Non nested classifications in τ-ARGUS

Luigi Virgili

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

Abstract: The aim of this report 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 τ-ARGUS to protect a set of non-nested hierarchical linked tables. The application to the set of tables stemming from Foreign Affiliates Trade Statistics *in-wards* supplied to Eurostat is presented. The analysis comprises both the non-nested hierarchies due to geography and economic classifications.

1 Introduction

The Regulation (CE) 716/2007 on Community statistics on the structure and activity of foreign affiliates *(Fats,* Foreign Affiliate Trade Statistics) requires EU member countries to produce statistics about both foreign affiliates resident in the country (*Inward Fats*) – that is, enterprises operating in the member state controlled by foreign entities - and not resident foreign affiliates (*Outward Fats*) – that is, member state controlled enterprises operating abroad.

The Regulation defines for both surveys which variables to include and the disaggregation level for geography and activity, as well as the frequency with which the data must be sent to the European Commission (Eurostat).

In Fats survey, in order to analyse the economy under different perspectives, units are grouped following different criteria. Such different criteria 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 report 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 τ-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 followed to treat non-nested linked tables we show its application to the Fats survey, with reference to *inward* Fats tables to be supplied to Eurostat.

The report is organised as follows: section 2 defines non-nested classifications. Section 3 analyses the whole breakdown 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. Section 4 describes the system of table in this work (Fats 2004). Section 5 analyses the protection sequence. Section 6 describes preparation of files for τ-ARGUS. Section 7 illustrates the results obtained from the application of the process explained above to the *Inward Fats* tables, year 2004. Section 8 compares results relative to different scenarios. Section 9 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.

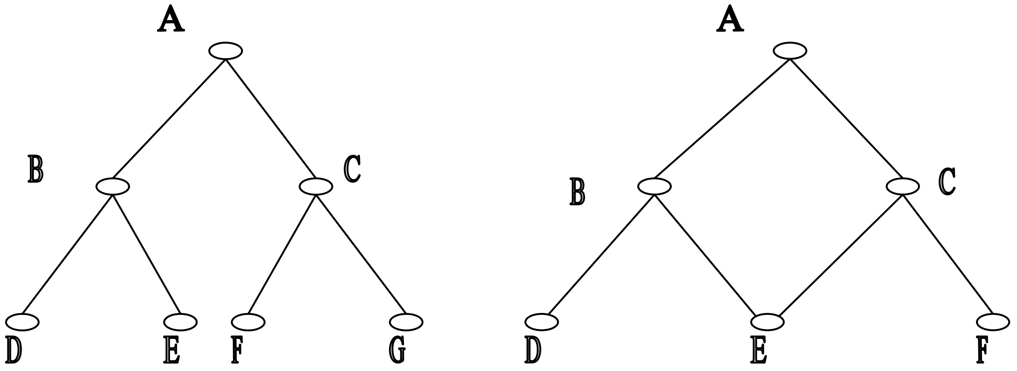

Fig 1 The diagram of a nested (left) and non-nested (right) hierarchical classification.

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 the breakdown

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 changed into 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 and it is based on two classifying variables: Geography and Activity, both considered under a double homogeneity criterion.

For Economic Activity the criteria are:

- homogeneity of the product and/or in the production process (NACE criterion);
- homogeneity of the technological level used in the production (using the NACE codes for *classes* and *groups*).

For Geography the criteria are:

- homogeneity meant as geographical vicinity and political and economic affinity;
- economic and fiscal homogeneity (offshore area countries).

The aggregates (hierarchical levels) defined by this classification system overlap generating intersection sets among the information sets to be published

3.1.1 Geography

The geography used to group the statistical units is hierarchical, the breakdown details is defined by Annex III of the Regulation. The main hierarchical levels (aggregates) defined by the first criterion are, in rank order: A1 (total of the countries included in the Regulation); A2 and Z9 (the compiling country and its complement to A1, respectively); D3 and D5 (the 25 European countries, excluding the compiling country, and ExtraEU25, the complement to Z9 with respect to D3, respectively). In this classification the set (A2, Z9) is a partition of A1 and the set (D3, D5) is a partition of Z9 (see Figure 2).

