UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Extreme value protection adjustment for different subpopulations in complex data sets.

Anna Oganian (National Center for Health Statistics)
annaoganyan7@gmail.com

Abstract

Public release of microdata sets with many attributes of different types is very important for research, policymaking and training. Statistical agencies strive to meet the demand for such data. However, disclosure risk protection of complex data sets is a challenging problem. In order to produce a good-quality public data, it is important to take into account the relationships between the variables. Statistical agencies are often concerned about disclosure risk associated with the extreme values of numerical variables. Thus, such observations are often top or bottom-coded, or they can be synthesized. Having a single rule for defining which records and values can be considered extreme, or in the case of top-coding, applying only one top-code threshold for all the records in the complex data set can lead to under- protection for some sub-populations and over-protection for the others. This is particularly the case when a particular subpopulation is different from other subpopulations in terms of a specific variable that is considered for disclosure limitation. For example, males and females of different race/ethnicity groups may have different definitions/notions of "extreme" for such variables as height or weight. In this presentation, we discuss varying definitions of extreme values for different subpopulations. Our approach is based on clustering the variables into groups according to some metric of closeness between the variables and then forming the rules for "extreme values" for different subpopulations using techniques of Association Rule Mining within the clusters of variables obtained in the previous step. For the top or bottom coding, we present how several different thresholds can be used for different subpopulations. We illustrate our method on a genuine multivariate data set of realistic size.

Extreme Value Protection Adjustment for Different Subpopulations in Complex Datasets

Anna Oganian[1], Mehtab Iqbal[2] and Goran Lesaja[3,4]

[1] National Center for Health Statistics
3311 Toledo Rd
Hyattsville, MD, 20782, U.S.A.
aoganyan@cdc.gov
[2] Clemson University
School of Computing
mehtabi@g.clemson.edu
[3] Georgia Southern University
Department of Mathematical Sciences
P.O. Box 8093, Statesboro, GA 30460, U.S.A.
[4] United States Naval Academy
Mathematics Department
121 Blake Road, Annapolis, MD 21402, U.S.A.
goran@georgiasouthern.edu

Abstract. Public release of microdata sets with many variables of different types is very important for research, policymaking and training. Statistical agencies strive to meet the demand for such data. However, the protection of complex datasets from disclosure risk is a challenge. In order to produce a good quality public dataset, it is important to take into account the relationships between the variables. Statistical agencies are often concerned about disclosure risk associated with the extreme values of numerical variables. Thus, such observations are often top or bottom-coded or synthesized. A single rule for defining which records and values can be considered as extreme can lead to under-protection for some subpopulations and over-protection for others. This disparity in protection is evident when a certain subpopulation is different from other subpopulations in terms of a specific variable subject to protection/modification. For example, males and females of different race/ethnicity groups may have different definitions of "extreme" for variables such as height and weight. In this paper, we discuss varying definitions of extreme values for different subpopulations. Our approach is based on clustering the variables into groups according to some metric of closeness between the variables. In the next step, we form the rules defining the notion of "extreme" for different subpopulations using techniques of Association Rule Mining within the clusters of variables obtained in the previous step. We illustrate our approach on a genuine multivariate dataset of a realistic size showing how several different thresholds can be used for different subpopulations, and then values above these thresholds can be synthesized using existing Statistical Disclosure Limitation (SDL) methods.

Keywords and phrases: Statistical disclosure limitation, extreme values, high dimensional dataset, subpopulations, synthetic data, association rule mining, hierarchical clustering, dimensionality reduction.

1 Introduction

Government statistical agencies in charge of conducting national surveys have an obligation by law to protect the privacy and confidentiality of their respondents, who can be individuals or enterprises. Such surveys often have a large number of variables of different types. Some examples of such surveys in the USA are the National Health Interview Survey [12], the Behavioral Risk Factor Surveillance System [5], the Current Population Survey [7], and the American Community Survey [2].

