tables, one for Age Group 16 to 25 and the other for Age Group 16 to 24, the resulting "implicit" table shows counts for a 25 year old that would have failed the disclosure rules due to small cell size. The single case can be identified as taking public transportation and being 25 years old.

The discovery above through table differencing is not too disclosive, however, it can then be followed by table linking; for example when MOT is a common variable throughout a set of tables. Table 2, extending upon the illustration in Table 1, shows a series of

*Corresponding author: Tom Krenzke, Westat, 1600 Research Boulevard, Rockville, MD 20850, USA. Tel.: +1 301 251 4203; E-mail: KrenzkT1@westat.com.

Table 1
Illustration of a table-differencing attack via counts on two published tables for a small area

Means of Transportation	Age Group 16 to 25	Age Group 16 to 24	Difference (Age 25)
Drove alone	30	20	10
Carpooled	10	10	0
Public transportation	20	19	1

Table 2
Illustration of a table-linking attack on a series of tables

Means of transportation	Sex		Vehicles available			Tenure				Total
	Male	Female	0	1	2 or more	Owned with mortgage	Owned without mortgage	Rented	Occupied without rent	
Drove alone	3	7	2	5	3	5	2	2	1	10
Carpooled	0	0	0	0	0	0	0	0	0	0
Public transportation	1	0	1	0	0	0	1	0	0	1
Total	4	7	3	5	3	5	3	2	1	11

implicit tables from differencing the Age Group 16 to 25 and 16 to 24 tables: MOT crossed with Sex, MOT crossed with Vehicles Available, and so on to MOT crossed with Tenure, where Sex, Vehicles Available, Tenure and so on are variables in the set of tables produced. A single case appears in the marginal total for MOT = public transportation in each table, from which a microdata record with MOT = public transportation, Sex = male, Vehicles Available = 0, ...Tenure = owned without mortgage can be constructed. Linking the tables to compile characteristics for the sliver is known as a table-linking attack. Even without table differencing, a variable (or a combination of variables) that is a common thread in the set of cross-tabulations with a marginal count equal to one for at least one of its categories can be susceptible to a table-linking attack.

Lastly, if a public use file of the source data for the tables without small area localities identified on the file is assessable, then the sliver case's characteristics (MOT = public transportation, Sex = male, Vehicles Available = 0...Tenure = owned without mortgage, Age = 25) from the set of tables can be matched using a record linkage approach (see [2], for a review) to the public use file with some probability of a correct match. Adding the small area locality to the public use file could enable an intruder to get much closer to identifying an individual, markedly increasing the risk of disclosure. The public use file may include hundreds of additional variables that could then be linked and added to the small area locality of the sliver case. When defining tables, it is important to form categories of variables that do not overlap to prevent slivers. Even so, when producing tables for small areas, low marginal counts for common variables throughout the set of tables (such as MOT in the example above), may cause violations of disclosure rules and an undesirable amount of information withheld. This paper introduces an approach used to produce small area tables without suppression of results due to low counts.

It should be noted that there are additional important risk elements to be addressed when a real-time query tool that generates tables on request from a microdata file is used. See [3] for more discussion and scenarios about data intruder attacks when producing tables in a real-time analytic query system.

This paper discusses a number of statistical disclosure control approaches that can be applied to address disclosure risks in a predesignated set of tables. Section 2 discusses the methods that have been developed to date. Section 3 presents a new approach called the model-assisted constrained hotdeck (MACH) which strives to reduce the variance introduced by perturbation while retaining relationships between variables. Section 4 provides an evaluation of the MACH approach. Section 5 shows how the MACH approach was used to reduce the disclosure risk in a U.S. Census Bureau set of tables known as the 2006–2010 Census Transportation Planning Products (CTPP) (based on data from the American Community Survey (ACS) from years 2006–2010) without suppressing a significant amount of tabular estimates. Section 6 presents a variance estimation approach to capture the impact of the SDC perturbation procedures. Lastly, Section 7 presents some concluding remarks.

2. Methods to address disclosure risk in tables

The SDC methods for tabular data are of two broad types. Post-tabular methods use cell suppression or modify table values after the tables are generated from the microdata [4], while pre-tabular methods apply

SDC treatments to the original microdata to produce a perturbed data set from which all the tables are produced. With cell suppression SDC, a primary table cell value is suppressed if a threshold rule is violated, and, subsequently, a complementary cell is suppressed to protect against the derivation of the primary cell value from other cell values [5]. The post-tabular approaches other than cell suppression include adding noise to counts in tables [6], controlled rounding of table counts that ensure internal cell counts sum to the original published marginal counts, controlled tabular adjustment [7], and the approach adopted for the Longitudinal Employment Household Dynamics (LEHD) OnTheMap system [8].

There are two broad classes of pre-tabular SDC methods: synthetic and perturbation approaches. Synthetic approaches involve producing fully synthetic datasets [9] or partially synthetic datasets that are mixtures of actual and multiply-imputed values (e.g. [10]). Synthetic approaches typically replace original values with draws from appropriate probability distributions in a way that aims to retain the essential statistical features of the original data, including multivariate associations. Most synthetic approaches rely on the multivariate relationship between the target variables (variables to be masked) and other variables in determining the synthetic values. When missing data patterns are non-monotone (e.g., some may refer to as "Swiss cheese") as typical in surveys, to maintain relationships between variables, a sequential approach [11] is used where synthetic values are drawn for Variable 1 from the posterior predictive distribution of observed data on predictor variables. Then synthetic values are drawn for Variable 2 from the posterior predictive distribution that includes the synthetic values from Variable 1 and observed values from other predictor variables. The process continues until all variables targeted for masking are synthesized. The process is circular and ends based on convergence rules or a predetermined number of cycles are conducted.

Perturbation approaches involve applying a controlled random treatment procedure to replace a subset of the original data values by other values, with the aim of introducing just enough noise or uncertainty into the microdata to reduce the disclosure risk to an acceptable level. Perturbation methods are sometimes referred to as blank-and-impute and can control change from the original values. Several perturbation approaches use the sequential approach outlined above to maintain multivariate associations.

For multiple tables for small areas, pre-tabular approaches were pursued in lieu of post-tabular approaches due to several reasons that relate to maintaining multivariate associations, operational practicability, applicability to a variety of types of variables (e.g., nominal, ordinal) and estimates (e.g., counts, means), ability to facilitate variance estimation, and ability to provide consistent results among the set of tables. Masking the microdata and then using the masked microdata in an already established production framework will produce multiple tables that have consistency in the margins and, therefore, are additive as tables are aggregated. Although not an exhaustive list, pre-tabular methods include data swapping [12], rank swapping [13], data shuffling [14], micro aggregation [15], additive and multiplicative noise masking (e.g., Post Randomization Method (PRAM) [16, 17]), and synthetic parametric and non-parametric approaches. A brief description of each method is provided below. A more detailed review of SDC methods is provided by [18]. At the end of the descriptions is a discussion of the motivations to pursue a new approach to perturbation that led to its implementation in a large-scale application.

Data swapping is applied to microdata prior to generating tables for the American FactFinder [19], which displays predefined tabular results primarily from the ACS and Decennial Census. A swapping method that has been used at the Census Bureau, described by [20], consists of swapping pairs of household records selected as having the greatest disclosure risk of being re-identified. The author of [20] discusses that the swapped records match on a set of demographic characteristics but were in different census blocks for the hundred percent (short form) data in the 2000 Census and in different block groups for the sample (long form) data. The tables are produced from the swapped data. In general, data swapping is also applied by swapping questionnaire items, in lieu of geographic areas. Key matching variables are selected for the purpose of pairing the records, and swapping variables are selected to have their values swapped between the pair. Rank proximity swapping constrains the swapping to pairs of records within a small range of data values for the target variable. Because of the constraining, it has the potential to maintain multivariate relationships between the targeted variable for swapping and unswapped variables.

Contrary to data swapping, in a sort on records, the data shuffling approach transfers values in a specific manner from record 1 to record 2, from record 2 to record 3, and so on. That is, the values of the target variables are "shuffled" among the data records.

