INTERNATIONAL HOUSEHOLD SURVEY NETWORK

Introduction to Statistical Disclosure Control (SDC)

Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen

Introduction to Statistical Disclosure Control (SDC)

Matthias Templ, Bernhard Meindl, Alexander Kowarik and Shuang Chen

August 2014

IHSN Working Paper No 007

Acknowledgments

This paper (or a revised copy of it) is available on the web site of the International Household Survey Network at www.ihsn.org.

Citation

Templ, Matthias, Bernhard Meindl, Alexander Kowarik, and Shuang Chen. "Introduction to Statistical Disclosure Control (SDC)." IHSN Working Paper No. 007 (2014).

Table of Contents

List of tables

List of figures

Listing

1. Overview

To support research and policymaking, there is an increasing demand for microdata. Microdata are data that hold information collected on individual units, such as people, households or enterprises. For statistical producers, microdata dissemination increases returns on data collection and helps improve data quality and credibility. But statistical producers are also faced with the challenge of ensuring respondents' confidentiality while making microdata files more accessible. Not only are data producers obligated to protect confidentiality, but security is also crucial for maintaining the trust of respondents and ensuring the honesty and validity of their responses.

Proper and secure microdata dissemination requires statistical agencies to establish policies and procedures that formally define the conditions for accessing microdata (Dupriez and Boyko, 2010), and to apply statistical disclosure control (SDC) methods to data before release. This guide, *Introduction to Statistical Disclosure Control (SDC)*, discusses common SDC methods for microdata obtained from sample surveys, censuses and administrative sources.

1.1   How to Use This Guide

This guide is intended for statistical producers at National Statistical Offices (NSOs) and other statistical agencies, as well as data users who are interested in the subject. It assumes no prior knowledge of SDC. The guide is focused on SDC methods for microdata. It does not cover SDC methods for protecting tabular outputs (see Castro 2010 for more details).

The guide starts with an introduction to the basic concepts regarding statistical disclosure in Section 2. Section 3 discusses methods for measuring disclosure risks. Section 4 presents the most common SDC methods, followed by an introduction to common approaches for assessing information loss and data utility in Section 5. Section 6 provides practical guidelines on how to implement SDC. Section 7 uses a sample dataset to illustrate the primary concepts and procedures introduced in this guide.

All the methods introduced in this guide can be implemented using sdcMicroGUI, an R-based, user-friendly application (Kowarik et al., 2013) and/or the more advanced R-Package, sdcMicro (Templ et al., 2013). Readers are encouraged to explore them using this guide along with the detailed user manuals of sdcMicroGUI (Templ et al., 2014b) and sdcMicro (Templ et al., 2013). Additional case studies of how to implement SDC on specific datasets are also available; see Templ et al. 2014a.

2. Concepts

This section introduces the basic concepts related to statistical disclosure, SDC methods and the trade-off between disclosure risks and information loss.

2.1   What is Disclosure

Suppose a hypothetical intruder has access to some released microdata and attempts to identify or find out more information about a particular respondent. Disclosure, also known as "re-identification," occurs when the intruder reveals previously unknown information about a respondent by using the released data. Three types of disclosure are noted here (Lambert, 1993):

- **Identity disclosure** occurs if the intruder associates a known individual with a released data record. For example, the intruder links a released data record with external information, or identifies a respondent with extreme data values. In this case, an intruder can exploit a small subset of variables to make the linkage, and once the linkage is successful, the intruder has access to all other information in the released data related to the specific respondent.

- **Attribute disclosure** occurs if the intruder is able to determine some new characteristics of an individual based on the information available in the released data. For example, if a hospital publishes data showing that all female patients aged 56 to 60 have cancer, an intruder then knows the medical condition of any female patient aged 56 to 60 without having to identify the specific individual.

- **Inferential disclosure** occurs if the intruder is able to determine the value of some characteristic of an individual more accurately with the released data than otherwise would have been possible. For example, with a highly predictive regression model, an intruder may be able to infer a respondent's sensitive income

information using attributes recorded in the data, leading to inferential disclosure.

2.2 Classifying Variables

2.2.1 Identifying variables

SDC methods are often applied to *identifying variables* whose values might lead to re-identification. Identifying variables can be further classified into direct identifiers and key variables:

- **Direct identifiers** are variables that un-ambiguously identify statistical units, such as social insurance numbers, or names and addresses of companies or persons. Direct identifiers should be removed as the first step of SDC.

- **Key variables** are a set of variables that, in combination, can be linked to external information to re-identify respondents in the released dataset. Key variables are also called "implicit identifiers" or "quasi-identifiers". For example, while on their own, the gender, age, region and occupation variables may not reveal the identity of any respondent, but in combination, they may uniquely identify respondents.

2.2.2 Sensitive variables

SDC methods are also applied to *sensitive variables* to protect confidential information of respondents. Sensitive variables are those whose values must not be discovered for any respondent in the dataset. The determination of sensitive variables is often subject to legal and ethical concerns. For example, variables containing information on criminal history, sexual behavior, medical records or income are often considered sensitive. In some cases, even if identity disclosure is prevented, releasing sensitive variables can still lead to attribute disclosure (see example in Section 3.2).

A variable can be both identifying and sensitive. For example, income variable can be combined with other key variables to re-identify respondents, but the variable itself also contains sensitive information that should be kept confidential. On the other hand, some variables, such as occupation, might not be sensitive, but could be used to re-identify respondents when combined with other variables. In this case, occupation

is a key variable, and SDC methods should be applied to it to prevent identity disclosure.

2.2.3 Categorical vs. continuous variables

SDC methods differ for categorical variables and continuous variables. Using the definitions in Domingo-Ferrer and Torra (2005), a *categorical* variable takes values over a finite set. For example, gender is a categorical variable. A *continuous* variable is numerical, and arithmetic operations can be performed with it. For example, income and age are continuous variables. A numerical variable does not necessarily have an infinite range, as in the case of age.

2.3 Disclosure Risk vs. Information Loss

Applying SDC techniques to the original microdata may result in information loss and hence affect data utility[1]. The main challenge for a statistical agency, therefore, is to apply the optimal SDC techniques that reduce disclosure risks with minimal information loss, preserving data utility. To illustrate the trade-off between disclosure risk and information loss, Figure 1 shows a general example of results after applying two different SDC methods to the European Union Structure of Earnings Statistics (SES) data (Templ et al., 2014a). The specific SDC methods and measures of disclosure risk and information loss will be explained in the following sections.

Before applying any SDC methods, the original data is assumed to have disclosure risk of 1 and information loss of 0. As shown in Figure 1, two different SDC methods are applied to the same dataset. The solid curve represents the first SDC method (i.e., adding noise; see Section 4.2.2). The curve illustrates that, as more noise is added to the original data, the disclosure risk decreases but the extent of information loss increases.

In comparison, the dotted curve, illustrating the result of the second SDC method (i.e., micro-aggregation; see Section 4.2.1) is much less steep than the solid curve representing the first method. In other words, at a given level of disclosure risk—for example, when disclosure risk is 0.1—the information loss resulting from the second method is much lower than that resulting from the first.

_____

1   Data utility describes the value of data as an analytical resource, comprising analytical completeness and analytical validity.

Therefore, for this specific dataset, Method 2 is the preferred SDC method for the statistical agency to reduce disclosure risk with minimal information loss. In Section 6, we will discuss in detail how to determine the acceptable levels of risk and information loss in practice.