Records that have extreme values are often a subject of concern regarding the disclosure risk associated with these values. To address this risk, numerical variables can be top-coded, synthesized, or other methods of disclosure limitation can be applied to protect them. The definition of "extreme", however, depends on the context and can vary among different subpopulations. If the values of the thresholds which define extremes of numeric variables are determined independently of the values of other variables, then some subpopulations can be under-protected. For example, if the data protector decides to protect extreme values of income and top-codes or synthesizes all the values above $\$150{,}000$, then such a threshold may be inadequate for those who work part-time, or those who were unemployed part of the year, those whose education level is low, and for certain occupation classes [16]. In other words, such individuals might not be adequately protected because the income of $\$150{,}000$ may be very rare within these groups, much rarer than in other groups, and hence might have a higher risk of re-identification. Thus, from the disclosure risk perspective, it would be desirable to determine appropriate thresholds for extreme values for certain groups (subpopulations) different from those for the rest of the population. However, such subgroups should first be identified. In some cases, as in the example above, this task may be intuitive and easy, but in datasets with a large number of variables, such a task is not always trivial.

1.1 Contributions and plan of the paper

In [14], a multivariate top-coding procedure was proposed to provide a similar level of protection in different subpopulations when top-coding is a statistical disclosure limitation method of choice. In this paper, we extend the approach of [14] to be applied with other SDL methods, such as synthesis of extreme values. Similarly as in [14], on the first step we cluster the variables in groups to reduce the complexity of the problem. Next, we use techniques of Association Rule Mining (ARM) [3] to determine the rules for extreme values for different subpopulations. In Section 2, our general procedure and its application to the synthesis of extreme values is described. In Section 3 the method is illustrated

on a genuine dataset of realistic size. Concluding remarks are given in Section 4.

2 The description of the method

Let D be a microdata file with p variables. Let $T = \{T_1, \cdots, T_k\} \in D$, $k < p$ denote numerical variables with extreme observations that need to be protected. For multivariate datasets, the number of possible combinations of categories of its variables can be extremely high, and each such combination may define a potential subpopulation or group of individuals. Thus, identification of subpopulations that might require adjusted computation of thresholds for extreme values of numerical variables can be computationally very demanding. To make it feasible, variables may be clustered into groups in such a way that the groups are formed around each numerical variable T_i, and the search of the aforementioned subpopulations is done within the vertical partition of the data where each partition corresponds to a cluster of variables around T_i. In [14], a simple method of clustering the variables was proposed where the variables at a distance less than a certain value h were included in the cluster around T_i. The distance is measured by the squared canonical correlation between T_i and other variables, the higher the correlation - the closer the variables. The advantage of canonical correlation is that it can be computed for variables of different types, numerical and categorical. The cut-off value h depends on the preferences of the data protector, intuitively representing a trade-off between accuracy/utility and computational complexity of the procedure.

To accomplish the search of the subpopulations within the clusters, we use association rule mining. There are several reasons for using ARM. First, definitions of extremes of numerical variables associated with different subpopulations can naturally be expressed as an association rule. As stated in [3], an association rule is an expression of the form $X \rightarrow Y$, where X and Y express conditions on the variables of the following form:

$$V_i = cat_{i_l} \wedge \cdots V_f \in [l_f, u_f] \cdots \wedge V_j = cat_{j_m}, \tag{2.1}$$

where $V_i, V_j, \cdots V_f$ are the variables from the dataset D, $cat_{i_l}, \cdots cat_{j_m}$ are the categories of the categorical variables, and $[l_f, u_f]$ are specific intervals within the domains of the corresponding continuous variables. In this paper, we call the rule's antecedent, X, a "LHS of the rule", and the consequent of the rule, Y, a "RHS of the rule".

The association rules defining extremes of numerical variables can be represented in the following form:

$$(V_i = cat_{i_l}) \wedge \cdots (V_j = cat_{j_m}) \rightarrow T_i < threshold. \tag{2.2}$$

For example,

$$(Sex = Female) \wedge (Height < 65 \text{ inches}) \rightarrow (Weight < 200). \quad (2.3)$$

Association rules are characterized by their support and confidence. According to the original definition and notation used in [3], a support of X, the antecedent of the rule, is defined as the proportion of records in the database D that satisfy the expression X:

$$Supp(X) = |\{r \in D | X \subseteq r\}|/|D|,$$

where r denotes a record in D and $|\cdot|$ indicates cardinality.

