

Distr.
GENERAL

CES/SEM.47/9
20 February 2002

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**
(17-19 April 2002, Geneva, Switzerland)

Topic II: Secure communications and data confidentiality

THE CASC PROJECT

Invited paper

Submitted by Statistics Netherlands ¹

Abstract: In this paper we will give an overview of the 5th framework CASC (Computational Aspects of Statistical Confidentiality) project. This project can be seen as a follow up to the 4th Framework SDC-project. However, the main emphasis is more on building practical tools. The further development of the ARGUS-software will play a central role in this project. Besides this software development, several research topics have been included in the CASC-project. These research topics, both for the disclosure control of microdata as well as tabular data, aim at obtaining practical results that might be implemented in future version of ARGUS and find its way to the end-users.

Keywords: Statistical Disclosure Control, μ -ARGUS, τ -ARGUS, microdata, tabular data.

I. INTRODUCTION

1. Statistical Disclosure Control is a topic in official statistics that is becoming increasingly important. The growing need for information puts increasing pressure on the National Statistical Institutes (NSIs) to publish more detailed statistical information. The NSIs are traditionally very well equipped to carry out large censuses and large scale surveys. These sources of information contain very detailed information about enterprises and individuals. The computing power of modern computer systems is no longer a barrier to compose very large and detailed tables, which was the traditional output of the NSIs.

2. The modern information system, online databases and information systems on the internet make it possible to publish thee large tables, when earlier the mere physical limit of the paper publications would restrict the amount of detail in these publications. For the users of statistical information, policy makers, researchers etc. this is a very positive development. The NSIs will meet these requests for information using the new possibilities.

¹ Prepared by Anco Hundepool (ahnl@krypton.vb.cbs.nl).

3. However there is another side of the coin. When the NSIs are collecting the information needed to compose these large statistical databases, they have, for obvious reasons, promised the respondents to guarantee the confidentiality of the information provided to the NSIs. When the information is collected via a voluntary survey or through a legal based compulsory survey/ census it is vital for the NSIs to safeguard the confidentiality. Not only to comply with the legal obligations, but even more important to maintain the confidence of the respondents. If the respondents have the feeling that their sensitive information is no longer safe in the hands of the NSIs they can close their doors.

4. The CASC project, on the one hand, can be seen as a follow up of the SDC project of the 4th Framework. It will build further on the achievements of that successful project. On the other hand, it will have new objectives. It will concentrate more on practical tools and the research needed to develop them. For this purpose, a new consortium has been brought together. It will take over the results and products emerging from the SDC project. One of the main tasks of this new consortium will be to develop further the ARGUS software, which has been put in the public domain by the SDC project consortium and is therefore available for this consortium. The main software developments in CASC are μ -ARGUS, the software package for the disclosure control of microdata while τ -ARGUS handles tabular data.

5. The CASC project will involve both research and software development. As far as research is concerned the project will concentrate on those areas that can be expected to result in practical solutions, which can then be built into (future version of) the software. The CASC project has been designed, therefore around this software twin ARGUS. This will make the outcome of the research readily available for application in daily practice of the statistical institutes.

II. CASC PARTNERS

6. At first glance, the CASC project team appears rather large. However, there is a clear structure in the project, defining which partners are working together for which tasks. Sometimes groups working closely together have been split into independent partners only for administrative reasons.

Institute	Short	Country
1. Statistics Netherlands	CBS	NL
2. Istituto Nazionale di Statistica	ISTAT	I
3. University of Plymouth	UoP	UK
4. Office for National Statistics	ONS	UK
5. University of Southampton	SOTON	UK
6. The Victoria University of Manchester	UNIMAN	UK
7. Statistisches Bundesamt	StBA	D
8. University La Laguna	ULL	ES
9. Institut d'Estadística de Catalunya	IDESCAT	ES
10. Institut National de Estadística	INE	ES
11. TU Ilmenau	TUIIm	D
12. Institut d'Investigació en Intel·ligència Artificial-CSIC	CIS	ES
13. Universitat Rovira i Virgili	URV	ES
14. Universitat Politècnica de Catalunya	UPC	ES

