Balancing Use of Weights, Predictions, and Locality Effects in a Model-Assisted Constrained Hot Deck Approach for Perturbation

Tom Krenzke[1], Jianzhu Li[1], Laura Zayatz[2]
[1]Westat, 1600 Research Blvd, Rockville MD, 20850
[2]U.S. Census Bureau, 4600 Silver Hill Rd, Washington, D.C. 20233

Abstract

This paper focuses on applying a random perturbation approach that protects microdata for the purpose of releasing data to the public. The classical challenge is to balance the need to reduce disclosure risk and retain data utility. An approach has been developed that provides the data producer flexibility to achieve the balance. Hot deck cells are formed from sampling weights, model predictions and/or covariates, the locality of the target records, and categorized bins of the target variable. Expanding or contracting the bin sizes allows the data producer the flexibility to control the distance between original and perturbed values. An evaluation was conducted to study the impact of the bin categories, sampling weights, model predictions and locality effects using the American Community Survey 2005-2009 sample data.

Key Words: Confidentiality, statistical disclosure control, microdata, data utility

1. Introduction

During 2009-2011, under contract to the National Academies of Science and the National Cooperative Highway Research Program (NCHRP), Westat worked with the U.S. Census Bureau to investigate statistical disclosure control (SDC) treatments to perturb American Community Survey (ACS) data prior to generating the Census Transportation Planning Products (CTPP) for the American Association of State Highway and Transportation Officials (AASHTO). By "perturb," we mean that an SDC treatment is used to modify data values through a controlled approach with a random mechanism. The CTPP includes data derived from specific transportation questions related to commuting times, distance from home to work, and mode of travel. Of greatest importance to transportation planners, however, was the unique ability to provide this information for cities and towns of different sizes, as well as for tracts and block groups, or combinations of these groups into traffic analysis zones (TAZ).

Previously, the tabulations were subject to suppression procedures to protect against the disclosure of identifiable information. To eliminate the effects of data suppression on small-area data and retain the necessary attributes to support the desired micro modeling at the TAZ level, a research study was undertaken (NCHRP 2011) to develop data perturbation procedures on microdata that would be used to produce local area pre-specified tabular estimates that would not violate the Census Bureau's confidentiality law. The results have attribute variables that are sufficiently perturbed to pass the Census

Bureau's Disclosure Review Board rules. The research team developed, refined, and thoroughly evaluated the credible data perturbation techniques that were identified through the initial critical assessment of plausible approaches.

Subsequently, the resulting approach, a constrained hot deck, was further developed under sponsored research with the Census Bureau. The resulting model-assisted constrained hot deck (MACH) constrains the amount of change in the target variable by forming hot deck cells using "bins" created on the target variable itself, model predictions, locality, and sampling weights. The MACH approach is used to replace observed data for the purpose of reducing disclosure risk. During the CTPP production stage, using the Census Bureau's information technology (IT) systems on site, we processed the perturbation programs on the full 2006-2010 ACS 5-year sample. The resulting perturbed ACS microdata were used in generating the CTPP tables. This report contains details of the perturbation approach.

While developing the MACH approach, an evaluation was conducted to provide guidelines for decisions when weights, covariates, and localities are used in the perturbation process. When weights are highly variable, the weights should have more influence on the perturbation. Similarly, if the set of covariates has moderate to high correlation with the target variable then the model predictions should have more influence on perturbation. Likewise, if the size of the locality is small, the units within the locality may be more homogeneous than if the size of the locality is large. Therefore, if the localities are very different from each other, then locality should have more influence. In this paper, the research team provides the evaluation results.

2. Perturbation Approach

Perturbation approaches have to meet several goals, such as reducing the risk of disclosure and maintaining data utility (including univariate and multivariate distributions), and doing so in an operationally efficient manner. An objective of the MACH procedure is to change the value of the published categories by only one or two categories by changing the value of the continuous version of the variable. The MACH approach is relevant to ordinal variables with at least three levels. A variation of the approach is a model-assisted unconstrained hot deck, which we hereafter refer to as the semi-parametric (SP) approach. The SP approach can be applied to any variable type.

The MACH approach constrains the amount of change on the target variable, similar to rank swapping (Greenberg 1987). A nice feature of this approach is that it can control the amount of perturbation by expanding the bin widths. Expanding or contracting the bin sizes allows the data producer the flexibility to control the distance between original and perturbed values.

The MACH and SP approaches are model-assisted in that they use the model parameters from the model selection process to generate predicted values to use in forming hot deck cells. Hot deck imputation begins by forming cells, or groups of data records, from auxiliary variables. Within the cells, missing data for a record is filled-in using a donor's value randomly drawn from within the cell. The hot deck approach strives to retain the structure of the observed data. More details about hot deck imputation are provided in Andridge and Little (2010).

