HHS Public Access

Author manuscript

Priv Stat Databases. Author manuscript; available in PMC 2020 March 23.

Published in final edited form as:

Priv Stat Databases. 2016 ; 6896: 69–80. doi:10.1007/978-3-319-45381-1_6.

Propensity score based conditional group swapping for disclosure limitation of strata-defining variables

Anna Oganian[1,2], **Goran Lesaja**[2]

Anna Oganian: aoganyan@cdc.gov; Goran Lesaja: goran@georgiasouthern.edu

[1]National Center for Health Statistics, 3311 Toledo Rd, Hyatsville, MD, 20782, U.S.A.

[2]Georgia Southern University, Department of Mathematical Sciences, P.O. Box 8093, Statesboro, GA 30460-8093, U.S.A.

Abstract

In this paper we propose a method for statistical disclosure limitation of categorical variables that we call Conditional Group Swapping. This approach is suitable for design and strata-defining variables, the cross-classification of which leads to the formation of important groups or subpopulations. These groups are considered important because from the point of view of data analysis it is desirable to preserve analytical characteristics within them. In general data swapping can be quite distorting ([12, 18, 15]), especially for the relationships between the variables not only within the subpopulations but for the overall data. To reduce the damage incurred by swapping, we propose to choose the records for swapping using conditional probabilities which depend on the characteristics of the exchanged records. In particular, our approach exploits the results of propensity scores methodology for the computation of swapping probabilities. The experimental results presented in the paper show good utility properties of the method.

Keywords

Statistical disclosure limitation (SDL); group swapping; strata; subpopulations; propensity scores

1 Introduction

Statistical agencies have an obligation by law to protect privacy and confidentiality of data subjects while preserving important analytical features in the data they provide. Privacy and confidentiality are not guaranteed by removal of direct identifiers, such as names, addresses and social security numbers, from the microdata file. Re-identification of individuals in the data is still possible by linking the file without direct identifiers to external databases. That is why in addition to the removal of direct identifiers, released microdata are typically modified, in order to make disclosure more difficult; that is, statistical disclosure limitation (SDL) methods are applied to the data prior to their release. The goal of such a modification is two-fold: to reduce the risk of re-identification and at the same time to preserve important distributional properties of the original microdata file. Although it is not possible to know all the uses of the data beforehand, some of the relationships of interest to the user may be known. For example, some surveys oversample particular groups of individuals with the goal of obtaining better estimates for these groups. This requires special sample design and

allocation of additional funds to obtain bigger samples for these groups. It would be particularly undesirable and counterproductive if SDL methods significantly change the estimates within these groups and/or considerably increase their standard errors. So every scenario of data release is different and disclosure limitation methods should be chosen accordingly. In this paper, we have focused on the situation of data release when the data protector has to modify categorical variables that define strata or subpopulations, but at the same time wants to minimize the distortion to the analytical structure within these strata.

To accomplish his/her task, the data protector can choose from among a wide variety of methods which can be divided in two groups: masking methods which release a modified version of the original microdata, and synthetic methods which generate synthetic records or values for specific variables from the distribution representing the original data.

A few examples of masking methods are: additive or multiplicative noise [1, 13, 19, 18, 14], in which noise is applied to numerical data values to reduce the likelihood of exact matching on key variables or to distort the values of sensitive variables; microaggregation, a technique similar to data binning (see [8, 9, 3, 24]) and data swapping [2], in which data values are swapped for selected records. There are many variants of swapping, some examples are [2, 20, 16, 7, 22]. Data swapping is popular among government agencies since it preserves marginal distributions, and it is often implemented as a simple random swapping [7] in which a prespecified percentage of randomly selected records is swapped with some other randomly selected records for specific variables.

To measure the utility of masked data, the data protector can use either analysis-specific utility measures, tailored to specific analyses, or broad measures reflecting global differences between the distributions of original and the masked data [12, 26, 17]. One example of an analysis-specific measure tailored for regression analysis is an overlap in the confidence intervals for the regression coefficients estimated with the original and masked data [12]. An example of broad measure is the propensity score measure proposed in [26]. It compares favorably with others and it is suitable for data sets with mixed attributes [4, 26]. Below we will review this measure in more detail because it is used as a part of our Conditional Group Swapping method described in Section 2.

