Cell suppression in linked tables from structural business statistics using Tau Argus 3.3.0: a conceptual framework

Alessandra Capobianchi¹, Luisa Franconi²
Istat, Direzione Centrale per le Tecnologie e il Supporto Metodologico - PSM/C
Via Cesare Balbo 16, 00184 Rome - Italy
e-mail: ¹capobian@istat.it ²franconi@istat.it

Abstract

The conceptual scheme to protect via cell suppression the set of linked tables stemming from the Community structural business statistics is presented. Such scheme comprises of the translation of the legal framework into a set of tables, the analysis of the type of dependences between the hierarchical tables and their subsequent protection. In this last phase of the process a manual updating procedure has been developed to take into consideration the constraints in suppression due to the links present in the set of tables. The scheme has been used with the software Tau-Argus to produce the protected Community SBS tables.

Keywords: [Linked table, Tau-Argus, Structural business statistics]

1. Introduction

The current version of the software Tau-Argus (version 3.3.0) available at the website of the Essnet project http://neon.vb.cbs.nl/casc/..%5Ccasc%5Ctau.htm is under development. At the moment it does not allow a completely automatic protection of a set of hierarchical linked tables with the spanning variable present at different levels of the hierarchy in different tables as it is the case for the Structural Business Statistics. This possibility should be possible in the future when a more comprehensive approach, described in De Wolf and Giessing (2008), will be implemented into the software. Until that time a partially manual procedure has been developed at Istat to perform the protection of linked hierarchical tables with different hierarchical levels. The implemented procedure takes into account the current limitations of the software Argus but, at the same time, takes advantage of its strengths.

The conceptual scheme at the basis of the procedure described in this paper needs both the knowledge of the software Tau-Argus and, most essential, the understanding of all the possible constraints stemming from the protection of a set of linked hierarchical tables as complex as the Community Structural Business Statistics (SBS). Such conceptual scheme can be divided into three main parts. The first is the translation of the legal framework into a set of tables that can be processed by Argus; this first step requires careful study of the legislation and clear mind in table building and definition. This first part is described in Section 2 for the main tables concerning SBS. The second part of the scheme needs a careful analysis of the set of tables from the point of view of their links and inter dependences; the output of this second phase, described in Section 3, is the definition of the order of processing of the set of linked tables to be
protected by the software Argus. Finally, in the third step we manually use a strength of the software i.e. its capability of setting *a priori* information on protection status to constrain the protection process to maintain coherence between common cells in different tables. This final part is presented in Section 4. The conclusions are reported in Section 5.

2. Release of Community Structural Business Statistics: from the Legal Framework to a Set of Tables to be Processed

Community Structural Business Statistics (SBS) are collected within the framework of Council Regulation (EC, EURATOM) No. 58/97 of December 1996 and according to the definitions and breakdowns specified in a series of Commission and Council Regulations implementing it. Such Council Regulation establishes a common framework for the collection, compilation, transmission and evaluation of Community statistics on the structure, activity, competitiveness and performance of businesses in the European Union. The relevant Council Regulation applies to all market activities (except agriculture) normally included in industry, construction, the distributive trades and services.

The statistical units used for the compilation of structural business statistics are listed in Section I of the Annex to Council Regulation (EEC) No 696/93 on the statistical units for the observation and analysis of the production system in the European Community. The main statistical unit is the enterprise, the smallest combination of legal units that is an organisational unit producing goods or services and which benefits from a certain degree of autonomy in decision-making, especially for the allocation of its current resources. Some statistics on specific variables are to be produced also for KAU, the kind-of-activity unit that groups all the parts of an enterprise contributing to the performance of an activity at the class level (four digits) of NACE Rev. 1 and corresponds to one or more operational subdivisions of the enterprise. Finally, to be able to have a geographical distribution of the production some statistics are reported for local unit that is an enterprise or part thereof (e.g. a workshop, factory, warehouse, office, mine or depot) situated in a geographically identified place.

