Working Paper
ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)

CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION

STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Ottawa, Canada, 28-30 October 2013)

Topic (ii): New methods for protection of microdata

Comparison of Perturbation Approaches for Spatial Outliers in Microdata

Prepared byNatalie Shlomo, University of Manchester, United Kingdom and Jordi Marés, Artificial Intelligence Research Institute (IIIA) and the Spanish National Research Council (CSIC), Spain

Comparison of Perturbation Approaches for Spatial Outliers in Microdata

Natalie Shlomo[*] and Jordi Marés[**]

[*] Social Statistics and CCSR, University of Manchester, United Kingdom, Natalie.Shlomo@manchester.ac.uk
[**] Artificial Intelligence Research Institute (IIIA) and the Spanish National Research Council (CSIC), Barcelona, Spain, jmares@iiia.csic.es

Abstract: We describe an evaluation study to compare statistical disclosure limitation methods for spatial outliers in microdata. The test dataset for this evaluation study is based on the transportation products of the 2006-2008 combined PUMS from the American Community Survey (ACS) of the United States. The spatial variables are trajectories defined as vectors of coordinates where the first component is the coordinate of place of residence (origin) and the second component is the coordinate of workplace (destination). Variables based on geographical spatial coordinates are particularly prone to disclosure risks since they can easily be visualized when disseminating statistical information through the use of maps. The first stage of the study is an outlier detection algorithm, which accounts for multivariate data and is robust to deviations from normality assumptions. Once outliers are identified, the second stage of the study is to recommend a targeted disclosure limitation method to confidentialize the outliers. We compare the perturbative methods of record swapping and hot-deck with respect to disclosure risk and data utility.

Acknowledgement: The project was funded by the Census Statistical Disclosure Control project at Westat, Inc. through the sponsorship of the U.S. Bureau of the Census.

1 Introduction

Outlier detection and error localization are important components of statistical data editing which is typically carried out during the data processing stage of a survey. Remaining outliers in survey data arising from rare and unique attributes appear as sample uniques on a set of cross-classified identifying variables. Although probability sampling is a good non-perturbative disclosure limitation method and provides a priori protection in the data, sample uniques based on rare attributes have a high probability of being a population unique and therefore pose a high disclosure risk. To ensure the confidentiality of outliers in survey data, disclosure limitation strategies should include localized targeted perturbation on the microdata prior to its tabulation. Variables based on geographical spatial coordinates are particularly prone to being detected as outliers since they can easily be visualized when disseminating statistical information through the use of maps. Outliers also arise on other (non-spatial) variables which can be categorical or continuous. In general, outliers arise from multivariate relationships between spatial and non-spatial characteristics.

In this paper, we present a small evaluation study which focuses on sample uniques arising from rare and unique attributes in the US American Community Survey (ACS) transportation outputs. This could be extended to other outputs based on migration patterns. The spatial variables are trajectories defined as vectors of coordinates where the first component is the coordinate of place of residence (origin) and the second component is the coordinate of workplace (destination). An example of a rare and unique attribute in the ACS transportation outputs are overly long commutes to work on a non-typical means of transportation (MOT), such as cycling. The outlier can be caused either by the trajectory or by the mode of transport. The objective of the study

was to inform and guide decisions about best practices that could be used for future dissemination strategies on these and other similar types of datasets by the US Census Bureau.

The first stage of the study is to recommend a procedure for detecting outliers. The multivariate aspect of the survey data makes the task of identifying outliers challenging. Once outliers are identified, we propose a targeted disclosure limitation method to confidentialize the outliers. This implies that only pre-tabular methods of disclosure limitation that are applied directly on the microdata prior to tabulation will be considered. All resulting tabular outputs derived from confidentialized microdata are assumed protected against disclosure risks. The method needs to account for different types of variables: spatial, categorical and continuous. Constraints to consider when developing a pre-tabular perturbative method of disclosure limitation are: determining which variable(s) to perturb, preserving the logical consistency of the record after the perturbation and ensuring that the disclosure limitation method preserves the sufficient statistics of the dataset. Therefore, a disclosure limitation method specifically targeted for outliers, whether they are caused by spatial or non-spatial variables or both, is particularly challenging due to the need to ensure the statistical integrity of the data.

