The dissemination process of the *Frame-SBS*: legislative and methodological aspects linked to increase information detail[1]

Carlo Boselli, Sabrina Brunetti, Mara Cammarrota, Viviana De Giorgi, Annamaria D'Urzo, Marco Ricci, Roberta Pazzini, Giovanni Seri, Giampiero Siesto, Luigi Virgili [2]

Abstract

The availability of administrative sources on enterprises and the results of direct surveys conducted by Istat allowed the Institute to obtain the microdata of the Frame-SBS register. It has been used to comply with the EU regulation on structural business statistics n. 295/2008 (SBS) and to improve the information released for national purposes. The paper describes the Regulations on data confidentiality, the methods applied to protect data, the IT aspects, and the channels used to disseminate the information.

Keywords: administrative data, structural business statistics, statistical data confidentiality, economic indicators.

[1] In this paper, the activities of the Istat task force: "Dissemination and confidentiality" are described. It was created with the task of improving the information released on structural business statistics (SBS), in compliance with the SBS Regulation. It also includes issues regarding statistical confidentiality. Although the article is the result of a joint work, Chapter 1 and Paragraph 6.2 has been drafted by Mara Cammarrota; Chapter 2 and Conclusions by Giampiero Siesto; Chapters 4 and Introduction by Luigi Virgili; Chapter 3 by Viviana De Giorgi; Introduction to Chapter 5 and Paragraph 5.1 by Annamaria D'Urzo; Introduction to Chapter 5 and Paragraph 5.2 by Sabrina Brunetti; Paragraph 5.3 by Marco Ricci; Paragraph 6.1 by Carlo Boselli; Paragraph 6.3 by Roberta Pazzini; Appendix A by Giovanni Seri;. The opinions expressed are those of the authors and do not reflect those of Istat.

[2] Carlo Boselli (cboselli@istat.it); Sabrina Brunetti (brunetti@istat.it); Mara Cammarrota (cammarro@istat.it); Viviana De Giorgi (degiorgi@istat.it); Annamaria D'Urzo (adurzo@istat.it); Marco Ricci (marricci@istat.it); Roberta Pazzini (pazzini@istat.it); Giovanni Seri (seri@istat.it); Giampiero Siesto (siesto@istat.it); Luigi Virgili (virgili@istat.it), Italian National Institute of Statistics - Istat.
The views and opinions expressed are those of the authors and do not necessarily reflect the official policy or position of the Italian National Institute of Statistics – Istat.

Introduction

In the activity to enhance the information available in the administrative sources Istat has built the *Frame-SBS* microdata register starting from the reference year 2012. It contains, for each unit of the business register on the active enterprises (Asia), the structural information (economic activity, geographical location, number of persons employed and employees) and some economic variables (revenues, cost of buying goods and services, personnel costs, value added and other more detailed variables) that result from the estimating process applied to the data from different administrative sources (Chambers of Commerce, Revenue Agency and INPS (National Institute of Social Insurance). *Frame-SBS* also provides information that allow classification of the statistical units according to several criteria, such as membership in groups and carrying out trade activities with foreign countries. With reference to the year 2013, the target population consists of 4,297,482 enterprises, of which 4,286,955 (99.7%) with less than 100 persons employed and 10,527 units with 100 and more persons employed (0.3%). The economic information related to the enterprises with less than 100 persons employed is obtained mainly from administrative sources, taking into account the information carried out from sample survey on small and medium enterprises and on the exercise of arts and professions (PMI), used in instrumental ways for the construction of the imputation models of other variables. Information related to the enterprises with 100 and more employed is obtained from the total survey of the system accounts of the enterprises (SCI).

The *Frame-SBS* responds to the EU Regulation on structural business statistics n. 295/2008 (SBS), and lets the Institute try to increase the information released.

A task force has been charged to examine/test the possibility to release data by breaking it down further specifically about the combination of economic activity and size classes of persons employed. The calculation of indicators on the distribution of variables in different domains (medium and variability indices is consistent with the SBS Regulation) and also with the confidentiality treatment.

The topics investigated in this paper are the following: Chapter 1 the provisions of applicable regulatory requirements; Chapter 2 the SBS Regulation n. 295/2008 and the sources of information used to build the *Frame-SBS*; section 3 operational choices for widening the detail of dissemination of data and the constraints imposed by the confidentiality processing; Chapter 4 the methodological aspect related to the disclosure limitation method applied by the generalised software τ-Argus; Chapter 5 the IT aspects; Chapter 6 the channels through which the Institute disseminates statistical data tables (I.Stat) and elementary data (ADELE Laboratory) and the description of the microdata files section of the English Istat website.

1. Legal framework of personal data protection[3]

The dissemination of data on structural business statistics complies with personal data protection provisions and Regulation 295/2008 (Chapter 2).

According to art. 8 of the Code regarding personal data protection (Legislative Decree 30 June 2003, n. 196), data can be disseminated only in aggregate form in order to guarantee the confidentiality of respondents. Also the Legislative Decree n. 322 of 1989 (art. 9, Paragraph 1) states that data coming from relevant statistical surveys included in the National Statistical Programme (NSP) can be disseminated only in aggregate form.

The confidentiality breach occurs when, using released data, an intruder gets confidential information about a statistical unit (for example, a survey respondent). Confidential information is defined as any information that the surveyed units do not want to get public and that the statistical offices and the offices of the National Statistical System (Sistan) have pledged to keep anonymous for legal constraints (but also to maintain a relationship of trust with surveyed units). It includes sensitive data and judicial data as well, while

[3] Due to delays in the publication of this paper, it should be noted that the legal framework of personal data protection presented is not updated with the General Data Protection Regulation 679/2016. This new Regulation imposes stringent obligations and introduces new responsibilities aimed at ensuring greater security measures to protect personal data. In fact, the regulation introduces clearer rules regarding information and consent, defines the limits to the automated processing of personal data, and also establishes strict criteria (and penalties) in cases of violation of personal data. Moreover, the new Regulation states that, in some defined cases, the data subject shall have the right to obtain from the controller the erasure data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay (the so called right to erasure).

the public variables are not considered confidential. When the statistical information to be released involves confidential variables, it is necessary to assess if there is a risk of disclosure.

The Italian regulation specifies that the risk of confidentiality breach has to regard the damage implied by disclosure. The level protection depends on the type of data to be released. For example, sensitive data relating to individuals (such as health information) should have a lower risk of identification compared to other confidential data (such as economic data).

Legislative Decree n. 322 of 1989 and Legislative Decree n. 196 of 2003 state that the exchange of personal data within the Sistan is possible if it is necessary for requirements provided by the NSP or to allow the pursuit of institutional goals. Moreover the Directive “Criteria and procedures for communication of personal data in the National Statistical System” of the Steering Committee and coordination of statistical information (Comstat) (Directive n. 9 of 20 April 2004) provides that a body or statistical office belonging to Sistan may request personal data. The supply of personal data with identification variables is, however, limited to cases of absolute necessity and impossibility to achieve the goal without the identification data. The Sistan bodies have to submit their request through the Contact Centre Istat, that is the web system for the acquisition and on-line processing of requests for statistical information and dissemination services.

As regards subjects who don’t belong to Sistan, Article. 7 of the Code regarding the protection of personal data (Decree n. 196/2003) states that it is possible to communicate individual data files without direct identifiers and which are protected by the application of different statistical methods that make it highly unlikely the indirect identification of statistical units.

Istat produces different types of anonymised microdata files:

- Standard files regard a number of surveys conducted on individuals and households. They may be requested by a variety of users, but restricted to study and research purposes;

- Microdata files for research (MFR) are developed in relation to statistical surveys regarding individuals and households as well as enterprises, and are created specifically for the purposes of scientific research. They contain a higher level of information detail with respect to standard files.
- mIcro.STAT are public use files which can be downloaded directly from the Istat website. They are obtained by applying (more) protective measures to the files for research and the information content of mIcro. STAT is a subset of the MFR.
- The requirements and conditions for the release of these files depend on the subjects requested them and they are subject to the signing of an agreements.

The communication of personal data to researchers from universities or institutes or research bodies or scientific societies members who don't belong to Sistan is allowed in Research data centre with the following conditions:

a) data come from a survey carried out by a Sistan subject;

b) data provided do not include identification data;

c) researchers who access to the Research data centre must respect the rules governing statistical confidentiality and protection of personal data;

d) access to the Research data centre is controlled and monitored;

e) access to different data stores from the subject of the communication is not permitted;

f) appropriate measures are taken to ensure that the input and retrieval of data are inhibited to researchers;

g) the release of the results of calculations performed by researchers using the Research data centre is allowed only after prior verification by the staff of the Research data centre.