Figure 1: **Disclosure risk versus information loss obtained from two specific SDC methods applied to the SES data**
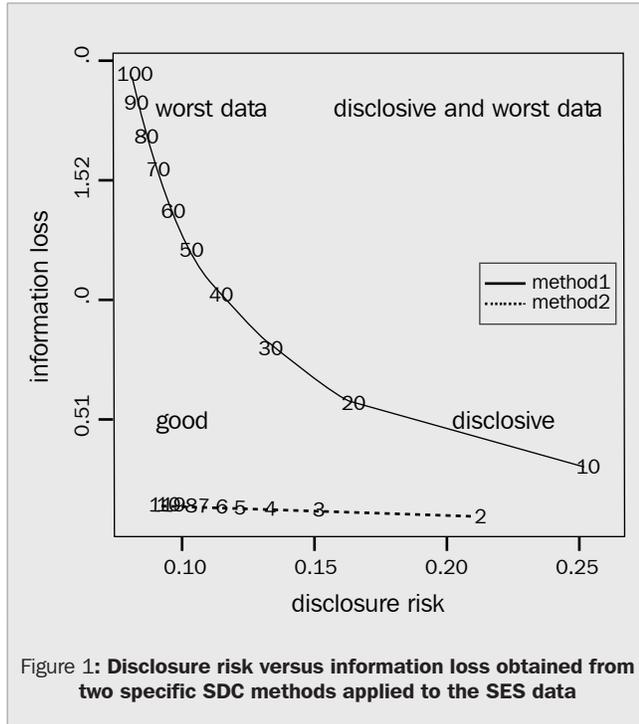
3. Measuring Disclosure Risk

Disclosure risks are defined based on assumptions of disclosure scenarios, that is, how the intruder might exploit the released data to reveal information about a respondent. For example, an intruder might achieve this by linking the released file with another data source that shares the same respondents and identifying variables. In another scenario, if an intruder knows that his/her acquaintance participated in the survey, he/she may be able to match his/her personal knowledge with the released data to learn new information about the acquaintance. In practice, most of the measures for assessing disclosure risks, as introduced below, are based on key variables, which are determined according to assumed disclosure scenarios.

3.1 Sample uniques, population uniques and record-level disclosure risk

Disclosure risks of categorical variables are defined based on the idea that records with unique combinations of key variable values have higher risks of re-identification (Skinner and Holmes, 1998; Elamir and Skinner, 2006). We call a combination of values of an assumed set of key variables a pattern, or key value. Let f_k be the frequency counts of records with pattern k in the sample. A record is called a *sample unique* if it has a pattern k for which $f_k = 1$. Let F_k denote the number of units in the population having the same pattern k. A record is called a *population unique* if $F_k = 1$.

In Table 1, a very simple dataset is used to illustrate the concept of sample frequency counts and sample uniques. The sample dataset has eight records and four pre-determined key variables (i.e., Age group, Gender, Income and Education). Given the four key variables, we have six distinct patterns, or key values. The sample frequency counts of the first and second records equal 2 because the two records share the same pattern (i.e., {20s, Male, >50k, High school}). Record 5 is a sample unique because it is the only individual in the sample who is a female in her thirties earning less than 50k with a university degree. Similarly, records 6, 7 and 8 are sample uniques, because they possess distinct patterns with respect to the four key variables.

Table 1: Example of frequency count, sample uniques and record-level disclosure risks estimated with a Negative Binomial model

	Age group	Gender	Income	Education	f_k	Sampling weights	Risk
1	20s	Male	>50k	High school	2	18	0.017
2	20s	Male	>50k	High school	2	92	0.017
3	20s	Male	≤50k	High school	2	45.5	0.022
4	20s	Male	≤50k	High school	2	39	0.022
5	30s	Female	≤50k	University	1	17	0.177
6	40s	Female	≤50k	High school	1	8	0.297
7	40s	Female	≤50k	Middle school	1	541	0.012
8	60s	Male	≤50k	University	1	5	0.402

Consider a sample unique with $f_k = 1$. Assuming no measurement error, there are F_k units in the population that could potentially match the record in the sample. The probability that the intruder can match the sample unique with the individual in the population is thus $1/F_k$ assuming that the intruder does not know if the individual in the population is a respondent in the sample or not. The disclosure risk for the sample unique

is thus defined as the expected value of $1/F_k$, given $f_k = 1$. More generally, the *record-level disclosure risk* for any given record is defined as the expected value of $1/F_k$, given f_k.

In practice, we observe only the sample frequency counts f_k. To estimate the record-level disclosure risks, we take into account the sampling scheme and make inferences on F_k assuming that F_k follows a generalized Negative Binomial distribution (Rinott and Shlomo, 2006; Franconi and Polettini, 2004).

3.2 Principles of k-anonymity and l-diversity

Assuming that sample uniques are more likely to be re-identified, one way to protect confidentiality is to ensure that each distinct pattern of key variables is possessed by at least *k* records in the sample. This approach is called achieving *k*-anonymity (Samarati and Sweeney, 1998; Sweeney, 2002). A typical practice is to set *k* = 3, which ensures that the same pattern of key variables is possessed by at least three records in the sample. Using the previous notation, 3-anonymity means $f_k \geq 3$ for all records. By this definition, all records in the previous example (Table 1) violate 3-anonymity.

Even if a group of observations fulfill *k*-anonymity, an intruder can still discover sensitive information. For example, Table 2 satisfies 3-anonymity, given the two key variables *gender* and *age group*. However, suppose an intruder gets access to the sample inpatient records, and knows that his neighbor, a girl in her twenties, recently went to the hospital. Since all records of females in their twenties have the same medical condition, the intruder discovers with certainty that his neighbor has cancer. In a different scenario, if the intruder has a male friend in his thirties who belongs to one of the first three records, the intruder knows that the incidence of his friend having heart disease is low and thus concludes that his friend has cancer.

Table 2: Example inpatient records illustrating k-anonymity and l-diversity					
	Key variables		f_k	Sensitive variable	Distinct l-diversity
	Gender	Age group		Medical condition	
1	Male	30s	3	Cancer	2
2	Male	30s	3	Heart Disease	2
3	Male	30s	3	Heart Disease	2
4	Female	20s	3	Cancer	1
5	Female	20s	3	Cancer	1
6	Female	20s	3	Cancer	1

To address this limitation of *k*-anonymity, the *l*-diversity principle (Machanavjjhala et al., 2007) was

introduced as a stronger notion of privacy: A group of observations with the same pattern of key variables is *l*-diverse if it contains at least *l* "well-represented" values for the sensitive variable. Machanavajjhala et al. (2007) interpreted "well-represented" in a number of ways, and the simplest interpretation, *distinct l-diversity*, ensures that the sensitive variable has at least *l* distinct values for each group of observations with the same pattern of key variables. As shown in Table 2, the first three records are 2-diverse because they have two distinct values for the sensitive variable, *medical condition*.

3.3 Disclosure risks for hierarchical data

Many micro-datasets have hierarchical, or multilevel, structures; for example, individuals are situated in households. Once an individual is re-identified, the data intruder may learn information about the other household members, too. It is important, therefore, to take into account the hierarchical structure of the dataset when measuring disclosure risks.

It is commonly assumed that the disclosure risk for a household is greater than or equal to the risk that at least one member of the household is re-identified. Thus household-level disclosure risks can be estimated by subtracting the probability that no person from the household is re-identified from one. For example, if we consider a single household of three members, whose individual disclosure risks are 0.1, 0.05 and 0.01, respectively, the disclosure risk for the entire household will be calculated as $1 - (1-0.1) \times (1-0.05) \times (1-0.01) = 0.15355$.

3.4 Measuring global risks

In addition to record-level disclosure risk measures, a risk measure for the entire file-level or global risk micro-dataset might be of interest. In this section, we present three common measures of global risks:

- **Expected number of re-identifications.** The easiest measure of global risk is to sum up the record-level disclosure risks (defined in Section 3.1), which gives the expected number of re-identifications. Using the example from Table 1, the expected number of re-identifications is 0.966, the sum of the last column.

- **Global risk measure based on log-linear models.** This measure, defined as the number

of sample uniques that are also population uniques, is estimated using standard log-linear models (Skinner and Holmes, 1998; Ichim, 2008). The population frequency counts, or the number of units in the population that possess a specific pattern of key variables observed in the sample, are assumed to follow a Poisson distribution. The global risk can then be estimated by a standard log-linear model, using the main effects and interactions of key variables. A more precise definition is available in Skinner and Holmes 1998.