We describe the data for the evaluation study in Section 2 based on the Public-Use file of the American Community Survey in the US. The outlier detection algorithm is discussed in Section 3 and the perturbation methods that were used are described in Section 4. Section 5 contains results and a brief discussion is provided in Section 6.

2 Data

In the investigation of outlier detection strategies and perturbation of spatial outliers, we carry out a small-scale simulation study based on an artificial population produced from the 2006-2008 combined PUMS from the American Community Survey (ACS). From this file, we selected records based on those living in the State of California, were employed and did not work outside of the US. There were 438,850 individuals in the file. Latitude and longitude of place of residence and workplace were generated by adding random distances around a radius of the centroid of the individual's residence and workplace PUMAs (public-use microdata area) which is the lowest unit of geography on the PUMS and are formed to be greater than 100,000 in the population. The geographical coordinates of the centroid of each PUMA were calculated through the GIS systems at Westat, Inc. This effectively dispersed individuals across different areas of the PUMA for place of residence and place of work. It is important to remember that this is a test dataset and that the place of residence and place of work is a result of additive random noise.

The primary geographic variables for the place of work for the evaluation were the state, and the associated latitude and longitude workplace coordinates. Variables of interest for the detection of outliers and their disclosure protection specifically for workplace trajectories included economic and demographic variables, and variables of interest such as the mode of transport, distance travelled and travel time. We calculate a travel distance computed by the latitude and longitude coordinates. The distance is computed 'as the crow' flies as they say, and therefore is computed using a straight line, and does not account for traffic flow patterns, etc.

3 Outlier Detection

An outlier is defined in terms of its deviation or inconsistency from other units in the dataset. Ben-Gal (2005) and the references therein provide a good introduction on outlier detection methods. Outlier detection methods are divided between univariate and multivariate methods and also can take parametric or non-parametric forms. Parametric forms assume an underlying model and define outliers as those observations that deviate from the model. Non-parametric forms are typically based on the measurement of observations using local distance metrics. Univariate statistical methods for outlier detection are based on confidence intervals of the normal distribution where multiple testing corrections, e.g., Bonferroni method, should be used for simultaneous testing of several outliers at once. The sample mean and sample variance are influenced by the presence of outliers and can be replaced by more robust estimators such as the median and the median absolute deviation. In multivariate outlier detection, interactions of different variables are examined where variables may be continuous or categorical. Statistical methods for multivariate outlier detection mainly rely on the Mahalanobis Distance metric where large values indicate outliers. This metric can be replaced by robust estimates of the multidimensional distribution parameters, for example mean vector replaced by median vector and the covariance matrix by the minimum covariance determinant (MCD) (Rousseeuw, 1985).

The MCD estimator for a dataset $\{x_1,...,x_n\} \in \Re^p$ where each x_i is a vector of p variables is defined by that subset $\{x_{i1},...,x_{ih}\}$ of h observations whose covariance matrix has the smallest determinant among all possible subsets of size h. The value h can be thought of as the minimum number of points which must not be outlying. The MCD location and scatter estimate: $\mathbf{M_{MCD}}$ and $\mathbf{S_{MCD}}$ are the arithmetic mean and a multiple of the sample covariance matrix of that subset:

$$M_{MCD} = \frac{1}{h}\sum_{j=1}^{h} x_{ij} \quad \text{and} \quad S_{MCD} = \kappa_{ccf}\kappa_{sscf}\frac{1}{h-1}\sum_{j=1}^{h}(x_{ij}-M_{MCD})(x_{ij}-M_{MCD})^T.$$

The constant factor κ_{ccf} is the consistency correction factor and ensures that for multivariate normal data, the sample covariance estimator **S** is equal to $\mathbf{S_{MCD}}$. The constant factor κ_{sscf} is a small sample correction factor to ensure that **S** is unbiased for small samples (Todorov and Filzmoser, 2009). The recommended choice for h is [(n+p+1)/2] where [.] is the greatest integer function. The idea is that points that are outliers will not be involved in the location and shape calculations. If h=n then $\mathbf{M_{MCD}}$ and $\mathbf{S_{MCD}}$ reduce to the sample mean and sample covariance matrix. Rousseeuw and van Driessen, 1999 provide an algorithm for calculating the MCD.