The Istat Research data centre (Laboratory for the Analysis of elementary data, ADELE) is located in Rome and in Istat's regional offices. In ADELE researchers can elaborated data coming from surveys carried out by Istat. Moreover some integrated files regarding enterprises are available in ADELE.

To access the Laboratory ADELE the applicants must belong to a university or other research institution and they have to submit a project indicating the data that they intend to develop and the objectives of the research. In the laboratory, it is not provided methodological/technical support to users. The authorisation shall be signed by the President of Istat.

Before entering the ADELE Laboratory, the users sign a contract that obliges him to the maintenance of statistical confidentiality. At the end of the project the ADELE staff, who verifies compliance with the rules for protecting confidentiality, evaluate the results of the calculations.

The procedure of the laboratory, how to access and output release practices are shared in the simple lines between European countries, and included in a process of harmonisation at the international level.

Article. 7 of the Code regarding the protection of personal data (Legislative Decree 30 June 2003, n. 196) also provides for a further possibility of data communication by Sistan subjects to researchers working on behalf of universities, other public institutions and agencies pursuing research purposes, as part of joint projects.

2. Regulation on Structural Business Statistics (SBS), administrative and statistics sources, series to be provided and development of the Community regulations (FRIBS)

Beginning from the reference year 2008 the Structural Business Statistics (SBS) have been covered by European Regulation (EC, Euratom) n.295/2008, adopted on 11 March 2008 by the Council and Parliament. Its objective is to establish a common framework for the collection, processing, transmission and evaluation of the Community statistics on the structure, activity, competitiveness of enterprises in the Community. The Regulations n. 250/2009 and 251/2009 of 11 March 2009 implement the SBS Regulation for the definition of the variables, the technical format for the transmission of data and for the series to be produced and transmitted to Eurostat.

The SBS Regulation develops nine annexes, each of them points out the series and the data breakdown to be transmitted.

As regards the annexes directly transmitted by Istat, the main domains of estimation for the annexes 1-4 (respectively for services, industry, distributive trade and construction activities) are the following:

a) 4-digit Nace rev. 2 without any distinction by size class of persons employed;

b) 3-digit Nace rev. 2 by size class of persons employed (0-9, 10-19, 20-49, 50-249, 250+ in industry and construction; 0-1, 2-9, 10-19, 20-49, 50-249, 250+ for trade and services);

c) 2-digit Nace rev. 2 at (3-digits for trade) by administrative region at level of Nuts2.

Preliminary data of annexes 1-4 have to be transmitted within 10 months from the end of the reference year at 3-digit Nace, whereas the final data have to be transmitted with the abovementioned details within 18 months.

The annex 8 (business services) regards statistics on the enterprises with specific economic activities and 20 and more persons employed; the information requested are on turnover by product type and customer nationality. Some activities are investigated annually (Nace 582, 62, 631, 731 and 78) and others every two years (Nace 691, 692 and 702 in even years; Nace 7111, 7112, 712, and 732 in odd years).

For the annex 9 (business demography) data are required up to 4-digits Nace, broken down by legal status and class of employees.

Data for the others annexes (5-insurance services, 6 credit institutions, 7-pension funds) are transmitted by other organisations.

Table 1 shows the main series to be transmitted by Istat, and to be treated jointly for statistical data confidentiality, by identifying domains with primary confidentiality (i.e. with a number of enterprises less than 3 units) and those one whose secondary confidentiality is assigned due to the hierarchical classification of economic activity and the breakdown of the data (e.g. size classes of persons employed). For more information on the confidentiality and its treatment with specific software Tau-Argus see Chapter 4.

The main variables requested for the Annexes 1-4 are the number of enterprises (code 11100), turnover (12100), production value (12120), gross margin on goods for resale (12130), value added at factor cost (12150), gross operating surplus (12170), purchases of goods and services (13110), purchases of goods and services for resale in the same condition as received (13120), change in stocks of goods and services (13210), change in stocks of goods and services purchases for resale (13211), personnel costs (13310), wages and salaries (13320), number of hours worked by employees (16150), gross investment in tangible goods (15110), number of persons employed (16110) and number of employees (16130).

Until the reference year 2011, the information sources used by Istat to fulfil the SBS Regulation have been the sample survey on small and medium-sized enterprises and on the exercise of arts and professions (PMI, enterprises with less than 100 persons employed) and the census survey on the system of company accounts (SCI, enterprises with 100 or more persons employed).

Starting from the reference year 2012, the main aggregates on the enterprises with less than 100 persons employed are estimated on the basis of a micro-data file (called Frame) that integrates several administrative sources (Financial statements from the Chambers of commerce; Sector studies survey, Tax statements data from the Fiscal Authority; labour costs data from the Italian National Institute for Social Security). The Frame aggregates are obtained by summing variables at the individual level in the domains of interest (economic activity, size classes of persons employed, regions, etc.). By contrast, the estimates of the PMI survey for variables not available from administrative sources are obtained by multiplying the variables by the final weight, thus obtaining meaningful data only for planned domains. The estimates obtained through Frame/PMI survey are added to SCI survey data to finally build the micro-data file named *Frame-SBS*.

The field of observation of PMI and SCI surveys, and hence of the *Frame-SBS*, is wider than SBS Regulation request, as it includes the Nace activities: P (Education), Q (Human health and social work activities), R (Arts, entertainment and recreation) and division 96 (Other personal service activities) whose data are disseminated through the data warehouse I.Stat.

From a technical point of view, the joint processing of *Frame-SBS* brings to macro level domain data, and then the the Tau-Argus data confidentiality procedure is run. Afterwards an IT process leads to the series requested by the Regulation annexes. Such series are then checked via the EBB Tool, a software developed by Eurostat, which defines both longitudinal and cross-sectional data quality checks. The SBS data production process ends with the transmission of the series to Eurostat via the web application eDAMIS.

In order to allow Eurostat to calculate aggregates at EU level, the SBS data are clear transmitted, i.e. by hiding nothing and indicating the confidential domains.

From the 1st January 2021 will enter in force in the European Community the Framework Regulation Integrating Business Statistics (FRIBS) n.2019/2152: the aims is to define a harmonised framework for the collection, transmission and dissemination of European statistics on the structure, activity, competitiveness, global transactions and the performance of enterprises. FRIBS Regulation will take over from the Regulations on Structural Business Statistics (SBS), Short-Term Statistics (STS), Inward and Outward Foreign Affiliates Trade Statistics (IFATS, OFATS), Foreign Direct Investment Statistics (FDI), Global Value Chains and International Sourcing (GVC), Innovation Statistics (CIS), Research and Development Statistics (R&D), Statistics on the Information Society (ISS), International Trade in Goods Statistics (ITGS) and marketed production of manufactured goods (Prodcom).

Table 1- Main series of the Annexes 1-4 and 8 to transmit to Eurostat for the SBS Regulation

Serie	Annexes and description
	Services, Industry, Distributive trade and Construction
1A 2A 3A 4A	Annual enterprises statistics (Nace at 4 digits)
1B 2B 3B 4B	Annual enterprises statistics by size classes of persons employed (Nace at 3 digits) *Size classes of persons employed: 0-1, 2-9, 10-19, 20-49, 50-249, 250+, total for series 1B and 3B* *Size classes of persons employed 0-9, 10-19, 20-49, 50-249, 250+, total for series 2B and 4B*
1C 2C 3C 4C	Annual regional statistics by Nuts2 (Nace at 2 digits for Services, Industry and Construction; Nace at 3 digits for Distributive trade)
1P 2P 3P 4P	Annual preliminary statistics on the enterprises (Nace at 3 digits)
	Services
1E	Annual enterprises statistics for special aggregates
	Industry, Constructions
2D 4D	Annual kau* statistics (Nace at 4 digits)
2E 4E	Multiannual enterprises statistics – Intangible investment (Nace at 4 digits)
2F 4F	Multiannual enterprises statistics – Sub-contracting (Nace at 4 digits)
2G 4G	Multiannual enterprises statistics – Size classes of turnover (Nace at 4 digits)
	Industry
2H 2J	Annual enterprises statistics on the environmental expenditure broken down by environmental domains (Nace at 2 digits)
2I 2K	Annual enterprises statistics on the environmental expenditure broken down by size classes of persons employed (Nace at 2 digits) *Size classes of persons employed: 0-49, 50-249, 250+, total*
	Distributive trade
3D	Annual enterprises statistics by size classes of turnover (Nace at 3 digits)
	Construction
4H	Multiannual enterprises statistics – Sub-contracting by size classes of persons employed (Nace at 3 digits) *Size classes of persons employed: 0-9, 10-19, 20-49, 50-249, 250+, total*
	Business services
8A	Annual enterprises statistics for activities of Nace rev.2 (62, 582, 631, 731 and 78) broken down by product type
8B	Annual enterprises statistics for activities of Nace rev.2 (62, 582, 631, 731 and 78) broken down by residence of client
8C	Annual enterprises statistics for activities of Nace rev.2 (691, 692 and 702) broken down by product type
8D	Annual enterprises statistics for activities of Nace rev.2 (691, 692 e 702) broken down by residence of client
8E	Biennial enterprises statistics for activities of Nace rev.2 (732, 711 and 712) broken down by product type
8F	Biennial enterprises statistics for activities of Nace rev.2 (732, 711 e 712) broken down by residence of client

* Kau = Kind of Activity Unit

The main objectives of the Fribs Regulation are on the one hand to rationalize the complex regulatory framework for European business statistics and define a new architecture for complying the compilation of business statistics (using more integrated manner information from administrative sources in order to reduce the statistical burden) and on the other hand to improve the quality of statistics on service sector, globalisation and entrepreneurship.

