Chapter 2
Basic Concepts

Abstract This section introduces the basic concepts related to statistical disclosure. It presents definitions for certain groups of variables such as sensitive variables or key variables. They are crucial for any other chapter, since SDC methods differ depending on the variables chosen. In addition, basic intruder scenarios are described such as identity, attribute and inferential disclosure. The chapter ends with a discussion about the trade-off between disclosure risk and information loss. The more the disclosure risk is reduced the higher the information loss and the lower the data utility. The concept of risk-utility maps that reports this trade-off is explained based on real data.

2.1 Types of Variables

Figure 2.1 shows different kinds of variables that are important for SDC, depending on the kind of data set. For typical data sets in the bio-medical area or data from census or registers, the distinction between direct identifiers, indirect identifiers, sensitive variables and non-confidential variables is important. For surveys with complex designs, sampling weights and cluster structures are crucial as well. In the following, the different kinds of variables are explained.

2.1.1 Non-confidential Variables

For these variables, it can be assumed that no information is available in external data bases or at least that linking to external information is not possible for the non-confidential variables. In the context of SDC, they are not of great importance, since SDC methods are not applied on them. In this book, they are only considered for few aspects on data utility in Sect. 5.

M. Templ, *Statistical Disclosure Control for Microdata*,
DOI 10.1007/978-3-319-50272-4_2

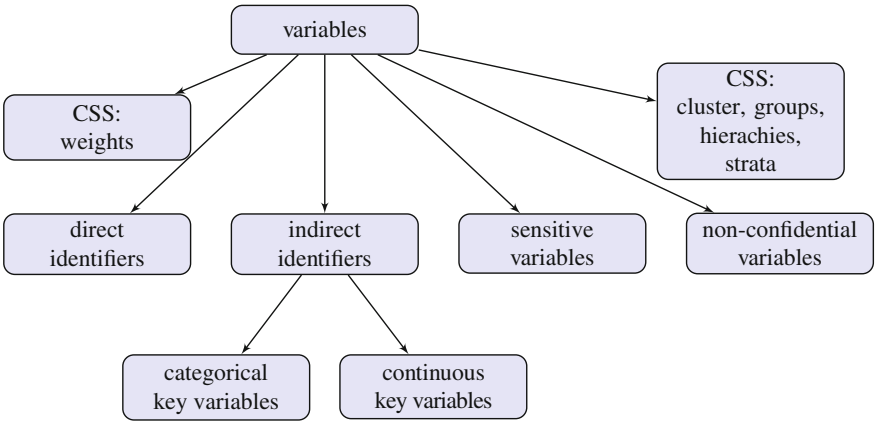

Fig. 2.1 Groups of variables and information in a data set. In addition to census, register or surveys without complex sample designs, some additional information is given for complex survey samples (CSS), e.g., the information on sampling weights, cluster structures and hierarchies in the data set. In special cases, sensitive variables can also be indirect identifiers, see Sect. 2.1.3

2.1.2 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 unambiguously identify statistical units, such as social insurance numbers or names and addresses of companies or persons. Removing direct identifiers is the first step of SDC.

Key variables are, in this book, defined as 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 "indirect identifiers", "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.

The decision on variables serving as key variables—the disclosure scenario—is always crucial and the discussion on it is continued in Sect. 2.2.1.

2.1.3 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 of 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 behaviour, 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 Sect. 2.2.2). A variable can be both identifying and sensitive. For example, income variables 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.

2.1.4 Linked Variables

Often, it is important that the cell value of the same observation of another ("linked") variable should automatically be suppressed as well when a cell is suppressed in a particular key variable. We also use the term *ghost variables* when such variables are linked to some categorical key variables. After applying local suppression the ghost variables should have the same suppression pattern as to those variables to that they are linked. A practical example of ghost variables and how to deal with them is given in Chap. 4, Sect. 4.2.2.7.

2.1.5 Sampling Weights

Sampling has an effect on the disclosure risk and are taken into account for estimation of the disclosure risk (see Chap. 3). It is obvious that the risk is higher for a whole target population (e.g. census) than for sample data. Sampling introduces additional uncertainty about whether the unit identified in statistics is the particular unit in the population, since it is usually not known if the particular respondent participated in the survey. However, if a unit is included in the sample and this unit is unique in the population, this unit might have the same risk, as if the sample were equal to the whole population, i.e. the same risk like as the sampling weight is 1.