As described in [14], perturbations occur among the target variables using the conditional distribution of the target variables, given the non-targeted variables (which could include the sample weights). Then the rank-ordered perturbed values are mapped to the rank-ordered original values.

Micro aggregation is a data coarsening non-random approach in which records are sorted by the values of the target variable, and then the perturbed values are computed as the average of at least three records. An approach has been created for the multivariate setting [21].

Adding noise (or multiplicative noise) is another approach for numeric data. For example, noise can be added to item y for record i as follows: $\tilde{y}_i = y_i(1+fz)$, where f is a constant between 0 and 1, and z is a draw from the standard normal distribution. The standard deviation of the added noise is a function of f and y_i, where the level of noise is allowed to vary relative to the magnitude of y_i.

A special case of noise addition to nominal variables is the PRAM approach. The PRAM consists of setting up the probability distribution of the variable, conditional on the observed value for the record. PRAM can be extended to multivariate distributions.

While data swapping is appealing due to its ease of implementation and retaining unweighted univariate distributions, a concern is a disconnection between swapped and un-swapped variables, which may cause an intolerable attenuation of multivariate associations. This concern has been explored in [22], for example. In general, a limited number of key matching variables are considered, and a limited number of variables are swapped. In addition, a large amount of swapping households across localities could potentially impact the data utility for small areas.

Rank proximity swapping, data shuffling, and micro aggregation are applicable only to ordinal variables with several levels. While the data shuffling paper referenced above reports improvement upon rank proximity swapping in the retention of multivariate associations, the patented approach was not explored for the CTPP, although potentially applicable. For micro aggregation, it was not readily clear how randomization, the sample weights and other predictor variables could be incorporated, and therefore the approach was not explored further. The PRAM approach was not considered due to its focus on nominal variables, and it does not take into account predictor variables in the noise addition process. A synthetic parametric approach to masking was attempted in the CTPP application as discussed in [23]. The approach involved building models for each variable to be masked and then making draws from the predictive distribution. The performance of the approach was inferior to other approaches mainly due to the unconstrained nature of the synthetic data generator for the variable included in the set of tables for the small areas. With more time, the models and the approach could have been improved. Other non-parametric approaches that use classification and regression trees are available [24] that have potential to be more time-efficient, however the unconstrained nature of the approach would still impact the resulting tabular estimates for the small areas.

In the next section, a new approach (MACH) is described that addresses several needs that none of the pre-tabular approaches fully address. The approach is applicable to nominal variables (including binary variables) and can constrain the amount of change to the values among the ordinal variables with three or more levels. In doing so, it can handle several categorical versions of the same variable in a consistent manner. For example in the CTPP, categorical household income had 5-levels, 9-levels, and 26-levels, and continuous income. The version with the most detailed categories (or continuous) was masked, while being able to ensure change for the more coarsened versions. The approach can take into consideration a large pool of predictor variables in the SDC model to maintain multivariate associations, while borrowing strength across geographic areas. Lastly, it is able to use the sample weights to limit distortion in estimates.

3. Framework of the MACH approach

The MACH approach is similar to swapping, however a key advantage to the MACH is the bounding and control on the amount of noise being added while retaining the association between the target variables and non-target variables by selecting the best predictors, from a large pool of variables, analytically for each target variable (variable to be perturbed).

The main goals of the perturbation process are: 1) to retain the univariate and multivariate distributions of the original data as closely as possible; and 2) to perturb the values to the least extent possible while giving acceptable levels of disclosure control. Simply put, the key to achieving the goals is the formation of the hotdeck cells. The MACH process forms hotdeck cells (c), and then values of the target variable are exchanged among all records within each cell. The general pertur-

bation model for the MACH is expressed as follows: $\tilde{y}_{i(c)} = y_{i(c)} + \varepsilon_{i(c)}$, where, subscript (c) denotes the c^{th} class (hotdeck cell) defined from the set of factors $\{I(s), y_{g'}, \boldsymbol{x}, \hat{\boldsymbol{y}}_{g''}, \boldsymbol{w}_{g'''}$, where $I(s)$ is the set of indicators for being selected for perturbation, $y_{g'}$ denotes g bins formed on the target variable y, $\boldsymbol{x}$ are the auxiliary variables, $\hat{\boldsymbol{y}}_{g''}$ are the g'' groups formed from model predictions, $\boldsymbol{w}_{g'''}$ are the g''' groups formed from the sampling weights and where $\varepsilon_{i(c)} = \tilde{y}_{i(c)} - y_{i(c)}$ results from the random error associated with case i for a random with-replacement draw within the c^{th} class. The bolding pattern represents vectors. The noise $\varepsilon_{i(c)}$ is bounded by the bins $(y_{g'})$ as part of the hotdeck cells.

The hot deck cells (c) are constructed based on a number of variables in order to provide several desirable features for the perturbed data set. The variables used in forming the hot deck cells were as follows:

- Target selection flag can be used to only allow target records to be replaced by values from other target records. Using the target selection flag in the hotdeck cell creation will result in the unweighted distribution of the target variable to be retained.
- Bins on the target variable, which are formed to constrain the difference between the original and perturbed values. This component, described in Section 4.2.2, applies only to ordinal variables.
- Locality is included to take account of the unique characteristics of geographic areas.
- Prediction groups are formed from predicted values from a stepwise linear regression analysis conducted for each target variable. The predicted values (in grouped form) were incorporated into the cell formation in order to retain covariances between the target variable and its main predictors. This procedure is described in Section 4.2.1.
- Auxiliary variables, if any, can be included to address the need to take account of variables such as skip patterns.
- Weight groups, which are formed by classifying records based on the sizes of the records' sampling weights. Weight groups are used to reduce the mean square error (MSE) of resulting estimates and to protect against distortion that could occur when a data value is replaced by the value from another record with a vastly different sampling weight.

The hotdeck cells are defined by cross-tabulating the above variables. Initially there can be small hotdeck cells, which are identified and combined using an automated algorithm.

3.1. The construction of prediction groups

The general approach for creating prediction groups for use in forming the hot deck cells was influenced by a sequential imputation procedure that was initially designed for handling non-monotone (Swiss cheese) missing data patterns in complex questionnaires [25]. The approach uses the predictive mean as a distance metric to form prediction groups to use in constructing hot deck cells, along similar lines to a method described by [26].

The construction of prediction groups begins with model selection and parameter estimation for each of the target variables that had been selected for perturbation. A stepwise linear regression model is used throughout. Since the modeling served only as input to the construction of hot deck cells, the simplicity and computing economy of the linear regression model is preferred over more complex modeling.

The models can be constructed separately for different geographic areas. The models are based on the fully complete original dataset since the joint distribution among the variables in that dataset is the target distribution to be retained as closely as possible by the perturbation process. The modeling identifies the best predictors for each target variable and the regression coefficients were estimated.

After the modeling is completed, model predictions are computed for the first target variable using the survey values for the predictor variables. The same procedure is used for the subsequent target variables but with one important difference. For the second variable, the original survey values for the first variable are replaced by perturbed values; for the third variable, the survey values of the first two variables are replaced by perturbed values; and so on, throughout all the variables to be perturbed. Also, after each variable is perturbed, any recodes, interaction terms, or indicator variables are recreated using perturbed values so the perturbed values could be used in the prediction equation for the subsequent target variables in the sequence. This replacement procedure is applied in order to retain the covariance structure as closely as possible.

The target variables can be classified into two groups: ordinal variables (including continuous variables) with at least three unique values and nominal variables (including binary variables).

3.2. Ordinal target variables

The MACH approach adopted aims both to constrain the amount of change in the target values by us-

ing bins created on the target variable itself, and to retain covariances with the most important predictors by the use of the model predictions. The bins are formed as categorized values of the target variable that satisfy the following conditions:

- Each bin has to contain more than one value of the pre-designated categories of the target variable to be published.
- If the distribution of the target variable has spikes (e.g., self-reported travel times have spikes at 5-minute intervals), then at least two spikes are included in a bin. If this condition has not been applied, the approach would likely result in many values being unchanged by the perturbation process.
- To the extent possible, the value of the target variable has to have the possibility to either increase or decrease.