3. New information detail

From the reference year 2013 new information details are available in the data warehouse I.Stat:

1. a wider breakdown by economic activity combined with the size classes;
2. position and variability indices for the main variables and for some indicators for the breakdowns (hereafter domains) by 1, 2, 3 and 4 digits of economic activities.

As for point 1: the size classes 0-1 and 2-9 persons employed have been introduced for manufacturing and construction sectors; and the same size classes are also used for 4-digits domains of economic activity. For such breakdowns the only core variables are disseminated, i.e. the variables: number of enterprises, number of persons employed, number of employees, sales proceeds (turnover), other revenues and income, cost of purchased goods, cost of purchased raw materials, cost of purchased services, costs of third parties assets, other operating costs, personnel costs, total wages and salaries, value added and gross operating surplus. Such variables, thanks to the use of administrative data, are consistent for so fine details.

As regards point 2: the series transmitted to Eurostat have been enriched with the first, second and third quartile and the standard deviation[4]. The availability of individual data for all enterprises, in fact, allows comparisons of competitiveness within and among sectors at the micro-level (enterprise). Position indices, for variables that describe asymmetric economic phenomena and that are, have the properties to not being influenced by extreme values. Besides, the index of variability helps to analyse the heterogeneity within the domain.

The quartiles and the standard deviation are calculated on the following variables: number of persons employed, number of employees, turnover, personnel costs, value added, gross operating surplus; and on the following ratios: turnover per person employed, personnel costs per employee, value added per person employed, vertical integration[5], wage adjusted labour

[4] The interquartile range is computable as a difference.

[5] Value added divided by turnover (percent).

productivity[6]. Ratios are calculated excluding null values at the denominators. The number of 1) enterprises with persons employed, 2) enterprises with employees, 3) enterprises with non-null turnover are available.

In order to ensure the confidentiality rules for data dissemination - the quartiles to be disseminated have been computed as the arithmetic average of the five (six) values around the actual quartiles if the number of units is odd (even). Then, by following the rule of a minimum number of values (threshold) to calculate on the indices is set up: when the number of units of one domain is less than 50, indices are confidential.

Indices application programme has been developed in SAS language: the horizontal indices file of indices[7] has been translated into Oracle environment, with the following information: domain, variable name, variable value, state of confidentiality[8].

To conclude, the indices here described at the micro-level are different from those obtainable by processing the aggregated tables SBS variables and available in the data warehouse I.Stat. This is mainly due to the fact that the relationship between aggregates in a specific domain represents a mere evaluation in the domain. For example, the ratio of the total value added and the total number of persons employed in one domain (macro-level index), provides a numerical value that can be far from the average or median in the same domain (micro-level index). In fact, 1) the different calculation algorithm, 2) the asymmetry of the distribution of economic variables, 3) the possible presence of high extreme values, 4) a high frequency of null values, and 5) the exclusion of null variables from denominator makes the comparison between macro and micro level indices non feasible. Anyway, in both cases the indices suit to compare sectors competitiveness (Istat, 2015).

[6] Ratio between value added per persons employed and personnel cost per employee (percent).

[7] One single record per domain.

[8] Either confidential to Eurostat or the number of enterprises is less then threshold.

4. Methodological aspects in *Frame-SBS* data protection

4.1 Breakdown of non-nested tables into nested ones

The aim of the national statistical Institutes (NSI) to disseminate surveys results at the most available details has to be performed respecting the surveyed units confidentiality. In Italy, the regulation is represented by Decreto legislativo 322/89 and the "Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell'ambito del Sistema statistico nazionale"[9].

Legal constraints turn into methodological rules and statistical procedures to reduce (within prescribed limits) the identification risk.

Below it is described the procedure applied to protect *Frame-SBS* data (year of reference 2013). Paragraph 4.1.1 analyzes the hierarchical classifications (nested and non-nested) used for tabular data. Subsequently, the document addressed the issue of the *Frame-SBS* data preparation. In particular, it addressed the disclosure limitation methods using generalised software.

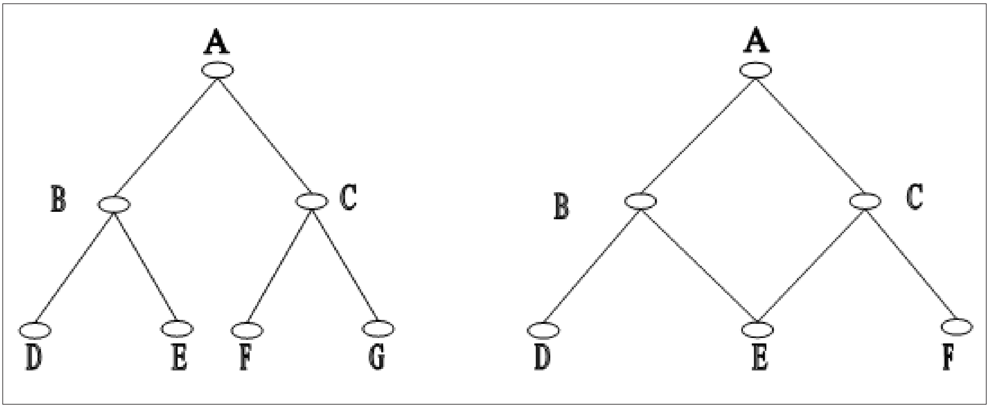
4.1.1 Non-nested hierarchical classification

A classification is called hierarchical when it splits the data along a tree structure as shown in Figure 1. 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. (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.

[9] Paragraf 1.

Figure 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, a unit can belong to more than one class (or higher hierarchical level) (see Figure 1- right)

Reporting the aggregates represented in Figure 1 as levels of spanning variables is how the following two schemes are obtained:

Scheme 1

A	
-B	
--D	
--E	
-C	
--F	
--G	

Scheme 2

A	
-B	
--D	
--E	
-C	
--E	
--F	

The root (A) is the total vertex representing the levels of variables (o subtotals). The number of dashes (“-“) represent the hierarchical levels.

In the scheme 2, the variable categories “E” occurs twice and the additivity is not respected. To protect this data, it is necessary to split the table into two linked tables that contain all the variables levels as represented below:

Scheme 3	
A	
-B	
--D	
--E	
-C	

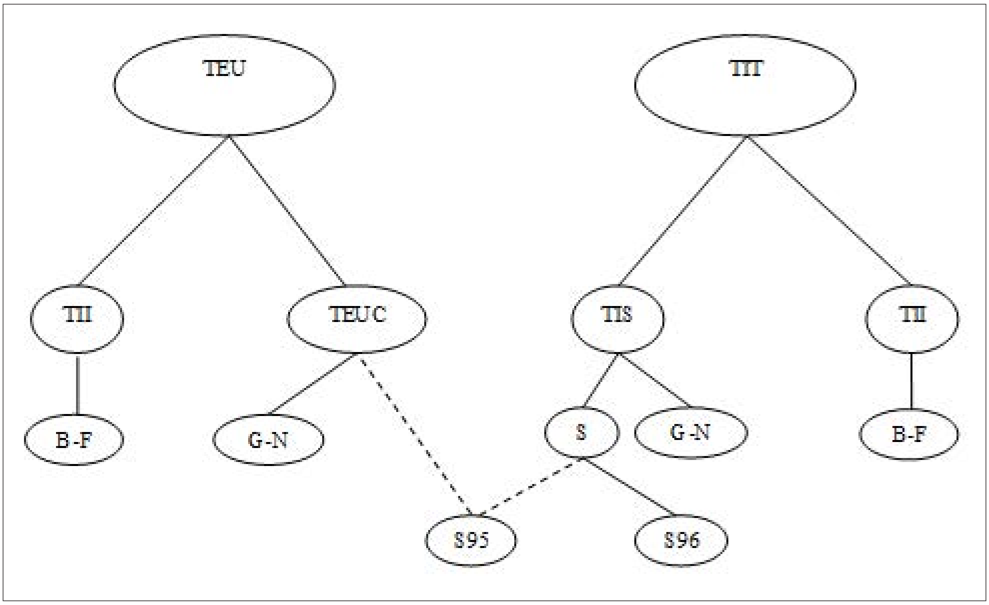
Scheme 4	
C	
-E	
-F	

Tabular data organised according to the schemes 3 and 4 can be protected one by one, making sure to assign the same confidentiality flag to the common cells (C, E).

