Summary of the CASC project

1. Introduction

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; and researchers have the capacity to perform complex statistical analysis on their powerful PCs and they desire detailed microdata. Therefore there is a growing pressure on the statistical offices to publish more and more detailed information. But National Statistical Institutes (NSIs) have to preserve the balance between their task as data providers and their legal and moral obligation to preserve the privacy of respondents, who have trusted their individual information to them. *Without respondents there is no statistical information*.

The 5th Framework program CASC project is a major effort to bring the standards of Statistical Disclosure Control (SDC) to a higher level. This project aims at a combination of research and the development of practical tools. Achieving practical results that can be applied easily in the practical production process was the central issue in this project. The ARGUS-software twins, both for microdata as well as tabular data, are the more visible results of the CASC project. But also many research papers have been published as a result of the CASC-project.

New information systems make the publishing of large tables a possibility, where previously the physical limit of the paper publications would restrict the amount of detail that could be published. In addition to tables, also the microdata themselves are becoming a more and more attractive form of output, as researchers nowadays are very well capable of performing complex analysis on these data themselves.

Even though this is a very positive development, there is another side of the coin. When NSIs are collecting the information needed to compose these large statistical databases, they must guarantee the confidentiality of the information provided to them by the respondents; NSIs must also safeguard respondents' confidentiality. In addition to this being a legal obligation, it is also vital for maintaining the confidence of respondents.

In this report we will give a summary of the results of the CASC project. The work in CASC falls into in two major topics: SDC of microdata and SDC of tabular data.

2. Starting point

The CASC project was not the first project in this field. In the 4th Framework project we have seen already the SDC project, a consortium of partners in The Netherlands,

United Kingdom and Italy. This project aimed more at the research side of Statistical Disclosure Control but already the first versions of the ARGUS software emerged from it. Positive side effects of the SDC project were the growing awareness of major project partners that European cooperation was a very good way of working. This led to long and intense discussions on the proposal of a new project for the 5th Framework. It was a common feeling between the partners that we should aim at practical results, leading to methods that could be easily used in the daily practice in the statistical institutes and similar institutes. So the software developments i.e. μ-ARGUS and τ-ARGUS became the key issue on this project. We accepted only research that could lead to practical extensions of ARGUS. Eventually the CASC-project was submitted and after some discussions with Eurostat and a successful negotiation phase this project was approved. January 1st 2001 became the official starting point for CASC, a three year project, which in the end was extended with six months, only to allow us to organise a final meeting/conference under nice circumstances in Barcelona.

3. Results

3.1 Introduction

The aim of this paper is to give an overview of all the results achieved during the CASC-project. We will only summarise the major results here, as all the deliverables of CASC are available via the CASC website (http://neon.vb.cbs.nl/casc/default.htm). Many reports, but also both μ-ARGUS and τ-ARGUS can be downloaded from this website. As the software is the outcome of this EU-funded project, it can be used free of charge.

3.2 Microdata

For *microdata* we have investigated methods like micro-aggregation, rank swapping, Sullivan's masking method, qualitative microaggregation and post-randomisation. Besides this, we have developed further the Franconi-Benedetti risk models to assess the disclosure risk per record and per household. This all has been built in μ-ARGUS in addition to the Dutch (simple) risk approach. At a certain stage in the project we have drawn the conclusion that Sullivan's masking method was theoretically interesting, but an efficient implementation that also could be used in daily practice was too complicated. The application requires too much technical knowledge by the data protectors. So we discontinued the research here and devoted the remaining part of the project to the start of the development of record-linkage software that eventually can be an extension of μ-ARGUS.

Josep Domingo-Ferrer of the University Rovira i Virgili coordinated the research on microdata, while the development of the μ-ARGUS software implementing many research results was the responsibility of Anco Hundepool, CBS.

3.2.1 Masking methodology

3.2.1.1 Model-Based Methods for Disclosure Limitation

Istat and the University of Plymouth have collaborated in developing methods for disclosure limitation of microdata, especially focusing on business surveys. A

“model-based" approach has been pursued, and several models for the release of safe microdata files have been investigated.

In Franconi and Stander (2002) we outline a new method for disclosure limitation of business microdata. The method builds regression models for the continuous variables to be protected. Some of the fitted values from these models are then shrunk before being released. In addition, a method based on principal components analysis for defining broader categories for releasing geographical area is proposed. In experiments on microdata from the Community Innovation Survey of manufacturing and service sector enterprises, the methodology developed offers more protection than microaggregation and very often leads to smaller error.

In Polettini, Franconi and Stander (2002) it is argued that any microdata protection strategy, such as those involving parametric probability models, matrix masking, coarsening, microaggregation, noise injection, and perturbation, can be based on a formal reference model. The method proposed by Franconi and Stander (2002) is examined in considerable detail, yielding valuable insights and suggestions for further research.

The model-based protection procedure of Franconi and Stander (2002) is extended in Franconi and Stander (2003) by allowing the model to take account of the spatial structure underlying the geographical information in the microdata. The Gibbs sampler is used to perform the computations required by this spatial approach. In experiments, it was found that, although the spatial method often induced higher inferential errors, it almost always provided more protection. Moreover, the aggregated areas from the spatial procedure can be somewhat more spatially smooth, and hence possibly more meaningful, than those from the non-spatial approach. The application of these model-based protection procedures to more spatially extensive data sets was also discussed.

In Burridge (2003) a disclosure limitation method that explicitly preserves certain information contained in the data is proposed. The data are assumed to consist of two sets of information on each respondent: public data and specific survey data. The public data are altered in some way to preserve confidentiality, whereas the specific survey data are disclosed without alteration. The method proposed is a model-based approach that utilizes the information contained in the sufficient statistics obtained from fitting a model to the public data by conditioning on the survey data. Deterministic and stochastic variants of the method are considered.

In Polettini (2003) a new disclosure limitation procedure based on simulation is proposed. The key feature of the method is to protect microdata by drawing artificial units from a probability model estimated from the observed data. Such a model is designed to maintain selected characteristics of the empirical distribution. The simulated data reproduce on average the sample characteristics. The model has a semi-parametric component, based on the maximum entropy principle, and a parametric component, based on regression. The use of the maximum entropy principle leads to a distribution which is consistent with the given information but is maximally non-committal with regard to missing information. Extensive experimentations demonstrated the success of the method.

A further summary of some of the work on model-based methods for statistical disclosure limitation performed under the CASC project was given in Burridge et al. (2002) and in Stander, Franconi and Burridge (2003).

Deliverable 5-D5 (part I) proposes a graphical method for joint assessment of information loss and disclosure risk for the data to be released, while Deliverable 5-D5 (part II) discusses practical issues and solutions for the release of a EU-wide microdata file of enterprises.

Finally, in Polettini and Stander (2003) we supply a comment on a theoretical basis for perturbation methods proposed by Muralidhar and Sarathy.

References

Burridge, J. (2003) Information preserving statistical obfuscation. *Statistics and Computing*, **13**, 312--327.

Burridge, J., Franconi, L., Polettini, S. and Stander, J. (2002) A Methodological Framework for Statistical Disclosure Limitation of Business Microdata. CASC Deliverable 1.1-D4.

Franconi, L. and Seri, G. (2003) Proposal for the creation of a micro-data file for research for Business Surveys – Part 2. CASC Deliverable 5-D5 (internal report).

Franconi, L. and Stander, J. (2002) A model-based method for disclosure limitation of business microdata. *The Statistician*, **51**, 51--61.

