This document provides an introduction to statistical disclosure control (SDC) and guidelines on how to apply SDC methods to microdata. Section 1 introduces basic concepts and presents a general workflow. Section 2 discusses methods of measuring disclosure risks for a given micro dataset and disclosure scenario. Sec- tion 3 presents some common anonymization methods. Section -1 introduces how to assess utility of a micro dataset after applying disclosure limitation methods.

1. Concepts

A microdata file is a dataset that holds information collected on individual units; examples of units include people, households or enterprises. For each unit, a set of variables is recorded and available in the dataset. This section discusses concepts related to disclosure and SDC methods, and provides a workflow that shows how to apply SDC methods to microdata.

1.1. Categorization of Variables

In accordance with disclosure risks, variables can be classified into three groups, which are not necessarily disjunctive:

Direct Identifiers are variables that precisely identify statistical units. For exam- ple, social insurance numbers, names of companies or persons and addresses are direct identifiers.

Key variables are a set of variables that, when considered together, can be used to identify individual units. For example, it may be possible to identify individuals by using a combination of variables such as gender, age, region and occupation. Other examples of key variables are income, health status, nationality or political preferences.

Key variables are also called implicit identifiers or quasi-identifiers. When discussing SDC methods, it is preferable to distinguish between categorical and continuous key variables based on the scale of the corresponding variables.

Non-identifying variables are variables that are not direct identifiers or key vari- ables.

For specific methods such as l-diversity, another group of sensitive variables is defined in Section

1.2. What is disclosure?

In general, disclosure occurs when an intruder uses the released data to reveal previously unknown information about a respondent. There are three dierelt types of disclosure:

Identity disclosure:

- In this case, the intruder associates an individual with a re- leased data record that contains sensitive information, i.e. linkage with ex- ternal available data is possible.
- Identity disclosure is possible through direct identifiers, rare combinations of values in the key variables and exact knowl- edge of continuous key variable values in external databases.
- For the latter, extreme data values (e.g., extremely high turnover values for an enterprise) lead to high re-identification risks, i.e. it is likely that responends with ex- treme data values are disclosed.

Attribute disclosure:

- In this case, the intruder is able to determine some cl1arac- teristics of an individual based on information available in the released data.
- For example, if all people aged 56 to GO who identify their race as black in region 12345 are unemployed, the intruder can determine the value of the variable labor status.

Inferential disclosure:

- In this case, the intruder, though with some uncertainty, can predict the value of some characteristics of an individual more accu- rately with the released data.

If linkage is successful based on a number of identifiers, the intruder will have access to all of the information related to a specific corresponding unit in the released data. This means that a subset of critical variables can be exploited to disclose everything about a unit in the dataset.

1.3. Remarks on SDC Methods

In general, SDC methods borrow techniques from other fields. For instance, multi- variate (robust) statistics are used to modify or simulate continuous variables and to quantify information loss.

Distribution-fitting methods are used to quantify disclosure risks. Statistical modeling methods form the basis of perturbation algo- rithms, to simulate synthetic data, to quantify risk and information loss. Linear programming is used to modify data but minimize the impact on data quality. Problems and challenges arise from large datasets and the need for efficient algo- rithms and implementations. Another layer of complexity is produced by complex structures of hierarchical, multidimensional data sampled with complex survey dcsigns.

Missing values are a challenge, especially for computation time issues; structural Zeros (values that are by definition Zero) also have impact on the application of SDC methods.

Furthermore, the compositional nature of many components should always be considered, but adds even more complexity. SDC techniques can be divided into three broad topics:

- Measuring disclosure risk (see Section 2)
- Methods to anonymize micro-data (see Section 3)
- Comparing original and modified data (information loss) (see Section /l)

1.4. Risk Versus Data Utility and Information Loss

The goal of SDC is always to release a safe micro dataset with high data utility and a low risk of linking confidential information to individual respondents. Figure 1 shows the trade-off between disclosure risk and data utility. We applied two SDC methods with different parameters to the European Union Structure of Earnings Statistics (SES) data [see Tenipl ct al., 2014-¡1, for more on anonymization of this dataset].