Data collected based on complex survey designs typically include at least one vector on sampling weights since each individual may have an unequal chance of being selected. For instance, single-parent households will be given a greater chance of being selected in a survey with the aim of estimating low incomes, and weights can adjust for this. Moreover, weights can also serve other purposes, such as calibration of survey samples to exactly fit known population frequencies and helping to correct for non-response.

In addition, the weights themselves can lead to identifying units. For example, in stratified simple random sampling designs, the weights might be the same for each individual in the strata. If the strata variable is a key variable that is anonymized, the sampling weights may make this anonymization useless because we know, based on the sample weights and some knowledge on the sampling design, which unit is in the same strata.

In summary, the sampling weights have to be considered for disclosure risk estimation (see Chap. 3) and in addition, if they report confidential grouping information, they also might be anonymized by SDC methods for continuous data (see Chap. 4).

2.1.6 Hierarchies, Clusters and Strata

Data may include hierarchies or clusters. This is often also related to sampling designs. In the Austrian Structural Earning Statistics (SES), for example, enterprises are drawn in the first stage and employees are drawn from the selected enterprises in the second stage. In the European Statistics on Income and Living Conditions (EU-SILC), for every member of a household information is retrieved (cluster design). The estimation of disclosure risk is then different from the one without having such clusters. For more details, see Sect. 3.6.

A SDC anonymization method may be applied independently on subgroups (defined by stratification variables). For example, if a grouping procedure such as microaggregation (see Sect. 4.3.1) is applied to economic data, a steel producer's continuous values might be aggregated with a enterprise that has its production in agriculture. However, for data utility aspects, it is often better to apply microaggregation in each economic branch independently.

2.1.7 Categorical Versus Continuous Variables

SDC methods differ between categorical variables and continuous variables. A categorical variable takes values over a finite set. For example, gender is a categorical variable. A continuous variable is numerical; arithmetic operations for real numbers can be performed with it. For example, personal income and turnover of enterprises are continuous variables.

2.2 Types of Disclosure

Three types of disclosure are noted here (see also, Lambert 1993).

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.

2.2.1 Identity Disclosure

Identity disclosure occurs when the intruder successfully associates/links a known individual with released data. For example, the intruder links a released data observation 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.

Even direct identifiers such as names, addresses or social security numbers are deleted from the data set, still extreme values will include high risk of disclosure. To give an extreme example, *occupation* = *president of USA* will surely lead to a disclosure. Also an observation with a value on income of 1.000.000.000€ is of high risk because there will not be many people among the population with such an income. Even more, rare attribute combinations of indirect identifying variables available in public databases are potential of high risk. For example, if an individual having the unique combination of attributes *state* = *AT*, *ethnicity* = *Korean*, *age* = 50, *gender* = *female* and *occupation* = *university lecturer* is in the release sample, this person most probably is also unique in the population and therefore of high risk of disclosure. As soon as the individual is (re-)identified, the intruder knows the possible sensitive values of any variables of this person.

2.2.2 Attribute Disclosure

Attribute disclosure occurs when the intruder is able to determine some new characteristics of an individual based on the information available in the released data. Membership disclosure can be seen as a specific type of attribute disclosure and for disclosure of microdata this is the most relevant case of the more general definition of attribute disclosure.

Table 2.1 serves as an example of (membership) attribute disclosure. In this toy data set, three categorical key variables are defined and the variable *religion* serves as sensitive variable. All in all, four observations are present in this data set. Identity disclosure is not possible since all persons have the same entries. However, we immediately know that each person in this sample with the combination *race* = *black*, aged 50–60, living in region *ZIP* = 1234 is *roman/catholic*. This leads to a disclosure for the sensitive variable "*religious view*".

Table 2.1 Small example showing a case of successful attribute disclosure

OBS	Key variables			Sensitive variable
	Race	Age	Region	Religion
1	Black	50–60	1234	Roman/catholic
2	Black	50–60	1234	Roman/catholic
3	Black	50–60	1234	Roman/catholic
4	Black	50–60	1234	Roman/catholic

Another example: If a hospital publishes data showing that all female patients aged 56 to 60 have cancer, an intruder then knows the medical condition (=cancer) of any female patient aged 56 to 60 staying in this hospital without having to identify the specific individual.

2.2.3 Inferential Disclosure

Inferential disclosure occurs when the intruder is able to determine the value of some sensitive characteristic of an individual more accurately with the released data than it would have been possible otherwise. Inferential disclosure happens when individual's sensitive characteristics can be well predicted from a good model applied on the released data.