- **Benchmark approach.** This measure counts the number of observations with record-level risks higher than a certain threshold and higher than the main part of the data. While the previous two measures indicate an overall re-identification risk for a microdata file, the benchmark approach is a relative measure that examines whether the distribution of record-level risks contains extreme values. For example, we can identify the number of records with individual risk r_i satisfying the following conditions:

$$r_i \geq 0.1 \text{ and } r_i \geq 2 \cdot [\text{median}(\mathbf{r}) + 2 \cdot \text{MAD}(\mathbf{r})]$$

Where $\mathbf{r}$ represents all record-level risks, and $\text{MAD}(\mathbf{r})$ is the median absolute deviation of all record-level risks.

3.5 Special Uniques Detection Algorithm (SUDA)

An alternative approach to defining disclosure risks is based on the concept of special uniqueness. For example, the eighth record in Table 1 is a sample unique with respect to the key variable set {Age group, Gender, Income, Education}. Furthermore, a subset of the key variable set, for example, {Male, University}, is also unique in the sample. A record is defined as a *special unique* with respect to a variable set K, if it is a sample unique both on K and on a subset of K (Elliot et al., 1998). Research has shown that special uniques are more likely to be population uniques than random uniques (Elliot et al., 2002).

A set of computer algorithms, called SUDA, was designed to comprehensively detect and grade special uniques (Elliot et al., 2002). SUDA takes a two-step approach. In the first step, all unique attribute sets (up to a user-specified size) are located at record level. To streamline the search process, SUDA considers only

Minimal Sample Uniques (MSUs), which are unique attribute sets without any unique subsets within a sample. In the example presented in Table 3, {Male, University} is a MSU of record 8 because none of its subsets, {Male} or {University}, is unique in the sample. Whereas, {60s, Male, ≤50k, University} is a unique attribute set, but not a MSU because its subsets {60s, Male, University} and {Male, University} are both unique subsets in the sample.

Once all MSUs have been found, a SUDA score is assigned to each record indicating how "risky" it is, using the size and distribution of MSUs within each record (Elliot et al., 2002). The potential risk of the records is determined based on two observations: 1) the smaller the size of the MSU within a record, the greater the risk of the record, and 2) the larger the number of MSUs possessed by the record, the greater the risk of the record.

For each MSU of size k contained in a given record, a score is computed by $\prod_{i=k}^{M}(ATT - i)$, where M is the user-specified maximum size of MSUs, and ATT is the total number of attributes in the dataset. By definition, the smaller the size k of the MSU, the larger the score for the MSU.

The final SUDA score for the record is computed by adding the scores for each MSU. In this way, records with more MSUs are assigned a higher SUDA score.

To illustrate how SUDA scores are calculated, record 8 in Table 3 has two MSUs: {60s} of size 1, and {Male, University} of size 2. Suppose the maximum size of MSUs is set at 3, the score assigned to {60s} is computed by $\prod_{i=1}^{3}(4 - i) = 6$, and the score assigned to {Male, University} is $\prod_{i=2}^{3}(4 - i) = 2$ The SUDA score for the eighth record in Table 3 is then 8.

	Age group	Gender	Income	Education	f_k	SUDA score	Risk using DIS-SUDA method
				Table 3: Example dataset illustrating SUDA scores			
1	20s	Male	>50k	High school	2	0	0.00
2	20s	Male	>50k	High school	2	0	0.00
3	20s	Male	≤50k	High school	2	0	0.00
4	20s	Male	≤50k	High school	2	0	0.00
5	30s	Female	≤50k	University	1	8	0.0149
6	40s	Female	≤50k	High school	1	4	0.0111
7	40s	Female	≤50k	Middle school	1	6	0.0057
8	60s	Male	≤50k	University	1	8	0.0149

In order to estimate record-level disclosure risks, SUDA scores can be used in combination with the Data Intrusion Simulation (DIS) metric (Elliot and Manning, 2003), a method for assessing disclosure risks for the entire dataset (i.e., file-level disclosure risks). Roughly speaking, the DIS-SUDA method distributes the file-level risk measure generated by the DIS metric between records according to the SUDA scores of each record. This way, SUDA scores are calibrated against a consistent measure to produce the DIS-SUDA scores, which provide the record-level disclosure risk. A full description of the DIS-SUDA method is provided by Elliot and Manning (2003).

Both SUDA and DIS-SUDA scores can be computed using sdcMicro (Templ et al., 2013). Given that the implementation of SUDA can be computationally demanding, sdcMicro uses an improved SUDA2 algorithm, which more effectively locates the boundaries of the search space for MSUs in the first step (Manning et al., 2008).

Table 3 presents the record-level risks estimated using the DIS-SUDA approach for the sample dataset. Compared to the risk measures presented in Table 1, the DIS-SUDA score (Table 3) does not fully account for the sampling weights, while the risk measures based on negative binomial model (Table 1) are lower for records with greater sampling weights, given the same sample frequency count. Therefore, instead of replacing the risk measures introduced in Section 3.1, the SUDA scores and DIS-SUDA approach can be best used as a complementary method.

3.6 Record Linkage

The concept of uniqueness might not be applicable to continuous key variables, especially those with an infinite range, since almost every record in the dataset will then be identified as unique. In this case, a more applicable method is to assess risk based on record linkages.

Assume a disclosure scenario where an intruder has access to a dataset that has been perturbed before release, as well as an external data source that contains information on the same respondents included in the released dataset. The intruder attempts to match records in the released dataset with those in the external dataset using common variables. Suppose that the external data source, to which the intruder has access, is the original data file of the released dataset. Essentially, the record linkage approach assesses to what extent records in the perturbed data file can be correctly matched with those in the original data file. There are three general approaches to record linkage:

- **Distance-based record linkage** (Pagliuca and Seri, 1999) computes distances between records in the original dataset and the protected dataset. Suppose we have obtained a protected dataset A' after applying some SDC methods to the original dataset A. For each record r in the protected dataset A', we compute its distance to every record in the original dataset, and consider the nearest and the second nearest records. Suppose we have identified r_1 and r_2 from the original dataset as the nearest and second-nearest records, respectively, to record r. If r_1 is the original record used to generate r, or, in other words, record r in the protected dataset and r_1 in the original dataset refer to the same respondent, then we mark record r "linked". Similarly, if record r was generated from r_2 (the second-nearest record in the original dataset), we mark r "linked to the 2nd nearest". We proceed the same way for every record in the protected dataset A'. Finally, disclosure risk is defined as the percentage of records marked as "linked" or "linked to the 2nd nearest" in the protected dataset A'. This record-linkage approach based on distance is compute-intensive and thus might not be applicable for large datasets.

- Alternatively, **probabilistic record linkage** (Jaro, 1989) pairs records in the original and protected datasets, and uses an algorithm to assign a weight for each pair that indicates the likelihood that the two records refer to the same respondent. Pairs with weights higher than a specific threshold are labeled as "linked", and the percentage of records in the protected data marked as "linked" is the disclosure risk.

- In addition, a third risk measure is called **interval disclosure** (Paglica and Seri, 1999), which simplifies the distance-based record linkage and thus is more applicable for large datasets. In this approach, after applying SDC methods to the original values, we construct an interval around each masked value. The width of the interval is based on the rank of the value the variable takes on or its standard deviation. We then examine whether the original value of the variable falls within the interval. The

measure of disclosure risk is the proportion of original values that fall into the interval.

3.7 Special Treatment of Outliers

Almost all datasets used in official statistics contain records that have at least one variable value quite different from the general observations. Examples of such outliers might be enterprises with a very high value for turnover or persons with extremely high income. Unfortunately, intruders may want to disclose a statistical unit with "special" characteristics more than those exhibiting the same behavior as most other observations. We also assume that the further away an observation is from the majority of the data, the easier the re-identification. For these reasons, Templ and Meindl (2008a) developed a disclosure risk measures that take into account the "outlying-ness" of an observation.

The algorithm starts with estimating a Robust Mahalanobis Distance (RMD) (Maronna et al., 2006) for each record. Then intervals are constructed around the original values of each record. The length of the intervals is weighted by the squared RMD; the higher the RMD, the larger the corresponding interval. If, after applying SDC methods, the value of the record falls within the interval around its original value, the record is marked "unsafe". One approach, RMDID1, obtains the disclosure risk by the percentage of records that are unsafe. The other approach, RMDID2, further checks if the record marked unsafe has close neighbors; if m other records in the masked dataset are very close (by Euclidean distances) to the unsafe record, the record is considered safe.

