WP. 36
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 UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (vi): Case studies

DISCLOSURE PROTECTION OF NON-NESTED LINKED TABLES IN BUSINESS STATISTICS

Supporting Paper

Prepared by Luigi Virgili and Luisa Franconi (Istat), Italy

Disclosure protection of non-nested linked tables in Business Statistics

Luigi Virgili and Luisa Franconi

Division for Information Technology and Methodology, Istat, Roma, via C. Balbo, 16 Italy, e-mail virgili@istat.it, franconi@istat.it .

Abstract: The study of globalised economy requires more and more complex way to aggregate information resulting in the production of non-nested classification systems. This is the case of European structural business statistics where units are aggregated according to different criteria. The aim of this paper is to present the rationale followed to disentangle non-nested hierarchies, reduce them to a nested case and set a general procedures that can be used by a standard software package like Argus to protect a set of non-nested hierarchical linked tables. The application to the set of tables stemming from Foreign Affiliates Trade Statistics supplied to Eurostat is presented.

1 Introduction

Business statistics most of the time involve the release of a set of linked hierarchical tables stemming from the classification of economic activities (in Europe NACE rev.2). In recent years we have experienced the production of more and more complex tables where, for example, to analyse the economy under different perspectives, units are grouped according to the Economic Activity variable following two different criteria: similarity of the product and/or the producing process (NACE criterion) and similarity of the technological level used. This is the case of the FATS (Foreign Affiliates Trade Statistics), where EU member states produce statistics about both enterprises resident in the country but controlled by foreign entities (*Inward* Fats) and member state controlled enterprises operating abroad (*Outward* Fats).

Grouping statistical units by different criteria on the same variable leads to the definition of different (non-nested) classifications in which categories of one do not correspond directly to the classes of the others. When, like in the Fats, more than one classification criterion is used it makes sense to speak of a classification system. Moreover, it is obvious that such classification system leads to a set of linked tables i.e. tables that contain the same responses classified by at least one common variable. If a non-nested classification is present the application of a standard software for disclosure protection requires the use of specific procedures. The aim of this paper is to present the rationale followed to disentangle non-nested hierarchies, reduce them to a nested case and set a general procedures that can be used by a standard software package like Argus to protect a set of non-nested hierarchical linked tables. τ-ARGUS (freely available at http://neon.vb.cbs.nl/casc/tau.htm) is a software program developed through a series of European projects (Giessing, 2001) designed to protect

statistical tables. It implements two algorithms that allow the protection of the tables: hypercube and modular; for more details see Hundepool *et al* (2009). To clarify the rationale we show its application to the Fats survey, with reference to inward Fats tables to be supplied to Eurostat.

The paper is organised as follows: section 2 defines non-nested classificatins. Section 3 analyses the whole protection process: the study of the classification system used in Fats, the definition of the release plan and disclosure scenario, the need for breaking down the tables in order to obtain a nested classification system and the definition of the protection sequence. Section 4 illustrates the application of the process explained above to the *Inward Fats* tables. Section 5 gives summary conclusions.

2 Non-nested hierarchical classification

The classifications required by the Fats Regulation are non-nested and hierarchical. A classification is called hierarchical when it splits the data along a tree structure that represents a hierarchy. The hierarchical levels correspond to different levels of detail and can be subtotals or, with respect to a tree structure, vertices (the distance between a vertex and the root defines the rank of the level). More details can be found in de Wolf (2007). The NACE classification which groups economic activities is an example of a hierarchical classification. We call a table hierarchical if at least one of its classifying variables is hierarchical.

A classification is called nested when its categories are mutually exclusive, that is a unit (or a hierarchical level) can only belong to one, and only one, category. More rigorously, with reference to a tree structure, in a hierarchical classification a child can only have one father (see Figure 1). For example, in the NACE classification a unit can only belong to one *class*, which can only belong to one *division*, and so on.

