A new shiny GUI for sdcMicro

Bernhard Meindl*, Alexander Kowarik*, Matthias Templ**, Matthew Welch***, Thijs Benschop****

* Methods Division, Statistics Austria, Vienna, Austria

** ZHAW - Zurich University of Applied Sciences, Winterthur, Switzerland

*** World Bank, Development Data Group, Washington, USA

**** Humboldt University Berlin, Germany

Abstract. The application of many anonymization methods is complex and requires knowledge of the methods and access to suitable tools for implementation. For users comfortable with using R, the package **sdcMicro** [1] provides a tool for the application of a comprehensive suite of methods commonly used and described in literature on disclosure control. Users not familiar with **R** [2], but who have an immediate need for tools to anonymize data, would benefit from a friendly Graphic User Interface (GUI) for the sdcMicro package. To provide a GUI environment for the non-**R** user we have developed a **Shiny** [3] application called sdcApp, which is included in the sdcMicro package. Users of the GUI are able to implement the most widely used anonymization methods present in the **sdcMicro** package without requiring in-depth knowledge of **R**. In addition to the anonymization methods implemented in the sdcMicro package, the GUI offers a comprehensive set of risk and utility measures. This includes functions to measure, visualize and compare risk and utility throughout the anonymization process. The GUI also helps users by producing reports on the methods used in formats suitable for internal and external audiences and saves the underlying **R** code to ensure reproducibility. For users of other statistical packages, the GUI supports importing and exporting microdata in several formats (STATA, SAS, SPSS, R). Like **R**, **sdcMicro** is open source and available in the CRAN Repositories and on **GitHub**. This paper describes the features available in the new GUI and its potential to lower barriers to the application of disclosure methods in agencies where users have limited knowledge of **R**.

1 Anonymization with sdcMicro

sdcMicro [1] is an **R** [2] package which provides the necessary methods for the preparation of anonymized microdata suitable for dissemination. In 2010, a graphical user interface (GUI) for **sdcMicro** was developed, which was available in a separate package called **sdcMicroGUI**. The new GUI included in **sdcMicro** is a major improvement over this interface in terms of user-friendliness and available functionality. The GUI is developed using the technologies provided by the **Shiny** (Chang et al. 2017) package. **Shiny** allows the power and possibilities of **R** together with web technologies to be combined in order

to create dynamic user interfaces that are run in a web browser. With the support and funding of the World Bank and the United Kingdom Department for International Development (DfID) a new GUI based on **Shiny** was developed. The new GUI is available from CRAN as part of the **sdcMicro** package.

The new GUI has the potential to open the power of **R** and **sdcMicro** to non-**R** users. In addition to all the main functionality of the **sdcMicro** package, a number of features from other **R** packages for measuring utility and visualizations are also included in the GUI. The aim being to allow users to complete the whole SDC process without leaving the sdcApp environment and without the need to revert to command line **R**.

The available anonymization methods include:

- Methods to anonymize categorical variables such as *k-anonymity*, *suppression of values in high-risk records*, *postrandomization* and *shuffling*
- Methods to anonymize numerical variables such as *adding noise*, *microaggregation*, *rank swapping* and *top/bottom coding*.

In addition to the anonymization methods, the GUI also allows users to visualize the risk and compare the original to the anonymized data, identify high risk combinations as well as to export data and create reports. To facilitate and ensure reproducibility, the application also allows the current SDC problem to be exported for use at another time. The exported problem includes the data, settings and results. Help is provided in the form of tooltips and provides users with additional information on methods and the setting of parameters.

The package **sdcMicro** is currently developed on **Github** as part of the *sdcTools* framework. The code is open source and may be modified and changed. The authors welcome any contributions, feature-requests or bug-reports to the software.