The same geographical classification, hierarchical aggregates, used for B1 (see Annex4) is used for B2 with the difference that in B2 all the elementary units (countries) are present; hence, each hierarchical level can be rebuilt as the sum of all the next lower levels. In a tree-diagram representation this means that for every *vertex* all the *children* are represented.

The decomposition of the aggregate D5 in Table B1 is not exhaustive. In a tree-diagram representation this means that only some of the *children* of the *vertex* D5 are represented (those defined as *main* in Annex4).

Thus the *Fats* geographical classification used for B1 is not complete: in B1 there exists an implicit partition of D5: the *principal countries* (hereafter *Principals*) individually released, and the complement (hereafter $Principals^c$) not published.

The second criterion used to group the category *countries* defines the aggregate offshore (named C4). This aggregate is a subset of D5 whose elements (countries) are both in the *Principals* and the $Principals^c$ aggregates. Therefore the set offshore is non-nested (or *transversal*) with respect to the partition *Principals* and $Principals^c$.

In a tree-diagram these last two sets represent the two fathers of C4, as shown in Figure 2, where: *Principals* are the main countries in CHECK and $Principals^c$ is the complement to D5 of the aggregate *Principals*; *others* is the set of the countries not in the offshore area which are in the *Principals* aggregate and $others^c$ is the set of the countries not in the offshore area which are in the $Principals^c$ aggregate.

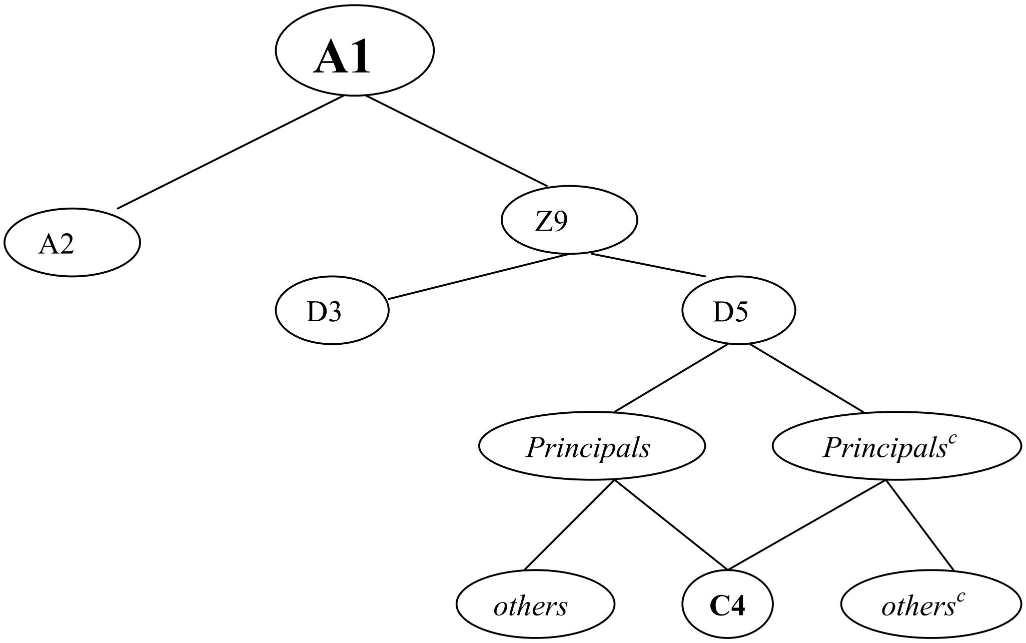

Fig 2 Tree-diagram for the *geo-economic* variable

3.1.2 Economic Activity

In the first criterion the classification by economic activity is identified by the non-homogeneous details of the NACE classification (Annex III of *Fats* Regulation). Details are called *non-homogeneous* because only some of the *divisions* are broken down to the *group* level; for example, the NACE *division* with code 40 is not further disaggregated but *division* 41 is broken down to the *group* detail level. Moreover, the Regulation defines *ad-hoc* categories obtained aggregating some *groups* within certain *divisions*. For example, *division* 35 is disaggregated into 351, 353 and 35b. Detail 35b, which is the sum of groups 352, 354 and 355, is not in the NACE classification but is an *ad-hoc* aggregate.

