rivista di statistica ufficiale

n. 1-2-3
2017

In this issue:

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 1-2-3
2017

In this issue:

Istat

rivista di statistica ufficiale
n. 1-2-3/2017

ISSN 1828-1982

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Editorial Preface

For the year 2017, the *Rivista di statistica ufficiale* is published in a single issue combining three quarterly numbers and presenting three extended original articles which, as usual, regularly underwent a double-blind reviewing process.

The title of the first article, authored by Elena Catanese and Marcello D'Orazio, is "*A new approach for multipurpose stratification in Agriculture Surveys*". This scientific contribution consists in an analysis of the stratification criteria used when dealing with a set of continuous auxiliary variables. If, on the one hand, it is possible to apply several stratification criteria when dealing with a single auxiliary variable, on the other hand few methods are available when it comes to the treatment of several auxiliary variables.

For this purpose, this study explores an innovative procedure comparing it to a multivariate method. The application of both approaches to the samples of three surveys related to the agricultural sector enables, then, to analyse their respective advantages and disadvantages.

To understand fully the other two articles it is important to know that during the last years the Italian National Institute of Statistics - Istat strongly increased to centrally collect administrative archives and use them for statistical production purposes. The integration of administrative and survey data aims at overcoming most of the limits of the traditional survey-based estimation strategy.

The two studies focus on Istat's commitment and on the use of administrative sources for producing Structural Business Statistics - SBS.

The availability of administrative sources on enterprises and the results of the direct surveys carried out by Istat allowed to obtain the microdata of the *Frame-SBS* register.

Since 2013, this register has been used both to meet the needs of the EU Regulation, in order to be compliant with it, and to improve the information released for national purposes.

More specifically, the second article "*The dissemination process of the Frame-SBS: legislative and methodological aspects linked to increase information detail*" describes the activities and the results of a large Task force within Istat, in charge of improving the information released on Structural Business Statistics.

This study also includes issues related to statistical confidentiality, focussing on the methods applied to protect data, on the IT aspects and on the channels used to disseminate information.

Orietta Luzi, Fabiana Rocci, Roberto Sanzo and Roberta Varriale are the authors of the third scientific article, whose title is "*A quality evaluation framework for the statistical register Frame-SBS*".

The transition to a production strategy essentially based on the use of administrative data requires the development of innovative methodological approaches, and determines the need of new tools to evaluate the quality of both data and statistical process.

This innovation involves a real tailoring of the current approaches used for quality measurement and assessment. For these reasons, this paper aims both at proposing a first pattern of indicators for the measurement and the documentation of the *Frame-SBS* quality, and at implementing a quality control system. The latter is necessary in order to monitor this register on a regular basis, by identifying possible process and data weaknesses, and by supporting quality improvements.

Nadia Mignolli

Coordinator of the Editorial board

A new approach for multipurpose stratification in Agriculture Surveys

Elena Catanese, Marcello D'Orazio [1]

Abstract

The stratified random sampling is frequently used in sample surveys on businesses and farms because of its efficiency and practical advantages. The stratification of the target population is a crucial step; it is based on the information available in the sampling frame being related to the phenomena under investigation. The task is not straightforward in multipurpose surveys, where different phenomena are simultaneously investigated. Several stratification criteria can be applied when dealing with a single auxiliary variable while few methods are available to deal with several auxiliary variables.

This work introduces a relatively new and simple procedure to stratify the sampling frame in the presence of a set of continuous auxiliary variables. Its main advantages and disadvantages are highlighted. The procedure is compared with one of the reference methods by applying both to the design of some sample surveys in the agricultural sector.

Keywords: stratified random sampling, multivariate stratification.

1 Elena Catanese (catanese@istat.it); Marcello D'Orazio (madorazi@istat.it), Italian National Institute of Statistics, Via C. Balbo, 16, Rome, Italy.
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.

1. Introduction

Traditionally sample surveys on enterprises and farms are based on one stage stratified sampling; it consists in partitioning the sampling frame into non-overlapping *subpopulations* (or *strata*) and selecting an independent sample in each subpopulation (*stratum*). If the strata consist of homogenous units with respect to the investigated phenomena then the stratified sampling permits to reduce the sampling error and to derive reliable estimates for each subpopulation (Cochran, 1977). In agriculture surveys, usually homogeneous strata can be achieved by partitioning the farms according to geographical information, type of farming (specialist crops, specialist livestock, mixed) and some measures of farm's size (e.g. size of areas with crops, livestock, etc.). The variables for stratification purposes should be chosen among those available in the sampling frame (e.g. Register of active farms, Administrative register, the previous Census data); the more the auxiliary variables are correlated with the target variables, the higher will be the benefits in using them in stratification. In general, for practical purposes it is common to organise the stratification in a way that the estimation domains are obtained by a simple aggregation of the elementary strata.