Propensity score measure

First, let us recall the definition of a propensity score. The propensity score is the probability that an observation i is assigned to a particular group, call it a treatment group, given covariate values $\boldsymbol{x}_i$. We denote $T = 1$ if a record is assigned to a treatment group and $T = 0$ otherwise. As [21] shows, T and $\boldsymbol{x}$ are conditionally independent given the propensity score. Thus, when two groups have the same distributions of propensity scores, the groups should have similar distributions of $\boldsymbol{x}$. This theory was used in [26] to measure data utility of disclosure protected data. In particular, [26] suggested the following approach. First, merge (by "stacking") the records from groups A and B that are being compared in their distributions. Then add an indicator variable T that equals one for all records from B and zero otherwise. Secondly, for each record i in the merged set, compute the propensity score, that is the probability of being in B given $\boldsymbol{x}_i$ - the values of the variables for this record. Propensity scores can be estimated via a logistic regression of the variable T on functions of

all variables $\boldsymbol{x}$ in the data set. Thirdly, compare the distributions of the propensity scores in groups A and B. If the propensity scores are approximately the same for all the records in groups A and B, then the the distributions of $\boldsymbol{x}$ in these groups are approximately the same This is an implication of the conditional independence of T and $\boldsymbol{x}_i$ given the propensity score (see [21] and [26]). The propensity score distance measure proposed in [26] is

$$U_p = \frac{1}{N}\sum_{i=1}^{N}[\hat{p}_i - c]^2 \tag{1}$$

where N is the total number of records in the merged data set, $\hat{p}_i$ is the estimated propensity score for unit i, and c equals the proportion of B units in the merged data set.

1.1 Contribution and plan of this paper

In this paper, we have focused on a non-synthetic approach for disclosure limitation suitable for categorical strata-defining variables, the cross-classification of which leads to the formation of important groups for a data analyst. We present the Conditional Group Swapping method designed to minimize the distortion incurred by swapping, to the relationships between the variables, particularly those that involve categorical strata-defining variables. The idea of the method is described in Section 2. The results of the numerical experiments are reported in Section 3. Section 4 provides a concluding discussion and sketches lines for future work.

2 Propensity score based conditional group swapping

In this section we describe the algorithm of our Conditional Group Swapping approach, hereafter, abbreviated as CGS. Below are the main steps of the method.

1. Compute pairwise distances between all the strata using the propensity score metric (1) described in Section 1. Note that the interpretation of the absolute value of this metric is not relevant here. The goal is to identify the pairs of the closest strata.

2. Compute swapping probabilities, that is the probabilities of moving records from one stratum to another, for the records in two closest strata. This will be done as follows. Suppose the distance between stratum A and stratum B is the smallest among all pairwise distances. Let n_s be the desired swapping rate, that is the number of records that will be moved from one stratum to another. To compute the swapping probabilities, first combine together all the records from A and B (by "stacking") and add an indicator variable T, $T = 1$ for all the records from stratum B and 0 for all the records from stratum A. Next, for every record i in the combined set compute the probability that this record is assigned to stratum B given the values of the variables for this record, $\boldsymbol{x}_i$. In other words we compute the propensity scores, denoting them as $P_{AB}(I \rightarrow B|\boldsymbol{x}_i)$.

3. Select n_s records from stratum A with the probabilities proportional to their propensity scores $P_{AB}(i \rightarrow B|\boldsymbol{x}_i)$ and change their stratum indicator to B. For example, if in stratum A there are residential hospital records and in stratum B -

multi-service hospitals, then for the selected n_s residential hospitals we will change their hospital type indicator to multi-service.

4. Select n_s records from stratum B with probabilities proportional to $1 - P_{AB}(i \to B|\mathbf{x}_i)$ and "move" them to stratum A. The records that arrived from stratum A on the previous step will be excluded from the selection.

5. Repeat steps 3 and 4 for another pair of strata with the next closest distance.

6. Repeat step 5 until there are no strata that have not been swapped.

3 Numerical experiments

The procedure described above was implemented and evaluated on several data sets. We experimented with genuine and simulated data. In this section we present only the results obtained on two genuine data sets. Simulated data results were very similar, so we omit them for brevity of the exposition. Below is the description of the two genuine data sets we used.