Franconi, L. and Stander, J. (2003) Spatial and non-spatial model-based protection procedures for the release of business microdata. *Statistics and Computing*, **13**, 295--305.

Polettini, S. (2003) Maximum entropy simulation for microdata protection. *Statistics and Computing*, **13**, 307--320.

Polettini, S., Franconi, L. and Stander, J. (2002) Model based disclosure protection. In: Domingo-Ferrer, J. (Ed.), *Inference Control in Statistical Databases: From Theory to Practice. Lecture Notes in Computer Science 2316*. Springer-Verlag, Berlin, pp. 83--96.

Polettini, S. and Stander, J. (2003) A comment on 'A theoretical basis for perturbation methods' by Krishnamurty Muralidhar and Rathindra Sarathy. *Statistics and Computing*, **13**, 337-338.

Seri, G. and Polettini, S. (2003) Proposal for the creation of a micro-data file for research for Business Surveys – Part 1. CASC Deliverable 5-D5

Stander, J., Franconi, L. and Burridge, J. (2003) Some model based approach to statistical disclosure limitation. 54th International Statistical Institute Session 2003. Berlin.

3.2.1.2 Microaggregation for numerical data

Multivariate microaggregation on unprojected data was identified as the best performing form of microaggregation in the first three months of the project (Domingo-Ferrer, Mateo-Sanz, Torra, 2001). To reach that conclusion, we had previously defined information loss and disclosure risk methods to compare continuous microdata masking. We then used the aforementioned measures to refine our comparison and identify rank swapping as a suitable complement to microaggregation. Thus, we decided it would make sense to have rank swapping included in μ-ARGUS, along with microaggregation. The next activity was to write portable and documented implementations of multivariate microaggregation and rank swapping on unprojected numerical data.

At this moment, the need arose to benchmark microaggregation and rank swapping against synthetic and hybrid microdata. Thus, we designed and implemented *hybrid microdata generation*, where hybrid means a (additive or multiplicative) mixture of original and synthetic microdata. We then compared the performance of this hybrid masking with multivariate microaggregation and rank swapping. It turned out that hybrid data and multivariate microaggregation performed similarly, according to our

measures and are outperformed by rank swapping (Dandekar, Domingo-Ferrer, Sebé, 2002). However, we observed that if we let the size of the hybrid data set be much larger than the size of the original data set, then hybrid data tend to score better and better.

The above and the rest of initial CASC work is documented in the book *"Inference Control in Statistical Databases" (Springer LNCS 2316)*, edited by Prof. J. Domingo-Ferrer, and in the book *"Privacy in Statistical Databases" (Springer LNCS 3050)*, edited by Prof. J. Domingo-Ferrer and Dr. V. Torra.

Agreeing on test microdata sets was perceived as essential by all CASC people working on microdata protection (microaggregation, additive noise, etc.). Thus, URV and Destatis joined their efforts to produce a joint deliverable "Reference Data Sets to Test and Compare SDC Methods for Protection of Numerical Microdata".

Toward the end of 2002, we reached the integration stage of our microaggregation and rank swapping software into μ-ARGUS. This required further software adaptation and debugging, especially as far as categorical microaggregation was concerned.

Work on microaggregation in the final year of CASC consisted of getting two Ph. D. theses ready, one of which was monographic on microaggregation. We also continued with the performance evaluation the security features of masking methods alternative to microaggregation:

- We pointed out some safety weaknesses of noise addition (Domingo-Ferrer, Sebé, Castellà, 2004);
- We proposed a method for synthetic data generation which is as fast as microaggregation, even though it very much damages subdomain analyses (Mateo-Sanz, Martínez-Ballesté and Domingo-Ferrer, 2004).
- Finally, we analysed the protection that several microdata masking methods provide for outliers. Microaggregation turns out to be the best option for protecting outliers and rank swapping follows closely behind (Mateo-Sanz, Sebé, Domingo-Ferrer, 2004).

Specific related publications produced under CASC are detailed below.

References

Dandekar, R. A., Domingo-Ferrer, J., Sebé, F. (2002), 'LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection', in 'Inference Control in Statistical Databases', ed. J. Domingo-Ferrer, *Lecture Notes in Computer Science 2316*, pp. 153-162.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Sebé, F. (2001), 'Watermarking for multilevel access to statistical databases', in IEEE ITCC'2001, Piscataway NJ: IEEE Computer Society, pp. 243-247.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Torra, V., 'Comparing SDC methods for microdata on the basis of information loss and disclosure risk', in Pre-Proceedings of ETK-NTTS'2001 (vol. 2), pp. 807-826.

Domingo-Ferrer, J., Torra, V. (2001), 'Disclosure protection methods and information loss for microdata', In: 'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies' Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland, pp. 91-110.

Domingo-Ferrer, J., Torra, V. (2001), 'A quantitative comparison of disclosure control methods for microdata', In: 'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies' Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland, pp. 111-134.

Domingo-Ferrer, J. (2002), 'Advances in inference control in statistical databases: an overview', in 'Inference Control in Statistical Databases', ed. J. Domingo-Ferrer, *Lecture Notes in Computer Science 2316*, pp. 1-8.

Domingo-Ferrer, J., Mateo-Sanz, J. M. (2002), 'Practical data-oriented microaggregation for statistical disclosure control', *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 1, pp. 189-201.

Domingo-Ferrer, J., Torra, V. (2002), 'Validating distance-based record linkage with probabilistic record linkage', in 'Topics in Artificial Intelligence', eds. Escrig, Toledo and Golobardes, *Lecture Notes in Computer Science 2504,* Springer, pp. 207-215.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Oganian, A., Torres, À., (2002), 'On the security of microaggregation with individual ranking: analytical attacks', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 477-492.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Oganian, A., Torres, À., (2002), 'A critique of the sensitivity rules usually employed for statistical table protection', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 545-556.

Domingo-Ferrer, J., Torra, V., (2002), 'Trends in aggregation and security assessment for inference control in statistical databases', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 5, pp. 453-458, 2002.

Domingo-Ferrer, J., Oganian, A., Torra, V. (2002), 'Information-theoretic disclosure risk measures in statistical disclosure control of tabular data', in Proc. Of the 14th Intl. Conference on Scientific and Statistical Database Management, ed. J. Kennedy, Los Alamitos CA: IEEE Computer Society, pp. 227-231.

Domingo-Ferrer, J., Sebé, F., Castellà, J. (2004), 'On the security of noise addition for privacy in statistical databases', in 'Privacy in Statistical Databases-PSD'2004', ed. J. Domingo-Ferrer and V. Torra, Lecture Notes in Computer Science 3050, Springer, pp. 149-161.

Domingo-Ferrer, J., Mateo-Sanz, J. M., Torres, À. (2003), 'Concepts for the evaluation of anonymized data', in *Anonymisierung wirtschaftsstatistischer Einzeldaten*, eds. G. Ronning und R. Gnoss, Wiesbaden: Statistisches Bundesamt, pp. 100-110.

Domingo-Ferrer, J., Torra, V. (2003) 'Disclosure risk assessment in statistical disclosure control of microdata via advanced record linkage', *Statistics and Computing*, vol. 13, no. 4, pp. 343-354.

Domingo-Ferrer, J., Torra, V. (2003) 'On the connections between statistical disclosure control for microdata and some artificial intelligence tools', *Information Sciences,*, vol. 151, pp. 153-170.