Council Regulation No 58/97 consists of one horizontal module (Annex 1), which includes a limited set of basic statistics covering all market activities and several sector-specific annexes, which include a more extensive list of variables describing the sector concerned. In this paper we focus our attention on the first four annexes covering the 'business economy' industry, construction and market services (NACE Sections C to K) not including the financial services (NACE Section J); see Table 1. The conceptual framework and reasoning used in this paper can be obviously applied to other set of linked tables as well.

<table>
<thead>
<tr>
<th>ANNEX</th>
<th>Annex NACE Sector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annex 1</td>
<td>horizontal module (C-F, G, H, I, K, new: 90, 921, 922)</td>
</tr>
<tr>
<td>Annex 2</td>
<td>industry (C-E)</td>
</tr>
<tr>
<td>Annex 3</td>
<td>distributive trades (G)</td>
</tr>
<tr>
<td>Annex 4</td>
<td>construction (F)</td>
</tr>
</tbody>
</table>
The data collection of SBS is carried out by the National Statistical Offices, and the aggregated data are transmitted to Eurostat, which takes on the work of calculating European totals and distributing tables through the SBS domain at Eurostat website. The SBS database contains different collections of tables which are loosely based on the structure of the legal text. The tables we focus on are those related to the non-financial business economy and, generally speaking relate to annual enterprise statistics, annual enterprise statistics broken down by size classes and annual regional statistics; see Table 2 for details of the classifying variables. In this table NACE is the Statistical Classification of Economic Activities in the European Community, Rev. 1.1 comprising four hierarchical levels and NUTS is the 2003 regional nomenclature of territorial statistical units comprising of four hierarchical levels. NUTS at level 2 corresponds, in Italy, to twenty-one regions.

Table 2: main classifying variables and response units for the collection of tables in Annexes 1-4 of the framework regulation.

<table>
<thead>
<tr>
<th>Classifying variables</th>
<th>Response Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>NACE 4</td>
<td>Enterprise</td>
</tr>
<tr>
<td>NACE3 and size class</td>
<td>Enterprise</td>
</tr>
<tr>
<td>NACE 2 and NUTS-2</td>
<td>Local unit</td>
</tr>
</tbody>
</table>

However, such general scheme comprising three types of tables presents some relevant differences according to the sector of reference:

- for sectors C-F the table at NACE 4 level is presented not only for enterprises but also for KAU;
- the variable size class presents two different classifications for sectors C-F and G-K;
- for sector G the regional table is released at NACE 3 level instead of NACE 2, moreover, only for this sector there is the additional table relating to NACE 3 by turnover in classes.

All these differences need to be taken into consideration when building tables for Tau-Argus as, for example, in the table relating to NACE 3, it is not possible to use the variable size class with two different definitions in the same table. Therefore the general scheme presented in Table 2 needs to be split into several tables that are homogeneous in the level of the classifying variables and homogeneous by response unit. In Table 3 we detail the correspondence between the tables we have created to be processed by Argus and the relevant Annexes. Table 3 reports also the names of the tables for the subsequent presentation in the paper.

This set of tables (from tab1.1 to tab3.3) for a series of response variables as defined in the legal framework constitutes the required set of tables to be sent to Eurostat to comply with the regulations. However, it might be possible that the National Statistical Office intends to publish for national release, on the same set of sectors, further tables with different levels of the classifying variables or with additional classifying variables. If this is the case then the same sort of reasoning should apply to such further statistics leading to further tables with homogenous level in the spanning
variables and homogeneous response units as those produced in Table 3. The definition of the set of tables to be processed is the preliminary step of the procedure. Once such set has been clearly defined we need to analyse the type of link existing between them and, as a consequence, an order of processing. These will be the theme of the next section.

Table 3: definition of spanning variables for each table to be processed by Argus in order to fulfil SBS regulations (Annexes 1-4). The response unit is the enterprise unless otherwise stated in brackets.