- The Titanic data is a public data set that was obtained from the Kaggle web-site [11]. This is a collection of records of 889 passengers of the Titanic, the British passenger liner that sank in the North Atlantic Ocean on April 15th 1912. The variables in this data set are: Survived - survival status (0=No; 1=Yes), Pclass - passenger class (1=1st; 2=2nd; 3=3rd), Sex - sex, Age - age in years, SibSp - number of siblings/spouses aboard, Parch - number of parents/children aboard, Fare - passenger fare, Embarked - port of embarkation (C= Cherbourg; Q=Queenstown; S= Southampton). The original file from Kaggle also contained names of the passengers, their ticket numbers and cabin number. These variables are irrelevant for our analysis, so they were excluded.

- The 1998 Survey of Mental Health Organizations (abbreviated as SMHO). This sample contains 874 hospitals. It is publicly available and can be obtained from the PracTools R package [25]. The 1998 SMHO was conducted by the US Substance Abuse and Mental Health Services Administration, which collected data on mental health care organizations and general hospitals that provide mental health care services. The goal of the survey was to provide estimates for total expenditure, full-time equivalent staff, bed count, and total scaled by type of organization. For this data it is desirable to preserve as much as possible the estimates of these variables within the strata defined by the type of hospital. There are five types of categories for the variable hosp.type: 1) Psychiatric , 2) Residential/Veterans hospitals, 3) General, 4) Outpatient/Partial care and 5) Multi-service/Substance abuse. Other variables in the data are: Exptotal - total expenditures in 1998, Beds - total inpatient beds, Seencnt - unduplicated client/ patient count seen during year, Eoycnt - end of year count of patients on the role, Findirct - money hospital receives from the statement health agency (1=yes; 2=no).

We applied the approach described in Section 2 to these data sets. For the Titanic data one of the relevant analyses is to check what sorts of people were likely to survive. In fact, since the sinking of the Titanic, there has been a widespread belief that the social norm of women and

children first gives women a survival advantage over men in maritime disasters, and that captains and crew members give priority to passengers. However, [5] presented an interesting study of historical records, spanning over three centuries, that suggests that in maritime disasters women and children die at significantly higher rates than male passengers and crew members. Their findings suggest that the events on the Titanic, where 20 percent of men and 70 percent of women and children survived, were highly unusual, if not unique. Besides gender, the class of the Titanic passengers was also related to their survival status.

Based on these considerations, we divided the data in six strata according to the cross-classification of the variables Pclass and Sex: 1) first class male passengers, 2) first class females, 3) second class males, 4) second class females, 5) third class males, 6) third class females.

The first step of the CGS procedure identified the following strata as closest: first class males and first class female, second class males and second class females, and third class males and third class females. For the measure of distance between the distributions of different strata (specifically, between the multivariate distributions of Survived, Age, Fare, SibSp and Parch for each stratum), we used the following model to estimate propensity scores: the main effects for the variables Survived, Age, Fare, SibSp and Parch and the interactions between Survived and Fare, Survived and Age, Survived and SibSp, Survived and Parch. We didn't include all the main terms and interactions because otherwise the totality of the estimated parameters would not be supported by the sample size.

Because the goal of our experiments is to test the potential benefits of using conditional probabilities for swapping and more specifically to estimate the effect of such probabilities on the quality of different statistical estimates, we compared the outcome of Conditional Group Swapping to the outcome of a similar approach which is characterized by uniform swapping probabilities. For the later approach the values of the variables of the records do not influence the probabilities of these records being swapped. We call it Random Group Swapping, hereafter, abbreviated as RGS. In a sense, RGS reflects the idea of the traditional approach for swapping. To make a fair comparison and to estimate the effect of using conditional swapping probabilities, RGS and CGS were implemented in the same way (as described in Section 2), except for the way how the probabilities of swapping were computed: for CGS they were proportional to the propensity scores, as described in Section 2, but for RGS they were uniform as we mentioned above.

We experimented with two swapping rates: $n_s = 20$ and $n_s = 40$ records exchanged between the strata. This corresponds respectively to about 15 and 35 percent of records swapped for each stratum. For each swapping rate, we generated 100 realizations of swapped data using Random and Conditional Groups Swapping.

Next, we compared the results of several statistical analyses based on the original and swapped data. One of them was logistic regression fitted to the complete Titanic data with Survived as the predicted variable and Pclass, Sex and Age as predictors. Hereafter, we will use R notation for the models. For the aforementioned regression it will be: Survived ~

Pclass+Sex+Age. Denote this model Reg1. We used this set of predictors in Reg1 because they were identified as being statistically significant based on the original data.

