Providing data with high utility and no disclosure risk for the public and researchers: an evaluation by advanced statistical disclosure risk methods

M. Templ
Vienna University of Technology & data-analysis OG
Austria
e-mail: matthias@data-analysis.at

Abstract

The demand of data from surveys, registers or other data sets containing sensible information on people or enterprises have been increased significantly over the last years. However, before providing data to the public or to researchers, confidentiality has to be respected for any data set containing sensible individual information. Confidentiality can be achieved by applying statistical disclosure control (SDC) methods to the data. The research on SDC methods becomes more and more important in the last years because of an increase of the awareness on data privacy and because of the fact that more and more data are provided to the public or to researchers. However, for legal reasons this is only visible when the released data has (very) low disclosure risk.

In this contribution existing disclosure risk methods are review and summarized. These methods are finally applied on a popular real-world data set - the *Structural Earnings Survey* (SES) of Austria. It is shown that the application of few selected anonymisation methods leads to well-protected anonymised data with high data utility and low information loss.

1 Introduction

A microdata file is defined as a data set on individual level. For each observation a set of variables is typically available. Concerning SDC, these variables can be split into three categories.

- **Direct Identifiers:** Variables that definitely identify a statistical unit. For example, the social insurance number, name of companies or people or addresses are considered as direct identifiers.
- **Key variables:** A set of variables that - when considered together - may be used to identify an individual unit. For example with the combination of gender, age, region and occupation some individuals may be identified. Other examples for (confidential) key variables could be income, health information, nationality or political preferences. For the description of the methods, it is advantageous to distinguish between categorical and continuous scaled key variables.
- **Non-confidential variables:** All variables that are not classified in any of the former two groups.

The goal of anonymizing a microdata set is to prevent that confidential information can be linked to a specific respondent. The ultimative aim is to release a safe microdata set that has both, low risk of linking confidential information to individual respondents and high data utility.

2 Measuring disclosure risk

Measuring risk in an microdata set is of course of great concern when having to decide on whether a microdata set is safe to be released. To be able to assess the disclosure risk it is required to make realistic assumptions on the information data users might have at hand to match against the microdata set. These assumptions are called 'disclosure risk scenarios'. Based on a specific disclosure risk scenario one must define a set of identifying variables (key variables) that can be used as input for the risk evaluation procedure.

Typically risk evaluation is based on the concept of "rareness/uniquenessss" in the sample and/or in the population. The interest is on units/individuals/observations that possess rare combinations of key variables. Those can be assumed to be identified easier and thus have higher risk. It is possible to cross tabulate all identifying variables and have a look at its cast. Patterns[1] with only very few individuals are in this sense considered risky if they have also low sampling weights, i.e. if the expected individuals with the same pattern is expected to be low in the population.

2.1 Frequencies Counts

Consider a random sample of size n drawn from a finite population of size N. Let $\pi_j, j = 1, \ldots, N$ be the (first order) inclusion probabilities, i.e. the probability that the element u_j of a population of the size N is chosen in a sample of the size n.

All possible combinations of categories in the key variables $X_1, \ldots, X_m$ can be calculated by cross tabulation of these categorical variables. Let $f_i, i = 1, \ldots, n$ be the frequency counts obtained by cross tabulation and let F_i be the frequency counts of the population which belong to the same category. If $f_i = 1$ applies the corresponding observation is unique in the sample. If $F_i = 1$ applies then the observation is unique in the population. Note that F_i is usually unknown since usually information on samples is collected and only few information about the population is known from registers and/or external sources.

2.2 The k-Anonymity Concept

Based on a set of key variables a desired characteristic of a protected microdata set might be to achieve k-anonymity [Samarati and Sweeney, 1998, Sweeney, 2002]. This means that each possible combination of the values of the key variables features at least k units in the microdata, meaning that all $f_i \geq 3\ , i = 1, ..., n\ .$ A typical value is $k = 3$.