4. Common SDC Methods

There are three broad kinds of SDC techniques: i) non-perturbative techniques, such as recoding and local suppression, which suppress or reduce the detail without altering the original data; ii) perturbative techniques, such as adding noise, Post-Randomization Method (PRAM), micro-aggregation and shuffling, which distort the original micro-dataset before release; and iii) techniques that generate a synthetic microdata file that preserves certain statistics or relationships of the original files.

This guide focuses on the non-perturbative and perturbative techniques. Creating synthetic data is a more complicated approach and out of scope for this guide (see Drechsler 2011, Alfons et al. 2011, Templ and

Filzmoser 2013 for more details). As with disclosure risk measures, different SDC methods are applicable to categorical variables versus continuous variables.

4.1 Common SDC Methods for Categorical Variables

4.1.1 Recoding

Global recoding is a non-perturbative method that can be applied to both categorical and continuous key variables. For a categorical variable, the idea of recoding is to combine several categories into one with a higher frequency count and less information. For example, one could combine multiple levels of schooling (e.g., secondary, tertiary, postgraduate) into one (e.g., secondary and above). For a continuous variable, recoding means to discretize the variable; for example, recoding a continuous income variable into a categorical variable of income levels. In both cases, the goal is to reduce the total number of possible values of a variable.

Typically, recoding is applied to categorical variables to collapse categories with few observations into a single category with larger frequency counts. For example, if there are only two respondents with tertiary level of education, tertiary can be combined with the secondary level into a single category of "secondary and above".

A special case of global recoding is top and bottom coding. Top coding sets an upper limit on all values of a variable and replaces any value greater than this limit by the upper limit; for example, top coding would replace the age value for any individual aged above 80 with 80. Similarly, bottom coding replaces any value below a pre-specified lower limit by the lower limit; for example, bottom coding would replace the age value for any individual aged under 5 with 5.

4.1.2 Local suppression

If unique combinations of categorical key variables remain after recoding, local suppression could be applied to the data to achieve k-anonymity (described in Section 3.2). Local suppression is a non-perturbative method typically applied to categorical variables. In this approach, missing values are created to replace certain values of key variables to increase the number of records sharing the same pattern, thus reducing the record-level disclosure risks.

There are two approaches to implementing local suppression. One approach sets the parameter k and

tries to achieve *k*-anonymity (typically 3-anonymity) with minimum suppression of values. For example, in sdcMicroGUI (Templ et al., 2014b), the user sets the value for *k* and orders key variables by the likelihood they will be suppressed. Then the application calls a heuristic algorithm to suppress a minimum number of values in the key variables to achieve *k*-anonymity. The second approach sets a record-level risk threshold. This method first identifies unsafe records with individual disclosure risks higher than the threshold and then suppresses all values of the selected key variable(s) for all the unsafe records.

4.1.3   Post-Randomization Method PRAM)

If there are a larger number of categorical key variables (e.g., more than 5), recoding might not sufficiently reduce disclosure risks, or local suppression might lead to great information loss. In this case, the PRAM (Gouweleeuw et al., 1998) may be a more efficient alternative.

PRAM (Gouweleeuw et al., 1998) is a probabilistic, perturbative method for protecting categorical variables. The method swaps the categories for selected variables based on a pre-defined transition matrix, which specifies the probabilities for each category to be swapped with other categories.

To illustrate, consider the variable *location*, with three categories: $location = 1$ "east", $location = 2$ "middle", $location = 3$ "west". We define a 3-by-3 transition matrix, where p_{ij} is the probability of changing category i to j. For example, in the following matrix,

$$\mathbf{P} = \begin{pmatrix} 0.1 & 0.9 & 0 \\ 0.2 & 0.1 & 0.7 \\ 0.9 & 0 & 0.1 \end{pmatrix}$$

the probability that the value of the variable will stay the same after perturbation is 0.1, since we set $p_{11} = p_{22} = p_{33} = 0.1$. The probability of *east* being changed into *middle* is $p_{12} = 0.9$, while *east* will not be changed into *west* because p_{13} is set to be 0.

PRAM protects the records by perturbing the original data file, while at the same time, since the probability mechanism used is known, the characteristics of the original data can be estimated from the perturbed data file.

PRAM can be applied to each record independently, allowing the flexibility to specify the transition matrix

as a function parameter according to desired effects. For example, it is possible to prohibit changes from one category to another by setting the corresponding probability in the transition matrix to 0, as shown in the example above. It is also possible to apply PRAM to subsets of the microdata independently.

4.2   Common SDC Methods for Continuous Variables

4.2.1   Micro-aggregation

Micro-aggregation (Defays and Anwar, 1998) is a perturbing method typically applied to continuous variables. It is also a natural approach to achieving *k*-anonymity. The method first partitions records into groups, then assigns an aggregate value (typically the arithmetic mean, but other robust methods are also possible) to each variable in the group.

As an example, in Table 4, records are first partitioned into groups of two, and then the values are replaced by the group means. Note that in the example, by setting group size to two, micro-aggregation automatically achieves 2-anonymity with respect to the three key variables.

To preserve the multivariate structure of the data, the most challenging part of micro-aggregation is to group records by how "similar" they are. The simplest method is to sort data based on a single variable in ascending or descending order. Another option is to cluster data first, and sort by the most influential variable in each cluster (Domingo-Ferrer et al., 2002). These methods, however, might not be optimal for multivariate data (Templ and Meindl, 2008b).

The Principle Component Analysis method sorts data on the first principal components (e.g., Templ and Meindl, 2008b). A robust version of this method can be applied to clustered data for small- or medium-sized datasets (Templ, 2008). This approach is fast and performs well when the first principal component explains a high percentage of the variance for the key variables under consideration.

The Maximum Distance to Average Vector (MDAV) method is a standard method that groups records based on classical Euclidean distances in a multivariate space (Domingo-Ferrer and Mateo-Sanz, 2002). The MDAV method was further improved by replacing Euclidean distances with robust multivariate (Mahalanobis) distance measures (Templ and Meindl, 2008b). All of

Table 4: Example of micro-aggregation: var_1, var_2, var_3, are key variables containing original values. var_1', var_2', var_3', contain values after applying micro-aggregation.

	var_1	var_2	var_3
1	0.30	0.40	4.00
2	0.12	0.22	22.00
3	0.18	0.80	8.00
4	1.90	9.00	91.00
5	1.00	1.30	13.00
6	1.00	1.40	14.00
7	0.10	0.01	1.00
8	0.15	0.50	5.00

	var_1	var_2	var_3	var_1'	var_2'	var_3'
7	0.10	0.01	1.00	0.12	0.26	3.00
8	0.15	0.50	5.00	0.12	0.26	3.00
2	0.12	0.22	22.00	0.15	0.51	15.00
3	0.18	0.80	8.00	0.15	0.51	15.00
1	0.30	0.40	4.00	0.65	0.85	8.50
5	1.00	1.30	13.00	0.65	0.85	8.50
6	1.00	1.40	14.00	1.45	5.20	52.50
4	1.90	9.00	91.00	1.45	5.20	52.50

these methods can be implemented in sdcMicro (Templ et al., 2013) or sdcMicroGUI (Kowarik et al., 2013; Templ et al., 2014b).

4.2.2 Adding noise

Adding noise is a perturbative method typically applied to continuous variables. The idea is to add or multiply a stochastic or randomized number to the original values to protect data from exact matching with external files. While this approach sounds simple in principle, many different algorithms can be used. In this section, we introduce the uncorrelated and correlated additive noise (Brand, 2002; Domingo-Ferrer et al., 2004).

Uncorrelated additive noise can be expressed as the following:

$$z_j = x_j + \epsilon_j$$

where vector x_j represents the original values of variable j, z_j represents the perturbed values of variable j and ϵ_j (*uncorrelated noise*, or white noise) denotes normally distributed errors with $Cov(\epsilon_t, \epsilon_l) = 0$ for all $t \neq l$.

