

**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**  
(Thessaloniki, Greece, 8-10 March 1999)

Working Paper No. 2  
English only

Topic (i): new applications of disclosure control methods

**STATISTICAL DISCLOSURE CONTROL OF THE STATISTICS NETHERLANDS  
EMPLOYMENT AND EARNINGS DATA**

Submitted by Statistics Netherlands<sup>1</sup>

**Invited paper**

**I. INTRODUCTION**

1. Statistical offices collect large amounts of data for statistical purposes. Respondents are only willing to provide the statistical offices with the required information if they can be certain that these offices will disseminate their data with the utmost care. This implies that the confidentiality of the data must be assured. Therefore the amount of detail in the publications has to be limited. At the same time researchers are no longer satisfied with the tables provided by statistical offices. There is a growing demand for more detailed information as all researchers have the possibility to analyse data on personal computers. Statistical offices thus no longer have the sole responsibility of analysing data. Statistical disclosure control theory is used to solve the problem how to publish and release as much detail as possible without disclosing individual information.

2. In the case of the new Statistics Netherlands Annual Survey on Employment and Earnings (ASEE), the responding firms 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 for firms to respond in the ASEE. By publishing the results of this survey no individual information may be disclosed. As the ASEE is a business survey, by law no microdata for research may be released. Statistics Netherlands therefore provides two kinds of information from the ASEE: tables and a public use microdata file. Public use microdata files contain much less detailed information than microdata for research.

3. In section II the data of the ASEE are described. The tables produced by Statistics Netherlands on the basis of the microdata of the ASEE have to be protected against the risk of disclosure. Therefore the software package  $\tau$ -ARGUS (Hundepool et al, 1998a) is applied on three basic tables. More information about  $\tau$ -ARGUS and how this package was applied on the ASEE can be found in section III. In section IV it is explained how a public use microdata file has been produced using the software package  $\mu$ -ARGUS (Hundepool et al, 1998b). The current state and some possible extensions for the ARGUS packages are discussed in section V. The software packages  $\tau$ -ARGUS and  $\mu$ -ARGUS have

---

<sup>1</sup> Prepared by Eric Schulte Nordholt.

emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework of the European Union. Lots of ideas for the present report came from Willenborg (1993), Citteur and Willenborg (1993), Willenborg and De Waal (1996) and Groot and Citteur (1997).

## II. THE ASEE DATA

4. The Division Socio-economic Statistics of Statistics Netherlands recently created a big new Annual Survey on Employment and Earnings (ASEE). Most data are no longer collected by paper forms but by EDI (Electronic Data Interchange). At the moment the percentage of firms that responds electronically is still modest, although increasing rapidly. Moreover, mainly the large firms switch to EDI quickly, so the number of employee records that is sent electronically to Statistics Netherlands is substantial. From the firms that switched to EDI in principle no longer samples of employees are received, but tapes with all the employee records. More information about the changes in the data collection process of the ASEE can be found in (Arnoldus, 1997). We thus have much more earnings information than before. In 1995, the first reference year of the survey, the ASEE data set contains approximately 1 500 000 records with detailed earnings information. The challenge is to enlarge this number of records in a few years to all 6 000 000 employees in the Netherlands.

## III. THE RELEASE OF TABLES WITH $\tau$ -ARGUS

5. Many tables will be produced on the basis of the ASEE. As these tables have to be protected against the risk of disclosure, the software package  $\tau$ -Argus (Hundepool et al, 1998a) is applied. Two common strategies to protect against the risk of disclosure are table redesign and suppressing individual values. It is necessary to suppress cell values in the tables because publication of (good approximations of) these values can lead to disclosure. These suppressions are called primary suppressions. By means of the dominance rule it is decided 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 concurrents. In  $\tau$ -Argus the default value for  $n$  is 3 and the default value for  $p$  is 70 %, but these values can easily be changed if the user of the package prefers other values. Using the chosen dominance rule  $\tau$ -Argus shows the user which cells are unsafe. In publications unsafe cell values are normally replaced by crosses ( $\times$ ).