The approaches are influenced by an imputation procedure, described in Judkins et al. (2007). Initially designed for handling non-monotone (swiss cheese) missing data patterns in complex questionnaires, the imputation process in general uses model predictions to form hot deck cells. Influenced by the Gibbs sampler (an iterative method for simulating posterior distributions in Bayesian analysis through sampling from alternating conditional distributions until convergence in distribution is achieved), the imputation process is done variable-by-variable, using previously imputed data in the model selection and estimation process, as well as in the prediction equation. The process proceeds sequentially through all variables needing imputation. The perturbation model can be expressed in general as follows:

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

where, subscript (c) denotes the c^{th} class (hot deck cell) defined from the set of factors $\{I(s), y_{g'}, \boldsymbol{x}, \hat{y}_{\boldsymbol{g}''}, \boldsymbol{w}_{\boldsymbol{g}'''}\}$, 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, $\hat{y}_{\boldsymbol{g}''}$ are the g'' groups formed from model predictions, $\boldsymbol{w}_{\boldsymbol{g}'''}$ are the g''' groups formed from the sampling weights and where $\varepsilon_{i(c)} \sim y_{i(c)} - \tilde{y}_{i(c)}$ resulting from the random error associated with case i for a random with-replacement draw within the c^{th} class. The bolding pattern represents vectors. The SP approach excludes the bins ($y_{g'}$) as part of the hot deck cells.

2.1 Challenges and Key Features

When developing the perturbation approach, there were a number of methodological challenges, and when addressed, the challenges transitioned to the following key features.

- **Variable types.** There are different types of variables to be perturbed (continuous, ordinal categorical, and unordered categorical). The SP approach is used for unordered categorical variables and binary variables, and uses the MACH for ordered categorical variables with at least three levels and continuous variables.
- **Several categorical versions.** The same variable may have multiple versions. For example, categorical household income may have 5-levels, 9-levels, and 26-levels, and continuous income. The version with the most detailed categories (or continuous) is used in the modeling and mapped to the other versions once perturbed.
- **Multiple file levels.** Data can include both household- and person-level data, where a two-stage approach can be employed. The household-level variables (e.g., household income) are perturbed first, then the values are transferred to each person within the household.
- **Weights.** Sampling weights can be quite variable, even within small areas, due to differential sampling rates, nonresponse follow-up sampling, and weighting adjustments. Therefore, groups based on the magnitude of the weights are formed in the process of identifying donors for cases that need to be perturbed.
- **Predictive strength of covariates**. There are many variables to consider when constructing hot deck cells for perturbation. Using a pool of predictor variables, the approach selects a set of predictors that has the highest prediction success (e.g., R^2 values). Predicted values are used to form prediction groups.

- **Automated collapsing of hot deck cells**. An algorithm has been implemented to conduct an automated collapsing of initial hot deck cells. Given a list of variables that form a nested stratification of the sample cases, cells with small sample sizes are combined with the preceding cell.
- **Highly related variables**. In many cases, there are target variables that may have associated recodes or derived variables, or may be either structurally linked through skip patterns in the questionnaire or highly correlated, such as earnings and poverty status. A mechanism to keep target and non-target variables consistent with each other is used. As primary target variables are perturbed, so do others to retain logical consistency.
- **Retain unweighted distribution**. Control over the unweighted one-way distributions is an option through a parameter in the process.
- **Random assignment to overlapping bins**. When bins are formed on the target variable, a target record with a value on the boundary of a bin could only have its value replaced by a lower value or an upper value, depending on if the original value is on the upper or lower boundary, respectively. The overlapping bins solution is discussed in the next section.
- **Ordering of cell variables**. The order of cell variables may matter due to the collapsing algorithm and, therefore, the capability of ordering the cell variables has been set in place.
- **Limit replacement of the same value**. The approach ensures that the donor is not the target record for replacement. If the same value results, an option is available to add noise (discussed later in this paper).

2.2 Process Flow

The main steps of the MACH process are shown in the flowchart in Figure 2-1. Each step is explained further in the following paragraphs.