[1] a pattern is defined as a specific combination of values of all key variables

k-anonymity is typically provided by recoding categorical key variables and by additionally suppressing specific values in the key variables of individual units.

An extension of k-anonymity is l-diversity [Machanavajjhala et al., 2007]. Consider for one group of observations with the same pattern in the key variables and let the group fulfill k-anonymity. A possible data intruder can therefore not identify an individual in this group. However, if all observations have the same entries in a sensitive variable (such as *cancer* in the variable *medical diagnosis*) then the attack is successful anyway.

2.3 Considering Sample Frequencies on Subsets: SUDA2

SUDA (**S**pecial **U**niques **D**etection **A**lgorithm) estimates a disclosure risk for each individual. SUDA2 [see, e.g., Manning et al., 2008] is a recursive algorithm for finding Minimal Sample Uniques. The algorithm generates all possible variable subsets of defined categorical key variables and scans them for unique patterns in the subsets of variables. The risk of an observation is then dependend on two aspects.

(a) The lower the amount of variables needed to receive uniquenesss, the higher the risk (and the higher the *suda score*) of the corresponding observation.

(b) The larger the number of minimal sample uniqueness contained within an observation, the higher the risk of the observation.

(a) is calculated for each observation i by $l_i = \prod_{k=MSUmin_i}^{m-1}(m-k) \quad , i = 1, ..., n$, for m the *depth* (the maximum size of variable subsets of the key variables), $MSUmin_i$ the number of minimal uniques of observation i and n the number of observations of the data set. Since each observation is treated independently, the l_i that belongs to one pattern are summed up to result in a common suda score for each of the observation belonging to this pattern (this summation is the contribution of (b)).

To result in the final SUDA score, the suda score are normalized due division by $p!$, with p being the number of key variables. The so called DIS suda score is then calculated from the suda and the so called DIS scores [we refer to Elliot, 2000, for details]. SUDA2 does not consider sampling weights and biased estimates may therefore result.

2.4 Considering Population Frequencies - The Individual Risk

To define if an individual unit is at risk, typically a threshold approach is used. If the individual risk of re-identification for an individual is above a certain threshold value, the unit is said to be at risk. To calculate the individual risks it is necessary to estimate the frequency of a given key in the population. In the previous section, Section 2.1, the population frequencies have already been estimated. However, one can show that these estimates almost always overestimate small population frequency counts [details can be found in Templ and Meindl, 2010] and should not be used to estimate the disclosure risk.

A better approach is to use so-called super-population models in which population frequency counts are modeled given a certain distribution. The whole estimation procedure of sample counts given the population counts can be modeled, for example, by using a Negative Binomial distribution [see, e.g., Rinott and Shlomo, 2006]. It is out of scope of the paper to explain the final measurement of individual risk in this contribution but it can be found in Franconi and Polettini [2004] and Templ and Meindl [2010].

2.5 Measuring the Global Risk

Although the individual risk have to be respected since a data intruder should not be able to identify individuals, often also a measure of the global risk is estimated to express the risk of the whole data set with one number.

2.5.1 Measuring the Global Risk Based on the Individual Risks

The first approach is to determine a threshold for the individual risk and to calculate the percentage of individuals that have larger individual risk than this threshold.

2.5.2 Measuring the Risk Using Log-Linear Models

The sample frequencies, considered for each of M patterns m, f_m , $m = 1, ..., M$ can be modeled by a Poisson distribution, and the global risk may be defined as [see Skinner and Holmes, 1998]

$$\tau_1 = \sum_{m=1}^{M} \exp\left(-\frac{\mu_m(1-\pi_m)}{\pi_m}\right) \quad , \text{ with } \mu_m = \pi_m \lambda_m \quad . \tag{1}$$