4.1.2 Hierarchical nested classification in data Frame-SBS protection

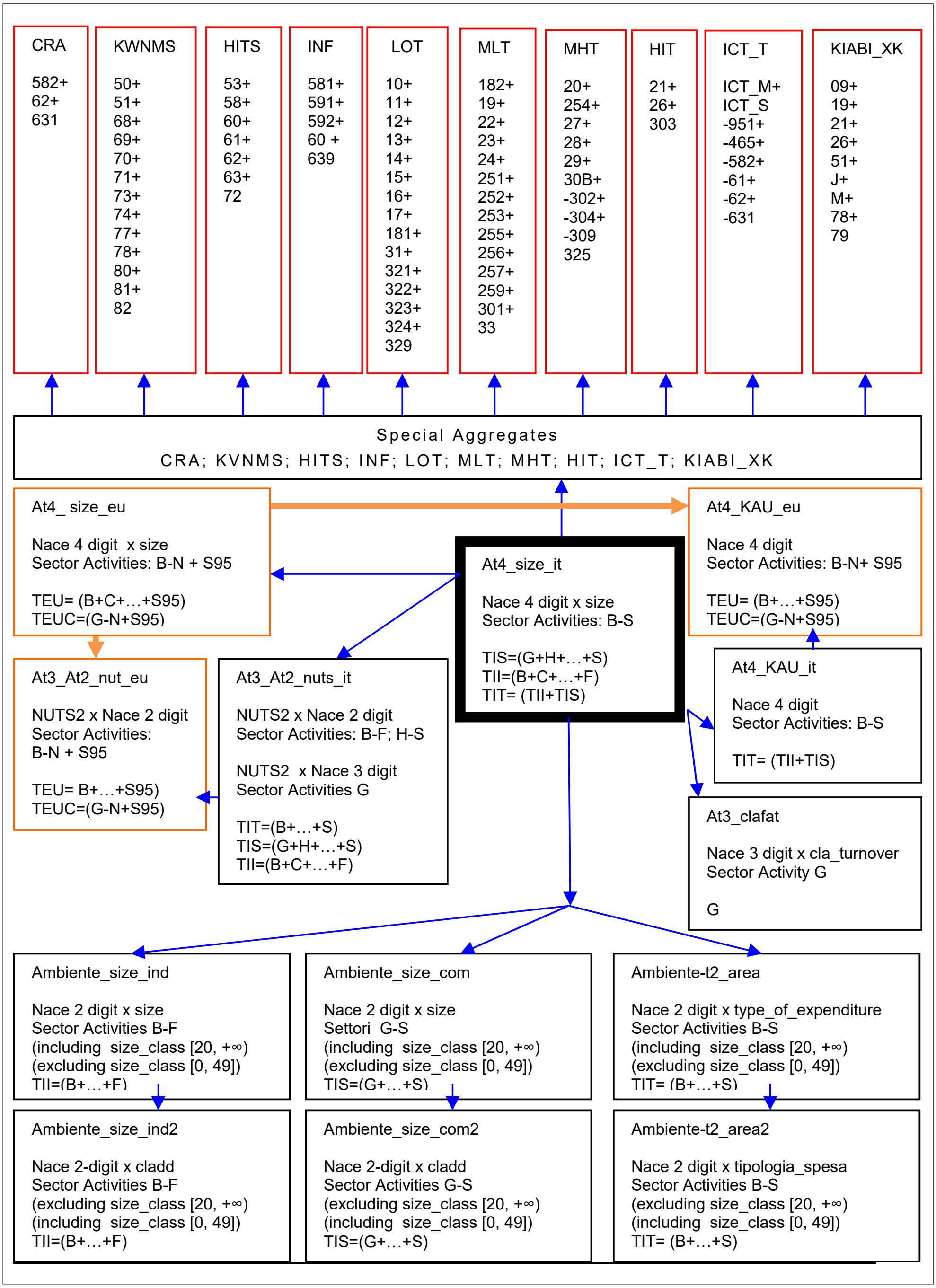
Figure 2 reports the diagram related to the aggregates TEU e TIT that Istat builds and releases.

Figure 2 - Diagram of TEU e TIT



TEUC level occur also in TIS: TEUC e TEU are subsets of TIS and TIT.

Figure 3 - Frame-SBS data in linked tables



However, it is impossible represent all of the aggregates as categories of a tabular classification variable without duplications. Similar cases occur considering other aggregates identified by the European Regulation, in particular considering special aggregates and environmental data.

The result of the break-down process, applied to Frame-SBS data reference of year 2013, is represented in Figure 3. Each box represents a hierarchical table. The levels of the spanning variables are nested. In each box, are showed: the name of tables, the classification variables, the sectors of economic activity and the Nace marginal totals. The tables are linked because common cells, the arrows showed these links between tables.

The tables in Figure 3 represent a superset of the data that has to be released. They contain both the aggregates provided by the series of European regulations, and the domains used in national released.

To link At4_size_it and special aggregates their levels are needed. For this reason, in the aggregate MHT occurs the category 30B, that it is an ad-hoc category (defined as the sum of Nace 302 + 304 + 309) also shown in the table At4_size_it.

The classes of persons employed provided by the Regulations for the environmental data (Ambiente_size_ind and Ambiente_size_com) contain the [0-49] category. It is not nested respect to the classification ([0,19], [0,9], [0,1], [2,9], [10,19], [20 and over), [20,49], [50.249], [250 and over)) used for both At4_size_it tables, At4_ size_eu, and environmental tables. This is the reason why environmental data is divided into six tables (instead of three).

Appropriate procedures described in Section 4.2 allow the protection of the set of tables shown in Figure 3, ensuring consistency in the final results.

4.2 Methodological and technical aspects

Tables represent the most common way of dissemination of statistical results. The cells (domains) are defined by the categories of the classification variables. The objective of the breach consist in linking statistical units with their information (for example turnover and enterprise). The hypothesis concerning the information available for the intruder is reported in the disclosure scenario.

For Frame-SBS aggregate data, the classification variables are considered as identifying keys, while the response variables are considered as confidential information.

The risk measure adopted is the linear measures proposed by Cox (Cox, 1981). They are definite by:

[1] $$S(X) = \sum_{i=1}^{\infty} x_i w_i$$

Where:

X is a cell;

x_i is the *i-th* contribution;

w_i is the *i-th* weight.

Cox proposed as risk measure all the linear combination, as described in [1], that are sub-additivity, that means that a cell resulting by the union of two or more cells that are not at risk will result not a risk.

By national law aggregated data are defined as "[...] combinations of categories which is associated with a frequency not less than a predetermined threshold [...] ". According to this definition, the risk rule adopted for table protection in Istat is the minimum frequency (or threshold) rule: a cell is at risk if it is referred to a number of contributors that is lower than a given parameter (n).

Referring to the [1], assuming $x_1 \geq x_2 \geq \ldots \geq x_N$ with N total numbers of contributors in X.

Defining w_i as :

$$w_i = -\frac{1}{x_i} \quad \forall i \in [1, \mathrm{N}]$$

$S(X)$ will result sub-additive. In fact in the [1] the sub-additivity will be obtained if and only if, $w_i \geq w_j$, $\forall i < j$. (see Cox, 1981)

Disclosure limitation implies a decrease of the information content with respect to the original data. Istat, applies methods that reduce the information released without changing the observed values: the contributions of sensitive cells are suppressed and replaced by the so called "confidentiality flag".

The protection process of the aggregated data does not end suppressing sensitive cells. In fact, as showed in the following example where the Table 2 contain one suppressed confidentiality cells, the (X1, Y1):

Table 2 - Intensity tabular data with one cell at risk (suppressed)

	Y_1	Y_2	Tot
X_1	a	5	6
X_2	3	5	8
Tot	4	10	14

In the Table 2 the value replaced with flag "a" can be recalculate.

