Anja Crößmann
Federal Statistical Office Germany
Research Data Centre

Wiesbaden, November 2009

ESSnet SDC Project

Task 5 Improvement of software for microdata
5. a (1) Standardised anonymisation of microdata sets

Report on testing µ-ARGUS 4.2

1. Introduction

Guaranteeing confidentiality and discretion of microdata is one of the most important duties of statistical surveys. Without the preservation of mutual trust between respondents and data producers data acquisition is in fact not thinkable. µ-ARGUS 4.2 applies state-of-the-art Statistical Disclosure Control (SDC) techniques to microdata and could be a helpful tool in process of data disclosure. The following report refers to testing µ-Argus version 4.2, which is intended of the ESSnet SDC project in task 5 "Improvement of software for microdata" and was carried out on the release of 5.a (1) "Standardised anonymisation of microdata sets". Main aim of the test will be to establish whether µ-ARGUS 4.2 can be used as an instrument for the standardised anonymisation of large microdata sets and to eventually identify problems with respect to the integration of the software into the production infrastructure. This question will be investigated using the German microcensus from 2001.

2. German microcensus

The German microcensus is an annual random sample survey with a sampling fraction of 1% of the population. It is the largest household sample survey in Germany. In the sample a total of about 390,000 households with 830,000 individuals are covered by overall eight hundred variables.

Designed as a multi-topic survey, the German microcensus integrates many different subject fields. Its basic annual programme provides, in particular, for the collection of individual data (age, sex, nationality etc.), data on family and household relationships and, in addition, data on a person's residence, (former) employment and economic status, educational and (former) vocational attainment, sources of living and data on the obligatory pension insurance as well as data on the amount of an individual's or household's net income. A quadrennial supplementary programme is used, first of all, to collect data on job and training commuters, on the housing situation, on health insurance and on a person's state of health and disablement status.

Access to German microdata sets can be granted through Public Use Files (PUF) and Scientific Use Files (SUF). PUF's are absolutely anonymised and standardised microdata, that are available to all those who are interested both within the country and abroad. Due to anonymisation, PUF's contain only selected variables. In most cases, detailed regional breakdowns cannot be made on the basis of PUF's. SUF's are de facto anonymised microdata (similar to the so called "Microdata file Under Contract", MUC), the research data centres offer to scientific community users for off-site use at research institutions which are governed by German law. These data have a far greater information potential than PUF's and they are well-suited for large part of the scientific data analyses.

3. Statistical disclosure control methods

Most of the variables of the German microcensus 2001 are categorically scaled, some are dummy variables with 'Yes/No' as possible answers and a few are metrically scaled. To assure, that there is no way to make regional or personal breakdowns on the basis of PUF's or SUF's, the main application of statistical disclosure control methods is a univariate (and in some cases bivariate) aggregation of variables with a high degree of subject-related details by global recoding, local suppression or top and bottom coding. According to this the test of using µ-ARGUS 4.2 as an instrument for the standardised anonymisation of large microdata will be basically limited to the relevant features, to check if the output is adequate to the output produced by recoding with SAS or SPSS.

4. Testing µ -ARGUS

4.1. Input

One problem may occur due to the number of cases implemented in the German microcensus. So we have to be sure that it is possible for µ-ARGUS 4.2 to work with such an amount of cases (about 830,000 records) and – if yes – if work is faster and more comfortable than recoding variables in the usual way by using a server-based SAS software. Unfortunately in µ-ARGUS 4.2 is no direct implementation of SAS-files available yet. Thus a second problem could be the interaction between SAS and µ-ARGUS 4.2. Before beginning the main test of the features these issues have to be checked and potentially a workaround has to be found to solve problems.

µ-ARGUS 4.2 is able to read most cross platform ASCII file types like .dat, .asc or comma separated files but there is no description given, how to let create a metadata describing ASCII file for files like this by an external program. With a high three-digit amount of variables in the microdata set compared to the earnings of µ-ARGUS manually specifying variables is disproportionate to the efforts. At least there should be an opportunity to handle with multiple variables at once by marking more than one variable at a time. In the process of specifying imported data files from SPSS furthermore there should be a button to mark all variables in the data set at once.

µ-ARGUS 4.2 seems not to be able to identify categorical variables even if missing values are specified in SPSS or in a manually written metadata file. Actually the source of this

problem could be version 13 of SPSS which was used in this test. Maybe the interaction with µ-ARGUS 4.2 works better with a newer version. Besides, µ-ARGUS 4.2 seems also not able to define missing values in SPSS imported variables automatically nor it is able to define empty cells as missing values. Because µ-ARGUS 4.2 requires specified missing values to accomplish local suppression, working with SPSS imported data is not possible. If it is technically possible, to manipulate SPSS metadata within µ-ARGUS, this feature should be implemented.

