STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint ECE/Eurostat Work Session on Statistical Data Confidentiality
(Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001)

Working Paper No. 3
English only

Topic I: Application of statistical disclosure control methodology and software in business statistics and social and demographic statistics

STATISTICAL DISCLOSURE CONTROL (SDC) IN PRACTICE: SOME EXAMPLES IN OFFICIAL STATISTICS OF STATISTICS NETHERLANDS

Invited paper

Submitted by Statistics Netherlands[1]

Summary*: The paper describes how two related software packages can be applied for producing safe data. The package* t*-ARGUS is used for tabular data and its twin* m*-ARGUS for microdata. The main techniques used to protect sensitive information are global recoding and local suppression. Bona fide researchers who need more information have the possibility to visit Statistics Netherlands and work on-site in a secure area within Statistics Netherlands. Some examples are given of official statistics that have benefited form statistical disclosure control techniques.*

Keywords: *microdata,* m*-ARGUS, software, statistical disclosure control, tables,* t*-ARGUS*

I. INTRODUCTION

1. 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 instead of on sufficiently large groups of individuals 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. The 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).

2. This paper discusses the available tools to protect data and the option for researchers to work on-site at Statistics Netherlands. The tables produced by Statistics Netherlands on the basis of the microdata of surveys have to be protected against the risk of disclosure. Therefore, the software package τ-ARGUS (Hundepool et al, 1998a) can be applied on the tables produced. More information about τ-ARGUS and how this package can be applied are given in section II. Section III explains how microdata for research and public use microdata files can be produced using the software package μ-ARGUS (Hundepool et al, 1998b). The option for bona fide researchers to work on-site on richer microdata files is explained in section IV. Some examples of official statistics that have benefited from statistical disclosure control

[1] Prepared by Eric Schulte Nordholt. The views expressed in this paper are those of the author and do not necessarily reflect the policies of Statistics Netherlands.

techniques are given in section V. Finally, a discussion about the current state and some possible extensions for the ARGUS packages in section VI concludes this paper.

II. THE RELEASE OF TABLES WITH t-ARGUS

3. Many tables are produced on the basis of surveys. As these tables have to be protected against the risk of disclosure, the software package τ-ARGUS (Hundepool, 1998a) can 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 dominance rule is 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.

4. 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 to within a sufficiently small interval. In practical situations every cell value is namely 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.

5. Preferably the secondary suppressions are executed in an optimal way, however the definition of optimal is an interesting problem. Often, the minimisation of the number of secondary suppressions is considered to be optimal. Other possibilities are to minimise the total of the suppressed values or the total number of individual contributions to the suppressed cells. The minimisation of the total of the suppressed values is of course only relevant if all cell values are non-negative. In τ-ARGUS the option of minimising the total of the suppressed values has been implemented as the default. In τ-ARGUS version 2.0 it is also possible to minimise the total number of individual contributions to the suppressed cells. If that criterion is desired a so-called cost variable that is equal to 1 for every record has to be used to execute the secondary suppressions in τ-ARGUS version 2.0. However, the option of minimising the number of secondary suppressions itself has not yet been implemented. For future versions of τ-ARGUS, the aim is to implement more options so that the different resulting groups of secondary suppressions can be compared.

6. If the process of secondary suppressions is directly executed on the most detailed tables available, large numbers of local suppressions will often result. Therefore, 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.

7. As 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, there is the risk that although each table is safe, 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. The current version of τ-ARGUS does not support linked tables. Although it has an option to protect such tables, this is not warranted in the current version. However, the aim is to extend τ-ARGUS in such a way that it is able to deal with 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 extensions of τ-ARGUS will be implemented in new versions of the package that will be released in the CASC (Computational Aspects of Statistical Confidentiality) project. The CASC project is funded under the Fifth Framework Programme of the European Union.

III. THE RELEASE OF MICRODATA FOR RESEARCHERS AND PUBLIC USE MICRODATA FILES WITH μ-ARGUS

8. Many users of surveys are satisfied with the safe tables released by Statistics Netherlands, however some users require more information. For many surveys microdata for researchers are released. The software package μ-ARGUS (Hundepool et al, 1998b) is of help in producing these microdata for researchers. For the microdata for researchers Statistics Netherlands uses the following set of rules:

1) Direct identifiers should not be released.