6. As apart from the cell values the marginal totals are given, it is necessary to suppress some more cells, because otherwise the suppressed cell values can be computed using the marginal totals. Even if it is not possible to recalculate the suppressed cell exactly, it is often possible to calculate an interval that contains the suppressed cell value. In practical situations like in the ASEE employees' tables 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 rather precisely and this is of course undesirable. Therefore, it is necessary to suppress additional cells to achieve 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  $\tau$ -Argus the default safety range has lower bound 70 % and upper bound 140 % of the cell value. A user can at will also change these default values in  $\tau$ -Argus. All extra suppressions are called secondary suppressions. A user of a table cannot see if a suppression is a primary or secondary suppression: normally all suppressed cells are indicated by crosses ( $\times$ ). Not showing why a cell has been suppressed helps preventing the disclosure of information.

7. Preferably the secondary suppressions are executed in an optimal way. An interesting problem is how to define optimal. An often chosen way of defining optimal is to minimise the number of secondary suppressions. 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 (as is the case with the ASEE employees' tables). In  $\tau$ -Argus the option of minimising the total of the suppressed values has been implemented as the default. It is also possible in  $\tau$ -Argus version 2.0 to suppress 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  $\tau$ -ARGUS version 2.0. However, the option of minimising the number of secondary suppressions itself is not yet implemented. For future versions of  $\tau$ -Argus it is the aim to implement more options so that the different resulting groups of secondary suppressions can be compared.

8. If the process of secondary suppressions were directly executed on the most detailed tables available, often large numbers of local suppressions would result. Therefore it is better to try to combine categories of the spanning (explanatory) variables. This table redesign by collapsing strata leads to 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 is safe or unsafe, but this can of course easily be checked afterwards. However, the remaining cells with larger numbers of firms tend to protect the individual information better. This implies that the percentage of unsafe cells tends to diminish by collapsing strata. Thus a practical strategy for the protection of the ASEE employees' tables is to start with combining rows or columns in the most detailed tables available. This can be executed within  $\tau$ -Argus easily. Small changes in the spanning variables can most easily be executed by editing manually in the recode box of  $\tau$ -Argus, while large changes can more efficiently be handled in an externally produced recode file which can be imported into  $\tau$ -Argus without any problem. After having finished this redesign process, the local suppressions can be executed with  $\tau$ -Argus given the parameters for  $n$ ,  $p$  and the lower and upper bound of the safety range.

9. To give an idea how the current version of  $\tau$ -Argus looks like, a window of this package is shown below in which the table 'number of employees by economic sector and firm size' is ready for processing using the default values for the parameters. As publishing the used parameters for the statistical disclosure control could help disclosing information, the parameters really used for this basic table are kept secret.

**Specify tables**

Explanatory variables: EconomicSector, FirmSize, Community

Cell item: Employees, EmployeesM, EmployeesF

Response variable: [ ]

Shadow variable: [ ]

Cost variable: [ ]

Dominance rule: Number: 3, Percentage: 70

Apply sample weights

Minimum number of records: 3

gro...	min #	#	%	resp var	shadow var	cost var	weight var	exp var 1	exp var 2	exp var 3	exp var 4
1	3	3	70	Employ...	Employees	Employ...		Economi...	FirmSize		

10. As many tables are produced on the basis of the ASEE microdata and the used software package for the statistical disclosure control is based on individual tables, we face the risk that all tables are safe, but that by combining tables unsafe cells result which disclose individual information. This can be the case when the tables have spanning and response variables in common. The current version of  $\tau$ -Argus is not able to deal with linked tables. However, the aim is to extend  $\tau$ -Argus in such a way that it is able to deal with an important sub-class of linked tables, namely hierarchical tables. Intuitively, a hierarchical table is an ordinary table with its marginals, but with additional subtotals. The case of hierarchical tables is an important stepping stone to the more general case of linked tables.

11. The problem is thus how we handle linked tables now. As all tables have to be protected against the risk of disclosure, the current version of  $\tau$ -Argus is applied to three basic tables. This number of basic tables is much smaller than the total number of tables that is published. Many specific tables can be constructed from the protected basic tables and will thus automatically be safe as well. What remains is how to protect the different basic tables simultaneously. As the problem how to solve the statistical disclosure control problem for two or more tables simultaneously in an optimal way has not yet been implemented, we had to find some practical protection strategy.

12. For the ASEE the following three basic employees' tables had to be protected: number of employees by economic sector and gender, number of employees by economic sector and region and number of employees by economic sector and firm size. The first basic table is constructed by combining economic sectors in such a way that no cell suppressions are necessary. The other two basic tables are protected by first calculating all primary suppressions. Then the number of primary suppressions in these two other basic tables has been enlarged by suppressing cells that would otherwise break the protection of the earlier protected basic table. Finally, safe specific tables are derived from the safe basic tables. This strategy does not always lead to the optimal statistical disclosure control strategy in the sense that the number of secondary suppressions is not necessarily minimised. However, it looks a reasonable approach that can be executed without too much trouble and leads to tables with only safe cells.