Confidence of the rule is defined as the proportion of the records in D that contain X which also contain Y:

$$Conf(X \rightarrow Y) = Supp(X \cup Y)/Supp(X)$$

There are other measures of performance of the rules, for example lift of the rule [19] defined as follows:

$$Lift(X \rightarrow Y) = Supp(X \cup Y)/(Supp(X) * Supp(Y))$$

Lift measures the performance of an association rule at predicting the event given by the RHS of the rule (when the LHS is true) comparative to a random association, when the events given by RHS and LHS are independent of each other. In the later case no rule can be drawn involving these two events. Typically the rules with $lift > 1$ are considered as potentially interesting rules[19].

The standard Apriori algorithm described in [4] can be used to mine association rules where all the variables are categorical. In the case of mining association rules on both categorical and numerical variables, often called mining quantitative association rules, there is no "gold standard". However, several implementations exist, for example, a genetic-based algorithm called QuantMiner [17]. The availability of different ARM techniques and methods, as well as their efficient implementations designed to work well for large databases, is an advantage of ARM. These implementations are typically fast and can efficiently search for the groups of records for which the RHS of the rule in (2.2) is satisfied. This is another reason we decided to use this methodology in the context of SDL of extreme values.

2.1 Definitions of the extreme values using ARM

As mentioned earlier, the definition of an attribute's "extreme" may vary depending on the context where "context" is essentially a description of a group

of records sharing the same values of categorical variables, for example, part-time workers, Asian females, etc. On the other hand, for numerical variables regardless of the context, the perception of "extremes" is often associated with its high percentiles (or low percentiles). Thus, one way to approach this problem is to identify the subpopulations for which high percentiles (or low percentiles) of the variables subject to disclosure limitation are different from the corresponding percentiles computed on the entire population. In particular, high percentiles can be much smaller (or low percentiles can be much larger) than the corresponding percentiles computed on the entire population, which may increase the risk of disclosure for the individuals from these subpopulations. In the rest of the paper, we focus our discussion on high percentiles since the protection of low percentiles is a straightforward extension of the protection of high percentiles.

In [14], we proposed a procedure that can be used to determine the subpopulations with different P-th percentiles, lower than the rest of the population, where P is a high percentile rank such as 95-th or 99-th. If top-coding is the preferred disclosure limitation method, then the procedure described in [14] can be used to identify top-code thresholds.

If synthesis is the preferred method for extreme value protection, various definitions of extremes can be used. Because synthetic methods for disclosure limitation can be used to protect any range of values, there is no reference in the SDL literature to specific percentiles as thresholds beyond which values should be synthesized, as in the case of top-coding. Indeed, the data protector can synthesize high percentiles of numeric variables or can use some common rules defining outliers. For example, one approach may be to use Tukey's fences [20] and synthesize the outliers. In the latter case, high outliers for a subpopulation sub are the values above the threshold $Thresh_{sub} = Q3_{sub} + k * IQR_{sub}$, where $Q3_{sub}$ is the third quartile of the corresponding subpopulation, IQR_{sub} the inter-quartile range of the corresponding subpopulation, and k can be chosen to be equal to 1.5, so $Thresh_{sub}$ indicates a possible outlier. Another option would be to let k be equal to 3 so that $Thresh_{sub}$ indicates data that is "extreme" (see [20]). In this paper, we use $k = 3$ to illustrate our method. However, the data protector can choose $k = 1.5$ if a more conservative definition of an outlier is preferred in a certain scenario of data release.

Hence, for synthesis, the following procedure can be used:

Algorithm 2.1

1. Compute threshold $Thresh_{pop} = Q3_{pop} + 3IQR_{pop}$ for the variable T_i using all the records in the dataset (subscript pop refers to the entire population). It will serve as a benchmark.

2. Mine the following type of association rules on the vertical partition of the data that corresponds to the cluster of variables $Clust_i$:

$$X \to T_i < Thresh_{pop} - \Delta. \tag{2.4}$$

The LHS of the rule X in (2.4) represents any combination of the variables/ categories from $Clust_i$ in the form given by (2.1). The RHS of the rule in (2.4) is the expression that makes the implication (that is, the rule) true. Δ is a minimal difference between the threshold $Thresh_{pop}$ computed on the entire dataset and a particular subpopulation that might need an adjustment to ensure protection of extreme values. The parameter Δ can be chosen by the data protector for practical reasons – in order not to have too many adjustments which are not very different from the benchmark $Thresh_{pop}$.