<table>
<thead>
<tr>
<th>Tables</th>
<th>Classifying variable</th>
<th>Annex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tab1.1</td>
<td>NACE 4</td>
<td>Annex 1A, Annex 2A, Annex 4A and Annex 3B</td>
</tr>
<tr>
<td>Tab1.2</td>
<td>NACE 4 (KAU)</td>
<td>Annex 2E and Annex 4E</td>
</tr>
<tr>
<td>Tab2.1</td>
<td>NACE 3 by size class1</td>
<td>Annex 2D and Annex 4D</td>
</tr>
<tr>
<td>Tab2.2</td>
<td>NACE 3 by size class2</td>
<td>Annex 3C and Annex 1B</td>
</tr>
<tr>
<td>Tab3.1</td>
<td>NACE 2 by NUTS-2 (local unit)</td>
<td>Annex 1C, Annex 2F and Annex 4F</td>
</tr>
<tr>
<td>Tab3.2</td>
<td>NACE 3 by NUTS-2 (local unit)</td>
<td>Annex 3E</td>
</tr>
<tr>
<td>Tab3.3</td>
<td>NACE 3 by Turnover in classes</td>
<td>Annex 3C</td>
</tr>
</tbody>
</table>

size class1 = 1-9, 10-19, 20-49, 50-249, 250 +
size class2 = 1, 2-9, 10-19, 20-49, 50-249, 250 +

3. The Analysis of the Set of Linked Tables Referring to SBS

As it is clear from Table 3 it is the hierarchical variable *main economic activity* as defined by the classification NACE rev. 1.1 that expresses the link between the set of tables stemming from SBS.

The current version of the software Tau-Argus still does not allow a completely automatic procedure to protect a set of tables linked via a hierarchical spanning variable present at different levels in different tables. This possibility should be possible in the future when a more comprehensive approach, described in De Wolf and Giessing (2008) will be implemented into the software. Until that time a partially manual procedure has been adopted at Istat to perform the protection of linked hierarchical tables with different hierarchical levels. In the next section we describe the analysis of the links between the tables and how such analysis will imply the definition of an order of processing.
3.1 Definition of the Order to Process the Set of Tables

In order to allow Argus to perform the suppression, for example by mean of the modular approach described in De Wolf (2002), care needs to be taken in understanding the link between the tables to be able to define the order in which the set of tables needs to be processed by the program. For each table that will undergo the protection we take notice of the hierarchical level of the corresponding NACE classification. Higher levels of the hierarchy (e.g. four digit NACE codes known as classes) correspond to more detailed breakdown of the information; lower level of the hierarchy (e.g. two digit NACE codes known as divisions) correspond to less detailed breakdown of the information. More detailed cells of the table will contribute to the construction of marginal cells in other tables; for such reason it is necessary to protect i.e. suppress such more detailed tables before the others in order to be able to take into consideration the suppression in the corresponding marginal cells.

The analysis of the levels of the hierarchy of the spanning variable defining the links between the tables implies the definition of a scheme of relationships that provides the order of the processing of the tables from the most detailed level of the hierarchy to the most aggregated. In the analysis of such scheme care needs to be taken on the response units as well as the sectors involved. The reasoning just described lead to the definition of the scheme of relationships for the of Community SBS pictured in Figure 1. We now described each relationship in such figure in turn.

The most detailed tables is Tab1.1 relative to all enterprises at national level classified according to classes of NACE (4 digit NACE codes) so this will be the starting table. Table 1.2 will be analyse later. The next table to be processed should present the hierarchical level immediately higher than the starting table. In SBS Community statistics there are two tables (Tab2.1 and Tab2.2) that present a 3 digit NACE code as classifying variable. As such tables relate to different sectors of the economy they indeed refer to different microdata subsets and therefore no link exist between the two but only between each one and the table 1.1. To analyse the next level of the hierarchy, the 2 digit NACE code, we need to split the problem according to homogeneous classes of employees and to different sectors. Tab3.1 presents marginal derived from tab2.1 and tab2.2 for sectors C to K excluding G even though the response unit is different. For sector G Tab3.2 is linked only to tab2.1 and 3.3 is linked to tab3.2. As for tab1.2 it differs from tab1.1 only for the response unit: the KAU instead of the enterprise. However, only in a very limited number of cases the number of KAU is slightly larger than the number of enterprises, otherwise the two tables coincide perfectly. For this reason it has been decided to apply Tab1.1 pattern of suppression to Tab1.2.

