WP.
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Confidentiality
(Bilbao, Spain, 2-4 Dezember 2009)

Topic (iv): Tools and software improvements

AN INTERACTIVE GRAPHICAL USER INTERFACE FOR MICRODATA PROTECTION WHICH ALLOWS REPRODUCIBILITY

Submitted by the Department of Statistics and Probability Theory, Vienna University of Technology, and the Department of Methodology, Statistics Austria[1]

ABSTRACT

The graphical user interface (GUI) of R package `sdcMicro` serves as an easy-to-handle tool for users which are not familar with the native R command line interface. Furthermore, interactions are provided between objects obtained from the anonymization process. This allows an automated recalculation and displaying of a summary of the frequency counts, the individual risk, the information loss and the data utility after any operation on the data is applied.
The code for every anonymization step carried out within the GUI is saved in a script which can easily be modified and re-used in order to provide reproducibility.
The tool is open-source and can be downloaded from the comprehensive R archive network (CRAN).

I. `sdcMicro` SOFTWARE

1. `sdcMicro` (Templ 2009) (Templ 2008) is a highly flexible package to generate anonymized microdata files (for applications, see, e.g., Meindl and Templ 2007). It includes all methods of the popular closed-source μ-Argus software (Hundepool, Van deWetering, R., Franconi, Capobianchi, DeWolf, Domingo-Ferrer, Torra, Brand, and Giessing 2008) plus several new ones (see, e.g., Templ 2008 Templ and Meindl 2008a Templ and Meindl 2008b) and it is open source and distributed via CRAN (`http://cran.r-project.org`).

II. `sdcMicro`'S NEW GUI

2. Figure 1 displays the main window of the GUI. Direct access to all available functions is offered. It is designed to give a brief summary of the frequency calculation and risk estimation and provides two sets of buttons for operations to use with categorical and numerical variables.

[1]Prepared by Matthias Templ (templ@tuwien.ac.at), Thomas Petelin (erdschock@gmail.com).