For simplicity, the inclusion probabilities are assumed to be equal, $\pi_m = \pi$, $m = 1, ..., M$. τ_1 can be estimated by log-linear models including the main effects and possible interactions. The model is

$$\log(\pi_m \lambda_m) = \log(\mu_m) = \mathbf{x}_m \beta \quad .$$

To estimate the μ_m's, the regression coefficients β have to be estimated, for example, by using the iterative proportional fitting.Global risk measure 1 is then given by $\hat{r}_1 = \sum_{i=1}^{n} \mathbf{I}(f_i = 1) e^{-(1-\pi)\hat{\mu}}$ (corresponding to the risk $P(F_i = 1|f_i = 1)$) and the second one by $\hat{r}_2 = \sum_{i=1}^{n} \mathbf{I}(f_i = 1) e^{1-(1-\pi)\hat{\mu}}/((1-\pi)\hat{\mu})$ (corresponding to the risk $E(1/F_i|f_i = 1)$.

2.6 Measuring Risk for Continuous Key Variables

Applying the concept of uniquenesss and k-anonymity on quantitative variables results that every observation in the data set is unique. Hence, this approach will fail for continuous key variables.

If detailed information about a value of a continuous scaled variable is available, one may be able to identify (by linking information) and eventually gain further information about an individual. For continuous key variables it is assumed that an intruder has information about a statistical unit

2.6.1 Distance-Based Record Linkage

By using distance based record linkage methods the aim is to find the nearest neighbors between observations from two data sets. Domingo-Ferrer and Torra [2001] has shown that these methods outperform probabilistic methods. Generally, it is evaluated if the original value falls within an interval centered on the masked value. Such an interval might be based on the standard deviation of the variable [see also Mateo-Sanz et al., 2004].

Almost all data sets from Official Statistics consists of statistical units whose values in at least one variable are quite different from the main part of the observations. This leads to the fact that these variables are very asymmetric distributed. Such outliers might be enterprises with a very large value for turnover, for example, or persons with extremely high income or even multivariate outliers. Other disclosure risk methods that are not used in this contribution take the "outlyingness" of an observation into account [for details, see, Templ and Meindl, 2008].

3 Application to the Sturctural Earnings Survey

The Structural Earnings Survey (SES) is conducted in almost all European countries and it includes variables on earnings of employees and other variables on employees and employment level (e.g. region, size of the enterprise, economic activities of the enterprise, gender and age of the employees, ...).

Generally such linked employer-employee data are used to identify the determinants/differentials of earnings but also some indicators are directly derived from the hourly earnings like the gender pay gap or the Gini coefficient [Gini, 1912]. The most classical example is the income inequality between genders as discussed in Groshen [1991], for example.

A correct identification of factors influencing the earnings could lead to relevant evidence-based policy decisions. The research studies are usually focused on examining the determinants of disparities in earnings.

The Austrian SES 2006 survey data consists of 199.909 observations obtained from a two-stage design - in the first stage of the design, the enterprises are chosen with certain inclusion probabilities depending on the enterprise size and location, in the second stage employee's in the selected enterprises are chosen with different inclusion probabilities [for more information have a look at Geissberger, 2009].

3.1 Disclosure Risk and Information Loss for SES

The following variables are chosen as key variables:

Categorical key variables: *size of enterprise* (5 ordered categories), *age* (66 ordered categories), *location* (3 categories), *economic activity* (53 categories)

Continuous key variables: *hourly earnings*, *earnings*

risk, IL	orig	+rec1	+rec2	+rec3	+supp	mdav	add	corr	sh
R:2-a	2.49	0.47	0.24	0	0	0	0	0	0
R:3-a	5.65	1.12	0.56	0.01	0	0	0	0	0
R:ind	2.48	0.67	0.52	0.05	0.05	0.05	0.05	0.05	0.05
R:suda	0.87	0.15	0.1	0	0	0	0	0	0
R:glob	0.83	0.14	0.08	0	0	0	0	0	0
R:glob	1.35	0.23	0.13	0	0	0	0	0	0
R:num	100	100	100	100	100	99.73	7.86	61.86	12.26
IL1	-	-	-	-	-	0	11.29	0.11	1.02
IL:eig	-	-	-	-	-	0	5.68	0.06	1.77
IL:lm	0	0.24	0.24	0.03	0.03	0.04	240.29	0.2	8.53

