Working Paper No. 38
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)**

Joint ECE/Eurostat work session on statistical data confidentiality
(Luxembourg, 7-9 April 2003)

Topic (vi): Software tools for statistical disclosure control

THE STATISTICAL PROTECTION OF THE EUROPEAN STRUCTURE OF EARNINGS SURVEY DATA

Contributed paper

Submitted by Statistics Netherlands[1]

[1] Prepared by Eric Schulte Nordholt (else@cbs.nl).
Remarks:
The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands. The author thanks Derek Bird (ONS), Richard Clare (Eurostat) and Luisa Franconi (Istat) for their contributions to this paper.

THE STATISTICAL PROTECTING OF THE EUROPEAN STRUCTURE OF EANINGS SURVEY DATA

Eric Schulte Nordholt

Summary:

The twin ARGUS software has been applied to earnings data from several EU countries. In particular, the application of m-ARGUS to Dutch microdata from the Structure of Earnings Survey (SES) identified that the collapsing of the regional variable (at the NUTS 1 level) was sufficient to anonymise the microdata to an extremely high degree. Very similar results were obtained in Italy and the UK. Nevertheless, some individuals (businesses and employees) could still be identified. In order to deal with the small number of disclosive cells that remained, t-ARGUS was applied to the Dutch tabular data in order to remove any remaining risks of disclosure. In this paper the software package t-ARGUS is described that can be applied for producing safe tabular data. The main techniques used to protect sensitive information are global recoding and local suppression. The packages m- and t-ARGUS are products developed in the SDC (Statistical Disclosure Control) project under the Fourth Framework Programme of the European Union. A new version of the package (that includes recent research results) has been released in the CASC (Computational Aspects of Statistical Confidentiality) project that is funded under the Fifth Framework Programme of the European Union. In 2003 again a new version of t-ARGUS will be released in the CASC project. In this paper, methods are described that have been developed to protect tables, through various means that either alter the data or restrict access to them. After combining categories of the spanning variables of the tables, the remaining sensitive cells are suppressed using t-ARGUS. In this way, statistical disclosure control techniques help in keeping the right balance between data confidentiality and data access. Eurostat followed a very good strategy to plan the statistical protection of the European Structure of Earnings Survey data well in advance. This way modern technology could be applied and best practices could be exchanged in the Expert Group preparing the protection of the SES. The preparations will help managing the dissemination at the European level.

Keywords: CASC; European Structure of Earnings Survey; Microdata; m-ARGUS; SDC; Software; Tabular data; t-ARGUS.

1. Introduction

National Statistical Institutes (NSIs) conduct surveys about many different topics. To reach this aim they have developed a fully-equipped statistical production process. The information from statistics becomes available for the public in tabular and microdata form. Historically, only tabular data were available and NSIs had a monopoly on the microdata. Since the eighties the PC revolution leaded to the end of this monopoly. Now, other users of statistics also have the possibility of using microdata. These microdata can be conveyed with floppy's, CD-roms and other means. Recently other possibilities of getting statistical information have also become more popular as remote access and remote execution. With these techniques researchers can get access to data that remain in a statistical office or can execute set-ups without having the data on their own PC. For very sensitive information some NSIs have the possibility to let bona fide researchers work on-site within the premises of the NSI.

The task of statistical offices is to produce and publish statistical information about society. The data collected are ultimately released in a suitable form to policy makers, researchers and the general public for statistical purposes. The release of such information may have the undesirable effect that information on individual entities is disclosed. 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 jeopardize the privacy of the entities concerned. Statistical disclosure control theory is used to solve the problem of how to publish and release as much detail in these data as possible without disclosing individual information (Willenborg and De Waal, 1996 and 2001). Practical applications can be found in Schulte Nordholt (1999 and 2000).