13. In practice two complications make our statistical disclosure control process of linked tables a bit more difficult. Firstly, not only cell values and totals are published, but also lots of subtotals. Therefore, the process must be executed at the level of the basic subtable. Secondly, if there is a choice where to put a secondary suppression cross it is considered to be better to put it in a cell that was suppressed last year as well. Otherwise, every year a basic subtable may be safe, but combining such tables of consecutive years could lead to disclosure of individual information. Lots of cell values namely do not differ very much from year to year and also the main contributors to those cells are often the same ones. Thus good estimates can be made for suppressed cell values if the same cell is not suppressed the year before or after.

14. Below, an example is shown of the results of the applied statistical disclosure control strategy to the basic subtables of the ASEE. In Table 1 codes and names are found of the economic sectors according to the NACE 1. These economic sectors are the row categories in Tables 2a and 2b where the safe table 'number of employees by economic sector and province' can be found. Tables 2a and 2b form one table but are split into two parts because of graphical reasons. This table comes from the basic table 'number of employees by economic sector and region'. In this safe table six crosses are placed to indicate the suppressions. The crosses that represent secondary suppressions are placed in such a way that the suppressed primary cell values cannot be recomputed accurately using the marginal totals. Accurately is here defined as within the chosen - but not published - safety range.

Table 1. Codes and names of the economic sectors according to the NACE 1 used in Tables 2a and 2b

Economic sector	
Code	Name
01-05	Agriculture and Fishing
10-14	Mining and quarrying
15-37	Manufacturing
40-41	Electricity, gas and water supply
45	Construction
50-52	Wholesale and retail trade; repair of motor vehicles, motor cycles and personal and household goods
55	Hotels and restaurants
60-64	Transport, storage and communication
65-67	Financial intermediation
70-74	Real estate, renting and business activities
75	Public administration and defence; compulsory social security
80	Education
85	Health and social work
90-93	Other community, social and personal service activities
01-93	All activities

Table 2a. First segment of a table with number of employees ( $\times 1000$ ) by economic sector and province, 31 December 1995

Economic sector	Province					
	Groningen	Friesland	Drenthe	Overijssel	Flevoland	Gelderland
Code						
01-05	1.7	4.5	2.5	3.6	2.1	9.2
10-14	×	0.2	1.9	0.4	×	0.4
15-37	33.7	31.8	27.9	78.5	8.1	115.6
40-41	×	1.5	0.8	3.0	×	4.6
45	10.8	12.8	9.8	27.0	3.8	40.8
50-52	25.8	28.1	22.0	58.8	14.9	109.3
55	4.2	5.3	4.2	10.3	2.1	18.3
60-64	12.5	9.3	5.7	19.6	2.8	32.3
65-67	3.6	8.2	3.0	7.4	1.3	20.1
70-74	28.5	22.8	17.6	44.8	12.0	84.5
75	13.8	12.5	10.7	20.8	6.1	38.9
80	19.0	13.7	8.0	28.3	5.7	45.7
85	30.2	28.2	21.5	48.8	8.0	91.4
90-93	6.3	7.0	5.0	11.6	2.7	23.7
01-93	193.5	185.8	140.5	362.8	69.9	634.8

Source: ASEE, 1995.

Table 2b. Second segment of a table with number of employees ( $\times 1000$ ) by economic sector and province, 31 December 1995

Economic sector	Province						The Netherlands
	Utrecht	North Holland	South Holland	Zeeland	North Brabant	Limburg	
Code							
01-05	3.1	11.3	26.1	1.9	11.8	6.5	84.3
10-14	×	1.2	2.2	0.0	0.5	0.7	9.2
15-37	46.3	118.1	143.6	23.2	190.8	92.9	910.5
40-41	×	7.4	8.8	1.6	5.4	3.9	42.3
45	26.4	47.1	78.5	7.5	56.1	20.9	341.4
50-52	85.6	178.1	217.4	19.4	148.7	57.4	965.5
55	14.1	39.3	33.0	5.0	25.1	14.1	175.0
60-64	30.3	90.1	100.3	7.2	43.4	21.0	374.4
65-67	24.7	52.3	47.8	2.8	22.9	10.7	204.7
70-74	76.3	155.8	191.6	12.2	103.7	45.1	794.9
75	27.4	65.2	106.3	8.7	40.2	22.3	372.8
80	34.8	59.8	81.6	7.2	52.7	24.3	380.8
85	67.6	127.2	158.9	17.0	101.7	54.1	754.5
90-93	20.2	51.5	47.3	3.7	26.0	11.6	216.7
01-93	459.9	1004.4	1243.4	117.5	829.1	385.3	5627.0

Source: ASEE, 1995.