Stratifying a population is relatively simple when it is performed using categorical variables (e.g. geographical regions), while eventual continuous auxiliary variables need to be categorised in advance. Multipurpose surveys pose additional problems; the main difficulty consists in creating strata of homogenous units with respect to the different phenomena to study. Moreover, the sampling frame may provide several auxiliary variables correlated with the target ones but uncorrelated with each other; thus increasing difficulties in choosing the stratification variables.

This paper tackles the problem of stratifying a population in presence of a set of continuous auxiliary variables, by exploring a relatively new procedure, illustrated in Section 3. This new procedure is compared with a multivariate method proposed by Ballin and Barcaroli (2013) by applying both to design the samples of three agriculture surveys; the main findings are summarised in Section 4. The main features and notation of stratified random sampling design are provided in Section 2.

2. Stratified Sampling

The main decisions in stratified sampling regard (i) how to stratify the population and how many strata to create; (ii) which selection scheme to adopt in each subpopulation (simple random sampling, systematic, probability proportional to size, etc.); and, finally, (iii) the size of the whole sample and the corresponding partitioning among the strata (so called *allocation*); these decisions are strictly related.

Stratification allows for different independent selection schemes in each subpopulation; the common practice in business and agriculture surveys consists in applying the *simple random sampling without replacement* in all the strata, because of its practical and theoretical advantages. The sample size is decided according to the desired precision for the main survey estimates (expressed in relative terms: desired sampling error divided by the quantity to estimate, denoted usually as CV). For instance, in the European Union (EU) the desired CVs in estimating the total amount for the main variables through national agriculture surveys (e.g. Farm Structure Survey) are explicitly listed in the EU regulations. The allocation of sample among the strata may follow different rules: equal allocation, proportional allocation, Neyman allocation, power allocation etc. The choice is related to the desired precision characterizing the final survey estimates and to the stratification strategy.

2.1 Main characteristics of stratified random sampling

Let U be the finite population under investigation, consisting of N units. At first, U is divided into H non-overlapping subpopulations or strata ($U = U_1 \cup U_2 \cup \ldots \cup U_H$) whereas N_h denotes the number of units in the stratum h and, consequently, $N = \sum_{h=1}^{H} N_h$. Then, a simple random sample without replacement, s_h of n_h ($n_h \leq N_h$) units is selected independently stratum by stratum; the overall sample size is $n = \sum_{h=1}^{H} n_h$. An estimate of the total amount of the target variable Y in U, $t_y = \sum_{k \in U} y_k$, is provided by:

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} y_k \qquad [1]$$

and the associated sampling variance is:

$$V\left(\hat{t}_y\right) = \sum_{h=1}^{H} V\left(\hat{t}_{yh}\right) = \sum_{h=1}^{H} N_h \left(N_h - n_h\right) \frac{S_{yh}^2}{n_h} \qquad [2]$$

It is usually expressed in relative terms: $CV_{t_y} = \sqrt{V\left(\hat{t}_y\right)} \Big/ t_y$, known as *relative standard error*.

By fixing in advance the desired relative error in estimating the total amount of a continuous variable Y, it is possible to determine the required sample size n_{opt} (see e.g. formulas in Section 5.9 in Cochran, 1977). The partitioning of n_{opt} among the strata can follow different criteria; in *proportional allocation* the stratum sampling fraction is set equal to the stratum relative size, $n_h = n_{opt} N_h / N$ so to ensure equal inclusion probabilities to all the units in U; in *optimal allocation* (or *Neyman allocation*) the sampling fraction is higher in more heterogeneous strata, being $n_h \propto N_h S_h$; *power allocation* (cf. Särndal et al., 1992, pp. 470-471) is a compromise between the Neyman and an allocation ensuring constant precision for each of the strata estimates. The derivation of n_{opt} and its allocation between the strata requires information concerning the Y variable, usually not known in advance. For this reason it is considered an auxiliary variable X, known for all the units in the population and being highly correlated with Y.

