**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**
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 23
English only

Topic II: Impact of new technological developments in software, communications and computing on SDC

# THE CASC-PROJECT

**Contributed paper**

Submitted by Statistics Netherlands[1]

**Summary:** *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 of 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, m-ARGUS, t-ARGUS, microdata, tabular data*.

## I. INTRODUCTION

1. Statistical Disclosure Control is a field in statistics that has attracted much attention in recent years. Decision-makers demand more and more detailed statistical information. Researchers at universities and similar institutes have the capacity to perform complex statistical analysis on their powerful PCs and they desire detailed microdata. Therefore the need for statistical offices to publish more and more detailed information is growing. The other side however is that statistical offices have a legal or moral obligation to protect the confidentiality of information provided to them by respondents. This confidentiality is vital also to guarantee the future co-operation of respondents.

2. This imposes a large obligation on the shoulders of statistical offices to minimise the risk of disclosure from the information that they make available from their censuses and surveys. The question then arises how the information available can be modified in such a way that the data released can be considered statistically useful and do not jeopardise the privacy of the entities concerned. The aim of Statistical Disclosure Control (SDC) is to diminish the risk that sensitive information about or from individual respondents can be disclosed from a data set. The data set can be either a microdata set or a table. A microdata set consists of a set of records containing information on individual respondents or economic entities. A table contains aggregate information of individual entities.

3. 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 further develop the ARGUS-software, which has been put in the public domain by the SDC-project consortium and is

[1] Prepared by Anco Hundepool, Methods and Informatics Department (ahnl@krypton.vb.cbs.nl).

therefore available for this consortium. μ-ARGUS is the version for microdata while τ-ARGUS handles tabular data.

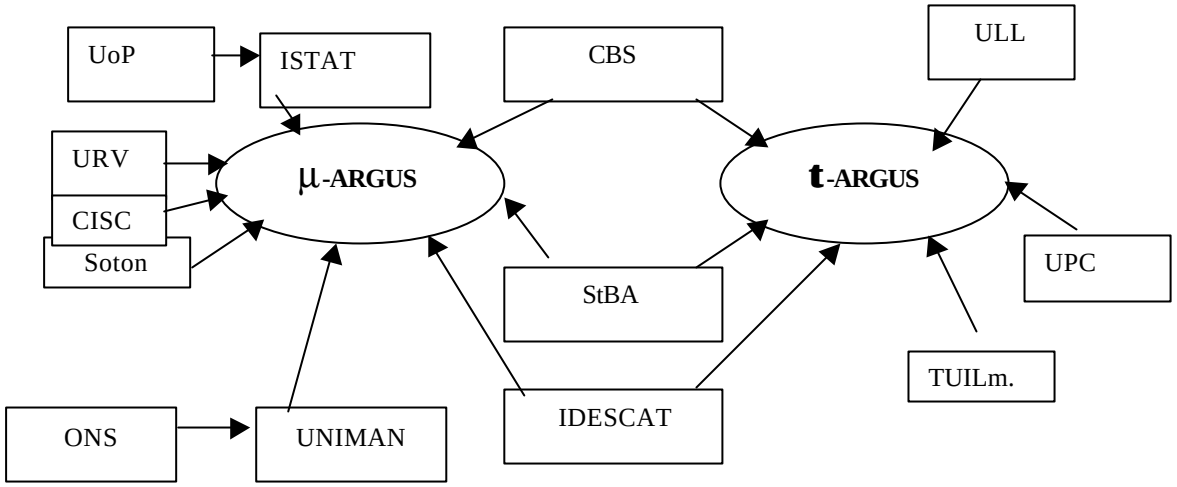
4. The CASC-project will involve both research and software development. As far as research is concerned the project will concentrate on those areas that can be expected to result in practical solutions, which can then be built into the software. Therefore the CASC-project has been designed round this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

## II. CASC-PARTNERS

5. At first sight the CASC-project team had become rather large. Sometimes groups working closely together have been split into independent partners only for administrative reasons. But for all the partners there are well defined tasks.

| Institute | Short | Country |
|---|---|---|
| 1. Statistics Netherlands | CBS | NL |
| 2. Istituto Nationale 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 Bundesambt | StBA | D |
| 8. University La Laguna | ULL | ES |
| 9. Institut d'Estadistica de Catalunya | IDESCAT | ES |
| 10. Institut National de Estadística | INE | ES |
| 11. TU Ilmenau | TUIlm | D |
| 12. Institut d'Investigació en Intelligència Artificial-CSIC | CIS | ES |
| 13. Universitat Rovira i Virgili | URV | ES |
| 14. Universitat Politècnica de Catalunya | UPC | ES |

**Graphical Overview of the project team**




Additional tester: INE

## III. ARGUS SOFTWARE DEVELOPMENT

6. 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 play 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 (future) version of ARGUS or aim at testing the methodology used in the CASC-project.

7. 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 its field.

- Extension of the existing τ-ARGUS software to allow for the first time secure treatment of hierarchical and linked tables. This will include novel solutions to the particularly large and complex optimisation 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.
  Besides this 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.
- Extension of the existing τ-ARGUS software to allow disclosure control of business microdata. Therefore several new techniques such as masking techniques, micro aggregation, noise addition and PRAM will be studied by the different partners in this project and will be implemented during the course of the project. Also disclosure risk models will be implemented.

## IV. METHODOLOGY RESEARCH FOR MICRODATA

### Introduction

8. 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. 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. These 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.

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

**Methodology for business microdata.**

10. 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.

**Measurement of risk and information loss**

11. A project aim is 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

**Introduction**

12. 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.

13. 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, as to avoid unnecessary high losses in the information content of the protected tables, is a very complex numerical optimisation problem.

14. 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.

15. This makes the optimisation problem to find the optimum suppression pattern still much harder. Even for renowned researchers in the field of numerical optimisation this is a very hard problem. Nevertheless we aim at a solution for this hard problem. 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.

16. Besides these complex optimisation 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.

17. 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 optimisation procedures and are of vital value for the researchers in numerical optimisation techniques to find the best solutions.

**Main objectives in tabular data research**

18. 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;
- The third main task will be to speed up the linear programming methodology in ARGUS as emerged from the 4$^{th}$ 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.

19. 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

20. Much attention will be paid within the project team to the testing of the results. Lessons drawn from its predecessor, the SDC project, have learnt that you cannot only rely on voluntary testers. Therefore testing has been incorporated in the project. Both the building of test-sets as well as the actual testing, not only the software tools, but also of the methodology, is an essential part of the CASC-project.

## VII. CONCLUSION

21. 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.

22. 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.

23. 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.