For example, with a highly predictive regression model, an intruder will be able to infer a respondents sensitive income information using attributes recorded in the data, leading to inferential disclosure.

In practice, a model would be fit onto the released data and external information on an individual would be used to predict attributes of this individual. For example, assume that *age*, *gender*, *region*, *economic status*, *economic status* and *income* are available in released data. The intruder also has information on the first five mentioned variables on a particular person, say *A*. He can fit a regression model with income as response on the released data and he will receive the fitted regression coefficients. A linear combination of these coefficients with the values on person A predicts the income of this individual. Depending on the quality of the model, the income might be fitted accurately enough. If this is the case, the intruder successfully disclosed the income of person A.

As an example, we use the EU-SILC data set from R package **laeken** (Alfons and Templ, 2013). This data set is synthetically generated. However, to show inferential disclosure, we assume that the values are real except the income components. Let's also assume that an intruder is interested in the variable employee cash or near cash net income (variable `py010n`) of a person with the following attitudes:

```
intrudersKnowledge <- data.frame("hsize" = 3,
                                 "db040" = "Tyrol",
                                 "age"= 34,
                                 "pl030" = "2",
                                 "pb220a" = "AT",
                                 "eqIncome" = 16090)
intrudersKnowledge

  ##   hsize db040 age pl030 pb220a eqIncome
  ## 1     3 Tyrol  34     2     AT    16090
```

Note that this is already more information than typically available, since the intruder even knows the equivalized income of persons in the household.

He is interested in knowing the employee cash or near cash net income of a person. The intruder might try to find a good model using personal income on employee cash or near cash net income as response and certain other variables as predictors. For simplicity, assume that he is just using his predictors without any interaction terms (the inclusion of interaction terms do not improve the predictive power for this data set anyhow). Since the employee cash net income variable is right-skewed, the log is taken. Moreover, assume the intruder knows that the income is not 0, so we only take observations with cash net incomes greater than zero.

```
data(eusilc)
mod1 <- lm(log(py010n) ~ hsize + db040 + age +
                         pl030 + pb220a + eqIncome,
           data=eusilc[eusilc[,"py010n"] > 0, ])
s1 <- summary(mod1)
s1$r.squared

  ## [1] 0.3682565
```

We see that the predictive power of the model is not very high ($R^2 \sim 0.37$).

Let us assume that the real (unknown!) value of `py010n` for person 1 is 9500€. We predict this value from our given data set and estimated model.

```
exp(predict(mod1, intrudersKnowledge))

  ##        1
  ## 7242.003
```

We get an estimated value of 7.242003×10^3 on employee cash or near cash net income. We can ask ourselves if this value is far enough from the true unknown value of 9500€. However, we should also take the model uncertainty into account.

```
exp(predict(mod1, intrudersKnowledge, interval = "prediction"))

  ##        fit      lwr      upr
  ## 1 7242.003 1826.026 28721.73
```

We see that the prediction interval is rather large, inferential disclosure is hardly possible with this scenario.

Exercises:

Question 2.1 **Choice of variables (I)**
Have a brief look at a popular data set, the EU-SILC data in R-package **laeken** (Alfons and Templ 2013). Read the help for this data set by typing in R:

```
install.packages("laeken")
data(eusilc)
?eusilc
```

Determine which of the variables should be defined as
(a) direct identifiers (if any)
(b) categorical key variables
(c) continuous key variables
(d) sensitive variables

Question 2.2 **Choice of variables (II)**
Please have a brief look at another popular data set, the Structural Earnings Statistics in R-package **laeken** (Alfons and Templ 2013). Read the help for this data set by typing in R:

```
data(ses)
?ses
```

Determine which of the variables should be defined as
(a) direct identifiers (if any)
(b) categorical key variables
(c) continuous key variables
(d) sensitive variables

Question 2.3 **Frequencies of key variables**
Use again the EU-SILC data set from package **laeken**. Is a re-identification of observation 8 possible assuming region (*db040*), age, gender (*rb090*) and economic status (*pl030*) as categorical key variables. Is re-identification of observation 3 easily possible?

If you do not have any experience in R, please read Chap. 1 first, before starting this exercise.

2.3 Disclosure Risk Versus Information Loss and Data Utility

Applying SDC techniques to the original microdata will result in information loss and hence affect data utility. Data utility describes the value of data as an analytical resource, comprising analytical completeness and analytical validity.