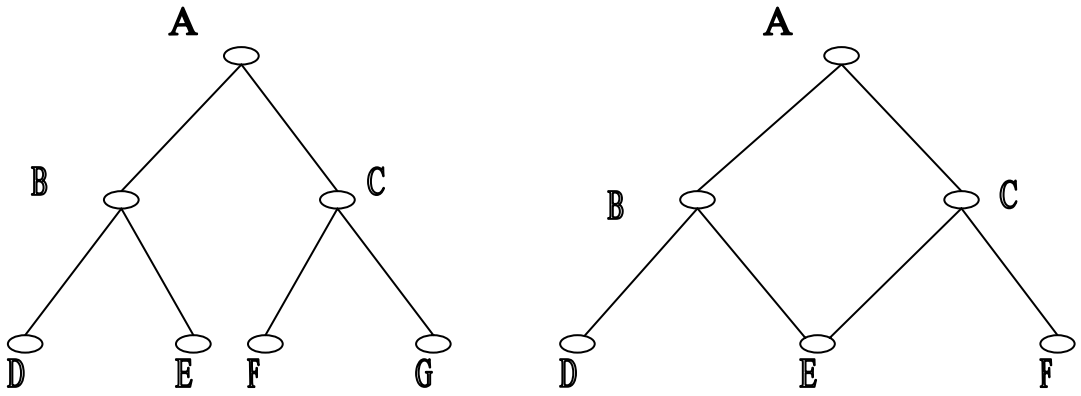

Fig 1 The diagram of a nested (left) and non-nested (right) hierarchical classification.

A classification is non-nested if its classes are not mutually exclusive. In this case the classes are *overlapping* and a unit (or hierarchy) can belong to more than one class (or higher hierarchical level). With reference to a tree, a classification is non-nested if a child can have more than one father (see Figure 1). The classifications used for business statistics are standard ones, rigorously defined, and, usually, do not include overlapping categories. However, in some cases it makes sense to group the statistical units along a variable, such as, Economic Activity, following a different classification criterion. Commonly, the classification criteria, with the exception of the most detailed categories (*classes*), cannot be put in direct correspondence with the NACE categories, hence there is *overlapping* between their hierarchical levels.

3 Rationale of data protection for a set of non-nested linked tables

To clearly address all the issues related to the protection of a set of linked tables a global study of such tables and their classification, the analysis of the release plan and hypothesis on the disclosure scenario are needed. This is presented in 3.1 and 3.2. Then, in order to protect a set of non-nested linked tables two problems need to be solved: the first one relates to the non-nested classification which needs to be reconducted to a nested one (see section 3.3) and the second is the protection of a set of linked tables by a standard software like τ-ARGUS: this implies passing protection information from one table to another using special features of the software (see section 3.4).

3.1 Analysis of the tables: the classification system in the Fats survey

Every year member states supplies Eurostat with two sets of tables (here after B1 and B2). In B1 the observed data are aggregated with respect to the two classifying variables economic activity and geography; in B2 the data are classified only by geography. The two series are *linked* because of geography.

The classification system underlying the Fats survey is non-nested hierarchical as both geography and economic activity are considered under a double homogeneity criterion. For geography the criteria are: (i) homogeneity as geographical vicinity and political affinity (EU25, EU27 etc.); (ii) economic and fiscal homogeneity (*offshore* area countries). For economic activity the criteria are: (i) homogeneity of the product and/or in the production process (NACE criterion); (ii) homogeneity of the technological level used in the production (using the NACE codes for *classes* and *groups*). The aggregates defined by this classification system overlap generating intersection sets among the information sets to be published. In this paper we focus our attention on those stemming from the economic activity. The complete description will be part of an Essnet on SDC case study (Virgili, 2009).

The criterion based on technological homogeneity defines aggregates HIT (High-technology), MHT (Medium-high-technology), MLT (Medium-low-technology) and

LOT (Low-technology), which are obtained aggregating specific NACE classes not nested with higher NACE hierarchical levels. For example, the aggregate HIT is defined, conformingly with the NACE classification, as the union of the NACE aggregates 24.4, 30, 32, 33 and 35.3, which are non-homogeneous in the detail level and belong to disjoint sets (*subsections*). It is evident that HIT is composed by aggregates that belong to different hierarchical levels. Furthermore, HIT is *transversal* to the NACE *sections* and *subsections*. In fact, while in the NACE classification the *subsections* DG, DM and DL have same rank and are disjoint, the aggregate HIT is composed of subsets of three *subsections* (DG, DM and DL). Figure 2 shows the corresponding tree-diagram completed with the levels (aggregates) with hierarchy higher than the Industry compartment (thus referred to the ObservedTotal, that is, the sum of Industry and Services inclusive of the NACE Section J), where X, 24b, and 35b are respectively the complements set of: DL, DM, DG to the Industry compartment; the union of the (24.1, 24.2, 24.3, 24.5, 24.6, 24.7) that is complementary of *group* 24.4 to *division* 24; the union of the (35.2, 35.4, 35.5) . BUS is the total of Business Economy (Industry + Service –J).

