WP.30
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Manchester, United Kingdom, 17-19 December 2007)

Topic (iii): Applications (including practical implementation of SDC methods, actual issues within NSIs and software)

THE ANONYMISATION OF THE CVTS2 AND INCOME TAX DATASET AN APPROACH USING R-PACKAGE "SDC MICRO"

Supporting Paper

Prepared by Bernhard Meindl (Statistics Austria) and Matthias Templ (Statistics Austria and University of Technology, Austria)

The anonymisation of the CVTS2 and income tax dataset

An approach using R-Package *"sdcMicro"*

Bernhard Meindl*, Matthias Templ*,**

* Department of Methodology, Statistics Austria, Guglgasse 13, 1110 Vienna, Austria. (bernhard.meindl@statistik.gv.at) and

** Department of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, 1040 Vienna, Austria. (templ@statistik.tuwien.ac.at)

Abstract. The demand for microdata for research and teaching purposes gets higher and higher. To overcome this fact we not only provide secure workstations where researchers can deal with "original" data but also provide more and more anonymised microdata. When using flexible software tools for anonymisation we can provide high quality anonymised data which can be generated in less time. In this paper we show such anonymisation processes on two different data set, the continuing vocational training survey (CVTS2) and the income tax data from 2005.

1 Introduction

Since the demand to publish micro data for researchers grows, it is necessary to take actions that published data will not lead to correct re-identification of either an individual or an enterprise. To assure the anonymity of micro data, different anonymisation methods such as global recoding, local suppression or microaggregation are used. Special emphasis should be placed on trying to keep the (multivariate) structure of the data and changing the original data as little as possible while guaranteeing a low individual risk of re-identification. With new and modern software it is possible to use anonymisation methods very effective.

1.1 Terms of Use

In Statistics Austria we provide two different variants of anonymised data sets. An SDS (standardised data set) is basically an anonymised micro data set for research and teaching purposes whereas an ADS (task-related dataset) is generated for special research purposes for only one research project or research collaboration with other organisations. However, users have to agree on the terms of use for both variants of anonymised data. From the anonymisation point of view, a SDS is somewhat between a public use file and a scientific use file. The complete terms of

use can be found at: http://www.statistik.at/web_de/services/mikrodaten_fuer_forschung_und_lehre/datenangebot/standardisierte_datensaetze_sds/index.html.

2 Software used

To protect the CVTS data and the Austrian income tax data we used the R-package sdcMicro [17]. R [12] is an open-source, free-available, high-level programming language for statistical computing and graphics. The main advantages of package sdcMicro are the reproducibility of any results, the flexible usage of the package, the import/export facilities, the richness of methods implemented, the graphical power for comparison of original data and perturbed data, the easy usage of the package and the fast calculation of results. Furthermore, one can easily use the whole power of R since the package runs in R.

Since all the results from anonymisation can be reproduced by running a script or parts of the script with the code, a user can do the anonymisation approach very flexible and in an explorative manner and can interactively communicate with all the objects in the workspace of R. It's a bit like playing with the data instead of writing a "batch file" which then must be processed.

3 CVTS2 Data

It was the objective to generate a SDS for the CVTS2 dataset (continuing vocational training survey). The goal of this survey is to gain information on internal measures enterprises have taken and (partly) payed for on advanced vocational training for employees. The raw data consisted of 2613 enterprises for which a total of 197 variables has been recorded. It has to be noted that the sample of the CVTS survey is chosen in a way that only enterprises with at least 9 employees may be drawn into the sample. Due to imputation and plausibility checks we observed some abnormalities such as ratios being greater than 100%, most of which can be explained though. Further information on the CVTS2 data can be found at Statistics Austria's webpage at: http://www.statistik.at/web_de/statistiken/bildung_und_kultur/erwachsenenbildung_weiterbildung_lebenslanges_lernen/index.html.

The difficulty in generating a SDS for this data was the large number of categorical variables and the fact that any combination of these variables might be used by an attacker to correctly identify an enterprise. Thus it was necessary to assess different scenarios of statistical disclosure control by considering different subsets of the available categorical variables as key variables in order to receive an impression of the disclosure and re-identification risk.