3. Choose the rules with the confidence larger than $Supp(T_i < Thresh_{pop} - \Delta)$, that is the support of the right-hand side of (2.4). Note, that $Supp(Thresh_{pop} - \Delta)$ is computed on the entire dataset and thus, it can be done before the mining process begins. Denote this set of rules as S.
4. Filter the set S, that is, only those rules which refer to the subpopulations satisfying the following expression

$$Thresh_{sub} < Thresh_{pop} - \Delta, \tag{2.5}$$

are retained. The LHS of the rules satisfying (2.5) describe subpopulations that may need special protection.

5. Compute $Thresh_{sub} = Q3_{sub} + k * IQR_{sub}$ – the actual thresholds for the identified subpopulations. Synthesize the outliers above these thresholds using the preferred synthetic method.

On step 3 of the above algorithm our choice of the level of confidence is based on the recommendations in the literature [19] to choose rules which have $Lift > 1$. Because $Lift(X \to Y) = Conf(X \to Y)/Supp(Y)$, $Conf(X \to Y)$ should be greater than $Supp(Y)$. Thus, for the rules given by expression (2.4) we choose the ones with confidence greater than $Supp(T_i < Thresh_{pop} - \Delta)$. Because $Thesh_{pop}$ is the threshold beyond which the observations can be classified as outliers, $Supp(T_i < Thresh_{pop} - \Delta)$ tends to be very high (depending on the value of Δ), which helps to reduce the number of rules. In our experiments, described in Section 3, it was about 0.95.

3 Numerical experiments

To illustrate Algorithm 2.1 described in the previous section, we applied it to a genuine multivariate dataset that was downloaded from the UCI Machine Learning Repository [8]. This is a sample drawn from the Public Use Microdata Samples (PUMS) person-level 1990 US Census file. In the paper, we will refer to

this file as "Census". We used 66 numerical and categorical variables and 1.2 million records in our experiments. A full description of the variables can be found in [6]. We experimented with several variables, but due to space limitations we present the results for the variable $Income1$ - wages or salary earned by the individuals in 1989. The results with other numerical variables are very similar in nature and are therefore not shown here.

The first step of our procedure, grouping the variables, led to a cluster of nine variables around $Income1$, these are $Class$ - class of worker with 10 categories, $IndustryClass$ - industry class with 13 categories, $Occupclass$ - occupation class with 8 categories, $Relat1$ - the respondent's relationship to the householder with 13 categories, $Disable1$ - work limitation with three categories, $Rlabor$ - current employment status with 7 categories, $Hour89$ - numerical variable denoting usual hours worked per week the year before the interview, $Week89$ - weeks worked the year before the interview, and $Yearsch$ - educational attainment with 18 categories.

To make interpretation of the rules easier and reduce the number of rules, we converted the variables $WEEK89$ and $HOUR89$ to categorical $W89$ and $H89$, respectively. In particular, if $WEEK89 < 26$ then $W89 = 1$ and 0 otherwise. $W89 = 1$ represents those individuals who worked less than half of the year, including seasonal workers and those who were unemployed for at least one-half of 1989. Categorical $H89 = 1$ if $HOUR89 < 30$ and 0 otherwise. $H89 = 1$ can be thought as an indicator for part-time workers (similar definition is given in [1]).

Below, we list the rules that are representative of this dataset. They have attribute categories that appear most frequently. It should be noted that the rules presented below are not our recommendations on extreme value thresholds definitions for this particular dataset nor any similar dataset. The rules and results presented in this section are for the illustration of our method only.

For the groups of individuals that fit the description that appears in the LHS of the rules based on Algorithm 2.1, the threshold that defines outliers for $Income1$ according to the IQR rule appears on the RHS of the rules. It is worth noting, that the threshold for outliers for $Income1$ computed on the entire dataset is \$92,000 which can be different from the corresponding thresholds for various subpopulations, as seen below.