2 Description of the GUI

To use the application, the user needs to install the latest version of the **R** software from the CRAN website. **R** is available for free for the following platforms; Linux, Windows and Mac OS X. After installing **R**, the user installs the **sdcMicro** add-on package as well as other packages used by **sdcMicro** (so-called dependencies). This is done automatically by running the command install.packages("sdcMicro") in the **R** console. After installing **sdcMicro**, the package **sdcMicro** needs to be loaded with the command library(sdcMicro). The GUI is then launched with the command sdcApp(). The application launches in the default web browser of the system. The user can interact with the application by using control inputs such as buttons, drop-down menus, sliders, radio buttons or text input. No further interaction with the **R** console is required.

The GUI consists of seven tabs that can be navigated using the top navigation bar. Table 1 gives an overview and description of the tabs in the GUI.

Screen tab name	Description
About/Help	Help and general settings
Microdata	Load and prepare dataset
Anonymize	Anonymization methods
Risk/Utility	Risk and utility measures
Export Data	Export datasets or reports
Reproducibility	View/Export R script
Undo	Undo steps in anonymization process

Table 1: Overview of screen tabs in the **sdcMicro** application

After the tool has been started, the application opens in the tab *'About/Help'*, which is shown in Figure 1. This screen provides basic information on using the application as well as the option to specify where output will be saved. Also, being an open source project, information and links are provided for providing feedback and contributing code to the project through **GitHub**. All data are loaded locally into **R** and the web browser is only used to communicate with **R**. An internet connection is not required during the anonymization process.

The first step after opening the application is to either load microdata or load a previously saved problem instance. Data from different statistical software packages (such as SAS, STATA or SPSS) are supported. For testing and demonstration purposes, the application includes two test datasets (testdata and testdata2). Figure 2 shows the upload screen for csv files with the options to be set. After loading the dataset into the application, the *'Microdata'* tab shows the loaded dataset and allows the user to explore, manipulate and prepare the data for the anonymization process. This is shown in Figure 3. The available functions for data exploration depend on the variable type. Examples are tabulations, histograms, mosaic plots and standard summary statistics. Examples of data preparation are variable type conversion or setting missing values to **R** system missing values. This tab also provides functionality to deal with datasets with a hierarchical structure, such as household surveys as well as to use only a subset of the full dataset for quicker testing.

After loading and preparing the dataset, the user can navigate to the *'Anonymize'* tab to select the key variables and create the anonymization problem instance. This is shown in Figure 4. In an interactive table the user can select categorical and continuous key variables, the sampling weight, a hierarchical identifier, variables suitable for the PRAM method as well as variables to be removed from the dataset before release. If an invalid choice is made, e.g., the user selects a variable both as key variable and sampling weight, the application provides feedback in the form of a pop-up window stating the variable and the error. A number of parameters can also be set; these include setting the parameter alpha used for the k-anonymity calculation as well as setting a seed for the random number generator for use in probabilistic methods. In the right sidebar on this screen, the user can browse summary information of the variables, such as frequency counts, histograms and summary statistics.

Once the user has made the selection of variables and clicked the button *'Setup SDC problem'*, the *'Anonymize'* tab shows a summary view of the SDC problem, including the variable selection and selected risk measures. In the left sidebar the anonymization

methods can be selected. They are grouped by variable type. For categorical variables, the methods global recoding, local suppression to achieve k-anonymity and PRAM can be selected. For numerical variables, the methods top-/bottom-coding, microaggregation, rank swapping and noise addition are available.

After selecting a method, the user is presented with a three column page as shown in Figure 5: on the left the methods that can be selected, on the right a summary overview of the current problem is shown, the main part presents options and parameter settings for applying the currently selected method. The parameters and options available to the user in the GUI are the same as those available from the command line in **sdcMicro**. For each method a brief description is included and for most parameters a help button provides more information.

In order to compare risk and measure utility (information loss) before and after applying a specific method, the tab *'Risk/Utility'* presents the user with a range of detailed risk measures and selected utility measures; as shown in Figure 6. All the risk and utility measures are automatically updated after an anonymization method is applied.

On the tab *'Export Data'* the user can browse the anonymized data and export the dataset in the data format of their choice. At any point in the anonymization process, the current data can be exported to perform analyses using the users' software of choice. This is useful, for example, for computing benchmark indicators and for the assessment of information loss. This can be particularly convenient if code for generating indicators and tables is already available in another software format. From the same tab the application will also generate both internal and external reports of the anonymization process.

