The CASC Project: Integrating Best Practice Methods for Statistical Confidentiality

Sarah GIESSING[1], and Anco HUNDEPOOL[2]
[1]*Federal Statistical Office of Germany; 65180 Wiesbaden*
e-mail: sarah.giessing@statistik-bund.de
[2]*Statistics Netherlands, PO-Box 4000, 2270 JM Voorburg*
e-mail: ahnl@krypton.vb.cbs.nl

Abstract: The paper will introduce into the topic, describing the position of Statistical Disclosure Control (SDC) in the statistical production process and explaining the need for a standard tool for SDC in the European Statistical System and the benefits that can be expected. It will focus on the description of current best practice methods and tools in the field, explaining how they have been studied and were identified as such. The paper will outline in which way technology transfer for statistical confidentiality can be promoted by integration of these methods and tools into one software package, focussing on what is needed to create standard software for statistical confidentiality. It is a special objective of this paper, to explain, how the use of such a software within the European Statistical System will be promoted, as to hopefully achieve a successful, widespread transfer of technology.

1. Introduction

Statistical Disclosure Control/Limitation 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 micro-data. 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.

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 (set of) table(s).

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.

The project CASC (Computational Aspects of Statistical Confidentiality) receives funding under the European Plan for Research in Official Statistics (EPROS). It has started January 2001 and should be finished 36 months after. The project is meant to be a follow up for the SDC project which has been carried out 1996 to 1999 in the sense that it will build further on the achievements of that project, and take over the results and products emerging from the SDC-project.

It is the main objective of the CASC project, to integrate best praxis tools and methods into the ARGUS software, e.g. the software μ-ARGUS for creation of safe micro-data files and the software τ-ARGUS for tabular data protection. In addition to that, the project will also involve research parts. The project will concentrate on those areas that can be expected to result in practical solutions which can then be built into the software. This will make the outcome of the research readily available for application in the daily practice of the statistical institutes. In this sense, the project aims at enhancing the transfer of technology and know-how in statistical confidentiality.

The authors are members of the CASC project consortium. Anco Hundepool is the project manager; Sarah Giessing is member of the steering committee.

Section 2 of this paper will describe the situation for microdata protection. Section 3 is concerned with tabular data protection. Both these section consider the current situation and user's needs. They focus on the identification of best practice in the field. Besides listing the respective main goals of the CASC project, they illustrate the main innovation plans. Section 4 briefly describes research parts of the project, while section 5 addresses particular issues of technology transfer. Section 6 briefly introduces the other papers of the confidentiality session of the ETK NTTS conference. The paper closes with a summary in section 7.

2. The situation for Micro-data protection

Concerning the protection of micro-data files, one has to take into account that the situation differs widely between data from social surveys and business data. In the following section, we will describe the current situation and discuss it in relation to the needs of users. We will explain in which way the goals of the CASC project correspond to these user needs. The reader will be referred to best praxis studies in this field, the main results of which will be introduced in section 2.3 . Section 2.4 will then illustrate that the decisions taken by the CASC team regarding the new disclosure limitation methods that will be further developed and will be integrated into μ-ARGUS are judicious and in accordance with the results of these studies.

2.1 Current situation and user's needs

Micro-data from social surveys

Regarding personal data, many NSIs and Eurostat release safe micro-data files. Based on results of research on the re-identification risk for particular micro-data sets, statistical institutes have developed some disclosure control rules for micro-data sets. These rules determine how much the information content of a micro-data file has to be reduced in order to be considered safe. Typical modifications are

- limitation of the geographical information,
- reduction of the classification detail or limitation of the upper codes for certain variables.

These kinds of modifications belong to the class of so called non-perturbative disclosure control measures. It is relatively easy to implement these kind of rules in the usual statistical data production process. However, to find a satisfactory set of additional measures (local suppression) is a more complex task. The software μ-ARGUS can be used to apply this kind of rules to a data set and makes it easy to compare the different choices of parameters.

The weak point in the field of micro-data protection of social survey data seems to be the definition of the disclosure rules. In the discussion at the 2001 Joint ECE/Eurostat work session on statistical confidentiality the need for routine checks of the re-identification risk associated to the release of particular data sets was expressed. Along this line, [9] identifies the need to implement individual risk assessment methodology into software, that should be shared among national statistical institutes.

Micro-data from business statistics

In [1] it has been established that non-perturbative disclosure control methods as they are currently used for micro-data protection of social survey data are not sufficient to create safe micro-data files for business data. Applying this kind of measures only, would yield micro-data files either with an unacceptably high disclosure risk associated to its release or would reduce the information content of the file to such an extent, as to make its analysis impossible.

In order to increase the use of their business data in spite of this, statistical institutes try to find other options, alternative to the release of safe micro-data files. In some NSI's so called Data Analysis Centres have been established, where external researchers can be permitted access to confidential data. Other options are to allow for remote access to a data-set. With this option, statistical institute staff runs programs developed by an external researcher for the analysis of the data set. After a thorough check of the confidentiality the results are returned to the researcher. However, it has turned out that both methods are relatively costly. Moreover, in some member states, due to the legal situation, the establishment of a data analysis centre can be difficult.

So, it is expected that at least as an addition to these options there is also a strong demand for micro-data files that have been protected using perturbative methods – at least if analytical validity can be proven for these file. Especially for perturbative methods it is essential, of course, to verify this.

2.2 Main Goals of the CASC project regarding the development and integration of new methodology for micro-data protection in μ-ARGUS

The aims of the CASC project concerning micro-data protection are consistent to the above described situation and user's needs in the following way:

- Methodology based on perturbative techniques for the protection of micro-data from business surveys will be explored, refined and integrated into μ-ARGUS. It has been established in advance, that the proposed methodology can be considered as 'good' or even 'best' practice, see sections 2.3 and 2.4 below.
- Methodology to assess the disclosure risk for micro-data at the individual record level will be researched. It will be feasible to integrate the results of this research into μ–ARGUS.