In multipurpose surveys the same sample should provide estimates for several target variables fulfilling the given precision requirements (CVs); for this reason, the decisions concerning the overall sample size and the corresponding allocation should be approached in a multivariate framework by setting up a convex mathematical programming problem; Chromy (1987) and Bethel (1989) provide solutions to this problem.

In the traditional approach to stratified sampling, the sample size and its allocation among the strata are derived given a certain stratification of U, i.e. once decided H and the corresponding partitioning of U into non-overlapping strata ($U = U_1 \cup U_2 \cup \ldots \cup U_H$). In such an approach the first step should be necessarily the stratification of the target population.

2.2 Stratification of the Population in the Univariate Case

Consider a single continuous target variable Y; an efficient stratification should try to derive strata as homogeneous as possible in terms of Y values. Unfortunately, these values are not known in advance and consequently the stratification is carried out on an auxiliary variable X strictly related with Y and whose values are known for all the units in the population. When X is a categorical variable (e.g. geographical regions, NACE in case of business surveys or Farm Typology in agriculture surveys) the stratification is straightforward: strata are formed by units with the same (or similar) X category. On the contrary, a categorisation step is needed when X is continuous.

Different criteria are available for categorizing a continuous X variable for stratification purposes. A widespread criterion is the *cumulative* $\sqrt{f}$ *rule* (Dalenious and Hodges, 1959). Unfortunately it performs poorly when X shows a highly skewed distribution; a frequent scenario in business and agriculture surveys, where most of the continuous variables presents high positive skewness. This situation is commonly solved by separating the few large units in a specific stratum; all these units are included with certainty in the sample (so called *take-all* stratum; cf. Hidiroglou and Lavallée, 2009); in practice, given that these units contribute at large extent to the total amount in the target population, separating them in a stratum that is censused allows to reduce the whole sample size. Hidiroglou (1986) proposes an iterative

algorithm to identify the threshold b_c for creating the take-all stratum (all the units with $x_k > b_c$); the procedure requires setting the desired CV. Once identified the take-all stratum, the remaining units can be further stratified according the *cum* $\sqrt{f}$ *rule* (or other criteria).

Lavalleé and Hidiroglou (1988) introduce a unified procedure for both identifying the take-all stratum and stratifying the remaining units. Once set the desired level of precision (CV), the procedure provides a stratification that minimizes the overall sample size. The partitioning of the sample between the take-some strata follows the *power allocation* criterion. Unfortunately, the Lavalleé and Hidiroglou procedure is based on an iterative algorithm which may not converge to a global minimum; this problem can partly be solved by applying the Kozak (2004) algorithm.

To overcome the problem of an allocation performed on a variable, *X*, assumed to be correlated with the unknown target one (*Y*), Rivest (2002) suggests to consider the anticipated moments of *Y* given *X* in the Lavalleé and Hidiroglou procedure. Baillargeon and Rivest (2009) introduce the possibility of separating very small units in a stratum that is not sampled (*take-none* stratum); these units have a negligible contribution to the total amount of the interest variable, which usually holds true in presence of highly positive skewed distributions. The same authors, provide an important contribution for applying the various methods by developing the software package "stratification" (Baillargeon and Rivest 2011, 2014) freely available for the R environment (R Core Team, 2016).

The stratification problem can also be tackled in a model-based framework (cf. Särndal et al., 1992, Section 12.4). In particular, if it is assumed a linear super-population model with $E_\xi(y_k) = \beta x_k$ and $V_\xi(y_k) = \sigma_0^2 x_k^\gamma$ ($\gamma > 0$, large γ denotes more pronounced heteroscedasticity), then the stratification can be performed by grouping units with similar values of the model variance $V_\xi(y_k)$. In this context the optimal sampling design (i.e. that minimizes the anticipated variance) is the one which ensures inclusion probabilities proportional to the model standard deviation:

$$\pi_k = n \frac{\sqrt{V_\xi(y_k)}}{\sum_{k \in U} \sqrt{V_\xi(y_k)}} = n \frac{x_k^{\gamma/2}}{\sum_{k \in U} x_k^{\gamma/2}} \qquad [3]$$

where n is the expected sample size. A simple fixed-size design which maintains $\pi_k \propto x_k^{\gamma/2}$ is a stratified random sampling design where: (i) the H strata are formed by applying the e*qual aggregate* σ*-rule* (cf. Särndal et al., 1992, Section 12.4), i.e. the strata are formed by grouping homogeneous units with respect to the $x_k^{\gamma/2}$; (ii) the sample is allocated equally among the strata, $n_h = n/H$; and, (iii) the combined ratio estimator is used for estimating the total amount of Y in the population. Usually γ lies in the interval $(0,2]$; in most establishment surveys $1 \le \gamma \le 2$ (cf. Särndal et al., 1992, Section 12.5); when $\gamma = 2$ the optimal model based design provides the same inclusion probabilities of *probability proportional to size* (PPS) sampling, i.e. $\pi_k = n\ x_k / t_x$.