The second 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 3 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 is the complement set of the union (DL, DM, DG) to the Industry compartment; 24b results from the union of (24.1, 24.2, 24.3, 24.5, 24.6, 24.7) and is the complement of 24.4 to *division* 24; 35b results from the union of (35.2, 35.4, 35.5) and is the complement of the union of (35.3, 35.1) to *division* 35. BUS is the total of Business Economy (Industry + Services –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.)

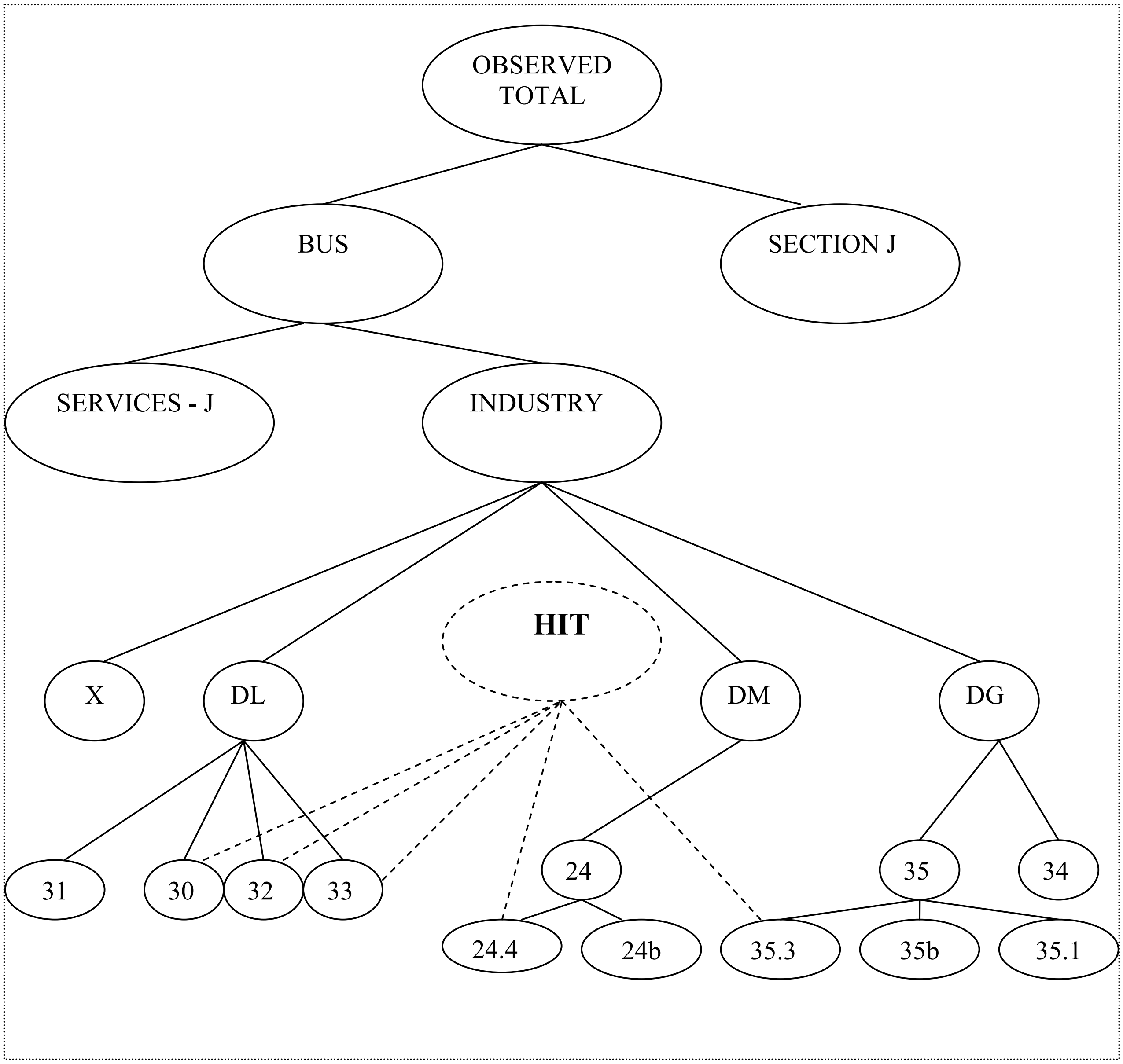

Fig 3 Tree-diagram for the whole ObservedTotal and the non-nested aggregate HIT in Fats in table B1

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.

The rationale of the procedure leads a non-nested classification system into a nested one is to transform each transversal aggregate (ex. HIT or C4) into several *ad hoc* tables each one referring to a different classification criterion: for example table B1 is split into two tables referring to the two aggregates BUS and HIT each one related to a different homogeneity criterion. This rationale is described in detail for both economic activity and geography in section 3.3.1

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. and some of their cells are also present in the base table (see Fig 3). 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 Breakdown of the spanning variable economic activity in table B1 into base table and HIT table. The overlapping categories present in both base and HIT tables are marked; the category HIT is present as subtotal in table B1 but all its components are split among more *subsections*.

The geographical classification used in the *Fats* tables for Eurostat is nested, with the exception of the aggregate offshore, transversal respect to the partition P*rincipals* and *Principals*^{c} (see Figure 2). To make this table nested it is enough to split the original table into two tables each one referring to a single homogeneous geographical criterion (see figure 4)

Original Table		**NoC4 Table**		**C4 Table**
D3 (Eu25 excluding A2)		D3 (Eu25 excluding A2)		C4 (offshore)
AT		AT		LI
BE		BE		HK
…		…		…
D5 (ExtraEU25)		D5 (ExtraEU25)		
USA		USA		
JP		JP		