To satisfy the third condition, two alternative sets of bins (A and B) with overlapping categories are formed, as illustrated in Fig. 1. The histogram in that figure depicts a frequency distribution for self-reported travel time in minutes (y), with spikes at multiples of five. The two rows below the histogram illustrate two sets of overlapping bins A and B, and the third row gives the published categories for the y variable. Prior to forming the hot deck cells, the sample of target records for that variable is split into two halves, with one-half using set bin set A and the other half using set B. A review of Fig. 1 shows that each of the three conditions is satisfied by the combination of bins A and B.

The size of the bins can be constructed to control the magnitude of the possible difference between the original and perturbed values. The wider the bin, the greater is the reduction in risk but also the greater is the reduction in data utility. The choice of bin size, therefore, involves a compromise between these considerations.

Bin width and prediction groups are related. The wider the bins, the greater the potential for the model predictions to influence the perturbations. When the model predictions are weak, the prediction groups have little effect on controlling the perturbations, but the bin formation keeps the differences between original and perturbed values within bounds. When the model is strong, the prediction groups result in data values exchanged between records with similar predictions, thereby retaining multivariate relationships with the predictor variables.

3.3. Nominal and binary target variables

Since bins are not applicable for nominal (e.g., industry classification) and binary (e.g., minority status) target variables, the MACH is applied unconstrained for these variables. Indicator variables are created for each category of the target variable, and a separate linear regression is computed for each indicator variable. Predictions are made for each of the indicators, resulting in each record to be perturbed having a profile of predictions across the set of indicators. The predictions do not sum to 1 and any one may be outside the range 0–1, but that is not a concern for the hot deck construction. Subsequently, a k-means clustering algorithm is performed on the vector of predicted values for each of the indicators. The algorithm produces a set of clusters, analogous to the prediction groups discussed above for ordinal variables.

3.4. Data replacement and weight calibration

When using the target selection flag in the hotdeck cell creation, all the target records in a hot deck cell are subjected to perturbation. A without-replacement draw from the empirical distribution is conducted to replace each target record's value by the value of another record in the cell. This is carried out for all records in the cell in a single step by exchanging original survey values in a random manner that ensures that a donor is not the same as the target record. Each record of the n records in a cell was indexed with a random sequential number from 1 to n. A random draw was conducted for the first donor, say 3, then sequentially proceeded with record 3 as a donor for record 1, 4 as a donor for 2, 5 as a donor for 3, and continued until n is a donor for n-2, then 1 is a donor for n-1, and 2 is a donor for n.

With this approach, the aim is to exchange data values within a locality. However, with the automated cell-collapsing routine, in practice data values could be exchanged across locality boundaries. When all target variables are perturbed, raking [27] is applied to ensure consistency with specified totals for larger geographic areas by calibrating the sampling weights.

4. Evaluation

A detailed review of select synthetic and perturbation approaches considered for the CTPP is given in [23], which established a constrained hotdeck as

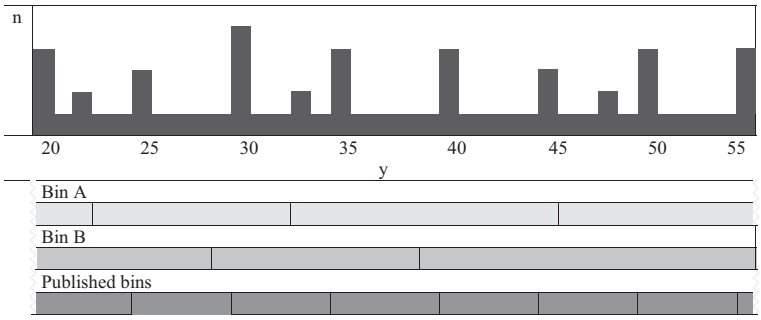


Fig. 1. Illustration of bin formation.

a credible and promising approach. The MACH approach was developed out of that review. In an additional evaluation, the impact of constraining by bins, and using weight groups, prediction groups and localities was investigated in terms of variance and multivariate associations among variables. The research team used the 5-year accumulation of 2005–2009 ACS data for workers 16 years old and older in North Carolina households. The evaluation was processed using the MACH and a blank-and-impute unconstrained version of the MACH (no bins). The target variable for evaluation is person earnings.

There were eight types of Core-Based Statistical Areas (CBSAs) used in the evaluation. Each CBSA was classified into types relating to the variation in the weights, quality of models, and variation among localities (e.g., tracts) within the CBSA. Computations were done at the national level to assign each of the nation's 953 CBSAs to low or high levels separately for each of the three factors: variation in the weights, quality of models, and variation among localities. The three factors were combined to assign a CBSA type for each CBSA as shown in Table 3.

The variation in the weights was computed as the coefficient of variation (CV), which ranged from 39 percent to 76 percent for low CV areas, and 76 percent to 136 percent for high CV areas. The model R^2 was computed from results of a stepwise regression, which ranged from 40 percent to 59 percent (40%) for the low group and 59 percent to 84 percent for the high group. Lastly, the variation between localities was computed by the variation among the tract-level mean person earnings.

The perturbation approaches were applied under various treatment scenarios. Four key factors were defined by use of bins on the target variable or not, number of prediction groups, number of weight groups, and size of locality. There were two sizes of the locality: blocks combined to have at least 300 ACS sample cases (L300) and at least 1000 ACS sample cases (L1000). These zones can have fewer than 300 (or 1000) sample cases due to the exclusion of group quarters for the evaluation. The average number of workers in the L300 entities is 464 and 1,365 for L1000.

The treatments were defined to arrive at hotdeck cells with similar sizes between the same treatment combinations. The experimental design is given in Table 4. There were five replications for each of the 16 treatments to make 80 processing runs. Partial replacement with a rate of 25 percent was assigned using simple random sampling.

There were two measures used to evaluate the performance on data utility. First, the interquartile range (IQR) of differences across all table cells between the raw and perturbed data were produced by each CBSA type and overall. We computed the difference in cell means for a given variable as follows: $D_{\bar{y}} = \tilde{y} - \bar{y}$, where $\tilde{y}$ = estimated mean from the perturbed data, and $\bar{y}$ = estimated mean from the original data. The cell mean differences were produced by CBSA type for earnings for two-way cross-tabulations involving the following variables: poverty status (3 levels), minority (2 levels), industry (7 levels), sex (2 levels), occupation (7 levels), years in the United States (U.S.) (5 levels), age of youngest child (3 levels), mode of data collection (3 levels), years of schooling (7), and Census tract. The two-way tables are a mix of those involving tracts and those not involving tracts. The IQR results provide an indication of variation due to perturbation.

The use of propensity scores as a global utility measure of the retention of multivariate associations is described in [28]. The perturbed and original data files were stacked and $T = 1$ was assigned to the perturbed records and $T = 0$ was assigned to the original records. A weighted logistic regression model was processed on T using main effects, and also with interaction terms associated with perturbed variables. Table 5 provides the terms in the model.

Table 3
Number of CBSAs by CBSA Type for Earnings

CBSA type	Variation in weights	Model R^2	Variation between localities	Number of CBSAs in North Carolina
1	Low	Low	Low	7
2	Low	Low	High	3
3	Low	High	Low	3
4	Low	High	High	12
5	High	Low	Low	5
6	High	Low	High	2
7	High	High	Low	3
8	High	High	High	6

Table 4
Experimental design

Perturbation approach	Treatment number	Number of bins	Number of prediction groups	Number of weight groups	Locality size (n)	Resulting expected cell size[1]
MACH	1	9	2	2	464	12.9
	2	9	2	2	1365	37.9
	3	9	2	4	464	6.4
	4	9	2	4	1365	19.0
	5	9	4	2	464	6.4
	6	9	4	2	1365	19.0
	7	9	4	4	464	3.2
	8	9	4	4	1365	9.5
Unconstrained	1	1	6	6	464	12.9
	2	1	6	6	1365	37.9
	3	1	6	12	464	6.4
	4	1	6	12	1365	19.0
	5	1	12	6	464	6.4
	6	1	12	6	1365	19.0
	7	1	12	12	464	3.2
	8	1	12	12	1365	9.5

[1]Computed as the locality size divided by the product of the number of weight groups, number of prediction groups, and the number of bins.