2.3 Identification of best practice methods for the protection of micro-data from business statistics

As pointed out in sec. 2.1, application of micro-data protection methodology so far has focused on non-perturbative methods which – if applied as the only technique – are not sufficient to create safe business micro data files. As a result, there are no mature 'good practice' methods available based on perturbative techniques alone or on a method mix. This however, does not mean that no methods are available at all or that none of them performs well. A variety of methods have been proposed and studied, mostly during the last decade. Some techniques were put into practice for some special application. But, due to the fact that it is a relatively new research topic, none of these methods so far has reached such a degree of maturity that it could have been established as standard technique which could then be labeled 'best practice'. However, some of the proposed approaches are very promising; and it can be expected that clustering the best of these methods into one software package will yield a usable software with a good prospect to be established as standard tool.

It is essential, of course, for the success of this project that as preparation step, research must be done to single out promising methods and also the non-promising ones. For the main part, this work has been carried out already, with results reported in the following papers:

1. presents methodology and results of an empirical, comparative study on statistical disclosure control methods for micro-data protection.
2. is a dissertation on methods for disclosure risk assessment and disclosure risk limitation for business data.
3. compares results on the estimated level of safety for different definitions and estimation models of the disclosure risk.

These papers give some facts establishing that those approaches that were selected to be further developed and implemented into μ-ARGUS are indeed promising and can be expected to meet the need of users.

2.4 New disclosure limitation methodology for μ-ARGUS

The following section will describe the proposed methods briefly, considering particularly the results of the above mentioned studies relating to them.

Microaggregation techniques
When microaggregation is applied to a data set, records are clustered into small aggregates or groups of size at least *k* (c.f. [3], [5]). Rather than publishing a variable for a given individual, the average of the values of the variable over the group to which the individual belongs is published. In literature and practice several variants of microaggregation have been considered.

The empirical study [4] points out that only multivariate microaggregation of unprojected data performs well. This technique is however not yet state of the art. Consequently, development and implementation of such algorithms will be one of the tasks of the project. Another microaggregation development in CASC refers to categorical data. Microaggregation is a new approach for this kind of data, and the research carried out will result in practical algorithms drawing on recent advances in data aggregation and data fusion.

Random noise masking
Most masking methods combine transformations with adding noise (see e.g. [15], [21], [10]. Studies for personal data show that not all of these techniques are sufficient for protecting privacy ([16], [17]). Therefore, only one of these approaches proposed by Sullivan (see [21], [10]) will be extended and the results will be implemented in μ-ARGUS. The idea of Sullivan's algorithm which is applicable to both discrete and continuous data is to mask data in such a way that their analytical validity is maintained to a wide extent. In particular, univariate distributions, variance and co-variance structures shall be preserved.
Based on an exhaustive literature research, [1] singles out Sullivan's approach as a very promising method. In addition to that, [1] reports results of an empirical study of this approach, proving that this technique may result in analytically valid, safe micro data files as far as small and medium sized businesses are concerned.

Model based disclosure limitation
Based on the idea expressed in [9] that it is important to preserve the individual profile of each unit present in a micro-data file, disclosure limitation methodology will be developed, using on certain techniques of imputation methodology.

3. The situation for Tabular data protection

After introducing into the current situation regarding tabular data protection, section 3.1 will explain in which way the European Statistical System would benefit from a standard tool for tabular data protection, focusing particularly on what is needed for creating such a standard tool. Section 3.2 will refer to the results of best practice studies on tabular data protection. The reader will find that main goals of the CASC project concerning tabular data protection as listed in section 3.3 meet the users needs, and that our intention for the

further development of τ-ARGUS is in fact to integrate best praxis tools. Section 3.4 will describe in which way this overall-aim is addressed in detail.

3.1 Current situation and user's needs

Particularly in the case of magnitude data, most National Statistical Institutes and Eurostat apply non-perturbative techniques to create tables that can be released safely. Non-perturbative disclosure limitation techniques for tabular data are table redesign (reducing the level of detail in a table) and cell suppression. When cell suppression is used, table cells are suppressed if the data disseminator considers them too revealing. Additional cells must be suppressed in order to prevent these so called "primary suppressions", or "sensitive" cells, from exact disclosure or too narrow an estimation of them from the additive relationship between the cells of the table.

Need for automated secondary cell suppression systems

These additional, so called "secondary" suppressions, are currently still often assigned manually. Manual procedures for secondary cell suppression however are time consuming and error-prone, therefore resulting in unacceptably high costs, as well as causing the dissemination to be somewhat out of date. Moreover, if manual procedures have been used for disclosure control, cells regarded as confidential by the disseminator might be disclosed exactly, when for instance certain techniques of linear programming are applied to the published table. This may partly be due to errors and partly be due to the fact that in complex cases manual procedures cannot use the full inter-relationship structure of the published tables. In order to reduce the effort required for the manual cell suppression procedure, table redesign may be applied. If this is overdone however, substantial amounts of information will get lost, diminishing the usability of the data substantially.

In particular, non-optimal cell suppression patterns as resulting typically from manual suppression procedure in national level tables may have a damaging impact on European level tables. In the context of the structural business survey, Eurostat has complained in numerous papers about the large number of national level cells indicated confidential, resulting in a substantial part of European level information that cannot be published. It can be expected that the situation will improve substantially when NSI's select secondary suppressions according to a sophisticated algorithm.

Applicability to large tables, and tables with complex structures

Data collected within government statistical systems usually must be provided as to fulfil requirements of many users, users differing widely in the particular interest they take in the data. Statisticians try to cope with this range of interest in their data, by providing the data at several levels of detail. They use elaborate hierarchical classification schemes for categorisation of the respondents on various levels of detail. A respondent will often belong to various categories of the same classification scheme. Within a hierarchical classification scheme all respondents belonging to the same low-level category will also belong to the same categories on the levels above. In fact, within a hierarchical classification scheme any lower level category will belong to one and only one category on the level above.