The squared Mahalanobis distances based on p variables that follow a multivariate normal distribution with mean $\bar{x}$ and sample covariance matrix **S** follows a χ^2_p distribution. The usual cutoff value for testing outliers based on the Mahalanobis distances is a quantile of the χ^2 distribution, such as $D_0 = \chi^2_p(0.975)$. However, this cutoff will not be valid if robust estimators are used in the robust Mahalanobis distances. In this case, Maronna and Zamar (2002) propose to use a transformation of the cutoff value as follows: $D_0 = \chi^2_p(0.975)\frac{median\{RD_1,..,RD_n\}}{\chi^2_p(0.5)}$ where RD_i is the robust distance for record i.

We compared the Mahalanobis distances and robust Mahalanobis distances to some standard methods for outlier detection used in selective statistical data editing. From this study, it was clear that standard methods of selective editing often assume normal distributions and generally perform best in the case of univariate distributions. It was clear that Mahalanobis distances work best to detect outlier trajectories to work by taking into account the spatial characteristics of the multivariate data. To calculate the Mahalanobis distances we use two continuous variables: minutes to work and distance travelled. Due to the large number of outliers, the multivariate distribution of these variables deviated from the Normal assumption and therefore we determined that the robust Mahalanobis distances is the preferred option for spatial outlier detection.

After calculating the robust Mahalanobis distances based on minutes to work and distance travelled, we carried out a large number of robust regression analyses to determine which explanatory variables are highly predictive of the distance travelled. The dependent variable for the regression analysis was the robust Mahalanobis distances and the independent variables were categorical variables transformed into dummy variables. The aim was to determine homogenous sub-groups within which we would calculate the robust Mahalanobis distances to test for outliers. The independent variables that were highly associated with the robust Mahalanobis distances were: mode of transport, sex, earnings and occupation. The sub-groups were collapsed to ensure that a minimum of 20 individuals were present in each sub-group. We deleted all individuals from the dataset where there was a missing value in either minutes to work or distance to work for this study.

A SAS program was written that includes a pre-developed SAS macro called 'Robcov' Version 1.3-2 (20 March 2007)[1] that was written by Michael Friendly. The program calculates the robust Mahalanobis distances within sub-groups based on the minimum covariance determinant (MCD) estimator. The program first calculated the cross- classification of the variables defining the sub-groups and then proceeds to collapse neighboring sub-groups to ensure that there are at least 20 individuals in each sub-group. The program then labels the final sub-groups sequentially 1, 2, 3, etc. The macro works through each sub-group separately and calculates the Mahalanobis distances and robust Mahalanobis distances. A flag is assigned to each record, giving a value of 1 if the robust Mahalanobis Distance squared is larger than the critical value, otherwise it has a value of 0.

4 Perturbation Methods

Once outliers have been detected in the survey microdata, they need to be targeted for a pre-tabular disclosure limitation method. Pre-tabular methods include some or all of the following: coarsening and suppression of variables, adding noise to continuous variables, misclassification of categorical variables, parametric or non-parametric delete-impute methods. These methods are well documented (see for example, Rubin, 1993, Willenborg and DeWaal, 2001, Reiter, 2005 Oganian and Karr, 2006, Shlomo, 2007). We therefore focus on pre-tabular perturbative methods that can be specialized and adapted to the protection of spatial coordinates.