3.2 Use of the Ordering in the Protection Process

The protection process using secondary cell suppression starts from the most detailed table tab1.1 which is protected by Argus according to the rules established by the Member State. In order to communicate to Argus the information related to the protection of this starting table we ask the software to save the “status” information relative to each single cell of the protected tab1.1. Such information on the status will to be used to protect in a coherent way the common cells present both in tab1.1 and in
Such ad hoc history file will be given to Argus after reading tab2.1 so that the constraints derived from the protection of tab1.1 will be considered in the protection of the current table tab2.1.

This process of updating the history file with the new constraints derived from the protection of the previous table (according to the order established in Section 3.1) is the manual part of the procedure and it is carried out until the most aggregated table has been processed. This will be further described in Section 4.

4. The Protection of a Set of Linked Tables

The order generated by the analysis of the links between the tables as described in the previous section aims at identifying common cells in subsequent tables; such common cells need to present a coherent suppression pattern in the whole set of tables. In Section 4.1 we remind how a priori information related to protection is passed to Argus. In Section 4.2 we show how such flexibility of the software can be used to impose coherent suppression patterns to a set of tables.

4.1 A priori Information in Argus

Argus allows, before the protection of a table, the setting of a priori information (see section 3.2.1 of Argus manual) for some cells selected by the user of the software. This information is organised in an history file (a file .hst) that can be read by the software before the protection phase. In such file, for predetermined cells of the table, the user can assign one of the following protection “status”: “U” the cell has to be considered unsafe by Argus, “S” the cell is safe and “P” the cell is protected i.e. has been published. For each of these status Argus will perform an action in the protection phase of the software as detailed in Table 4.

<table>
<thead>
<tr>
<th>Alphanumeric code</th>
<th>Meaning</th>
<th>Action to be taken by Argus</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Unsafe</td>
<td>The cell has to be protected.</td>
</tr>
<tr>
<td>S</td>
<td>Safe</td>
<td>The cell is not at risk; it can be used as secondary suppression.</td>
</tr>
<tr>
<td>P</td>
<td>Protected</td>
<td>The cell cannot to be used as secondary suppression.</td>
</tr>
</tbody>
</table>

This way of passing a priori information to Tau-Argus before the protection of a table has been used in our procedure to build the necessary constraints typical of a set of linked tables. We now consider a certain common cell appearing in two subsequent tables of our ordered set of SBS tables. If such a cell has undertaken a secondary suppression in the first table then, by mean of an history file, we are able to pass this a priori information (manually suppress) to Argus before protecting the second table. In the same way, if a common cell has been released in the first table as safe, then, we can ask Argus to avoid using such a cell as secondary suppression in the second table by imposing the status P (protected) in the corresponding a priori information. All such status setting of common cells in successive tables of our ordered set of tables can be done in a structured way by mean of an external software that updates the values of the
common cells in the two tables. In the following section more details of such process of passing information between tables is presented.

4.2 Updating the **history** File

The protection process of the set of linked tables starts from tab1.1, the starting tables. Once a table has been protected by Tau-Argus, the software creates a .save file, containing all the information related to the table and its protection information. In particular, for each cell in the table, the following information are present: the cell identification, the cell value and the cell “status” (we call this the output status). Five different output status are currently allowed by Tau-Argus; they are presented in the first column of Table 5. A table is safe when the cells with output status 5, 9 and 11 are suppressed. The first table is now protected; we pass to the second table, tab2.1. For clarity of the process we call tab1.1 the previous table and tab2.1 the current table. The current table will present some marginal cells in common with the previous table just protected. This means that all the suppression applied to the previous table in such common marginal cells will have to be replicated in the current table. This will be done by creating an history file for the current table containing a priori information that will impose to Argus the constraints stemming from the protection of the previous table. Different types of constraints may arise for each of the common cell. The constraints may derive from observing that the common cell in the two tables has been already suppress in the previous table and therefore will need suppression also in the current table. However, a distinction need to be made between primary and secondary suppressed cells. Cells that are unsafe (i.e. primary suppression, output status 5) in the previous table will be recognised as primary suppression also in the current table; therefore there is no need to set any status as they will be recognised by Argus as unsafe. Common cells that are secondary suppressions in the previous table are safe cells in the current table so we need to set them to Manually unsafe (a priori information corresponding to status U) in the history file for the current table. In this way Argus will apply the suppression beforehand in coherence with the previous table. Common cells that present the status manually unsafe, (i.e. output status 9), in the previous table will remain unchanged as the cells will have to be suppressed also in the current table. Notice that all the common cells that have been chosen by Argus as secondary suppression in a certain table will be marked as manually unsafe for all the subsequent table where present. Using an external software we read the .save file of the previous table and select the common cells appearing also in the current table. Then we convert all the output status of such common cells of interest to us into the corresponding a priori information for the current table as shown in Table 5.