7. Although Statistics Netherlands is the main contractor, the management of this project is a joint responsibility of the steering committee. This steering committee consists of 5 partners, representing the 5 countries involved and also bearing a responsibility for a specific part of the CASC project:

CASC Steering Committee

Institute	Country	Responsibility
Statistics Netherlands	Netherlands	Overall manager Software development
Istituto Nazionale di Statistica	Italy	Testing
Office for National Statistics	UK	
Statistisches Bundesamt	Germany	Tabular data
Universitat Rovira i Virgili	Spain	Microdata

III. ARGUS SOFTWARE DEVELOPMENT

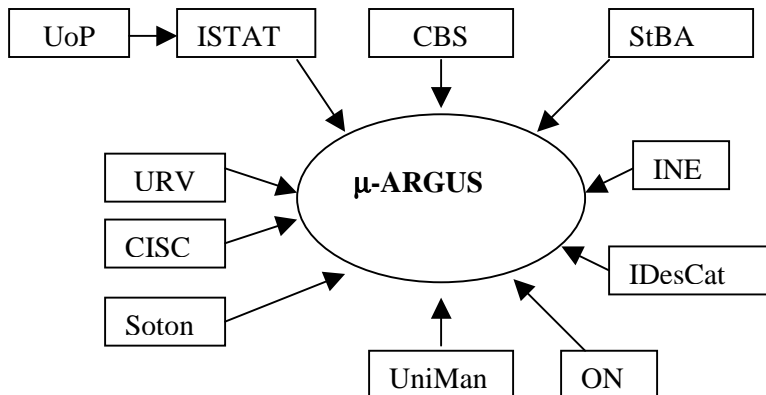
III.1 Software concepts

8. As the CASC project aims at practical solutions for disclosure control, we have given the development of the ARGUS software a central role in the project. The ARGUS software will be the binding factor between the different parts of the project. Research topics have only been included if they aim at results that either can be implemented in a (future) version of ARGUS or aim at testing the methodology used in the CASC project.

9. The starting point for the CASC project was the ARGUS-twins resulting from the SDC project. However, as these twins had been developed with Borland C++, we have decided to convert the software to a more modern, up-to-date version of C++, i.e. Visual C++. But for the user-interfaces we use Visual Basic as a programming tool. This is an easier platform for the development of user-interfaces, still meeting the needs of ARGUS. For the more crucial routines taking care of the heavy calculations we use Visual C++, which will lead to more efficient code. Some methods in ARGUS lead to complex computation problems, which justify the choice for C++.

10. The routines built into Visual C++ will be compiled into an OCX-component, which can easily be used in the Visual Basic user-interface programme. This guarantees a more flexible software concept and provides better options for the inclusion of additional routines for disclosure control and even third party solutions. A first example is the link between ARGUS and the German GHQUAR/GHMITER software. However, the aims with ARGUS in the CASC project are that ARGUS should be expanded into a control centre that will offer the user the choice of SDC solutions. This also makes the comparison between the different solutions within one framework much easier.

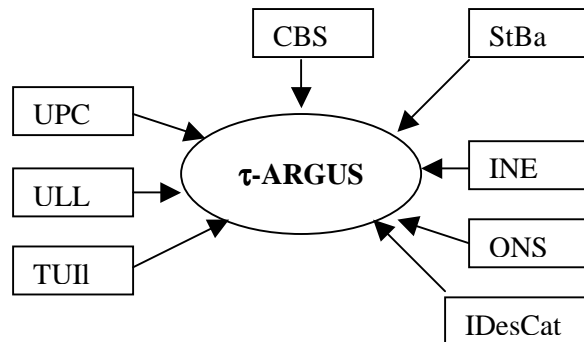
III.2 μ -ARGUS



11. Although the current version of ARGUS (3.0) has adequate routines for the disclosure protection of social micro data, we lack solutions for economic micro data. This is due to the different characteristics of these data. The distribution of the variables is much more skewed, making it much easier to re-identify the individual records. Also, the legal restrictions in many countries are much more restrictive for the access of economic micro data. The following enhancements to ARGUS software will provide techniques and methods not yet available to SDC practitioners, and will be a major step forward in this field.