15. The table 'number of employees by economic sector group and province' is another table of 'number of employees by economic sector and region'. The variable economic sector group is derived

from the variable economic sector by collapsing the 14 economic sectors into four groups (01-05, 10-45, 50-74 and 75-93). In the table ‘number of employees by economic sector group and province’ no cell suppressions were necessary. Therefore the secondary suppressions in Tables 2a and 2b are placed in such a way that the table ‘number of employees by economic sector group and province’ will not help disclosing these suppressions. As in Tables 2a and 2b only suppressions are found in economic sectors of the economic sector group 10-45, we actually applied the statistical disclosure control strategy to the subtable ‘number of employees by economic sector group 10-45 and province’.

#### IV. THE RELEASE OF A PUBLIC USE MICRODATA FILE WITH $\mu$ -ARGUS

16. Many users of the ASEE are satisfied with the released safe tables of Statistics Netherlands. However, some users need more information. As the possibility of microdata for research is by law not allowed for business surveys, a public use microdata file has been produced with the software package  $\mu$ -Argus (Hundepool et al, 1998b). Two criteria have to be fulfilled for this public use microdata file: every category of an identifying variable needs to occur frequently enough and every bivariate value combination of two identifying variables needs to occur frequently enough. These criteria depend on the number of individuals in the population with those characteristics. A common threshold value for the number of individuals in the population in a category of an identifying variable is 200 000; a common threshold value for the number of individuals in the population in a bivariate value combination of two identifying variables is 1000.

17. One of the most important users of the ASEE is the Dutch Central Planning Bureau (CPB). One of their tasks is to calculate the effects of proposed policy decisions. The CPB needs microdata of the ASEE to be able to quickly produce estimates of the effects of the proposed decisions. To calculate these estimates they only need a limited number of variables in a limited number of categories. Therefore, a public use microdata file is made with  $\mu$ -Argus to meet their basic needs. For special requests they sometimes need more information than is available in the public use microdata file. Then possibilities exist to work on richer microdata files in the Statistics Netherlands premises. Also other bona fide researchers have the possibility to work on-site on Statistics Netherlands microdata files. Just like all employees of Statistics Netherlands, all these people who work on-site have to swear an oath to the effect that they will not disclose individual information of respondents.

18. Data of the ASEE were matched with microdata from the Social Security Files and the Labour Force Survey (LFS). This matching was executed because not only earnings itself, but also the structure of earnings is a main target of analysis. The Social Security Files were taken into account to enlarge the matching probability with the LFS. The LFS happens to be the only one of these three sources that contains data on level of education and occupation. Thus, Statistics Netherlands got census like information without having to set up a separate and elaborate survey. Schulte Nordholt (1998) describes in more detail how these three sources were matched and how some of the remaining missing information was imputed. Actually, the produced public use microdata file does thus not only contain information from the ASEE, but also from the Social Security Files and from the LFS. However, the ASEE remains the main source for these earnings microdata.

19. The software package  $\mu$ -Argus is used to identify and protect the unsafe combinations in the wanted microdata file with information on earnings. Two statistical disclosure techniques to protect microdata files are global recoding and local suppression. In 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. This is done to obtain a uniform categorisation of each identifying variable. Examples of identifying variables in this data set are sex, economic sector and level of education. If an identifying variable is really desired in many categories, this implies that other identifying variables can have fewer categories. Ideally, all identifying variables have so few categories that no more unsafe combinations in the microdata exist and local suppressions are not 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 done as bad imputations give misleading information to users and good imputations could lead to disclosure of individual information of respondents. Local suppressions thus limit the analysis possibilities as no longer rectangular data files to analyse result. However, in the practice of producing a public use microdata file it is hard to limit the level of detail in the identifying variables and one will often need some local suppressions to meet the statistical disclosure control criteria for a public use microdata file. Therefore, after the recoding of the identifying variables interactively with  $\mu$ -Argus the remaining unsafe combinations had to be protected by suppressing some values. The software package  $\mu$ -Argus determined the necessary local suppressions automatically and optimally, i.e. the number of values that have to be suppressed is minimised. That way it was possible to quickly produce a public use microdata file.