Further suppressions (secondary suppression) are necessary to ensure that risk cells (suppressed) will not be breached:

Table 3 - Intensity tabular data with primary and secondary suppressions

	Y_1	Y_2	Tot
X_1	a	b	6
X_2	c	d	8
Tot	4	10	14

In Table 3 cells (X1, Y1), (X1, Y2), (X2, Y1), (X2, Y2) are suppressed, and values are replaced with the Flags a, b, c, d. In the example, it is assumed that (X1, Y1) is the only one cell at risk and that the letters b, c, d are used as a flag for the secondary suppressions.

Table 3 appears as a protected tables. However, even in this case, it is possible, by a linear equation system, carry out intervals for any cell suppressed:

a+b=6;

c+d=8

a+c=7

b+d=4

with a,b,c,d, ≥ 0.

The system can be solved and a *feasibility interval* can be derived for all the cells suppressed.

For risk rule based on the concentration the *feasibility intervals* are defined by rule and parameters. Table 4 shows the Upper protection level (UPL) related to some concentration rules:

Table 4

Risk rule	Upper Protection level
Dominance(n-k)	$(100/k)(x_1+x_2+....x_n)$-X
Ratio(p%)	$(p/100)x_1$-(X-x_1-x_2)
Priori-posteriori(p,q)	$(p/q)x_1$-(X-x_1-x_2)

In Table 4, "xi" represents the *i-th* contribution, "X" indicates the total (or the sum) of all contributions; "N" is the number of contributors on which it evaluates the dominance rule; "K" represents the percentage (maximum) of the total contribution which may be held by the first "n" contributors (threshold); "p" and "q" are probabilities.

The range of protection is obtained by adding and subtracting from the true value of the cell at risk the higher level of protection shown in Table 4.

The protection achieved by cell suppressions is considered appropriate, according to the adopted rule, if the protection intervals contain the feasibility intervals.

It is not possible to identify a Upper (Lower) Protection level based on the parameterisation of the threshold rule. However, it is possible to set (*a priori*) a minimum level of protection as percentage of the cell a risk. This solution (implemented in τ-Argus) ensures the size of the interval protection, but cannot ensure about its symmetry around the suppressed value.

Secondary suppressions are identified by minimizing a cost function (according to the protection required or set). The aggregated data protection process results in an optimisation problem solved by an algorithm implemented in the generalised software τ-Argus[10] (http://research.cbs.nl/casc/tau.htm)

[10] The algorithms commonly used in the protection of linked hierarchical tables (such as those represented in Figure 3, Section 4.1), are the Hitas (or modular) and the Optimal. The latter can have very long processing times for tables with high complexity, in relation to the number of hierarchical levels, the number of cells at risk and the level of protection set.

The complexity of calculation increases in case of linked tables: the track of suppression for a table becomes input in the protection of the linked tables. In this protection process the degree of freedom decease table by table. The final result also depends on the order of tables protection.

In the hypothesis of all categories are aggregations of the finest details, the general rule is to proceed from the "particular to the general" starting from the most detailed tables (in the common classification variables) and continuing until the less detailed. The tool for this procedure is the so-called history file (or a priori file) that allows keep the same status (protected or released) for common cells. The history file is obtained by τ-Argus function.

4.2.1 Application rules of confidentiality with τ-Argus to Frame-SBS 2013 data

In the case of Frame-SBS data, the disclosure scenario assumes that the intruder is able to place statistical units within the domains defined by the classification variables. This assumption means that the assessment of the disclosure risk has to be carried out for each cell that has to be released.

The risk rule adopted is the minimum frequency rule (k); the parameter k is set equal to three: cells with a number of contributors strictly less than three are defined at risk.

The algorithm used to identify of secondary cells suppression is the modular (or Hitas) (De Wolf 2002). The cost function is the amount of value added: suppression is determined by minimizing the total value of deleted contributions.

Frame-SBS data are protected using τ-Argus software. It allows to select the risk rule, configure the security level, and select the algorithm for secondary suppressions. With reference to the tables in Figure 3 of the Paragraph 4.1, the protection sequence is:

Figure 4 - Protection sequence

1. *At4_size_it*
2. *At4_size_eu*
3. *At3_clafat*
4. *At3_At2_nuts*
5. *At3_At2_nut_eu*
6. *At4_KAU_eu*
7. *At4_KAU_it*
8. *Aggregati speciali*
9. *Ambiente_size_ind*
10. *Ambiente_size_com*
11. *Ambiente_at2_area*
12. *Ambiente_size_ind2*
13. *Ambiente_size_com2*
14. *Ambiente_at2_area2*

The finest domains, Nace 4-digit and size classes of persons employed, are the first to be protected. The resulting history file is used to constrain the confidentiality flag in common cells to be protected.

Frame-SBS data to be released is obtained by selecting the τ-Argus singleton option, relative to cells with only one contributor. It ensures that two suppressed cells protecting each other are not both related to a single contributor (thus excluding the possibility that a "self-recognition" causing a direct breach of confidentiality).

One track of suppression (primary and secondary) is adopted for all response variables. The audit programme, available in τ-Argus, allows to asses *a posteriori* the protection levels achieved.

For transparency introduction of missing values resulting for reasons of confidentiality is communicated to the user.

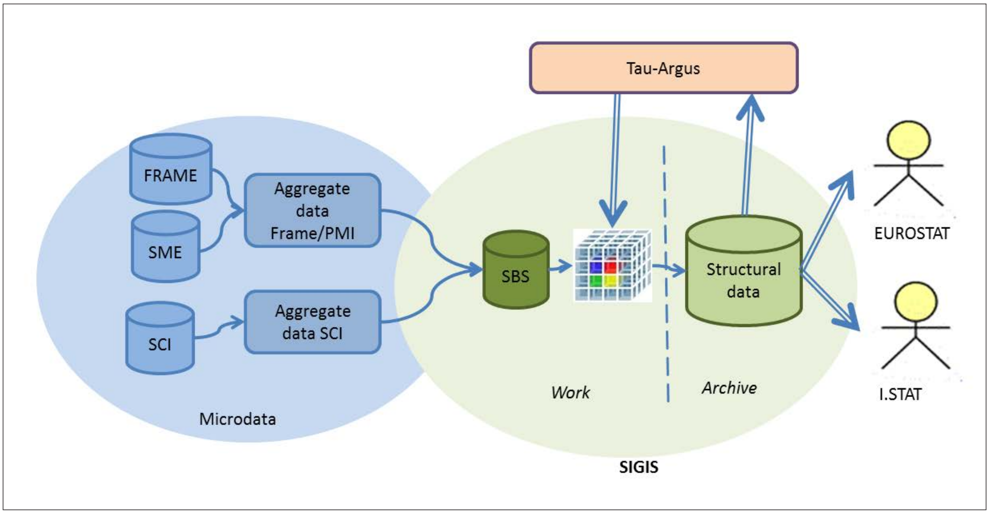
The result of the whole protection process by τ-Argus is a set of tables, exhaustive of the all domains to be published, whose cells contain confidentiality flags.

5. IT view of SBS data

The computerisation of the SBS output management is achieved through Sigis, Information System for the Management of the Structural Indicators, which provides the preparation activity support functions of the structural indicators and ensures their storage on two different areas: *work* and *archive*.

The SBS process in Sigis provides five distinct steps of processing, as illustrated in figure 5. The first phase of the process consists in the aggregation of survey microdata of production processes Frame/PMI and SCI composing SBS, this phase is implemented in data production environment. In the second phase the calculation of the SBS aggregate is carried out, this phase is realised in the work area of Sigis, common to all the structural surveys; the aggregates at the maximum detail of Frame/PMI and SCI add up providing the base of the SBS data. The data results of this step are accessible to production managers in read-only mode for commercials processing and verification. In the third step the preparation of SBS indicators on dissemination is carried out, in order to aggregate the variables in accordance with the regulations for the different levels of diffusion; operations carried out in this phase consist in rolling-up the dimensions of analysis for the different variables, that is an aggregation according to the hierarchy of each dimension starting from the end of the level. In the fourth step the integration for the management of confidentiality is made; at this step the data are arranged for being handled by the methodology group in charge of the confidentiality management; the confidentiality procedure receives from Sigis data files and returns the same with the appropriate indication of confidentiality. In the fifth and last phase the extraction of the structural indicators is carried out, in this phase the information to be used for dissemination to Eurostat and I.Stat is arranged.