We also fitted logistic regressions within each stratum: Survived ~ Age + Fare. Denote this model Reg2

Next, we compared confidence intervals of regression coefficients for these regressions based on the original and swapped data. There were five regression coefficients for Reg1, including intercept, coefficient for Age, coefficients for dummy variables Pclass=2, Pclass=3 and for Sex=male and three regression coefficients for Reg2 (intercept and coefficients for Age and Fare).

As a measure of comparison we used the relative confidence interval overlap similar to the one used in [12]. Let $(L_{\text{orig},k}, U_{\text{orig},k})$ and $(L_{\text{swap},k}, U_{\text{swap},k})$ be the lower and upper bounds for the original and masked confidence intervals for the coefficient k. Let $L_{\text{over},k} = \max(L_{\text{orig},k}, L_{\text{swap},k})$ and $U_{\text{over},k} = \min(U_{\text{orig},k}, U_{\text{swap},k})$. When the original and masked confidence intervals overlap, $L_{\text{over},k} < U_{\text{over},k}$ and $(L_{\text{over},k}, U_{\text{over},k})$ represent the lower and the upper bounds of the overlapping region. When these confidence intervals do not overlap, $L_{\text{over},k} > U_{\text{over},k}$ and $(L_{\text{over},k}, U_{\text{over},k})$ represent the upper and the lower bounds of the non-overlapping region between these intervals. The measure of relative confidence interval overlap for the coefficient k is defined as follows:

$$J_k = \frac{1}{2}\left[\frac{U_{\text{over},k} - L_{\text{over},k}}{U_{\text{orig},k} - L_{\text{orig},k}} + \frac{U_{\text{over},k} - L_{\text{over},k}}{U_{\text{swap},k} - L_{\text{swap},k}}\right] \quad (2)$$

When confidence intervals overlap, $J_k \in (0, 1]$ and $J_k = 1$ when the intervals exactly coincide. In case one of the confidence intervals is “contained” in the other, the relative confidence interval measure will capture such a discrepancy, and $0 < J_k < 1$. When intervals don’t overlap, $J_k \leq 0$. In this case, J_k measures non-overlapping area (between the intervals) relative to their lengths. We also report an average confidence interval overlap over all the coefficients defined as $J = (1/p)\sum_{i=1}^{p} J_k$.

Table 1 presents the results of the experiments. The first column of the table is the type of analysis, Reg1 or Reg2, for which the outcome is compared between the original and masked data. The second column is the swapping rate. Columns “Average” and “Range” display average confidence interval overlap J and the range of variation of individual confidence interval overlaps J_k over all 100 realizations and all the coefficients. Range of variation is reported for the central 90% of the distribution of J_k. Column “# non-over” displays the fraction of times the intervals didn’t overlap over all the realizations and coefficients. For example, 100⁄500 means that 100 out of 500 intervals didn’t overlap (i.e. the number of times $J_k < 0$). For Reg1 the number of computed intervals is 500 = 100 realizations × 5 coefficients; and for Reg2 it is 1800 = 3 coefficients × 6 strata × 100 realizations

As can be seen from the table, the average confidence interval overlap J for Reg1 are relatively high for CGS (0.88 and 0.65 for $n_s = 20$ and 40 respectively). Moreover, the values

of J are considerably higher for CGS than for RGS. The range of variation is also narrower for CGS. The lower bounds for the range of variation correspond to the worst cases of the confidence interval overlaps. These smallest overlaps are still quite larger for CGS than for RGS. The upper bounds of the range of variation are similar for both methods, although still slightly larger for the CGS.

Regarding individual coefficients overlap measures J_k, we observed that they were similar in values for different coefficients, except the coefficient for Sex. In particular, the average J_k values over 100 realizations were smaller for Sex than for other coefficients (it was equal to 0.5 for CGS). Confidence intervals for Sex overlapped for all 100 realizations of CGS for the swapping rate 20. However, confidence intervals for Sex never overlapped for RGS. There is an explanation to that. In particular, in both cases swapping was done between the strata which were identified as closest to each other. The closeness was estimated for the multivariate distribution of Survived, Age, Fare, SibSp and Parch. The closest strata happened to be the ones that have the same passenger class Pclass but different Sex, *e.g.*, 1st class male and 1st class females, and so on. So, it was Sex that was actually swapped for the selected records. The selection probabilities of RGS are independent of the values of the variables, so it is not surprising that the relationships between Sex and other variables, and in particular Sex and survival status, are particularly affected. On the other hand, when swapping probabilities are proportional to the propensity scores, as in the CGS method, the relationship between Sex and survival status is taken into account (through propensity scores), so the swapped and original data confidence intervals for Sex are much more similar.