The tab *'Reproducibility'* shows a commented and downloadable **R** script that is ready to run in the **R** console and can be used to recreate the SDC problem and recreate all the steps. This is shown in Figure 7. This functionality also provides users who, at some point, wish to move to using the **sdcMicro** package from the command line, an easy way to learn the methods and commands available. On the same tab, the user can also export and reload the **R** workspace. The workspace contains the data as well as all settings, selections and results. This feature is useful as a backup as well as to restore and continue working at a later point.

Finally, the tab *'Undo'* as shown in Figure 8 allows the user to undo the last anonymization step. This is useful when exploring the best methods and parameters in a trial-and-error fashion. In order to revert to a previous state more than one step back, the user can import a previously saved problem instance.

3 Conclusion

The recently developed Shiny-based GUI makes the complete set of SDC functions and tools included in the **R** package **sdcMicro** available to a wider group of users that are

not proficient in **R**. The GUI includes the full functionality of the **sdcMicro** package, including all the methods, parameters and options, as well as additional functionality from other **R** packages, such as visualization of results and data manipulation. This makes it a complete solution for the anonymization of microdata. The GUI is user-friendly and provides the functionality to import and export to all the major statistical package file formats. The Shiny-based interface lowers the barriers to entry to **sdcMicro** for non-**R** users and brings the ability to apply the most widely used SDC methods to a larger audience. Unencumbered from having to know how to program the methods, it is our hope that this will allow more agencies to apply appropriate SDC methods which will lead to greater and safer release of microdata.

The feedback on the recently developed Shiny-based GUI has been positive and has been used in a number of training sessions by the authors. Development of the interface is ongoing and improvements will continue to be made.

References

[1] Matthias Templ, Alexander Kowarik, and Bernhard Meindl. Statistical disclosure control for micro-data using the R package sdcMicro. *Journal of Statistical Software*, 67(4):1–36, 2015.

[2] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[3] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. R package version 1.0.0.