This paper discusses the available methods to protect sensitive information. The tables produced by statistical offices on the basis of the microdata of surveys have to be protected against the risk of disclosure. If the microdata are safe the tables produced are automatically safe as well. However, the microdata used for producing tables are often not quite safe and therefore some additional measures have to be taken. For the Structure of Earnings Survey 2002 μ-ARGUS will be applied to protect the microdata to a large extent. The software package τ-ARGUS can thereafter be applied to the tables produced in order to protect the tabular data released. The experiments conducted with μ-ARGUS are described in section 2. More information about τ-ARGUS and how this package can be applied is given in section 3. Finally, in section 4 some conclusions are drawn. In the appendix a flowchart is shown which describes the future process of protecting the European Structure of Earnings Survey data.

The software package τ-ARGUS and its twin μ-ARGUS to protect microdata have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework Programme of the European Union. The Computational Aspects of Statistical Confidentiality (CASC) project can be seen as a follow-up of the SDC project. The CASC project, which has been designed around the twin ARGUS software, is funded under the Fifth Framework Programme for Research, Technological Development and Demonstration (RTD) of the European Union. It builds further on the achievements of the SDC project. In the CASC project fourteen partners from five different European countries (Germany, Italy, the Netherlands, Spain and the United Kingdom) work closely together. One of the main tasks of this consortium is to further develop the ARGUS-software (Hundepool et al, 2002a and b) which has been put in the public domain by the SDC project consortium. The CASC project involves both research and software development and concentrates on those areas that can be expected to result in practical solutions, which can then be built into the software. More information about the CASC project can be found in Hundepool (2001).

2. Protecting microdata

2.1 Introduction

This section deals with the assessment of different proposals for the anonymisation of the European Structure of Earnings Survey (SES) microdata. It describes the methodology used to anonymise the data, the results obtained in three countries and the initial conclusions from this work. This section can be considered as a summary of the work of the Expert Group on SES data confidentiality.

A SES microdata file contains a record for each employee in the sample and certain information on the local unit / enterprise in which the employee works. The aim is to produce a microdata file that is anonymised so that it is extremely unlikely that local units / enterprises and employees can be identified. Identification occurs when it is possible to recognize the name of the enterprise or individual. Such recognition is made possible by means of identifying variables. For enterprises in the SES these are mainly region, size of enterprise and economic activity. For individual employees, the identifying variables include age, occupation, education and citizenship. The more detail that is available, the easier it is to identify a unit. For example, with the NACE division, the size of the enterprise and the NUTS 1 region, it might be possible to identify many enterprises in the sample. In contrast, with no NACE, size or regional breakdown, it is impossible to identify a business.

To lower the information detail and therefore reduce the identification risk, global recoding is applied. This simply combines some categories of an identifying variable. In the previous example, global recoding can be used to restrict the NACE detail to 1-digit, to collapse some of the size bands, or to consider SES data only at the national level.

To quantify the identification risks the following steps are performed:

(i) choose a criterion to define a unit at risk of being identified;
(ii) using this criterion, count the number of units in the microdata file that are at risk;
(iii) if the number of units at risk is too high, apply (further) global recoding and return to step (ii) above; otherwise the degree of anonymisation obtained is acceptable.

The criterion used to define a unit at risk of being identified is the threshold rule (described in subsection 2.2 below). In order to assess different global recoding schemes (each with different restrictions on the identifying variables), experiments were carried out on 1995 SES-type micro-data sets. Subsection 2.3 gives background information about the analyses undertaken. Subsection 2.4 presents the numerical results. In subsection 2.5 further measures are proposed. Subsection 2.6 contains the conclusions from the work undertaken to date.

The treatment of confidential data represents an important element in Eurostat's plans for handling the SES data for the year 2002 – from the initial receipt of the microdata, right through to the final dissemination of the tabular results. The appendix illustrates the SES 2002 flowchart from start to finish, including the steps Eurostat is taking to protect the confidentiality of the SES data.

2.2 Criterion to define units at risk: the threshold rule

When business units are classified according to the identifying variables (for instance NACE, size, region) the total number of units are distributed among the cells of the resulting matrix. The risks of identification are clearly greatest for those cells with only one or two units. The threshold rule relates to the number of units that fall within a particular cell. In practical terms this means that for each cell (representing a combination of the identifying variables) at least three units must be present either in the sample, or in the population. The sample threshold and population threshold need to be distinguished. To check how many units are at risk with a sample threshold, a table is constructed with all the identifying variables of interest; then a count is made of how many units are in cells with less than (say) three units. If a population threshold is preferred, the weighted combination of scores is checked, using the sampling weights.