The structure between the categories of hierarchical variables also implies sub-structure for the table. It is possible to 'partition' a hierarchical table into a set of sub-tables without substructure. Many cells of the original, hierarchical table will appear in more than one of

the sub-tables. Therefore, the sub-tables must not be protected separately. Otherwise, it might happen that the same cell is suppressed in one sub-table, because it is used as secondary suppression, while within another table it remains unsuppressed. A user comparing the two sub-tables would then be able to disclose confidential cells. So, usable software for secondary cell suppression must be able to deal with these complex structures.

Applicability to multiple tables

Due to technological advance, it is much easier nowadays for users of statistical data to compare and analyse suppression patterns in different tables. For overlapping (linked) tables, this fact increases the risk of disclosure to happen actually. So using proper procedures for co-ordination of suppression patterns in multiple (linked) tables is becoming an issue of growing importance. Although statistical institutes are aware of this problem, due to its complexity, this issue can be expected to gain practical relevance only when software for automated protection of linked tables is available.

Table-to-table protection in the context of data base query systems

Ideally a table-to-table protection procedure should be applied to the full set of tables ever to be published from this data source. This, however, seems less and less a realistic option. Nowadays, the process of releasing data turns to be more and more user demand driven and less pre-planned – to the extent even of providing public use data base query systems. This does cause serious trouble with cell suppression. The situation can be improved to some extent, when data are 'pooled' to keep track of suppressions in those tables which have already been published, while still others get newly created. However, this approach may result in the situation that the tables that are requested first will be published, preventing the subsequent (perhaps more relevant) tables from dissemination.

Co-ordination of suppression patterns

A need for co-ordination of suppression patterns arises for instance for those tables published periodically, monthly, quarterly, or annually. A part of the sensitive cells may be sensitive in every period. As simple illustration, assume a table without substructure, containing some sensitive cells which are 'forever' sensitive. Assume further, that there is more than one feasible suppression pattern. If nothing is done, it is then very likely that the suppression pattern changes from period to period, which might be undesirable and also cause a risk of disclosure, when the variation in the cell values of the secondary suppressions for different periods is only small.

Specific problems arise when data are published on different levels of a regional classification (on the national and on the supra national (EU) level, or on the regional and national level) but secondary suppressions are to be assigned by different agencies, actually (e.g. NSI's and Eurostat, or regional and national statistical institutes).

Table perturbation techniques

Users of a published table by making use of the linear relations between published and suppressed cells of the table would be able in principle to derive upper and lower bounds for the true value of any suppressed entry. The suppression procedures of proper cell suppression software will therefore ensure that no suppression pattern will be considered to be feasible unless disclosure of all these bounds does not cause any risk of disclosure for individual respondent data. Considering this, a data disseminator might as well publish

these bounds along with the protected table and could also publish perturbed values to replace suppressed original cell entries. These perturbed values should be located between the upper and lower bounds, matching subtotals and totals of the protected table. Such a technique would particularly be supported by Partial Cell Suppression, where the secondary suppression are selected in such a way as to minimise the intervals given by the resulting bounds of the suppressed cells [19].
As emphasised at the 2001 joint ECE/Eurostat work session on statistical confidentiality, such a facility would meet especially the needs of the users in the statistical institutes of the transition countries. Also, Eurostat has recently suggested perturbation-techniques for the protection of European level tables [7].

3.2 Best practice studies

Between 1996 and 1998 the German Federal Statistical Office carried out a comparison of software packages that protect tables by cell suppression. It was considered as basic requirement for such a system that the underlying algorithm used for the selection of secondary suppressions is able to avoid a certain disclosure risk for the protected table. In order to avoid this kind of risk it must be able at least of considering all the relations between the cells of simple two dimensional tables at once. So, although several statistical institutes may have implemented procedures similar to those used for manual selection of suppression patterns into software tools, the study involved only those systems satisfying this basic requirement. These were the systems of Statistics Canada CONFID and its commercial variant ACS, the German system GHQUAR, τ-ARGUS and software used by the US Bureau of the Census. Details and results of this study have been published in [11],[12] and [13].
The study concludes that all five systems solve the secondary cell suppression problem properly. The resulting suppression patterns are acceptable, with respect to both key criteria, disclosure control and information loss. Except for τ-ARGUS, all these systems had been in regular use for some time and were all considered to be good practice, although certain differences were observed. Regarding the disclosure risk for instance, the American and Canadian systems are superior concerning the assessment of disclosure risk and the avoidance of certain residual disclosure risks.
As a third key criterion, the study considered a system's applicability to large and complex structured tables. With respect to this criterion, GHQUAR has outstanding qualities. In the recently appeared new version of GHQUAR (GHMITER) the software is applicable to tables practically of any size and complexity of structure without requirement of further user interaction. Quite in contrast, τ-ARGUS is only applicable to simple, unstructured tables. On this set of tables, however, it performs extraordinarily well, giving the impression, that the underlying algorithm for the selection of secondary suppressions might have the potential to reduce the information loss due to cell suppression substantially as compared to the other systems [8]. Compared to GHQUAR, this reduction might amount to 30 to 50 % .
Generally, the study observed a trade-off effect between information loss and software speed. That means, those systems performing better regarding information loss, do require much longer computation times. This effect becomes substantial in large tables. So, for large applications, one has to choose between excellence and efficiency.
As very important feature of τ-ARGUS, the study rated the design of the software.