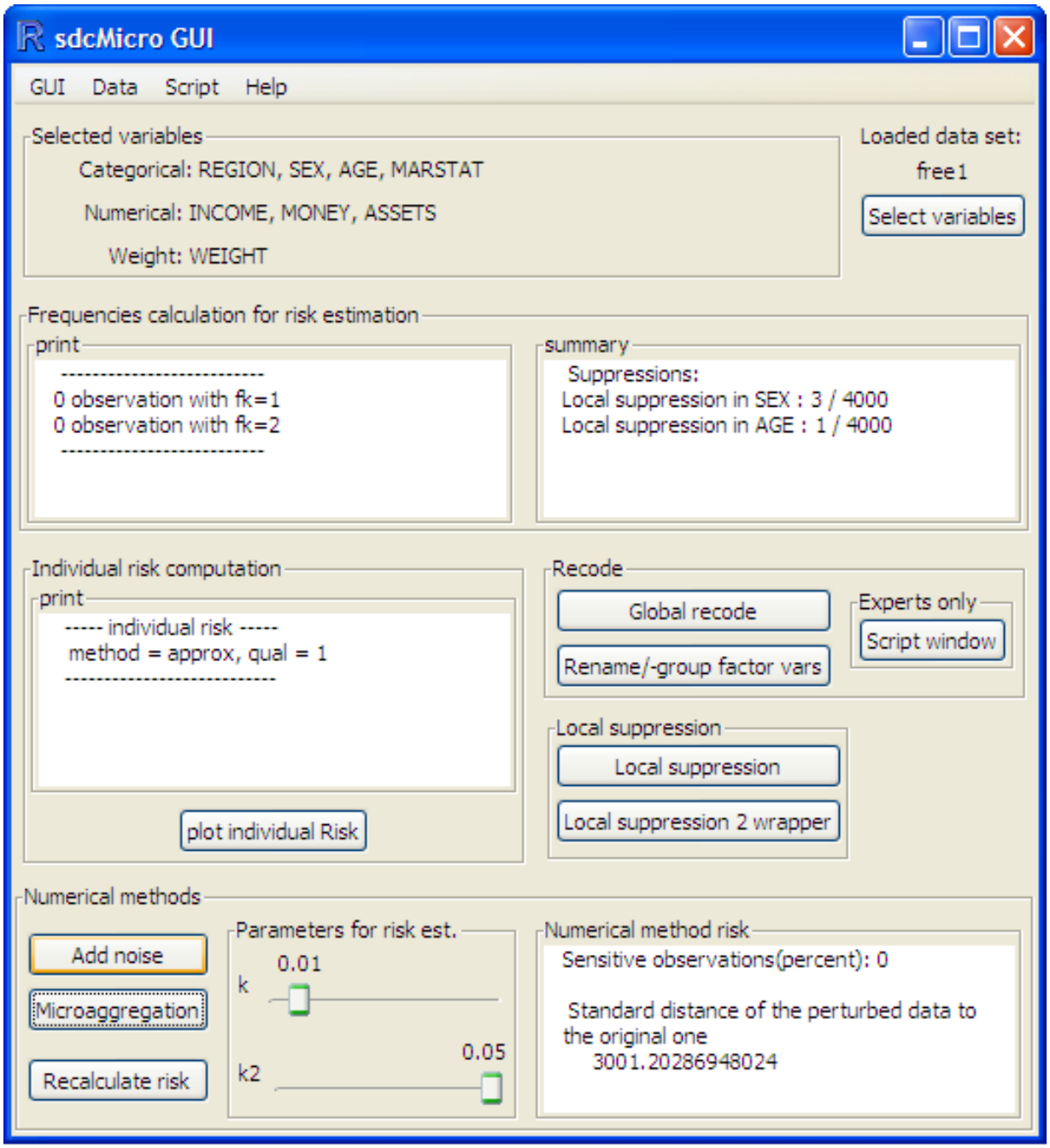

FIGURE 1. Main window of the `sdcMicro` GUI. Some anonymisation steps are already carried out. The corresponding script is shown in the second figure in this paper.

3. Data can be either loaded from the hard disk or choosen from the R workspace. The GUI provides interactive combo boxes to select variables by clicking on the `select variables` button in the main window of the GUI.

4. To reach anonymity of categorical variables recoding of variables is applied in an explanatory manner (Templ 2008). The main goal is to reach both, a low re-identification risk and a minimum modification of the data. This is usually done by trying out several recodings, considering suggestions

from subject matter specialists. The easy-to-use GUI is designed for the explorative use of microdata protection methods in general. All operations, such as recoding, can be applied easily within the GUI.

5. Some specific values which are unique in the sample may be suppressed in an optimal way (minimizing a cost function, see the `sdcMicro` manual (Templ 2009), for example). The GUI provides access to functions which supports local suppression by clicking on `local suppression` or `local suppression 2 wrapper` on the main window. The probability of suppressions in different variables can be adjusted, i.e. the user can define how important a variables is (using a continuous scale) – the less important the higher the probability of local suppressions in the variable.

6. Whenever recoding or suppression is applied, any summary displayed in the main window of the GUI is automatically refreshed/recalculated to see the effect of any operation on the data on the fly. The GUI therefore considers interactions between recoding, local suppression and individual risk computation (see, e.g., Franconi and Polettini 2004) but also considers interactions between reidentification risk, information loss and methods for perturbation of continuous scaled variables.

7. When perturbing continuous scaled key variables the aim is to keep the structure of the dataset while minimizing the risk of re-identification by applying perturbation to the data. Within the GUI several methods for adding noise and microaggregation can be applied to the data. After a method is applied, the risk of reidentification is displayed. Further information on risk estimation can be found in Templ and Meindl 2008a.

8. It is possible to apply any operation to the data which are not explicity supported by the GUI within the `Experts only` button. After pressing this button the user can apply any R code, but working in the environment of the GUI (more information on environments in R, see R Development Core Team 2008).

9. Every action carried out within the GUI is saved in a script, i.e. all parameter values and functions applied are saved. Thus, within such a script it is possible to reproduce every result without any clicks in the GUI (just by loading the script). Modifications can be done in the script and the script can then be easily re-run. So, the actual status could be easily saved and/or and old script can be easily loaded to reproduce output or modify some steps or alter the output. Another possibility is that the user can delete single steps from the script or execute it to a certain point to start his work from there.

10. The `sdcMicro` package and it's graphical user interface runs under all known operating systems. Only MAC OSX users with 64 bit machines may have problems since R is only available within an experimental version for that platform.

III. OPEN-SOURCE INITIATIVE

11. Not all methods are included in the GUI till now, e.g., post randomization (PRAM) or model-based risk estimation can only be carried out in the `Experts only` mode. However, the GUI comes with the general public licence and it is therefore open-source. Implementing additional functionality is not very hard to achive and, e.g. the implementation of method PRAM in package `sdcMicro` can be easily ported to the GUI within a few hours. Nevertheless, while the copyright of the existing code is by the author (intellectual rights must be respected. Side mark: it is also restricted to use the code commerically), the code is open to everybody to cooperate in the development of the package or to extend the GUI or implement new functionalities.