Which of the two rules are applied (a sample or population threshold) depends on the legal position in each member state. If there is no legal requirement, the member state can select the threshold which best fits the circumstances. If there is a risk that intruders have response knowledge, a threshold on the sample will prevent them from identifying units. A population threshold takes into account the fact that, especially for small enterprises, even if the number of sampled units is small, the total number of small units is large and identification is difficult.

2.3 Different global recoding schemes for variables identifying enterprises: background

The results of different recoding schemes applied to different sets of microdata are presented in subsection 2.4. The identifying variables involved in these schemes were: region, economic activity and size of the local units / enterprises. The ultimate objective with global recoding is to try to identify the 'best' scheme which successfully anonymises the microdata without sacrificing too much detail for one

or more of the identifying variables. Some of the results that follow concern local units / enterprises, whereas others refer to individual employees working inside those enterprises. This is because different member states have different views on statistical disclosure control.

The results of each scheme show the number of cells where units (or individuals) are at risk because the number of units is less than the threshold. The maximum number of units at risk is the number of cells at risk multiplied by the threshold minus one. Conventionally, the threshold is taken to be 3 units. So, if the number of cells below the threshold (i.e., with 2 or fewer units) is 10, then the maximum number of units at risk is 20.

For the results presented in subsection 2.4, only identifying variables related to the business are considered. Identifying variables for individuals (age, sex, length of service, occupation, educational level, etc.) are considered in subsection 2.5.

Finally, the analyses were undertaken by three member states: the Netherlands, Italy and the UK. For the first two countries, the analyses were carried out using the 1995 SES microdata. For the UK, use was made of the 2001 New Earnings Survey (NES). Not all the analytical work undertaken is presented here. Subsection 2.4 provides a selection of the results.

2.4 Results of different global recoding schemes for variables identifying enterprises

In common with the larger member states, the Netherlands, Italy and the UK produce a regional breakdown (at the NUTS 1 level) for the SES. Because region is a sensitive identifying variable for a business, the results of retaining a full NUTS 1 breakdown was assessed in the first and second global recoding schemes below.

First scheme

Region:	the full NUTS 1 breakdown was retained;
Size of the enterprise:	was restricted to 3 bands:10-49, 50-249, 250+;
Economic activity:	the 2-digit NACE Rev.1 breakdown was retained.

The number of enterprises at risk when this scheme is applied to the Italian SES data (using a population threshold of 3) is very high: 158 cells out of 1050 were at risk. The UK work on the 2001 NES data (with a sample threshold of less than 4 individuals) identified 207 cells at risk. The test on the Dutch employee data from the 1995 SES (using a sample threshold of 3 for individuals) also showed that a very high number of records were at risk.

Second scheme

Region:	the full NUTS 1 breakdown was retained;
Size of the enterprise:	was restricted to 3 bands:10-49, 50-249, 250+;
Economic activity:	was restricted to 1-digit NACE Rev.1.

Under this scheme, the NUTS 1 breakdown was retained plus the same size breakdown. However, global recoding was applied by limiting the NACE breakdown to the 1-digit level. The anonymisation was not very successful for Italy and the Netherlands as the results for these countries continued to identify a high number of cells at risk. In the UK, the identified cells at risk fell to 22, indicating a significant improvement compared with the first scheme.

Third scheme

Region:	restricted to the national level (no NUTS 1 breakdown);
Size of the enterprise:	10-49, 50-249, 250-499,500-999,1000+ (ie all sizes);
Economic activity:	the 2-digit NACE Rev.1 breakdown was retained.

In order to make progress, the regional breakdown was dropped and only national data was used. However, no restrictions were placed on the size or NACE breakdowns. This scheme enormously improved the position in all three countries. The number of cells at risk when this scheme is applied to

the Italian data is equal to 22 (see Table 1). For the UK data the number of cells at risk was 6 (see Table 2). The Dutch analysis identified 9 individuals at risk (see Table 3).