With the comprehensiveness the German microcensus has manually preparing the whole data file in a way µ-ARGUS 4.2 can work with is very time consuming. Overall it takes more time to get the data into µ-ARGUS 4.2 than to anonymise the data in the conventional way by recoding with a server-based SAS or SPSS system.

Finally execution of data input could be made by a SAS macro written by Johan Heldal from Statistics Norway which grants the possibility to extract metadata out of common SAS data. The macro compiles a semicolon separated file and a compatible rda-file. The original data file of the German microcensus 2001 is nearly 150 MB big, consists of 84 variables which need to get anonymised and 787,465 records. To specify a combination with only one variable and a threshold of 10,000 records nearly takes 30 minutes. Because of this, combinations with more than one variable are not tested. Note that µ-ARGUS 4.2 is not server-based in our case but installed in a local drive of an Intel Core 2 Duo with 2.3 gigahertz and 1 GB RAM. While exploring file the CPU-load is about 50%, so it does not even use the whole range. Apparently µ-ARGUS 4.2 is not able to handle with such an amount of variables and records.

4.2. Process of statistical disclosure

A big deficit is the stability of the program. Even with a special produced slight dataset it crashes several times while testing. Mostly it crashes by specifying combinations or produces an error that the exploring file length is wrong at a specific point in the file. Manually revision of the file does not solve the problem and more detailed description of the error is nowhere to find.

Recoding itself works fine; the handling is very intuitive and fast. It would be fine, if frequencies below the threshold would not just be labelled with the number of unsafe records in each dimension but with a specific colour. Furthermore it would improve the handling if it would be possible to scroll the main window in the background while the window of the respective modifying method is active. The opportunity to rename and recode in one step would be another good improvement (something like: 01: 05-10; Label: Test).

4.3. Output

Finally the program compiles a file with the extension .saf which seems to be a usual text file; there is no continuing explanation in the manual. Neither SAS nor SPSS nor STATA are able to read such files without more or less complex importing like a usual txt-file.

Again Johan Heldal's external macro offers a way how to import .saf in a comfortable way into SAS.
If it would be possible to get an anonymised semicolon separated output file that could be used in every kind of analysis software without much effort, implementing this opportunity would be much more convenient than the current solution.

5. Conclusion

In summary µ-ARGUS 4.2 in its current state of the art seems not useable for purposes of large data sets. Especially for purposes of traditional disclosure control methods like global recording, local suppression or top and bottom coding, the efforts to get the data ready and µ-ARGUS started is disproportionate to the gains of the program. Obviously the program seems not fully developed yet to work with data files with a dimension like the German microcensus. Even if it would be able to run processes in an acceptable span of time, prearrangement takes too much time to detach conventional recoding methods. Furthermore the manual is not detailed enough to be helpful in terms of problem solving.

Report on several problems while testing µ-ARGUS 4.2:

- No direct import of SAS data files possible
- No description how to let create metadata files automatically
- Insufficient import of SPSS data files
- Insufficient processing with large data files
- Stability
- Continuing processing of output file not directly possible

Proposals to improve µ-ARGUS 4.2:

- Software:
  - Implementation of the possibility to import SAS system data files
  - Possibility of handling with multiple variables at once by marking more than one variable at a time
  - Improving of the SPSS data files import to make sure µ-ARGUS is able to identify missing values
  - Implementation of the possibility to manipulate missing values of SPSS imported data within µ-ARGUS
  - Highlight unsafe records in a specific colour
  - Possibility to work in main windows while another windows of the program is active
  - Possibility to create csv-files as output files
  - Possibility to recode and relabel in one step
  - With authorisation by the programmer, Johan Heldals helpful SAS macro should be added to the zip-file
  - Open extern metadata systems for µ-ARGUS

- Manual:
  - Description how to export metadata out of SAS and – if possible – out of SPSS and STATA

- Detailed description of errors µ-ARGUS 4.2 could produce while processing
- Detailed description how to continue with the output file, the easiest way to import it in SAS, SPSS and STATA

References

A. Hundepool, A. van de Wetering, R. Ramaswamy, L. Franconi, S. Polettini, A. Capobianchi, P-P. de Wolf, , J. Domingo-Ferrer, V. Torra, R. Brand, and S. Giessing 2009: µ-ARGUS version 4.2 User's Manual, http://neon.vb.cbs.nl/casc/mu.htm