τ-ARGUS is the only system included in the study, where portability and transferability had been an issue in the design of the software. It is WINDOWS based and has a comfortable user-interface. Due to this design, it has a potential to serve as a platform for more than one cell-suppression algorithm, facilitating in this way the transfer of other products, or of the know-how of other products. As a platform for more than one cell suppression method, it has the potential to serve the needs of *all* users, and also to account for the dynamic nature of method-development in statistical confidentiality.

A second study on cell suppression software was carried out by Eurostat. The results of this study have not yet been published. However, as a consequence of the study, Eurostat started a co-operation with the developer of GHQUAR, developing a GHQUAR user-interface for the application of GHQUAR to European level tables.

3.3 Main Goals of the CASC project regarding the development and integration of new methodology for tabular-data protection in τ-ARGUS

Concerning tabular data protection, it is the objective of the project to develop a software package suitable to be established as standard tool for disclosure control of aggregated data, meeting the needs of all users, as described in section 3.1. This implies that the software package must be able to deal with tables of any size and complexity of structure, facilities must be offered to deal with specific problems of particular situations in a flexible, user-friendly and comfortable way. The software must be easily accessible and usable, and most of all, the package must be able to strike a good balance between quality and quantity. That is, the package should be able of offering, depending on the particular situation, (size and complexity of the particular application) the best suppression patterns (in terms of information loss due to suppression) efficiently achievable (in terms of computing resource requirements).

Specific aims to address this overall goal are the following:

(1) Refine and support the integration of desirable qualities and facilities of existing software systems for tabular data protection as identified in the best-practice study [11] into τ-ARGUS.
(2) Integration of the most recent version of the GHQUAR software, which will ensure wide applicability of the package to (linked) tables of any size and complexity of structure.
(3) Significant improvement of the cell suppression algorithms based on linear programming as already supplied along with the package, and supply of supplementary heuristic methodology.
(4) Provide information on the performance of the various algorithms for secondary cell suppression to be included in the final package. This information shall support the transfer of the package and will as well be useful for guiding internal decisions to be made during the project.
(5) Development and integration of table-perturbation methodology as described in section 3.1.
(6) Gain expertise with any newly implemented facilities for control of the selection of secondary suppressions and impart this expertise with potential users. In particular the 'European dimension' of the secondary cell suppression problem shall be addressed,

e.g. how to ease and sustain approaches of co-ordinating suppression patterns within Europe, as suggested e.g. by Eurostat for application to data of the structural business survey, c.f. [6].

3.4 Innovation for τ-ARGUS

In order to reach the above mentioned goals, further development of τ-ARGUS is necessary. In this section we will briefly describe the main issues of this innovation.

Extension of the kernel of τ-ARGUS
As a means of optimising the information content of protected data, algorithms will be supplied supporting the release of intervals for suppressed values, and of perturbed values to replace suppressed original ones.

Improvement of the cell suppression algorithms based on linear programming: In order to extend τ-ARGUS as to address the above mentioned user's need for a tool applicable to large, complex structured tables, it will be necessary to significantly improve the cell suppression algorithm of the current version and alternatively to include new algorithms: In the current version, the selection of secondary suppression within τ-ARGUS is carried out using (integer) linear programming (ILP) methodology. This method will be further improved in order to make it applicable to hierarchical tables at reasonable expense of computing time. Due to the enormous computational burden of the method, it will probably be impossible even with the improved version, to apply it to large multidimensional real-life tables.

Integration of GHQUAR: Besides some alternative heuristic strategies to improve the practicability of the ILP method, the GHQUAR hypercube software will be integrated into τ-ARGUS as an alternative tool for secondary cell suppression. In the best praxis study mentioned in section 3.2 above, GHQUAR has proven to be a most powerful tool for application to large, multi-dimensional tables.

Table-perturbation methodology: As a means of optimising the information content of protected data, algorithms will be supplied supporting the release of intervals for suppressed values, and of perturbed values to replace suppressed original ones. This meets the above mentioned user requirement for table-perturbation techniques.

Methodological strategies for implementing the extension of τ-ARGUS
In the current version, τ-ARGUS cannot handle tables with hierarchical substructure. There also is no option for table-to-table protection of linked tables, and if, due to a decentralised organisation structure like within in the European or the German statistical system, a potential user of the software is unable of providing the micro-data set on which the table he wants to protect is based, he cannot use the system. All these facilities will be offered by new versions, which will of course require modification and new concepts in the data structures. New data structures need to be designed for tabular data input, modifications will be made in the structure for the micro-data input, and the design of files containing meta-information, codelists, and other structural information on the tables has to be modified or newly invented.

In addition to that, strategies will have to be implemented to apply a suppression algorithm for unstructured tables to hierarchical tables, to carry out table-to-table protection, and to extend table-to-table protection efficiently in the context of data base query systems. Note that any of the strategies outlined below is already part of the new (GHMITER-) version of GHQUAR, or has already been implemented as prototype implementations of utility routines in the GHQUAR context.

Hierarchical tables: From the pure methodological point of view, it is actually the best strategy for protection of hierarchical tables, not to treat it as hierarchical table, but to turn it into one single(!) non-hierarchical very high dimensional table in advance. This is always possible and would be relatively simple to implement. The challenge of this strategy is to speed up the suppression algorithm for single, unstructured tables sufficiently, as to be powerful enough for application to real-life sized tables.
Because of the computational hardness of the secondary cell suppression problem, it is, however, clear in advance that it will not be an option to protect extremely large tables of several hundred thousand cells or more using linear programming methodology within such a single table approach. A common alternative approach is, to split the table into sub-tables and protect the sub-tables separately. Doing so, one must of course take into account that these sub-tables of the same table do have cells in common.