[1] http://euclid.psych.yorku.ca/SCS/sasmac/robcov.html (current as of February 26, 2013)

Wang and Reiter (2011) focus on the confidentiality protection of a geographical variable defined by latitude and longitude. They propose the perturbation of the latitude and longitude by replacing original values with imputations based on random draws from a posterior predictive distribution using multiple imputation techniques. Domingo-Ferrer and Trujillo (2011) discuss the confidentiality protection of trajectories that have a time-stamp using clustering k-anonymity techniques and data swapping. Krenzke et al. (2011) discuss the protection of origin-destination tables from the American Community Survey planned transportation products through the perturbation of attributes. They compare disclosure limitation methods under the delete-impute approach using a semi-parametric model-assisted method, a parametric model-based method and a constrained hot-deck method.

In this study, there are several objectives to the perturbation process that have been developed and evaluated. A general goal of the perturbation process was to change the residence of the person, while retaining the workplace. The reason for not changing the workplace was that if a workplace is changed, the new area may not offer the industry or occupation of the individual. In addition, we attempted to retain the relationship between distance and length of commute (minutes).

We calculated the distance to work using the following SAS statement where POW denotes place of work:

DistanceToWork = **geodist** (latitude,longitude,POW_latitude,POW_longitude,'DM');

To reduce the size of the test dataset, we included only those non-outliers that had a high level of consistency between the distance travelled to work and the minutes to work, resulting in 283,423 records. We then ran the outlier detection program (see Section 3) and based on the robust Mahalanobis distance algorithm identified 60,007 outliers (21.2%). From this set of outliers, we removed outliers with a mode of transport of 'other' resulting in 59,080 (20.8%) outliers.

4.1 The Coherence Function

The coherence function quantifies the level of consistency and plausibility for an individual trajectory from place of residence to place of work. The function takes into account the mode of transport, the distance travelled to work and the number of minutes of travel. As an example, an individual who traveled 100 miles in 2 hours by car should have a large coherence function, while another individual who traveled 50 miles in 2 hours by walking would have a small coherence function. In other words, the coherence function aims to quantify the plausibility of the observed values when taken in a multivariate setting.

To calculate the coherence function, we first define the maximum and minimum velocity for each mode of transport based on the set of non-outliers in the dataset after the determining outliers (see Section 3). We calculate for all non-outliers having the same mode of transport, the number of miles travelled divided by the number of minutes to travel, i.e. the number of miles per minute, and multiply this value by 60 to obtain a standard velocity measure to work of miles per hour. We then find the maximum/minimum computed velocity and define this as the maximum/minimum coherent velocity for that mode of transport. The formula for the velocities and the mean velocity at their initialization are:

$$maxVelocity_m = \max\left\{\frac{dist_r}{time_r} \times 60 \mid \forall r \in data_{nonOutliers}, r_{modeTransport} = m\right\}$$
$$minVelocity_m = \min\left\{\frac{dist_r}{time_r} \times 60 \mid \forall r \in data_{nonOutliers}, r_{modeTransport} = m\right\} \quad (1)$$
$$meanVelocity_m = \left\{\frac{maxVelocity_m + minVelocity_m}{2}\right\}$$

Once we obtained these velocities for each mode of transport, we calculated the coherence function for each of the individuals in the dataset by computing the deviation of the distance travelled from the distance travelled according to the mean velocity in (1), and compared that deviation to the deviation of the maximum distance travelled (based on the non-outliers) as calculated in the maximum velocity in (1) from the distance travelled according to the mean velocity in (1). Figure 1 shows the algorithm which outlines the calculation of the coherence function.