Table 5
Effects Used in the U Statistic Logit Model

Target variable	Main effects	Interactions
Earnings	Age, earnings, poverty	Earnings with [L300 (many), Means of transportation (9 levels), Poverty status (3 levels), Minority (2 levels), Industry (7 levels), Occupation (7 levels), Years in the U.S. (5 levels), Mode of data collection (3 levels), Age] and L300 with Age

The following statistic U should be close to zero if the perturbed data and original data were indistinguishable.

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

where

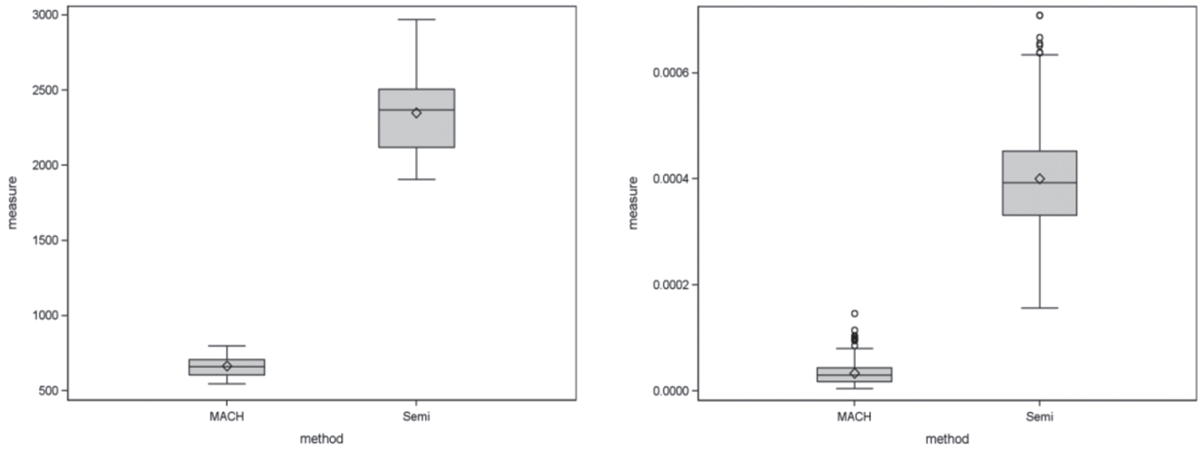
N = number in the stacked file

$\hat{p}_i$ = propensity score (logistic regression prediction) for record i

c = proportion of units from the perturbed data file (e.g., 1/2)

Figure 2 shows differences between the MACH and the unconstrained approach. Treatments 1 through 8 were combined for each approach. The IQR and U results show much less variation for MACH approach than for the unconstrained approach. Due to these results, we focus the remainder of the analysis on the MACH approach.

For each CBSA type, we processed a one-way ANOVA with treatment as a main effect. Statistical significance was determined by the adjusted Tukey pairwise comparison tests and $\alpha = 0.05$. Figure 3(a) provides box plots for each treatment within each of the CBSA types for the IQR statistic, and Fig. 3(b) provides box plots for the U statistic. An explanation is provided for any significant pairwise comparison. For CBSA type 2, we would expect that the attributes for this CBSA type would result in potential help from smaller localities, which is exactly what happened. Treatments 1, 5 and 7 have lower IQRs than Treatment



IQR measure U measure

Fig. 2. Comparison of IQR and U statistics between the MACH and the unconstrained approach.

2 in which the main difference was the smaller geography used in the hot deck cell creation. For CBSA type 3, we would expect that the attributes for this CBSA type would result in potential help from the model. Treatment 7 has lower IQRs than Treatment 8. The main difference was lower level of geography used in the hot deck cell creation. Interesting that the model did not seem to help; however, smaller geography did.

For CBSA type 4, we would expect that the attributes for this CBSA type would result in potential help from the model and smaller geography. Treatments 1, 3, 5, and 7 have lower IQRs than Treatments 2 and 4. The main difference was due to smaller geography used in the hot deck cell creation. Also, Treatments 2 and 4 have only two prediction groups, while Treatments 6 and 8 have four and were not significant different from the low-level locality treatments. Lastly, Treatment 8 has lower IQRs than Treatment 2. The main difference is that Treatment 8 had more prediction and weight groups. Smaller geography helped quite a bit, with some benefit from the model predictions. For CBSA type 7, we would expect that this CBSA type would result in potential benefit from weight groups or prediction groups. While treatments have some evidence of significant U measure results, there was not enough evidence to find a specific significant result. For CBSA type 8, we would expect that this CBSA type would result in benefits from weight or prediction groups, or small geography. While the overall statistical test was significant for the U measure, there was not enough evidence to find a significant difference between specific treatments.

The evaluation showed that there was little benefit from controlling the weights, which suggests a strategy of using a small number of weight groups (perhaps just two) and using them last in the hot deck cell formation, if at all. The predictions of continuous variables did not contribute a great deal when covariates were weak and when constraining bins were used. The benefits of the predictions come more into play when strong covariates exist, for nominal or ordinal target variables with few categories, and when the constraining bins needs to be wide to reduce disclosure risk. Lastly, if the between-locality variance for the target variables is high, it is beneficial to restrict locality size for the application of the perturbation procedure to be as small as possible while generating enough perturbation variance to satisfy disclosure risk thresholds. If between-locality variability is high, then locality should be one of the first components in the ordering of the hot deck formation in order that locality has less chance of being collapsed due to small cells.

5. Application

The CTPP are a set of tables, prespecified by transportation planners, for each of over 140,000 combinations of census blocks termed traffic analysis zones (TAZs) and for other geographies and for journey-to-work flows by mode of transportation. The TAZs are the building blocks that transportation planners use to define the areas of interest to them. Overall the CTPP contains millions of tables involving approximately 30 variables, particularly the responses to the ACS transportation questions. The table cells contain frequencies (e.g., counts of workers for modes of travel), means