Table 1: Disclosure risk and information loss on SES

Table 1 shows the resulting disclosure risk and information loss of the SES data. The the columns in Table 1 corresponds to the following data

orig: original data (key variables)

rec1: (recoding) the variable *economic activity* is recoded to 14 reasonable categories.

rec2: (recoding) additionally, the variable *size of employment* is recoded into three reasonable categories (10-49, 50-249, 250-...).

rec3: (recoding) additionally, age is discretised into six reasonable categories.

supp: (suppression) additionally, local suppression is applied so that no observation violates 3-anonymity.

mdav: microaggregation (method mdav, see e.g. Domingo-Ferrer and Mateo-Sanz [2002]) with aggregation level 3.

add: additive noise (noise parameter equals 10, see Templ et al. [2012])

corr: correlated noise (defaults of Templ et al. [2012])

sh: shuffling [Muralidhar and Sarathy, 2006]

The rows of Table 1 corresponds to disclosure risk and information loss measures - **R:2-a** (**R:3-a**): percent of observations violating 2(3)-anonymity, **R:ind**: percent of observations with individual risk below 0.01, **R:suda**: percent of observations having suda dis score lower than 0.1, **R:glob1**, **R:glob2**: global risks from log-linear models, **R:num**: distance-based disclosure risk, **IL1**: information loss IL1, **IL:eig**: information loss based on differences in the eigenvalues and **IL:lm**: model-based estimation information loss. The mentioned measures of information loss are briefly explained in the following.

IL1: $IL1 = \frac{1}{p}\sum_{i=1}^{p}\frac{|x_{ij}-x'_{ij}|}{\sqrt{2}S_j}$, scaled distances between original and perturbed values for all p continuous key variables.

IL:eig: The relative absolute differences between the eigenvalues of the covariance standardized continuous key variables of the original and the perturbed variables.

IL:lm: $|(\bar{\hat{y}}_w^o - \bar{\hat{y}}_w^p)/\bar{\hat{y}}_w^o|$, with $\bar{\hat{y}}_w$ the (Horwitz-Thompson) weighted mean of exponentials of the fitted values from the model $log(earningsHour) \sim age + Location + Sex + education + Occupation + economicActivity + Length + Size)$ (using weighted least squares estimation considering the sampling weights) obtained from the original (index o) and the perturbed data (index p).

Table 1 let us to the following interpretation. The original unmodified SES data contains about 5.35 % of observations that violate 3-anonymity and about 2.48 % of risky observations (using the individual risk approach). For the original data, the global model-based risk is 0.83 (and 1.35) which is quite similar to the percentage of observations having high dis suda score (0.87). Of course, the risk on continuous key variables is 100 % and the information loss on that variables is zero. When recoding *economic activity* into less categories, the risk reduces by almost the factor of 5. When additionally recoding the variable *age* the risk reduces dramatically. After applying local suppression additionally, the risk for all risk methods zero, expect the individual risk.

The risk on continuous variables is evaluated for any method independently. It is very low for adding additive noise to the data but in the same time the information loss is inacceptable large. The information loss is very small for adding correlated noise, but the risk is still high. For microaggregation, the information loss is (almost) zero, but the risk is high. However, always three observations are aggregated and therefore anonymisation might be fine but the disclosure risk method is not suitable for microaggregation. The performance of shuffling is good, but the model based estimates differ more than 8 % after shuffling the data.

Probably the most interesting information loss measure - the measure which accounts for fitting a linear model on the data (IL:lm) reports that the information loss very low expect for the adding additive noise method and shuffling.