Multiple linked tables: Usually, some of the tables in the set of multiple tables published from the same source (response data from a survey) will be overlapping. When secondary cell suppression is carried out for tables having cells in common individually, then it is not unlikely that there will be cells unsuppressed in one table, while in others the same cells are complementary suppressions. Any user given access to all tables, will be able to disclose these values and may hence be able to recalculate sensitive cells.
Of course, there are possibilities of preventing this situation. One could protect a higher dimensional 'full' table, containing the set of linked tables.
Due to huge computer resource requirement this will often turn out to be impossible in practice. In this situation, new versions of τ-ARGUS will offer to apply a table-to-table protection procedure. Within a repeated procedure, each table would be protected separately, but ARGUS would keep track of any new suppressions belonging to overlap sections. A secondary suppression resulting from protecting one table will be treated like a primary suppression when protecting another table that contains this cell as well. The procedure would be repeated until no more cells get newly suppressed in the overlap sections.

Table-to-table protection in the context of data base query systems: Upcoming versions of τ-ARGUS will be able of setting up a 'data pool'. One might attempt to use such a data pool as data basis for public- or scientific use data base query systems. It shall be stressed here, however, that it is not at all within the scope of the project to implement such a data base query system. Nor do we claim that users or suppliers of data base query systems will be substantially happy with the level of detail or the proportion of unsuppressed low level cells in the data pool.
The data pool as corresponding to a particular micro data basis will contain one and only one record for each cell of any table already protected. This record will contain an entry regarding the suppression status of the cell. When a new table has to be protected, the

software will for any cell of this table investigate the data pool. Assume now, the data pool does already contain an entry for this cell. If the cell has already been used as secondary suppression in one of the tables processed earlier, then within a backtracking procedure, in the new table it will be treated like a primary suppression. If, on the contrary, the suppression status for the cell is 'unsuppressed' according to the data pool entry, then the software will attempt to avoid to select this cell as complementary suppression in the new table to some extent. Strategies like that will be implemented by 'freezing' the previously published cells, making them ineligible for suppression. The weaker variant - instead of 'freezing' these cells completely – would be to give them a low probability to be selected as secondary suppression. Running the 'freeze-variant' the user may sometimes be forced to abandon the new table, at least part of it. Running the weaker variant instead, he must be prepared to allow for an inconsistency between the suppressions patterns of a table already published and the new table, and hence put up with a risk of disclosure.

4. Research parts of the CASC project

As statistical data protection is a relatively new field in statistics, with most of the methods and techniques suggested and developed within the last decade, it is an extremely dynamic field. So, it is both natural and essential for the success of the project that it does not consist of software development only, but also involves a substantial portion of research.

Research tasks will involve both further development of the disclosure limitation methodology, as mentioned in section 2.4, and 3.4 above, and assessment of the quality. Quality assessment for statistical disclosure limitation methodology must basically focus on how the method performs regarding its two competing objectives which are

- to diminish the risk that sensitive information about or from individual respondents can be disclosed from a data set, e.g. minimise the disclosure risk, and on the other hand
- to maintain the analytical validity and completeness of the data to the extent possible, e.g. minimise the loss of information.

That means, the quality assessment must deal with these two issues: disclosure risk assessment and measurement of the information loss. Certainly, nobody will attempt to develop a method for disclosure limitation, without considering the performance of this method on these central issues. In this sense, research on quality assessment is a natural component of the research regarding methodology development. Apart from this 'natural' quality assessment research, the CASC project involves additional research components concerning quality assessment for statistical disclosure control:

4.1 Disclosure risk assessment

Micro-data protection

Disclosure risk for micro-data can be measured at either the file level or the record level. Record-level measures are useful for use in conjunction with disclosure limitation methods which are applied at the record level, for example local suppression. The main research objectives of the CASC project in this context are

- to refine methods as proposed in [20] based on log-linear models, and

- to investigate the application of record-linkage ideas(as suggested in [22], and [2]) to record-level measures of disclosure risk.
  The application of record-linkage approaches to assess disclosure risk, is based on the assumption that an intruder has an external data set containing as key variables some of the same variables that are present in the released masked data set. The intruder is assumed to try to link the masked data set with the external data set using suitable record-linkage techniques. In order to assess the disclosure risk for a micro-data file, a statistician would have to play the role of the intruder. The number (or percentage) of correct matches is then a measure for the disclosure risk.

Tabular data protection

Similar to the case of micro-data protection, cell suppression as disclosure limitation technique for tabular data involves two levels: In the first step the disclosure risk connected to each individual cell of the tables is assessed. Cells are suppressed when the data disseminator considers that they reveal too much information. In the second (table level) step, in order to prevent these so called ”primary suppressions”, or ”sensitive” cells, from exact disclosure or from being closely estimable from the additive relationship between the cells of the table, additional cells (so called ”secondary” or ”complementary” suppressions) must be suppressed.

- In the current version of τ-ARGUS, for disclosure risk assessment on the cell level the user is offered to state a minimum number of respondents and to specify a dominance rule. Use of dominance-rules is certainly ‘good practice’. However, it has been proven to be ‘better practice’ to replace these rules by similar ones. For instance to use the so called ‘p%-rule’ instead of a (2,k)-dominance rule. Upcoming version of τ-ARGUS will offer this.
- Regarding disclosure risk limitation on the table level, the specification method for the protection level and for modelling the user knowledge, resulting in upper and lower bounds for the cells, will be adapted to best praxis methodology.
- The problem of disclosure limitation for cell combinations will be addressed specifically. Methodology to solve this problem is offered in [14] and [18]. From these, we will select the most suitable alternative.

4.2 Research regarding information loss

All risk control methods degrade the data to some extent. It is of course essential for the development of disclosure limitation strategies, to bare this in mind.

Micro-data protection

The degradation of the data does of course reduce the ability of data users to conduct the analyses they need for their legitimate purposes. These effects fall into two categories:

- Reduction of analytical completeness: Some control methods, typically the recoding of taxonomic schemes into coarser categorisations, mean that analyses that could have been conducted with not-recoded data cannot be done.
- Loss of analytical validity: The loss of analytical validity is harder to define, but in some ways more critical because of its insidious nature. Technically, loss of validity can be said to occur when a disclosure control method has changed a dataset to the

point where a user reaches a different conclusion compared with the same analysis on the original dataset.