3.2 Release plan and disclosure scenario

Every time that a release of statistical information is evaluated in the light of disclosure protection it is necessary to consider two issues: the release plan, i.e. all correlated data previously released or that are planned to be released at a later time and the disclosure scenario. The evaluation of the release plan must be done regardless of the type of release (tables, graphics, datasets, etc). Moreover the evaluation should consider different release levels:

- **single release:** when several linked tables from the same survey are to be released, the protection of each table should take into account the protection of the tables linked to it;
- **subsequent releases of the same survey:** the release should be evaluated taking into account all the Institute's planned releases for such survey (national, Eurostat, OECD publications, web system, etc);
- **releases of different surveys containing correlated information**: all the Institute's planned releases concerning data correlated to those to be released, for example it is known that same statistics for Fats stems from the same units and may share the same variables of structural business statistics (Fats, SBS, etc);
- **correlated data released by other entities**: publications containing data correlated to those to be released (Central Banks, administrative archives, business demography, etc.)

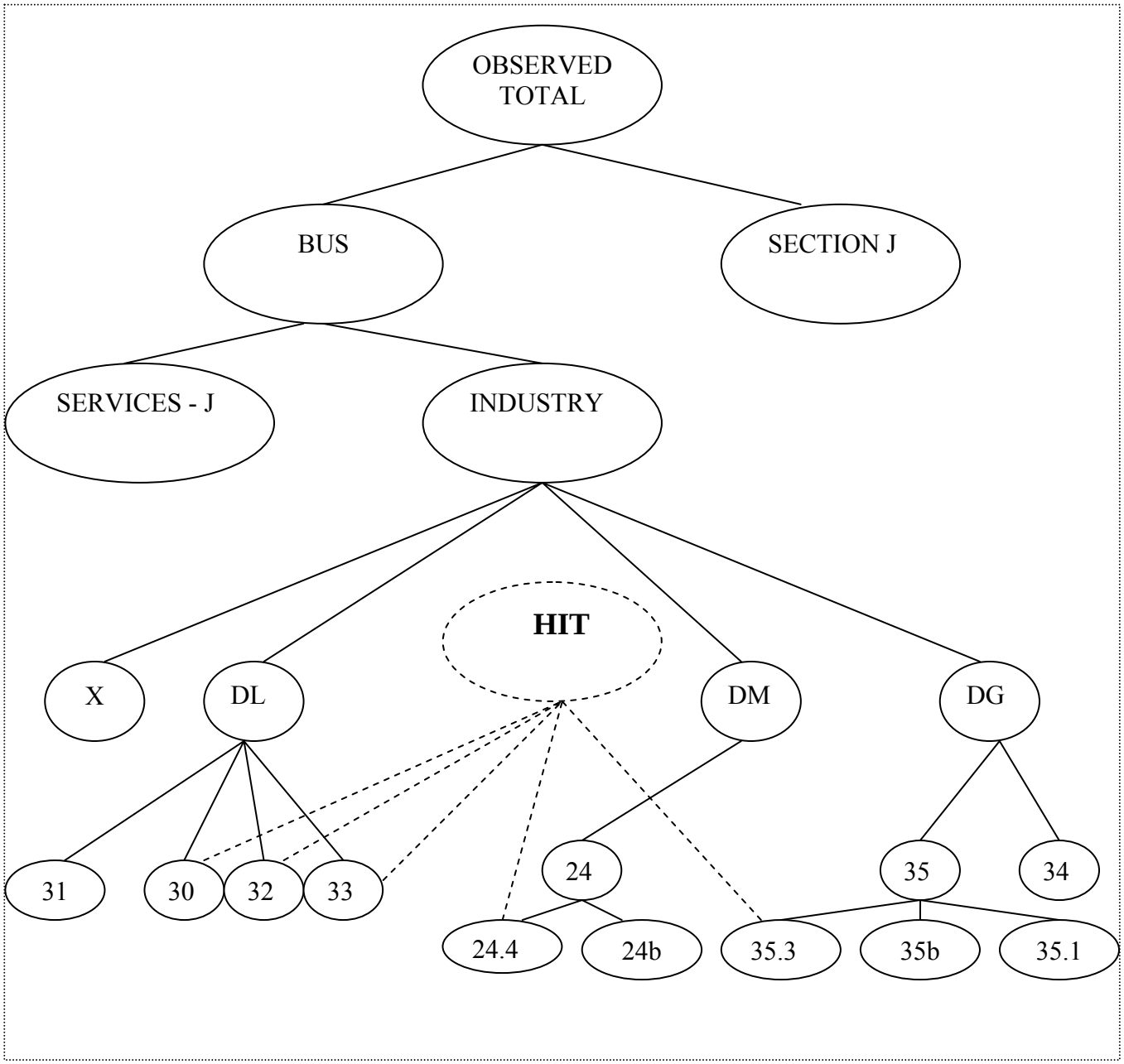

Fig 2 Tree-diagram for the whole ObservedTotal and the non-nested aggregate HIT in Fats.

Furthermore, while it is possible to suppress the data before they are released, past releases cannot be modified and constitute a constraint on the information to be released. This implies that, in order to have the most degrees of freedom (and, consequently, highest efficiency) for the protection of the information to be released,

the entire release plan, for as much as possible, should be considered since the beginning. The data released by other entities also cannot be modified and must be taken as constraints. So the judgement of the protection process should take into account since the beginning the release plan in its entirety and not only regarding the single survey data to be released.

As for the disclosure scenario this concerns the ability and possibility of the intruder to extrapolate new information from the data already released including also information released by entities other than the Institute. This last aspect requires an estimate of the resources that will be used to unveil the privacy of the released tables, and the hypothesis of what is actually usable In this work we have considered two different scenarios: the first relates to the publication of the Industry and Services aggregates, the second to availability of the ObservedTotal. Eurostat Fats publications do not give separate figures for the Industry and Services compartments. However, as already mentioned, every time that statistical data are protected it is necessary to consider the whole release plan as well as the existence of possible external