2.3 Stratification of the Population based on Several Variables

The stratification task becomes more complex in a multipurpose survey with many target variables not necessarily related one with each other. In this case there may be a high number of auxiliary variables X related differently with the various target ones; a stratification based just on a single X variable may not be efficient for all the target variables. According to Kish and Anderson (1978) the advantages of using several stratification variables are greater in multipurpose surveys, but potential gains depend on the (i) the relationship between the stratification variables and the target ones, and (ii) intercorrelations among the stratification variables.

The 'traditional' strategy to carry out stratification in presence of a large set of continuous X variables consists in: 1) selecting the X variables highly correlated with most of the target ones; 2) performing univariate stratification on each of the selected Xs, and, then 3) deriving the final stratification by cross-classifying units according to the chosen categorised X variables. Parsimony should be the guiding principle in step (1), the chosen variables should not be related one with each other (or weakly related); moreover, in the presence of a set of highly correlated X variables, it would be preferable to select just the one with the highest relative variability, to avoid any lack of information. This strategy can determine too many strata with many of them too small in terms of size.

In literature there are other proposals to perform stratification in the multivariate framework. For instance, in the bivariate case Kish and Anderson (1978) suggest to apply the *cum* $\sqrt{f}$ *rule* independently on each of the X variables; then the final stratification as a combination of the two results. An extension of the model based stratification to the bivariate case can be found in Roshwalb and Wright (1991).

When dealing with more than two stratification variables Hagood and Bernert (1945) suggest to perform the stratification on a subset of the first principal components computed starting from the set of the *X*s. Pla (1991) considers just the first component. Kish and Anderson (1978) warn against the use of principal components because the final strata cannot be readily interpretable; moreover the principal components analysis (PCA) considers just intercorrelations among the stratification variables and not their relationship with the target variables. Barrios *et al* (2013) note that the PCA is not suitable for high skewed variables with few units exhibiting very high values and performing it on the log-transformed variables may not solve the problem.

Benedetti *et al* (2008) suggest a unique procedure tackling both stratification and sample allocation in a multivariate framework. This procedure requires setting the desired CVs for proxies of the target variables, then a tree-based technique identifies finer and finer partitions of the units by minimizing at each step the overall sample size.

A similar approach is suggested by Ballin and Barcaroli (2013). Their sequential procedure starts with a very fine stratification and then iteratively collapses the strata with the objective of minimizing the overall sample size, given the target precision (CVs) required for a set of proxy variables (can be the same auxiliary variables used to create the initial fine partition) under the optimal Bethel allocation. The proposed procedure makes use of a genetic algorithm and is implemented in the package "SamplingStrata" (Barcaroli, 2014) available for the R environment. The procedure is very effective in achieving a small sample size given the target CVs, however the identified final stratification, obtained in subsequent collapsing steps of the intermediate strata, is not readily interpretable. Moreover the procedure requires a subjective choice for the initial stratification; a possible starting point can be the stratification obtained by cross-classifying the chosen X variables

conveniently categorised. Finally, the Ballin and Barcaroli (2013) procedure requires setting a high number of input parameters and a high number of iterations are necessary to achieve valuable final results, thus implying a non-negligible computational effort.

In a recent article Ballin *et al* (2016) explore the problem of stratification of a sampling frame in the multivariate setting by using the functionalities of the R environment. The paper compares the 'traditional' approach to multivariate stratification and the Barcaroli and Ballin (2013) one.

