4. Measuring Data Utility and Information Loss

Measuring data utility of the microdata set after disclosure limitation methods have been applied is encouraged to assess the impact of these methods.

4.1. General applicable methods

Anonymized data should have almost the same structure of the original data and should allow any analysis with high precision.

To evaluate the precision, use various classical estimates such as means and co- variances. Using function `dUtility()`, it is possible to calculate different measures based on classical or robust distances for continuous scaled variables. Estimates are computed for both the original and perturbed data and then compared. Following are three important information loss measures:

- **IL1s** is a measures that can be interpreted as scaled distances between original and perturbed values for all p continuous key variables.
- **eig** is a measure calculating relative absolute differences between eigenvalues of the co-variances from standardized continuous key variables of the original and perturbed variables. Eigenvalues can be estimated from a robust or classical version of the co-variance matrix.
- lm is a measure based on regression models, similar to a residual.

Note that these measures are automatically estimated in sdcMicro when an object of class sdcMicr0Obj is generated or whenever continuous key variables are modied in such an object. Thus. no user input is required. We note however that only the former two measures are automatically presented in the GUI in tab Continuous) as IL1 and Difference Eigenvalues respectively.

4.2. Specific tools

In practice, it is not possible to create an anonymized file with the same structure as the original file. An important goal, however, should always be that the difference in results of the most important statistics based on anonymized and original data should be very small or even zero. Thus, the goal is to measure the data utility based on benchmarking indicators, which is in general a better approach to assess data quality than applying general tools. The first step in quality assessment is to evaluate what users of the underlying data are analyzing and then try to determine the most important estimates, or benchmarking indicators.

Special emphasis should be put on benchmarking indicators that take into account the most important vari- ables of the micro dataset. Indicators that refer to the most sensitive variables within the microdata should also be calculated.

The general procedure is quite simple and can be described in the following steps:

1. Selection of a set of (benchmarking) indicators

2. Choice of a set of criteria as to how to compare the indicators

3. Calculation of all benchmarking indicators of the original micro data

4. Calculation of the benchmarking indicators on the protected micro data set

5. Comparison of statistical properties such as point estimates, variances or overlaps in confidence intervals for each benchmarking indicator

6. Assessment as to whether the data utility of the protected micro dataset is good enough to be used by researchers

- If the quality assessment in the last step of the sketched algorithm is satisfactory, the anonymized micro dataset is ready to be published. If the deviations of the main indicators calculated from the original and the protected data are too large, the anonymization procedure should be restarted and modified. It is possible to either change some parameters of the applied procedures or start from scratch and completely change the anonymization process.

- Usually the evaluation is focused on the properties of numeric variables, given unmodified and modified microdata. It is of course also possible to review the impact of local suppression or recoding that has been conducted to reduce individual re-identification risks.

- Another possibility to evaluate the data utility of numerical variables is to define a model that is fitted on the original, unmodified microdata.

- The idea is to predict important, sensitive variables using this model both for the original and protected micro dataset as a first step.

- In a second step, statistical properties of the model results, such as the differences in point estimates or vari- ances, are compared for the predictions, given original and modified microdata, thcn the resulting quality is assessed. If the deviations arc small enough, one may go on to publish the safe and protected micro dataset. Otherwise, adjustments must be made in the protection procedure. This idea is similar to the information loss measure lm described in Section 4.1.

- In addition, it is interesting to evaluate the set of benchmarking indicators not only for the entire dataset but also independently for subsets of the data. In this case, the microdata are partitioned into a set of h groups.

- The evaluation of benchmarking indicators is then performed for each of the groups and the results are evaluated by reviewing differences between indicators for original and modified data in each group.

4.3. Workflow

Figure 3 outlines the most common tasks, practices and steps required to obtain confidential data. The steps are summarized here:

1. The first step is actually to make an inventory of other datasets available to users, to decide on what an acceptable level of risk will be, and to identify the key users of the anonymized data to make decisions on anonymisation to achieve high precision on their estimates on the anonymized data.

2. The first step in anonyimization is always to remove all direct identification variables and variables that contain direct information about units from the microdata set.

3. Second, determine the key variables to use for all risk calculations. This decision is subjective and often involves discussions with subject matter spe- cialists and interpretation of related national laws.

4. The most important methods are included in the sdcMicroGUI, such as basic risk measurement, recoding, local suppression, PRAM (post- randomization), information loss measures, shuffling, microaggregation and adding noise. Other methods listed in the figure for the sake of completeness are included in the sdcMicro R package and in the sim- Population R package. for the simulation of fully synthetic data, choosing key variables is not nec- essary since all variables are produced synthetically, see for example Alfons ct al. [2011].

5. After the selection of key variables, measure disclosure risks of individual units. This includes the analysis of sample frequency counts as Well as the application of probability methods to estimate corresponding individual re- identification risks by taking population frequencies into account.

6. Next, modify observations with high individual risks. Techniques such as recoding and local suppression, recoding and swapping, or PRAM can be applied to categorical key variables. In principle, PRAM or swapping can also be applied without prior recoding of key variables; a lower swapping rate might be possible, however, if recoding is applied before. The decision as to which method to apply also depends on the structure of the key variables. In general, one can use recoding together with local suppression if the amount of unique combinations of key variables is low. PRAM should be used if the number of key variables is large and the number of unique combinations is high; for details, see Sections 3.1 and 3.3 and for practical applications Tcmpl ct al. [2014a]. The values of continuously scaled key variables must be perturbed as well. In this case, micro-aggregation is always a good choice (sec Section More sophisticated methods such as shuffling (see Section 3.6 often provide promising results but are more complicated to apply.

7. After modifying categorical and numerical key variables of the microdata, estimate information loss and disclosure risk measures. The goal is to release a safe micro dataset with low risk of linking confidential information to individuals and high data utility. If the risks is below a tolerable risk and

   the data utility is high, the anonymized dataset is ready for release. Note that the tolerable risk depends on various factors like national laws and sen- sitivity of data, but also subjective arbitrary factors play a role and the risk depends on the selected key variables the disclosure scenario. If the risk is too high or the data utility is too low, the entire anonymization process must be repeated, either with additional perturbations if the remaining re- identification risks are too high, or with actions that will increase the data utility.

In general, the following recommendations hold:

Recommendation 1: Carefully choose the set of key variables using knowledge of both subject matter experts and disclosure control experts. As already men- tioned, the key variables are those variables for which an intruder may possible have data/ information, e.g. age and region from persons or turnover of enter- prises. Which external data are available containing information on key variables is usually known by subject matter specialist.

Recommendation 2: Always perform a frequency and risk estimation to evaluate how many observations have a high risk of disclosure given the selection of key variables.

Recommendation 3: Apply recoding to reduce uniqueness given the set of cat- egorical key variables. This approach should be done in an exploratory manner. Receding on a variable, however, should also be based on expert knowledge to combine appropriate categories. Alternatively, swapping procedures may be ap- plied on categorical key variables so that data intruders cannot be certain if an observation has or has not been perturbed.

Recommendation 4: If recoding is applied, apply local suppression to achieve k-anonymity. In practice, parameter It” is often set to 3.

Recommendation 5: Apply micro-aggregation to continuously scaled key variables. This automatically provides k-anonymity for these variables.

Recommendation 6: Quantify the data utility not only by using typical estimates such as quantiles or correlations, but also by using the most important data-specic benchmarking indicators