- Micro-aggregation has been proven to be a promising method. Studies by Josep Domingo (2002) support this and the inclusion of multivariate micro-aggregation routines is planned.
- PRAM (Post RAndoMisation) is a technique under investigation at Statistics Netherlands. The basic idea is to add noise to the data. The distribution of the noise will also be made available to the potential data users. The combination of the distorted data and the known noise allows the user to make good estimates of population tables etc. This might be a drawback for the inexperienced users, but the advantage is that much more detailed information can be released. Good additional (tabulation) software could overcome this.
- Matrix Masking. Masking techniques have been studied already for some time, but we think that the time is ready for the inclusion in μ -ARGUS. Ruth Brand (1999) studied this method, building further on the work by Sullivan (1989).
- Disclosure risk models will be implemented. Based on the work of Skinner, the Italian partner, Luisa Franconi, has prepared a practical implementation for ARGUS. This gives a more sophisticated approach for the estimation of the disclosure risk of the individual records.

III.3 τ -ARGUS



12. The current version of τ -ARGUS (2.0) has been built to implement the optimization approach by J.J. Salazar and M. Fischetti. This version is suitable for the disclosure protection of simple 3-dimensional tables. Although this version works well, most practical tables have a more complex structure. In most tables at least one of the spanning variables has a hierarchical structure. This implies that there are many more sub-totals. This makes it much easier to recalculate individual cells. Therefore, the mathematical optimization problems are much more complex.

13. We will include novel solutions to the particularly large and complex optimization problems associated with SDC, which will allow minimal loss of information from treatment of multi-dimensional tables, and will use algorithms based on state of the art programming tools. Solving these problems is a challenging task, for which two groups of mathematical programming experts have joined the CASC team. The main approach will be by Fischetti/Salazar by extending the current library with facilities allowing for more complex data structures. The main problem here is that some computationally efficient shortcuts can no longer be used. Jordi Castro investigates the use of networks. Currently only solutions based on networks are known for two-dimensional tables, but extension to three dimensions are possible. This might lead to computationally quicker solutions.

14. However, it is to be expected that these complex mathematical solutions will work well for small- and medium-sized problems, but we also need solutions for very large tables. For this purpose, we have made a method, HITAS, which uses the current available solutions for unstructured tables. The large structured table is then broken down into many small, unstructured tables and protected. By applying this method top-down we end up with a protection of the whole table. Sometimes some backtracking is necessary, but that problem is taken care of. In this way rather large tables can be protected. As soon as the Salazar solution for a hierarchical table becomes available, we could include this solution finding an optimal balance between the number of smaller sub-tables to be protected and the complexity of these tables.

15. A second solution for very big and complex tables is provided by the German GHQUAR/GHMITER programme. This method based on hypercubes can solve tables with hierarchies up to 7 dimensions and also allows for linked tables. The drawback can be that the solution provided will be over-protective, because the problem is not solved to optimality. However, this will at least open a solution for the very big tables, which – for the time being – cannot be solved optimally.

16. In addition, we aim at extending τ -ARGUS into a control-centre for tabular disclosure control, making other techniques (new and existing) more easily available through τ -ARGUS. This will offer the users a range of solutions with all their own qualities and characteristics. The user can much more easily compare the different solutions, choosing a quick but not optimal solution or a more slow, but optimal solution. We can now also investigate the drawbacks of the non-optimal solutions.

17. Finally we will include an audit programme into τ -ARGUS. The idea behind this audit programme is that margins could be calculated for each solution, whatever method is used.