Domingo-Ferrer, J., Torra, V. (2004), 'Selecting potentially relevant records using re-identification methods', *New Generation Computing*, vol. 22, no. 3, pp. 239-252.

Domingo-Ferrer, J., Torra, V. (2004), 'Disclosure risk assessment in statistical data protection', *Journal of Computational and Applied Mathematics*, vol. 164-165C, pp. 285-293.

Mateo-Sanz, J.M., Martínez-Ballesté, A., Domingo-Ferrer, J. (2004), 'Fast generation of accurate synthetic microdata', in 'Privacy in Statistical Databases-PSD'2004', ed. J. Domingo-Ferrer and V. Torra*, Lecture Notes in Computer Science* 3050, Springer, pp. 298-306.

Mateo-Sanz, J.M., Sebé, F., Domingo-Ferrer, J. (2004), 'Outlier protection in continuous data masking', in 'Privacy in Statistical Databases-PSD'2004', ed. J. Domingo-Ferrer and V. Torra*, Lecture Notes in Computer Science* 3050, Springer, pp. 201-215.

Oganian, A., Domingo-Ferrer, J. (2001) 'On the complexity of optimal microaggregation for statistical disclosure control', *Statistical Journal of the United Nations Economic Comisión for Europe*, vol. 18, no. 4, pp. 345-354.

Oganian, A., Domingo-Ferrer, J. (2003) 'A posteriori disclosure risk measure for tabular data based on condicional entropy', SORT-Statistics and Operations Research Transactions, vol. 27, no. 2, pp. 175-190.

Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J. M., Torra, V. (2002), 'Post-masking optimization of the tradeoff between information loss and disclosure risk in masked microdata sets', in 'Inference Control in Statistical Databases', ed. J. Domingo-Ferrer, *Lecture Notes in Computer Science 2316*, pp. 163-171.

3.2.1.3 Microaggregation for categorical data

An objective of CASC work on microaggregation was to come up with categorical microaggregation methods. This was performed by drawing on fuzzy techniques (Domingo-Ferrer, Torra 2002a and 2002c). Whereas straightforward distance-based record linkage is suitable for analysing disclosure risk associated to numerical microaggregation, we had to design an extension of distance-based record linkage for categorical data (Torra, Domingo-Ferrer, 2003).

Work performed in the final part of CASC, including the one reported here, is documented in the book *"Privacy in Statistical Databases" (Springer LNCS 3050)* edited by J. Domingo-Ferrer and V. Torra. Specific references produced under CASC are given below.

References

Domingo-Ferrer, J., Torra, V. (2002a) 'Towards fuzzy c-means based microaggregation', in *Soft Methods in Probability, Statistics and Data Analysis*, eds. P. Grzegorzewski, O. Hryniewicz, M. A. Gil, Heidelberg: Physica-Verlag, 2002, pp. 289-294.

Domingo-Ferrer, J., Torra, V. (2002b) 'Aggregation techniques for statistical confidentiality', in Aggregation Operators: New Trends and Applications, eds. T. Calvo, G. Mayor and R. Mesiar, Berlin: Physica-Verlag, pp. 260-271.

Domingo-Ferrer, J., Torra, V., (2002c), 'Extending microaggregation procedures using defuzzification methods for categorical variables', in Proceedings of the First International Symposium on Intelligent Systems (IS'2002), IEEE, pp. 44-49.

Domingo-Ferrer, Torra, V., (2003), 'Fuzzy microaggregation for microdata protection', *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 7, no. 2, pp. 153-159.

Domingo-Ferrer, J., Torra, V. (2003), 'Median based aggregation operators for prototype construction in ordinal scales', *International Journal of Intelligent Systems*, vol. 18, no. 6, pp. 633-655.

Torra, V. (2004), 'Microaggregation for categorical variables: a median-based approach', in 'Privacy in Statistical Databases-PSD'2004', ed. J. Domingo-Ferrer and V. Torra*, Lecture Notes in Computer Science* 3050, Springer, pp. 162-174.

Torra, V., Domingo-Ferrer, J. (2003) 'Record linkage methods for multidatabase data mining', in *Information Fusion in Data Mining*, ed. V. Torra, Berlin: Springer, pp. 99-130.

Torra, V., Miyamoto, S. (2004), 'Evaluating fuzzy clustering algorithms for microdata protection', in 'Privacy in Statistical Databases-PSD'2004', ed. J. Domingo-Ferrer and V. Torra, *Lecture Notes in Computer Science* 3050, Springer, pp. 175-186.

Torra, V., Domingo-Ferrer, J., Mateo-Sanz, J. M., Ng, M. (2005) 'Regression for ordinal variables without underlying continuous variables, *Information Sciences* (to appear)

Valls, A., Torra, V., Domingo, J. (2003) 'Semantic-based aggregation for statistical disclosure control', *International Journal of Intelligent Systems*, vol. 18, no. 9, pp. 939-951.

Valls, A., Torra, V., Domingo-Ferrer, J. (2002) 'Aggregation methods to evaluate multiple protected version of the same confidential data set', in *Soft Methods in Probability, Statistics*

and Data Analysis, eds. P. Grzegorzewski, O. Hryniewicz, M. A. Gil, Heidelberg: Physica-Verlag, 2002, pp. 355-362.

3.2.1.4 Sullivan's masking

According to the original CASC project plan, it was foreseen to include at least three alternative methods that seemed to be suitable for production of safe micro-data files from business statistics. In the course of the project, it became evident that one of these alternative approaches, Sullivan's algorithm, would probably turn out to be of less practical relevance. A corresponding CASC research report (deliverable 1.1-D1, (Brand, 2002)) was rather discouraging, emphasising that "... an algorithm as complex as the one proposed by Sullivan can only be applied by experts. Every application is very time-consuming and requires expert knowledge on the data and the algorithm". The report insinuated that it would be hard for National Statistical Institutes to practically apply the method properly. Also, first results of a comparison of the performance of Sullivan's method to alternative methods for microdata protection indicated that the gains of Sullivan's method over microaggregation in terms of data quality could be expected to be at most moderate, while the effort for using Sullivan's method is much higher. This conclusion gave the impression that for practical applications it was likely that the method would be of minor relevance.

On the other hand, it became obvious that it would be a great benefit for μ-ARGUS to offer some kind of record linkage software for disclosure risk assessment which, however, was not foreseen in the project plan. Therefore, CASC partners decided to stop further work on Sullivan's algorithm. They agreed it would be better instead to use the part of the project budget originally meant to be spent on further research on applicability of Sullivan's method to pay work on implementation of a record linkage tool.

References

Brand, R. (2002), 'Tests of the Applicability of Sullivan's Algorithm to Synthetic Data and Real Business Data in Official Statistics, deliverable 1.1-D1 of the CASC-project, unpublished report.

Domingo-Ferrer, J., Torra, V. (2001), 'A Quantitative Comparison of Disclosure Control Methods for Microdata', In: 'Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies' Doyle, Lane, Theeuwes, Zayatz (Eds), North-Holland

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

3.2.1.5 Record linkage tool

The software package μ-ARGUS offers a variety of methods for producing safe microdata files (see Hundepool et al., 2003). Users of the package are offered a choice between methods, or they must select suitable parameters when applying a method to a data set. When choosing an anonymization method, two aspects have to be taken into account: information loss and data protection. An anonymization method may perform well with respect to information loss, but it may not protect the data sufficiently according to the user's requirements.