Fourth scheme

Region: restricted to the national level (no NUTS 1 breakdown);
Size of the enterprise: was restricted to 3 bands:10-49, 50-249, 250+;
Economic activity: the 2-digit NACE Rev.1 breakdown was retained.

To anonymise the data further, the enterprise size variable was restricted to three bands (the top 3 size bands 250-499,500-999,1000+ being collapsed into one band: 250+ employees). This scheme produced a further reduction in the number of cells at risk to just 2 cells in Italy and the UK (see Tables 1 and 2, respectively). While the risk in the Dutch analysis remained at 9 individuals (see Table 3), this should be seen relative to the total number of individuals of 106 755, so the number of problematic cases is below 0.01 %.

Table 1. *Italian data from the 1995 SES. The number of cells where the threshold rule on population frequencies for enterprises is not met. The variable region is recoded to the national level, economic activity is at 2-digit level of NACE and the enterprise size bands are collapsed from 5 to 3 bands corresponding to the third and fourth schemes above.*

Global Recoding	Variable: Size of enterprise (in bands of employees)					Total number of cells at risk
	10-49	50-249	250-499	500-999	1000 +	
Third scheme	0	1	2	6	13	22
Fourth scheme	0	1	1			2

Table 2. *UK data from the 2001 NES. The number of cells where the threshold rule on sample frequencies for individuals is not met. The variable region is recoded to national level, economic activity is at 2-digit level of NACE and the enterprise size bands are collapsed from 5 to 3 bands corresponding to the third and fourth schemes above.*

Global Recoding	Variable: Size of enterprise (in bands of employees)					Total number of cells at risk
	10-49	50-249	250-499	500-999	1000 +	
Third scheme	2	0	2	2	0	6
Fourth scheme	2	0	0			2

Table 3. *Dutch data from the 1995 SES. The number of individuals for whom the threshold rule on sample frequencies is not met. The variable region is recoded to national level, economic activity is at 2-digit level of NACE and the enterprise size bands are collapsed from 5 to 3 bands corresponding to the third and fourth schemes above.*

Global Recoding	Variable: Size of enterprise (in bands of employees)					Total number of individuals at risk
	10-49	50-249	250-499	500-999	1000 +	
Third scheme	5	4	0	0	0	9
Fourth scheme	5	4	0			9

The results in subsection 2.4 show that the NUTS 1 variable is a very sensitive identifying variable and its retention dramatically increases the number of problematic records. SES microdata can be very significantly anonymised by using data at the national level without collapsing either the size or NACE breakdown. Collapsing the top 3 size bands further reduces the risk of identification to negligible proportions.

To further lower the information detail on business units, it is proposed to exclude from the anonymised microdata the following SES variables.

- Form of economic and financial control;
- Total number of employees in local unit (optional);
- Size of the group of enterprises (optional);
- Country of residence of the entity controlling the group (optional).

2.5 Further anonymisation measures and implementation of possible rules

The analyses above made use only of the identifying variables for businesses. In addition five identifying variables for individuals were seen to be: sex, age, level of education, occupation and citizenship. With regard to the identifying variables for individuals, the following global recoding was undertaken (on top of the global recoding for businesses) on the Italian, UK and Dutch data. Sex was not recoded and level of education was considered in 7 categories based on the ISCED 97 classification. Citizenship (an optional variable) was sacrificed. For age and occupation different recodings were undertaken on the data of the different countries.

Italian data

Age:	collapsed into 6 bands (14-19, 20-29, 30-39, 40-49, 50-59, 60-);
Occupation:	restricted to 9 categories of the ISCO 88 classification.

UK data

Age:	collapsed into 6 bands (14-19, 20-29, 30-39, 40-49, 50-59, 60-) and smaller age bands (-16, 16-19, 20-24, 25-29, …, 55-59, 60-64, 65-);
Occupation:	restricted to the 1 digit level and also to the 2 digit level of the ISCO 88 classification.

Dutch data

Age:	five year class intervals (-19, 20-24, 25-29, …, 55-59, 60-64, 65-);
Occupation:	restricted to 5 occupational groupings (according to the ISCO 88 classification): elementary, low level, middle level, high level and academic occupations.