An attempt will be made to categorise the effects on analytical power of the full range of disclosure control techniques and to examine the feasibility of developing methods for measuring the scale of such effects.

Tabular-data protection

Benchmarking: Any of the algorithms for selection of secondary suppressions developed in the course of the project will be run on a set of real-life test tables, calculated on the basis of a specially created, masked micro-data file. Performances with respect to certain key issues (information loss in terms of number and/or total value of suppressions, computing time trade-off effect) will be recorded. This information will be useful for guiding internal decisions to be made during the project. Later on, it shall support the transfer of the package, supplying potential users with information on the performance of the package or of particular algorithms included, in advance of procuring it.

Co-ordination of suppression patterns: Addressing the need for co-ordination of suppression patterns as explained in section 3.1, we will test the feasibility of several approaches to improve the situation with a particular view on the practicability of any methods suggested. The methods will be applied to several suitable real-life data sets. Methods turning out to be promising shall be supported by the software package. Special options may be included in the software facilitating their use.

5. Technology transfer in the field of statistical confidentiality

In the first place, the CASC project will address problems of technology transfer by integrating best praxis methodology along with results from the research part of the project into the ARGUS software as outlined in the previous sections. While doing so, we certainly will have to supply a lot of user guidance and default options for 'beginners' and less experienced users.

In this context, it is another important issue to make the software very flexible, including special options for skilled users. As it must be expected that the project capacity (chiefly: the capacity available for software development) will not suffice to implement all the user-friendly facilities that may prove to be useful, we will at least make the design of the system open enough to allow for the experienced user supplying his own procedures (outside the system) or using manual intervention, to run the suppression procedure according to his individual needs.

Project tasks particular addressing issues of technology transfer are the testing of methodology and software, the exploitation plan, and the user training as will be outlined below.

5.1 Testing of ARGUS

A central feature of the CASC testing task regarding technology transfer is the fact, that it will involve nine potential user institutes, five NSI's and four other official statistical institutes.

Methodology testing is supposed to assure that the methods supplied in the package meet

the needs of the users. The users will check whether the methods are applicable to their data sets, and perform satisfactorily.
Software testing aims at testing the capabilities of the Argus software and at establishing its correct implementation. This task involves a check of the software-reliability, of its user-friendliness, performance and portability from the users perspective.

5.2 Exploitation plan

Except for the testing, other TTK issues, such as post-technology-transfer maintenance and updating, dissemination strategies, provision of training and technical assistance have also been addressed by the CASC project consortium during the project set-up. This exploitation plan as provided by the CASC project proposal, will be specifically supported by means of the AMRADS project. Within the framework of take-up actions and support measures of the IST programme, the project AMRADS (Accompanying Measure to R&D in Statistics) receives funding from the European Commission. Members of the CASC group are participating in this project. As part of the AMRADS contract, AMRADS partners will assist in creating infrastructure to facilitate the dissemination of CASC results.

Dissemination of the results
The new techniques developed within the CASC project will require a careful introduction. It will be a major responsibility to educate the staff of the NSI's. Information technology will be used in order to disseminate information, as well as implementing a helpdesk. Members of the AMRADS consortium will physically create an infrastructure comprising a repository of documentation, a website, thematic help-desks and the SODECE. The SODECE, which will be located in Luxembourg (with access to Eurostat's diverse expertise), will provide facilities to host demonstration and training.

The core technology involved in setting up this infrastructure will be:
- A Microsoft Windows based platform with the latest service pack
- Internet Information Server (for HTTP and FTP services)

These activities will be supported by a newsgroup, a dedicated web page, a well integrated FTP service, so an easy-to-use user web interface will be implemented.

The aim of the newsgroup will be the implementation of an easy helpdesk mechanism, as well as providing a discussion environment to the users of the ARGUS software. CASC project partners will manage the incoming user messages, providing the appropriate answers and supporting the discussions.

A Web Server supported by a FTP server will provide the mechanism for disseminating information. An attractive and user-friendly set of web pages will guide the user through the information activities and to gaining access to the helpdesk system. Of course, the connection from the most common client browsers (typically Internet Explorer and Netscape Navigator) will be guaranteed.

Exploitation
The internal exploitation is already guaranteed by the participation of so many NSI's in this

project. This is the best guarantee for the application of the results in the daily processes of these NSI's. The willingness of these NSI's to participate in the testing phase with real-life situations will ensure that the results will have a real value for the end-users and will readily lead to exploitation. Also it is to be expected that due to the courses on SDC and the presentation of the results the external uses (other NSI's and other data producing agencies) can be expected.
At this stage, it is difficult to estimate the size of the commercial market for these products. If we will gain a reasonable installed base for these products, this will ensure the need for the NSI's and other institutes to guarantee future developments in this field.

Involvement of end-users
The end-users of the methodological and practical advances are primarily the national and European Statistical Institutes, which are facing the problems of Statistical Confidentiality in their daily work. The outcome of this project will facilitate this. To ensure the proper implementation of the products in the daily practice, several potential end-users are participating already in this project. This will guarantee essential feed-back from the potential end-users to the development-team.