Lenz (2003) proposes three procedures for record linkage to be used in order to assess the disclosure risk of a protected micro-data file. One of the procedures is

based on methods of multi-objective optimization, the others on simple heuristics. Those three procedures have been implemented in SAS, and their performance on real-life data sets was compared. It turned out that the second heuristic procedure (see Lenz (2003a), section 4.1, Procedure 2) performs reasonably well compared to the optimization approach, while the computational effort for the heuristic is much lower (see Lenz, 2003 b). A C++ implementation of this procedure (deliverable 1.2-D6) has been delivered in December 2003 for integration in μ-ARGUS

References

Lenz, R. (2003 a): A graph theoretical approach to record linkage. Appears in: Proceedings of the joint UN-ECE/Eurostat work session on statistical data confidentiality, Luxembourg (2003)

Lenz, R. (2003 b): Disclosure of confidential information by means of multi objective optimisation. Comparative analysis of enterprise (micro) data conference, London (2003)

3.2.2 Risk assessment

The work on risk estimation was carried out by two teams. One team, led by Chris Skinner, University of Southampton, measured the disclosure risk both at the file level and the record level; the other team, led by Luisa Franconi, Istat, concentrated on measuring risk at both the individual level as well as the household level.

3.2.2.1 The British approach.

Research was undertaken to develop methodology for the assessment of disclosure risk for microdata. This research was concerned with the case of unperturbed microdata, that is where no disclosure limitation methods have been employed, and did not attempt to assess the disclosure protection provided by various methods of disclosure limitation. Disclosure risk was defined as identity disclosure, which in broad terms was conceived of as the probability that a microdata record may be identified by record linkage between the microdata file and external data sources on known individuals.

Measures of disclosure risk for microdata were developed at both the file level and the record level. One record-level measure was defined in terms of the probability of population uniqueness. An additional measure was defined by extending this measure to a definition in terms of the probability that an observed match is correct. Both measures depend on the specification of a log-linear model for an assumed set of key variables and the choice of model was also considered.

Empirical evaluations of different versions of the new record-level measure were undertaken using real survey data, including data from the UK General Household Survey. The measures were found to be effective. For example, evidence was found that the measures could discriminate between records of different levels of risk, in particular records, which are very likely to be population unique could be identified by consideration of records with high values of the measure.

The basic measures considered assumed that variables were recorded in the microdata in the same way as in the external sources. This is an unnecessarily conservative assumption, since in practice there will always be differences in the way the variables are recorded, as a result of measurement error, definitional differences or differences over time. Research was therefore also undertaken to extend both file-level and record-level measures to the case of misclassification. The extended definitions assumed that the misclassification process was known. Both

theoretical and empirical evidence was provided that the extended measures had useful properties.

3.2.2.2 The Italian approach

Work carried out under the CASC project includes a thorough revision and systematisation of the theory developed by Benedetti and Franconi (1998), with some notable achievements concerning approximation and numerical evaluation of the individual risk of disclosure. This work also benefits from the implementation, and subsequent testing, of the risk methodology in μ-ARGUS, as performed under work packages 1.2 and 6 (see CASC project Deliverables 1.2-D1, 1.2-D2, 6-D1 and 6-D4). The results of this research work are presented in Polettini (2003), Franconi and Polettini (2004) and finally in Polettini (2004). In these papers we motivate and define the Benedetti-Franconi (B-F) individual risk, show that it can be expressed in terms of the integral representation of the Gauss Hypergeometric function (see Abramowitz and Stegun, 1965), and finally discuss an approximation to this quantity based on the series representation and contiguity relations valid for the Gauss Hypergeometric function. As already mentioned, such an approximation can be exploited to produce more stable numerical evaluations of the B-F individual risk of disclosure.

The first paper also presents a useful analysis of different attack models and scenarios and a helpful discussion about the practical issues in the application of the individual risk methodology such as the setting of thresholds. The second and third papers add further results for practical application of the methodology and also discuss the so called household risk, which is a simple extension of the individual risk measure to files of dependent records, as is the case for files of households.

A parallel achievement of the project is the implementation of the individual and household risk measures into the final release of μ-ARGUS. A risk section has also been contributed to the μ-ARGUS user manual.

Additional contributions to risk estimation provided under the CASC project include Di Consiglio, Franconi and Seri (2003), and Stander (2003). In the first of these a thorough simulation experiment for assessing the performance of the Benedetti-Franconi risk methodology used in μ-ARGUS is presented. It was found that the Benedetti-Franconi risk methodology performed well, except possibly in the cases of rare combinations of key variables. The second contribution takes the form of a summary and discussion.

Polettini and Stander (2004) builds on previous work of Benedetti and Franconi to define a Bayesian hierarchical model for risk estimation. A super population approach, similar to the one adopted by other authors, is developed. For each combination of values of the key variables the posterior distribution of the population frequency given the observed sample frequency is derived. Knowledge of this posterior distribution enables suitable summary statistics to be obtained that can be used to estimate the risk of disclosure. One such summary is the mean of the reciprocal of the population frequency or B-F risk as used in μ-ARGUS, but others, such as the mode, are also investigated. Extensive experiments based on an artificial sample of the Italian 1991 Census data, drawn by means of a widely used sampling scheme, suggested that the approach works reasonable well and have led to potential alternative strategies.

References

Capobianchi, A., Polettini, S. and Lucarelli, M. (2001) Strategy for the implementation of individual risk methodology into μ-ARGUS: independent units. CASC Deliverable 1.2-D1.

Capobianchi, A., Polettini, S. and Lucarelli, M. (2002) Strategy for the implementation of individual risk methodology into μ-ARGUS: case of independent and hierarchical units. CASC Deliverable 1.2-D2.

Di Consiglio, L., Franconi, L. and Seri, G. (2003) Assessing individual risk of disclosure: an experiment. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7--9, 2003.

Franconi, L. and Polettini, S. (2004) Individual risk estimation in μ-ARGUS: a review. In: Domingo-Ferrer, J. and Torra, V. (Eds.), *Privacy in Statistical Databases. Lecture Notes in Computer Science 3050*}. Springer-Verlag, Berlin, pp. 262--272.

Merola, G. (2002). Testing of μ-ARGUS version 3.1. CASC Deliverable 6-D1 (internal report).

Merola, G. (2003). Testing of μ-ARGUS version 3.2. CASC Deliverable 6-D4 (internal report).

Polettini, S. (2003) Some remarks on the individual risk methodology. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7--9, 2003.

Polettini, S. (2004) Revision of Deliverable 1.2-D3 “Guidelines for the protection of social micro-data using individual risk methodology" by S. Polettini and G. Seri. Internal CASC report.

Polettini, S. and Stander, J. (2004) A Bayesian hierarchical model approach to risk estimation in statistical disclosure limitation. In: Domingo-Ferrer, J. and Torra, V. (Eds.), Privacy in Statistical Databases. Lecture Notes in Computer Science 3050. Springer-Verlag, Berlin, pp. 247--261.

Stander, J. (2003) Discussion of Topic (v): Risk Assessment. In *Proceedings of the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality*, Luxembourg, April 7--9, 2003.

