CASC PROJECT

Computational Aspects of Statistical Confidentiality

Report on the testing of μ-Argus Version 3.2

Work Package 6 of the CASC project

May 14, 2003

Deliverable 6-D4

Giovanni Maria Merola

ISTAT, MPS/D

Via C. Balbo, 16

00184, Roma

Italy

Introduction

μ-Argus is a computer program that applies state-of-the-art Statistical Disclosure Control (SDC) techniques to microdata, being developed within the European Project CASC (Computational Aspects of Statistical Confidentiality) with the final goal of creating a tool usable by NSI's as well as other agencies. This testing of μ-Argus is Work Package 6 of the CASC project and was carried out on the intermediate release 3.2. After the final corrections and improvements, this version will become the final release which will be able to serve as a standard tool for applying different SDC methods to microdata usable by different NSIs.

Testing is a crucial phase in the development of software; this testing follows a previous one carried out on version 3.1 in 2002 (Deliverable 6-D1). These testing have been designed for evaluating μ-Argus especially with respect to its integration in the data production process of statistical institutions. That is, having in mind the final user and his/her needs. Hence, the testing has been designed to check for bugs and limitations, collect suggestions for improvements, assess the clarity of the documentation, verify the portability (input/output formats, platform required etc.) and user friendliness of μ-Argus. In this way, scores and comments expressed by testers become a vital source of information to help developers not only to fix bugs and malfunctioning, but also to improve the program towards meeting the requirements of the end users. For this reasons testers were selected among potential users of the software, privileging the quality of testing over the quantity of tests run.

Description of the software

μ-Argus 3.2 (see A. Hundepool (2001) The CASC Project. *Proceedings of the ECE/Eurostat Work Session on Statistical data Confidentiality*, 2001 Skopje, Macedonia.) is a program for applying Statistical Disclosure Control (SDC) to microdata, that is to individual records containing confidential information, that runs under different Microsoft Windows platforms (95/98 and NT/2000). The program and the manual are freely available on the CASC web-page at http://neon.vb.cbs.nl/casc/.

The program is designed in such a way that SDC can be applied in steps, comparing the effect of different methods and different parameters values in the same run. The first step consists of reading in the data and the metadata. A new feature of version

3.2 is the possibility of reading data stored in comma separated format (csv). The metadata can be edited from within the program using the *modify metadata* option. The combinations of variables to be considered are defined in the *specify combination* window. Also from this window, it is possible to view combinations of key variables displayed as tables. In the second step different criteria for selecting records at risk and for disclosure control can be chosen, such as sampling threshold, individual risk and values for PRAM. The third step is the actual protection of the microdata. In the previous version of μ-Argus three different methods were available: global recoding, top/bottom coding, local suppression, PRAM and perturbation. The following methods have been added to the new version: numerical microaggregation, rank swapping and Sullivan method for noise addition.

The last step is the production of the output which consists of the protected data files and a report in HTML format. The report gives statistics on the data protected and details on the methods and parameter used for the protection of the data.

Outline of the testing set up and questionnaire

Version 3.2 of μ-Argus was tested by 9 testers: 5 of them were officially assigned; the other four were volunteer testers external to the CASC project, three of which from "non-EU" countries statistical institutes. Another tester from the Bulgarian Statistical Office volunteered his comments.

Testers are supposed to have some knowledge of SDC for microdata and access to the existing literature. A set of test data was provided in order to have standardized evaluations but testers were asked to try the program on other real data produced by their institution.

The questionnaire was designed as a MS Excel worksheet and it was sent to testers through email. A preliminary version of the questionnaire was sent to all testers for comments and, a week later, the reviewed version was mailed out, together with instructions and installation files. The completed questionnaire with the results of the testing had to be returned a month later.

The questionnaire is structured in 7 sections, following an ideal microdata protection operation path using the program. The sections were: 1) tester's identity; 2) data importing; 3) data specification; 4) data modification; 5) output of protected files; 6) documentation and 7) general remarks. Some questions required an evaluation with four possible scores (I=insufficient, S=sufficient, G=good and E=excellent), other a yes/no answer and the rest just comments. All questions admitted the possibility of adding comments and suggestions. Further details can be found in the questionnaire with the instructions in Appendix A.

Overview of the testing results

In this section we give a summary of the testing results for each section of the questionnaire. Detailed summaries of scores and comments for each question can be found in the summary sheet in Appendix B.

The document in Appendix B reports summary scores and comments for each question as well as individual tester's testing reports: columns F to K give the frequencies of the scores given to each question, scores less frequent than 3 are highlighted in Yellow, frequencies of 3 or 4 are highlighted in Orange and higher frequencies are highlighted in Red. Column L gives a "satisfaction" indicator obtained either assigning the value zero or one to the scores "No" and "Yes" (depending on whether the Yes or No answer imply satisfaction) or assigning the numerical values 0, 1/3, 2/3 and 1 to the scores I, S, G and E, respectively. In this way, the value of the satisfaction index ranges from 0 to 1; the closer it is to 1, the closer the mean judgement is close to satisfaction. Column M shows the number of missing answers and column N gives summaries of the individual comments. Columns from O to W show the scores for each tester. The comments given by the testers are shown in columns from X to AF.