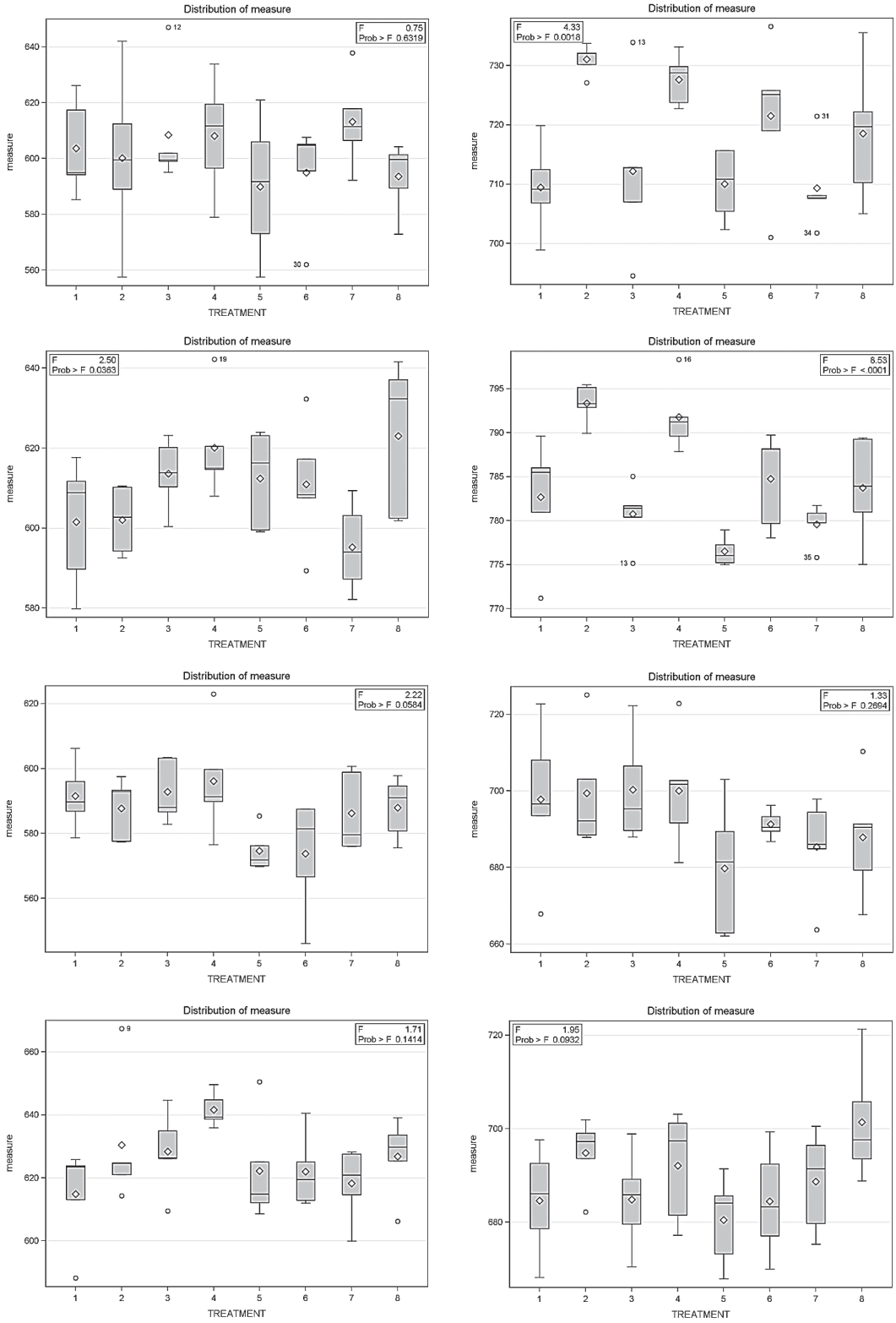


Fig. 3a. IQR for Earnings for Treatments within CBSA Type (types 1 and 2 in the top row, types 3 and 4 in the second row, etc.).

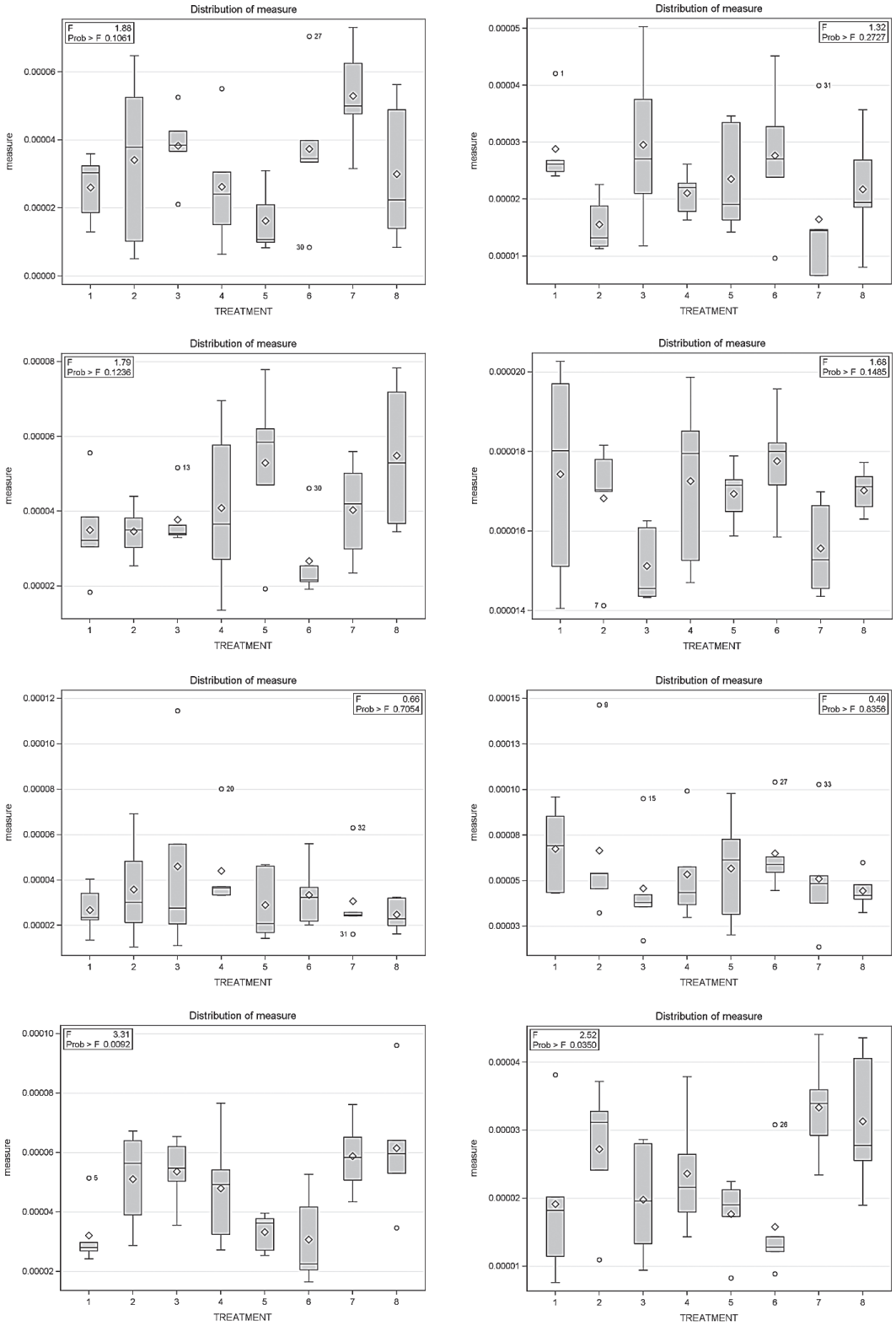


Fig. 3b. U for Earnings for Treatments within CBSA Type (types 1 and 2 in the top row, types 3 and 4 in the second row, etc.).

and medians (e.g., commuting times), together with their margins of error. A feature of the CTPP tables is that there are different analytic versions for some of the underlying survey variables that are used in different tables; for example, household income is sometimes analyzed as a continuous variable and other times as a categorized variable, with several alternative versions of the categorization. All recodes were applied uniformly to all TAZs.

The data source for the 2006–2010 CTPP moved from the Census Long Form that was used with the 2000 and previous censuses to the 5-year ACS sample. Since the ACS 5-year sample size is about one-half of the 2000 Census Long Form sample size, the 2006–2010 CTPP is potentially subject to much greater disclosure risk than earlier versions of the CTPP. The ACS sample in some TAZs had only 20 to 25 workers. There were about 7 million TAZ-to-TAZ flows in the journey-to-work tables, and only about 10 million workers in the ACS 5-year sample. As a result, almost all TAZ-to-TAZ flows had very sparse tables.

For the CTPP based on the 2000 Census, the Census Bureau applied a Rule of Three threshold rule that required at least three cases in a cell before an estimate was reported for that cell. The application of this rule resulted in a sizable amount of suppression. Since the application of the Rule of Three with the 2006–2010 ACS would lead to substantially more suppression due its smaller sample size. To reduce the risk of a successful table-differencing attack with the 2006–2010 CTPP, the prespecified universes and variables were reviewed prior to the table production with an objective of preventing slivers from occurring. Strict limitations were placed on the occurrence of overlapping classifications for variables and on the production of two tables that differed only to the extent that their universes were slightly different. Even though the risk from table differencing was reduced to a certain extent by these limitations, the potential attacks of table differencing, linking explicit tables and record linkage, still existed.

The MACH approach used for the 2006–2010 CTPP generated an ACS dataset that comprised a mixture of original ACS data and randomly perturbed values. This dataset was then used to generate all CTPP tables. By "original ACS data," we refer to the microdata that had already undergone 1) ACS standard data perturbation (including data swapping for households and perturbation for group quarters) and 2) imputation for item nonresponses (including stochastic imputation applied specifically for CTPP purposes for records missing workplace locations below the place level). By "CTPP perturbed data," we refer to the original ACS data that were subsequently perturbed by the approaches discussed below. The CTPP perturbation was carried out on the demographic variables but not on the geographic variables, which were subject to imputation (workplace locations) and the regular ACS data-swapping procedures. The following sections describe how the CTPP perturbation procedures were implemented, beginning with a risk assessment, and continuing with a description of the perturbation process, and the quality checks that were performed on the CTPP perturbed data file.