2) The indirect identifiers are subdivided into extremely identifying variables, very identifying variables and identifying variables. Only direct regional variables are considered to be extremely identifying. Each combination of values of an extremely identifying variable, a very identifying variable and an identifying variable should occur at least 100 times in the population.

3) The maximum level of detail for occupation, firm and level of education is determined by the most detailed direct regional variable. This rule does not replace rule 2, but is instead an extension of that rule.

4) A region that can be distinguished in the microdata should contain at least 10 000 inhabitants.

5) If the microdata concern panel data direct regional data should not be released. This rule prevents the disclosure of individual information by using the panel character of the microdata.

9. In the case of most Statistics Netherlands' business statistics the responding enterprises are obliged by a law on official statistics to provide their data to Statistics Netherlands. This law dates back to 1936 and was renewed in 1996 without changing the obligation of enterprises to respond. No individual information may be disclosed when the results of these business surveys are published. The law rules that no microdata for research may be released from these surveys. Statistics Netherlands can therefore provide two kinds of information from these surveys: tables and public use microdata files. Public use microdata files contain much less detailed information than microdata for research. The software package μ-ARGUS (Hundepool et al, 1998b) is also of help in producing public use microdata files. For the public use microdata files Statistics Netherlands uses the following set of rules:

1) The microdata must be at least one year old before they may be released.

2) Direct identifiers should not be released. Also direct regional variables, nationality, country of birth and ethnicity should not be released.

3) Only one kind of indirect regional variables (e.g. the size class of the place of residence) may be released. The combinations of values of the indirect regional variables should be sufficiently scattered, i.e. each area that can be distinguished should contain at least 200 000 persons in the target population and, moreover, should consist of municipalities from at least six of the twelve provinces in the Netherlands. The number of inhabitants of a municipality in an area that can be distinguished should be less than 50 % of the total number of inhabitants in that area.

4) The number of identifying variables in the microdata is at most 15.

5) Sensitive variables should not be released.

6) It should be impossible to derive additional identifying information from the sampling weights.

7) At least 200 000 persons in the population should score on each value of an identifying variable.

8) At least 1 000 persons in the population should score on each value of the crossing of two identifying variables.

9) For each household from which more than one person participated in the survey we demand that the total number of households that correspond to any particular combination of values of household variables is at least five in the microdata.

10) The records of the microdata should be released in random order.

10. According to this set of rules the public use files are protected much more severely than the microdata for research. Note that for the microdata for research it is necessary to check certain trivariate combinations of values of identifying variables and for the public use files it is sufficient to check bivariate combinations. However, for public use files it is not allowed to release direct regional variables. When no direct regional variable is released in a microdata set for research, then only some bivariate combinations of values of identifying variables should be checked according to the statistical disclosure control rules. For the corresponding public use files all the bivariate combinations of values of identifying variables should be checked.

11. The software package μ-ARGUS is of help to identify and protect the unsafe combinations in the desired microdata file. Thus rule 2 for the microdata for researchers and the rules 7 and 8 for the public use microdata files can be checked with μ-ARGUS. Global recoding and local suppression are two data protection techniques used to produce safe microdata files. In the case of global recoding several categories of an identifying variable are collapsed into a single one. This technique is applied to the entire data set, not only to the unsafe part of the set, so that a uniform categorisation of each identifying variable is obtained.

12. If a certain identifying variable is desired in many categories, it means that other identifying variables can have fewer categories. Ideally, all identifying variables would have so few categories that no more unsafe combinations in the microdata would exist and local suppressions would not be necessary. When local suppression is applied, one or more values in an unsafe combination are suppressed, i.e. replaced by a missing value. These missing values could be imputed, but this is normally not attempted as bad imputations give misleading information to users and good imputations could lead to disclosure of the individual information of respondents. Local suppressions thus limit the possibilities of analysis, as there are no longer rectangular data files to analyse. However in practice, when producing protected microdata (microdata for researchers or public use microdata files) it is hard to limit the level of detail in the identifying variables and one often needs some local suppressions to meet the data protection criteria. Therefore, after the recoding of the identifying variables interactively with μ-ARGUS the remaining unsafe combinations have to be protected by the suppression of some of the values. The software package μ-ARGUS automatically and optimally determines the necessary local suppressions, i.e. the number of values that have to be suppressed is minimised. In this way it is possible to quickly produce microdata for researchers and public use microdata files.