sources. This first scenario is justified by the fact that most of Istat publications of business data (SBS, Fats, *etc*.) produce such totals. For the second scenario, we notice that the ObservedTotal is not released with the Fats data, however, it could be deduced, at least for some marginal cells (geo-economic areas), from other publications disseminated by other institutes. Given the choice of the safety rule and the parameters setting, the protection level of the released data is based on the intrusion scenario adopted and on the level of disclosure risk that one is willing to take.

3.3 Breakdown of non-nested tables into nested ones

This section discusses the breakdown of the Fats tables with respect to two connected aspects: the classification system from one side, the release plan and disclosure scenario on the other side. The former is necessary to get nested tables starting from non-nested ones and depends on the classification system used. The latter is partly arbitrary and depends also on the assumptions made on the disclosure scenario and the level of risk that one is willing to take.

3.3.1 Breakdown by classification

In Fats, the non-nested table B1 can be broken down by the variable economic activity into five nested linked tables: the base table, that groups the statistical units by the NACE classification, which is *non-homogeneous* in the levels, and other four tables, called *technological* tables. These latter tables are built using the *non-homogeneous* NACE classification to determine the technological aggregates. Their marginals are the technological levels to which they refer, HIT, MHT, etc. (see Fig 2) and some of their cells are also present in the base table. Therefore, the technological tables are linked and overlapping with the base table but they are not linked among themselves because they are defined on disjoint aggregates.

TABLE B1

BUS
Service – J
IND
C
D
DA
…
…
DL
31
30
32
33
DG
24
24b
24.4
DM
34
35
35.b
35.1
35.3
HIT…
…

BASE TABLE

BUS
Service - J
IND
C
D
DA
…
…
DL
31
30
32
33
DG
24
24b
24.4
DM
34
35
35.b
35.1
35.3

HIT TABLE

HIT
30
32
33
24.4
35.3

Fig 3 Classification of the spanning variable economic activity in table B1, base table and HIT table: the overlapping categories present in both base and HIT tables are marked in grey; the category HIT is present as subtotal in table B1 but all its components are split among more *subsections*.

3.3.2 Breakdown with respect to the intrusion scenario

As stated in section 3.2 we have considered two different scenarios: the first correspond to the availability of Industry and Service aggregates and the second correspond to the availability of the observation total. As for the first, the hypothesis that contributes to the definition of the table breakdown is that the values of the totals for the Industry and Services compartments are available for some or all the geographical categories considered. Operationally, the protection of the tables considering also such aggregates can be done by adding a hierarchical level for the Industry and Services subtotals.