$W89 =$ Part-year worker $\rightarrow Income1 <$ \$16,400
$H89 =$ Part-time worker $\rightarrow Income1 <$ \$22,300
$Relat1 =$ Persons in group quarters $\rightarrow Income1 <$ \$29,300
$Relat1 =$ Other relative of the householder $\rightarrow Income1 <$ \$51,200
$Relat1 =$ Son/stepson/daughter/stepdaughter of the householder $\rightarrow$
$Income1 <$ \$40,000

$Occupclass =$ Service $\rightarrow Income1 <$ \$41,300
$Occupclass =$ Farming $\rightarrow Income1 <$ \$51,300
$Relat1 =$ Husband/wife of the householder $\wedge$
$Yearsch =$ High school diploma or less $\rightarrow Income1 <$ \$47,100
$Class =$ Employee of private not for profit company $\wedge$
$Yearsch =$ High school diploma or less $\rightarrow Income1 <$ \$50,000
$Industryclass =$ Trade $\wedge\, Yearsch =$ High school diploma or less $\rightarrow$
$Income1 <$ \$50,000
$Industryclass =$ Agriculture $\wedge\, Yearsch =$ High school diploma or less $\rightarrow$
$Income1 <$ \$41,200

Before discussing the rules, it should be noted that the main purpose of the proposed procedure is the reduction of the computational and procedural burden for the data protector. Indeed, going through a large number of possible combinations of the relevant variables in a big dataset in order to find rarely observed values of numerical variables for certain groups of records or subpopulations can be daunting. Thus, our automated procedure is meant to bring such special cases to the data protector's attention. However, the decision about whether to use these rules to synthesize or apply top-codes to protect extreme observations depends on many factors, such as a particular scenario of data release, SDL practice at a particular institution, and preferences of data protectors. In any case, such decisions are usually made together with the subject area specialists.

Some of the rules presented above seem intuitive or common sense. One example of such rules are those that have income on the RHS and usual hours worked per week or number of weeks worked in 1989 on the LHS. These two variables are positively correlated with income. The rules suggest lower thresholds for income for part-time workers and individuals who worked only a part of the year, compared to the rest of the population.

Another example of the rules that are intuitive are those that involve $Relat1$ (relationship of the respondent to the householder) on the LHS of the rules. For instance, when $Relat1 = son/daughter$, then the threshold for $Income1$ may be lower compared to other groups of individuals. According to the documentation on the 1990 Census data files [21], in most cases, a householder is a person, or one of the persons, in whose name the home is owned, being bought, or rented. Higher-income respondents may be expected to be householders themselves, rather than living with a parent-householder, which may be one of the reasons for these respondents to be lower-income. Similar reasoning may be applied to the rules that involve other relatives of the householder and their respective extreme values.

Another characteristic that is related to income is the occupation of the respondent ($Occupclass$ variable). The rules identified some occupation classes with lower values of $Income1$ in this dataset. For example, $Occupclass = Service$ which includes cooks, waiters and waitresses, housekeepers, cleaners, maids and housemen, hairdressers, welfare service aides, and some others, has a lower threshold for extreme values of income, which agrees with the literature on the subject [16]. Also, according to the rules, $Occupclass = Farmers$ has a lower threshold of $Income1$ as well.

As expected, rules that included the variable $Yearsch$, educational attainment, indicated that if educational attainment is less than high school, then $Income1$ is limited, especially for certain categories of individuals in the dataset.

We conclude this discussion by emphasizing that the focus of this paper is not the discussion and analysis of particular rules, rather it is the development and description of the methodology to obtain such rules. A deeper analysis of the rules obtained by our procedure should be done by the data protector and the subject area specialist for each particular dataset as well as the scenario of data release.

After the rules were obtained, we synthesized the extreme values of $Income1$ according to the rules that we obtained. We used the R package Synthpop [13], in particular, a method based on Classification and Regression Trees (CART) [15] which is a default method for synthesis in Synthpop. Because we synthesize only the extreme values which are defined by the association rules described above, CART models were estimated using only the records with such extreme values, as it is recommended in [15], in order to maximize the utility of the resulting masked data. Below, we call masked data the Census dataset with extreme values substituted by the synthesized ones. To assess the quality of the masked data, we computed several utility metrics and compared some of the univariate statistics of $Income1$ based on the original and masked data. Regarding univariate statistics, the mean of the masked $Income1$ is $\$21{,}644$ and the mean of the original data is $\$21{,}627$. The standard deviation for the masked data is $\$22{,}038$, and $\$22{,}007$ for the original data. Medians and interquartile ranges are exactly the same in the masked and the original data, being $\$17{,}000$ and $\$21{,}000$ respectively. Regarding utility metrics, we computed a confidence interval overlap measure [10] which represents a probability overlap in the confidence intervals for the regression coefficients computed on the masked and the original data. The regression model for this measure has $Income1$ as a predicted variable and all other numerical variables (from the same cluster) as its predictors. Average confidence interval overlap is 90% which is good because maximum probability overlap is 95%. We also computed a generic measure of data utility based on propensity scores [22] , [18]. Generic measures of data util-