The process will continue following the order established in Section 3 until all the tables have been protected.
Table 5: Status transition table for the common cells in two subsequent tables. Output status in the previous table, meaning and corresponding status to be applied in the a priori information for the current table.

<table>
<thead>
<tr>
<th>Output status of the common cell in the previous table</th>
<th>Meaning</th>
<th>A priori information of the common cell of the current table</th>
<th>Action taken by Argus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Safe</td>
<td>P</td>
<td>This cell will not be selected as secondary suppression</td>
</tr>
<tr>
<td>5</td>
<td>Unsafe</td>
<td>—</td>
<td>Argus will recognise a primary suppression</td>
</tr>
<tr>
<td>11</td>
<td>Secondary suppression</td>
<td>U</td>
<td>Set to manually unsafe in the current table i.e. to be protected</td>
</tr>
<tr>
<td>14</td>
<td>Missing value</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>9</td>
<td>Manually unsafe</td>
<td>U</td>
<td>Set to manually unsafe in the current table i.e. cell to be protected</td>
</tr>
<tr>
<td>10</td>
<td>Manually safe</td>
<td>P</td>
<td>This cell will not be selected as secondary suppression</td>
</tr>
</tbody>
</table>

5. Conclusions and Further Work

The procedure defined in this paper has allowed the protection of the SBS tables described in Section 2. Notice that the described conceptual scheme does not depend on a particular type of setting of the software Argus but all the measures of risk and all possible parametrisation are indeed possible. The study of the link between the tables presented in Section 3 stemming from the analysis of the Annexes of the framework regulation is common to all Member States who need to protect their SBS to be sent to Eurostat. Also, the analysis done to identify common cells in subsequent tables in Section 4 will be useful for most of the Member States. What will change from one Member State to the other is the actual status of each of these common cells but the structured approach implemented will easily guide the changes.

In our experience there has been no need in the final stage of the protection to apply the backtracking procedure as defined in the Handbook on Statistical Disclosure Control (Hundepool, et al. 2009) but such a possibility needs to be taken into account.

With the entry into force of the new SBS regulation pertaining changes due to the adoption of the new classification of economic activities, NACE rev.2, more work need to be done. In fact, for the first year of release of the new NACE, some tables need to be released with the new as well as the old classification of economic activity. Such release will impose further studies of the link between the new and the old linking variable as well as an updating of the analysis presented in Section 3. Finally a protection pattern that is as much as possible harmonised between different years need to be carefully tuned so that coherence is maintained not only within a year (the set of linked tables) but also for successive years.
Acknowledgement
One of the author was partially supported by the European project Essnet on Statistical Disclosure Control, founded by Eurostat.

References

Figure 1: Analysis of the relationship between SBS tables

Tab1.1
Classifying variable: Nace 4 digit
Sectors C-K
Response unit: enterprise

Tab1.2
Classifying variable: Nace 4 digit
Sectors C-F
Response unit: KAU

Tab2.1
Classifying variables: Nace 3-digit
Employees in classes (1-9)
Sectors C-F
Response unit: enterprise

Tab2.2
Classifying variables: Nace 3-digit
Employees in classes (1; 2-9)
Sectors G-H-I-K
Response unit: enterprise

Tab3.1
Classifying var: Nace 2-digit
NUTS2
Sectors C-K excluding G
Response unit: local unit

Tab3.2
Classifying var: Nace 3-digit
NUTS2
Sector G
Response unit: local unit

Tab3.3
Classifying var: Nace 3-digit
Turnover in classes
Sector G
Response unit: