sdcApp – a "shiny" new GUI for sdcMicro

Bernhard Meindl (bernhard.meindl@statistik.gv.at)[1], Alexander Kowarik[2], Matthias Templ[3]

Keywords: anonymization, graphical user interface

1. MOTIVATION

Anonymization of micro data is often a non-trivial task. The R [1] package sdcMicro [2] allows to create safe micro data sets that meet confidential requirements. However, the usage might be difficult for non-experts in R. Thus, a graphical user interface (GUI) for sdcMicro targeted for users with low or no expertise in R has been made available on CRAN in form of the package sdcMicroGUI [3] since 2010.

sdcMicroGUI allowed to apply to create problem instances that could be used within sdcMicro. However, there were some issues with sdcMicroGUI that could not be easily fixed. Instead, the idea was born to create a new interface from scratch using modern technologies. Thanks to funds from the World Bank Group (http://www.worldbank.org) and the Departement for International Development DfID (https://www.gov.uk/government/organisations/department-for-international-development), this new interface is directly included in sdcMicro since version 5.0.0 of the package. This makes it also easier for (inexperienced) users to start working on data anonymization because no additional packages other than sdcMicro and its dependencies need to be installed to make the GUI work.

The new point-and-click graphical user interface is highly interactive, can import and export data, considers reproducibility of any result, includes "undo" functionality, can deal with hierarchical data structures, sampling weights and missing values, supports instant help, present summaries and plots at any stage of the anonymisation, runs in any popular internet browser, and can work also with large data sets since it calls computationally optimized methods for statistical disclosure control.

2. METHODS

The new graphical interface is based on shiny [4]. Shiny allows to build dynamic web-applications and provides the required functionality to get data from and send data to the active R process. It also allows to deploy the GUI locally. This means that once the interface is started with function *sdcApp()*, a new webpage in opened in the default web browser of the system. The user can then interact with the interface by using standard control inputs such as buttons, drop-down menus, sliders, radio buttons or file-upload fields.

In this browser window, the main sections are accessible with tabs listed in the top-navigation bar. The first step in the anonymization process is always to either upload microdata (from various formats) in tab "Microdata" or by importing a previously exported problem instance from tab "Undo".

[1] Statistics Austria and data-analysis OG

[2] Statistics Austria and data-analysis OG

[3] Zurich University of Applied Sciences (ZHAW)

Once micro data have been uploaded (which are only stored locally of course), users can apply a range on methods to the currently active dataset. As soon as data are loaded, the left navigation menu changes and anonymisation methods can be applied. The possibilities include the conversion from numeric variables to factors, the exploration of variables as well as modifying factors or setting values to missing. It is also possible to deal with hierarchical data (e.g persons living within households) or subsetting data for (rapid) testing.

Once any modifications on the (raw) input data have been applied, the next logical step is to create a problem instance by switching to Tab "Anonymization". Here, the user is presented with an interactive table, in which the important categorical (and optionally) numerical key variables can be selected by clicking into the table. Also, variables that exist in the micro data but should be excluded from the anonymized data set can be specified. Furthermore it is possible to select a variable that holds sampling weights or variables that can later be post-randomized. In case of any wrong choice, the user is presented with automatic feedback with pop-up windows stating the specific error which allows the user to easily solve the problem. One of the nice things on this page is that on the right hand side of the screen, users can obtain summary information in form of a plot and a numerical summery of any possible variables. The plot and the statistics presented to the user depend on the scale of the variable. This helps users to identify variables that should be selected as (numerical) key variables.

If a problem instance has been created, the layout of Tab "Anonymize" changes to a three column layout. In the left sidebar, a navigation using buttons for possible anonymization steps is shown while in the centre the currently selected content based is presented. Finally, on the right hand side important information about the current problem is displayed. The implemented anonymization procedures that can be selected include obtaining k-anonymity based on the selected set of categorical key variables or the postrandomization of factor variables by either using a random transition matrix or a custom-defined matrix can be specified interactively.

In case numeric key variables have been selected, it is also possible to apply top-/bottom coding, microaggregation, rank swapping to those variables or to add noise. A big improvement of the new GUI implementation is that the GUI supports all choices on parameters for each method as the command-line version supports.

A lot of measures of risk and utility can be viewed in Tab "Risk/Utility". The content of this section updated automatically whenever anonymization methods are applied to the current problem. This part of the interface also allows comparing variables before and after the anonymization.

Furthermore, in Tab "Export Data" it is not only possible to write the current state of the anonymized dataset to various file formats but also create and export an anonymization report that summarises (in varying detail) the anonymization process.

It was taken great care while developing the new interface that even if users specify the anonymization steps by interactively working in the browser, the process is reproducibly. Thus, in Tab "Reproducibilty" the code that has actually been run is shown along with some comments. The code can also be saved to a file for documentation purposes. Users can also export the current problem instance with all modifications up to this point to a file on disk which can be imported at a later time and the anonymization process can be restarted.

Another nice feature is available from Tab “Undo”. Here users can – if possible – undo the last anonymization step. This is useful for example if some parameters that have been selected turn out to be not optimal. If it is not possible to undo a step (for example because no anonymization method has been applied), it is only possible to import a previously exported problem instance from this Tab.

3. RESULTS AND CONCLUSIONS

The new shiny-based GUI is a huge improvement in usability compared to the graphical user interface that has been available in R package sdcMicroGUI and it represents – together with the methods in sdcMicro - a state-of-the-art tool for statistical disclosure control on micro data. It also contains several enhancements that were missing in sdcMicroGUI. One of these enhancements is that the GUI is now also fully working on small devices such as netbooks. This is due to the fact that for the GUI a responsive css-theme based is used. Due to the fact that the new interface is based on shiny which is actively developed, supported and used in many organisations, it is much easier to maintain and possibly extend the functionality of the GUI even further in the future. The new GUI has been extensively tested during development and has shown to be very robust.

The aim when developing the GUI was to provide an interactive way to perform anonymization of micro data that can be available in different file formats. The GUI features an implicit navigation guide in the sense that depending on the current state (eg. micro data have been loaded or not, a problem instance has been defined or not, …) the content of the Tabs changes. Also, throughout the interface, the user is provided with pop-up messages that appear when the mouse arrow is moved over ?-signs. These pop-up windows contain, for example, additional information about specific parameters or explain the steps the user has to perform to be able to continue with the anonymization process.

In case the user wants to restart the anonymization process from scatch, this is also easily possible by either resetting the current problem instance (keeping the underlying micro data) or by removing the micro data itself which automatically also removes the current problem instance. The anonymization process itself is also very interactive since all information available within the GUI is automatically updated whenever the underlying problem instance changes. Therefore, whenever a user switches for example to the Tab “Risk/Results”, he will always be presented with measures and values based on the actual state of the anonymization process.

To conclude, the new shiny-based user interface of sdcMicro that is directly included with the package helps data analysts that are non-experts in R to apply anonymization techniques available from package sdcMicro easily.

REFERENCES

[1] R Core Team, R: A language and environment for statistical computing. (2016), URL: https://www.R-project.org.

[2] Matthias Templ and Alexander Kowarik and Bernhard Meindl, Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro, Journal of Statistical Software 67(4) (2015), 1-36.

[3] Alexander Kowarik and Matthias Templ and Bernhard Meindl and Francois Fonteneau, R-Package "sdcMicroGUI" (2010), URL: https://cran.r-project.org/package=sdcMicroGUI.

[4] Winston Chang and Joe Cheng and JJ Allaire and Yihui Xie and Jonathan McPherson, shiny: Web Application Framework for R (2016), URL: https://cran.r-project.org/package=shiny