ity are designed to assess general discrepancies between the distributions of the original and masked data. The propensity score utility measure for this data is 0.05, which can be interpreted as a relatively good utility ([22]). For reference, the possible range of values for this measure is 0 to 0.25, smaller values meaning better utility, with 0 corresponding to the case when original and masked data are identical.

To assess the disclosure risk, we computed the absolute differences between the original and synthesized values of $Income1$. Average differences ranged between \$14,000 and \$50,000 among different subpopulations where extreme values were synthesized. We also computed the average percentage change:

$$|orig_i - syn_i|/|orig_i| * 100\%, \tag{3.1}$$

where $orig_i$ stands for the i-th original value of $Income1$ and syn_i - for its corresponding synthetic value. The average percentage change was about 40% for this dataset. Disclosure risk assessment based on the differences between the original and masked data has been used in the SDL literature before, for example [11] reports smaller change for their methods - about 10% (which means higher risk). Some existing SDL methods, like rankswapping [9], even have a disclosure risk parameter which is based on differences between the original and masked data. Hence, depending on the practice at a particular institution and the scenario of data release, data protectors can achieve different levels by varying method parameters and applying more or less intensely the protection method. For example, in the case of synthesis using CART, the data protector can specify a minimum size of a final node that a CART model can produce. This can help to decrease the disclosure risk even further - details can be found in [13].

4 Concluding remarks

In this paper, we discuss an approach to identify subpopulations that may need special consideration while protecting extreme values of numerical variables. This procedure may be used as an aid for data collecting organizations in the disclosure review process as an alternative, or in addition, to their regular procedures. Such procedures often involve identification of risky combinations of variables, which are typically based on intuition and knowledge of a particular dataset. In big datasets, these procedures may be complicated and computationally involved as they require the computation of many tabulations to identify potentially rare combinations of variables that may pose increased disclosure risk. Thus, an automated procedure to identify such cases may be helpful, especially when the data protector intends to release datasets with many variables of different types, as is the case for many national surveys.

5 Aknowledgments

We would like to thank John Pleis and Lauren Rossen from the National Center for Health Statistics for the careful review of the paper and their valuable suggestions.

Disclaimer

The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the National Center for Health Statistics, Centers for Disease Control and Prevention.

References

1. ACA definition of full-time employee, https://www.shrm.org/hr-today/public-policy/hr-public-policy-issues/Documents/15%20ACA%20Definition%20of%20Full-Time%2003-10.pdf
2. ACS: American community survey. United States Census Bureau, https://www.census.gov/programs-surveys/acs
3. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. pp. 207 – 216 (June 1993)
4. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB. pp. 487 – 499. Santiago, Chile (September 1994)
5. BRFSS: Behavioral risk factor surveillance system. Centers for Disease Control and Prevention (CDC), https://www.cdc.gov/brfss/index.html
6. Census: US census (1990) data set. UCI Machine Learning Repository (2017), https://archive.ics.uci.edu/ml/datasets/US+Census+Data+%281990%29
7. CPS: Current population survey. United States Census Bureau, https://www.census.gov/programs-surveys/cps.html
8. Dheeru, D., Karra Taniskidou, E.: UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences (2017), http://archive.ics.uci.edu/ml
9. Hundepool, A., DomingoFerrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., de Wolf, P.: Statistical Disclosure Control. Wiley (July 2012)
10. Karr, A.F., Kohnen, C.N., Oganian, A., Reiter, J.P., Sanil, A.P.: A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. The American Statistician 60(3), 224–232 (2006)
11. Karr, A.F., Oganian, A.: Masking methods that preserve positivity constraints in microdata. J. Statist. Planning Inf. 141(1), 31–41 (2010)
12. NHIS: National Health Interview Survey. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS), https://www.cdc.gov/nchs/nhis/index.htm
13. Nowok, B., Raab, G., Dibben, C.: Synthpop: Bespoke creation of synthetic data in R. Journal of Statistical Software 74, 1–26 (2016), https://www.jstatsoft.org/article/view/v074i11
14. Oganian, A., Iacob, I., Lesaja, G.: Multivariate Top-Coding for Statistical Disclosure Limitation. In: J, D.F., K., M. (eds.) Privacy in Statistical Databases, Lecture Notes in Computer Science. vol. 12276, pp. 136–148. Springer-Verlag (2020)