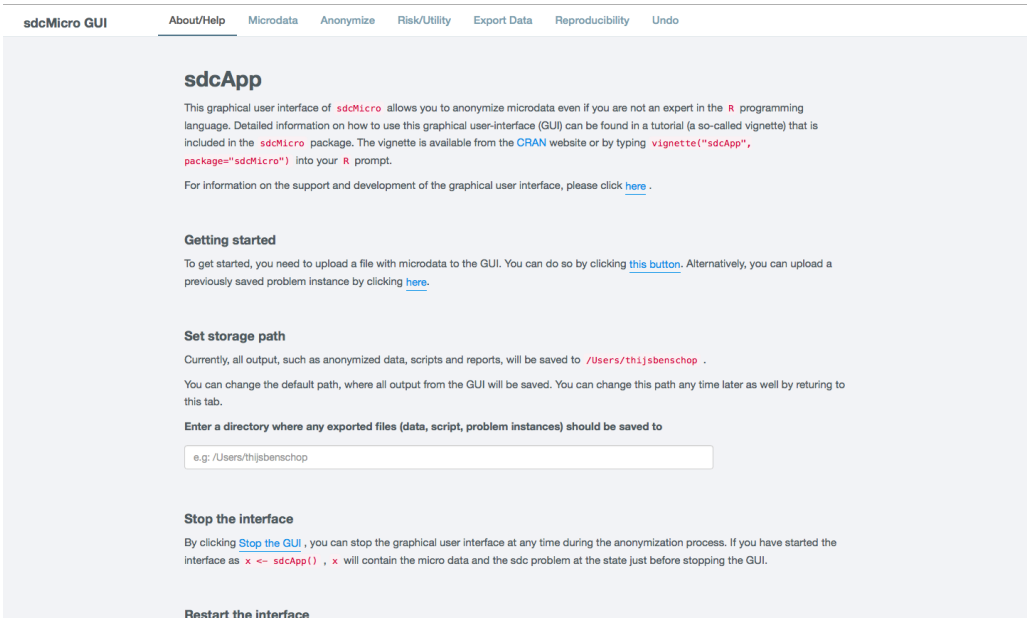

Figure 1: About/Help tab after launching the GUI

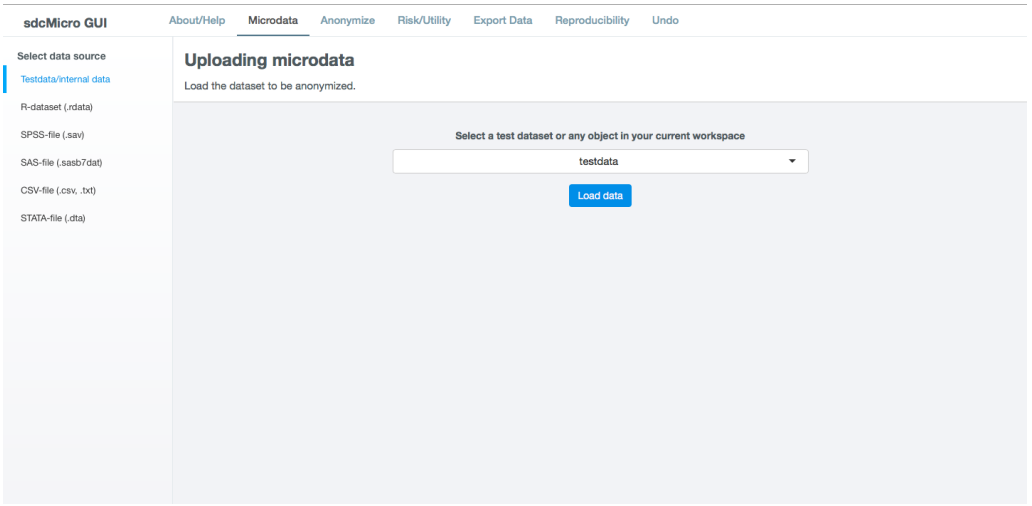

Figure 2: Microdata tab to load a csv-file