Future developments
As the development of the ARGUS software and all the supporting actions have been subsidised by the European community, we will consider these results (μ-ARGUS and τ-ARGUS) as free software. We will be happy if the results of our work will be heavily used in the NSI's and other data producing institutions. By assisting the NSI's in the use of ARGUS we will do our best to ensure that we will come to a de facto standard for this kind of software. Through this project, but also by participating in the AMRADS-project, we will ensure that the results will be widely introduced. Once there is a larger group using ARGUS, the participation of so many NSI's is a guarantee for this, there will emerge a growing need for these tools.
We cannot make decisions now, how to organise the further development of ARGUS when this project is over. However, if a commercial market for this software will exist and the future development of ARGUS is guaranteed by this, the project team might consider this as an option and devote the resulting software to the group who guarantees this further development. Most likely, this team will consist of (a group of) the current project team. During the course of the project, the project team will take a decision of the status of the software. Either a new consortium will take over the development or we will denote the software to the public domain.

Scientific prospects
The composition of the project team (including several academic partners) will result in many scientific papers. Within the field of statistics, we will expect results on e.g. micro-aggregation, noise addition masking and model-based masking. The research in this project is not only in the field of statistics, but also significant progress is expected in the field of operational research. This with respect to the problems of secondary cell suppression for tabular data. These papers will be presented at relevant scientific meetings and made available through the CASC Web-site. Where possible, we will submit these papers to scientific journals for publication. This guarantees that the results will be widely available and known.

5.3 User training, working groups, workshops

Courses in the framework of TES are an obvious vehicle for the user education. In particular, the currently available TES-course will be extended with modules to educate the NSI-staff in the use of the techniques implemented during this project in both μ-ARGUS as well as τ–ARGUS. As far as is appropriate, training will revolve around the SODECE. Training of trainers would be preferred because of the multiplier effects. It will also be considered to offer follow-up courses, in order to teach already experienced users to use more specialized tools of the software that would be probably not of much interest for new users who did not yet experience the need for such facilities. These training courses will be delivered by members of the CASC project team. The staff for a training course will include at least one ARGUS software expert and experts in micro-data and tabular-data protection, respectively.

There are already good platforms for the discussion of statistical disclosure control (the UN-ECE/EUROSTAT work sessions and the EU-sponsored SDP-conferences), which we will actively use for the dissemination of the results of CASC.

If feasible, it might be a good idea to offer in addition a workshop for special, homogenous user groups. Such a group could be the group of NSI and Eurostat structural business statistics (SBS) directors. For several years now, the discussion of questions and possible solutions related to statistical disclosure limitation has been a permanent issue at the regular SBS meetings. So, in a late stage of the CASC project, it would be good to bring the ARGUS solutions close to them, offering a workshop in context with an SBS meeting.

6. User needs, other research: Discussion of invited ETK-NTTS papers

Finally we will give a summary of the other papers of this NTTS/TTK session on statistical confidentiality.

- Statistical Disclosure Control from a users' point of view, *Eric Schulte Nordholt*
- Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk, *J. Domingo-Ferrer*
- On properties of multi-dimensional statistical tables, *L. Cox*
- Applicability of Latin Hypercube Sampling Technique to create multi variate synthetic micro Data, *R. A. Dandekar*
- Methods for Data Directed Microaggregation in One Dimension, *G. Sande*
- Safe Dissemination of Census Results by Means of Interactive Probabilistic Models, *J. Grim*

The first paper (by *Eric Schulte Nordholt*) describes the application of Statistical Disclosure Control software (i.e. ARGUS) in the practical daily situation of a Statistical Office. It shows the need for these kind of software tools. It discusses both the disclosure control of microdata as well as tabular data. Although much progress has been achieved, the final solution has not yet been reached. It also pays attention to the needs of the transition countries. It clearly shows the need for further training and support to introduce these methods in the transitions countries. However, also other European countries could

benefit from these kind of training programs.

Josep Domingo Ferrer (a member of the CASC steering committee) compares and gives an overview of different SDC-methods in his paper [4]. He focuses on the trade-off between information loss and the risk of disclosure. Disclosure risk measures and record linkage techniques have been used to measure the identification risk of the different methods. For continuous variables he compares additive noise, data distortion by probability distribution, resampling, microaggregation, lossy compression, rank swapping. For categorical data top/bottomcoding, global recoding and rost randomisation (PRAM). He concludes that for continuous variables rank swapping and multivariate microaggregation performs best while top-coding gives satisfactory results for categorical variables.

Larry Cox reports on his recent studies on higher dimensional tables. He shows that there exist unexpected restrictions/relations in these higher dimensional tables, which are not present in two-dimensional tables. From the viewpoint of Statistical Disclosure Control, these restrictions imply additional problems for data protectors, as they restrict the freedom of choosing a secondary suppression pattern.

Ramesh Dandekar proposes an interesting alternative for the dissemination of micro data files. His proposal is, to generate a synthetic file while preserving necessary statistical properties. A good synthetic file has many benefits from the SDC point of view. It might serve as an alternative, but could also be used in conjunction with OnSite Data Archive centres. Researchers could use the synthetic file to test their hypotheses and only run their final setup's on the original data in the DAC. A side effect could be that these techniques will reduce the burden of a DAC on the NSI's. Although a DAC is a valuable option, its operation requires a big effort for the organising institute.

Dandekar uses Latin Hypercube Sampling to produce a synthetic data set that reproduces many of the essential features of an original data set while providing disclosure protection and demonstrates this on a larger dataset. He shows that this method has certain advantages, although we must be realistic and must investigate in further research.