Figure 5 - SBS process in Sigis



5.1 Sigis architecture

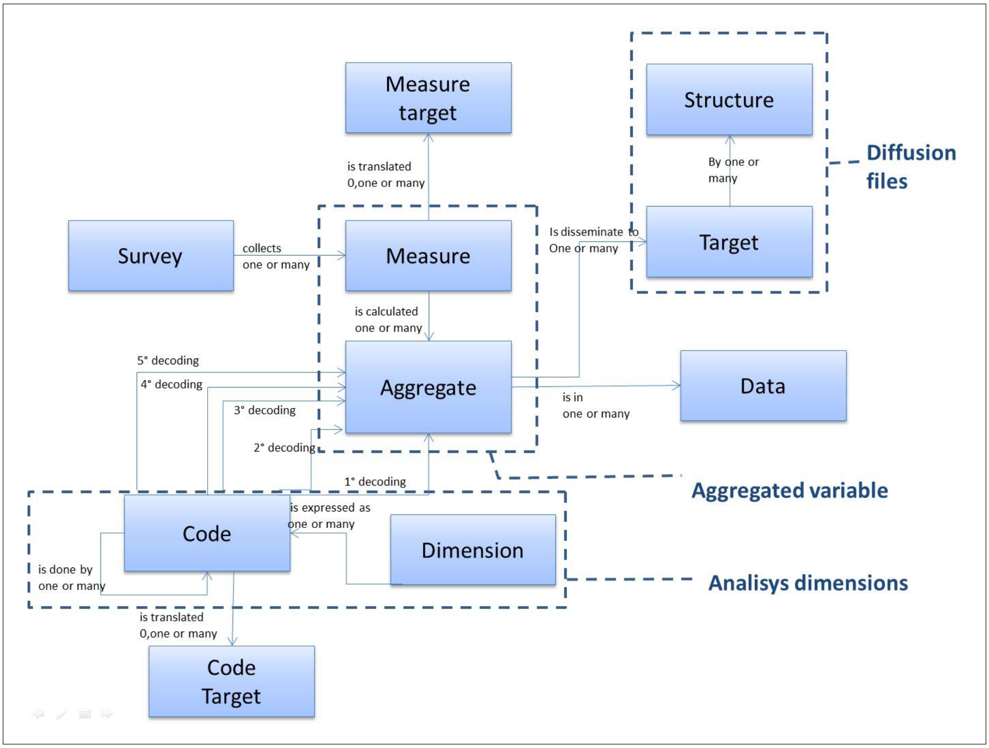
The management information system of the structural indicators, Sigis, consists of two environments:

a) a work area for the data loading functions, roll-up for the analysis dimensions for each variable and the integration with the confidentiality system used for SBS; in particular the confidentiality system is a user when the system receives the data files on which to apply the rules of confidentiality and also a provider of information when it releases the files in which the status of confidentiality is assigned;

b) an archive area where the aggregated data is stored with an indication of confidentiality, to be accessed to perform data extractions for the different dissemination systems (I.Stat, Eurostat).

Data loading is implemented through Pl-Sql procedure in Oracle platform; the loading procedures are customised for each current survey in Sigis. The data model on which Sigis is based is the following:

Figure 6 - Data model Sigis



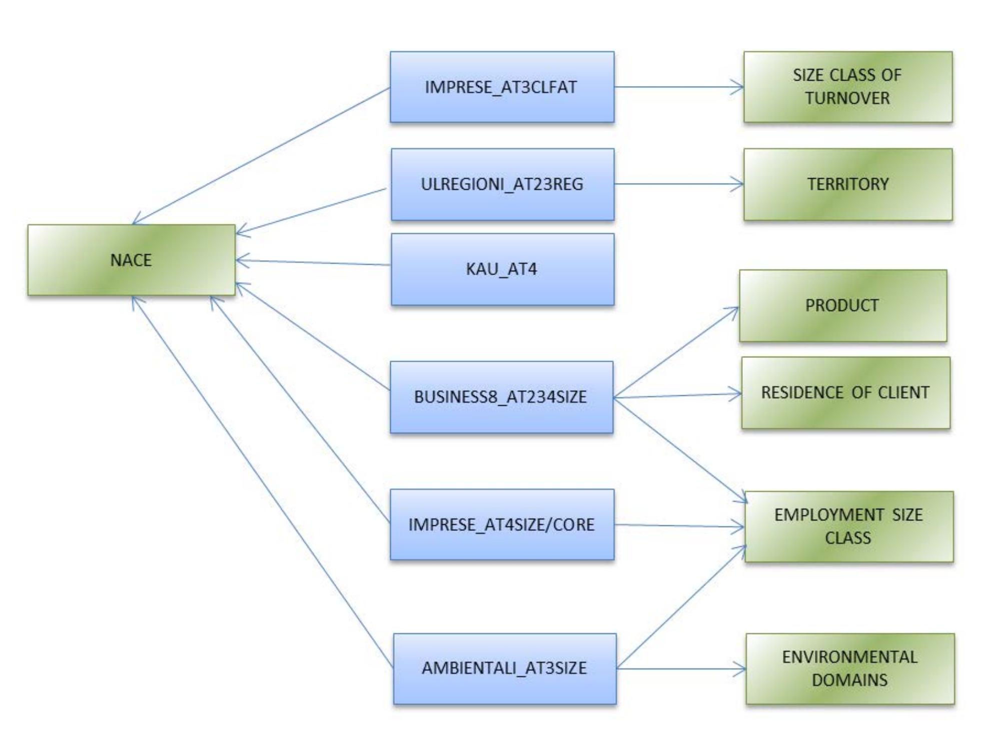
A survey collects one or more aggregated variables according to different dimensions of analysis that, in the present context, arrive to a maximum of five crossings. Each item of a dimension can be defined as a union of other voices, storing this information the value can be automatically calculated. The aggregate can be arranged for one or more diffusion files, for each diffusion file are stored : the type of the structure, the separator character of the fields and, for each column, the information necessary for the identification of the content. In each output file it is possible to insert different aggregates, and each aggregate may be contained in several files, for that reason there is an ad hoc table to store the list of the aggregates for each output file. Each variable and each code of the dimensions can be contained in various diffusion file according to a different decoding with respect to the encoding used for storage in the system, this information is stored respectively in Measure Target and Target Code. The aggregate is loaded every time the information for a new reporting period is released or when a review of an already diffused period is carried out; any information that would be appropriate can be jot down, in particular if it is confidential or not.

5.2 SBS data

SCI aggregated data united with the aggregates of Frame/PMI, through the operations of sum on the same scale generate SBS data.

The schema of the SBS data is a star type schema and is described below.

Figure 7 - Dimensional model SBS



The green color tables represent the dimensions of analysis while the blue color tables represent aggregated dataThe dimensions of analysis considered to create the aggregate are the following:

a) Nace,

b) size class of turnover,

c) product,

d) residence of client,

e) class of person employed,

f) territory,

g) environmental domains.

5.3 Integration for the management of confidentiality

τ-Argus software is a Sigis's target, it gets from Sigis file as input data but it returns to Sigis the same number of files with indication of the cells to suppress.

τ-Argus software is the recipient "T" and the files for τ-Argus are files "csv" delimited by ",", in the appropriate table stores the aggregate, as requested by da τ-Argus.

For the download data it is possible use the function above described with appropriate parameter, the data exchange with τ-Argus is done with SAS files, after the writing csv files with Sigis, a sas programme converts files according to the required format; the sas programme read the Sigis table for the files number, the files name and the files structure.

The τ-Argus output for the SBS data are stored in sas files, the same files number, a sas programme reads the files, reads the Sigis tables for the files name to read and it converts the sas files in csv files. Then a Pl-Sql programme reads csv files and stores all data in a database's table and according the τ-Argus output it updates the field relating to the confidentiality.

6. Dissemination and communication of data

6.1 Aggregated data dissemination: tables and indicators

The expansion of the set of information relating to SBS data has required an updating of the data production chain for I.Stat.

This update has required several steps, including the revision of the documentation plan for I.stat inherent to the structure and competitiveness of enterprises, the consequent updating of Sigis, the I.stat data warehouse updating and the implementation of the changes introduced on the website of I.Stat.

In general, this activity has been divided, in an organisational point of view, in the four steps of the Deming cycle, or PDCA (Plan, Do, Check, Act). Below, the description of the four steps:

1. in the “Plan” step, the task force has planned upgrades to be introduced in the set of SBS data
2. in the “Do” step, these updates were translated in a first modification of the documentation plan for I.stat;
3. in the “Check” step, the changes, of the documentation plan for I.stat, were shown to all the units involved in the data dissemination chain for I.Stat, and some improvements have emerged to be made to documentation plan, regarding the assignment of codes to the new position indices and the methods for developing and displaying queries on I.Stat;
4. in the “Act” step, the Sigis, the data warehouse of I.Stat and the I.Stat website have been up-dated, according to the provisions in the documentation plan. Previously, the data warehouse and the I.Stat website, have been upgraded within a safe area not accessible to external users, for a further step of testing by the production sector.