Another important point is that we decided to calculate and publish ratios for most of the numeric variables. Doing so, absolute values can not be recycled anymore. This approach was described as well by Brandt and Hafner [2]. However, most multivariate statistical methods are invariant against transformations and the same results can be obtained as for the original values. But, of course, certain aggregations of the data are not suitable for the transformed subset of the data.

3.1 Anonymisation of the CVTS2 data

Before actually starting to apply anonymisation methods, 29 variables which were either direct identifiers or were including non relevant information were deleted from the data set. Then ratios for most numeric variables in the dataset were calculated as already explained above. Furthermore, we recoded several variables such as the *economic classification of the enterprise* or the *number of employees* into broader categories. All operations have been done using `R` and package `sdcMicro` which make the entire anonymisation procedure flexible, fast and reproducible.

We started by comparing different scenarios and several combinations of possible key variables by having a look at the corresponding individual risks for re-identification [8] as well as the number of unique combinations of the characteristics in the key variables. It is in fact very convenient to compare different scenarios using `sdcMicro` because the user only has to specify the desired key-variables and re-run the code. After comparing several possibilities we decided to use the following key variables which are listed below:

- **economic classification of the enterprise:** 10 categories
- **number of employees:** 4 categories
- **generated revenues for vocational training:** 2 categories
- **expenses for vocational training:** 2 categories

It should be noted that all of these key variables have been already changed in an explorative manner by global recoding techniques or have been generated from other variables.

Then we looked at the number of unique combinations of the key variables, the number of observations with a given combination of the key variables that occurs only twice as well as the individual risk for re-identification. We observed 58 unique observations and 50 observations whose combination of values of the key variables occurred exactly two times.

`sdcMicro` provides a method to plot the individual risk and to interactively change the threshold value similar to the μ-Argus [10] plot method. This is helpful to determine suitable threshold values for the local suppression methods that need to be applied to the key variables. Figure 1 shows the individual risk for re-identification along with its empirical distribution function for the original, non perturbated data.

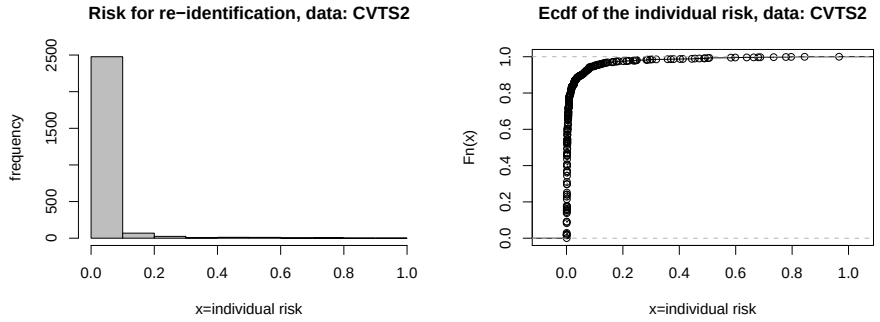

Figure 1: individual risk (left) and empirical distribution (right) in original data.

We want to provide k-anonymity ([15, 13, 14]) for this dataset. This means that for any combination of key variables at least k observations must exist in the data set sharing that combination. The `sdcMicro` function `localSupp()` can be used to suppress values in the key variables. We find 3-anonymity in combination with the other anonymisation methods applied to be a sufficient for publishing for this dataset since the CVTS2 data are not as interesting from an attacker's point of view as for example the income tax data described later.

We started with the variable *beiträge* and set the threshold value to 0.25. This resulted in suppressing 19 values in this key variable. Afterwards, the risk of re-identification of enterprises is plotted again and a new threshold values is determined. Using the threshold value of 0.167, `localSupp()` was applied to the key variable *einnahmen*. This led to a suppression of 25 values in this variable. The same procedure was used to suppress values in the key variable *a299tot*. After choosing a suitable threshold value (0.104) and applying `localSupp()` we note that 12 values were suppressed in this key variable.