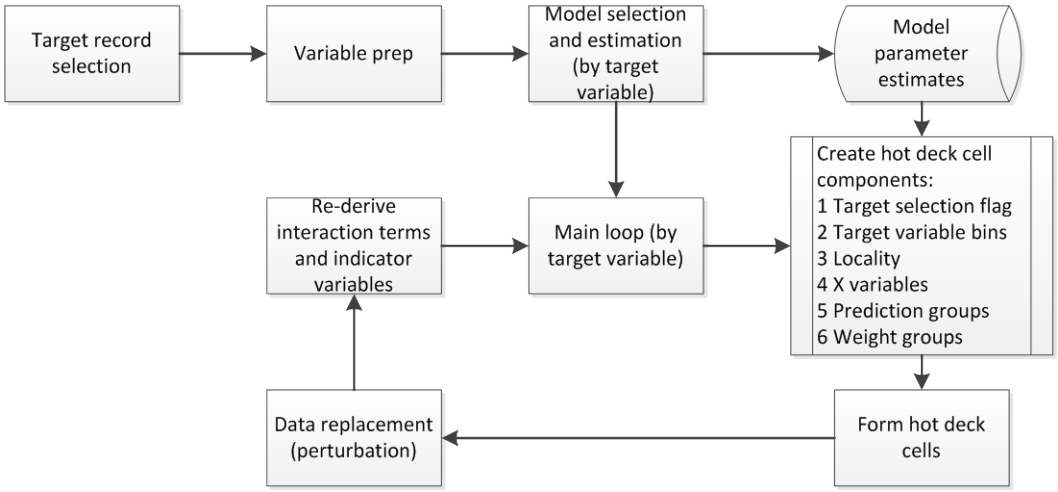

Figure 2-1: Flowchart of Perturbation Program Component

Target record selection. The first step is to set the target selection flag. A risk measure (e.g., Skinner and Shlomo 2008) resulting from an initial risk analysis can be used as a measure of size for the target selection of records for each variable.

Variable prep. The Variable Prep step compiles the pool of predictor variables and prepares recodes for the models. For unordered categorical variables, indicator variables

are created. Select interaction terms to be added to the pool of candidate predictor variables are identified as well.

Model selection and estimation. Next, the model selection approach is processed for all target variables. The purpose is to identify the predictors for each target variable, and to estimate the model parameters for generating predicted values, which are used when creating hot deck cells. The process is done once since the joint distribution among the variables is given, conditional on the fully complete original data.

The model selection process occurs for each model area (e.g., counties). The modeling is done differently for variables of the ordered categorical, continuous (OC) type than for the unordered categorical, binary (UC) type. For OC variables, a stepwise linear regression is processed while bringing in significant predictors to gain predictive power from the model. A clustering procedure is done for UC variables, which fits a separate linear regression for each category of the variable, and subsequently conducts a k-means clustering algorithm on the vector of predicted values for each level. The algorithm produces g clusters to be used in the hot deck cell formation.

To facilitate the discussion of the perturbation approach, we set up a scenario with three target variables. Item 1 is an ordinal variable that is continuous, for which the MACH is applied with additive noise if the perturbed value does not change. This approach is useful if the value absolutely needs to be changed to protect from a record linking attack on a continuous variable. Item 2 is a binary variable with two levels, such that the SP approach is appropriate. Item 3 is an UC variable where the SP approach is used. One cycle through the variables is conducted. Let y_{ki} denote the k^{th} variable to be perturbed for record i, where k is the item number, and y represents the data values. The subscript j identifies indicator variables associated with UC variables (e.g., industry). The bolding pattern represents vectors. The model selection for Item 1, Item 2, and Item 3 is essentially as shown below.

$E(y_1|y_2, \boldsymbol{y_3}, \mathbf{X})$ = f($y_2, \boldsymbol{y_3}, \mathbf{X}, \boldsymbol{\beta}$),
$E(y_2|y_1, \boldsymbol{y_3}, \mathbf{X})$ = f($y_1, \boldsymbol{y_3}, \mathbf{X}, \boldsymbol{\beta}$),
$E(y_{3j}|y_1, y_2, \mathbf{X})$ = f($y_1, y_2, \mathbf{X}, \boldsymbol{\beta}$),

for j = 1,2,…J for J categories

The models are processed to allow predictors to enter the model during the stepwise modeling steps if significant at the $\propto$ = .05 level. Predictors not significant at the .05 level exit the model. Within the candidate predictor pools, interactions can be selected with a UC predictor.

Initial hot deck cells. Hot deck cells are used as part of the perturbation process, and are formed using the following information:

- The target selection flag to retain the unweighted empirical distribution,
- The bins (either Bin A or Bin B) on the target record in order to control the amount of change (MACH only),
- Locality to benefit from homogeneous areas,
- Key auxiliary variables to address highly related variables,
- Groups of predictions from predictive models with strong covariates, and

- Coarsened values of the sample weights to reduce the mean square error in perturbed estimates. Groups of weights are created from a ranking of the weights with an equal number of sampled cases within each group.

When forming hot deck cells, small cells are identified and combined in an automated manner. The following describes select key components of the hot deck cell formation.