20. To give an idea how the current version of  $\mu$ -Argus looks like, a window of this package is shown below in which the 'Table of Combinations with Unsafe Cells' is ready to help to decide the user which variables have to be recoded. In this example it is clear that the variable community leads to a lot of unsafe cells and therefore this variable is the first candidate to recode. Small changes in the identifying variables can most easily be executed by editing manually in the recode box of  $\mu$ -Argus, while large changes can more efficiently be handled in an externally produced recode file which can be imported into  $\mu$ -Argus without any problem. After this global recoding the remaining unsafe combinations will be suppressed by  $\mu$ -Argus to obtain a public use microdata file. No other public use microdata files may be made from the same data set as otherwise the statistical disclosure control measures could be circumvented by combining information. Before releasing a public use microdata file one has thus to think carefully which variables to include in this file and how to recode the identifying variables included in the file. One can namely produce such a file only once.

# unsafe cells	Var 1	Var 2
3	Community	
2	LevelOfEdu...	
3101	Community	EconomicS...
15	Community	Sex
3645	Community	LevelOfEdu...
4551	Community	Occupation
283	EconomicS...	LevelOfEdu...
456	EconomicS...	Occupation
3	Sex	LevelOfEdu...
2	Sex	Occupation
365	LevelOfEdu...	Occupation

## V. DISCUSSION

21. The software packages  $\tau$ -ARGUS and  $\mu$ -ARGUS have emerged from the Statistical Disclosure Control (SDC) project that was carried out under the Fourth Framework of the European Union. These software packages appeared to be of great help in the practice of statistical disclosure control. Lots of statistical disclosure control problems of socio-economic statistics can be solved using the ARGUS packages. A few of these problems concern employment and earnings data. In this paper it is described how these problems were solved.

22. The new manuals (Hundepool et al, 1998a and Hundepool et al, 1998b) will be of great help for the users of the ARGUS packages. However, there always remain things to desire. In the case of  $\tau$ -ARGUS it would be of great help if linked tables and more in particular hierarchical tables could be dealt with in a more automatic way. Also, more research is needed how consecutive years of the same survey can be protected against disclosure. Finally, it would be good to have more options available how to execute the secondary suppressions. In the case of  $\mu$ -ARGUS it is important to make the difference in the package clearer between protecting microdata for research and protecting public use microdata files. As  $\mu$ -ARGUS can be used with lots of different protection criteria, it is important to help the users how different strategies can be executed using this package. 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 be released in the years to come.

**References**

- Arnoldus, F., 1997. Electronic supply of data for labour statistics. In: Netherlands Official Statistics, Volume 12, autumn 1997, pp. 60-68.
- Citteur, C.A.W. and L.C.R.J. Willenborg, 1993. Public use microdata files: current practices at national statistical bureaus. In: Journal of Official Statistics, Volume 9, No. 4, 1993, pp. 783-794.
- Groot, A. and C.A.W. Citteur, 1997. Accessibility of business microdata. In: Netherlands Official Statistics, Volume 12, winter 1997, pp. 18-32.
- Schulte Nordholt, E., 1998. Imputation, the alternative for surveying earning patterns. In: Netherlands Official Statistics, Volume 13, spring 1998, pp. 14-15.
- Statistics Netherlands, 1998a.  $\tau$ -Argus, users manual, version 2.0. Authors: Anco Hundepool, Leon Willenborg, Lars van Gemerden, Agnes Wessels, Matteo Fischetti, Juan-José Salazar and Alberto Caprara.
- Statistics Netherlands, 1998b.  $\mu$ -Argus, users manual, version 3.0. Contributors: Anco Hundepool, Leon Willenborg, Agnes Wessels, Lars van Gemerden, Sergey Tiourine and Cor Hurkens.
- Willenborg, L.C.R.J., 1993. Discussion statistical disclosure limitation. In: Journal of Official Statistics, Volume 9, No. 2, 1993, pp. 469-474.
- Willenborg, L.C.R.J. and A.G. De Waal, 1996. Statistical disclosure control in practice, Lecture Notes in Statistics 111 (Springer-Verlag, New York).