In addition to confidence interval comparisons, we also computed the element-wise ratios of original and swapped data means and covariance matrices for numerical variables Age, Fare, SibSp and Parch within each stratum. The results of these comparisons are presented in Table 2. We can see that the ratios of original and masked means are very similar for CGS and RGS. The range of variation is, however, larger for Random Swapping, which is an indicator of larger disturbances introduced by Random Swapping. In column “# sign change” we display the fraction of times an element in the covariance matrix changed in sign. These changes occurred predominately for the variables with covariances close to zero. These sign changes happened more often for RGS than for CGS.

For our second data set, SMHO, we fitted a logistic regression of Find-irct (hospital receives money from the statement health agency) on all other variables, denote it Reg3 and a regression of Exptotal (total expenditures in 1998) on all other variables, denote it Reg4. Both regressions were fitted to the complete data. Within strata, analyses included regressions: Findirct on all other variables (Reg5) and Exptotal on all other variables (Reg6). Hospital type was not included in the predictor set of Reg5 or Reg6 because it was the same value for all the records in a particular stratum. The results are presented in Table 3. Just as with the Titanic data, we also computed mean and covariance matrices ratios based on the original and swapped data within each stratum. These comparisons are presented in Table 4.

As can be seen from Tables 3 and 4, original and swapped confidence intervals overlap at a higher level for CGS than for RGS, and discrepancies in means and covariance matrices are

smaller for CGS. Interestingly, when fitting logistic regression Reg5 within strata, we noticed that in 3 out of 100 realizations of swapped data using CGS the regression coefficient for variable Bed was not estimable within the Outpatient stratum. A closer examination of the swapping results of these three realizations showed that the values for the variable Bed in stratum Outpatient were predominately zeros in the original data. So, most of the times CGS led to the selection of those records that had non-zero values for the variable Bed to be moved to another stratum. In those cases the Outpatient stratum received records with low counts for Bed from another stratum, but for those exceptional 3 cases, the incoming records had all zeros for Bed, resulting in a non-estimable coefficient. However, with the exception of those three cases, when CGS was used, the original and masked confidence intervals for Bed had larger overlap and the means and covariance matrix were better preserved for stratum Outpatient. In fact, CGS led to the choice of records with low counts for Bed which fits well the description of the Outpatient hospital stratum, while RGS on several occasions moved records with large values for Beds, which is inconsistent for Outpatient stratum.

4 Concluding discussion and future work

In this paper we presented a Conditional Group Swapping method suitable for categorical variables which define strata or subpopulations. This swapping method is designed with the goal to reduce the damage incurred by the disclosure limitation to the relationships between the variables within the strata and in the overall data. Our experimental results showed that the method has the potential to better preserve inferential properties, such as confidence intervals for the regression coefficients specific to particular strata and for the overall data, than Random Swapping. For numerical variables the means and covariance matrices within the strata are less distorted as well.

We believe that in practice CGS should not be the only method that is applied to the data, especially if there are continuous variables in the data. Similar to other swapping approaches, CGS can be used together with other SDL methods. For example, one can apply Conditional Group Swapping to strata-defining variables and then add multivariate noise **E** to the continuous variables within each stratum s with strata-specific parameters:

$$\mathbf{X_m^s} = \mathbf{X_o^s} + \mathbf{E^s} \tag{3}$$

where $\mathbf{X_o^s}$ and $\mathbf{X_m^s}$ are the original and masked (continuous) data in stratum h, $\mathbf{E} \sim N\left(\mathbf{0}, c\mathbf{\Sigma}_{\text{orig}^s}\right)$, $\mathbf{\Sigma}_{\text{orig}^h}$ is the covariance matrix of the original data in stratum h, c is the parameter of the method. Such noise preserves the correlation structure within the strata. Conditional Group Swapping is designed with the same goal, so the combination of these two methods may work in synergy. Investigation of the best combinations of Conditional Group Swapping with other methods is one of the directions of our future research.

Another direction for future research is the investigation of the risk associated with the method. We believe that the risk assessment is more comprehensive and practically useful when done for the final version of the masked data, which, as we noted above, will result from the application of our Conditional Group Swapping together with other SDL methods.