Bin formation. The formation of "bins" applies only to variables perturbed through the MACH approach. The hypothetical example in Figure 2-2 illustrates the assignment of bins. The figure depicts a frequency distribution, with spikes at multiples of 5. Within Figure 2-2, below the histogram, the rows illustrate two sets of overlapping bins (Bin A and Bin B) and published categories for the *y* variable. The bins are formed while striving to achieve the following objectives:

- Ensure that the bins contain more than one value of the published categories; and
- Ensure that if there are spikes, then at least two spikes are included in a bin; otherwise, the approach results in values unchanged for many cases.

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

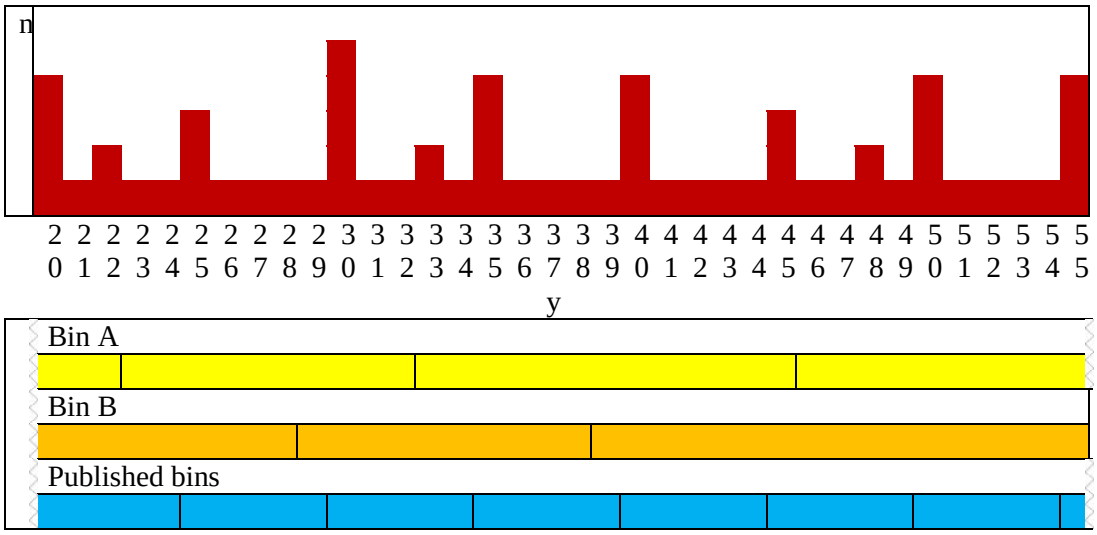

Figure 2-2: Illustration of Bin Formation

Prediction groups. After the model parameters are estimated for all variables, and after the bin formation occurs (MACH approach only), then the sequential prediction steps occur that lead to the formation of groups of predictions from prediction models to be used as a component of the hot deck cell formation. The prediction equation is created from the estimated regression parameters and predictions are computed using either original or perturbed data if already available. Ignoring any interaction terms for simplicity, for Item 1, the prediction equation is,

$$\hat{y}_{1i} = \beta_0 + \beta_2 y_{2i} + \sum_{j=1}^{J} \beta_{3j} y_{3ji} + \sum_{l=1}^{L} \beta_l x_{li}$$

where J is the number of categories for Item 3, and L is the number of other predictor variables.

Then after predictions for the target variable are generated, prediction groups are formed on $\hat{y}_{1i}$. The groups are formed from a ranking of the predictions, with an equal number of sampled cases within each group. Let $\tilde{y}_{1i}$ represent the perturbed value drawn at random without replacement within the hot deck cells. The predictions for Item 2 (OC binary variable) use the perturbed values for Item 1, as follows:

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

For Item 3, there were J categories in this UC variable from which J corresponding indicator variables were formed. Let the prediction equation for the j^{th}category of Item 3 be represented as follows using the perturbed values for the previous two variables, as follows:

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

where j = 1, 2, … J.

For the UC variable, a clustering program (SAS Proc FastClus) is used to form clusters (prediction groups), using the J sets of predicted values $\hat{y}_{3ji}$. Let $\tilde{y}_{3ji}$ represent the perturbed value drawn within the hot deck cell. In general, after a UC variable is perturbed, indicator variables are recreated using the perturbed values.

Perturbation. Within each final hot deck cell, a without replacement draw from the empirical distribution is conducted. The replaced value is obtained through a random draw without replacement from the empirical distribution within the hot deck cell. The values are perturbed only if they are flagged for replacement. When the target selection flag is used, then all records targeted for replacement are used to donate their values to others. This approach retains the overall empirical distribution of the target variable. This is similar to data swapping, which swaps values between pairs while the MACH exchanges values amongst several records. Mechanically, to ensure the donor is not the same as the target record, each record in the cell is indexed with a random sequential number from 1 to n (the number of records in the cell). A random draw is conducted for the first donor, say 3, then sequentially proceeds with 3 as a donor for 1, 4 as a donor for 2, 5 as a donor for 3, and continues until n is a donor for n-2, then 1 is a donor for n-1 and 2 is a donor for n.