While adding uncorrelated additive noise preserves the means and variance of the original data, co-variances and correlation coefficients are not preserved. It is preferable to apply *correlated noise* because the co-variance matrix of the errors is proportional to the co-variance matrix of the original data (Brand, 2002; Domingo-Ferrer et al., 2004).

The distribution of the original variables x often differs and may not follow a normal distribution. In this case, a robust version of the correlated noise method is described in detail by Templ and Meindl (2008b). The

method of adding noise should be used with caution, as the results depend greatly on the parameters chosen.

4.2.3 Shuffling

Shuffling (Muralidhar and Sarathy, 2006) generates new values for selected sensitive variables based on the conditional density of sensitive variables given non-sensitive variables. As a rough illustration, assume we have two sensitive variables, *income* and *savings*, which contain confidential information. We first use age, occupation, race and education variables as predictors in a regression model to simulate a new set of values for *income* and *savings*. We then apply reverse mapping (i.e., shuffling) to replace ranked new values with the ranked original values for *income* and *savings*. This way, the shuffled data consists of the original values of the sensitive variables.

Muralidhar and Sarathy (2006) showed that, since we only need the *rank* of the perturbed value in this approach, shuffling can be implemented using only the rank-order correlation matrix (which measures the strength of the association between the ranked sensitive variables and ranked non-sensitive variables) and the ranks of non-sensitive variable values.

5. Measuring Information Loss

After SDC methods have been applied to the original dataset, it is critical to measure the resulting information loss. There are two complementary approaches to assessing information loss: (i) direct measures of distances between the original data and perturbed data, and (ii) the benchmarking approach comparing statistics computed on the original and perturbed data.

5.1 Direct Measures

Direct measures of information loss are based on the classical or robust distances between original and perturbed values. Following are three common definitions:

- **IL1s**, proposed by Yancey, Winkler and Creecy (2002), can be interpreted as the scaled distances between original and perturbed values. Let $X = \{x_{ij}\}$ be the original dataset, $X' = \{x'_{ij}\}$ is a perturbed version of X, and x_{ij} is the j-th variable in the i-th original record. Suppose both datasets consist of n records and p variables each. The measure of information loss is defined by

$$\text{IL1s} = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij}-x'_{ij}|}{\sqrt{2}S_j}$$

where S_j is the standard deviation of the j-th variable in the original dataset.

- A second measure is the relative absolute **differences between eigenvalues** of the co-variances from standardized original and perturbed values of continuous key variables (e.g., Templ and Meindl, 2008b). Eigenvalues can be estimated from a robust or classical version of the co-variance matrix.

- **lm** measures the differences between estimates obtained from fitting a pre-specified regression model on the original data and the perturbed data:

$$\text{lm} = \left|\frac{\bar{\bar{y}}_w - \bar{\bar{y}}'_w}{\bar{\bar{y}}_w}\right|$$

where $\bar{\bar{y}}_w$ denotes the estimated values using the original data, $\bar{\bar{y}}'_w$ the estimated values using the perturbed data. Index w indicates that the survey weights should be taken into account when fitting the model.

5.2 Benchmarking Indicators

Although, in practice, it is not possible to create a file with the exact same structure as the original file after applying SDC methods, an important goal of SDC should be to minimize the difference in the statistical properties of the perturbed data and the original data.

Using this assumption, an approach to measuring data utility is based on benchmarking indicators (Ichim and Franconi, 2010; Templ, 2011).

The first step to this approach is to determine what kind of analysis might be conducted using the released data and to identify the most important or relevant estimates (i.e., *benchmarking indicators*), including indicators that refer to the sensitive variables in the dataset.

After applying SDC methods to the original data and obtaining a protected dataset, assessment of information loss proceeds as follows:

1. Select a set of benchmarking indicators

2. Estimate the benchmarking indicators using the original microdata

3. Estimate the benchmarking indicators using the protected microdata

4. Compare statistical properties such as point estimates, variances or overlaps in confidence intervals for each benchmarking indicator

5. Assess whether the data utility of the protected micro-dataset is high enough for release

Alternatively, for Steps 2 and 3, we can fit a regression model on the original and modified microdata respectively and assess and compare statistical properties, such as coefficients and variances. This idea is similar to the information loss measure lm described in Section 5.1.

If Step 4 shows that the main indicators calculated from the protected data differ significantly from those estimated from the original dataset, the SDC procedure should be restarted. It is possible to either change some parameters of the applied methods or start from scratch and completely change the choice of SDC methods.

The benchmarking indicator approach is usually applied to assess the impact of SDC methods on continuous variables. But it is also applicable to categorical variables. In addition, the approach can be applied to subsets of the data. In this case, benchmarking indicators are evaluated for each of the subsets and the results are evaluated by reviewing differences between indicators for original and modified data within each subset.

6. Practical Guidelines

This section offers some guidelines on how to implement SDC methods in practice. Figure 2 presents a rough representation of a common workflow for applying SDC. [2]

Pre-processing steps are crucial, including discussing possible disclosure scenarios, selecting direct identifiers, key variables and sensitive variables, as well as determining acceptable disclosures risks and levels of information loss.

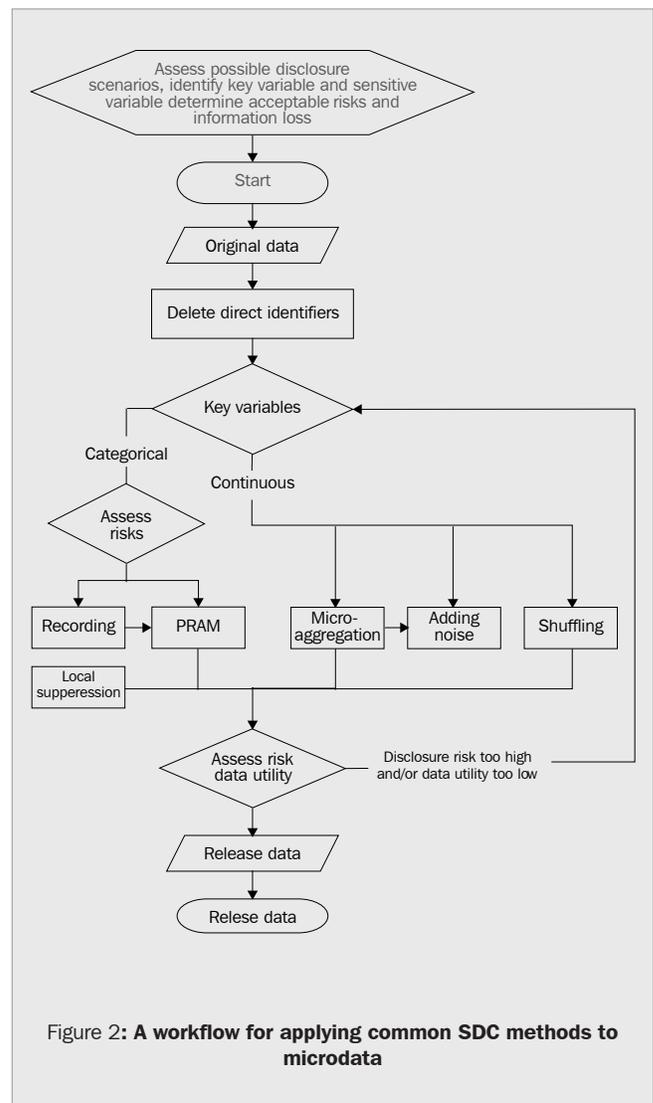The actual SDC process starts with deleting direct identifiers.

For categorical key variables, before applying any SDC techniques, measure the disclosure risks of the original data, including record-level and global disclosure risks, and identify records with high disclosure risks, such as those violating *k*-anonymity (typically 3-anonymity). Every time an SDC technique is applied, compare the same disclosure risk measures and assess the extent of information loss (for example, how many values have been suppressed or categories combined).