Indeed, if there are continuous variables in the data and they are not masked, then re-identification risk can be high regardless of the protection of categorical variables, because the values of continuous variables are virtually unique. CGS method is not suited for continuous variables, however, as mentioned above, it can be used in combination with additive noise. So, we carried out several experiments with this combination, in particular we applied it to both out data sets. The value $c = 0.15$ was used as a parameter of noise (see [17, 12] for recommendations for c). Changes in utility were insignificant, in particular the average confidence interval overlaps decreased by about 3 to 5 percent, and the range of variation was almost the same.

Next, we estimated the re-identification disclosure risk, defined as an average percentage of correctly identified records when record linkage techniques [10, 6] are used to match the original and masked data. Specifically, we assume that the intruder tries to link the masked file with an external database containing a subset of the attributes present in the original data (see [17]).

The re-identification disclosure risk for the Titanic data masked with multivariate noise and CGS was low: about 4% of all records were correctly identified for $n_s = 20$ and about 3% for $n_s = 40$. For SMHO data the risk was even lower, it was about 2% for $n_s = 20$ and 1.5% for $n_s = 40$.

As we mentioned above, these experiments do not represent a comprehensive risk analysis, however, they give an idea of the magnitude of risk. Thorough investigation of the disclosure risk for the combination of Conditional Group Swapping together with different SDL methods is the topic of our future research.

Acknowledgments

The authors would like to thank Alan Dorfman and Van Parsons for valuable suggestions and help during the preparation of the paper. The findings and conclusions in this paper are those of of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

References

1. Brand R. Microdata protection through noise, In Domingo-Ferrer J, editor, Inference Control in Statistical Databases, Lecture Notes in Computer Science, 2316:97–116, SpringerVerlag, 2002.
2. Dalenius T and Reiss SP. Data-swapping: A technique for disclosure control Journal of Statistical Planning and Inference, 6:73–85, 1982.
3. Defays D and Anwar N. Micro-aggregation: A generic method. In Proceedings of the 2nd International Symposium on Statistical Confidentiality, 69–78, Luxembourg, O ce for Official Publications of the European Community, 1995.
4. Drechsler J. Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation, Springer, 2011.
5. Elinder M and Erixson O. Gender, social norms, and survival in maritime disasters. Proceedings of the National Academy of Sciences of the United States of America, 109(33): 13220–13224, 2012. [PubMed: 22847426]
6. Fellegi IP and Sunter AB. A theory for record linkage, Journal of American Statistical Association, 64:1183–1210, 1969.

7. Gomatam S,Karr AF, Chunhua L and Sanil A. Data swapping: A risk-utility framework and web service implementation. Technical Report 134, National Institute of Statistical Sciences, Research Triangle Park, NC, 2003.
8. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Lenz R, Longhurst J, Schulte-Nordholt E, Seri G, and DeWolf P-P. Handbook on Statistical Disclosure Control (version 1.2). ESSNET SDC project, 2010 http://neon.vb.cbs.nl/casc.
9. Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Schulte Nordholt E, Spicer K, K., and DeWolf P-P. Statistical Disclosure Control, Wiley, 2012.
10. Jaro MA. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association, 84:414–420, 1989.
11. Kaggle. The Home of Data Science. http://www.kaggle.com
12. Karr AF, Kohnen CN, Oganian A, Reiter JP, and Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician, 60(3):224–232, 2006.
13. Kim JJ. A method for limiting disclosure in microdata based on random noise and transformation. In Proceedings of the ASA Section on Survey Research Methodology, 303–308, 1986.
14. Lin YX. Density Approximant Based on Noise Multiplied Data, Privacy in Statistical Databases 2014, LNCS 8744:89–104, 2014.
15. Mitra R, and Reiter JP. Adjusting Survey Weights When Altering Identifying Design Variables Via Synthetic Data Privacy in Statistical Databases 2006, LNCS 4302:177–188, 2006.
16. Moor R. Controlled data swapping techniques for masking public use microdata sets. U.S. Census Bureau, 1996.
17. Oganian A. Security and Information Loss in Statistical Database Protection. PhD thesis, Universitat Politecnica de Catalunya, 2003.
18. Oganian A and Karr AF. Combinations of SDC methods for microdata protection Privacy in Statistical Databases 2006, LNCS, 4302:102–113, 2006.
19. Oganian A and Karr AF. Masking methods that preserve positivity constraints in microdata. Journal of Statistical Planning and Inference, 141(1):31–41, 2011.
20. Reiss SP, Post MJ and Dalenius T. Non-reversible privacy transformations. In Proceedings of the ACM Symposium on Principles of Database Systems, 3 29–31, 139–146, 1982.
21. Rosenbaum PR and Rubin DB. The Central Role of the propensity score in observational studies for Causal Effects. Biometrika, 70:41–55, 1983.
22. Takemura A. Local recoding and record swapping by maximum weight matching for disclosure control of microdata sets. Journal of Official Statistics,18, 275289, 2002.
23. Templ M. Statistical disclosure control for microdata using the R-package sdcMicro. Transactions on Data Privacy, 1(2):67–85, 2008.
24. Torra V. Microaggregation for categorical variables: a median based approach In Privacy in Statistical Databases 2004, LNCS 3050: 162–174, 2004.
25. Valliant R, Dever JA and Kreuter F. Package ‘PracTools’: Tools for Designing and Weighting Survey Samples. https://cran.r-project.org/web/packages/PracTools/PracTools.pdf, 2015.
26. Woo M-J, Reiter JP, Oganian A, and Karr AF. Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality, 1(1):111–124, 2009.