5.1. Risk assessment

The Census Bureau's Disclosure Review Board (DRB) established two general threshold rules that served as a basis for determining the allowable level of disclosure risk for the CTPP: 1) without perturbation, means and aggregates must be based on at least three unweighted records for every table cell, and 2) counts must be based on at least three unweighted records for any marginal for means of transportation (MOT) and for any cell in one-way tables within a journey-to-work flow excluding MOT. This second threshold was introduced to reduce the disclosure risk from table linking and is required because MOT is a common variable throughout the CTPP tables.

At the outset, the CTPP tables were divided into two sets. Set A comprised tables that were not subjected to the above threshold rules, such as when the table did not involve MOT for residence or workplace-based TAZ-level tables (e.g., counts for categorized age and sex). Set B comprised all tables that were targeted accordingly to the original DRB threshold rules. Set A tables were generated from the original ACS data. Set B tables were generated from the CTPP perturbed data. Some rounding rules were applied for all tables.

The Set B tables were processed for each TAZ to determine the cells that violated the threshold rules. For each table, each of the variables defining a violating cell was identified and flagged on the survey records that fell in those cells. The flag indicated whether the variable was associated with a cell containing a single record, two records, or three or more records. A record's final flag for a given variable indicated the smallest number of records in the cells with which the record was associated across all tables. Based on the final flag, each record was then assigned to a risk stratum for the purpose of assigning perturbation rates for the given variable. Since data values that had been

swapped, perturbed, or imputed in the standard ACS random procedures did not require further perturbation, they were separated off and excluded from the perturbation process.

The rest of the data records were allocated to one of three risk strata: a record was assigned to Stratum 1 if it was a singleton for any table involving the target variable, to Stratum 2 if it was a doubleton for any table involving the target variable, and to Stratum 3 if it was one of three or more records for all tables involving the target variable. The DRB assigned perturbation rates to each risk stratum for each target variable. In Strata 2 and 3, records were sampled with equal probability to be part of the perturbation process, that is, to allow their values on the target variable to be perturbed. Since it was desired that more of the values for the target variable for the records with many variables in Stratum 1 should be perturbed, these high-risk records were sampled at higher rates. This oversampling was achieved by sampling target records in Stratum 1 using probability proportional to size sampling, where the measure of size for a record was set equal to the number of its target variables that were assigned to either Stratum 1 or 2. Variable-specific target selection flags were assigned to all records. These flags indicated the target records that were selected for the application of the perturbation process for a particular variable.

5.2. Application of the MACH approach to the CTPP underlying microdata

The perturbation was applied at two levels. Household-level variables (e.g., household income) were perturbed first in the household-level file, and then the values were transferred to the records for each person within the household in the person-level file. Within each file level, the ordinal variables were perturbed first since the constrained perturbation approach used for ordinal variables performed better than the unconstrained approach used for nominal variables.

In general, the hotdeck cells were formed by the cross-classification of the target selection flag, bins (for ordinal variables), locality, prediction groups, and weight groups. In view of the small sample sizes in TAZs, the locality used in the hot deck cell formation was larger than TAZ and depended upon the target variable. For example, the residence Public Use Microdata Area (PUMA) was used for travel time (journey to work), and the workplace tract was used for industry classification. Since the perturbation procedure exchanges the ACS responses for the target records within a hot deck cell, using locality in the cell formation implies that the target variable's unweighted empirical univariate distribution is unaltered at the locality – but not at the TAZ – level.

In the construction of prediction groups a stepwise linear regression model was used throughout, with the candidate predictor variables obtained from ACS household and person variables, block-level variables derived from 2010 census (e.g., percentage of Blacks, percentage of occupied housing units), and area-level variables derived from the ACS (e.g., median household income, average household size). The areas for which the ACS area-level predictor variables were created were traffic analysis districts (TADs), which are CTPP geographic-defined entities that are formed by combining TAZs to areas that have population sizes of at least 20,000. Some interaction terms were included in the pool of candidate predictor variables. The models were constructed separately for each Core-Based Statistical Area (CBSA) and the remainder of each state. After the modeling was completed, model predictions were computed as described in Section 3.1 and the data replacement was done as described in Section 3.4.

When all target variables were perturbed, raking was applied to ensure consistency with specified totals for larger geographic areas, first calibrating the household weights, and then the person weights. The control totals were estimated using the original ACS variables. They were computed based on the full sample weights and also on each of set of replicate weights. The resulting estimated control totals were different across the replicate samples. The full sample and each of the replicates was raked independently to recapture the sampling error component of the variance. At the household level, the household weights were calibrated to the estimated total households in traffic analysis districts (TADs), and by PUMA for select ACS household variables (e.g., the number of vehicles in the household). The person weights were raked to total workers in TADs, and by PUMA for select ACS variables (including target variables). Most totals were generated for places where respondents lived; however, some were based on where they worked, such as MOT and industry classification.

5.3. Evaluation of perturbation applied to the CTPP

Two main types of checks were conducted on the perturbed data set: 1) checks on the impact of perturbation on data utility and 2) checks on the level of dis-

closure risk after perturbation. The checks on data utility compared ACS estimates computed prior to perturbation with those computed after perturbation. The set of estimate consisted of weighted cell means, weighted cell quantiles, weighted cell counts, and standard errors at the TAZ and county levels (for flows as well). In addition, the impact of perturbation on measures of association was examined using the Cramer's V statistic, Pearson product moment correlation, and also on measures of multivariate associations using the U statistic. With the relatively large MSEs for ACS estimates for small areas such as block groups and TAZs, the results of the before/after perturbation comparisons showed that the changes due to perturbation were very small relative to the MSEs except for a few extreme cases. This finding is consistent with an earlier evaluation of travel model output comparisons reported in Appendices R and S of [23].

A disclosure risk measure was created to check on the level of disclosure risk remaining after perturbation. The measure included components that took into account the ACS sampling rate, the effects of mobility and workplace changes over time, the effects of data swapping and imputation in the original ACS data, and the effects of the perturbations in the CTPP data. The DRB reviewed the results and accepted the level of risk.

6. MSE estimation with perturbed data with application to CTPP

Perturbation has the effect of decreasing the accuracy the survey estimates. The accuracy of an estimate based on a perturbed data set should capture both the variance due to sampling error and the MSE due to perturbation given the sample. Appropriate methods have been developed in [29,30] for variance estimation for use with fully synthetic and partially synthetic public-use datasets. However, little research has been conducted to develop methods for MSE estimation for use with perturbed datasets. For practical application with the CTPP data and the very large number of estimates produced, the MSE estimator needed to be computationally straightforward, yet have adequate precision. We present two such MSE estimators that are also robust to the perturbation approach. These estimators take advantage of information from both the original data and the perturbed data, and can be computed while the table products are being prepared and processed.

The MSE estimators proposed for handling the effects of perturbation together with sampling variance start out from the computation of the sampling variance for an estimate calculated from the original ACS dataset. In the case of the unperturbed ACS, sampling variances are computed using a Successive Difference Replication (SDR) approach [31–33] that approximates the variances of the estimates under the survey's complex design and weighting adjustment process. The approach is designed to be used with systematic samples for many types of statistics. Let $\hat{\theta}$ represent the ACS estimate of a statistic, θ, using the ACS full sample weight, and $\hat{\theta}_k$ be the ACS estimate of θ using the kth set of ACS replicate weights (see [34] Chapter 12, for the formation of the replicate weights). Then, with the 80 replicates used with the ACS, the variance of $\hat{\theta}$ is estimated by using the SDR formula:

$$v_{ACS}\left(\hat{\theta}\right) = \frac{4}{80}\sum\nolimits_k \left(\hat{\theta}_k - \hat{\theta}\right)^2. \tag{1}$$

With large samples, this variance estimate has approximately 53 degrees of freedom [35]. For the smaller geographic areas for which CTPP estimates are produced, and also for subgroup estimates, the degrees of freedom can be much smaller than 53.