We then observed that there were still enterprises left that had unique combination of the key variables or a combination which only occurred two times in the dataset. Thus, we manually set the values of the variable *beiträge* for those enterprises that had a unique combination of key variables to missing. As a result, 14 suppressions had to be done. Additionally, the variable *einnahmen* was set to missing for all enterprises that had a combination of the key variables that occurred only

twice. This resulted in suppressing one additional value. Summarizing this process, a total of 33 values had to be suppressed in variable *beiträge*, 26 variables had to be suppressed in variable *einnahmen* and 12 variables had to be suppressed in variable *a299tot*. After these suppressions, each combination of the values in the key variables occurs at least three times and we note that the goal of 3-anonymity is reached.

In Figure 2 the individual risk for re-identification is plotted along with its empirical distribution function after using local suppression for anonymisation. It is obvious, that we achieved a clear reduction in the individual risk of re-identification compared to Figure 1 as the different scaling of the x-axis indicates.

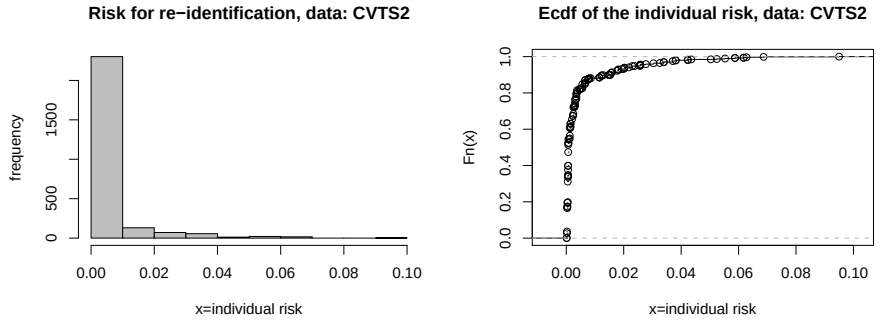

Figure 2: individual risk (left) and empirical distribution (right) in anonymised data.

After dealing with categorical variables and indirect identifiers, we took additional precautions by microaggregating (references on microaggregation can be found in [1], [7], [4], [3], [5]) the available numeric variables. This effectively means that for each numeric variable the values are grouped by a proximity measure into small groups consisting of 4 values. Then the values within each group are averaged and the aggregate are finally released in the anonymised dataset. This assures that each numeric value occurs at least 4 times in the anonymised dataset. In this case we have used a version of individual ranking [4] which can also be applied on data with missing values.

4 Austrian Income Tax Data

The Austrian income dataset for 2005 consists of 5.919.739 rows. The data have already been aggregated from pay slip level to person-level. Thus, exactly one row exists in the raw data for each person that has been liable to pay taxes in 2005. The dataset consists of a total of 74 variables, however, only 17 variables have been included in the published micro data SDS file.

The categorical variables give information about the social status, sex, the federal state and the age of the individual as well as of the number of pay slips considered, the number of weeks employed, whether the person was employed part- or fulltime and the economic classification of the enterprise the individual was predominantly employed at. The quantitative variables included give information for example about the gross wages, social security contributions or information about the amount of other payments a given person has generated.

Further information on the income tax data can be found at the Statistics Austria web page located at: `http://www.statistik.at/web_de/statistiken/oeffentliche_finanzen_und_steuern/_steuerstatistiken/lohnsteuerstatistik/index.html`

The final anonymised dataset is available for download at the following web page: `http://www.statistik.at/web_de/services/mikrodaten_fuer_forschung_und_lehre/datenangebot/standardisierte_datensaetze_sds/index.html#index8`

4.1 Anonymisation of the Austrian Income Tax Data

We will now describe the anonymisation methods applied to the raw data set. Our anonymisation approach is quite different than the one used to generate the german public and scientific use file, respectively of income tax data. More on the german contribution in this field can be found on [11].

First, many quasi direct identifiers and variables that should not be included in the final anonymised data set have been deleted from the raw data. Among the variables deleted are the tax-id which is unique for each person as well as regional information on the individuals such as the exact address. Then, the anonymisation methods described later are applied to the resulting dataset which consisted of all in all 17 variables. Eight variables are scaled categorically while eight variables are quantitative. Furthermore, the sampling weight resulting from drawing a subset of the original data is attached to the dataset as an additional variable.