…		…		
C4 (offshore)		LI		
LI		HK		
HK		…		
…				

Fig 4 Breakdown of the spanning variable geography into NoC4 and C4. The overlapping categories present in both NoC4 and C4 tables are marked; the category offshore is present as marginal (total) in table C4. The countries LI and HK (together with others not mentioned in B1) form the aggregate offshore.

Hence, it is required to identify two classifications for the elementary units *countries* with different hierarchical aggregates. One for the enterprises with headquarters in offshore countries (and marginal total equal to C4) and the other for all the countries (including those with headquarters in offshore) classified by the aggregates D3, D5, *etc*., as defined in the Regulation. Note that C4 is the marginal (total) in the last table.

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. 3). 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 ObservedTotal 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.

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. Paragraph 4.1 analyses the breakdown of Table B1; Paragraph 4.2 analyses the structure of the series B2.

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. Moreover, in order to take into account some previous protections and define the history files needed to complete the protection process of three of the four technological tables, two more tables were defined.

The (intuitive) names of the fourteen tables are: B1_ind_C4, B1_ind_NoC4, B1_serv_J_C4, B1_serv_J_NoC4; B1_serv_NoJ_C4; B1_serv_NoJ_NoC4; HIT, LOT; MHT, MLT, HIT_NoC4, LOT_NoC4; MHT_NoC4 and MLT_NoC4. In the naming convention “ind” and “serv” refer to Industry and Services; “C4”*,* “NoC4”, as mentioned before, indicate whether the aggregate offshore is included or not; “J” and “NoJ” identify the table relative only to Section J, Financial Intermediations, and the table relative to Services with exclusion of *Section* J, respectively.

In the case under consideration, the overlapping concerns the countries in offshore, which are also in the D5 aggregate (and therefore are duplicated), and the categories relative to the technological that belong also to the categories of *Section* D, which is internal to the Industry compartment.