3. A New Procedure for Stratification in a Multivariate Setting

Recently D'Orazio and Catanese (2016) suggested a new procedure to tackle the problem of stratification in presence of a series of auxiliary variables, supposed to be related to the target ones. The procedure follows the same reasoning of the model-based stratification in the univariate case, but the stratification is performed on the inclusion probabilities obtained by applying the *Maximal Brewer Selection* (MBS; also known as *Multivariate Probability Proportional to Size*, MPPS) (Kott and Bailey, 2000). In particular, the stratification is obtained by applying the *equal aggregate* σ*-rule* to the probabilities

$$\pi_k^* = \min\left\{1,\ \max_j\left[\pi_{1,k},,\ldots\pi_{j,k},\ldots,\pi_{J,k}\right]\right\},\quad k=1,2,\ldots,N \qquad [4]$$

Where

$$\pi_{j,k} = n_j \frac{x_{j,k}^{\gamma/2}}{\sum_{k\in U} x_{j,k}^{\gamma/2}},\qquad j=1,2,\ldots,J\ ;\quad k=1,2,\ldots,N \qquad [5]$$

In practice, to derive the π_k^*, it is necessary to set in advance the "target" sample size n_j for each of the J ($J \geq 2$) auxiliary variables being considered (cf. Kott and Bailey, 2000). A simplifying choice consists in setting $n_j = n_0$ ($0 < n_0 < N$) for $j=1,2,\ldots,J$ (i.e. a constant value). As a consequence the stratification is performed directly on the values:

$$z_k = \max_j\left[\frac{x_{1k}^{\gamma/2}}{\sum_{k=1}^N x_{1k}^{\gamma/2}},\ldots,\frac{x_{jk}^{\gamma/2}}{\sum_{k=1}^N x_{jk}^{\gamma/2}},\ldots,\frac{x_{Jk}^{\gamma/2}}{\sum_{k=1}^N x_{Jk}^{\gamma/2}}\right],\quad k=1,2,\ldots,N \qquad [6]$$

Where the constant γ depends on the heteroscedasticity. Usually $0 < \gamma \leq 2$ but in most establishment surveys a narrower interval $1 \leq \gamma \leq 2$ can be considered. Särndal et al. (1992) claim that $\gamma = 1$ is a good compromise choice, another suggestion favors $\gamma = 3/2$.

Once stratified the units in the desired H strata, the overall sample size and the corresponding allocation is determined by applying the Bethel algorithm.

3.1 A Simulation Study

The efficiency of the D'Orazio and Catanese (2016) procedure (DC hereafter) is investigated through a series of simulations carried out with real data of agricultural holdings. Moreover, a comparison with the procedure suggested by Ballin and Barcaroli (2013) (BB hereafter) is performed. The simulation study considers three sample surveys carried out on Italian Agriculture holdings: (i) the annual Early Estimates for Crop Products Survey (EECPS); (ii) the livestock survey (LS), carried out twice a year; and (iii) the Farm Structure Survey (FSS), carried out every three years, where both livestock and crops are investigated. In practice DC and BB stratification procedures are compared in terms of the overall final sample size needed to achieve the desired CVs, once fixed the total number of strata H; different values of H are considered. The data used for stratification and allocation purposes are those collected in the 2010 Census occasion.

In the present work we focus on specific NUTS1 or NUTS2 regions; in particular, for the EECPS we consider the sampling frame of the 282 017 agricultural holdings having at least one hectare of crops (target population) in the south and islands of Italy (one NUTS1 region). As far as LS is concerned, the considered sampling frame includes all the Italian farms having at least one head of bovine animals, pigs, sheep and goats and includes 173 617 farms. Finally, in the FSS case all the farms of Veneto (one NUTS2 region) are considered (119 384 farms); here, according to EU regulations, there are 5 target variables in terms of crop aggregates and 3 for livestock. The Table 1 provides the list of the variables observed in the 2010 Agriculture Census that in the simulations were used for stratification purposes (X) or as survey target variables (Y) (a variable can be used both for stratification and as proxy of the target one); for each Y variable it is also reported the associated desired CV.

Table 1 - Auxiliary and target variables used for stratification and design purposes

EECPS			*LS*			*FSS*		
X (areas in ha)	*Y (areas in ha)*	*CVs*	*X (No. animals)*	*Y (No. animals)*	*CVs*	*X*	*Y*	*CVs*
Cereals	Durum Wheat	0.03	Bovines	Bovines	0.010	Cereals	Cereals	0.05
	Barley	0.03		Cows	0.015	Industrial crops	Oil seed crops	0.05
	Oats	0.03	Pigs	Pigs	0.020	Harvest. green	Harvest. green	0.05
Legumes	Legumes	0.03	Sheep	Sheep	0.020	Perm. grassland	Perm. grassland	0.05
Harvest. green	Harvest. green	0.03	Goats	Goats	0.050	Vineyards	Vineyards	0.05
Vegetab.	Tomatoes	0.03				Bovines	Dairy cows	0.05
Potatoes	Potatoes	0.03					Other bovines	0.05
						Pigs	Pigs	0.05
						Poultry	Poultry	0.05

Table 2 summarizes the main results of the simulation study in terms of the achieved overall optimal sample size, given *H*: a stratification and allocation procedure that achieves the same target CVs with a smaller sample size is obviously the preferred one.