For continuous key variables, disclosure risks are measured by the extent to which the records in the perturbed dataset that can be correctly matched with those in the original data. Therefore, the disclosure risk is by default 100% for the original dataset. After applying any SDC method, disclosure risk measures are based on record linkage approaches introduced in Section 3.6. The risk measure should be compared and assessed together with information loss measures, such as IL1s and differences in eigenvalues introduced in Section 5.1.

For both categorical and continuous key variables, information loss should be quantified not only by direct measures, but also by examining benchmarking indicators. Data are ready to be released when an acceptable level of disclosure risk has been achieved with minimal information loss. Otherwise, alternative SDC techniques should be applied and/or the same techniques should be repeated with different parameter settings.

In this section, we provide some practical guidelines on common questions, such as how to determine

Figure 2: **A workflow for applying common SDC methods to microdata**

key variables and assess levels of risks, and how to determine which SDC methods to apply.

6.1 How to Determine Key Variables

Most disclosure risk assessment and SDC methods rely on the selected key variables, which correspond to certain disclosure scenarios. In practice, determining key variables is a challenge, as there are no definite rules and any variable potentially belongs to key variables, depending on the disclosure scenario. The recommended approach is to consider multiple disclosure scenarios and discuss with subject matter specialists which scenario is most likely and realistic.

---

2   Note that the figure only includes SDC methods introduced in this guideline, excluding methods such as simulation of synthetic data.

A common scenario is where the intruder links the released data with external data sources. Therefore, an important pre-processing step is to take inventory of what other data sources are available and identify variables which could be exploited to link to the released data. In addition, sensitive variables containing confidential information should also be identified beforehand.

6.2 What is an Acceptable Level of Disclosure Risk versus Information Loss

Assessment of data utility, especially the benchmarking indicators approach, requires knowledge of who the main users of the data are, how they will use the released data and, as a result, what information must be preserved. If a microdata dissemination policy exists, the acceptable level of risk varies for different types of files and access conditions (Dupriez and Boyko, 2010). For example, public use files should have much lower disclosure risks than licensed files whose access is restricted to specific users subject to certain terms and conditions.

Moreover, a dataset containing sensitive information, such as medical conditions, might require a larger extent of perturbation, compared to that containing general, non-sensitive information.

6.3 Which SDC Methods Should be Used

The strength and weakness of each SDC method are dependent on the structure of the dataset and key variables under consideration. The recommended approach is to apply different SDC methods with varying parameter settings in an exploratory manner. Documentation of the process is thus essential to make comparisons across methods and/or parameters and to help data producers decide on the optimal levels of information loss and disclosure risk. The following paragraphs provide general recommendations.

For categorical key variables, recoding is the most commonly used, non-perturbative method. If the disclosure risks remain high after recoding, apply local suppression to further reduce the number of sample uniques. Recoding should be applied in such a way so that minimal local suppression is needed afterwards.

If a dataset has large number of categorical key variables and/or a large number of categories for the given key variables (e.g., location variables), recoding and suppression might lead to too much information loss. In these situations, PRAM might be a more advantageous approach. PRAM can be applied with or without prior recoding. If PRAM is applied after recoding, the transition matrix should specify lower probability of swapping.

In addition, for sensitive variables violating *l*-diversity, recoding and PRAM are useful methods for increasing the number of distinct values of sensitive variables for each group of records sharing the same pattern of key variables.

For continuous variables, micro-aggregation is a recommended method. For more experienced users, shuffling provides promising results if there is a well-fitting regression model that predicts the values of sensitive variables using other variables present in the dataset (Muralidhar and Sarathy, 2006).

7. An Example Using SES Data

In this section, we use the 2006 Austrian data from the European Union SES to illustrate the application of main concepts and procedures introduced above. Additional case studies are available in Templ et al. 2014a. The SES is conducted in 28 member states of the European Union as well as in candidate countries and countries of the European Free Trade Association (EFTA). It is a large enterprise sample survey containing information on remuneration, individual characteristics of employees (e.g., gender, age, occupation, education level, etc.) and information about their employer (e.g., economic activity, size and location of the enterprise, etc.). Enterprises with at least 10 employees in all areas of the economy except public administration are sampled in the survey.

In Austria, a two-stage sampling is used: in the first stage, a stratified sample of enterprises and establishments is drawn based on economic activities and size, with large-sized enterprises having higher probabilities of being sampled. In the second stage, systematic sampling of employees is applied within each enterprise. The final sample includes 11,600 enterprises and 199,909 employees.

The dataset includes enterprise-level information (e.g., public or private ownership, types of collective agreement), employee-level information (e.g., start date of employment, weekly working time, type of work agreement, occupation, time for holidays, place of work,

gross earning, earnings for and amount of overtime, etc.), and information from registers (e.g., age, occupation, education, enterprise size, size of enterprise, sectors and economic activities classifications).

7.1   Determine Key Variables

No direct identifiers, such as social insurance number, name or exact address, are included in the dataset. Therefore, we proceed directly to determining key variables according to disclosure scenarios.

Two disclosure scenarios are identified: one for enterprise re-identification and the other for employee re-identification. In the enterprise scenario, an intruder could use publicly available business registers to re-identify an enterprise sampled in the survey. These registers usually contain information on name, address, number of employees, economic activities, location, etc. Among these, the following variables are also included in SES data: *size*, *location*, *economic activity*[3]. We select these three variables as the enterprise-level key variables.

On the employee level, we assume that personal-level information, such as age and sex, can be combined with enterprise information to identify individuals. Moreover, the intruder will be particularly interested in high-earning employees. We therefore include the following employee-level key variables: *age*, *sex*, *earnings*, and *overtime earnings*. A more detailed process for determining key variables in the SES data is available in Ichim and Franconi (2007).

7.2   Risk Assessment for Categorical Key Variables

After selecting the key variables, we assess record-level risk measures for categorical key variables (i.e., size, location, economic activity, age and sex). We use sdcMicro and/or sdcMicroGUI package[4] (Kowarik et al., 2013; Templ et al., 2014b) to calculate disclosure risks, including the number of records that violate 2-anonymity or 3-anonymity, number of records with risks higher than the main part of the data, and the expected number of re-identifications.

Listing 1: **Record-level and global risk assessment measures of the original SES data**

```
Number of observations violating
− 2−anonymity: 11212
− 3−anonymity: 23682
_____
Percentage of observations violating
− 2−anonymity: 5.61 %
− 3−anonymity: 11.85 %
_____
0 observations with higher risk than the main part
Expected number of re−identifications:
8496.45 [4.25 %]
_____
```

The output in Listing 1 from sdcMicroGUI indicates a large number of records possessing unique patterns of selected categorical key variables (about 5.61% of the total observations violated 2-anonymity). All in all, 4.25% of the records are expected to be re-identified. In addition, the global risk using log-linear models[5] is estimated to be 2.22% in the original data.

7.3   SDC of Categorical Key Variables

To reduce the number of sample uniques (in this example, the goal is to achieve 3-anonymity), we start by recoding the economic activities from 2-digit to 1-digit codes. The recoding is based on expert knowledge about which economic activities are similar and can be combined. We also recode the age of the employees into six age groups (less than or equal to 15; 16 to 29; 30 to 39; 40 to 49; 50 to 59; and greater than or equal to 60). After performing the recoding of key variables, we re-calculate the sample frequency counts and obtain the following.

Listing 2: **Frequency calculation after recoding**

```
Number of observations violating
− 2−anonymity: 12
− 3−anonymity: 22
_____
Percentage of observations violating
− 2−anonymity: 0.01 %
− 3−anonymity: 0.01 %
_____
0 observations with higher risk than the main part
Expected number of re−identifications:
51.01 [0.03 %]
_____
```

---

3   Statistical classification of economic activities in the European Community (NACE).

4   Detailed guides on how to use sdcMicro (Templ et al. 2013) and sdcMicroGUI (Kowarik et al. 2013; Templ et al. 2014b) are available in separate documents.

5   Global risk measure based on log-linear models can be calculated using sdcMicro (Templ et al. 2013), but is not available in sdcMicroGUI.

We see that record-level disclosure risks decreased dramatically after recoding. For example, only 0.03% of the records is expected to be re-identified, compared to 4.25% in the original data (Listing 1). Meanwhile, our measure of global risk using log-linear models has dropped to 0.