One naive approach for estimating the precision of the perturbed estimates is to apply the ACS formula directly to the perturbed data, ignoring any bias arising from the perturbation. This naive variance estimator v_1 is

$$v_1\left(\tilde{\theta}\right) = \frac{4}{80}\sum_k \left(\tilde{\theta}_k - \tilde{\theta}\right)k \tag{2}$$

where $\tilde{\theta}$ and $\tilde{\theta}_k$ are computed from the perturbed data set with the raked full sample and the k sets of raked replicate weights. (Note that after data perturbation the raking process is applied to the full sample weight and to each set of replicate weights.) This variance estimator does not directly address perturbation variance and ignores potential perturbation bias. However, it may prove adequate if the amount of perturbation is small.

A simple approach for taking account of the effect of perturbation is to add a term that estimates the perturbation MSE to the estimated sampling variance. The overall MSE of the perturbed estimator can be expressed as

$$MSE(\tilde{\theta}) = E_p E_s(\tilde{\theta} - \theta)^2 = E_s(\hat{\theta} - \theta)^2 + E_p(\tilde{\theta} - \hat{\theta})^2 + 2E_p E_s(\tilde{\theta} - \hat{\theta})(\hat{\theta} - \theta) \tag{3}$$

where E_p and E_s denote expectations with respect to perturbation and sampling, respectively. The first term on the right-hand side of Eq. (3) is the variance of the

Table 6
Degrees of freedom and corresponding t-value by contribution of perturbation error to overall variance and number of multiple perturbations (assume ACS has 50 degrees of freedom [df])

Number of perturbed datasets	Percentage contribution of perturbation error to total variance													
	5%		10%		15%		20%		30%		40%		50%	
	df	t	df	t	df	t	df	t	df	t	df	t	df	t
1	49	2.01	38	2.02	27	2.05	19	2.10	10	2.23	6	2.57	4	3.18
3	53	2.01	51	2.01	46	2.01	38	2.02	25	2.06	17	2.12	11	2.20
5	54	2.01	55	2.00	53	2.01	48	2.01	36	2.03	26	2.06	18	2.10

unperturbed mean, which is estimated by Eq. (1). The second term of Eq. (3) is the perturbation MSE, which, since $\hat{\theta}$ is known, can be simply estimated by $(\tilde{\theta} - \hat{\theta})^2$. The covariance term will generally be close to zero and, as an approximation, is ignored if the perturbation bias is zero or does not depend on the sample. In the application to CTPP, the covariance term was assumed to be negligible. A further empirical study would help determine if the assumption is valid, especially for small sample estimates. Thus, an estimator of the overall MSE of the perturbed estimate $\tilde{\theta}$ was estimated by mse_2:

$$mse_2\left(\tilde{\theta}\right) = v_{ACS}\left(\hat{\theta}\right) + \left(\tilde{\theta}_i - \theta\right)^2 \tag{4}$$

Note that this variance estimator does not present any additional disclosure risk provided that the microdata underlying the tabulations are not released; without the microdata, an intruder cannot derive the unperturbed estimate $\hat{\theta}$ by separating the two components in Eq. (4).

A limitation of mse_2 in Eq. (4) is that the second term, $\left(\tilde{\theta} - \hat{\theta}\right)^2$, is based on a single degree of freedom. This limitation can be addressed by repeating the production of perturbed datasets multiple times, estimating the perturbation MSE for each dataset, and averaging these perturbation MSE estimates. The resultant MSE estimator mse_3, is given by

$$mse_3\left(\tilde{\theta}\right) = v_{ACS}\left(\hat{\theta}\right) + \frac{1}{m}\sum_{i=1}^{m}\left(\tilde{\theta}_i - \theta\right)^2 \tag{5}$$

where $\tilde{\theta}$ is the estimate of θ computed from the ith perturbed dataset $(i = 1, 2, \ldots, m)$ using the raked full sample weights. Increasing the number of replicates m increases the precision of the MSE estimator but adds significantly to the computational burden. The variance of the second term in the estimator mse_3 is m^{-1} times that of the corresponding term in mse_2.

A small-scale simulation study was conducted to investigate the effect of the perturbation on the accuracy of estimates of the mean travel times for workers who drove alone in subareas of two test sites (Olympia, Washington, and the combination of Henry and Clayton counties in Atlanta, Georgia), using the three MSE estimators described above with 5-year ACS sample data from 2005–2009. The simulation study examined the effect of the perturbation at two levels of geographic aggregation (small and large) formed by combining TAZs until there were at least 300 sampled workers who lived in the area (termed CTAZ300) and until there were at least 50 sampled workers who lived in the area (CTAZ50). The Olympia test site contained 22 CTAZ300 areas and 87 CTAZ50 areas. The Atlanta test site contained 33 CTAZ300 areas and 105 CTAZ50 areas. The MACH approach excluded the modeling process, and the number of iterations was limited to 400 iterations for mse_2 and 2,000 iterations for mse_3 (5 imputations occurred 400 times). The variables used in forming the hot deck cells included the target selection flag, a single set of bins on travel time, locality (PUMA), auxiliary variables (MOT and categories of time leaving home), and weight groups. In each iteration, the ACS sample was perturbed and estimates were generated. For the variance estimator v_3, each iteration involved generating five independently perturbed datasets. The simulation results showed that, on average, the MSE estimates were about 17 percent larger than v_1 for small CTAZs and 8 percent larger for large CTAZs in Olympia, and 12 percent larger for small CTAZs and 10 percent larger for large CTAZs in Atlanta. In general, the average TAZ size for the CTPP is slightly larger than the small combined TAZs used in the simulation.

An issue to be resolved is how many replications of the perturbation process should be used, balancing the increased precision of the MSE with more replicates against the added computational burden. To investigate this issue, we examined the relationship between the percentage of total MSE contributed by the perturbation MSE, the approximate number of effective degrees of freedom, and the associated 95 percent t-value. Table 6 gives the effective degrees of freedom of mse_3 computed by the Satterthwaite approxi-

mation [36], assuming the two terms in mse_3 are independent, for $m = 1$, 3, and 5 sets of perturbations. Suppose that the perturbation accounts for 20 percent of the overall MSE. Using single perturbation would only give 19 degrees of freedom and result in a corresponding t-value of 2.10, whereas using five multiply-perturbed datasets would give approximately 48 degrees of freedom with a t-value of 2.01. Assuming the bias associated with the perturbation is not large, in expectation the use of five multiply-perturbed datasets reduces the width of the confidence interval by about 5 percent. If the perturbation accounts for 50 percent of the total MSE, then there would by approximately only 4 degrees of freedom (d.f.) for a single perturbation dataset and 18 d.f. for five multiply-perturbed datasets, resulting in a confidence interval about 50 percent wider with the single perturbation dataset.

Ignoring perturbation error results in confidence intervals that are too narrow, which leads to actual coverage rates lower than the nominal level, especially when the perturbation error contributes a relatively large percentage to the overall variance. Ignoring a perturbation error that accounts for 20 percent of overall MSE, with 50 d.f., and an assumed 95 percent confidence level, the actual coverage rate is 91 percent. If the perturbation error is about 50 percent of overall MSE, the actual coverage rate is only 77, far less than the nominal level of 95 percent.

7. Concluding remarks

The MACH approach can be used in general when reducing the disclosure risk in publishing microdata and tables. It constrains the perturbation deviations for targeted ordinal variables, takes advantage of the predictability of regression models, and larger datasets allow one to limit the donor pool to geographic clusters and to cases with similar weights. The evaluation of the procedure found that the major benefit of the MACH is unsurprisingly the constraining aspect. The MACH approach was programmed as a SAS macro called *SDCPert* under contract to the U.S. Census Bureau. Its' application to the CTPP satisfied the Census Bureau's DRB rules, provided an operationally feasible approach for generating a massive number of tables for small areas, and satisfied transportation analysts in terms of data quality.