Table 1.

The Titanic data results: original and masked con_dence interval overlaps.

		CGS			RGS		
	rate	**Average**	**Range**	**# non-over**	**Average**	**Range**	**# non-over**
Regl	20	0.88	[0.6, 0.99]	0/500	0.52	[−0.53, 0.95]	100/500
	40	0.65	[−0.15, 0.96]	51/500	0.16	[−1.86, 0.92]	106/500
Reg2	20	0.85	[0.60, 0.98]	1/1800	0.76	[0.42, 0.97]	1/1800
	40	0.79	[0.41, 0.97]	0/1800	0.69	[0.28, 0.95]	3/1800

Table 2.

The Titanic data results: ratios of means and ratios of covariance matrices based on the original and masked data.

		CGS			RGS		
	rate	**Average**	**Range**	**# sign change**	**Average**	**Range**	**# sign change**
Mean ratio	20	1.003	[0.91, 1.10]	N/A	1.002	[0.81, 1.21]	N/A
	40	1.004	[0.87, 1.15]	N/A	1.02	[0.72, 1.37]	N/A
Cov. ratio	20	1.02	[0.61, 1.67]	208/9600	1.03	[0.51, 1.62]	238/9600
	40	1.5	[0.52, 1.68]	238/9600	0.99	[0.34, 1.68]	364/9600

Table 3.

The SMHO data results: original and masked con_dence interval overlaps.

		Conditional Swap			Random Swap		
	rate	**Average**	**Range**	**# non-over**	**Average**	**Range**	**# non-over**
Reg3	20	0.91	[0.72, 0.99]	0/900	0.72	[0.2, 0.99]	2/900
	40	0.84	[0.84, 0.99]	0/900	0.64	[−0.29, 0.97]	100/900
Reg4	20	0.94	[0.81, 1]	0/900	0.84	[0.58, 0.97]	0/900
	40	0.92	[0.77, 0.99]	0/900	0.71	[0.26, 0.95]	11/900
Reg5	20	0.85	[0.56, 0.99]	5/2500	0.64	[−0.25, 0.97]	196/2500
	40	0.82	[0.45, 0.98]	14/2500	0.47	[−0.7, 0.95]	393/2500
Reg6	20	0.81	[0.45, 0.98]	0/2500	0.72	[0.26, 0.96]	25/2500
	40	0.73	[0.15, 0.97]	71/2500	0.61	[−0.06, 0.94]	158/2500

Table 4.

The SMHO data results: ratios of means and covariance matrices based on the original and masked data.

		Conditional Swap			Random Swap		
	rate	**Average**	**Range**	**# sign change**	**Average**	**Range**	**# sign change**
Mean ratio	20	0.99	[0.85, 1.09]	N/A	0.94	[0.56, 1.17]	N/A
	40	0.96	[0.63, 1.12]	N/A	0.92	[0.38, 1.32]	N/A
Cov. ratio	20	1.03	[0.57, 1.46]	0	0.87	[0.05, 2.07]	276/8000
	40	0.90	[0.34, 1.49]	334/8000	0.79	[0.016, 2.88]	332/8000