13. Small changes in the identifying variables can be executed most easily by the manual editing in the recode box of μ-ARGUS, while large changes can be handled more efficiently in an externally produced recode file which can be imported into μ-ARGUS without any problem. After this global recoding the remaining unsafe combinations will be suppressed by μ-ARGUS to obtain protected microdata. No other protected microdata may be produced from the same data set, as the data protection measures could be circumvented by combining information. Therefore, before releasing protected microdata one has to plan carefully which variables to include in these files and how to recode the identifying variables included in the file. One can produce such a file only once.

14. In the field of microdata several new techniques will be investigated in the CASC (Computational Aspects of Statistical Confidentiality) project. The CASC project is funded under the Fifth Framework Programme of the European Union. New methodologies like post randomisation (PRAM), micro-aggregation and noise-addition will be implemented in new versions of μ-ARGUS that will be released in the near future. This will allow for experimenting with these techniques. To measure the quality of the methods applied, disclosure risk and information loss models will be implemented too.

IV. WORKING ON-SITE IN A SECURE AREA WITHIN STATISTICS NETHERLANDS

15. Some researchers need more information than is available in the released microdata for researchers or public use microdata file. As the releasing of richer data is not allowed, it is then possible for individual researchers to perform their research on richer microdata on the premises of Statistics Netherlands. Bona fide researchers have the opportunity to work on-site in a secure area within Statistics Netherlands. Researchers can at will choose between the two locations of Statistics Netherlands: Voorburg in the west of the Netherlands and Heerlen in the south of the Netherlands. The possibility to export any information is however only possible with the permission of the responsible statistical officer. They can apply standard statistical software packages and also bring their own programmes. Like all employees of Statistics Netherlands, these people who work on-site have to swear an oath to the effect that they will not disclose the individual information of respondents (Kooiman, Nobel and Willenborg, 1999).

16. The researchers who work on-site on economic data have to take the rules of Statistics Netherlands' Centre for Research of Economic Microdata (CEREM) into account. The most important rules are:

- researchers must be associated with a recognised research institute (e.g. a university);
- there must be a research proposal that conforms to current scientific standards;
- the researcher and his superior have to sign a confidentiality warrant;
- the researcher obtains only access to the data needed for his project;
- the data do not contain information on names and addresses of the enterprises;
- data related to the two most recent years will not be supplied;
- it is forbidden to let data or not safeguarded intermediate results leave the premises of Statistics Netherlands;
- all prospective publications will be screened with respect to risk of disclosure;
- all publications will be in the public domain;
- a public register contains the researcher's name(s), the research project, the publication(s) and the databases provided.

17. The facility is not free of charge. As a rule the researcher has to pay the cost for the supply of the required data. In addition, there is a tariff for using the on-site facility.

V. EXAMPLES OF OFFICIAL STATISTICS

18. In September 2000 Statistics Netherlands introduced a new organisation structure. Most data are now produced in the Divisions of Business Statistics and Social and Spatial Statistics.

19. In the Division of Business Statistics the most important surveys are the Production Statistics. Lots of tables are produced on the basis of these surveys. It is not an easy task to develop a consistent protection strategy for these tables. Some specially developed modules are used to tackle this problem. The idea is to integrate some of these modules into τ-ARGUS so that many users can profit from them. Some bona fide researchers want to perform special research projects and work on-site in a secure area within Statistics Netherlands on the microdata of the Production Statistics. For some projects these

microdata have to be matched with other surveys. In those cases Statistics Netherlands does the matching and then the resulting data set without the direct identifiers can be analysed by the bona fide researchers.

20. In the Division of Social and Spatial Statistics many different smaller surveys are conducted. The biggest of these surveys is the Annual Survey on Employment and Earnings (ASEE). In Schulte Nordholt (2000) it is described how payroll data for the ASEE are collected. The ASEE data sets contain large numbers of records and a lot of information concerning earnings.

21. The problem is how to handle linked ASEE tables using the current version of τ-ARGUS. As all of the tables have to be protected against the risk of disclosure, the current version of τ-ARGUS is applied to three basic tables. These are far fewer tables than are published, however, many specific tables can be constructed from the protected basic tables which will automatically also be safe. What remains is how to simultaneously protect the different basic tables. As the problem of how to solve the suppression problem in an optimal way for two or more tables simultaneously is not warranted in the current version, it was necessary to find a practical protection strategy.