Additionally, top coding of the income variable is recommended. This could imply fixing a threshold value 'x' and assigning this value to all incomes higher than 'x'. For example, suppose that 2 % of the employees in the SES have an income higher than 100 000 Euro. Then, for these individuals, their income will be put equal to 100 000 Euro.

The rules and the global recoding proposed in this document have already been implemented in the software package µ-Argus, one of the products of the CASC (Computational Aspects of Statistical Confidentiality) project in the Fifth Framework Programme for Research, Technological Development and Demonstration (RTD) of the European Union. It is therefore possible to use the capabilities of µ-Argus for a fast production of the microdata files. The rules proposed constitute a statistical way to avoid disclosure of confidential information. To protect the confidentiality of respondents further software can be implemented e.g. to the tabular data produced from the SES microdata.

2.6 Conclusions

The analyses in subsection 2.4 analyses show that an extremely high degree of anonymisation of SES microdata is achievable (virtually 100 % in the Italian, Dutch and UK studies). Similar analyses could usefully be done in other countries to see if similar results are obtained.

For business units, the three most important identifying variables are seen to be the region (NUTS 1), the size (in bands of employee numbers) and economic activity (NACE 2 digits). The analyses carried out show that the NUTS 1 variable is the most disclosive (in combination with the other two variables). A very high degree of anonymisation of the business units is achievable by simply dropping the regional variable. As a NUTS 1 breakdown is not required for the Candidate Countries and several EEA countries, anonymisation problems for these countries may be relatively slight. But this hypothesis needs to be tested.

A further minor improvement in anonymisation can be secured by additionally collapsing some of the size bands. But as the results above indicate, this seems hardly worthwhile, given the loss of some of the size bands. There is a 'trade-off' between complete anonymisation and the loss of important analytical possibilities.

It is important to stress that the loss of the regional variable in order to anonymise SES microdata does not mean that the NUTS 1 data is permanently sacrificed. SES tabular analyses can be produced which give a regional breakdown, but the simultaneous presence of other identifying variables in the tables (like size and economic activity) will need to be collapsed or removed if any of the cells are disclosive.

Anonymisation of the business units goes a very long way towards anonymising the employees sampled in those units (because the identifying variables for individuals are less likely to be disclosive in the absence of the firm's details). Subsection 2.5 contains the results of the work to anonymise the residual risks to individual employees. Different experiments were conducted on data of different countries. It became clear that if one variable (e.g. age) is collapsed in smaller bands, another variable (e.g. occupation) has to be sacrificed more. Nevertheless, three countries came up with very similar recoding strategies. When producing tables remaining risks have to be checked as is described in the next section. The better the data are protected with μ-ARGUS, the less work remains to be done with t-ARGUS when tabular data are released..

3. The Release of Tabular data

Many tables will be produced on the basis of the microdata of the European Structure of Earnings Survey. As these tables have to be protected against the risk of disclosure, the software package τ-ARGUS will be applied. Two common strategies to protect against the risk of disclosure are table redesign and the suppression of individual values. It is necessary to suppress cell values in the tables because publication of (good approximations of) these values may lead to disclosure. These suppressions are called primary suppressions.

A simple example explains how one unsafe cell in a table could be protected by putting extra crosses in such a table. Suppose that the cell with the value 5 in Table 4 is unsafe (i.e. individual information could be disclosed from this cell value) and has to be suppressed.

Table 4. *Number of employees by Nace (two digit level) and Size of the enterprise (in bands of employees), original data.*

	10-49	50-249	250-	total
Nace 51	5	15	20	40
Nace 52	10	20	30	60
Nace 53	15	25	40	80
total	30	60	90	180

One suppression would lead to recalculations of the suppressed value from the totals and therefore in each row and column where a suppression exists at least one more cell value has to be suppressed. A possible suppression pattern is given in Table 5.

Table 5. *Number of employees by Nace (two digit level) and Size of the enterprise (in bands of employees), protected data.*

	10-49	50-249	250-	total
Nace 51	x	x	20	40
Nace 52	x	x	30	60
Nace 53	15	25	40	80
Total	30	60	90	180

One has to realise that all cell values are non-negative and can logically be not larger than the total minus the published cell values in the same row and column. This implies that a user who gets Table 5 can recalculate intervals for the cell values that have been changed into crosses. This will result in the intervals given in Table 6.

Table 6. *Number of employees by Nace (two digit level) and Size of the enterprise (in bands of employees), recalculated intervals for the protected cell values.*

	10-49	50-249	250-	total
Nace 51	$0 \le x \le 15$	$5 \le x \le 20$	20	40
Nace 52	$0 \le x \le 15$	$15 \le x \le 30$	30	60
Nace 53	15	25	40	80
total	30	60	90	180

It has to be decided if the intervals of Table 6 are large enough to protect the unsafe cell value 5 properly. If this is the case the suppression pattern of Table 5 protects the tabular data well, otherwise another suppression pattern has to be chosen. A more drastic approach would be to redesign the table first and then to check again if no individual information may be disclosed.

A dominance rule is often used to decide which cells have to be suppressed. This rule states that a cell is unsafe for publication if the *n* major contributors to that cell are responsible for at least *p* percent of the total cell value. The idea behind this rule is that in unsafe cells the major contributors can determine with great precision the contribution of their competitors. In τ-ARGUS the default value for *n* is 3 and the default value for *p* is 70 %, but these values can be changed easily if the user of the package prefers other values. Using the chosen dominance rule τ-ARGUS shows the user which cells are unsafe. In publications, crosses (×) normally replace unsafe cell values. In version 2.1 of τ-ARGUS other rules have also been implemented that can be used to decide which cells have to be suppressed. Examples of these new rules are the p-percent rule and the pq rule. The p-percent rule can be considered as a special kind of pq rule. Both new rules take into account prior knowledge about respondent's values. The most widespread used technique to identify sensitive cells is the dominance rule. The situation in the practice of official statistics is that in many cases a simple version of the dominance rule is used: only the *n* is specified. Popular choices for *n* are 3 or 5.

As marginal totals are given as well as cell values, it is necessary to suppress further cells in order to ensure that the original suppressed cell values cannot be recalculated from the marginal totals. Even if it is not possible to recalculate the suppressed cell value exactly, it is often possible to calculate it within a sufficiently small interval. In practical situations every cell value is often non-negative and thus cannot exceed the marginal totals in the row or column. If the size of such an interval is small, then the suppressed cell can be estimated with great precision, which is of course undesirable. Therefore, it is necessary to suppress additional cells to ensure that the intervals are sufficiently large. A user has to indicate how large a sufficiently large interval should be. This interval is called the safety range and in τ-ARGUS the default safety range has a lower bound of 70 % and an upper bound of 130 % of the cell value. However, it is possible for a user to change these default values at will. A user of a table cannot see if a suppression is a primary or secondary suppression: normally all suppressed cells are indicated by crosses (×). Not revealing why a cell has been suppressed helps to prevent the disclosure of information.

Preferably, the secondary suppressions are executed in an optimal way. However, different definitions of optimal exist. Several measures for the loss of information can be defined and then the loss of information according to the measure chosen should be minimised. Three possibilities are:

- the minimisation of the number of secondary suppressions;
- the minimisation of the total of the suppressed values;
- the minimisation of the total number of individual contributions to the suppressed cells.

The minimisation of the number of secondary suppressions is often considered to be optimal. The other two possibilities (minimisation of the total of the suppressed values or the total number of individual contributions to the suppressed cells) are used less frequently. The minimisation of the total of the suppressed values is of course only relevant if all cell values are non-negative. In τ-ARGUS version 2.1, all three of these options have been implemented and thus the different resulting groups of secondary suppressions can be compared.