Note that the breakdown used in these work is not the only possible. For example it is possible to consider Industry and Services as hierarchical levels and not as different tables. In the Italian cases it is common to release these economic aggregates as different tables and that is why in these work they have been considered in separate tables.

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. In this table all the countries are considered. However, in order to take into account the link between the two series B1 and B2, the partition *Principal* and *Principals*^{c} should be included as hierarchical level in the protection of B2. In this way, also for B2 the aggregate offshore should be considered non-nested with respect to the partition *Principal* and *Principals*^{c}.

Therefore, the most general procedure, to protect Table B2 is breaking down B2 in two linked nested tables: one for the aggregate offshore (C4) and its components and the other for all the geo-economic categories without the aggregate C4.

5 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 tables with the largest information content, Table B1 was protected before Table B2 and the base table was protected before the technological tables.

The table BUS, which contains the published total, was protected last.

Lastly the protection algorithm was applied to Table ObservedTotal so that the values of the aggregates ObservedTotal, BUS and J could be related. This table was not protected because it is completely constrained with the exception of the marginal (ObservedTotal), which will not be released. The aim of determining the suppressions for this table is to identify those values that, if available, would allow invalidating the whole protection process.

The protection sequence with respect to Geography was chosen so as to protect first the tables without the aggregate offshore and then those for the offshore tables.

6 Preparation of the file for τ -ARGUS

We choose to provide as input to ARGUS tables and not microdata. In this Section the steps needed to use τ-ARGUS are presented, with special reference to the table squaring and to the files that are need for a correct use of the software (for a general explanation on the structure of the tables, the files used, parameter settings and other variables used see the Manual, Statistic Nederland, 2008).

6.1 Table coherence checks

Once the table to be published has been broken down into a set of non-nested linked tables to be protected, it is necessary to make the individual tables compatible with the chosen software package.

To use τ-ARGUS it is required that the tables are readable and compatible with some constraints and checks.

The first step is to check that all the totals and subtotals (that is the hierarchical levels) are compatible among themselves; that is, the values of each hierarchical level must be equal to the sum of the values of the hierarchical level immediately below. In our case, for example, D3 (EU25) and D5 (extraEU25) must add up to Z9 which, together with A2 (country running the survey), must add up to the total A1.

As specified in section 3.1.1, by the Regulation, aggregate D5 in Table B1 is decomposed in *principal countries (Principals)*, which is subset of the units that form the aggregate. Hence, the sum of the values of the *principal countries* does not equal the corresponding value of the aggregate D5 for any of the variables in the table. Therefore, an artificial category for the countries in D5 that are not in *principal countries* must be created. This artificial category corresponding to *Principals*^{c} in the Figure 2 , will be referred in the input tables to as D6.

In the same way, the artificial aggregate C5 was created as the difference between the value of the aggregate C4 (offshore) and the sum of the values of the countries LI and HK as in many cases more countries contribute to the total C4 (see figure 4)

6.2 Table structure and choice of the cost variable

τ-ARGUS requires that each table is properly structured. The classifying variables must appear in the first columns. For example, the two classifying variables of a bivariate table must be in the first two columns, following the Eurostat standard.

Next to the classifying variables must be entered the *cost* variable, which is used to determine the secondary suppressions (if missing it is set equal to the response). Last goes the absolute frequency, which is the input for the *sensitivity rule* used in this work for identifying the cells at risk (primary).

Information loss is minimised with respect to the *cost* variable. The *cost* variable, with respect to which the information loss is minimised, can be chosen by the survey manager on the basis of the knowledge of the phenomenon under investigation and used for the whole set of data to be released. In this case, a single *suppression pattern* will we determined for all the variables, risking, not only not to minimise the information loss for every variable, but, also, to suppress one or more structural zero cells, invalidating, in part or completely, the protection process. Alternatively, this problem can be avoided by setting for each table the cost equal to the response variable. In this way, a *suppression pattern* will be identified for each released variable, avoiding the risk of suppressing empty cells. However, in this case it is possible that highly correlated variables present different suppression patterns, permitting to obtain relatively precise estimates of the suppressed values. A possible alternative solution is to use the same cost variable for highly correlated variables.