Algorithm 1 Coherence function algorithm

Input: d distance travelled, m minutes travelled, tr mode of transport.
Output: c coherence level.
$y_{test} \Leftarrow d$
$y_{max} \Leftarrow m * \frac{maxVelocity_{tr}}{60}$
$y_{mean} \Leftarrow m * \frac{mean(maxVelocity_{tr}, minVelocity_{tr})}{60}$
$\Delta_{test} \Leftarrow abs(y_{test} - y_{mean})$
$\Delta_{max} \Leftarrow abs(y_{max} - y_{mean})$
$c \Leftarrow 1 - (\frac{\Delta_{test}}{\Delta_{max}} * \alpha)$
if $c < 0$ **then**
  return 0
else
  return c
end if

Figure 1. Algorithm of the Coherence Function

The parameter α in the algorithm presented in the algorithm in Figure 1 is used to smooth the borders of the coherence function and introduce 'fuzziness' into the function. Otherwise, all outliers outside of the minimum and maximum velocities would have zero coherence and it would not be possible to determine whether it is improving as we maximize the objective function. A good practice would be to use α = 0.5 because, by doing that, we assign a value of 0.5 to the borders of the coherent area, allowing the outliers to have a value below 0.5, but different than 0. Figure 2 shows a graphical explanation.

The coherence function assigns high coherence to the individuals within the mode of transport whose travelled distance is close to the computed mean, and low coherence to the individuals whose travelled distance is far from this mean. Finally, this function will be used as an objective function to guide the perturbation algorithms where we aim to obtain a higher coherence for the outliers.

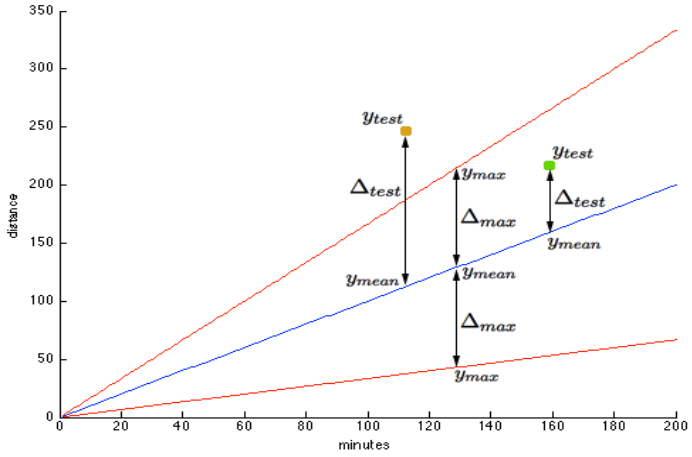

Figure 2. Graphical explanation of all terms involved in coherence function algorithm (Figure 1).

The upper, middle, and lower lines represent the max, mean, and min velocities respectively. In addition, the left point represents an outlier, and the right point represents a non-outlier.

4.2 Record Swapping Method

In this section, we present the algorithm used to swap place of residence for the outliers using the coherence as the objective function. Under this approach, we only use the outlier individuals identified in the dataset. We aim to pair outliers working in different workplaces in such a way that by swapping their place of residence, we increase the coherence function of at least one of the outliers and for the other outlier we either increase the coherence function or it remains the same. We carry out the algorithm iteratively, continuing to improve the overall coherence function in the data until it stabilizes.

The swapping algorithm is carried out separately within classes defined by the mode of transport, sex and age group. Any pair of outliers selected for swapping their place of residence will match on the three fixed control variables (mode, sex, and age group). Within each class, we split the outliers into groups defined by their workplace. We pair each outlier to all the other outliers in different workplaces and calculate their coherence function based on swapping the place of residence. If one of the two outliers receives more coherence than it already had in the previous stage and the other outlier does not have a decrease in its coherence, we swap the place of residence (i.e. the latitude and longitude of place of residence, PUMA and any other geographical information in the dataset). This process is carried out iteratively. As the number of iterations increases, the coherence also increases but at a lower rate for each iteration. The process stops when the improvement of the coherence is less than a pre-set threshold given as a parameter for the method.

Figure 3 presents a graphical explanation of the swapping algorithm. At the top of the figure, we present the initial state with two outliers, their corresponding workplaces, and the borders that mark the coherent area around each workplace. On the bottom half of the figure we present the final state after the swapped place of residences between the two outliers. As can be seen, the point x is not an outlier anymore, while point y is still an outlier but with a much better coherence. In the following iterations, the coherence can continue to improve until it reaches the

state of a non-outlier. The algorithm depends on a large pool of outliers since only the outliers are considered for swapping.