Table 2 - Overall sample size achieved with the alternative stratification strategies

EECPS			*LS*			*FSS*		
H	*DC*	*BB*	*H*	*DC*	*BB*	*H*	*DC*	*BB*
20	4 107	5 601				20	2 706	2 265
30	4 020	4 996				30	2 664	2 272
40	3 926	4 706				40	2 634	2 044
50	3 821	4 465	50	11 885	3 163	50	2 619	2 103
75	3 682	3 986	75	11 517	3 130	75	2 592	1 977
100	3 498	3 626	85	11 277	3 127	100	2 554	1 851
150	3 381	3 275	110	11 160	3 109	150	2 521	1 837

Results are not homogeneous with respect to the surveys: in designing the EECPS the DC procedure is very efficient and performs better than BB in almost all cases with the exception of $H = 150$. In the FSS case, the BB procedure performs always better than the DC and the distance in terms of final sample size increases as the total number of strata grows. Finally, the BB procedure outperforms DC in LS, in this case a finer stratification (i.e. increasing *H*) does not imply a reduction of the final sample size.

In general, it seems that summarizing a high number of variables with a unique score (Z) subsequently used for stratification purposes may not be a good solution when the variables have different nature as in FSS (areas and animals), but this is not the only reason, given that in LS all the X variables refer to animals. A possible explanation to this situation has to be searched by exploring at the correlations between the Xs. In particular, in the LS case (DC worst performance) it can be seen that 'Bovines' variable is negatively correlated with all the remaining ones (Table 3). This situation suggests to test the DC stratification strategy by applying it separately at two score variables: Z_1 derived starting just from 'Bovines' X variable, and Z_2 derived summarizing the remaining variables ('Pigs', 'Sheep', 'Goats') through the expression [6].

Table 3 - Spearman's correlation coefficients between stratification variables in the LS

	Pigs	*Sheep*	*Goats*
Bovines	-0.17	-0.43	-0.22
Pigs		0.03	0.03
Sheep			0.23

The procedure remains the same as in DC but final strata are derived by crossing the results of the univariate stratification performed independently on Z_1 and Z_2. This new strategy improves markedly the performances of the DC procedure in LS case, as shown in the Table 4; however the BB procedure still remains the best in terms of final overall sample size necessary to fulfill the CVs constraint, for any H.

Table 4 - Performances of the revised DC procedure in LS

H	*DC* *One Z variable*	*DC* *Two Z variables*	*BB*
50	11 885	4 528	3 163
75	11 517	4 258	3 130
85	11 277	4 173	3 127
110	11 160	4 085	3 109

4. Conclusions

The work deals with 'multivariate' stratification and allocation procedures in the presence of a set of continuous stratification variables. The procedure introduced in D'Orazio and Catanese (2016) is further investigated. In particular, while this method is very efficient when the variables are positively correlated or uncorrelated, results can be very poor if one of the initial auxiliary variables is negatively correlated to all the others. A very simple solution to tackle this issue is proposed here. The whole procedure is effective because permits to overcome the problem of choosing a small subset of auxiliary variables (2 or 3 in practice in most cases) to perform separately univariate stratification, by allowing to create only one composite variable starting from a set of several variables (7 in the EECPS case). Moreover the procedure is very simple and with a negligible computational effort. The results obtained when the procedure is applied to design the samples of three agriculture surveys seem promising and, as shown, a marked improvement in some cases can be achieved with a relative additional effort. In any case, the procedure proposed by Ballin and Barcaroli (2013) remains the best if the focus is the reduction of the overall sample size. The price to pay is a higher number of final strata and a non-negligible computational effort. It is worth noting that both the DC and BB procedures provide final strata which are not are not readily interpretable, this is an unpleased feature for subject matter experts and may create problems when, after data collection, strata collapsing should be performed to compensate for empty strata caused by unit nonresponse.