After each variable is perturbed, any interaction terms are recreated using perturbed values so the perturbed values could be used in the prediction equation for the next target variables in the sequence.

Suppose for Item 1, the additive noise procedure is conducted on any target record where Item 1 does not change value. That is, during the perturbation step, if left unchanged from the MACH procedure, noise is added to the original Item 1 value y as follows: $\tilde{y}_{1i} = y_{1i}(1 + fz)$, where f is a constant between 0 and 1, and z is a draw from the standard normal distribution. The noise is centered at 0 with a draw from the standard normal

distribution. The standard deviation of the added noise is the product of f and y_{1i}, which means the level of noise is allowed to vary relative to the magnitude of Item 1.

After processing, checks were conducted for an initial look at the impact of the perturbation. Frequencies, means, and correlations were generated before and after perturbation.

3. Evaluation

The objective of the analysis is to improve current perturbation approaches and to provide guidelines for decisions to be made in the future by the Census Bureau when weights, covariates, and localities are used in the perturbation process. The research team used the 5-year accumulation of 2005–2009 ACS data for workers 16 years old and older in the South region. Persons in group quarters were excluded. The evaluation was processed using the MACH and the SP approach. Since the MACH approach is only applicable for ordinal variables, the evaluation focused on two key ordinal items: age and 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 3-1. The number of CBSAs in the South region and in North Carolina is also shown.

Table 3-1: Number of CBSAs by CBSA Type for Age and Earnings

CBSA Type	*Variation in weights*	*Model R^2*	*Variation between localities*	*Age*		*Earnings*	
				Number of CBSAs in South region	*Number of CBSAs in North Carolina*	*Number of CBSAs in South region*	*Number of CBSAs in North Carolina*
1	Low	Low	Low	91	10	44	7
2	Low	Low	High	81	10	36	3
3	Low	High	Low	20	2	36	3
4	Low	High	High	20	3	96	12
5	High	Low	Low	68	8	54	5
6	High	Low	High	33	3	24	2
7	High	High	Low	32	1	52	3
8	High	High	High	52	4	55	6

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 for age (earnings) from 12 percent (40%) to 36 percent (59%) for the low group and 36 percent (59%) to 55 percent (84%) for the high group. Lastly, the variation between localities was computed by the variation among the tract-level mean age and person earnings.

The MACH and SP approaches were applied under various treatment scenarios. Four key factors were defined by use of bins on the target variable (MACH) or not (SP), number of prediction groups, number of weight groups, and size of locality. There were two sizes of the locality (CTAZ300, CTAZ1000). CTAZ300 are combined Traffic Analysis Zones (TAZs) that were formed to have at least 300 ACS sample cases. These zones can have fewer than 300 sample cases due to the exclusion of group quarters for the evaluation. For the South region and North Carolina, the average number of workers in the CTAZ300 entities is 421 and 464, respectively. Similarly, CTAZ1000 was initially formed to have at least 1,000 ACS sample cases. The average size of CTAZ1000 was 1,250 for the South, and 1,365 for North Carolina.

The treatments were defined to arrive at hot deck cells with similar sizes between the same treatment combinations between the MACH and the SP approaches. The experimental design is given in Table 3-2. 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. For each method, the variable earnings were perturbed first, and then age was perturbed using predictions based on perturbed earnings.

Table 3-2: 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	421	11.7
	2	9	2	2	1250	34.7
	3	9	2	4	421	5.8
	4	9	2	4	1250	17.4
	5	9	4	2	421	5.8
	6	9	4	2	1250	17.4
	7	9	4	4	421	2.9
	8	9	4	4	1250	8.7
SP	9	1	6	6	421	11.7
	10	1	6	6	1250	34.7
	11	1	6	12	421	5.8
	12	1	6	12	1250	17.4
	13	1	12	6	421	5.8
	14	1	12	6	1250	17.4
	15	1	12	12	421	2.9
	16	1	12	12	1250	8.7

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

It is important to develop measures for the resulting data utility so that the balance between risk and utility can be understood. We used two measures to help determine the best set of perturbation parameters (number of weight groups, prediction groups, locality size) under certain conditions (variation in weights, strength of covariates, and between locality variation). The measures compare the original ACS data and the perturbed ACS data. 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 two attributes (age and 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 number of tracts in the South region with workers not in Group Quarters (GQs) is 19,390). The two-way tables are a mix of those involving tracts and those not involving tracts. The median and Interquartile range (IQR) of the differences across all table cells between the raw and perturbed data were produced by each CBSA type and overall.