IV. METHODOLOGY RESEARCH FOR MICRODATA

IV.1 Introduction

18. It is foreseen in the CASC project that several new techniques for disclosure protection will be implemented. The need for these new techniques lies in the fact that the currently used methods like global recoding and local suppression serve very well the needs for social survey data but are inadequate for the disclosure protection of business microdata. New techniques investigated are micro-aggregation, noise addition, PRAM (Post-randomisation) and masking techniques. The research on PRAM is formally not part of the CASC project as this research is being carried out already as a PhD research at Statistics Netherlands. However, the results will be implemented in μ -ARGUS. Noise addition and masking techniques are studied and a special study into an alternative method for business data preserving the individual profile for each unit will be undertaken. Micro-aggregation will be studied as an alternative.

19. In addition to these new techniques for disclosure protection risk models will be investigated. These disclosure risk models help to assess the safety of a protected microdata file. A study on record level measures will result in a research report on noise addition. The latter will result in research that might be implemented in ARGUS during the CASC project, but will be implemented only after the foreseen scope of this project.

20. A simulation of the intruder will be investigated, when attempts will be made to undo the disclosure protection. Another important study is undertaken into the effects on the analytical power of the protected microdata file, i.e. how well are these protected microdata files suited for statistical analysis projects.

21. The different approaches for this topic are justified by the need for safe business microdata files, for which few solutions are available. The implementation of these methods in ARGUS will allow for an easy application of these methods, which will result in growing insight in the quality and the applicability of these methods. In the long run, we might reach a common opinion on recommendations for the generation of safe business microdata files. Eventually this offers the possibility of European harmonization.

IV.2 Methodology for business microdata

22. Research in this area is at an early stage, with, however, some applications successfully attempted. The project work will be focused on the practical need for users to have a secure methodological framework within which they can select suitable techniques to effectively treat small to medium size business microdata. Research topics include:

- Building of a new framework for business microdata that will maintain an individual profile for each unit.
- Development of matrix masking methods to allow their application to the complex data structures found in practice

- Further refinement of microaggregation techniques

IV.3 Measurement of risk and information loss

23. The project aims to incorporate realistic measures of risk, and when they become available measures of information loss, into the ARGUS software to make available to users. Users will then for the first time have the tools to make a properly informed choice between different methods of SDC treatment, which will balance risk of disclosure against cost in terms of information loss.

- Extension of record level measures to take account of the possible misclassification of key variables and of emerging ideas on record-linkage
- Setting a framework to work towards measures of information loss, with a first attempt to quantify the loss.
- A feature of elements of this proposal will be the incorporation of the measurement of risk and of loss of analytical validity into research on data perturbation techniques.

V. METHODOLOGY FOR TABULAR DATA

V.1 Introduction

24. τ -ARGUS for tabular data resulting from the SDC project covers the disclosure protection of simple unstructured tables up to dimension 3. A central role in the disclosure protection of tables is played by the dominance rule. This rule states which cells in a table are unsafe and therefore cannot be published. Alternatives for the dominance rule (the pq-rule) will be made available as well.

25. Due to the presence of marginals in a table, it is often easy to recalculate these suppressed cells. So additional cells must be suppressed to prevent this recalculation of the primary unsafe cells. It is not only enough to prevent exact recalculation but also to guarantee a safety range to protect the primary unsafe cells. The optimal selection of these secondary cells, to avoid unnecessary high losses in the information content of the protected tables, is a very complex numerical optimization problem.

26. Although in the τ -ARGUS version resulting from the SDC project a solution is available for unstructured tables, it cannot be applied to many tables in the daily life of a statistical office, because they have a hierarchical structure. These hierarchical structures imply many more (sub-) marginals, which can be used to recalculate these primary suppressed cells. Also the linked tables, having some marginals in common, must be treated simultaneously.