1) Preliminary Issues: Tester's identity and NSI's practices

This section was designed to get an idea of the tester's skills, the equipment they were using and the needs of their institution.

Most testers are statistical researchers with good knowledge of SDC for microdata. All testers had over 100 Mb of Ram, most 256Mb. Some were running windows98, others Windows NT 4.0 and the rest Windows 2000. Some testers showed interest for a UNIX (LINUX) version of the program.

Tester's institutions apply different SDC methods. Most of them apply local suppression, global recoding and top/bottom coding; some use also rounding, noise addition, sub-sampling and swapping. The data used for testing were mainly social but also business and price data were used. Samples of different sizes were used, from as few as 1,391 to as many as 250,000

The installation process gave a warning message about the registration of the mfc42d.dll. However, if the message was ignored the installation ended successfully.

2) Data Importing

This version of the program enabled importing data from "comma separated variables" (csv) files. These new formats could be read in without problems but some testers complained about problems regarding the metadata specification and not the reading. Therefore, the 59% of satisfaction registered should really be higher, since the complaints regard another section of the questionnaire. A message announcing the successful end of the process was requested.

The on-line help and the manual for this section scored mostly *Good*, some found that the topic was not located in the right place.

The majority of testers were satisfied by the formats that can be imported. Some testers would appreciate the possibility of importing files in SAS, Excel and other commercial software format. Some others requested the possibility of importing ASCII files with variables separated by tabs or blanks.

3) Data specification

Option *Metadata*

Some testers required a more flexible tool for managing data and metadata from within the program. Many testers reported run-time errors and program crashes when trying to modify the metadata from the *specify metadata* window. Some of the requirements for the data, such as the field length, are inconsistent with the csv format. Undo and Cancel buttons were requested by most testers. The satisfaction for this option was 44%. Better definition for categorical and numerical variables was required.

Option Combinations

This option was rated *Good* by most testers and summary window with the tables produced is well understood by most testers. The manual and on-line help were required some revision. The overall judgement for this section is 50% satisfaction, mainly because the metadata specification window is unsatisfactory.

4) Data modification

This part of the testing regards the selection of records at risk and data protection methods.

Show Table collection

This window was found clear and complete. Some minor changes were required.

Option Global recode

This procedure gave no problems. Some minor changes were required.

Option PRAM Specification

This option was found satisfactory but for the fact that the choice of the range percentage must be entered for each variable. The manual was found not very informative about the choice of the parameters.

Option Risk Specification

Some testers had doubts on the methodology (manual needs improvement?) the risk graph could be improved and a default threshold value was suggested. The possibility of accommodating different risk models was suggested as a possible development direction. A default value for the risk threshold would be liked by most testers and the documentation needs improvement.

Option Modify Numeric Variables

This collection of methods is rated *Good* by all testers and some improvements were suggested.

Option Numerical Microaggregation

This is a new addition to the program. There seem to be some doubts about what the "optimal method" does and there seem to be some numerical problems. It was suggested the implementation of a microaggregation algorithm that could deal with strata. The documentation needs improvement.

Option Numerical Rank Swapping

This is a new addition to the program. The method is not much used nor likely to be adopted by most. In its present implementation the method conflicts with other protections and does not allow enough flexibility. There are also some numerical problems. Documentation is satisfactory but more examples would be appreciated.

Option Sullivan Masking

This is a new addition to the program. The method is neither used nor likely to be adopted by any of the institutions. In the current implementation the method requires the external program Gauss. The testers that tried the method found several problems with the parameters definition and with the program Gauss. Some other testers could not run Gauss from within μ-Argus at all. The documentation for this method was rated sufficient by the few testers that evaluated it.

Overall the *data modification* section was rated 57% satisfactory. Real time feed-back of changes applied and possibility to retain options entered were required by most testers. Also new protection methods were required.

5) Output of Protected Files

The procedure for the output file was rated *Good* by most testers and the file created was as expected. However, some testers would like to have the suppressed records to be marked differently and others would like to be able to inspect the output file before it is written. Some other improvements were suggested.

Some testers would appreciate the possibility of producing output files in commercial packages formats; others would like to be able to produce the output file in a format different from that of the input.

The report window was found clear but more details were required.

6) Documentation

The on-line help was rated 59% satisfactory. There seem to be problems in opening the help file when the source data are not in the main directory. More examples were required. The search facility was also rated 59% satisfactory but more keywords were requested.

5 testers out of 9 prefer the present format of the manual, divided in two parts, while 3 find this format hard to read. However, all testers find the methodological notes useful. The manual was rated satisfactory but more examples and details on specific usage were requested.

7) Global statements about the program

Overall the organization of the program was rated *Good* but the following improvements were suggested: more general menus with non-appropriate options disabled; undo, cancel and help buttons; visual inspection of changes applied, before the protected file is saved. The majority of testers found the program to be adequate for carrying out SDC a their institution. However, more control on data protection and better documentation were required. Also some methodological improvement towards the full integration with NSI's practices is needed. One NSI would require running the program on a UNIX mainframe system.

Overall this version of μ-Argus was rated *Good* and most testers found the program to be very useful. As major problems outlined, stability and limited capacity of dealing with large files seem to be the main concerns.