Woo et al. (2009) propose using propensity scores as a global utility measure for microdata as follows. 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 3-3 provides the terms in the model. The following statistic U should be close to zero if the perturbed data and original data were indistinguishable.

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

where

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

The dataset was subset to North Carolina in order to reduce run time. North Carolina was chosen due to its diversity in terms of the CBSA type, that is, each of the eight CBSA types are represented.

Table 3-3: Effects Used in the U Statistic Logit Model

Target variable	*Main effects*	*Interactions*
Age	Age, earnings, poverty	Age with [CTAZ300 (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), Earnings] and CTAZ300 with Earnings
Earnings	Age, earnings, poverty	Earnings with [CTAZ300 (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 CTAZ300 with Age

3.1 MACH and SP comparison

Figure 3-1 shows differences between the MACH and the SP approach. Treatments 1 through 8 were combined for the MACH and Treatments 9 through 16 were combined for the SP. The IQR results provide some indication of the resulting variation of the estimates from the perturbed data. The IQR and U results show much less variation for MACH approach than for the SP approach. The results show a better fit (lower U statistic values) for the MACH (Treatments 1 to 8) than for the SP (Treatments 9 to 16) for each measure and for each variable. The MACH and the SP results are significantly different in each of the four scenarios. This result verifies the evaluation results reported in the NCHRP (2011). Due to these results, we focus the remainder of the analysis on the MACH approach.

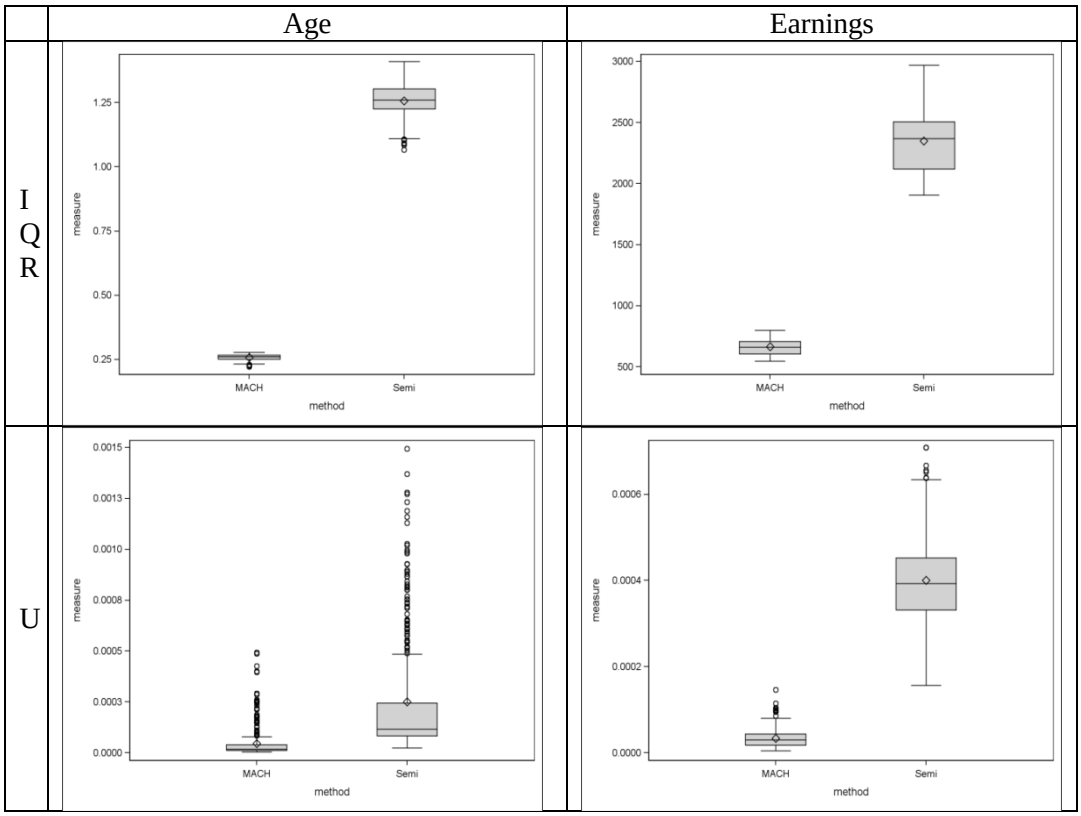

Figure 3-1: Box Plots of the Analysis Measure by Method and Target Variable

3.2 Treatment Comparisons

The main research question is which treatment or treatments are best used for different types of CBSAs. As given in Table 3-1, the CBSA types are a function of the predictability, weight variability, and between locality variance. As given in Table 3-2, the treatments are the number of prediction groups, number of weight groups, and locality.

