



CASC PROJECT

Computational Aspects of Statistical Confidentiality

June 2003

Methods for tabular data protection in τ -ARGUS

Author: Sarah GIESSING,

Institute: Federal Statistical Office of Germany

Deliverable No:5-D6-1



Methods for tabular data protection in τ -ARGUS

Deliverable 5-D6
Part 1

Sarah GIESSING,
Federal Statistical Office of Germany
65180 Wiesbaden
E-mail: sarah.giessing@statistik-bund.de

Data collected within government statistical systems is usually provided as to fulfil requirements of many users differing widely in the particular interest they take in the data. Data are published at several levels of detail in large tables, based on elaborate hierarchical classification schemes. In many cases, cells of these tables contain information on single, or very few respondents. In the case of establishment data, given the meta information provided along with the cell values (typically: industry, geography, size classes), those respondents could be easily identifiable. Therefore, measures for protection of those data have to be put in place. The choice is between suppressing part of the information (cell suppression), or perturbing the data.

The software τ -ARGUS offers to identify sensitive cells, a choice of algorithms to select secondary suppressions, a program to compute interval bounds for suppressed cells (audit), and a tool for controlled rounding. It is also foreseen to integrate software generating synthetic values instead of cell suppressions.

The first two chapters of the τ -ARGUS users manual [10] provide a brief introduction into the methods for cell suppression presented by the package. The following is supposed to give some guidance where to find more detailed methodological background information, and documents reporting on the performance of the algorithms offered by the package.

In τ -ARGUS there is a variety of options for a disseminator to formulate protection requirements. When cell suppression is used as disclosure limitation technique, in a first step *sensitive cells* will be suppressed. In a second step, other cells (so called '*secondary*' or '*complementary*' *suppressions*) must be suppressed along with these so called '*primary suppressions*' in order to prevent the possibility that users of the published table would be able to recalculate individual respondent data. For a broad introduction into the basic concepts of cell suppression methodology see [5].

The goal of secondary cell suppression is to find a valid suppression pattern satisfying the protection requirements, while minimizing the loss of information associated with the suppressed entries. The 'classical' formulation of the secondary cell suppression problem is a combinatorial optimisation problem, which is computationally extremely hard to solve. τ -ARGUS presents a variety of algorithms to find a valid suppression pattern even for sets of large hierarchical tables linked by linear interrelations. It is up to the user to trade-off quality vs. quantity, that is to decide how much resources (computation time, costs for extra software etc.) he wants to spend in order to improve the quality of the output tables with respect to information loss. The package offers a choice basically between four different approaches:

OPTIMAL Fischetti/Salazar methodology aims at the *optimal solution* of the cell suppression problem [8]. A feasible solution is offered at an early stage of processing, which is then optimised successively. It is up to the user to stop execution before the optimal solution has been found, and accept the solution reached so far. The user can also choose the objective of optimisation, i.e. choose between different measures of information loss. Note that the method relies on high performance, commercial OR solvers.

MODULAR The *HiTaS* method [7] subdivides hierarchical tables into sets of linked, unstructured tables. The cell suppression problem is solved for each subtable using Fischetti/Salazar

methodology [3]. Backtracking of subtables avoids consistency problems when cells belonging to more than one subtable are selected as secondary suppressions.

NETWORK The concept of an algorithm based on *network flow methodology* has been outlined in [1]. Castro's algorithm aims at a heuristic solution of the CSP for 2-dimensional tables. Network flow heuristics are known to be highly efficient. τ -ARGUS offers an implementation applicable to 2-dimensional tables with hierarchical substructure in one dimension. A license for a commercial OR solver will not be required to run the algorithm.

HYPERCUBE The *hypercube algorithm* GHMITER developed by R.D. Repsilber ([see 4,11]) is a fast alternative to the above three OR based methods. This heuristic is able to provide a feasible solution even for extremely large, complex tables without consuming much computer resources. The user, however, has to put up with a certain tendency for over-suppression.

In order to give a clue, at least, which of the alternative methods offered by the package might be likely to perform best in a given situation, a benchmark study has been carried out. Each algorithm has been applied to the data sets of a library of test instances. [6] compares the results of these tests with respect to key issues such as practical applicability, information loss, and disclosure risk. Section 4 of part 2 of this deliverable provides a summary of this comparison.

By solving a set of equations implied by the additive structure of a statistical table, and some additional constraints on cell values (such as non-negativity) it is possible to obtain upper and lower bounds for the suppressed entries of a table. τ -ARGUS offers to derive the bounds of these so called 'feasibility intervals'. Based on ideas of [2], a method for controlled tabular adjustment (CTA) has been implemented to supply users with synthetic values located within those intervals which could be used to replace suppressed original values in a publication. For discussion of these techniques see [7, sec. 2.4].

While most of the methods provided by τ -ARGUS are designed to be used for the protection of establishment data, the package also offers a tool for Controlled Rounding, a method particularly well suited for limitation of disclosure risk in tabulations of data on households and individuals. See [9].

REFERENCES

- [1] Castro, J. (2004), 'Network Flows Heuristics for Complementary Cell Suppression: An Empirical Evaluation and Extensions', In: '*Privacy in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 3050)
- [2] Dandekar, R.H., Cox, L. (2002), 'Synthetic Tabular Data – an Alternative to Complementary Cell Suppression, unpublished manuscript
- [3] De Wolf, P.P. (2002), 'HiTaS: A Heuristic Approach to Cell Suppression in Hierarchical Tables', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)
- [4] Giessing, S., Repsilber, D. (2002), 'Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)
- [5] Giessing, S. (2004 a), 'Methodological Background of Cell Suppression in τ -ARGUS: an Introduction for the Practitioner', unpublished manuscript
- [6] Giessing, S. (2004 b), 'Benchmark report: Performance of alternative algorithms for secondary cell-suppression implemented in τ -ARGUS 2.2', deliverable 3-D5 of the CASC project
- [7] Giessing, S. (2004 c), 'Survey on methods for tabular data protection in ARGUS', In: '*Privacy in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 3050)
- [8] Salazar Gonzalez, J.J (2002), 'Extending Cell Suppression to Protect Tabular Data Against Several Attackers', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)
- [9] Salazar Gonzalez, J.J, Lowthian, P., Young, C., Merola, G., Bond, S., Brown, D. (2004), 'Getting the Best Results in Controlled Rounding with the Least Effort', In: '*Privacy in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 3050)

- [10] Hundepool, A., van de Wetering, A., de Wolf, P.P., Giessing, S., Fischetti, M., Salazar, J.J., Caprara, A. (2002), *τ -ARGUS users 's manual, version 2.1*
- [11] Repsilber, D. (2002), 'Sicherung persönlicher Angaben in Tabellendaten' - in *Statistische Analysen und Studien Nordrhein-Westfalen*, Landesamt für Datenverarbeitung und Statistik NRW, Ausgabe 1/2002 (in German)