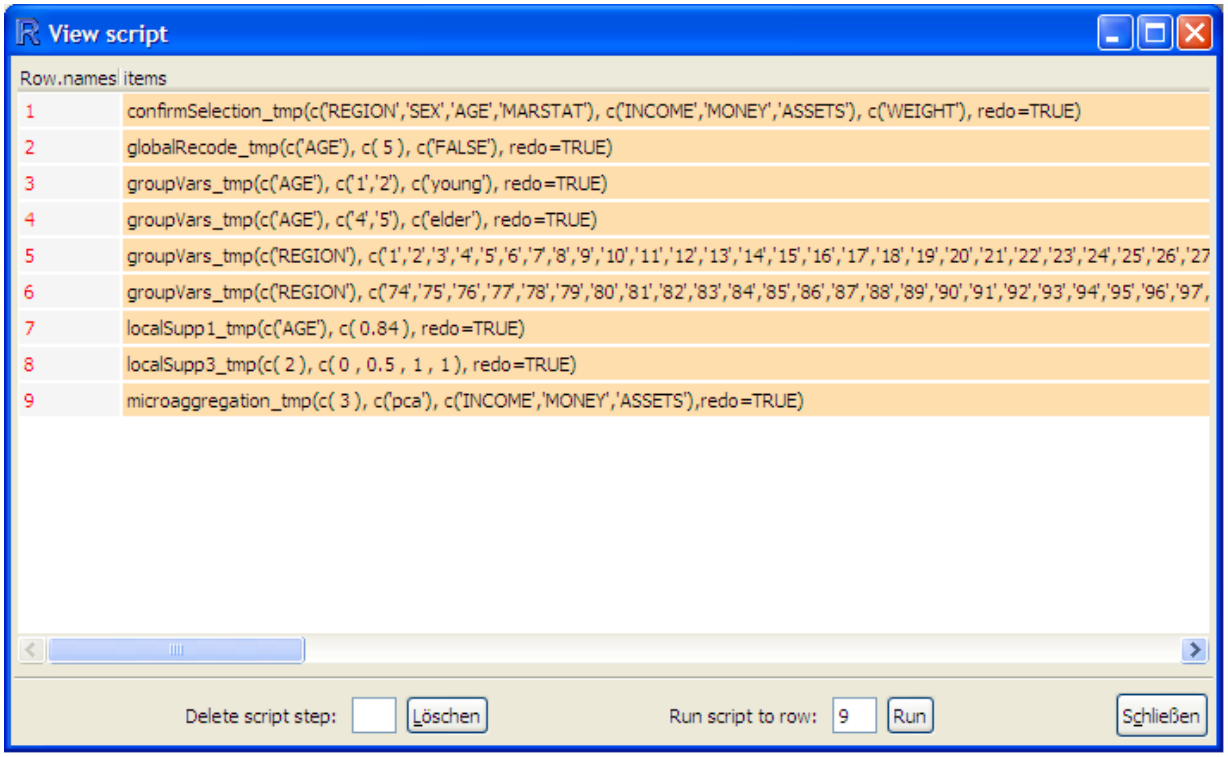

FIGURE 2. Script which is automatically produced by the GUI in order to provide reproducibility. The code can be modified, saved, run or run only up to a specific line of the code. The actual script includes the steps of anonymisation which were applied to the μ-Argus test data set.

IV. CONCLUSION

The GUI provides an extension to the package `sdcMicro` (Templ 2009). The developed GUI makes `sdcMicro` accessible to a wider range of users including ones which are not familar with the software system R. The user has access to all basic functions for microdata protection by using this GUI.
Interactivity is provided by automatic displaying of the main results which are updated after every operation carried out by the user automatically. Reproducibility is provided by storing all the users actions in a script which can then be saved, modified and/or reloaded.
It is also easy to extend the GUI, e.g. when new functions in the core `sdcMicro` package will be included.
Work in progess is to describe the internal object handling of the GUI to make it more easy for other users and researchers to enhance the GUI.

References

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

Hundepool, A., A. Van deWetering, R. R., L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing (2008). μ-ARGUS version 4.2 software and users manual. <http://neon.vb.cbs.nl/casc>.

Meindl, B. and M. Templ (2007). The anonymisation of the CVTS2 and income tax dataset. an approach using R-package sdcMicro. In *to appear in: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality. Monographs of Official Statistics.*

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

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

Templ, M. (2009). sdcmicro: Statistical disclosure control methods for the generation of public- and scientific-use files. version 2.6.0. software and users manual. published online. <http://cran.r-project.org/web/packages/sdcMicro/index.html>.

Templ, M. and B. Meindl (2008a). 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. ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3_10.

Templ, M. and B. Meindl (2008b). Robustification of microdata masking methods and the comparison with existing methods. *Privacy in Statistical Databases. Lecture Notes in Computer Science. Springer 5262*, 177–189. ISBN 978-3-540-87470-6, DOI 10.1007/978-3-540-87471-3_15.