If the process of secondary suppressions is directly executed on the most detailed tables available, large numbers of local suppressions will often result. So it is better to try to combine categories of the spanning (explanatory) variables. A table redesigned by collapsing strata will have a diminished

number of rows or columns. If two safe cells are combined a safe cell will result. If two cells are combined when at least one is not safe it is impossible to say beforehand if the resulting cell will be safe or unsafe, but this can easily be checked afterwards by τ-ARGUS. However, the remaining cells with larger numbers of enterprises tend to protect the individual information better, which implies that the percentage of unsafe cells tends to diminish by collapsing strata. Thus, a practical strategy for the protection of a table is to start by combining rows or columns. This can be executed easily within τ-ARGUS. Small changes in the spanning variables can most easily be executed by manual editing in the recode box of τ-ARGUS, while large changes can be handled more efficiently in an externally produced recode file which can be imported into τ-ARGUS without any problem. After the completion of this redesign process, the local suppressions can be executed with τ-ARGUS given the parameters for *n*, *p* and the lower and upper bound of the safety range.

Normally, many tables are produced on the basis of a survey and the software package used for the data protection is based on individual tables. Although each table is safe, there is a risk that the combination of the data in these tables will disclose individual information. This may be the case when the tables have spanning and response variables in common. Version 2.1 of τ-ARGUS supports an important sub-class of linked tables, namely hierarchical tables. A hierarchical table is an ordinary table with marginals, but also with additional subtotals. Hierarchical tables imply much more complex optimisation problems to be solved than single tables. Some approximation methods exist for finding optimal solutions for these problems. The extended version 2.1 of τ-ARGUS was released in the CASC project.

4. Conclusions

The software packages μ-ARGUS and τ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth RTD Framework Programme of the European Union. These software packages appear to be of great help in the practice of statistical disclosure control. Many of the protection problems of statistical data can be solved using the ARGUS packages.

The twin ARGUS software will be applied to European Structure of Earnings Survey data. In particular, the application of μ-ARGUS to the microdata will lead to (largely) anonymised microdata that can be used for research under contract (for those countries where the law allows so). Experiments with data from Italy, the Netherlands and the UK identified that the collapsing of the regional variable (at the NUTS 1 level) was sufficient to anonymise the microdata to an extremely high degree. Nevertheless, some individuals could still be identified. In order to deal with the small number of disclosive cells that remained, τ-ARGUS will be applied to the (largely) anonymised microdata in order to remove any remaining risks of disclosure. In this paper, methods have been described that have been developed (and implemented in the ARGUS packages) to protect tables, through various means that either alter the data or restrict access to them.

Statistical disclosure control techniques have thus helped in keeping the right balance between data confidentiality and data access. Eurostat followed a very good strategy to plan the statistical protection of the European Structure of Earnings Survey data well in advance. This way modern technology could be applied and best practices could be exchanged in the Expert Group preparing the protection of the SES. The preparations will help managing the dissemination at the European level.

References

Hundepool, A.J., 2001. Computational aspects of statistical confidentiality: the CASC-project. In: Statistical Journal of the United Nations Economic Commission for Europe, Volume 18, Number 4, 2001, pp. 315-320.

Hundepool, A., A. van de Wetering, P.P. de Wolf, S. Giessing, M. Fischetti, J.J. Salazar and A. Caprara, 2002a. τ-ARGUS, user's manual, version 2.1.

Hundepool, A., A. van de Wetering, L. Franconi, A. Capobianchi and P.P. de Wolf, 2002b. μ-ARGUS, user's manual, version 3.1.