As for the second scenario if the ObservedTotal is considered as released then the suppression of *Section* J implies that also the total BUS must be suppressed and vice versa (see Fig. 2). This is because each of these values can be obtained as the

difference between the ObservedTotal and the other. In this paper it is assumed that it is not possible to deduce the value of the observed total disaggregated by the geographical areas used in the Fats surveys but *a posteriori* checks were made to identify possible cells of the ObservedTotal which, if known, would allow the disclosure of the suppressed values of BUS and J.

3.4 The protection sequence and ranking criteria

In order to assure that a protected table cannot be unprotected by using information taken from a table linked to it, it is necessary to include in the protection process each and every table that is part of the release plan.
Given the complete set of tables, it is necessary to define an order of processing to protect the individual tables, and a tool to hold memory of the table to table protections realised. Each table is protected in the established order taking into account the suppressions previously determined on linked tables and the existence of constraints due to the intrusion scenario adopted. The tool to hold memory in τ-ARGUS is the history file that allows setting constraints on the data to be protected (see Statistics Netherlands, 2008). Using the history file it is possible to keep track of all the cells that have been suppressed (*secondary suppression*) and, also, of all the cells that have been deemed releasable (or *protected*); for more details see Capobianchi and Franconi (2009). The cells deemed releasable in previous protections cannot be suppressed in other tables and the cells suppressed must be constrained as non-releasable (*manually unsafe*), hence treated as if they were at risk (*primary*). In this way it is possible to protect a system of linked tables and in particular the system of table from the Fats survey.
The choice of the protection sequence is partly subjective and partly based on the structure of the tables to be protected. The general rule is to proceed from particular to general, that is, to start with the table that has the highest level of detail in the linking variable and continue in decreasing order of detail level. Hence the last table protected will be the one with the least detail. This rule, however, cannot always be followed. In fact, in several applications, like the Fats survey, the tables do not present a difference in the detail of the levels of the classifying variables. In particular, table overlapping denotes a partial equality of the cells and the same level of detail in the classifying variables. In this situation the choice of the protection sequence is up to the survey manager; in deciding such sequence it should be considered that the last tables will have, for the same number of cells at risk, a greater number of constraints and, therefore, a greater number of suppressions which results in a larger loss of information. In fact, the order (i.e. the position in the sequence) in with the table is processed has an effect on the total frequency of the suppressed cells and on the suppression *pattern*, that is, the distribution of the suppressions in the columns and the rows of the table being protected. In fact, the first table protected has only the constraints due to the intrusion scenario and the suppression pattern will be the minimal one determined by the algorithm. The second

table with is linked with the previous one, though, will also have the constraints deriving from the suppressions determined by the protection of the first; the third table will have the constraints deriving from the suppressions made on the first two, and so on. In general, the *n*-th table treated will have, beside possible *a-priori* constraints, also all the constraints due to the protection of the previous *n*-1 tables. In this work, in order to minimise the number of suppressions in the table with the largest information content, Table B1 was protected before Table B2 and the base table was protected before the technological tables.

4 The system of tables in this work

This section analyses the breakdown described above applied to the protection of the 2004 Fats tables supplied to Eurostat. The breakdown of Table B1 will be analysed in Paragraph 4.1; in paragraph 4.2 the structure of the series B2 will be analysed.

4.1 Table B1

Table B1 is the one that contains the most information since it classifies the units by the two variables geography (incomplete) and economic activity. Two classification criteria are used: a properly modified version of the NACE (non-homogeneous) for economic activity and a classification by geo-economic affinity for geography. The breakdown scheme that leads to the nested tables to which the protection algorithms can be applied is determined by the analysis of the classification system. In this work table B1 has been disaggregated with respect to economic activity into the compartments Industry, Services_NoJ (that is, excluding *Section* J) and *Section* J, which is financial intermediations. Each of these three tables has been further broken down with respect to geography, separating the table with the aggregate *offshore* (C4) from the table with the rest of the data (hereafter NOC4). The suffix "NOC4" will be added to the names of these tables to indicate that the geographical classification does not include the aggregate C4 which is non-nested with respect to the other aggregates (categories) of the strictly geographical classification. Furthermore, also the four tables obtained by disaggregating the technological aggregates with respect to C4 and NoC4 are considered. Finally, two more tables are created: ObservedTotal Table, which allows to relate the totals BUS, J and ObservedTotal (if known from other sources, this last aggregate would permit the disclosure of the protected tables) and Table 24_35, which helps keeping track of the protections made on *Divisions* 24 and 35 included in the technological aggregates MLT, MHT and HIT.