In this application, the survey manager chose the variable *number of employees*, V16110 (see Regulation) as *cost* variable. This choice was made also to avoid problems connected with using a *cost* variable with negative values.

6.3 Metadata files for τ-ARGUS

τ-ARGUS requires that, together with the table to be protected, a metadata file is entered. This file has extension *.rda* and contains the information necessary for Tau to interpret the table.

Furthermore, when, like in this work, the data to be protected are organised hierarchically, the hierarchies are saved in a special file which must be indicated in the metadata file. In appendix 1 an example of metadata file is shown with the corresponding hierarchy file, the .hrc file.

The hierarchies can also be defined directly in the metadata file, without using another file. In this case, though, the hierarchical classification variable must follow the hierarchy's tree structure. For example, for the NACE Industry compartment it is possible to associate the hierarchical level to every detail to which it refers. So, the NACE "DB17" detail is referred to *Section* D, *Subsection* DB, *Division* 17. With such a structure it is possible to enter in the metadata file all the hierarchical levels, associating every *position* to its hierarchical level. In the example, the first *position* indicates the highest hierarchical level (first level); the first two *positions* indicate the intermediate hierarchical level, that is DB (second level); the first four positions indicate the lowest hierarchical level (third level). In the example mentioned above the *positions* will be: 1, 1, 2. Therefore in the metadata file corresponding to the variable Economic Activity there must be a string like: <HIERLEVELS> 1 1 2 0. Where, the final zero indicates that the variable has maximum length equal to four.

Defining the hierarchies in the metadata file is less labour intensive but it is not always possible. In fact, whenever the hierarchical variable does not have a tree structure it is necessary to use the *.hrc* file. For example, the *Fats* table for Eurostat with both the Industry and Services compartment would require the use of the *.hrc* file.

6.4 Parameter setting and algorithm used

Sensitivity rule

In τ-ARGUS there are several parameters that can be set through a *dialogue window*. One of these is the choice of the sensivity rule (or rules), that is, the adopted definition of *cell at risk*. In this work it was uniquely chosen the *frequency rule,* by which all the cells with absolute frequency less than a given threshold are at risk. For the Fats table the threshold was set equal to three.

Minimum frequency range

Another parameter that must be set in τ-ARGUS is the *minimum frequency range* (Statistics Netherlands, Dec. 2008), which is the minimum width of the *existence intervals* that can be determined for the suppressed values from the released data. The *minimum frequency range* is the precision margin with which a suppressed value can be estimated from the released data.

Suppression Algorithm

τ-ARGUS allows choosing among different suppression methods (algorithms) through the *dialogue window*. However, as already said, only two of these algorithms have been tested and diffusely used: the Hypercube method, that implements the *hypercube* algorithm, and the method *modular*, that implements the HiTaS algorithm.

In this work it was used *modular* because it is more efficient than *hypercube* for hierarchical tables.

This module showed a limitation with some of the overlapping tables. In fact, the protection of such tables presents many constraints (cells set to *protected* or *unsafe*) deriving from the previous protection of overlapping tables. The limited number of degrees of freedom the algorithm HiTaS to find a wrong suppression *pattern* without issuing error messages. Therefore great care is needed when protecting tables that present a high percentage of cells constrained through the history file.

7 Evaluation of the result of the protection process

As different scenarios can be chosen the result of a protection process can be evaluated, for a given risk level, as a function of the information loss: the less information is lost the more efficient the protection process is.

The evaluation for tabular data can be either made with respect to either the total of the cost variable released (or suppressed) or the number of cells suppressed for every hierarchical level (combination of levels).

In the following the evaluation is made with respect to the number of suppressed cells.