3.2.3 μ-ARGUS

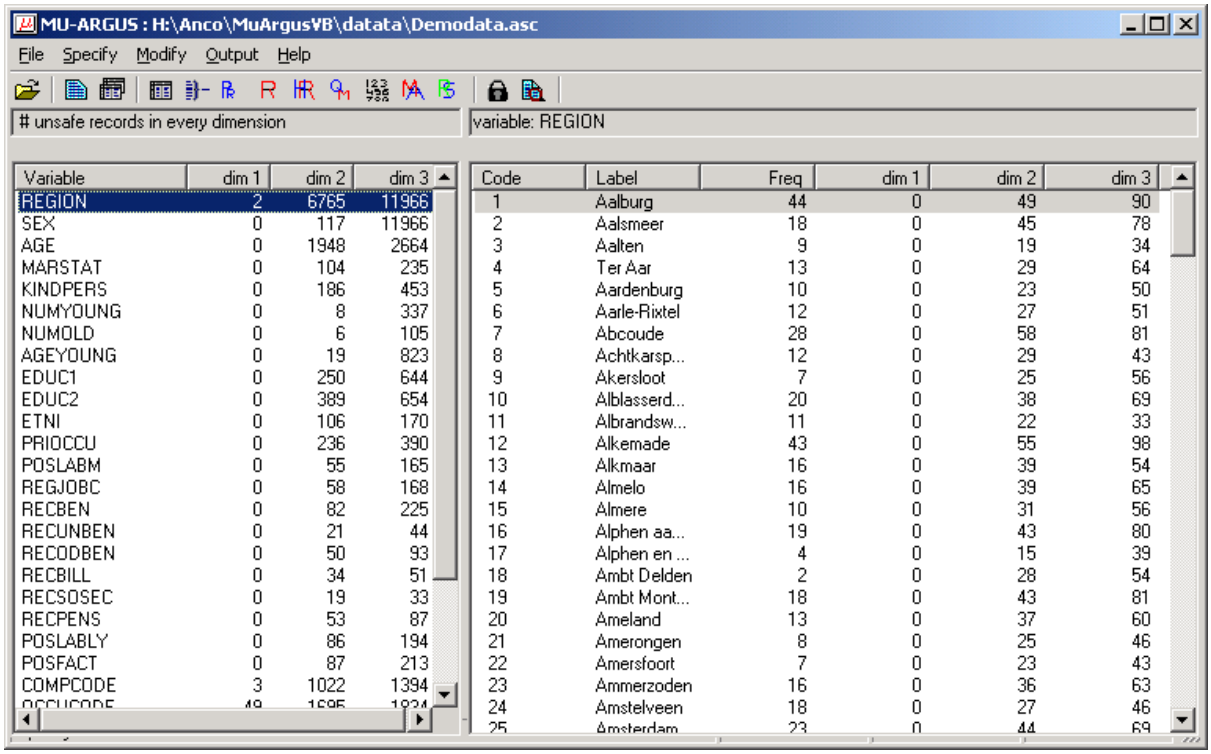
The μ-ARGUS software aims at the protection of microdata sets. The starting point for the development of μ-ARGUS was a view of safety/unsafety of microdata that is used at Statistics Netherlands. In fact the incentive to build a package like μ-ARGUS was to allow data protectors to apply the general rules for various types of microdata easily, and to relieve them from the chore and tedium that producing a safe file in practice can involve.

During the CASC-project we have rewritten the original prototype software using Visual Basic for the user-interface and the management of the process, while Visual

C++ is being used to program the more computationally intensive kernel. Several modules supplied by the other partners are linked to μ-ARGUS via a DLL.

The aim of statistical disclosure control is to limit the risk that sensitive information of individual respondents can be disclosed from data that are released to third party users. Identifying variables are those variables for which the scores are easily known to possible intruders, like municipality, sex etc. Based on frequency tables of identifying variables an approximation of the disclosure risk is calculated. Records are considered to be not safe, if a combination of its identifying variables does not occur frequently enough in the population.

Traditional methods to hamper the possible re-identification of records are global recoding and local suppression. Global recoding will reduce the amount of detail in the identifying variables, while local suppression will change individual codes into missing values. This main μ-ARGUS window gives the user an overview of all the unsafe combinations. This will give the user insight to which variables are most risky and need to be modified. Several methods of global recoding are available for hierarchical and non-hierarchical variables. Of course global recoding will reduce the number of unsafe combinations, but on the other hand it will reduce the amount of information in the resulting safe dataset. To find the balance between these two is the key-task of a data-protector. μ-ARGUS facilitates easy inspection of the results of various combinations of global recodings in order to choose the best one.



(μ-ARGUS central window)

To replace the very basic risk approach new risk estimators have been investigated in CASC (see section 3.2.2). Currently we have included the Italian approach (see section 3.2.2.2) in μ-ARGUS, which also can take into account the structure of the households. This risk estimation procedure will identify the records at risk. Also in this case global recodings can be applied to find an acceptable level of detail given your risk threshold.

Thanks to all the research work in the CASC project several methods are now available in μ-ARGUS:

1. Multivariate numerical microaggregation (grouping similar records together and replacing values of numerical variables by their mean) (see section 3.2.1.2),
2. Rank swapping (exchanging values between neighbouring records),
3. Categorical microaggregation (see section 3.2.1.3)
4. Post Randomisation. PRAM is a method of disclosure protection, which was not formally a part of the CASC project. However it is being investigated in The Netherlands. The general idea is that certain identifying variables are distorted with a given probability mechanism. So the datafile is protected by this transformation. Nevertheless, if we supply the parameters of the distortion mechanism, researchers can correct for this distortion on the level of the estimated aggregates. On the one hand side, this is more complex for researchers, but on the other hand we can supply them with a data file with more detailed coding schemes.
5. Top and bottom coding (replacing the tails of the distribution by e.g. the mean of the extreme values),
6. Rounding.
7. Noise addition. When a weight variable is present, this weight variable could reveal certain information on other variables, e.g. if these variables were used for the sampling frame. Adding noise to this weight variable could hide this relation.

All these latter methods have in common that they will distort the individual records and make the disclosure much harder. Nevertheless the resulting data files can be very well used by researchers for their analyses.

When you are satisfied with the specification of your data file, in the end a safe data file is written together with an extensive report of the modifications applied. This will document the process of Disclosure Protection.

References

Anco Hundepool et al (2004), μ-ARGUS 4.0 user manual, Statistics Netherlands, Voorburg.

3.3 Tabular data

3.3.1 Research on tabular data

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.

In τ-ARGUS there is now a variety of options for a disseminator to formulate protection requirements for respondent data. When cell suppression is used as disclosure limitation technique, in a first step *sensitive cells,* are identified and 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.

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 as emerging from the CASC project 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 many resources (computation time, costs for extra software etc.) they want 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 The Fischetti/Salazar methodology aims at the *optimal solution* of the cell suppression problem see Salazar (2000). 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 *Modular* method (Giessing, 2004) subdivides hierarchical tables into sets of linked, unstructured tables. The cell suppression problem is solved for each subtable using Fischetti/Salazar methodology (De Wolf 2002). 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 Castro, 2004. Castro's algorithm aims at a heuristic solution of the CSP for 2-dimensional tables. τ-ARGUS offers an implementation applicable to 2-dimensional tables with hierarchical substructure in one dimension. A license for a commercial OR solver is not be required to run the algorithm.

HYPERCUBE The *hypercube algorithm* GHM*ITER* developed by R.D. Repsilber (see Giessing, 2002 and Repsilber, 2002) 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 many 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. Giessing (2004 b) compares the results of these tests with respect to key issues such as practical applicability, information loss, and disclosure risk.

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 Dandekar (2002) 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 Giessing (2004C).