In a second step, a 1% random sample stratified by age, sex and federal states was drawn from the raw dataset. This resulted in a dataset including 59.279 individuals. This should be seen as a quite effective anonymisation method because even if an attacker manages generate a one to one match from a reference file to an individual from the sample using key variables, he cannot be sure if the possibly identified individual has even been drawn to the sample.

After generating a subset from the raw data it was necessary to define key variables. As already mentioned before, a total of 8 categorical variables was included in the data set. We used `sdcMicro` to compare different scenarios and several com-

binations of key variables by having a look at individual risks for re-identification and the number of unique combinations of the characteristics in the key variables. After comparing several possibilities, we considered the following five key variables:

- **social status:** 7 categories
- **federal state:** 10 categories
- **sex:** 2 categories
- **age classes:** 8 categories
- **economic classification of the enterprise:** 12 categories

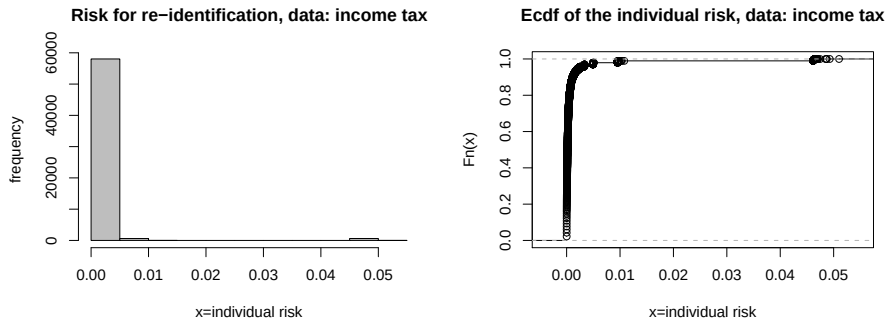

Figure 3: individual risk (left) and empirical distribution (right) in original data.

After deciding on the key variables, the individual risk for re-identification was calculated for all the key variables defined above. It turned out in this explorative approach that we should recode some variables in order to reduce the re-identification risk. Therefore, we recoded variable *social status* as well as the the employment state into less categories. It turned out that 623 observations had a unique combination of characteristics of the key variables and 604 individuals had a combination of values of the key variables that only existed twice. In Figure 3 the individual risk of re-identification is plotted for the data using the five key variables discussed above.

In the next step, values of certain key variables are set to missing for individuals with high risk of re-identification. To assess the individual risk and to find a suitable threshold value which determines the number of suppressions we used again the interactive plot method of `sdcMicro` to assess the individual risk of re-identification and to interactively change the threshold value and look at the resulting re-identification rates conditional on the chosen threshold value.

After choosing a suitable threshold we used the function `localSupp()` to suppress data in the variable *economic classification of the enterprise* the individuals were employed at with a quite low threshold of 0.01. As a result, a total of 631 values ($=\approx 0.01\%$) for this key variable was set to missing. As a result we obtain that the number of individuals that are unique in the dataset drops to 85 and the number of individuals with a combination of values in the key variables that occurs twice drops down to 256.

After this step, the relative risk is plotted again interactively in order to find a suitable threshold value for local suppression of values in the key variable *social status*. Applying the local suppression method with a threshold value of 0.01 results in no observation that has a unique combination of values in the key variables after setting 93 values in the variable *social status* to missing. However, there are still 154 observation left that have a combination of key values that occurs only twice.

We find that 3-anonymity in addition to microaggregation of numeric variables and the fact that the released data set itself is just a 1% random sample from the population, is adequate for this critical dataset. In order to guarantee 3-anonymity we had to set additional values in the key variables to missing. As already described we decide on a threshold value (0.01) and apply the function `localSupp()` to the key variable *"federal state"*. By setting 154 values of this variable to missing, 3-anonymity for this dataset is obtained.