15. Reiter, J.: Using cart to generate partially synthetic, public use microdata. Journal of Official Statistics 21(3), 441–462 (2005)
16. Ross, M., Bateman, N.: Meet the low-wage workforce. Tech. rep., Brookings (2019)
17. Salleb-Aouissi, A., Vrain, C., Nortet, C., Xiangrong Kong, X., Vivek Rathod, V., Cassard, D.: Quantminer for mining quantitative association rules. Journal of Machine Learning Research 14(61), 3153–3157 (2013), http://jmlr.org/papers/v14/salleb-aouissi13a.html
18. Snoke, J., Raab, G., Nowok, B., Dibben, C., Slavkovic, A.: General and specific utility measures for synthetic data. Journal of the Royal Statistical Society. Series A 181, 663–688 (2018)
19. Tufféry, S.: Data Mining and Statistics for Decision Making. Chichester, GB: John Wiley and Sons (2011)
20. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley (1977)
21. U.S. Department of Commerce Economics and Statistics Administration. BUREAU OF THE CENSUS: 1990 Census of Population and Housing. Public Use Microdata Samples. United States (1990)
22. Woo, M.J., Reiter, J.P., Oganyan, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1(1), 111–124 (2009)

A Appendix. Selected variables in the 1990 U.S. Census Public Use Microdata Samples (PUMS) person-level file

$Class$ - Class of worker. Categories: 0 N/a, Unemployed who never worked. 1 Employee of a private for profit company. 2 Employee of a private not for profit company. 3 Local government employee (city, county, etc.). 4 State government employee. 5 Federal government employee. 6 Self employed in own not incorporated business. 7 Self employed in own incorporated business. 8 Working without pay in family business or farm. 9 Unemployed, last worked in 1984 or earlier.

$IndustryClass$ - Industry class. Categories: 1 Agriculture. 2 Mining. 3 Manufacturing. 4 Transportation. 5 Wholesale trade. 6 Retail trade. 7 Finance. 8 Business. 9 Personal services. 10 Entertainment. 11 Professional. 12 Public administration.

$Occupclass$ - Occupation class. Categories: 1 Managerial. 2 Professional. 3 Technical. 4 Sevice. 5 Farming. 6 Precision. 7 Operators. 8 Military.

$Relat1$ - Relationship to the householder. Categories: 0 Householder. 1 Husband/wife 2 Son/daughter 3 Stepson/stepdaughter 4 Brother/sister 5 Father/mother 6 Grandchild 7 Other relative 8 Roomer/boarder/foster child 9 Housemate/roommate 10 Unmarried partner 11 Other non-related. 12 Institutionalized person. 13 Other person in group quarters.

$Disable1$ - Work limitation. Categories: 0 N/a. 1 Yes, Limited in kind or amount of work. 2 No, not Limited.

$Rlabor$ - Current employment status. Categories: 0 N/a 1 Civilian employee, at work. 2 Civilian employee, with a job but not at work. 3 Unemployed. 4

Armed forces, at work. 5 Armed forces, with a job but not at work. 6 Not in labor force.

$Hour89$ - Usual hours worked per week the year before the interview. This is a numerical variable ranging from 0 to 99.

$Week89$ - Weeks worked the year before the interview. This is a numerical variable ranging from 0 to 52.

$Yearsch$ - Educational attainment. Categories: 0 N/a. 1 No school completed. 2 Nursery school. 3 Kindergarten. 4 1st, 2nd, 3rd, or 4th grade. 5 5th, 6th, 7th, or 8th grade. 6 9th grade. 7 10th grade. 8 11th grade. 9 12th grade, no diploma. 10 High school graduate, diploma or GED. 11 Some college, but no degree. 12 Associate degree in college, occupational. 13 Associate degree in college, academic program. 14 Bachelors degree. 15 Masters degree. 16 Professional degree. 17 Doctorate degree.