We hypothesize that smaller cell sizes result in less variability (IQR) in the results and providing a better fit (*U* statistic). While the cell sizes shown in Table 3-2 are approximate, they are very useful to help determine if there is an important cell size factor that needs to be paid attention to in the evaluation. A regression was performed to test the hypothesis of no impact from cell sizes on the resulting measures (IQR, *U*) for the two variables (age, earnings), by method (MACH, SP). The results show the cell sizes are important in one of the four analysis scenarios for the MACH (IQR for age).

Therefore, for the analysis of IQR for age, we first ran an ANCOVA using cell size as a covariate, as well as the 8-level treatment factor. For CBSA types 1 through 7, the overall treatment main effect significance status was the same as an ANOVA without the use of cell size as a covariate. For these CBSA types, we processed a one-way ANOVA with treatment main effect. For CBSA type 8, the ANOVA showed the treatment effect as significant, while the ANCOVA showed treatment effect as not significant. Therefore, we treated the effect as not significant due to the influence of cell size.

Statistical significance was determined the adjusted Tukey pairwise comparison tests and $\alpha = 0.05$. Figure 3-2 provides box plots for each CBSA type. An explanation is provided below for any significant pairwise comparison.

For the IQR measure, significant results were found for 7 of the 16 analysis scenarios (2 target variables (age and earnings) and 8 CBSA types).

For CBSA type 1, we would expect that the attributes for this CBSA type would result in very limited impact from treatments, however, for age, there are significant effects. Treatments 1, 3, and 5 have lower IQRs than Treatment 4. The main difference between the treatments was a smaller locality size.

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 for age, where all the odd number treatments have lower IQRs than the even number treatments. In addition, Treatment 1 has lower IQRs than Treatment 7. Interestingly, the main difference was fewer prediction groups and fewer weight groups, perhaps allowing locality to have more influence. For earnings, Treatments 1, 5 and 7 have lower IQRs than Treatment 2. 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. For age, Treatment 5 has lower IQRs than Treatment 8. The main difference was fewer weight groups and smaller geography. Essentially, fewer weight groups and more localities points to locality being a key factor in this comparison. For earnings, Treatments 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. For earnings, 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.

