Report on the testing of τ-Argus Version 2.1.1

Work Package 6 of the CASC project

Deliverable 6-D2

Giovanni Maria Merola
Maurizio Lucarelli

ISTAT, MPS/D
Via C. Balbo, 16
00184, Roma
Italy

Report on the testing of τ-Argus version 2.1.1

Introduction

τ-Argus is a computer program that applies state-of-the-art Statistical Disclosure Control (SDC) techniques to tabular data, 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 τ-Argus is Work Package 6 of the CASC project and was carried out on the intermediate release 2.1.1

Testing is a crucial phase in the development of software; this testing has been designed for evaluating τ-Argus especially with respect to its integration in the data production process of statistical institution. 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 τ-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 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

τ-Argus 2.1.1 is a program for applying SDC to tabular data, that is, to tables carrying aggregates of individual records containing confidential information. It can run under different Microsoft Windows platforms (95/98 and NT/2000).

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 individual records (the *microdata*) and the description of the variables (the *metadata*). These must be stored in files in a fixed format (.asc and .rda, respectively). The metadata can be edited from within the program using the *specify metafile* option. The second step consists of specifying the tables to be protected, defining the spanning variables, the response variables and the risk criteria in the *specify tables* window; the risk criteria can be chosen among: minimum frequency threshold, dominance rule and *p-q rule*. The third step is the actual protection of the tables. In the present version of τ-Argus the following methods are available: variable recoding and secondary suppression. This last method can be carried out using three different routines: two exact LP-solvers, Xpress and Cplex, which require purchasing a license, and the Hypercube method, a heuristic solver, called GHMITER (R. D. Repsilber 1994. Preservation of Confidentiality in Aggregated Data, Second International Seminar on Statistical Confidentiality, Luxemburg) which can also deal with hierarchical variables (via the HiTas algorithm: Fischietti, M. and J.J. Salazar-González (1998). Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints. Technical Paper, University of La Laguna, Tenerife.) and is implemented freely by the CASC project members. The last step of SDC is the production of the output which consists of protected tables and a report, in HTML format.

Outline of the testing set up and questionnaire

Version 2.1.1 of τ-Argus was tested by 10 testers, three of which, external to the CASC project and from "non-EU" countries, on a volunteer basis. Unfortunately, some tests were returned incomplete and could not be considered.

Testers are supposed to have some knowledge of SDC for tabular data 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 was sent electronically. A preliminary version 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 table protection operation path using the program. The sections were: 1) tester's identity; 2) preliminary issues (installation and data importing); 3) data specification; 4) protection methods; 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 a comment. All questions admitted the possibility of adding comments and suggestions (for scores insufficient or sufficient this was actually required). Further details can be found in the questionnaire with the instructions in Appendix A.

Testing report

An overview of the testing results for each section of the questionnaire will be given in the following section. Detailed summaries of scores and comments for each question can be found in the summary sheet in Appendix B. In this document 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 zero to "No" and one to "Yes" or assigning the numerical values 0, 1/3, 2/3 and 1 to the scores I, S, G and E, respectively. In this way, its values range from 0 to 1, and the closer they are to 1, the closer the mean judgement is to Excellent or Yes and vice versa. Columns M to T show the scores for each tester. Column U contains a summary of the comments given by testers, which are shown in columns from V to AE (these include also the incomplete testing reports). Cells highlighted in violet in column U refer to points that received particularly low scores.

Overview of the testing results

1) 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 for testing and the needs of their institution.

Most testers were statistical researchers with good knowledge of SDC for tabular data who were already familiar with τ -Argus. All tests were run on reasonably fast processors with good amount of RAM (minimum 256MB), most of them running Windows NT/2000 and a few Windows 98. Some testers showed interest for a UNIX (LINUX) version of the program.

All the NSI's participating to the testing do release tabular data (mostly for business data but also for social data). It should be noted that some mentioned explicitly the release of frequency tables, which, however, are commonly released for Structural Business Statistics data. The most commonly applied protection rule is frequency threshold, but also the p-rule is used. Tables are mostly protected by suppression but also by recoding and rounding.

2) Preliminary issues (installation and data importing)

This section was designed to gather impressions on issues arising before working with the data. That is, installing the program and data importing. We feel that data and metadata importing is a crucial issue for users and, therefore, a few question insisted on this topic.

Some testers considered a problem the fact that the installation under Windows 2000 requires administrator's privileges; some other experienced problems with the installation because it did not permit the use of suppression routines. The installation of the optimization licences has created some difficulties. Once installed alright, the program could be launched without problems. The first impression on the graphical appearance is "Good".

Importing the data gave no problems, except when importing data in which "missing values" are specified as dots (SAS default), in which case a run-time error occurred. Most testers have the need for importing data in other formats, especially SAS and .csv but also MS Excel. Most testers would like to be able to import data in tabular form. There have not been problems with importing metadata. It was suggested to include a DDI compliant format for them. Many would like to have τ-Argus to produce summaries of the data and to check the correspondence to the metadata specifications. It was suggested the introduction of an import wizard and some lamented a run-time error if metadata importing is interrupted via the *cancel* button.

3) Data specification

Data specification is a crucial step for any statistical analysis; it should be easy and unambiguous in order to set all analysis on firm ground and make comparisons easy. This section deals with editing the metadata, defining the tables to be protected and specifying the parameters for the sensitivity rules to be applied. The whole process leads to computing the tables and identifying the sensitive cells.

The *specify metadata* window for editing the metadata was found clear and easy to use, however several problems were detected:

- The presence of missing values leads to run-time errors in many situations, for example when these are specified as dots; the program cannot handle missing values for hierarchical variables;
- The last variable in the list cannot be deleted;

- Adding new variables, saving changes, defining more than one weight variable.