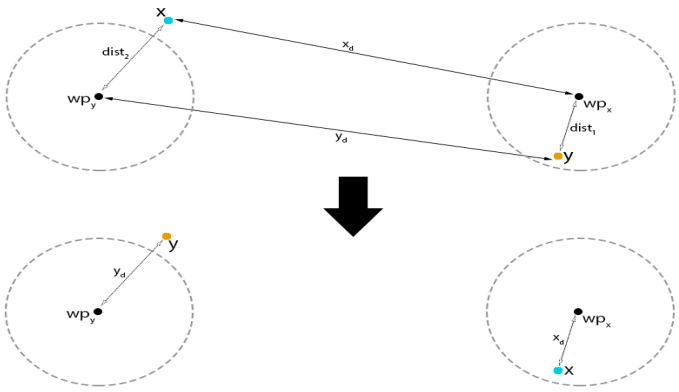

Figure 3. Graphical Explanation of the Record Swapping Algorithm

4.3 Hot Deck Method

The second proposed algorithm is based on imputing the place of residence (latitude, longitude, PUMA) of an outlier individual by the place of residence of a non-outlier that belongs to the same class. As defined for the swapping method, classes are created by mode of transport, sex and age group.

Figure 4 shows a graphical explanation of this method. Whilst for the swapping algorithm we were searching for a donor from other workplaces, in this case we focus only on a single workplace. We select a donor that is a non-outlier within the same class and having the same workplace as the outlier. We then impute the place of residence of the outlier with the donor's place of residence.

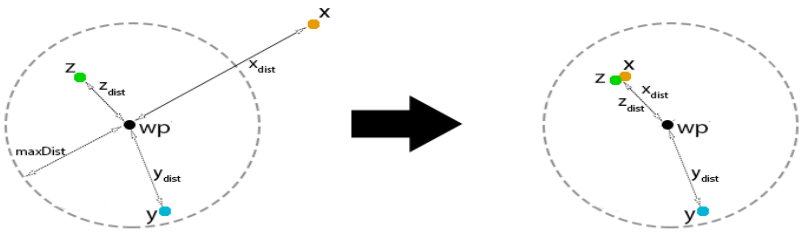

Figure 4. Graphical Explanation of the Hot Deck Algorithm

The key point of this method is the donor selection criteria. We propose two different approaches for selecting the donor. In the first approach we consider candidate donors whose distance to work are within the coherence range of distances given by the coherence function. We then select the donor that maximizes the coherence, i.e. the candidate donor whose distance to work is closer to the mean of the given outlier's coherent distances range taking into account the outlier's travelled minutes. This hot deck approach can only be applied to workplaces with more than one individual. If a workplace only has one individual in a given class it is not possible to find a donor in the same workplace. In that case, it is ignored and is marked as an isolated individual. Following the same reasoning, it can be deduced that the higher number of individuals in a workplace, the more likely it is to find good donors for the outliers. In the second approach we

do not use the coherence function. Instead, the donor is selected as the non-outlier in the same workplace having the most similar travelled minutes to work to the outlier's minutes to work, i.e. a nearest-neighbor approach. Using this approach, we allow outliers to receive place of residence where they have a distance to work that may still not be coherent for them but will improve their coherence by controlling for the minutes to work. Therefore, this approach is less strict than the first approach while improving the overall coherence of the dataset.