We notice, however, that 22 observations still violate 3-anonymity. We further apply local suppression, and suppress 4 values for the variable *size*, and 14 values for the variable *age*. Depending on the expected goals discussed before the SDC process, these steps may have already sufficiently reduced the disclosure risks.

An alternative here is to apply PRAM to the location variable, swapping values between categories using pre-specified probabilities. The higher the probabilities, the more perturbation of data, and thus the greater reduction of disclosure risks.

7.4 SDC of Continuous Key Variables

For continuous key variables (i.e., *earnings* and *overtime earnings*), we apply micro-aggregation (MDAV method) partitioning records into groups of 3 based on classical Euclidean distances (Domingo-Ferrer and Mateo-Sanz, 2002) and assigning the arithmetic mean to each group. By setting group size to three, 3-anonymity is achieved with respect to the earnings variables.

Alternatively, we can add correlated noise to *earnings* and *overtime earnings*. Here we set noise level at 150, defined as the percentage of co-variance of the continuous key variable in the original data.

We also applied shuffling by fitting a regression model that predicts overtime earnings and earnings using *location*, *gender*, *age*, *education*, *occupation* and *type of contract* as predictors.

After each SDC method is applied on the continuous key variables, we examine the risk and information measures. For example, Listing 3 presents the risk and information loss measures after adding noise. The measure of disclosure risk presented below uses the interval disclosure approach (described in Section 3.6). Recall that the interval disclosure approach constructs an interval around each masked value, and examines whether the original value falls within the interval. The upper bound of the risk measure shown in Listing 3 indicates the proportion of the original values that fall within the interval, assuming the worst-case scenario

where every original value that falls within the interval is a correct match with the masked value (i.e. the two values refer to the same respondent). Additionally, Listing 3 shows two direct measures of information loss, IL1 and differences in eigenvalues.

Listing 3: **Disclosure risks and information loss after applying microaggregation (MDAV, k=3) to continuous key variables**

```
Number of observations violating
Disclosure Risk is between:
        [0%; 61.42%] (current)

                        (orig: ~100%)
-Information Loss:
        IL1: 0.11
-Difference Eigenvalues: -0.64%

(orig: Information Loss: 0)
```

7.5 Assess Information Loss with Benchmarking Indicators

We use the benchmarking indicator approach to assess information loss and data utility resulting from SDC. We assume that one of the most interesting indicators for users of the SES data is the Gender Pay Gap, i.e., the difference in hourly earnings between men and women. Thus, our goal is to ensure that the estimate of the Gender Pay Gap using the perturbed data is very close to the estimate using the original data. We also assume that many users are interested in the relation between hourly earnings and sex, age, location, economic activity and education variables. Therefore, if we estimate log hourly earnings using the perturbed data and the original data, we should expect similar coefficients on the predictor variables.

To illustrate how to compare across SDC methods, Figure 3 shows the regression coefficients and confidence intervals estimated using the original data (in black), in comparison to the estimates from anonymized data (in grey). The SDC methods compared in this example include recoding of *economic activity* (recoded from 52 classes to 15) and *age* (recoded into 6 age groups), PRAM applied to *location* and shuffling (as described in Section 7.4).
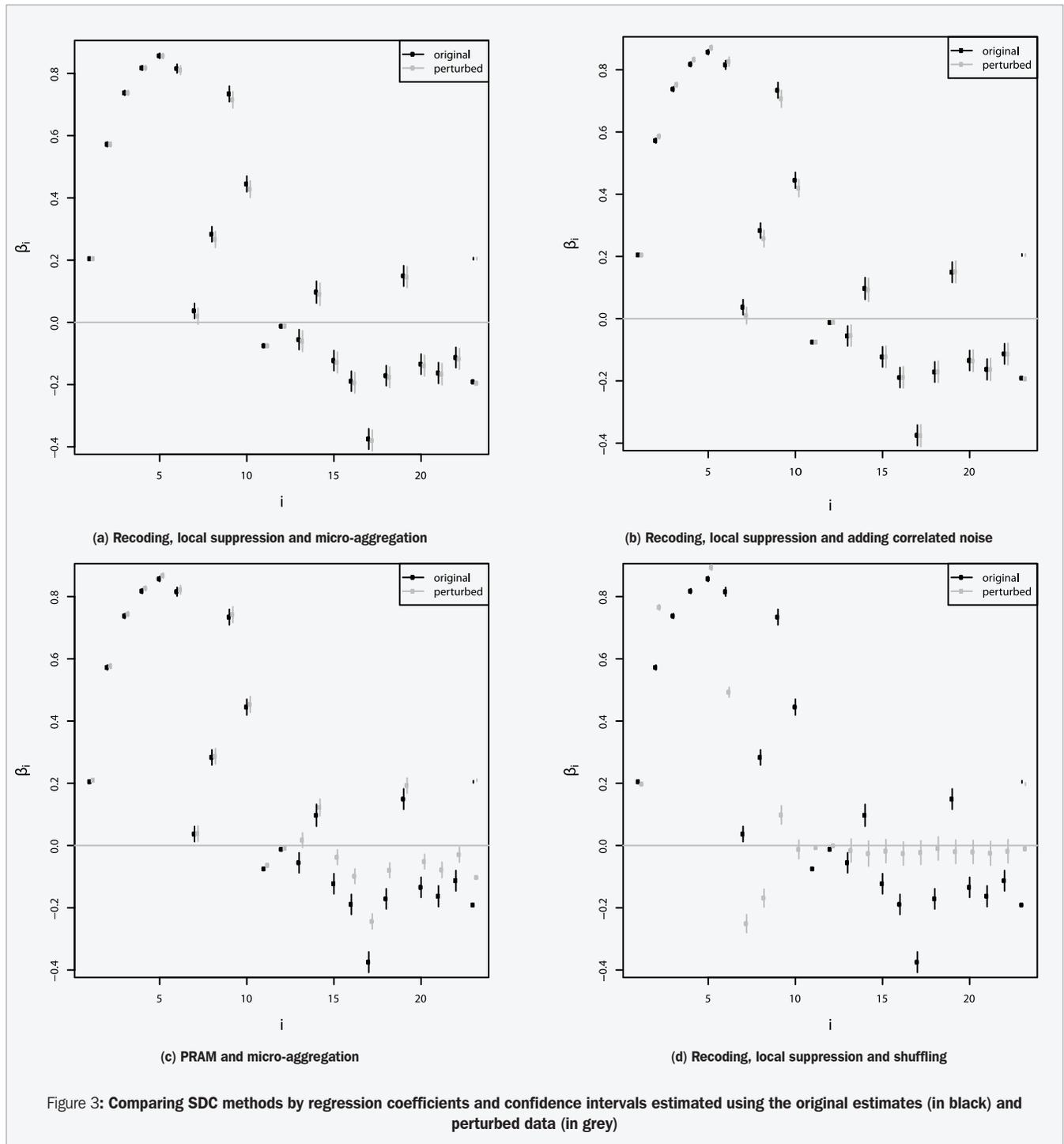
Assuming the level of disclosure risk is the same, recoding, local suppression and micro-aggregation seem to perform the best among the four results, as shown in Figure 3(a), especially since the confidence intervals obtained from the perturbed data cover almost completely the confidence intervals obtained

from the original data. Almost as good are the results after applying recoding, local suppression and adding correlated noise, as shown in Figure 3(b).

Most coefficients are preserved after applying PRAM and micro-aggregation, except the coefficient on *economic activity* (Figure 3c). This is not surprising,

since we swapped the categories of *economic activity* when we applied PRAM.

Fewer coefficients are preserved after applying recoding, local suppression and shuffling. This is because even if the relation between the variables specified in the shuffling model (i.e., between *earnings*

(a) Recoding, local suppression and micro-aggregation

(b) Recoding, local suppression and adding correlated noise

(c) PRAM and micro-aggregation

(d) Recoding, local suppression and shuffling

Figure 3: **Comparing SDC methods by regression coefficients and confidence intervals estimated using the original estimates (in black) and perturbed data (in grey)**

and *sex, age* and *education*) is well preserved, the relation between earnings and variables not included in the shuffling model (e.g., *location, economic activity*) might not be preserved. Specifying a better-fitting model for shuffling might generate better results in this case.