27. This makes the optimization problem of finding the optimum suppression pattern still much more difficult. Even for renowned researchers in the field of numerical optimization this is a very complex problem. Nevertheless, we aim to find a solution. The main approach is undertaken by J.J. Salazar, dealing with the research required to specify the new models before implementation and testing. A second supporting approach is based on network flow algorithms.

28. Besides these complex optimization approaches we will develop and implement heuristic methods, which aim at a much quicker solution. It is also to be expected that these methods will be able to solve much larger instances. The price for this will however be a non-optimal solution. It is known from previous investigations that τ -ARGUS is able to reduce the information loss for about 30 to 50%. For several tables this advantage of speed might prove to be adequate. Some of these methods are already available in a basic form (e.g. GHQUAR) but we will extend τ -ARGUS to facilitate the access to these heuristic methods.

Another approach is based on the non-hierarchical solutions already available, by breaking down the big hierarchical table into several sub-problems.

29. One of the outcomes of this project is the composition of a set of test-tables. These tables will play the role of test-bench for the optimization procedures and are of vital value for the researchers in numerical optimization techniques to find the best solutions.

V.2 Main objectives in tabular data research

30. The three main goals of innovation in the proposed project regarding tools for tabular data protection will be:

- Firstly to develop data-structures for τ -ARGUS that are able to represent the cell suppression problem for hierarchical structured and linked tables.
- Secondly, GHQUAR will be integrated into the restructured version of τ -ARGUS.
- Thirdly, to speed up the linear programming methodology in ARGUS as emerged from the 4th Framework project. This will make τ -ARGUS capable of solving the larger problems that result from the representation of real life tables with many sub-marginals in reasonable (computing) time.

31. It should be noted, assuming that the computational burden was not an issue at all, a straightforward change of the data-structure would do, to make the current linear programming approach applicable to hierarchical structured and linked tables too. However, in real life the computational burden is an issue indeed, making it quite a challenge to preserve the excellent performance of the linear programming approach with respect to information loss, while speeding it up sufficiently for moderate to larger sized applications. (It won't certainly be possible to bring it to the extremes, e.g. make it applicable to those X-large applications that can still be handled efficiently by GHQUAR).

VI. TESTING

32. In this chapter we make a distinction between the actual testing of the software and the testing of the methodology. It is very important to see how much protection is gained from the methods implemented. So one of our partners, Mark Eliot, is playing the role of simulating the intruder. This work can be seen as a test-case of the (newly) developed methodology.

33. The actual testing of the software gets serious attention too. Lessons drawn from its predecessor, the SDC project, have learnt that you cannot only rely on voluntary testers. Therefore, testing has been incorporated as a separate task in the project. Both the building of test-sets as well as the actual testing of the software tools is an essential part of the CASC project.

VII. CONCLUSION

34. The major objective of this project is that the results will be used in real life situations in official statistics. The composition of the project team has been designed in such a way that the primary users, i.e. the NSIs, are active members. Seven statistical offices (5 national and two regional) participate in the project, either actively in the various stages of the development or as testers of the results. This reflects the needs and the interest of the NSIs for these kinds of tools.

35. The side effects of this project will be that the research community on Statistical Disclosure Control in Europe will work together. This joint effort will bring the state-of-the-art to a higher level.

36. In order to disseminate the results of the CASC project the project team will maintain a WEB-site. (<http://neon.vb.cbs.nl/casc>). Research papers resulting from this project as well as other material of interest for this field will find a place there.

References

Josep Domingo-Ferrer (2001), Josep Mateo-Sanz and Vicenc Torra, "Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk", *Paper presented at the NTTS/TTK Meeting Crete, Greece, June 2001.*

Sullivan, G. R. (1989), The Use of Added Error to Avoid Disclosure in Microdata Releases, *unpublished Ph. D. Thesis, Iowa State University.*

Ruth Brand (2001), "Microdata protection through noise addition", Paper presented at the AMRADS workshop *Luxembourg 2001.*