5 Results

We first look at the remaining disclosure risk in the dataset following the methods of perturbation and the changing of place of residence by re-assessing the number of outliers in the dataset. It can be seen that the swapping method was able to correct fewer outliers than the hot deck methods. This is expected since it is very difficult to correct a large number of outliers based solely on the sub-set of outliers. However, as will be seen, the swapping method had the lowest information loss. The two approaches of the hot deck method are both able to correct approximately twice as many outliers compared to the swapping method. This is due to the fact that we are using non-outliers' data to correct the outliers and there is a much larger pool of donors. The last thing to notice is that in all cases of perturbation (swapping and hot-deck) there are non-outliers that became outliers. The outlier detection algorithm is based on the distribution of individuals within sub-groups (see Section 3). It is possible that based on the new set of place of residence following the perturbation, the distributions were changed in such a way as to be less consistent compared to the original distributions and this produced new outliers. However, the number of non-outliers that were defined as outliers following the perturbation was much less than those outliers that were corrected to non-outliers.

Table 1 presents the results of the outlier detection algorithm before and after each of the methods of perturbation. The percent of the remaining outliers after the perturbation was 78.7% under the record swapping, 60.0% under the hot-deck and coherence function and 60.2% under the hot-deck and nearest neighbor of minutes to work.

We next look at some utility measures for the perturbation methods. The number of individuals who had their PUMA changed in the dataset due to the perturbation were: Swapping Method: 56,562 individuals; Hot Deck Method (Minutes): 53,945 individuals; Hot Deck Method (Coherent): 53,181 individuals. We cross the PUMA geography (which is the lowest geography available in the PUMS) with other variables to assess bivariate counts before and after perturbation. We calculate two distance metrics where $freq_{ij}$ is the original count in row i and column j and $freq'_{ij}$ is the perturbed count:

- Absolute Distance Normalized by 2N: $dist(attr_1, attr_2) = \sum_{\substack{i \in Dom(attr_1) \\ j \in Dom(attr_2)}} | freq_{ij} - freq'_{ij} | / 2N$
- Hellinger's Distance Normalized by SQRT(N):
$dist(attr_1, attr_2) = \sqrt{.5 \sum_{\substack{i \in Dom(attr_1) \\ j \in Dom(attr_2)}} (\sqrt{freq_{ij}} - \sqrt{freq'_{ij}})^2} / \sqrt{N}$

Table 1. Number of Outliers Before and After Perturbation Methods

Original Outliers	Total	Outliers after Swapping		Outliers after HD (Coherence Measure)		Outliers after HD (Minutes)	
		Yes	No	Yes	No	Yes	No
Yes	59,080 (20.9%)	42,788 (92.0%)	16,292 (6.9%)	27,099 (76.2%)	31,981 (12.9%)	28,123 (79.3%)	30,957 (12.5%)
No	224,,343 (79.2%)	3,731 (8.0%)	220,612 (93.1%)	8,456 (23.8%)	215,887 (87.1%)	7,321 (20.7%)	217,022 (87.5%)
Total	283,423 (100%)	46,519 (100%)	236,904 (100%)	35,555 (100%)	247,868 (100%)	35,444 (100%)	247,979 (100%)

Bivariate counts were crossed with the following variables: AGE9 {recoded: young; middle; elderly} AGEP {individual ages top coded at 99}, SEX {male; female}, OCCUPATION {recoded: managers; professionals; salaried employees; others}, EARNINGS {recoded: low; medium; high; highest}, MODE OF TRANSPORT {recoded: car, motorcycle; bus, taxi, ferryboat; subway, train; cycle; walk; home, other}. Table 2 provides the results of the distance metrics. It can easily be seen that the hot deck methods perturb the bivariate counts more than the swapping method. This is also true when crossing PUMA with those variables that were used as the control variables (SEX, AGE9, MODE). This happens because in the swapping algorithm we are not changing the PUMA distribution of values inside a class. We are only changing the position of the values but the marginal frequency for each PUMA value is always the same. However, in the case of the hot deck method these marginal distributions are actually changed because we are copying the values of the donors and deleting the values of the receivers. These changes on the marginals impact on the bivariate frequencies for all variables, including the control variables.