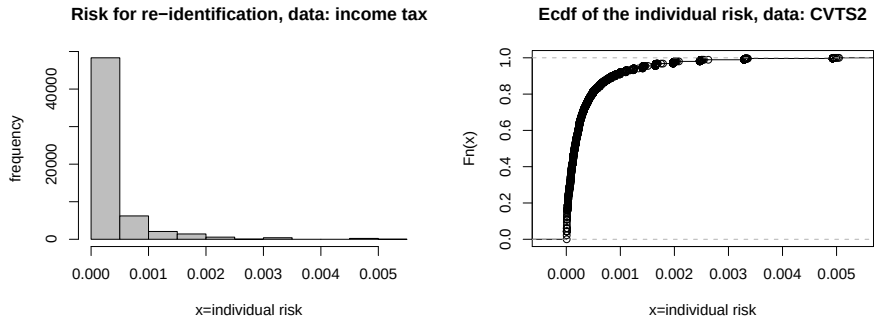

Figure 4: individual risk (left) and empirical distribution (right) in anonymised data.

In Figure 4 the individual risk after locally suppressing values in the key variables is plotted. The graph looks similar to Figure 3, however one should note the different scaling on the x-axis.

After dealing with categorical variables and indirect identifiers, we additionally

microaggregated all numeric variables available in the dataset. As for the CVTS2 data we used a version of individual ranking for the microaggregation procedure which guarantees that each numeric value exists at least 4 times in the SDS.

5 Conclusion

In both anonymised data sets the re-identification of individuals is very hard or even impossible. For example, each combination of extremely identifiers [9] such as the regional variable occurs more than 1926 times for the income tax data set which itself is only a 1% sample from the original data. Most of the proposed anonymisation rules of the handbook on SDC [9] are not considered since our data are too small to fit this criteria. Nevertheless, global recording, local suppression and microaggregation were applied to achieve sufficient anonymisation of the data, i.e. to provide both 3-anonymity and low re-identification risk. In the other hand, the perturbed data are of high quality and have nearly the same (multivariate) structure as the original data since only few selective recodings and local suppressions on categorical variables were made. It is also well known, that microaggregation does not destroy the (multivatiate) structure of the data (see e.g. in [16] or [6]). The software used had allowed the anonymisation in less time and in an explorative way.

References

[1] N. Anwar. Micro-aggregation - the small aggregates method. In *Internal report*. Luxembourg: Eurostat, 1993.

[2] M. Brandt and H. Hafner. Leitfaden zur Anonymisierung für die Erstellung eines Campus-Files aus den Einzeldaten der zweiten europäischen Erhebung zur beruflichen Weiterbildung. Technical report, Statistisches Bundesamt, Hessisches Statistisches Landesamt., 2007. Nr. 5, 10/2005.

[3] D. Defays and Anwar M.N. Masking microdata using micro-aggregation. *Journal of Official Statistics*, 14(4):449–461, 1998.

[4] D. Defays and P. Nanopoulos. Panels of enterprises and confidentiality: the small aggregates method. In *Proceedings of the 1992 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204. Statistics Canada, Ottawa, 1993.

[5] 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.

[6] J. Domingo-Ferrer, J.M. Mateo-Sanz, A. Oganian, and A. Torres. On the security of microaggregation with individual ranking: analytical attacks. *In-*

ternational Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(5):477–492, 2002.

[7] M. Elliot, A. Hundepool, E.S. Nordholt, J-L. Tambay, and T. Wende. Glossary on statistical disclosure control, 2005.

[8] L. Franconi and S. Polettini. Individual risk estimation in μ-ARGUS: a review. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 262–272, 2004.

[9] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte Nordholt, G. Seri, and P. De Wolf. *Handbook on Statistical Disclosure Control*, 2007.

[10] A. Hundepool, A. Van deWetering, Ramaswamy R., L. Franconi, A. Capobianchi, P-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing. μ-argus version 4.1 software and users manual, 2006.

[11] J. Merz, D. Vorgrimler, and M. Zwick. De facto anonymised microdata file on income tax statistics 1998. Technical report, SRI Intl. Tech. Rep., 2005. Nr. 5, 10/2005.

[12] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[13] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.

[14] P. Samarati. Achieving k-anonymity privacy protection using generalization and suppression. *Intl. Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.

[15] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI Intl. Tech. Rep., 1998.

[16] M. Templ. Software development for SDC in R. In *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer*, pages 347–359, 2006.

[17] M. Templ. *sdcMicro. Manual and Package.* Statistics Austria and Vienna University of Technology, Vienna, Austria, 2007. *http://cran.r-project.org/src/contrib/Descriptions/sdcMicro.html.*