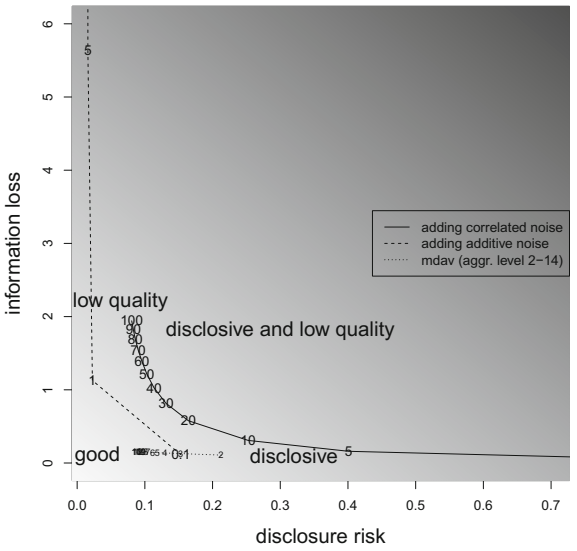

Fig. 2.2 Disclosure risk versus information loss obtained from two specific SDC methods applied to the SES data. Note that the information loss for the original data is 0 and the disclosure risk is 1 respectively, i.e. the curves theoretically start from (1, 0)

Therefore, the main challenge for a statistical agency 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, Fig. 2.2 shows a general example of results after applying two different SDC methods to the Structure of Earnings Statistics (SES) data (cf. Chap. 8). Please note that 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 information loss of 0. As shown in Fig. 2.2, three different SDC methods are applied to the same dataset with varying parameters. The solid curve represents the first SDC method (i.e., adding correlated noise; see Sect. 4.3.2). The curve illustrates that, as more noise is added to the original data, the disclosure risk decreases but the extent of information loss increases. This is also visible with method 2 (adding additive noise; see Sect. 4.3.2), but it is more extreme. Small noise is enough to decrease the quality of the data. In comparison, the dotted curve, illustrating the result of the third SDC method (micro-aggregation; see Sect. 4.3.1), is much less steep than the solid and dashed curves representing the first and second 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 and second method. Therefore, for this specific dataset and the tested methods, method 3 with parameter greater than 5 is the preferred SDC method for the statistical agency to reduce disclosure risk with high data utility.

To check if the anonymized data has a similar structure as the original data, certain metrics can be used as information loss measures. For example, the number of categories of a variable in the original data set can be compared with the number of categories in the anonymized data set. We come back to this issue in Chap. 5.