In the proposed procedure the stratification of the transformed variables is performed by using the equal aggregate σ-rule; improvements are likely to be achieved by using more advanced univariate stratification procedures like the Lavalle-Hidiroglou (2009) one.

In summary the new stratification procedure proposed in this work represents a valid fast and simple alternative to achieve an efficient stratification with a relatively small number of strata, when having a small sample size is not a stringent goal (e.g. when oversampling should be performed to prevent reduction of sample size due to nonresponse) and in presence of many target variables where no evident negative correlation among auxiliary variables is present.

References

Baillargeon, S., and L.-P. Rivest. 2009. "A general Algorithm for Univariate Stratification". *International Statistical Review*, Volume 77, Issue 3: 331-344.

Baillargeon, S., and L.-P. Rivest. 2011. "The Construction of Stratified designs in R with the package stratification". *Survey Methodology*, Volume 37, N. 1: 53-65.

Baillargeon, S., and L.-P. Rivest. 2017. "stratification: Univariate Stratification of Survey Populations". *R package version 2.2-6*. http://CRAN.R-project.org/package=stratification

Ballin, M., G. Barcaroli, E. Catanese, and M. D'Orazio. 2016. "Stratification in Business and Agriculture Surveys with R". *Romanian Statistical Review*, N. 2/2016: 43-58.

Ballin, M., and G. Barcaroli. 2013. "Joint Determination of optimal Stratification and Sample Allocation Using Genetic Algorithm". *Survey Methodology*, Volume 39, N. 2: 369-393.

Barcaroli, G. 2014. "SamplingStrata: An R Package for the Optimization of Stratified Sampling". *Journal of Statistical Software*, Volume 61, Issue 4: 1-24.

Barrios, E.B., K.C.P. Santos, and I.I.M. Gauran. 2013. "Use of principal component score in sampling with multiple frames". *Proceedings of the 12th National Convention on Statistics*. Republic of the Philippines, Mandaluyong City, October 1-2, 2013.

Benedetti, R., G. Espa, and G. Lafratta. 2008. "A tree-based approach to forming strata in multipurpose business surveys". *Survey Methodology*, Volume 34, N. 2:195-203.

Bethel, J. 1989. "Sample Allocation in Multivariate Surveys". *Survey Methodology*, Volume 15, N. 1: 47-57.

Chromy, J. 1987. "Design Optimisation with Multiple Objectives". *Proceedings of the Survey Research Methods Section of the American Statistical Association*: 194-199.

Cochran, W.G. 1977. *Sampling Techniques, 3rd Edition*. Hoboken, NJ, U.S.: John Wiley & Sons.

Dalenious, T., and J.L. Hodges. 1959. "Minimum variance Stratification". *Journal of the American Statistical Association*, Volume 54, Issue 285: 88-101.

D'Orazio, M., and E. Catanese. 2016. "A simple approach for stratification of units in multipurpose in business and agriculture surveys". *Istat working papers*, N. 10/2016. https://www.istat.it/it/archivio/185685

Hagood, M.J., and E.H. Bernert. 1945. "Component indexes as a basis for stratification in sampling". *Journal of the American Statistical Association*, Volume 40, Issue 231: 330-341.

Hidiroglou, M.A. 1986. "The Construction of a Self-Representing Stratum of Large Units in Survey Design". *The American Statistician*, Volume 40, N. 1: 27-31.

Hidiroglou, M.A., and P. Lavallée. 2009. "Sampling and Estimation in Business Surveys". In Pfeffermann, D., and C.R. Rao (*eds*.). Sample Surveys: Design, Methods and Applications, Volume 29A. *Handbook of statistics 29/A*. Amsterdam, The Netherlands: Elsevier.

Kish, L., and D.W. Anderson. 1978. "Multivariate and Multipurpose Stratification". *Journal of the American Statistical Association*, Volume 73, N. 361: 24-34.

Kott, P.S., and J.T. Bailey. 2000. "The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling". *Proceedings of the Second International Conference on Establishment Surveys (ICES- II)*, June 17-21, 2000, Buffalo, NY, U.S.

Kozak, M. 2004. "Optimal Stratification Using Random Search Method in Agricultural Surveys". *Statistics in Transition*, Volume 6, N. 5: 797-806.