The *specify tables* option is considered clear and easy to use. However, some functionality problems were spotted. Some additions to the safety rules comprised and the possibility of working with frequency tables was requested by many. The definitions of the parameters for the P-Q rule are not clear to all users. The minimum frequency rule cannot be disabled and it should be possible to apply it only to nonempty cells. The meaning of the safety ranges needs a better explanation.

The computation time of tables was rated from good to excellent and also defining more than one table. The information about unsafe cells with respect to the number of classifying variables considered was found useful by most testers. The documentation is clear but some topics need deeper explanation.

The *select table* window is clear and easy to use but some users would like more information on the tables.

The *view table* window was particularly well liked by testers. It was suggested that *Table summary* be shown directly on the table while information for each specific cell be given upon request. Also a navigating facility allowing visiting only sensitive cells was required.

The help for this section was rated good.

Protection Methods

Protection of unsafe cells is the primary purpose for which τ-Argus was developed. It comprises two methods: variable recoding and secondary suppression. Variables for which recode levels have been specified can be recoded manually choosing the levels, via a graphic three. Recoding is designed in such a way that changes can be visualized and undone easily. The computation of secondary suppressions is a hard optimization problem. It is implemented in three versions: using the Xpress or Cplex routines or the Hypercube method. As mentioned above, the first two routines are exact and equivalent in the results but require a commercial licence while the hypercube method is heuristic and uses freely available routines. Table protection is designed in such a way that users can view the results and easily undo what has been done.

The *recode* window was unanimously found extremely good and easy to use. Even the documentation was found exhaustive. The error caused by an erroneous specification of the codes should be better explained.

Suppression methods gave different types of problems:
Cplex and Xpress routines were not tested by many testers who did not have the licence. Some found difficulties installing these licences. Cplex was found faster than Xpress, both methods failing give run-time errors. The documentation for both libraries needs improvement.

Since the *Hypercube* method does not require licensing it should have been tested by all testers. Unfortunately, this method showed some problems and in some cases it resulted in a run-time error. The method requires the specification of the “ratio parameter” (sometimes referred to as “range ratio”) the meaning of which should be better explained. The value suggested by default for this parameter often leads to an error. The subsequent error window should help the user to choose the suppressions. However, some testers found this task a bit difficult given the large quantity of numbers involved. The implementation of this method surely needs improvement as does the documentation.

Overall this section was rated just sufficient, it is the part of the program that needs most improvement.

Output of Protected Tables
The output is the final product of the whole SDC process. It consists of creating protected tables to be released and a report on the protection of the data. Since the protected tables and report could be handed to people not familiar with τ-Argus and SDC, all items should be easy to understand and to explain to others.

The *save table* option was rated "Good", a part from the option *JJ-format*, which needs explanation. Excel, ASCII and Oracle were suggested as alternate formats for the output. A micro-file with the indices of suppressed records was also suggested as an additional output.

The *View Report* option was quite liked by all testers. One improvement suggested is to specify the rule by which cells are declared unsafe.

Overall this section was rated good.

Documentation
Documentation of a program can hardly ever satisfy all users. In the case of scientific programs, such as τ-Argus, a balance must be found between theoretical explanations and guidance to users. Furthermore the scores given may be influenced by the level of theoretical knowledge of the tester. In this section we sought an overall impression of the documentation, where specific critiques can be found in each section.

The on-line help was found clear and useful but in need of improvements. The index for the search facility should include more items and a help button on each window pointing to the right help page would be much appreciated.

The manual is clear and helpful, however it needs to be more comprehensive. Particularly, methodological explanations of the suppression methods and of the saving formats have been required.

Overall the documentation got mostly “Good” and “Excellent” marks, however some points need attention.

General Remarks
This section asks testers to express their overall opinion on the program. Questions regarding the organization of the menu options, intuitiveness of the ordering of the steps for carrying out the protection of a data file and overall functionality are included. Testers are also asked if they would advice their institution to adopt it “were it a commercial program”. This question was asked in order to have the tester change perspective.

Menus and options were found very clear and well organized. It was found that too often τ-Argus shuts down on run-time errors, obliging users to start the whole protection process all over again.

The sequence of the steps required to apply the protection methods was unanimously judged very good and does not need improvement. The information required for the application of the SDC methods is adequate as is the guidance provided (see comments on suppression methods though).

The guidance provided throughout the protection process seems adequate although some required a help button on each window.

The program did not have conflict with other running applications. There has not been registered significant slowing on the system while running the program, a part when computing secondary suppressions. There have been reported problems with the handling of errors in the metadata file and in the file for recoding. Some testers experienced slowness in computing more than one table or large tables.

One major problem is connected with fatal errors. There have been reported run-time errors under different circumstances. For this reason the overall functionality of the program has been rated less than good.

All testers concluded that the program could be adopted by their institution as a standard tool for SDC as long as the methods currently adopted are included in the program. All testers would then recommend buying it to their institution. The overall rating was "*Good*".

Conclusions

The testing of τ-Argus showed that all testers had a good impression of the software with respect to the way it was concealed and designed. However, it also revealed some important problems. On one hand some NSI's would need different sensitivity rules implemented, in order to be able to adopt the program as a standard tool at their institution. On the other hand, some problems have been reported in what is already implemented. Among the most critical problems seem to be: the handling of missing values and the implementation of the heuristic routines for the computation of secondary suppressions. Other problems connected with fatal errors of the program and relative to a better documentation need also to be addressed.