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 frequency tabulations of data on households and individuals. See Salazar et al (2004).

References

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)

Dandekar, R.H., Cox, L. (2002), 'Synthetic Tabular Data – an Alternative to Complementary Cell Suppression, unpublished manuscript

De Wolf, P.P. (2002), 'HiTaS: A Heustic Approach to Cell Suppression in Hierarchical Tables', In: '*Inference Control in Statistical Databases*' Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316)

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)

Giessing, S. (2004 a), 'Methodological Background of Cell Suppression in τ-ARGUS: an Introduction for the Practitioner', unpublished manuscript

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

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)

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)

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)

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*

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)

3.3.2 Optimisation models based on LP/MIP

During the CASC project the team at University of La Laguna (Tenerife, Spain) has conducted research on protecting sensitive cells in tabular data. On this topic, we

have analysed classical methodologies and proposed new variants. The most important contribution is that all methods implicitly guarantee the required protection for different sensitive cells and against different attackers, thus saving the effort of solving the Disclosure Auditing problem. We have defined a common framework including the main concepts of the Statistical Disclosure Control problem. We have then presented the well-known Cell Suppression Methodology, which replaces some unsafe cells by missing values, and computes other safe cells to be suppressed, but under the requirement of minimising the loss of information. A new methodology that replaces unsafe cells by intervals was also studied under the name of Partial Cell Suppression

In addition, we also developed Controlled Rounding Methodology, and even more we proposed a relaxed version called Partial Controlled Rounding Method. For each version a Mixed Integer Linear Programming model was described emphasizing the common definitions and features. By using the Duality Theory in Linear Programming (or Benders' Decomposition), it is possible to derive an improved model with a smaller number of variables and a bigger number of constraints. The advantage of the second type of models is that the number of variables is not dependent on the number of sensitive cells, and also there is no need for managing all the constraints explicitly since the relevant ones can be dynamically generated when required. These features are quite important because they suggest efficient algorithms using modern Mathematical Programming approaches.

We have also built an all-in-one methodology that is still a theoretical idea. However, in practice we have implemented some automatic tools for cell suppression and for controlled rounding which are now included in the latest version of τ-ARGUS, the main outcome of the CASC project.

Using the financial support from the CASC project, we have also produced the following reports:

J.J. Salazar González, "Extending Cell Supresion to Protect Tabular Data against Several Attackers", Inference Control in Statistical Databases (edited by J. Domingo-Ferrer) Lecture Notes in Computer Science 2316 (2002) 34-58

M. Fischetti, J.J. Salazar Gonzalez, "Partial cell suppression: A new methodology for statistical disclosure control", "Statistics and Computing" 13 (2003) 13-21

J. Riera Ledesma, J.J. Salazar González, "Algorithm for automatic data editing", “Statistical Journal of the United Nations ECE" 20 (2003) 255-264

J.J. Salazar Gonzalez, "Mathematical models for applying cell suppression methodology in statistical data protection", European Journal of Operational Research 154 (2004) 740-754

J.J. Salazar, P. Lowthian, C. Young, G. Merola, S. Bond, D. Brown, "Getting the Best Results in Controlled Rounding with the Least Effort", "Privacy in Statistical Databases" (edited by J. Domingo-Ferrer) Lecture Notes in Computer Science 3050 (2004) 58-72

J.J. Salazar-González, M. Schoch, "A new tool for applying Controlled Rounding to a Statistical Table in Microsoft Excel", “Privacy in Statistical Databases” (edited by J. Domingo-Ferrer) Lecture Notes in Computer Science 3050 (2004) 44-57

J. J. Salazar González, "Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data", "Mathematical Programming" (2005) to appear

J. J. Salazar González, "A Unified Mathematical Programming Framework for different Statistical Disclosure Limitation Methods", "Operations Research" (2005) to appear

J. Riera-Ledesma, J. J. Salazar González, "A branch-and-cut algorithm for the Error Location Problem in Data Cleaning", submitted for potential publication in "Management Science".

3.3.3 Optimisation models based on networks

The cell suppression problem (CSP) for statistical data protection is an NP-hard problem. That means that optimal procedures may be slow for some large tables, in practice. To overcome this drawback some heuristics based on network flows were developed in the past, mainly in US, and were used by the Bureau of the Census. They relied on the successive solution of minimum-cost network flows problems. But this procedure can still be quite inefficient for very large tables, as we observed in our preliminary results. Our main contribution is that we developed a new heuristic that solves shortest paths problems instead of minimum-cost network flows ones. The new heuristic sensibly combines and extends ideas of previous approaches. This resulted, in practice, in an improvement of two and three orders of magnitude for two-dimensional (2D) and two-dimensional with one hierarchical dimension (1H2D) tables with respect to previous network flows heuristics.

The heuristic we developed can be used for any table that can be modelled as a network. This is only possible for 2D and 1H2D tables. For 1H2D we also developed a fast procedure for obtaining the associated network. Both variants of the heuristic, for 2D and 1H2D tables, were implemented at UPC. With them we were able to solve 2D random instances of about 500000 cells and 3000 primaries in 40 seconds on a Pentium IV 1.8Ghz computer. 1H2D random tables of 300000 cells and 1000 primaries were protected in 12 seconds on the same computer. Executions performed on real data by CBS (Netherlands) and Destatis (Germany) showed that the heuristic is far more efficient than the optimal procedure and provides quite good results. Compared to other heuristics developed in the project, as HiTas (slower but better solutions) and Hypercube (faster but worse solutions), the network heuristic seems to provide a good trade-off between efficiency and quality of the solution. Moreover, the network heuristic always guarantees a feasible solution, while other heuristics do not. An additional benefit is that it does not require any external commercial solver, since all the required packages are included in the distribution developed for CASC. This is a key point for the usability and dissemination of this approach.