Considering the computing resources and statistical properties, the implementing the MSE estimator mse_2 with a single perturbed dataset was manageable and was the accepted choice for the CTPP tables to account for the effects of perturbation. The approach created extra work in generating the CTPP tables beyond that needed to compute the naive variance estimator. If computing resources are available, then the use of more than one perturbation dataset for estimating the MSE due to perturbation would be recommended over the use of a single perturbed dataset, at least for estimates for which the perturbation error is a sizable proportion of total MSE. More evaluation of the contribution of the perturbation MSE to total MSE is needed across a range of estimates and variables in order to decide on how many perturbed datasets are needed.

Acknowledgements

The research discussed herein was partly performed under NCHRP Project 08-79 by Westat with subcontractor Vanasse Hangen Brustlin and consultant Michael Larsen of George Washington University, and further research and development under contract to the Census Bureau. Due to data security requirements relating to the ACS data, much of the work was done at the Census Bureau. The authors gratefully acknowledge the many individuals who contributed to the research, development, and evaluation of the approach. Thanks goes to Nanda Srinivasan, the project officer for NCHRP 08-79, Guy Rousseau, the Transportation Research Board Panel chair, and the many panel members for their review and valuable comments and suggestions. At the Census Bureau, special thanks to the members of the DRB for facilitating timely discussions. The authors are indebted to Chad Russell for accommodating our various system needs and to Brian McKenzie, Liza Hill, Alison Fields, and David Raglin. At Westat, our thanks goes to the senior statistical advisory group, including David Judkins, J. Michael Brick, David Morganstein, as well as Graham Kalton, whose extra efforts helped improve this paper greatly.

References

[1] M. Elliot, Disclosure risk assessment, in: P. Doyle, J. Lane, J. Theeuwes and L. Zayatz, eds, Confidentiality, disclosure, and data access: Theory and practical applications for statistical agencies. Amsterdam, The Netherlands: Elsevier; 2001, 75–95.

[2] W. Winkler, Matching and record linkage. Washington, DC: U.S. Census Bureau, 1993.

[3] T. Krenzke, J. Gentleman, J. Li and C. Moriarity, Addressing disclosure concerns and analysis demands in a real-time online analytic system, *Journal of Official Statistics* **29**(1) (2013), 99–134.
[4] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E.S. Nordholt, K. Spicer and P.P. de Wolf, Statistical disclosure control. Chichester, UK: John Wiley & Sons, 2012.
[5] L.H. Cox, Suppression methodology and statistical disclosure control, *Journal of the American Statistical Association* **75** (1980), 377–385.
[6] B. Fraser and J. Wooten, A proposed method for confidentialising tabular output to protect against differencing. Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality; Nov 9–11; Geneva, Switzerland: Australian Bureau of Statistics, 2005.
[7] R. Dandekar, Maximum utility-minimum information loss table server design for statistical disclosure control of tabular data. In: Domingo-Ferrer J, Torra Lecture V, editors. Privacy in statistical databases. New York: Springer; 2004, 121–135.
[8] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke and L. Vilhuber, Privacy: Theory meets practice on the map. ICDE 2008. Proceedings of the 2008 IEEE 24$^{\text{th}}$ International Conference on Data Engineering; 2008 April 7–12; Cancun, Mexico. Washington, DC: IEEE Computer Society; 2008: 227–286.
[9] D.B. Rubin, Discussion: Statistical disclosure limitation, *Journal of Official Statistics* **9** (1993), 462–468.
[10] F. Liu and R.J.A. Little, Multiple imputation and statistical disclosure control in microdata. Joint Statistical Meetings Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association; 2012: 2133–2138.
[11] T.E. Raghunathan, J.M. Lepkowski, J. van Hoewyk and P. Solenberger, A multivariate technique for multiply imputing missing values using a series of regression models, *Survey Methodology* **27** (2001), 85–96.
[12] S. Fienberg and J. McIntyre, Data swapping: Variations on a theme by Dalenius and Reiss, *Journal of Official Statistics* **21**(2) (2005), 309–323.
[13] R. Moore, Controlled data swapping techniques for masking public use datasets. U.S. Census Bureau Statistical Research Division Research Report RR96/04. Washington, DC: U.S. Census Bureau. 1996. Available from https://www.census.gov/srd/papers/pdf/rr96-4.pdf.
[14] K. Muralidhar and R. Sarathy, Data shuffling – A new masking approach for numerical data, *Management Science* **52** (2006), 658–670.
[15] J. Domingo-Ferrer and J. Mateo-Sanz, Practical data-oriented micro aggregation for statistical disclosure control, *IEEE Transactions on Knowledge and Data Engineering* **14**(1) (2002), 189–201.
[16] J. Gouweleeuw, P. Kooiman, L. Willenborg and P.P. de Wolf, Post randomisation for statistical disclosure control: Theory and implementation, *Journal of Official Statistics* **14**(4) (1998), 463–478.
[17] P.P. de Wolf, J. Gouweleeuw, P. Kooiman and L. Willenborg, Reflections on pram. Statistical Data Protection, *Luxembourg: Office for Official Publications of the European Communities* (1998), 337–349.
[18] G. Matthews and O. Harel, Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy, *Statistics Surveys* **5** (2011), 1–29.
[19] U.S. Census Bureau. American FactFinder. 2015. Available from http://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml.
[20] S. Hawala, L. Zayatz and S. Rowland, American FactFinder: Disclosure limitation for the advanced query system, *Journal of Official Statistics* **20**(1) (2004), 115–124.
[21] A. Solanas and A. Martínez-Ballesté, V-MDAV: Variable group size multivariate micro aggregation. COMPSTAT 2006. Rome, Italy, 2006; 917–925.
[22] S. Kaufman, M. Seastrom and S. Roey, Do disclosure controls to protect confidentiality degrade the quality of the data? Proceedings of the Joint Statistical Meetings, Section on Government Statistics. Alexandria, VA: American Statistical Association, 2005; 1218–1225.
[23] T. Krenzke, J. Li, M. Freedman, D. Judkins, D. Hubble, R. Roisman and M. Larsen, Producing transportation data products from the American Community Survey that comply with disclosure rules. Washington, DC: National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences, 2011.
[24] J. Reiter, Using CART to generate partially synthetic public use microdata, *Journal of Official Statistics* **21**(3) (2005), 441–462.
[25] D. Judkins, A. Piesse, T. Krenzke, Z. Fan and W.C. Haung, Preservation of skip patterns and covariance structure through semi-parametric whole-questionnaire imputation, *Joint Statistical Meetings Proceedings of the Section on Survey Research Methods of the American Statistical Association* (2007), 3211–3218.
[26] R.R. Andridge and R.J.A. Little, A review of hot deck imputation for survey non-response, *International Statistical Review* **78**(1) (2010), 40–64.
[27] G. Kalton and I. Flores-Cervantes, Weighting methods, *Journal of Official Statistics* **19**(2) (2003), 81–97.
[28] M. Woo, J. Reiter, A. Oganian and A. Karr, Global measures of data utility for microdata masked for disclosure limitation, *Journal of Privacy and Confidentiality* **1**(1) (2009), 111–124.
[29] T.E. Raghunathan, J.P. Reiter and D.B. Rubin, Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics* **19** (2003), 1–16.
[30] J. Reiter, Inference for partially synthetic, public use microdata sets, *Survey Methodology* **29** (2003), 181–188.
[31] K.M. Wolter, An investigation of some estimators of variance for systematic sampling, *Journal of the American Statistical Association* **79** (1984), 781–790.
[32] R. Fay and G. Train, Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. Proceedings of the Section on Government Statistics. Alexandria, VA: American Statistical Association; 1995: 154–159.
[33] D.R. Judkins, Fay's method for variance estimation, *Journal of Official Statistics* **6**(3) (1990), 223–239.
[34] U.S. Census Bureau. Variance estimation. American Community Survey design and methodology report. 2009. Available from http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology_ch12.pdf.
[35] E.T. Huang and W.R. Bell, A simulation study of the distribution of Fay's successive difference replication variance estimator. Proceedings of the American Statistical Association, Survey Research Methods Section, [CD-ROM]. Alexandria, VA: American Statistical Association, 2009.
[36] F.E. Satterthwaite, An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**(6) (1946), 110–114.