Table 2. Distance Metrics under the Swapping Method, the Hot-Deck Method Based on the Coherence Function and the Hot-Deck Method based on nearest Neighbor of Minutes

Bivariate Variables Crossed with PUMA	Normalized Absolute Difference			Normalized Hellinger's Distance		
	Swapping	HD Minutes	HD Coherence	Swapping	HD Minutes	HD Coherence
AGE9	0	0.109	0.095	0	0.154	0.134
AGEP	0.048	0.119	0.107	0.059	0.166	0.147
SEX	0	0.113	0.095	0	0.161	0.140
OCCUPATION	0.094	0.134	0.125	0.164	0.215	0.203
EARNINGS	0.024	0.104	0.089	0.029	0.148	0.129
MODE	0	0.104	0.087	0	0.154	0.131

6 Conclusion

Based on the small evaluation study, the results are inconclusive. It is clear that record swapping provided the lowest information loss, especially with respect to the preservation of bivariate counts of the swapping variable and other control variables. However, the record swapping method only corrected 21.3% of the outliers in the dataset, while the hot-deck methods corrected approximately 40.0% of the outliers. On the other hand, the hot-deck method transformed more non-outliers to outliers than the record swapping. Record swapping provided higher utility in

terms of the distance between original and perturbed bivariate counts, especially for those counts that were formed from the control variables used to define the strata within which the perturbation was carried out. The recommendation would be to carry out both methods, starting with the record swapping and then proceeding to the hot-deck method on the remaining outliers. Ultimately, it would be up to the policy makers to determine the thresholds as to how much perturbation should be carried out in the data.

In this evaluation study we did not take into account the survey weights since these would have to be recalibrated in any case due to the change in place of residence. It is clear however that using the calibration variables as control variables, such as age group and sex, we are able to preserve the original sampling weights better under the record swapping method compared to the hot-deck method.

References

Ben-Gal, I. (2005). Outlier Detection in Maimon, O. and Rockach, L. (eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers.

Domingo-Ferrer, J. and Trujillo, R. (2011). Anonymization of Trajectory Data. UNECE/Eurostat work session on Statistical Data Confidentiality. October, 2011. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/32_Domingo-Trujillo.pdf

Krenzke, T., Li, J., Judkins, D. and Larsen, M. (2011). Evaluating a constrained hotdeck to perturb American Community Survey Data for the Census Transportation Planning Products. Proceedings of the Section on Government Statistics of the Joint Statistical Meetings.

Maronna, R.A. and Zamar, R.H. (2002). Robust Estimation of Location and Dispersion for High-Dimensional Datasets. *Technometrics,* vol. 44, no. 4, 307-317.

Oganian, A. and Karr, A. (2006). Combinations of SDC Methods for Micro-data Protection. Privacy in Statistical Databases-PSD2006 (eds. J. Domingo-Ferrer and L. Franconi), *Lecture Notes in Computer Science*, vol. 4302, pp. 102-113.

Reiter, J.P. (2005). Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society*, A, Vol.168, No.1, pp. 185-205.

Rousseeuw, P.J. (1985). Multivariate Estimation with High Breakdown Point. In W. Grossman, G. Pflug, I. Vincze, W. Wertz (eds.) *Mathematical Statistics and Applications*, Vol B. 283-297. Reidel Publishing, Detroit.

Rousseeuw, P.J. and Van Driessen, K. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, Vol. 41, No. 3, 212-223.

Rubin, D.B. (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-imputed Microdata. *Journal of Official Statistics*, 91, pp. 461-468.

Shlomo, N. (2007). Statistical Disclosure Control Methods for Census Frequency Tables. *International Statistical Review*, Vol. 75, No. 2 pp. 199-217.

Todorov, V. and Fizmoser, P. (2009). An Object-Oriented Framework for Robust Multivariate Analysis. *Journal of Statistical Software*, Vol. 32, Issue 3.

Wang, H. and Reiter, J. P. (2011). Multiple Imputation for Sharing Precise Geographies in Public Use Data. *Annals of Applied Statistics*.

Willenborg, L. and De Waal, T. (2001). Elements of Statistical Disclosure Control in Practice. *Lecture Notes in Statistics*, 155. New York: Springer-Verlag.