22. In practice, two complications make our data protection process for linked ASEE tables a bit more difficult. Firstly, it is not only cell values and totals that are published, but also many subtotals. Therefore, the process must be executed at the level of the basic subtable. Secondly, if there is a choice of where to put a secondary suppression cross it is considered to be superior practice to put it in a cell that was also suppressed the previous year. Otherwise, each year a basic subtable may be safe, but the combination of such tables from consecutive years could lead to the disclosure of individual information. Many cell values do not differ substantially from year to year and often the main contributors to these cells are the same, thus good estimates can be made for suppressed cell values if the same cell is not suppressed the year before or the year after.

23. Currently the ASEE data set contains about 50 % of all employees in the Netherlands. The challenge is to enlarge the number of records of earnings information to all employees in the Netherlands within the next few years. To reach this aim it will be of great help to use information from the Insured Persons Register, which contains a large number of records and in which the private sector is very well represented. A disadvantage of this register is that the number of variables is smaller than in the ASEE, but imputation techniques (see e.g. Schulte Nordholt, 1998) help to overcome this problem. Of course the enlarged survey gives new challenges in the field of statistical disclosure control.

24. Another interesting recent development is the production of matrices (aggregated microdata) that are published in Statline. Statline is a product of Statistics Netherlands to view the data on a user-friendly way and to give users the possibility to let them make their favourite tables. As users of Statline can produce any table from a matrix at will one must be careful what kind of information is included in these matrices. The number (of categories) of identifying variables per matrix can therefore only be limited and rounding is used to further protect the individual sensitive information. Such matrices are currently being produced for social security statistics, education statistics and labour statistics.

VI. DISCUSSION

25. The software packages τ-ARGUS and μ-ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth 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. A few of these problems were mentioned in this paper.

26. The manuals (Hundepool et al, 1998a and b) are of great help for the users of the ARGUS packages. However, there are always additional things to desire. In the case of τ-ARGUS it would be of great help if linked tables, and more in particular hierarchical tables, could be dealt with in a more automated way. This need has been recognised for some time; in fact a preliminary implementation of a linked table option is already available in the current version of τ-ARGUS. A dedicated computer program (using the optimisation DLL of τ-ARGUS) is being developed by the department of statistical methods of Statistics Netherlands to deal with hierarchical tables. More research is also needed into how

consecutive years of the same survey can be protected from disclosure. Finally, it would be good to have more options available on how to execute the secondary suppressions. In the case of μ-ARGUS, it is important to clarify in the package the difference between protecting microdata for research and protecting public use microdata files. As μ-ARGUS can be used with lots of different protection criteria, it is important to help the users to understand how different strategies can be executed using the package. Recently, research has been directed at a perturbation method by adding stochastic noise to microdata. It would be good to have an option in μ-ARGUS to perturb data as a protection technique.

27. It can be concluded that there is still a lot of research to be done in the field of statistical disclosure control. Hopefully, new versions of the ARGUS packages (that include results of the on-going research) will soon be released. The production of these new versions is part of the CASC (Computational Aspects of Statistical Confidentiality) project. The CASC project is funded under the Fifth Framework Programme of the European Union. To promote the results of the statistical projects under the Fourth Framework Programme of the European Union the AMRADS (Accompanying Measures in Research And Development in Statistics) project is also funded under the Fifth Framework Programme. Many courses and conferences will be organised, among other topics, about statistical disclosure control.

References

Hundepool, A.J., Willenborg, L.C.R.J., Van Gemerden, L., Wessels, A., Fischetti, M., Salazar, J.J. and Caprara, A. (1998a), *t-ARGUS, user's manual*, version 2.0.

Hundepool, A.J., Willenborg, L.C.R.J., Wessels, A., Van Gemerden, L., Tiourine, S. and Hurkens, C. (1998b), *m-ARGUS, user's manual*, version 3.0.

Kooiman, P., Nobel, J.R. and Willenborg, L.C.R.J. (1999), 'Statistical data protection at Statistics Netherlands' in *Netherlands Official Statistics*, Volume 14, spring 1999, pp. 21-25.

Schulte Nordholt, E. (1998), 'Imputation: methods, simulation experiments and practical examples' in *International Statistical Review*, Volume 66, Nr. 2, pp. 157-180.

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 De Waal, A.G. (1996), 'Statistical disclosure control in practice', *Lecture Notes in Statistics 111*, Springer-Verlag, New York.