In summary, Table B1 is broken down into fourteen linked and overlapping nested tables as Industry and Service have been considered separated. For more detailed information on the composition of these tables see Essnet case of study (Virgili, 2009).

4.2 Table B2

Series B2 presents the Fats survey data classified by the geographical variable. This classification includes all the possible geographical areas and the overall total is BUS. The protection of this table does not present many problem having only the geography as classifying variable.

4.3 Protecting the system of tables and recovering tables B1 and B2

The fourteen nested tables derived from B1 and the tables from B2 have been assigned a rank according to the criterion described in section 3.4 and have been protected using τ-ARGUS in the defined sequence. The Modular algorithm has been used to protect all the nested hierarchical table; for more details see de Wolf, (2002)

Through the history file a coherent suppression pattern was imposed on all the tables. Finally the structure of the original tables B1 and B2 was recovered from the system of tables.

5 Conclusion

Standard SDC software are not able to deal with non-nested tables in an automatic way. In this work a general procedure allowing the protection of non-nested tables using τ-ARGUS is described. Such procedure breaks down the non-nested classification into several hierarchical nested tables. Every single table can be protected following an appropriate sequence. By means of the history file in τ-ARGUS it is possible to protect all the tables and maintain coherent protection among different tables. This general process has been applied to the Fats survey aggregates to be supplied to Eurostat. Criteria for ranking the sequence are discussed as well as the general rationale to take into account both the release plan and the disclosure scenario. Until now, the protection of the tables from Fats regulation has been carry out by Eurostat taking in account the indication of member states. Currently, the protection of all tables is a duty of each singular Statistical Institute. This means that individual institute have the direct control over the whole release plan and it would be easier for them to consider the links between the different releases of the same survey and between different linked survey.

To use τ-ARGUS it is required that the tables are additive and all the totals are considered. Therefore, if there is an incomplete set of the units that form the aggregate, then an artificial category needs to be created (for example in B1 we need to create the complementary to *principal countries).* All such artificial aggregates remain most of the time unpublished. It would be of great value the possibility of setting those cells in the table as outside of the release plan; this would mean that such cells will never be at risk and will always have cost equal to zero. This option could have been used for all those tables for which it is necessary to compute the

complement to the total, which is not published, because only part of the values that add up to the total are shown.

Acknowledgments

This work was partially supported by the European project "Essnet on Statistical Disclosure Control". The views expressed are those of the authors and do not represent the policies of Istat.

References

Capobianchi, A. and Franconi, L. Cell suppression in linked tables from structural business statistics using Tau Argus 3.3.0: a conceptual framework, New Techniques and Technologies for Statistics, Brussels, 2009, available at http://epp.eurostat.ec.europa.eu/portal/page/portal/research_methodology/documents/S18P1_CELL_SUPPRESSION_IN_LINKED_TABLES_CAPOBIANCHI_FRAN.pdf

de Wolf, P.P.: HiTaS: A heuristic approach to cell suppression in hierarchical tables. *Inference Control in Statistical Databases*: From Theory to Practice. (Ed.) J. Domingo-Ferrer Lecture Notes in Computer Science, Vol. 2316, 2002,

de Wolf, P.P.: Cell suppression in a special class of link tables. P*roceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality,* Manchester, United Kingdom 17-19 December 2007. http://www.unece.org/stats/documents/2007/12/confidentiality/wp.21.e.pdf .

Giessing S.: New tools for cell suppression in τ-ARGUS: one piece of the CASC project work draft. P*roceedings of the Joint UNECE/Eurostat work session on statistical data confidentiality,* Skopje, The former Yugoslav Republic of Macedonia, 14-16 March 2001. http://www.unece.org/stats/documents/2001/03/confidentiality/2.e.pdf .

Hundepool *et al*.: Handbook on Statistical Disclosure Control, version 1.1, January 2009. http://neon.vb.cbs.nl/casc/SDC_Handbook.pdf

Statistics Netherlands: τ-ARGUS *User's Manual,* Version 3.3, December 2008 http://neon.vb.cbs.nl/casc/Software/TauManualV3.3.pdf

Virgili, L.: Non nested classifications in τ-ARGUS. Essnet case study: available at http://neon.vb.cbs.nl/casc/handbook.htm#casestudies (*In preparation*)