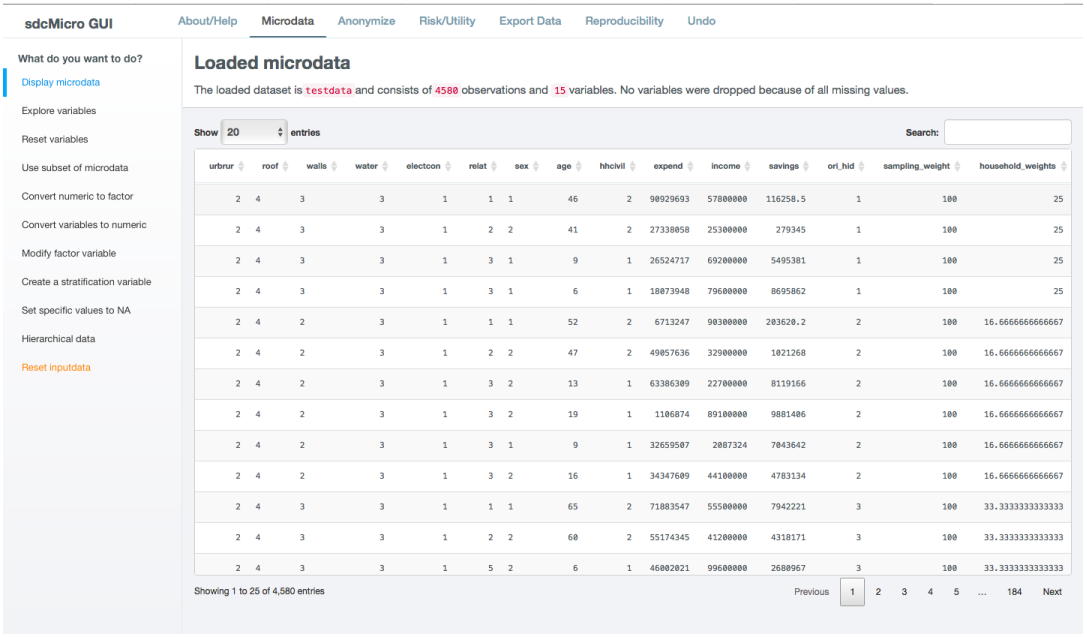

Figure 3: Microdata tab after loading the microdata

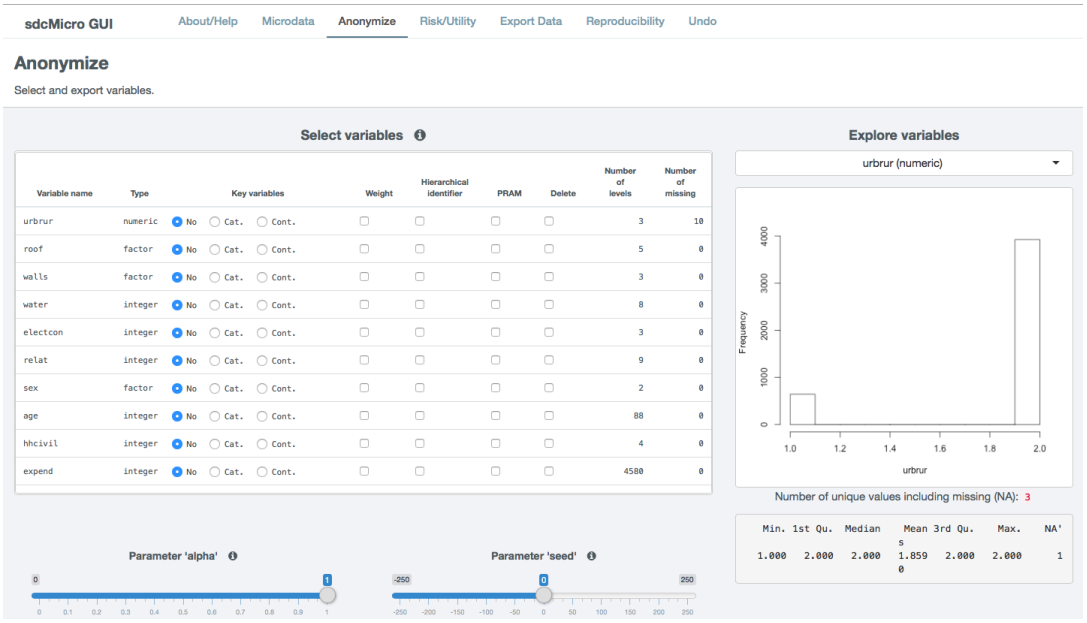