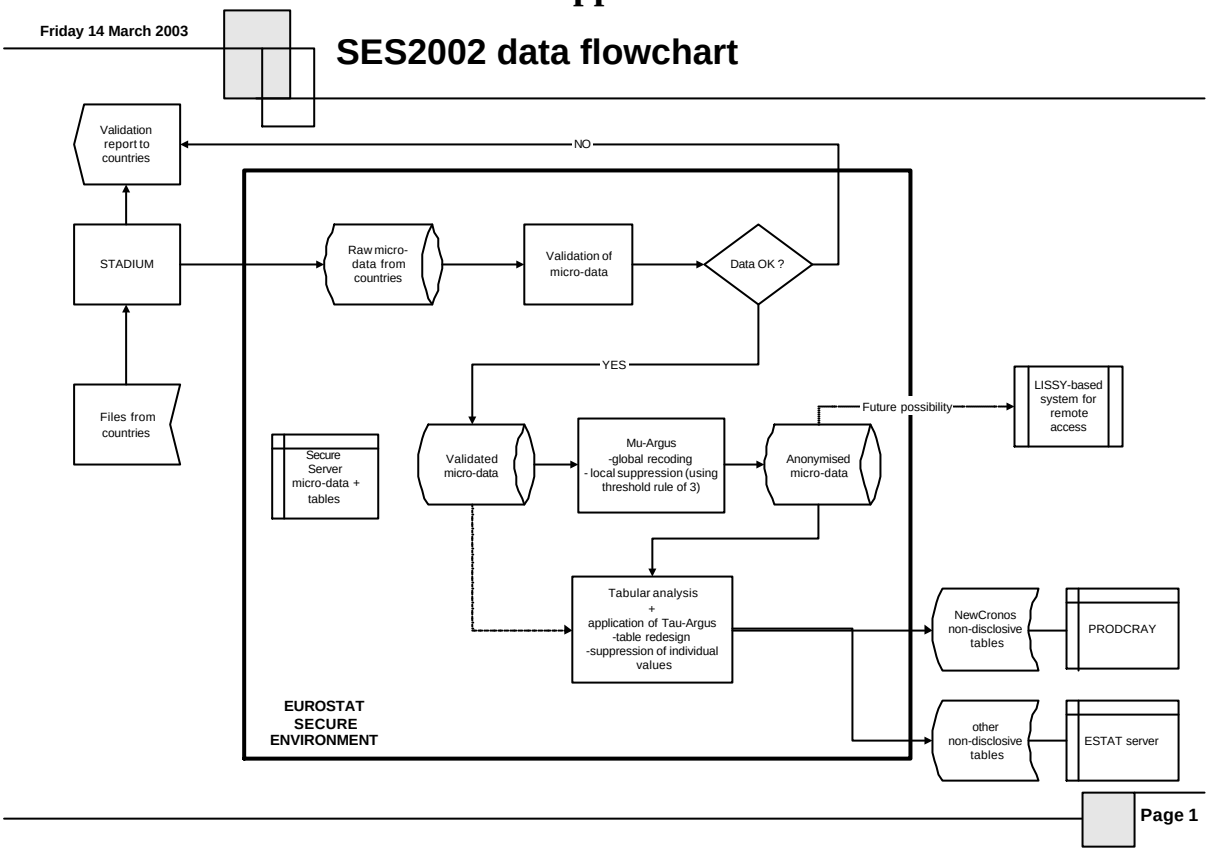
Schulte Nordholt, E., 1999. Statistical disclosure control of Statistics Netherlands employment and earnings data. In: Netherlands Official Statistics, Volume 14, spring 1999, Special issue on statistical disclosure control, pp. 34-38.

Schulte Nordholt, E., 2000. Statistical disclosure control of the Statistics Netherlands employment and earnings data. In: Statistical Data Confidentiality, Proceedings of the Joint Eurostat / UN-ECE Work session on Statistical Data Confidentiality held in Thessaloniki in March 1999, European Communities, 1999, pp. 3-13.

Willenborg, L.C.R.J. and T. de Waal, 1996. Statistical disclosure control in practice, Lecture Notes in Statistics 111 (Springer-Verlag, New York).

Willenborg, L.C.R.J. and T. de Waal, 2001. Elements of statistical disclosure control, Lecture Notes in Statistics 155 (Springer-Verlag, New York).

Appendix


Friday 14 March 2003
SES2002 data flowchart
Validation report to countries
NO
STADIUM
Raw micro-data from countries
Validation of micro-data
Data OK ?
YES
Files from countries
Future possibility
LISSY-based system for remote access
Secure Server micro-data + tables
Validated micro-data
Mu-Argus -global recoding - local suppression (using threshold rule of 3)
Anonymised micro-data
Tabular analysis + application of Tau-Argus -table redesign -suppression of individual values
NewCronos non-disclosive tables
PRODCRAY
other non-disclosive tables
ESTAT server
EUROSTAT SECURE ENVIRONMENT
Page 1