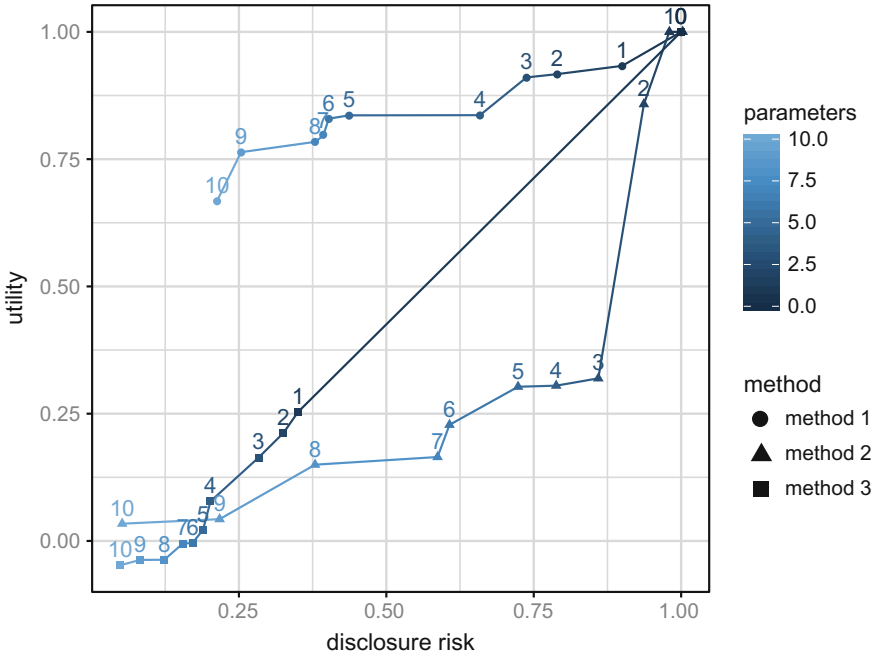

Fig. 2.3 Risk-utility map for three methods

Often not the information loss but the data utility is estimated. For example, a risk-utility map typically includes the data utility instead of the information loss. Theoretically, its name should then be a risk-information-loss map, however, we do not change the term risk-utility map in the following. The data utility can be estimated, for example, on results of regression analysis, i.e. to compare regression coefficients obtained from original and anonymized data. A typical risk-utility map looks like Fig. 2.3. The higher the parameter for perturbation, the lower the disclosure risk but also the lower the data utility. It is clearly visible that method 1 outperforms the other two methods, while method 2 is still better than method 3. For example, with parameter 7, the utility is still high but the disclosure risk reduced significantly. However, the data holder still has to decide if the loss in data utility is still too high while the disclosure risk did not reduce to a given specified threshold. Method 2 is not applicable for this data set. If the parameter for perturbation is low, the risk is too high. If the parameter value increases, the data utility is unacceptably low. Almost the same picture describes method 3.

However, if the utility measure is changed, or the disclosure risk method, the picture would change. In any case, the trade-off between risk and utility always exists and a compromise between disclosure risk and utility always has to be made. The lower the disclosure risk the lower the data utility.

More theory on data utility and information loss can be found in Chap. 5 and practical applications are included in Chap. 8.

However, not only the choice of the information loss criteria or data utility measure is crucial, but also the choice of the risk measure is important. This is discussed in Chap. 3.

Question 2.4 **Risk-utility maps**

Have a look at Fig. 2.2. Assume that the lawyers at your organisation determine the maximum tolerable risk of 0.07. In this case, less than 7% of the observations can be matched correctly. Please answer the following questions.

(a) Which method do you choose—the method corresponding to the lower curve (microaggregation), the method corresponding to the dashed line (adding additive noise) or the method represented by the upper solid-line curve (adding correlated noise)?
(b) The threshold on risk is determined at 0.09 and maximum information loss at 0.5. Your company wants to use adding additive noise to continuous key variables. Would you agree to choose this method?
(c) The risk threshold is 0.2 and the maximum information loss should be 1.5. Which method would you choose?

2.4 Release Types

Confidentiality aspects and the accepted level of disclosure risk depends on national laws, on the type of users and on the type of release of data.

Generally, data release methods can be classified in five different types.

2.4.1 Public Use Files (PUF)

In simple words, this type of data is accessible by the public mostly without any conditions, sometimes with easy-to-meet conditions, e.g. by registration with email, name and address details.

The quality of PUFs varies depending on the needs of the users. Sometimes a lot of effort is spent to produce a PUF that is close-to-reality but at low disclosure risk. Such data sets are then useful for (mostly) researchers to develop methods. But such PUFs are also very useful for teaching. However, sometimes the PUFs are produced with low quality because they are only needed by researchers to make their computer code run on the data set (see Sect. 2.4.4).

In general, PUFs are made easily accessible to

- let researchers develop methods on close-to-reality data;
- for remote execution tasks (Sect. 2.4.4);
- lecturers to support teaching with close-to-reality data sets.

The risk of identifying individual respondents in a PUF should converge to zero. Minimizing the risk of disclosure involves eliminating direct identifiers and modification of indirect identifiers (see Sect. 2.1.2), e.g. by recoding and local suppression (see Sect. 4.2). Also, outliers in continuous scaled variables might be removed and continuous variables might be perturbed (see Sect. 4.2). Other very promising methods include the simulation of synthetic data Alfons et al. (2011), Templ and Filzmoser (2014), Drechsler et al. (2008), Templ et al. (2017) (see Chap. 6 and also Sect. 8.7). Such methods should incorporate sampling designs, missing values, hierarchical and cluster structures (such as persons in households or employees in enterprises). Generally, close-to-reality data sets can be simulated synthetically including very low disclosure risk (Templ and Alfons 2010).

2.4.2 Scientific Use Files (SUF)

SUFs are microdata whereby the researchers need a licence or contract to access them, however, often only researchers can get a contract. In any case, the users of SUFs need authorization to access such data sets.

SUFs are commonly less restrictive than PUFs according to the disclosure risk, but the risk of disclusure still should be low. Of course, direct identifiers are removed as well and anonymization is applied to indirect identifiers. For a data provider, it is recommended that the potential users are asked to complete an application form to demonstrate the need to use a licensed file for a stated statistical or research purpose (Dupriez and Boyko 2010). This allows the data producer to learn which characteristics and data analysis are important for the users. This may also lead to an adaptation of the anonymization methods applied to optimize user needs.