Figure 4: Anonymize tab to setup an SDC problem

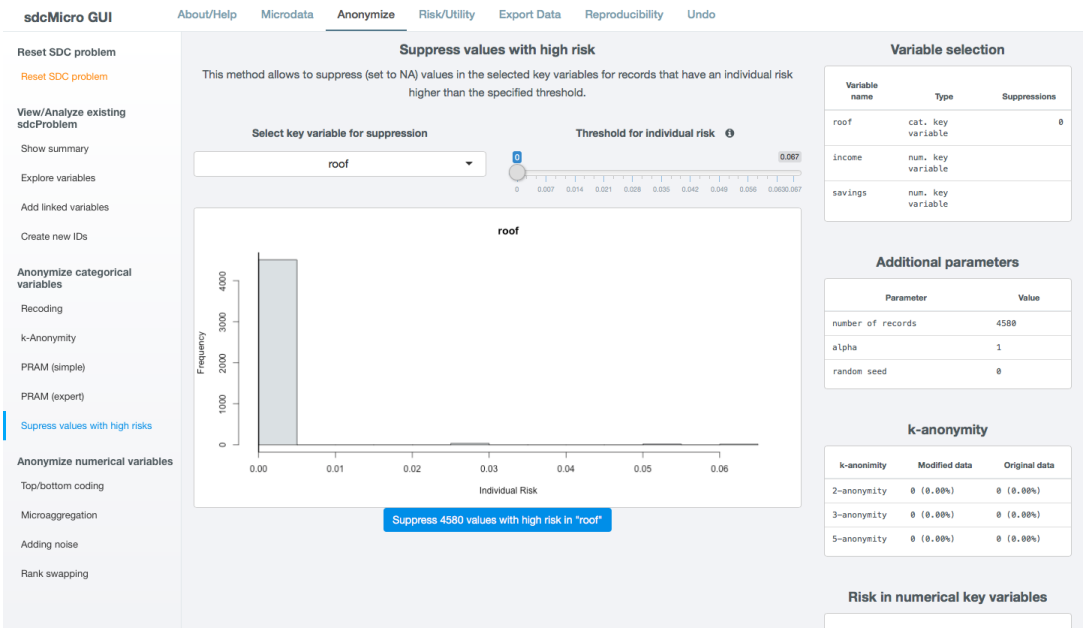

Figure 5: Anonymize tab to suppress values in key variables in high-risk observations