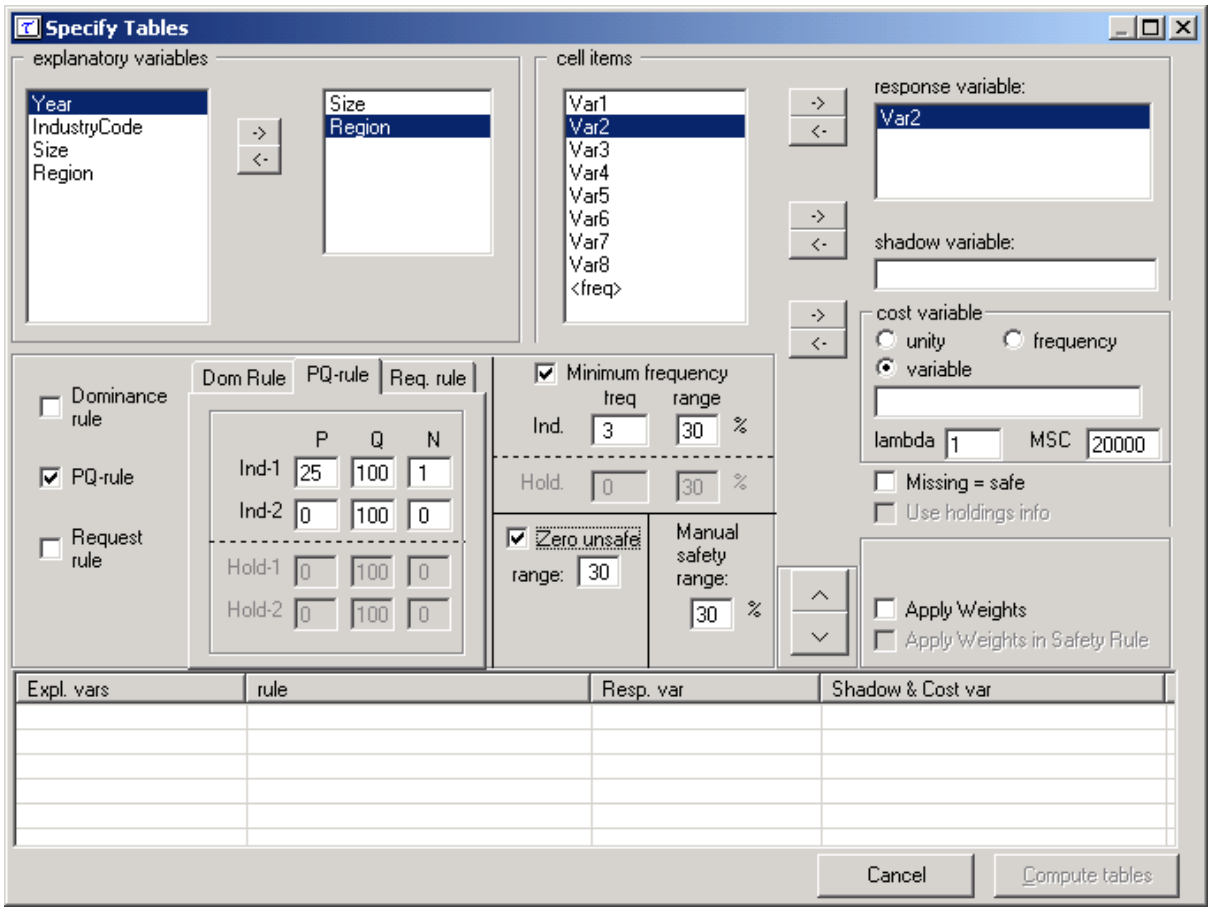
3.3.4 τ-ARGUS

Tables have traditionally been the major form of output of NSIs. There is also a longer tradition of studying the SDC-aspects of tabular data. Even in moderate sized tables there can be large disclosure risks. Protecting tables usually is done in two steps. The first step is to identify the sensitive cells. The next step is to protect these cells, because often the suppression of sensitive cells only can easily been undone, due to e.g. the marginal values in a table.

To find the sensitive cells in magnitude tables, traditionally the well-known dominance (n,k) rule is used. However there is a tendency to apply the prior-posterior rule. This rule has several advantages, see Loeve (2001). The basic idea is that it should be prevented that one contributor to a cell can make too narrow an estimate of another contributor's input to the same cell.

Mainly for the foreign trade statistics a special rule is used, the request rule. The tradition here is that only contributors actively asking for protection will be protected. Also a minimum frequency rule can be specified. For zero-cells (cells with only zero-value contributions) a special option has been provided to decide whether or not these cells are safe, as traditional rules like the (n,k) rule fail in this case.

For tables computed from sampled data, the sampling weights can be taken into account not only for computing the cell values, but also to influence the sensitivity rules used.

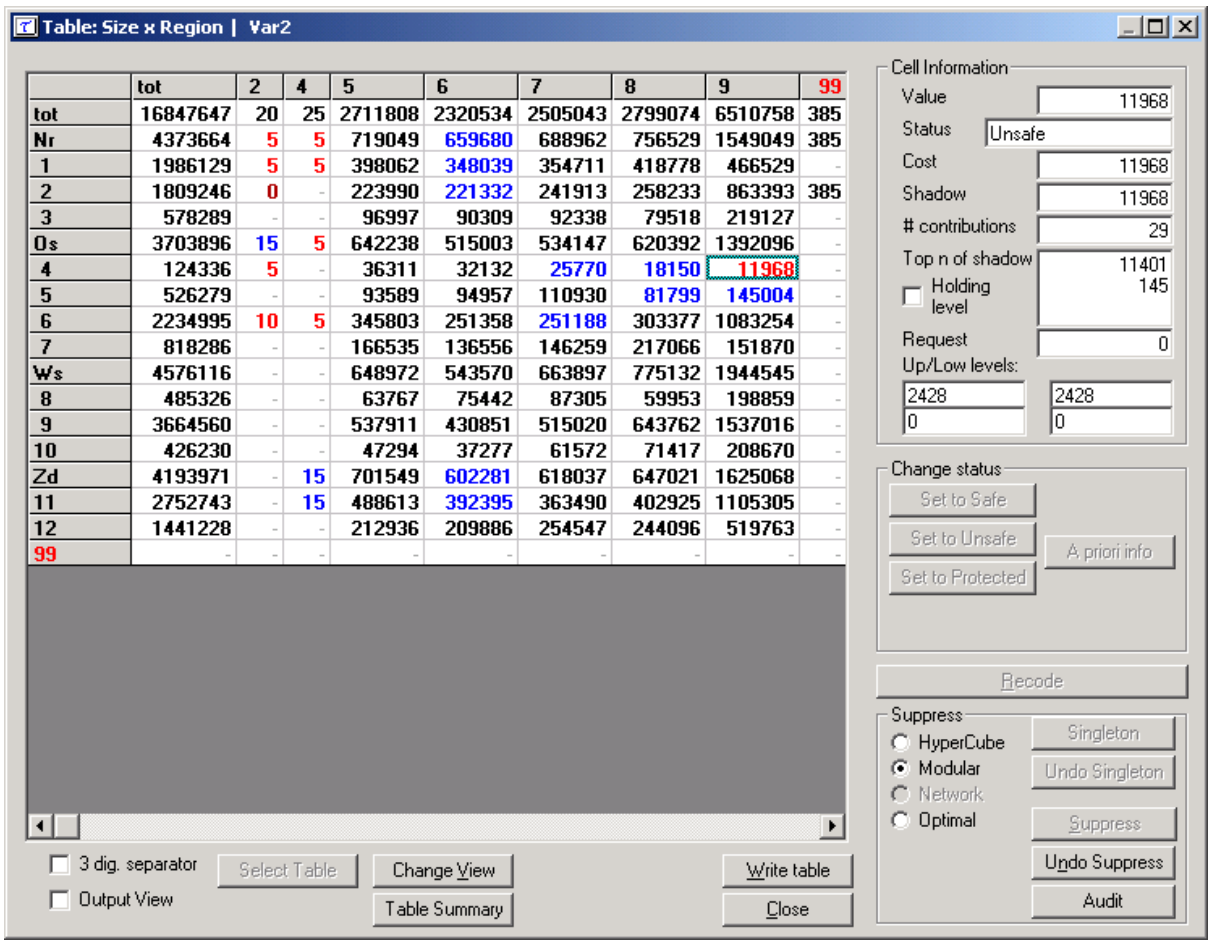


(τ-ARGUS window to specify tables with the different sensitivity rules)

Locating the sensitive cells is by far the easiest part of the job. To protect these cells additional cells (secondary) have to be found to make recalculation impossible. Not only the exact recalculation should be made impossible, but also the solution must

ensure large enough protection intervals around the original sensitive cells. This leads to very complex mathematical optimisation problems. These models have been discussed in the previous sections.

Before applying the secondary suppression routines, restructuring of a table can be performed as well. Collapsing categories of a variable tends to reduce the number of sensitive cells very well.