The paper by *Gordon Sande* describes an application of micro-aggregation. Gordon Sande has a long record of work in the field of Statistical Disclosure Control.
His current paper describes a technique that has drawn also much attention at Eurostat. It focuses on the release of microdata files. Especially for the economic microdata we face large problems as the entities on these data files are much easier to identify than that in social survey data. Basically, microaggregation will replace the value of a certain variable of respondent by grouping similar respondents together and replace the value by the average of this group. Most 'harm' is done in the tails of the distribution, where most of the protection is needed. In the middle of the distribution (where less protection is needed) the information is much better preserved. Several improvements over the original simple univariate micro-aggregation are discussed, like varying group size and certain optimisation techniques to find more optimal micro-aggregation solutions

Jiri Grim presents a dissemination tool for census data. It is obvious that the original data

cannot be made available. As an alternative he describes a method, where not (adapted) data is made available, but the system satisfies the needs of the data-users by supplying the information using probabilistic models

7. Summary

The paper has explained the need for a standard tool for statistical disclosure limitation in the European Statistical System and the benefits that can be expected. It has described current best practice methods and tools in the field, explaining how they have been studied and were identified as such.
The main objective of the paper was to illustrate in which way the CASC project addresses problems of technology transfer: In the first place, by integrating best praxis methodology along with results from the research part of the project, the main tasks of which have been outlined in this paper as well, into a portable user-friendly tool that can be provided on free cost basis for the main part.
Particular TTK issues, such as post-technology-transfer, maintenance and updating, dissemination strategies, provision of training and technical assistance have also been addressed by this paper. In this context, plans for user training, the establishment of a working group and organisation of a workshop have been mentioned.
In its function to serve as keynote-paper for the confidentiality session of the 2001 ETK NTTS conference, the paper also provided brief discussion of the other papers of this session.

Internet links:

First UN-ECE/EUROSTAT work session:
http://www.unece.org/stats/documents/1999.03.confidentiality.htm
Second UN-ECE/EUROSTAT work session:
http://www.unece.org/stats/documents/2001.03.confidentiality.htm
CASC-project homepage: http://neon.vb.cbs.nl/casc/

Acknowledgements

The authors are grateful to the other partners in the CASC team. This paper could not have been written without their contribution to the project proposal. Special thanks go to Ruth Brand and Josep Domingo-Ferrer for reviewing this paper.

References

[1] Brand, R., 'Anonymität von Betriebsdaten, Verfahren zur Erfassung und Maßnamen zur Verringerung des Reidentifikationsrisikos', dissertation (in German)

[2] Copas, J. B. and Hilton, F. J. (1990) Record Linkage: statistical models for matching computer records, (with discussion) *J. Roy. Statist. Soc., A,* 287-320

[3] Defays, D., Nanopoulos, P., (1993), Panels of enterprises and confidentiality: the small aggregates method, in Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa: Statistics Canada, 195-204

[4] Domingo-Ferrer, J., Mateo-Sanz, J.M., (2001), Comparing SDC Methods for Microdata on the Basis of Information Loss and Disclosure Risk, proceedings of the NTTS&ETK 2001 New Techniques and Technologies for Statistics, Exchange of Technology and Know-How conference, (to appear)

[5] Domingo-Ferrer, J., Mateo-Sanz, J.M., (2002), Practical Data-Oriented Microaggregation for Statistical Disclosure Control, IEEE Transactions on Knowledge and Data Engineering, (to appear, March 2002)

[6] Eurostat, working paper, Doc. Eurostat/D2/SBS-T/NOV99/03

[7] Eurostat (2000) "Confidentiality of EU15 Aggregates", working paper for the meeting of the working group "Structural Business Statistics" on 12 & 13 September 2000, Doc. Eurostat/D2/SBS/SEPT00/03/EN

[8] Fischetti, M and JJ. Salazar (1998) , 'Modelling and Solving Cell Suppression Problems in linearly-constrained tabular data', *Statistical Data Protection 1998, Lisbon, Portugal*

[9] Franconi, L.(1999), 'Level of safety in microdata: comparisons between different definitions of disclosure risk and estimation models', proceedings of the Eurostat/UN-ECE Work Session on Statistical Data Confidentiality 1999

[10] Fuller, W. A. (1993) Masking procedures for microdata disclosure limitation, Journal of Official Statistics, 9, 383-406

[11] Gießing, S. (1998), 'Looking for efficient automated secondary cell suppression systems: a software comparison', Research in Official Statistics Journal 2/98

[12] Gießing, S. (1999), 'A survey on packages for automated secondary cell suppression', proceedings of the Eurostat/UN-ECE Work Session on Statistical Data Confidentiality 1999

[13] Gießing, S. (1999), 'Vergleich der Software zur maschinellen Durchführung der Sekundären Geheimhaltung', In: *Forum der Bundesstatistik, Band 31/1999: Methoden zur Sicherung der Statistischen Geheimhaltung* (in German)

[14] Jewett, R. (1993), 'Disclosure Analysis for the 1992 Economic Census. Unpublished Manuscript. Economic Statistical Methods and Programming Division, Bureau of the Census, Washington, DC.

[15] Kim, J.J. (1986) 'A Method for Limiting Disclosure in Microdata based on Random Noise and Transformation', *Proceedings of the Section on Survey Research Methods 1986, American Statistical Association, Washington D.C.*, 303-308

[16] Moore, R.A. (1996a) 'Controlled Data-Swapping Techniques for Masking Public Use Microdata Sets', *Statistical Research Division RR96/04, US Bureau of the Census, Washington, D.C.*

[17] Moore, R.A. (1996b) 'Analysis of the Kim-Winkler Algorithm for Masking Microdata Files – How Much Masking is Necessary and Sufficient? Conjectures for the Development of a Controllable Algorithm', *Statistical Research Division RR96/05, US Bureau of the Census, Washington, D.C.*

[18] Robertson, D. (2000), 'Improving Statistics Canada's cell suppression software (CONFID)', *Proceedings of the Compstat 2000 conference 21.-25. August, Utrecht, Netherlands*

[19] Salazar, JJ and M. Fischetti (1998), 'Partial Cell Suppression: a new methodology for Statistical Disclosure Control, University La Laguna, Tenerife, Spain.

[20] Skinner, C. J. and Holmes, D. J. (1998) Estimating the re-identification risk per record in microdata. *J. Official Statist. 14*, 361-372

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

[22] Winkler, W. E. (1998) Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics*, *2*,87-104