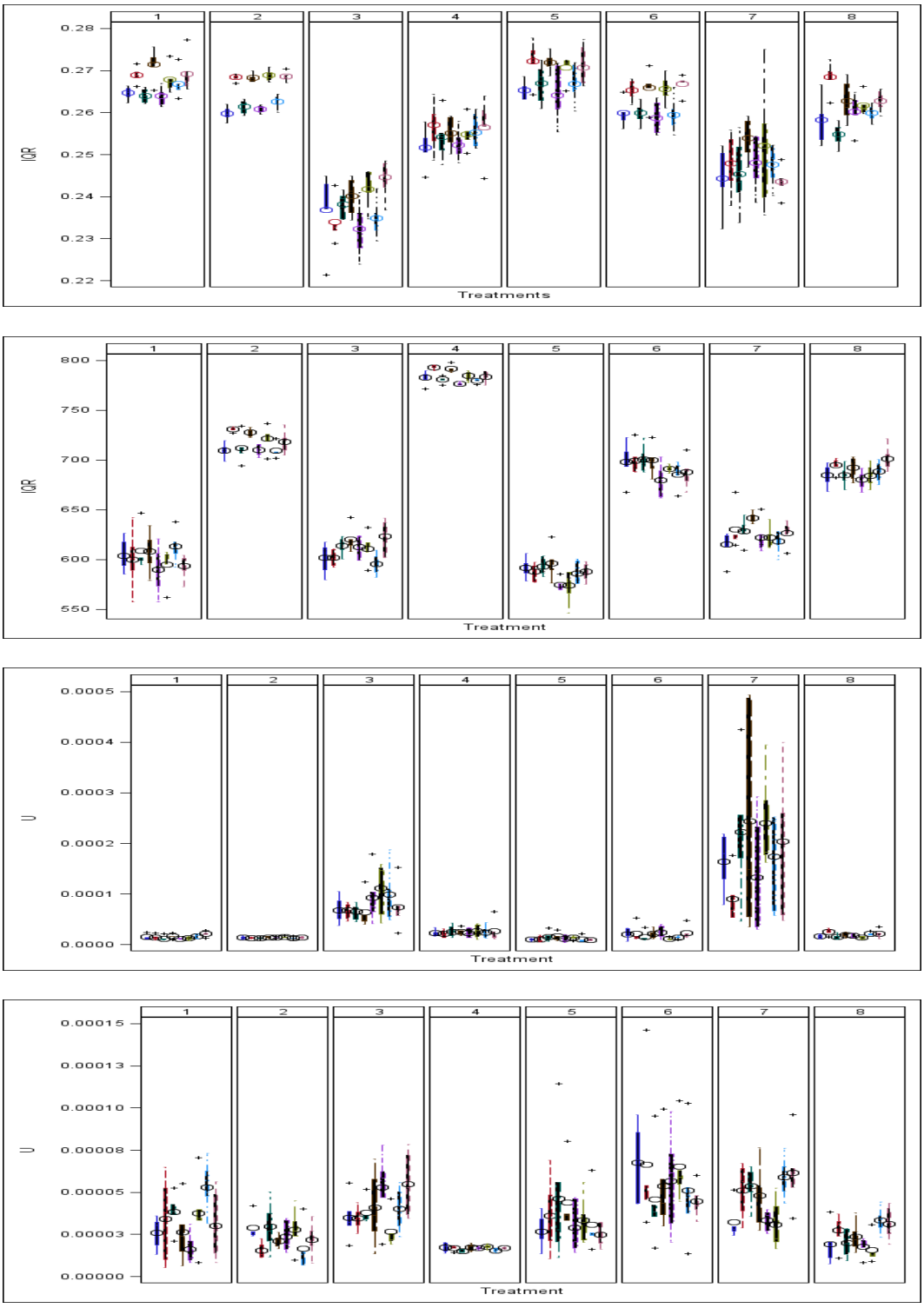

Figure 3-2: Top: IQR for Age for Treatments within CBSA Type
2nd: IQR for Earnings for Treatments within CBSA Type
3rd: U for Age for Treatments within CBSA Type
Bottom: U for Earnings for Treatments within CBSA Type

For CBSA type 6, we would expect that the attributes for CBSA type would result in potential help from the weight groups and smaller locality size. For age, Treatment 5 has lower IQRs than Treatments 4 and 8. The main difference was smaller geography and more prediction groups; however, contrary to expectations, Treatment 5 has a lower number of weight groups. Also, Treatment 7 has lower IQRs than Treatment 8. The main difference was smaller geography. Locality size helped quite a bit with some benefit from the predictions, but no benefit from the weight groups.

For the U statistic, there were not as many significant results. Among the 16 analysis scenarios, only three were significant.

For CBSA type 1, we would expect that the attributes for CBSA type would result in very limited help from treatments. For age, however, Treatment 5 has lower U values than Treatment 4. The main differences were lower geography and more prediction groups.

For CBSA type 7, we would expect that this CBSA type would result in potential benefit from weight groups or prediction groups. For earnings, while treatments have some evidence of significant 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. For earnings, while the overall statistical test was significant, there was not enough evidence to find a significant difference between specific treatments.

4. Summary

The MACH approach has the capability to take advantage of predictability, weights, and locality when perturbing data. The recent enhancements allow the unweighted empirical distribution to be maintained, allows perturbed values to increase or decrease even on bounds of constraining bins, uses model predictions as covariates, orders cell variables when forming hot deck cells, automates the collapsing of hot deck cells, limits the replacement of the same value, and links variables to retain logical consistency.

An extensive evaluation was conducted to determine if it is worth the effort to adapt the perturbation parameters to attributes of areas. Three specific attributes were studied in the analysis: low or high variability of the sampling weights, low or high predictability of covariates, and low or high between locality variance. The major benefit of the MACH is the constraining aspect, which limits the amount of change on the target variable. The research has shown the constrained approach to be superior to its unconstrained counterpart. The research also showed some benefit to conducting an investigation into the special attributes of the data, and that significant improvements can be made by a simple set up of the perturbation parameters. A recommended approach would be to do the following:

- Compute the between locality variance for the target variables. The research showed significant benefits to locality size being small, even when the between locality variance was low.

- Run some regression models to determine the R^2 level. Create more prediction groups when the R^2 value is moderate to high. Cutoffs for low, moderate and high could be determined.
- There was little benefit observed from the weight groups; therefore, we recommend keeping the number of weight groups at a low number (perhaps just two groups).

Because the inclusion of bins on the target value is a key factor in the performance of the MACH approach, future development of the MACH approach would include an automated approach to bin creation. Another potential development would include an automated approach to deciding on the priorities of weights, prediction and locality in the hot deck cell creation. Future research could involve a comparison with other approaches, such as data swapping, rank swapping, or additive noise. Also, an evaluation could be done on different approaches to forming the bins and how each impacts the results. It would also be useful to measure the proportion of data values that changed, and relate that to the utility results. Other metrics to measuring data utility could be developed and incorporated.

References

Andridge, R.R., and Little, R.J.A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1): 40-64.

Judkins, D., Piesse, A., Krenzke, T., Fan, Z., and Haung, W.C. (2007). Preservation of skip patterns and covariance structure through semi-parametric whole-questionnaire imputation. In Joint Statistical Meetings Proceedings of the Section on Survey Research Methods of the American Statistical Association.

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

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