2.4.3 Controlled Research Data Center

Microdata available in a controlled research data center are usually provided at computers at the site of the data provider and are not linked to the internet. Usually, no information can be downloaded via USB ports, CD-DVD or any other device. The accessible data sets may include high risk of disclosure although direct identifiers are removed. Users may only have access to one data set for which they also signed an agreement for use, have to specify why they need access and they typically have to report the goal of their research. The final output is checked by the staff of the data provider (usually experts in statistical disclosure control) and only output with very low risk of disclosure is finally given to the user for external use, e.g. in researchers publications. In any case, the costs for running a research data center might be high because of the staff and facilities needed.

2.4.4 Remote Execution

With remote execution researchers, send their code to the data provider and the data provider executes the code on the original files, checks the output and sends the output that does not disclose information back to the researchers. This is usually an iterative process. The researcher will check their results, and depending on the previous results, they might modify their code or write a more detailed code for analysis.

The costs of remote execution might be reduced by providing structural files. Such files have the same levels of variables as the original data sets, but simulated in a very simplified manner, often just by sampling variable-wise. The structure of structural files might differ a lot from the original data sets, but researchers can check if their code runs. As an alternative to structural files, close-to-reality synthetic PUFs might be provided so that the researchers can test their code before submitting the code to the data provider.

2.4.5 Remote Access

Remote access is the NSI's dream. It would only be necessary to have the data on a server, make a secure connection, install necessary software and finally check the output from the researcher once.

The researcher uses their own desktop computer to connect with a secure connection to the server of the data providing agency. The researcher then usually has access to raw unmodified microdata, but it is not possible to download the data or generated results. The researcher, however, can look at the data and work with it depending on the software installed on the server. Using the pre-installed software, the researcher can manipulate microdata without any restriction. The final results are checked by the data holder and those results which fulfil all criteria for confidentiality are sent to the researcher. Of course, additional attempts may contain the logging of the work of the researchers.

However, this dream hardly ever comes true because of legislative restrictions and practical limitations. For example, even displaying the unmodified raw microdata violates the Austrian law on data privacy. Thus, the researcher may only get access to modified microdata (scientific-use files). In a remote access environment, it is possible to report query results from the underlying original data, but certain queries can also disclose confidential information. It is also out-of-scope that a NSI will check every result that a researcher wants to publish against confidentiality, since the methods (out of thousands of available methods) may vary a lot. Also, it often cannot be precisely determined if a result is dislosive or not. Moreover, queries that disclose information can be programmed in such a manner that there is no chance to detect if such a query discloses information. In any case, strict penalties for any misuse of the data should be executed. Another practical problem is that only pre-defined

software products are available on the server, but in nowadays world researchers may need a great number of individual software packages that may depend on many other software products. Maintaining the server is thus time consuming. Beside all the mentioned risks, various countries offer (successful) remote access facilities and the users usually also have to sign a contract against misuse of the provided remote access system.

References

Templ, M., & Alfons, A. (2010). Disclosure risk of synthetic population data with application in the case of EU-SILC. In *Privacy in Statistical Databases*. Lecture Notes in Computer Science, pp. 174–186. Springer. ISBN 978-3-642-15837-7.

Templ, M., Meindl, B., Kowarik, A., & Dupriez, O. (2017). Simulation of synthetic complex data: the R-package simPop. *Journal of Statistical Software*, 1–38. Accepted for publication in December 2015.

Alfons, A., & Templ, M. (2013). Estimation of social exclusion indicators from complex surveys: The R package laeken. *Journal of Statistical Software*, *54*(15), 1–25.

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

Drechsler, J., Bender, S., & Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy*, *1*(3), 105–130.

Dupriez, O., & Boyko E. (2010). Dissemination of microdata files. Formulating policies and procedures. IHSNWorking Paper No 005, Paris: International Household Survey Network.

Lambert, D. (1993). Measures of disclosure risk and harm. *Journal of Official Statistics*, *9*, 313–331.

Templ, M., & Filzmoser, P. (2014). Simulation and quality of a synthetic close-to-reality employer-employee population. *Journal of Applied Statistics*, *41*(5), 1053–1072.

http://www.springer.com/978-3-319-50270-0

Statistical Disclosure Control for Microdata
Methods and Applications in R
Templ, M.
2017, XIX, 287 p. 37 illus., 27 illus. in color., Hardcover
ISBN: 978-3-319-50270-0