4 Conclusion

In this contribution, popular disclosure risk methods have been summarized. We stressed to measure the disclosure risk after the application of any SDC method to the data. Because of the limit of pages we only briefly focused on measuring the data utility and information loss, but it should be clear that the aim is both, to provide a data set with low disclosure risk and high data utility.

In the practical example, a very popular data set was used and the disclosure risk and data utility/information loss is evaluated. Hereby, the whole range of disclosure

risk methods has been applied to the data, which is done the first time to our knowledge. The results show that by application of few selected anonymisation methods, the disclosure risk dramatically decreases and in the same time, the information loss is considerable small.

All estimations/calculations have been made with the R-package **sdcMicro** Templ et al. [2012]. The SES data were provided by Statistics Austria.

References

J. Domingo-Ferrer and J.M. Mateo-Sanz. Practical data-oriented microaggregation for statistical disclosure control. *IEEE Trans. on Knowledge and Data Engineering*, 14 (1):189–201, 2002.

J. Domingo-Ferrer and V. Torra. A quantitative comparison of disclosure control methods for microdata. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 111–134, 2001.

M. Elliot. DIS: A new approach to the measurement of statistical disclosure risk. *Risk Management*, 2(4):39–48, 2000.

L. Franconi and S. Polettini. Individual risk estimation in μ-Argus: a review. In J. In: Domingo-Ferrer, editor, *Privacy in Statistical Databases, Lecture Notes in Computer Science*, pages 262–272. Springer, 2004.

T. Geissberger. *Verdienststrukturerhebung 2006, Struktur und Verteilung der Verdienste in Österreich*. Statistik Austria, 2009. ISBN 978-3-902587-97-8.

C. Gini. Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche. *Studi Economico-Giuridici della R. Università di Cagliari*, 3: 3–159, 1912.

E. Groshen. The structure of the female/male wage differential. *Journal of Human Resources*, 26:455–472, 1991.

A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1 (1), March 2007. ISSN 1556-4681. doi: 10.1145/1217299.1217302. URL `http://doi.acm.org/10.1145/1217299.1217302`.

A. Manning, D. Haglin, and J. Keane. A recursive search algorithm for statistical disclosure assessment. *Data Mining and Knowledge Discovery*, 16:165–196, 2008. ISSN 1384-5810. URL `http://dx.doi.org/10.1007/s10618-007-0078-6`. 10.1007/s10618-007-0078-6.

J.M. Mateo-Sanz, F. Sebe, and J. Domingo-Ferrer. Outlier protection in continuous microdata masking. *Lecture Notes in Computer Science, Vol. Privacy in Statistical Databases, Springer Verlag*, 3050:201–215, 2004.

K. Muralidhar and R. Sarathy. Data shuffling- a new masking approach for numerical data. *Management Science*, 52(2):658–670, 2006.

Y. Rinott and N. Shlomo. A generalized negative binomial smoothing model for sample disclosure risk estimation. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 82–93, 2006.

P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI International, 1998.

CJ. Skinner and DJ. Holmes. Estimating the re-identification risk per record in microdata. *Journal of Official Statistics*, 14:361–372, 1998.

L. Sweeney. k-anonymity: a model for protecting privacy. *Int J Uncertain Fuzziness Knowl Syst*, 10(5):557–570, 2002.

M. Templ and B. Meindl. Robust statistics meets SDC: New disclosure risk measures for continuous microdata masking. *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, 5262:113–126, 2008. ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3_10.

M. Templ and B. Meindl. Practical applications in statistical disclosure control using R. In J. Nin and J. Herranz, editors, *Privacy and Anonymity in Information Management Systems*, Advanced Information and Knowledge Processing, pages 31–62. Springer London, 2010. ISBN 978-1-84996-238-4. URL `http://dx.doi.org/10.1007/978-1-84996-238-4_3`. 10.1007/978-1-84996-238-4_3.

M. Templ, A. Kowarik, and B. Meindl. *sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package.*, 2012. URL `http://CRAN.R-project.org/package=sdcMicro`. R package version 4.0.0.