The steps above show the importance of the shared construction of the documentation plan for I.stat, relating to SBS data on the structure and competitiveness of enterprises. By the way, the documentation plan describes what data (also indicated with the term of “data types”), will be published, excluding data that are not covered in the dissemination purposes, and at the same time the methods of query construction, indicating how data are combined with the classification variables.

In this specific case, new data types have been introduced in the documentation plan: the position indexes (47 new data types). To improve the usability of the data, the 47 new data types were divided into 6 groups:

1. Distribution of turnover indicators
2. Value added distribution indicators
3. Gross operating surplus distribution indicators
4. Personnel costs distribution indicators
5. Persons employed and employees distribution indicators
6. Wage adjusted labour productivity distribution indicators

As it regards the methods of the queries construction, has been decided to create 6 queries, one for each individual group of data types, and to spread data types for the classification NACE with a level of depth to 4 digits.

In the documentation plan for I.Stat relating the structure and competitiveness of enterprises we have also introduced changes related to a broader breakdown of core variables for economic activity combined with the size of workers. The new query is structured with the size classes in the header and the NACE on side, while the data type is selectable by a pulldown menu.

Finally, the new queries were included in I.Stat within the theme “Enterprises” and sub-theme “Competitiveness - National Structure Business Statistics (data from 2008 onwards).” The query on the broader breakdown of core variables for economic activity, combined with classes of employees, is the first query of this environment and is called the “Main variables for classes of NACE and size classes.” Queries on position indices are included in a specific index of this environment regarding “distribution indicators.”

6.2 Access and comunication of individual data

The integration of data Frame with different sources (administrative, PMI and SCI) allows the production of two different microdata files: “Frame-SBS - Integrated system of administrative data and survey data for the estimation of economic aggregates on enterprises", and "TEC-FrameSBS -Structure and economic performance of exporting firms". Both files can be requested by Sistan subjects and are available for the scientific community at the ADELE laboratory. The data Frame are stored in the system ARMIDA (ARchive MIcroDAta) whose main objectives are to maintain validated metadata and microdata of surveys carried out by Istat, and to promote re-use of micro-data for statistical purposes by external users.

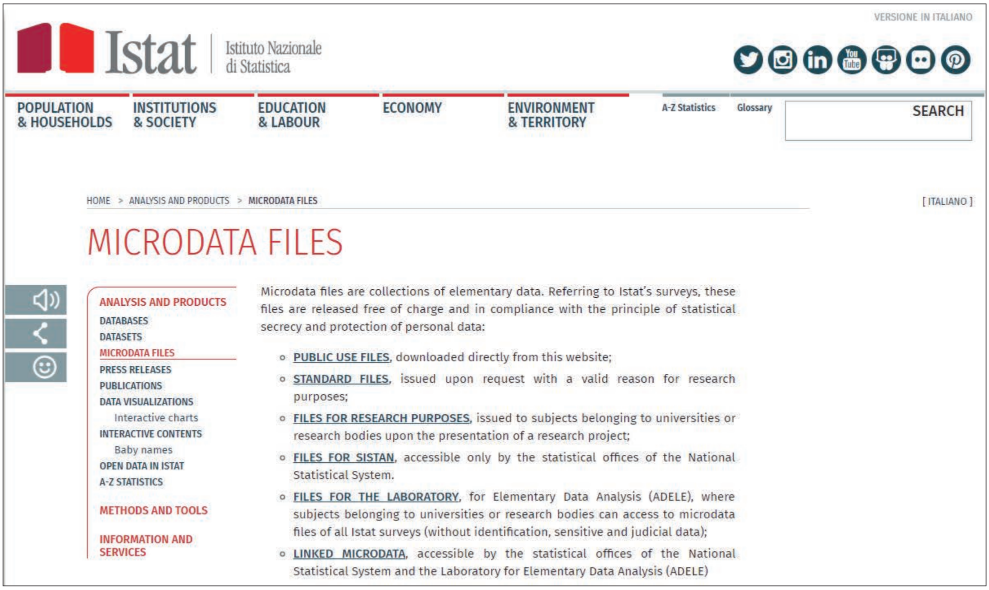
Frame-SBS file contains the twenty following variables:

1. Enterprise Code (Code Asia)
2. Region code
3. Number of persons employed
4. Number of employees
5. NACE 4-digit
6. Membership in enterprise groups
7. Artisan enterprise
8. Class of persons employed
9. Revenues from sale of goods and services
10. Other income
11. Costs for raw materials, supplies, consumables and goods
12. Cost of services
13. Costs for use of third party assets
14. Personnel costs
15. Salaries and wages (gross earnings)
16. Other operating expenses
17. Value added at factor cost
18. Gross operating surplus
19. Exports of goods (source: Istat-Coe)
20. Imports of goods (source: Istat-Coe)

The microdata "TEC-FrameSBS -Structure and economic performance of exporting firms" come from the integration of three different statistical sources: the statistical register of active enterprises (Asia), the register of operators that realize foreign trade of goods (Coe) and Frame-SBS files. The main variables of interest are: value added, labor costs, turnover, purchases of goods and services, the value of exports (total value and decomposed by geographical area and major groupings of products), the value of imports (total value and decomposed by area geographical and major groupings of products), the number of exported and imported products, many countries / regions export and import.

6.3 Istat web site description

Microdata files have a dedicated section on the English Istat website: starting from the home page, click "Analysis and products" on the footer menu and then "Microdata files" on the submenu. You will find the various types of files created by Istat and the conditions for accessing and using them.



MICRODATA FILES

Microdata files are collections of elementary data. Referring to Istat's surveys, these files are released free of charge and in compliance with the principle of statistical secrecy and protection of personal data:

- PUBLIC USE FILES, downloaded directly from this website;
- STANDARD FILES, issued upon request with a valid reason for research purposes;
- FILES FOR RESEARCH PURPOSES, issued to subjects belonging to universities or research bodies upon the presentation of a research project;
- FILES FOR SISTAN, accessible only by the statistical offices of the National Statistical System.
- FILES FOR THE LABORATORY, for Elementary Data Analysis (ADELE), where subjects belonging to universities or research bodies can access to microdata files of all Istat surveys (without identification, sensitive and judicial data);
- LINKED MICRODATA, accessible by the statistical offices of the National Statistical System and the Laboratory for Elementary Data Analysis (ADELE)

In order to protect the anonymity of respondents (persons, organisations), in the download area Istat just provides the metadata files and the methodological notes of each survey or data collection.

The microdata archive of the website is organised by typology:

- Public use files, collections of elementary data accessible directly from the Istat website and provided free of charge;
- Standard file, files containing anonymised data, issued upon request of any applicants for scientific purposes only;
- Files for research purposes, files with a high level of detailed information, issued only to subjects belonging to organisation recognised as a research entity upon the presentation of a research proposal;
- Files for Sistan, elementary data files requested by the statistical offices of the National Statistical System in order to implement the National Statistical Programme (PSN);
- Files for the Laboratory, for Elementary Data Analysis (ADELE), where subjects belonging to universities or research bodies can access to microdata files of all Istat surveys (without identification, sensitive and judicial data);
- Linked microdata, special datasets combining data coming from different surveys, available for access at the ADELE Laboratory.

Concerning the Linked microdata website section, since 2013 Istat has released metadata on TEC-FrameSBS, a database obtained linking information on exporting firms from TEC (Trade by Enterprise Characteristics) and the main economic variables from Frame-SBS (Structural Business Statistics).

Elementary data from Frame-SBS referred to years 2012 and 2013 are accessible at the Laboratory for Elementary Data Analysis (ADELE), a Research Data Centre (RDC) where researchers working for universities or research institutions or fellows of bodies can conduct, free of charge, their own statistical analyses on microdata from the Istat's surveys. The aim of the ADELE Laboratory is to meet those needs of scientific research that are not satisfied by conventional tools for accessing statistical information (such as publications, data tables, databases, microdata files).

On the ADELE Laboratory webpage you can browse the list of Istat surveys by theme. In particular the Frame-SBS variables list is found under the heading for “Industry and services”:

- Frame SBS - Integrated system of administrative and survey data for the estimation of structural business statistics (since 2012);
- TEC – FrameSBS (since 2013).