At the end of every protection cycle τ-ARGUS produces report files that summarise the results of the protections made. For the *Fats* tables for Eurostat the reports concern the protection of the single tables created through the breakdown process explained in section 3. However care is needed in using this information as in these tables data that will not be released are also present. In particular, the artificial aggregates D6 and C5 are not published. Furthermore, some of the data are duplicated. These duplications derive from the breakdown process of the non-nested aggregates and will be eliminated when rebuilding the table to be released. So, for example, the data relative to countries LI and HK, which are both in D5 and in C4, will be included only once in the table to be released to Eurostat.

Therefore, for the *Fats* tables the evaluation of the protection process will be made on the reconstructed tables to be released to Eurostat evaluating the number of suppressed cells relatively to the hierarchal level they are in.

7.1 Table B1

The 2004 *Fats* tables for Eurostat have been protected using scenario 1 of section 3.2 and following the procedure described above. The results are given in summary tables separated for Industry, Services, J, BUS and Total (see Appendix 3). All the duplications and the aggregates that are not released have been removed (with the exception of the totals for Industry and Services). The total number of suppressions is equal to the sum of the suppressions for each table. However, from the separate tables it is possible to evaluate the information loss for the Industry and Services compartments separately. It is also possible to focus the loss of information relatively to the BUS aggregate, which is the highest hierarchical level. Where not explicitly specified, the hierarchical levels in the summary tables are reported in table 1.

7.1.1 Industry

The protection process for the Industry compartment of Table B1 alone determined 304 suppression on a total of 875 cells (about 35%) leaving 571 (about 65%) releasable; 209 of the 875 suppressions (about 69%) were cells at risk (primary) and 95 (about 31%) were secondary suppressions.

In the suppression pattern at the fourth hierarchical geographical level (maximal detail) there are 4 suppressions for the first economic activities (hereafter NaceFats) level (Industry Total) and 13 (8 primary and 5 secondary) for the second NaceFats hierarchical level (NACE *Sections*).

Two of the four suppressions in the first hierarchical level NaceFats are identified by the chosen risk rule (primary), the remaining 2 are secondary suppressions.

In the technological aggregates there are 2 suppressions at the second NaceFats hierarchical level, one of which is a cell at risk and the other is a secondary suppression.

Code of the hierarchical levels	EconomicActivity (NaceFats)	Geografy
1	Sector total (industry, Services)	A1 (total)
2	Section	A2, Z9
3	Subsection	D3, D5
4	Maximum detail	Single country
20	Non-nested technological aggregate	Non-nested area offshore

Tab 1 Code of the hierarchical level*s* used in summary report (Appendix 3) and comparisons of results (Appendix 4)

7.1.2 Services

The Services compartment of Table B1 has 1212 cells, 768 of which (about 63%) publishable and 444 (about 37%) suppressed. 298 (about 67%) of the suppressed cells are at risk, while 146 (about 33%) are secondary suppressions.

In the third geographical detail level and at the second *NaceFats* hierarchical level there are 4 suppressions with status 11 (secondary).

In the maximum geographical detail in the maximal NaceFats level there are 7 suppressions, of which 3 are *cells at risk* and 4 secondary suppressions; in the second NaceFats level there are 42 suppressions of which 28 (66%) at risk and 14 (33%) secondary.

There are three suppressions in the offshore aggregate at the second NaceFats level, one is at risk while the other 2 were identified by the protection algorithm.

7.1.3 Section J

The table for *Section* J has 22 cells, 17 (77%) of which releasable and 5 to be protected (23%). 4 of the 5 suppressed cells are at risk and only one was identified as a secondary suppression.

7.1.4 BUS Totals

The table with the BUS totals has 37 cells, 30 of which releasable and 7 suppressed. 4 of the 7 suppressed cells are at risk and 3 secondary suppressions.

7.1.5 ObservedTotal (BUS +J)

This table has 38 cells (in BUS one country is present only in section J and not in the *Industry* and *Services* compartments) of which 26 releasable; altogether there are 12 cells that should not be released: 5 determined by the risk rule, and 7 suppressed to protect the cells at risk.. All the suppressed cells are in the maximum geographical detail level.

As mentioned before, the protection of this table does not aim to suppress cells of the ObservedTotal. In fact, these cells cannot be suppressed because they refer to data that are not released. The protection of this table aims to identify those totals that, if available, would allow breaking the privacy of the tables to be released.

7.2 Table B2

The protection algorithm did not find any cells that needed suppressing to protect cells at risk (determined by the sensitivity rule or through the *history* file**).**

8 Comparison of the result under different scenarios

In this Section the results obtained with the package τ-ARGUS under two different hypotheses are analysed.

The Case 1 considers the Fats data for Eurostat not linkable to other publications that give also the values of the Industry and Services totals. Hence, these totals do not need to be inserted explicitly as hierarchical levels in the protection stage.

The Case 2 collects the number of suppressions determined by τ-ARGUS when the Industry and Service totals are included as hierarchical levels. The hypothesis is that all the survey results are protected together and that the values of the Industry and Services totals published are linkable to the *Fats* tables for Eurostat. The suppressions determined at Industry or Services total level must be considered a constraint on any other dissemination.

The choice of comparing these two situations spawns from the fact that, until now, the *Fats* data for Eurostat have never included the Industry and Services total as hierarchical levels. Furthermore, the hypothesis that the data for these hierarchical levels are available is the strongest possible with respect to the possible information loss. The objective of this comparison is to evaluate such information loss.

The comparisons are made on the number of suppressions per hierarchical level determined for the protection of Table B1 not including the Technological aggregates, *Section* J (financial intermediations) and aggregate C4.

The two protection processes were run on τ-Argus under the same parameter settings, which are given in Appendix 2.

The two tables in Appendix 4 summarise the number of suppressions per hierarchical level.

The first table refers to the protection process for the table in which the totals for Industry and Services are not separated. Hence the suppressions are determined without considering the hierarchical level for the two compartments (NaceFats 1). The second table summarises the results obtained under the hypothesis that the hierarchical levels Industry and Services are available. In this case, as mentioned above, the data for the compartments Industry and Services have been protected separately. In order to make this table comparable with the first one, the suppressions at Industry and Services totals have not been included.

In order to make the results homogeneous with those presented in Section 7, in the following tables, where not explicitly specified, the hierarchical levels are identified by the same levels as specified in table 1.

The number of cells in the two tables compared is the same (in fact the NaceFats equal 1 is not considered in both the tables), like the number of cells at risk identified by the adopted risk rule.

As expected from the results it can be seen that the number of suppressions in the second case, 204 (11.1%), is higher than for the first case 176 (9.5%).

The analysis of the suppression by hierarchical level shows that in the second case there are three suppressions at BUS Total level (equal to about 9,7% of the corresponding cells) and only one for the first case (3.2%); all the suppressions at Total level refer to countries.

The summary tables for the highest geographical detail present the following number of suppressions (in brackets the corresponding percentage): 19 (18.8%) versus 12 (8.1%) for *Section*, 45 (10.4%) versus 40 (9.2%) for *subsection*, 104 (14.2%) versus

93 (12.7%) for the highest detail of Economic Activity and 168 (12.8%) versus 145 (11 %) for the marginal total.

For the geography hierarchical level (hereafter GeoFats) equal 3 the suppressions are distributed very similarly among the levels of Economic Activity with the exception of the hierarchical level *section*. For this level, differently from what expected, the suppression are 4 (25%) for Case 1 and none for case 2. For GeoFats 2 at the maximal Economic activity detail level there are 6 suppressions (4.4%) in Case 1 and 8 (5.9%) in Case 2.

No suppressions are needed for the A1 hierarchical level. All in all the scenario that considers Industry and Services as separated has not such a great impact on the number of suppressions except for the case of the section.

An overall analysis of the loss of information with respect to these results can also be made evaluating if suppressing some of the cells, with particular reference to the highest levels (or combinations of levels), is acceptable and compatible with the lower identification risk associated with the protected tables.

9 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 data provider. 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 such end a comparison of different scenarios has been performed in order to provide tool to judge which scenario was more appropriate considering the information loss incurred.

Finally a possible extension to the software is suggested. 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 τ-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