For Method 1 (in this example adding noise), the parameter varies between 10 (small perturbation) to 100 (perturbation is 10 times higher). Vl/hen the parameter value is 100, the disclosure risk is low since the data are heavily perturbed, but the information loss is very high, which also corresponds to very low data utility. When only low perturbation is applied to a dataset, both risk and data utility are high. It is easy to see that data anonymized with Method 2 (we used microaggregation with different aggregation levels) have considerably lower risk; therefore, this method is preferable. In addition, information loss increases only slightly if the parameter value increases; therefore, Method 2 with parameter value of approximately 7 would be a good choice in this case since this provides both, low disclosure risk and low information loss.

For higher values, the perturbation is higher but the gain is only minimal, lower values reports higher disclosure risk, Method 1 should not be chosen since the disclosure risk and the information loss is higher than for method 2. However, if for some reasons method 1 is chosen, the parameter for perturbation might be chosen around 40 if 0.1 risk is already considered to be safe. For data sets concerning very sensible information (like cancer) the might be, however, to high risk and a perturbation value of 100 or above should then be chosen for method 1 and a parameter value above 10 might be chosen for method

Figure 1: Risk versus information loss obtained for two specific perturbation meth- ods and different parameter choices applied to SES data o11 continuous scaled variables. Note that the information loss for the original data is O and the disclosure risk is 1 respecively, i.e. the two curves starts from (1,0). In real-world examples, things are often not as clear, so data anonymization spe- cialists should base their decisions regarding risk and data utility on the following considerations:

- What is the legal situation regarding data privacy? Laws on data privacy vary between countries; some have quite restrictive laws, some dont, and laws often differ for different kinds of data (e.g., business statistics, labor force statistics, social statistics, and medical data).

- How sensitive is the data information and who has access to the anonymized data file? Usually, laws consider two kinds of data users: users from universities and other research organizations, and general users, i.e., the public. In the rst case, special contracts are often made between data users and data producers.

- Usually these contracts restrict the usage of the data to very specic purposes, and allow data saving only within safe work environments. For these users, anonymized microdata les are called scientic use les, whereas data for the public are called public use les. Of course, the disclosure risk of a public use ile needs to be very low, much lower than the corresponding risks in scientic use les. For scientic use les, data utility is typically considerably higher than data utility of public use les.

- Another aspect that must be considered is the sensitivity of the dataset. Data 011 individuals medical treatments are more sensitive than an establishments turnover values and number of employees. If the data contains very sensitive in- formation, the microdata should have greater security than data that only contain information that is not likely to be attacked by intruders.

- Which method is suitable for which purpose? Methods for Statistical Disclo- sure Control always imply to remove or to modify selected variables. The data utility is reduced in exchange of more protection. While the application of some specic methods results in low disclosure risk and large information loss, other methods may provide data with acceptable, low disclosure risks.

- General recomendations can not be given here since the strenghtness and weakness of methods depends on the underlying data set used. Decisions on which variables will be modied and which method is to be used result are partly arbitrary and partly result from a prior knowledge of what the users will do with the data.

- Generally, when having only few categorical key variables in the data set, re- coding and local suppression to achieve low disclosure risk for categorical key variables is recommended.

- In addition, in case of continous scaled key variables, microaggregation is easy to apply and to understand and gives good results. For more experienced users, shufl-ling may often give the best results as long a strong relationship between the key variables to other variables in the data set is present.

- In case of many categorical key variables, postrandomization might be applied to several of these variables. Still methods, such as postrandomization (PRAM), may provide high or low disclosure risks and data utility, depending on the specic choice of parameter values, e.g. the swapping rate.

- Beside these recommendations, in any case, data holders should always estimate the disclosure risk for their original datasets as well as the disclosure risks and data utility for anonyrnized versions of the data.

- To achieve good results (i.e., low disclosure risk, high data utility), it is necessary to anonyrnize in an explanatory manner by applying different methods using different parameter settings until a suitable trade-off between risk and data utility has been achieved.

1.5. R-Package sdcMicro and sdcMicroGUI

SDC methods introduced in this guideline can be implemented by the R-Package sdcMicro. Users who are not familiar with the native R command line interface can use sdcMicroGUI, an easy-to-use and interactive application.