(τ-ARGUS window to show and protect a table)

When you are satisfied with the restructuring of your table you can protect it by selecting one of the suppression options. The table will be protected automatically. Depending on the size of the table and the number of (unsafe) cells, this can be a more or less time consuming process.

Traditionally τ-ARGUS could only read fixed format microdata files, but because of so many requests to be able to protect ready-made tables this has been included in τ-ARGUS as well. Also free format files can be used now.

After the protection of a table the protected table can be saved in various formats, together with an extensive report of the job done.

One of the latest developments is that τ-ARGUS can be used as a batch process as well. This enables you to use τ-ARGUS in an automated production process.

References

Anco Hundepool et al (2004), τ-ARGUS 3.0 user manual, Statistics Netherlands, Voorburg.

Anneke Loeve (2001), Notes on sensitivity measures and protection levels, Research paper 0129, Statistics Netherlands, Voorburg

3.4 Testing

Testing in the CASC project had two different goals:

1. Testing the quality of the methodology. How good are our methods both with respect to confidentiality preservation as well as data utility. With respect to tabular data protection, objectives and results of testing the methodology have already been reported on in section 3.3.1 above. With respect to micro data, to answer the question, whether data are sufficiently protected, the University of Manchester played the role of the intruder. They also addressed the second important question, i.e., can researchers still do sensible research on protected micro data files?
2. Testing the software. This is always an important but also often neglected part of software development. In order to take this seriously we have included partners in the project with the sole task of testing the software (mostly ARGUS) on their datafiles. This task was coordinated by Istat and has led to a considerable increase in the quality of the software.

3.4.1 Testing of methodology

The work for this topic was divided in three parts:

- Which data files are generally available in Europe for re-identification
- Do the methods implemented in μ-ARGUS really protect against disclosure
- A case study with UK SAR's

3.4.1.1 An evaluation of the availability of public data sources which could be used for identification purposes.

The research focused on a review of the availability of individual data in five European countries: Hungary, Italy, the Netherlands, Spain and the UK. The countries were selected in order to give some coverage across Europe and in relation to anticipated differences in data sources and data protection. There has been a substantial growth in the collection, storage and release of personal data across the public and private sector in Europe. In addition, far more personal information is collected and kept on restricted access databases across the private public and voluntary sector. Across Europe, public records and information available from the Internet contain a wide range of personal information. New databases and types of data are being constructed and made available each day. Data release policy should be made on the basis that a determined intruder will be increasingly likely to be able to obtain sufficient identification information to attempt to identify individuals within anonymised datasets. Data release policy, as well as considering which variables might be used to match against a given data set, should take into account of the sensitivity of the remaining variables. The sensitivity analysis should take account of public opinion about that information and the objective availability of such information elsewhere in the public domain. The decision over whether to include a given

variable or not should be based critically on the sensitivity of the information disclosed by that variable.

3.4.1.2 An assessment of the extent to which identification is impeded by the application of disclosure control methods in ARGUS.

An assessment of risk from identification attempts from matching information available in the public domain with data released by National Statistical Institutes. Using the UK General Household Survey and a sample survey data from a local authority, we attempted to mimic an intruder seeking to establish identification. Starting from two or three defined scenarios to define key variables we then assessed the availability of matching data. We compiled an identification database from a range of publicly available and restricted access databases and conducted a series of matching analyses. Data sources for matching attempts included occupational registers, electoral registers, GP lists, housing information. The results showed a low level of correct matching using simple key variables, but these rates of matching were improved by focusing on *Special Uniques*. It was found that protecting the data using the facilities of ARGUS reduced the level of matching significantly. However, some of those matches against perturbed records could be recovered by fuzzy matching techniques, although this introduced further false matches.

3.4.1.3 A Case Study of the Impact of Statistical Disclosure Control on Data Quality.

The 1991 UK Samples of Anonymised Records (SARs), which are publicly available sets of microdata from the UK Census, were used as a trial dataset in order to assess the impact of statistical disclosure control measures on data quality. A typical set of analyses was constructed through a literature review of published analyses using the SARs and through a user survey. These were selected on the basis of providing a good range of variables used and type of analyses conducted. The research allowed an empirical investigation of the feasibility of assessing the impact of disclosure control techniques on analytical power. An initial categorisation of the effect on those analyses on the application of SDC methods has been developed. The work is being taken forward to consider the plausibility of generalised metrics of analytical power, which will then be assessable for their relationship with disclosure risk impact. Further research is necessary to look at the relationship in detail.

References

Elliot, M. and Purdam, K. (2003) Analysis of Information Loss as a Case Study from a UK Survey,"

Federal Committee on Statistical Methodology Conference (FCSM) Research Conference, Washington DC, USA

Elliot, M. and Purdam, K. (2003) Data Quality and Disclosure Control, ISI Conference, Berlin 2003

Purdam, K. and Elliot, M. (invited contribution under review) Data Quality and the 2001 Census, Environment and Planning 2005

Purdam, K. (2005, forthcoming) The Nation's Data? The UK Census - Guaranteed Confidentiality But Only Limited Information, Information Communication and Society

Purdam, K, Mackey, E. and Elliot, M. (2005, forthcoming) Privacy, Identity and Data Use, Journal of Policy Studies.

Purdam, K. (2003) Privacy and Confidentiality and the 2001 Census Collection of Historical and Contemporary Census Data.

Purdam, K, Mackey, E. and Elliot, M. (2003) “Personal Data and Privacy” in Isaias, P. (ed) International Association for the Information Society, Conference Proceedings, Lisbon, Portugal.

Purdam, K, Mackey, E. and Elliot, M. (2003) “Whose Data Is It? Privacy, Data Value and the 2001 UK Census”, British Sociology Association, Annual Conference 2003

3.4.2 Testing of software

Testing is always an important job when developing software. Experiences from previous projects have taught us to take this matter seriously. You cannot only rely on good friends, who will do some testing for you on a rainy afternoon. So a special work package software testing had been added to the project and several partners were only participating in the project for the testing. This approach gave us much more control over the test efforts and this has proven to be a serious advantage.

The project CASC produced two pieces of software: τ-ARGUS and μ-ARGUS. τ-ARGUS carries out statistical disclosure control of tables and μ-ARGUS that of microdata. During the project the packages underwent one major user testing and several functional more limited ones. The user testing has been carried out by several of the NSI’s involved in CASC (ISTAT, IDESCAT-Statistics Catalonia, etc.) and a few others who volunteered (e.g. Israeli CBS and Stats New Zealand), coordinated by ISTAT.

The functional tests revealed a series of defects and malfunctions that, where possible, were mended or included in a list of “known defects.”

The user testing was designed to determine whether the packages were usable in a production stage by the European NSI’s. The tests were carried out on sets of real data chosen by the single entities and on a common set of data. Detailed results of these tests constituted a deliverable of the project. Succinctly, the tests revealed that both packages were well designed and their features corresponded to most of the needs of the NSI’s, but that especially Tau, had too limited capacity for having been adopted as only SDC tool. Mu was found a bit more difficult to understand than Tau and, since the practices for microdata protection greatly vary among different NSI’s, some testers pointed out that the package only partly performs what they need.

3.5 Future

All results of CASC can be found on the CASC website:
http://neon.vb.cbs.nl/casc/default.htm

A significant step forwards has been made in the research and development of Statistical Disclosure Control, giving Europe a leading role in this field. But the world is changing continuously, leading to new challenges in the future. If possible the CASC team will try to keep together and search for new opportunities to bring solutions for the new challenges in this field.