Aggregate data from Frame-SBS together with the PMI (Small and Medium Enterprise) (for variables not available from administrative sources) and SCI (Business accounts system) data are available on the Istat data warehouse I.Stat, under the heading for “Enterprises”. The tables contain data referred to years 2012 and 2013 at various levels of disaggregation. In previous years the tables stored the estimates of the PMI and SCI surveys. From the home page of the Istat website you can reach the I.Stat data warehouse by clicking on the special banner present in.

In the press release website archive, the statistics reports “Structure and competitiveness of the industrial and services enterprises” present the main economic results on enterprises, based on the Regulation (EC) n. 295/2008 concerning structural business statistics. Actually in Italy the traditional SBS estimation strategy has been completely reversed in 2012 with the development of the Frame SBS, as in this new statistical information system, administrative and fiscal data are used as primary source of information, whit the complementary use of the PMI and SCI data. The statistics report comes with aggregate datasets in xls format, methodological notes and glossary and it is found under the themes Enterprises, Industry and construction and Services.

As publication, since the 2014 edition the “Productivity and Competitiveness Report” (Rapporto sulla competitività dei settori produttivi - only in Italian) is based on Frame-SBS data for estimating the measurement of productive efficiency of enterprise.

The methodological innovations of the new statistical information system Frame-SBS have been debated also during some events, suchs as the workshop Nuove informazioni statistiche per misurare la struttura e la performance delle imprese italiane (New statistical information for measuring structure and performance in enterprises) and the workshop Micro dati per l'analisi della performance delle imprese: fonti, metodologie, fruibilità, evidenze internazionali (Microdata for the analysis of performance in enterprise). You can read or download slides and abstracts of the speeches either on the webpage or on Slideshare.

Finally all microdata information is quickly accessible from the home page thanks to the search box, a single-line text box with the search service provided by GSA (Google Search Appliance), in which the dynamic navigation allows the user to restrict the search results (*e.g.* by reference period, document typology, theme, tags, date of publication, *etc.*).

7. Conclusions

The Frame-SBS, for the reference years 2012 and 2013, has been allowed to obtain reliable estimates, even for small domains. It was considered possible to expand the information details (larger breakdowns, and statistics about the distribution in some domains), in compliance with current legislation on data protection.

The analysis of the data confirmed the possibility to increase dissemination details without reducing released information (because the number of suppression cells).

Starting from the breakdowns requested by the European Regulation on structural business statistics n. 295/2008 (SBS), the Task force has identified, for the *core* variables, the new breakdowns: the size classes (in terms of persons employed) [0-1] and [2-9] have been applied for all sectors of economic activity and the disaggregation of data according to NACE four-digit by size classes (persons employed) has been adopted.

The software τ-Argus has been used to protect data tables. The final solution was considered a viable solution (comparing number of cells suppressed and released).

For the *core* variables, for domains with at least 50 units at the level of Section and Nace at 2, 3 e 4 digits (with no size classes) quartile and standard deviation were computed. The Task force used quantiles because they are not depending on outlier.

The paper described the regulatory framework for data protection, the European Regulation on structural business statistics n° 295/2008 (SBS), IT aspects, the statistical disclosure control, and all the procedures to release data to Eurostat and data warehouse I.Stat.

One aspect that is critical is related to the opportunity to disseminate the data Frame more widely represented in particular populations (for example, small businesses enterprises, enterprises belonging to groups, etc.). To specify this kind of populations there are some definitional problems, and statistical disclosure control issues as well.

When the information required by Community regulations need to be changed, the procedures to carry out the dissemination have to be adjusted according to the new domains, in accordance with the privacy constraints.

Appendix A

About minimum size of subpopulations for releasing descriptive statistics associated to SBS aggregates

The Frame-SBS has allowed for increasing the information disseminated through the website I.Stat, in particular more detailed tables on main SBS variables will be available. As Frame SBS can be considered as a register, the aggregates released are obtained by sum of individual values. Descriptive statistics such as median or standard deviation can be associated to the aggregates released in order to give information about the distribution of values inside a cell. Usefulness of these increasing of information and the related problem of statistical confidentiality strongly depend on the size (number of enterprises) of the subpopulations belonging to each cell.

Type of variables can be also considered as a factor of interest. Variables such a s Turnover or Personnel costs can be addressed as proxy of the enterprise size while derived variables such as the indexes of Productivity (Value added per employee, Turnover per employee, etc.) are ratios with a reduced disclosive power.

Some descriptive statistics associated to the value of a cell could be informative on the value of a single enterprise if the number of enterprises contributing to the cell value is too low. As an example, consider the first and third quartile are released and their value is similar: the interquartile interval can then represent a very accurate estimate for the true value of a single units unless it can be consider that is unknown (or at least uncertain) the rank of the unit in the subpopulation of enterprises contributing to the cell value. This kind of uncertainty is usually assumed if the number of enterprises contributing to the cell value is higher of a given threshold. Of course, the case of derived variables is less problematic as these kind of information is hard to be informative on a single enterprise and the relative threshold can be set at a lower level.

For the release of SBS tabular data Istat usually adopt the threshold rule, the threshold being set to the minimum level of 3. Given the remarks stated above this threshold cannot be considered appropriate to preserve confidentiality if the information is increased by releasing the proposed descriptive statistics of each cell. Therefore it has been suggested to adopt the rule used by the Research centre (at Istat it is the ADELE Laboratory[11]) where descriptive statistics such as those proposed are released if they refer to a subpopulation of at least 50 units. In other words it has been suggested to adopt the threshold rule, the threshold being set to 50.

In order to test the appropriateness of the suggested rule the size (number of units/enterprises) of 1809 cells (domain of estimate or domain for aggregate values) have been computed. Cells have been defined by the combination of 2 and 3 digit NACE codes and size class (number of employees in classes). The outcome of the analysis is reported in Table A1 where the two threshold rules (3 and 50) are compared. Cells with less than 50 contributors are 569 but only 5 and 19 are relative to cells defined by only 2 or 3 digit NACE respectively despite of the size class. Out of these, 1 and 3 respectively are subject to primary confidentiality also for the SBS threshold 3 rule. It can be then stated that, despite of the size class the adoption of a threshold 50 rule does not considerably increase the number of confidential cells with respect to the usual SBS threshold 3 rule.

Table A1 - Number of cells defined by the combination of NACE (2 and 3 digits) and Size class and confidential according to the threshold rule (threshold 3 and 50)

DOMINI	N	Confidential cell by threshold	
		threshold 50	threshold 3
2 digits NACE	77	5	1
3 digits NACE	236	19	3
2 and 3 digits NACE and size class	1496	545	91
Total	1809	569	95

In the end, cells involving size class will be disseminated without descriptive statistics while cells defined only by NACE, 1 to 4 digits will be released enhanced by descriptive statistics if they refer to at least 50 enterprises.

[11] http://www.istat.it/it/informazioni/per-i-ricercatori/laboratorio-adele

References

Codice in materia di protezione dei dati personali, D.Lgs n. 196 of June 30, 2003, Gazzetta Ufficiale N. 174, Supp. n. 123 (July 29, 2003) annex A.3 (‘*Codice di deontologia e di buona condotta per i trattamenti di dati personali a scopi statistici e di ricerca scientifica effettuati nell’ambito del Sistema statistico nazionale*’). http://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2003-06-30;196!vig=.

de Wolf, P.-P. 2007. "Cell suppression in a special class of linked tables". In the *Joint UNECE/Eurostat work session on statistical data confidentiality*, Manchester, U.K., December, 17-19 2007.

de Wolf, P.-P. 2002. "HiTaS: a heuristic approach to cell suppression in hierarchical tables". In Domingo-Ferrer, J. (*ed*.). Inference Control in Statistical Databases: from theory to practice. *Lecture Notes in Computer Science*, Volume 2316. Heidelberg, Germany: Springer.

Giessing, S. 2001. "New tools for cell suppression in τ-Argus: one piece of the CASC project work draft". In the *Joint UNECE/Eurostat work session on statistical data confidentiality*, Skopje, Republic of North Macedonia, March, 14-16 2001.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Naylor, E. Schulte Nortol, G. Seri, and P.-P. de Wolf. 2010. *Handbook on Statistical Disclosure Control Version 1.2*.

Virgili, L., and L. Franconi. 2009. "Disclosure protection of non-nested linked tables in business statistics". In the *Joint UNECE/Eurostat work session on statistical data confidentiality*, Bilbao, Spain, December, 2-4 2009. http://www.istat.it/it/files/2013/12/Franconi_Virgili_wp.36.e.pdf.

Statistics Netherlands - CBS. 2008. *τ-Argus Version 3.3 User’s Manual*. The Hague, Heerlen, The Netherlands: CBS.