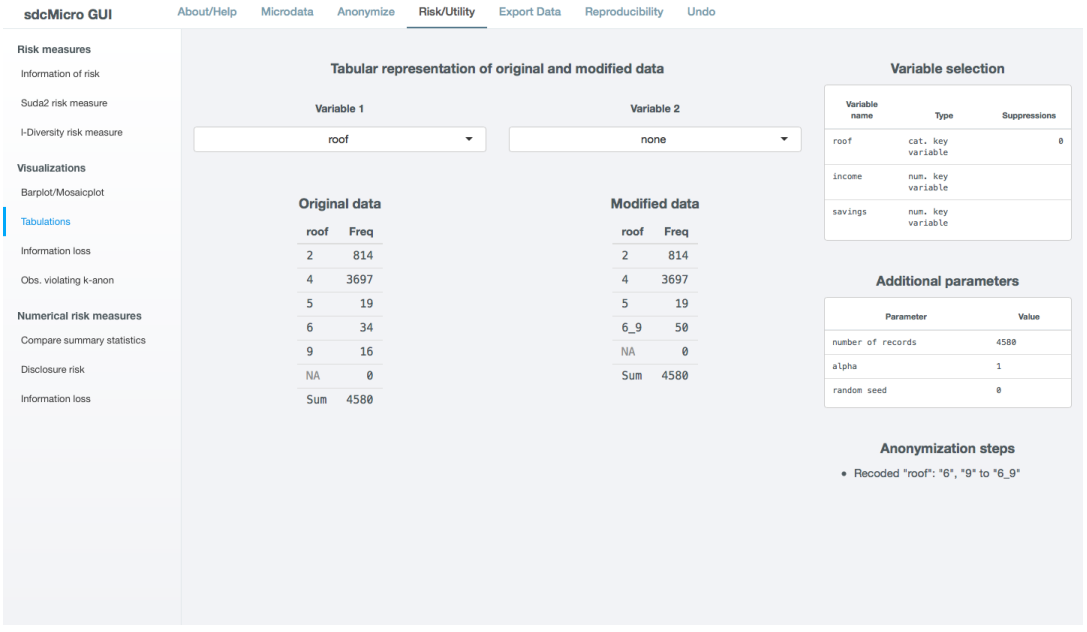

Figure 6: Risk and Utility tab - tabulations before and after anonymization

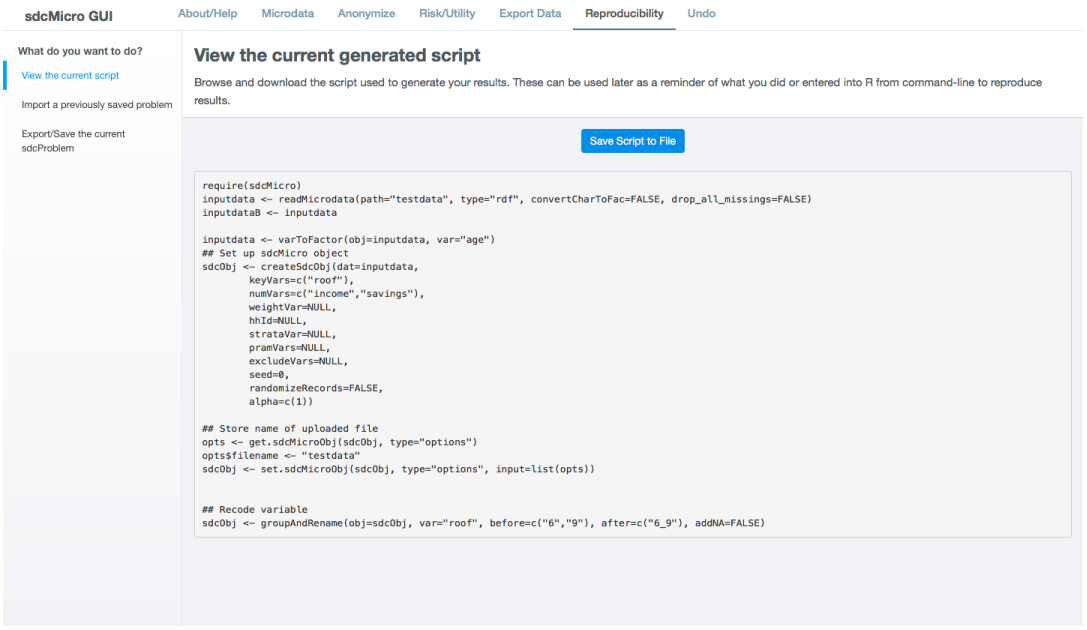

Figure 7: Reproducibility tab showing the **R** script

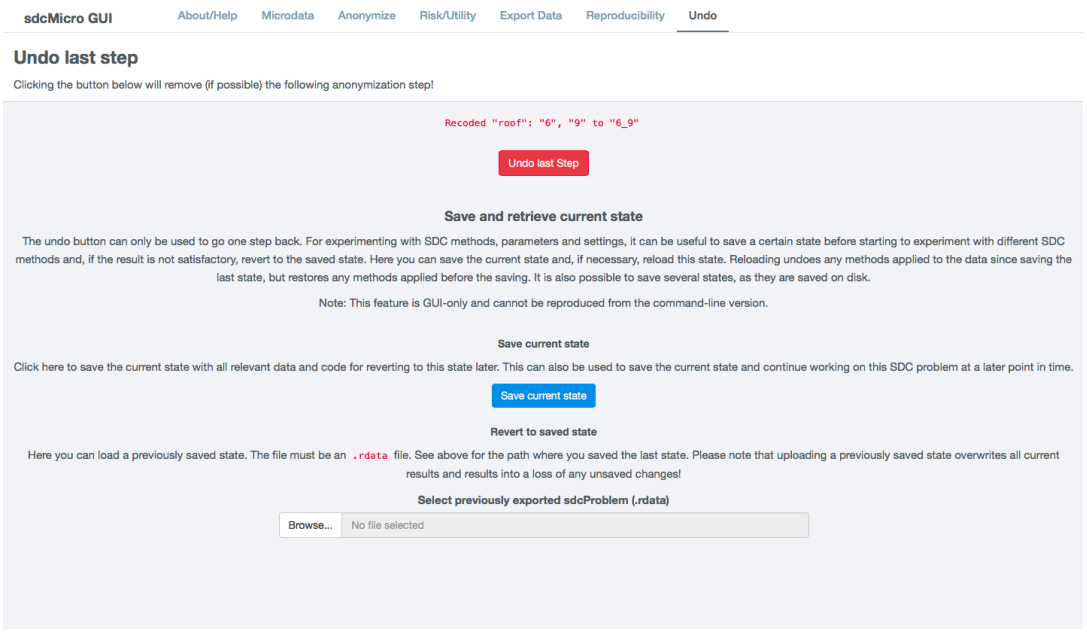

Figure 8: Undo tab allowing to undo the last anonymization method