

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint ECE/Eurostat Work Session on
Statistical Data Confidentiality**
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 12 (Summary)
English only

Topic (ii): software and computing developments

A SURVEY ON SOFTWARE PACKAGES FOR AUTOMATED SECONDARY CELL SUPPRESSION

Submitted by the Federal Statistical Office of Germany¹

Contributed paper

Summary

1. The paper presents brief descriptions of various cell suppression software packages, focussing on availability, costs, platforms, and the underlying methodology. The systems performances will be compared with respect to information loss, and computing time requirement; other key qualities of the software will be discussed.

I. INTRODUCTION

2. When cell suppression is used as a Statistical Disclosure Control Technique, table cells are suppressed, if the data disseminator considers them revealing too much. To prevent these so-called "primary suppressions", or "sensitive" cells from exact disclosure or a too narrow estimation from the additive relationship between the cells of the table, additional cells must be suppressed. The "Secondary Cell Suppression Problem" is to apply these complementary suppressions to the set of sensitive cells, in such a way as to ensure that the complementary suppressions:

- create the required uncertainty about the true values of the sensitive cells while still
- preserving as much information in the table as possible.

3. While it is relatively straightforward to derive the mathematical formulation of the secondary cell suppression problem as (Integer)-Linear Programming ((I)LP) problem, solving those large LP-problems for real-life sized statistical tables is far from easy.

¹ Prepared by Sarah Giessing.

A. Systems Included in the Comparison

4. The survey is on five existing software packages for table protection by cell suppression, four of them already in regular use, and one prototype.

Table 1: Systems included in the comparison

Name of software	Software development	Platforms / Programming Languages / availability / costs
GHQUAR	Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen (Dietz Repsilber)	FORTTRAN code, versions for IBM and SIEMENS mainframes, non-commercial system
USBCSUP	US-Bureau of the Census (Bob Jewett)	FORTTRAN code for DEC computer, non-commercial system
CONFID	Statistics Canada (Gordon Sande / Dale Robertson)	FORTTRAN (RATFOR) code for IBM mainframe, SUN SPARC workstation, non-commercial system
ACSSuprs	Sande and Associates, Inc. (Gordon Sande)	Improved version of CONFID, commercial system
τ -ARGUS Version 2.0	CBS Netherlands (second prototype version)	C++ code, WINDOWS-software (32 BIT), non-commercial-system, available for free at Statistics Netherlands, additional commercial software required (ca. 1000 - 2000 EURO)

B. Methodology

5. There is broad variety in the methodology applied:

- a very sophisticated algorithm for exact solution of the ILP-problem in τ -ARGUS 2.0,
- a heuristic LP-relaxation approach in CONFID and ACSSuprs,
- iterative procedures in USBCSUP and GHQUAR, which subdivide complex structured tables into subtables, and consider as feasible solutions suppression patterns of a certain structure only, thereby reducing the enormous computational burden effectively.

Considering the underlying methodology, we would expect, that in certain situations some programs will perform better than others. Whether such differences matter in practice is another question.

C. The Experiment: Comparison of Software Performance Using Original Tables from a Major Economic Census

6. For the comparison of secondary cell suppression software we chose seven two- and three-dimensional tables, some of them data relating to turnover, and some presenting data regarding numbers of employees. The most complex set of row relations resulted from an elaborate breakdown of industries (630 rows within a hierarchical system of six levels, created by intermediate sums).

7. CONFID, USBCSUP, and GHQUAR have been applied to these tables². Runs with τ -ARGUS 2.0 have not yet been performed (but probably will be, in time to present the results in the final paper).

² As the suppression algorithm is the same for CONFID and ACSSuprs, ACSSuprs runs were not performed.

D. Results of the Experiment

8. Overall, CONFID clearly gave the best performance regarding the number of suppressions, whereas USBCSUP suppressed the smallest total value, a direct result of realisation of the Census Bureau's notion of minimum information loss, which is to prefer small valued complements, rather than to minimize the number of suppressions. GHQUAR gave reasonable results with respect to both these criteria.

9. Close inspection of the results, gave the impression that the particular formulation of the secondary cell suppression problem:

- the particular notion of information loss (caused by suppression of cells), and
- the particular formulation of "required uncertainty" (to be created about the true values of the primary suppressions),

which will be discussed in the final paper in more detail, has a strong impact on the systems performances.

10. The computation times observed show that the GHQUAR is exceedingly faster than CONFID, as well as USBCSUP, the latter is however still much faster than CONFID, especially on three-dimensional tables.

E. Applicability and other key qualities

11. ACSSuprs, CONFID, and GHQUAR will be applicable to most real-life tables, single tables, at least, and if computing time resources are unrestricted, whereas it isn't easy to apply USBCSUP to three-dimensional tables, when substructure is present in each of the three variables.

The second version of τ -ARGUS has been developed for application to up-to-four-dimensional tables, without substructure (without f.e. hierarchies in the classifications, as created by intermediate sums).

12. The final paper will discuss some more key qualities for Cell Suppression systems:

- Availability of Audit Routines,
- User Intervention Facilities,
- Facilities for Table-to-Table Protection of sets of tables, linked by cells, which they have in common,
- Software Utility, etc.

F. Conclusions

13. The final paper will characterize:

- GHQUAR as a very efficient, powerful system, well suited for application to large statistical tables,
- CONFID / ACSSuprs as most reliable systems with respect to Data Security, and overall, maximising the information content of the output data,
- USBCSUP as a useful, flexible system, relatively easy to adopt,
- -ARGUS 2.0 as a modern, promising system, however still requiring further development, and improvement, first of all to make it applicable to tables with hierarchical substructure and to linked tables.