Acronyms

DIS	Data Intrusion Simulation
EFTA	European Free Trade Association
IHSN	International Household Survey Network
MAD	Median Absolute Deviation
MDAV	Maximum Distance to Average Vector
MSU	Minimal Sample Unique
NACE	Statistical Classification of Economic Activities in the European Community
NSO	National Statistical Office
OECD	Organisation for Economic Co-operation and Development
PRAM	Post-Randomization Method
RMD	Robust Mahalanobis Distance
SDC	Statistical Disclosure Control
SES	European Union Structure of Earnings Statistics
SUDA	Special Uniques Detection Algorithm

References

A. Alfons, S. Kraft, M. Templ and P. Filzmoser. Simulation of close-to-reality population data for household surveys with application to EU-SILC. Statistical Methods & Applications, 20(3):383-407, 2011. URL http://dx.doi.org/10.1007/s10260-011-0163-2.

R. Brand. Microdata protection through noise addition. In Inference Control in Statistical Databases, LNCS 2316, pp. 97-116, 2002. Springer-Verlag Berlin Heidelberg.

J. Castro. Statistical Disclosure Control in Tabular Data. In Privacy and Anonymity in Information Management Systems, Advanced Information and Knowledge Processing, pp 113-131, 2010. Springer London.

D. Defays and M.N. Anwar. Masking Microdata Using Micro-aggregation. Journal of Official Statistics, Vol. 14, No. 4, pp. 449-461, 1998.

J. Domingo-Ferrer and V. Torra. Ordinal, Continuous and Heterogeneous k-Anonymity Through Microaggregation. Data Mining and Knowledge Discoverty, 11, 195-212, 2005.

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, J.M. Mateo-Sanz, A. Oganian and A. Torres. On the security of microaggregation with individual ranking: analytical attacks. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):477-492, 2002.

J. Domingo-Ferrer, F. Sebé and J. Castellà-Roca. On the Security of Noise Addition for Privacy in Statistical Databases. In J.Domingo-Ferrer and V.Torra (Eds): PSD 2004, LNCS 3050, pp. 149-161, 2004.

J. Drechsler. Synthetic Datasets for Statistical Disclosure Control. Lecture Notes in Statistics. Volume 201, 2011. Springer, New York.

O. Dupriez and E. Boyko. Dissemination of Microdata Files: Formulating Policies and Procedures. IHSN Working Paper No. 005, 2010. International Household Survey Network.

E.A.H. Elamir and C. J. Skinner. Record level measures of disclosure risk for survey microdata. Journal of Official Statistics, 22 (3), 2006.

M. J. Elliot, C. J. Skinner and A. Dale. "Special Uniques, Random Uniques, and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk." Research in Official Statistics 1(2), pp. 53-67, 1998.

M. Elliot, A.M. Manning and R.W. Ford. A computational algorithm for handling the special uniques problem. International Journal of Uncertainty, Fuzziness and Knowledge Based System 10(5): 493-509, 2002.

M. Elliot and A.M. Manning. Using DIS to modify the classification of special uniques. Invited paper. Joint ECE/Eurostat work session on statistical data confidentiality. Luxembourg, 2-9 April 2003.

L. Franconi and S. Polettini. Individual risk estimation in mu-Argus: a review. In Domingo-Ferrer, J. Eds. Privacy in Statistical Databases, Lecture Notes in Computer Science. Pp: 262-272. 2004. Springer.

J. Gouweleeuw, P. Kooiman, L. Willenborg and P. P. de Wolf. Post randomization for statistical disclosure control: Theory and implementation. Journal of Official Statistics, 14(4):463-478, 1998.

D. Ichim. Extensions of the Re-identification Risk Measures Based on Log-Linear Models. In J.Domingo-Ferrer and Y.Saygin (Eds): PSD 2008, LNCS 5262, pp 203-312, 2008. Springer-Verlag Berlin Heidelberg.

D. Ichim and L. Franconi. Disclosure scenario and risk assessment: structure of earnings survey. In Joint UNECE/Eurostat work session on statistical data confidentiality, Manchester, 2007. DOI: 10.2901/Eurostat.C2007.004.

D. Ichim and L. Franconi. Strategies to achieve SDC harmonisation at European level: Multiple countries, multiple files, multiple surveys. In Privacy in Statistical Databases'10, pages 284-296, 2010.

A. Kowarik, M. Templ, B. Meindl and F. Fonteneau. sdcMicroGUI: Graphical user interface for package sdcMicro. 2013. URL http://CRAN.R-project.org/package=sdcMicroGUI. R package version 1.0.3.

M. A. Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. Journal of the American Statistical Association 84: 414–420.

D. Lambert. Measures of disclosure risk and harm. Journal of Official Statistics, 9 313-331, 1993.

A. Machanavajjhala, D. Kifer, J. Gehrke and M. Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from 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.

R. Maronna, D. Martin and V. Yohai. Robust Statistics: Theory and methods. 2006. Wiley, New York.

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

Pagliuca D. and Seri G. Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2, 1999.

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, pp. 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.

C.J. Skinner and D.J. 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. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):557-570, 2002.

M. Templ. Statistical disclosure control for microdata using the R-package sdcMicro. Transactions on Data Privacy, 1(2):67-85, 2008. URL http://www.tdp.cat/issues/abs.a004a08.php.

M. Templ. Estimators and model predictions from the structural earnings survey for benchmarking statistical disclosure methods. Research Report CS-2011-4, Department of Statistics and Probability Theory,Vienna University of Technology, 2011. URL http://www.statistik.tuwien.ac.at/forschung/CS/CS-2011-4complete.pdf.

M. Templ and P. Filzmoser: «Simulation and quality of a synthetic close-to-reality employer-employee population;» Journal of Applied Statistics, x (2013), S. 1 - 20.

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, 2008a. ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3 10.

M. Templ and B. Meindl. Robustification of microdata masking methods and the comparison with existing methods. Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer, 5262:177-189, 2008b. ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3 15.

M. Templ, A. Kowarik and B. Meindl. sdcMicro: Statistical Disclosure Control methods for the generation of public- and scientific-use files. Manual and Package., 2013. URL http://CRAN.R-project.org/package=sdcMicro. R package version 4.2.0.

M. Templ, A. Kowarik and B. Meindl. sdcMicro case studies. Research Report CS-2014-1, Department of Statistics and Probability Theory. Vienna University of Technology, 2014a. (not published).

M. Templ, B. Meindl and A. Kowarik. GUI tutorial. Research Report CS-2014-2, Department of Statistics and Probability Theory. Vienna University of Technology, 2014b. (to be published soon)

W. E. Yancey, W. E. Winkler and R. H. Creecy, "Disclosure risk assessment in perturbative microdata protection," Lecture Notes in Computer Science, vol. 2316, pp.135-152, Springer, 2002.

About the IHSN

In February 2004, representatives from developing countries and development agencies participated in the Second Roundtable on Development Results held in Marrakech, Morocco. They reflected on how donors can better coordinate support to strengthen the statistical systems and monitoring and evaluation capacity that countries need to manage their development process. One of the outcomes of the Roundtable was the adoption of a global plan for statistics, the *Marrakech Action Plan for Statistics* (MAPS).

Among the MAPS key recommendations was the creation of an International Household Survey Network. In doing so, the international community acknowledged the critical role played by sample surveys in supporting the planning, implementation and monitoring of development policies and programs. Furthermore, it provided national and international agencies with a platform to better coordinate and manage socioeconomic data collection and analysis, and to mobilize support for more efficient and effective approaches to conducting surveys in developing countries.

The IHSN Working Paper series is intended to encourage the exchange of ideas and discussion on topics related to the design and implementation of household surveys, and to the analysis, dissemination and use of survey data. People who whish to submit material for publication in the IHSN Working Paper series are encouraged to contact the IHSN secretariat via info@ihsn.org.

www.ihsn.org
E-mail: info@ihsn.org