Lavalleé, P., and M.A. Hidiroglou. 1988. "On the Stratification of Skewed Populations". *Survey Methodology*, Volume 14, N. 1: 33-43.

Pla, L. 1991. "Determining Stratum Boundaries with Multivariate Real Data". *Biometrics*, Volume 47, N. 4: 1409-1422.

The R Development Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Wien, Austria: The R Foundation.

Rivest, L.-P. 2002. "A generalization of the Lavallée and Hidiroglou algorithm for stratification in business surveys". *Survey Methodology*, Volume 28, N. 1: 191-198.

Roshwalb, A., and R.L. Wright. 1991. "Using information in addition to book value in sample designs for inventory cost estimation". *The Accounting Review*, Volume 66, N. 2: 348-360.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. *Model Assisted Survey Sampling*. New York, NY, U.S.; Heidelberg, Germany: Springer.

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)
	Annual enterprises statistics by size classes of persons employed (Nace at 3 digits)
1B 2B 3B 4B	*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.

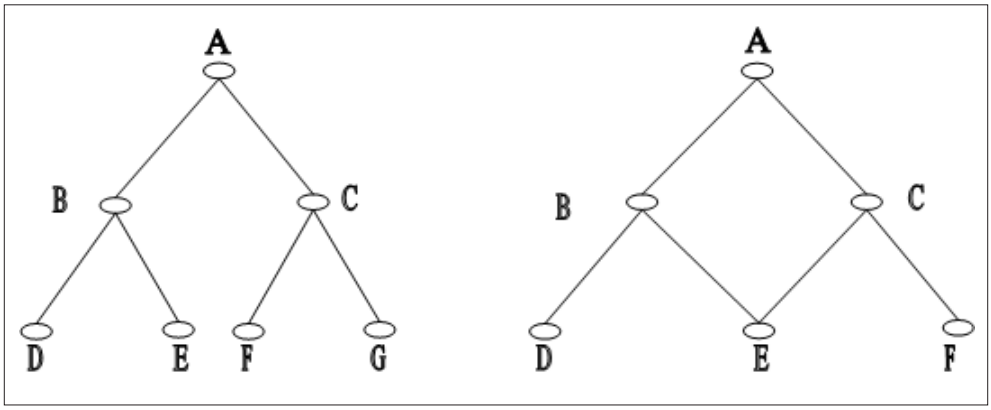
4.1.1 Non-nested hierarchical classification

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	

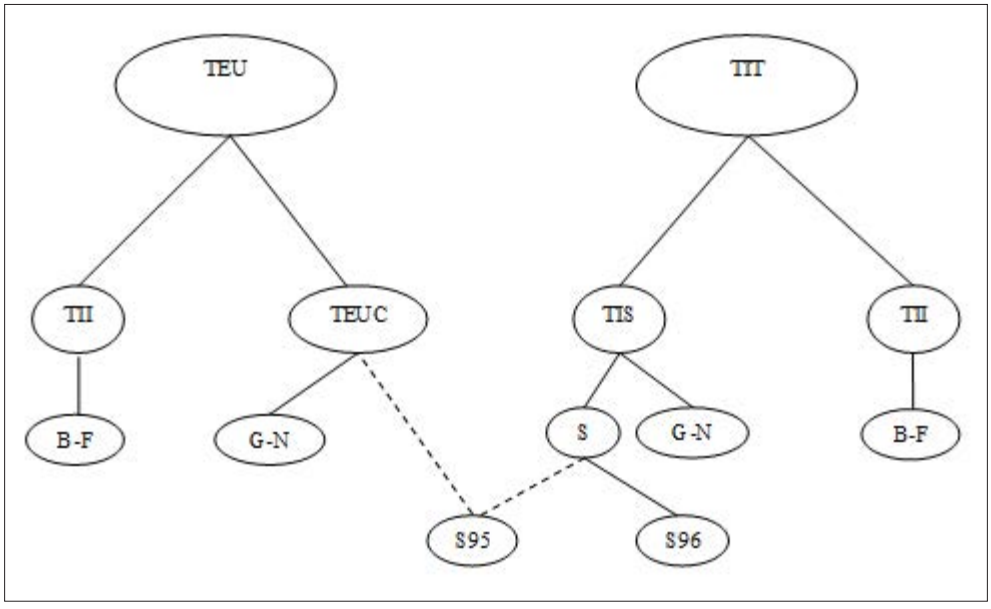
Scheme 4	
C	
-E	
-F	

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

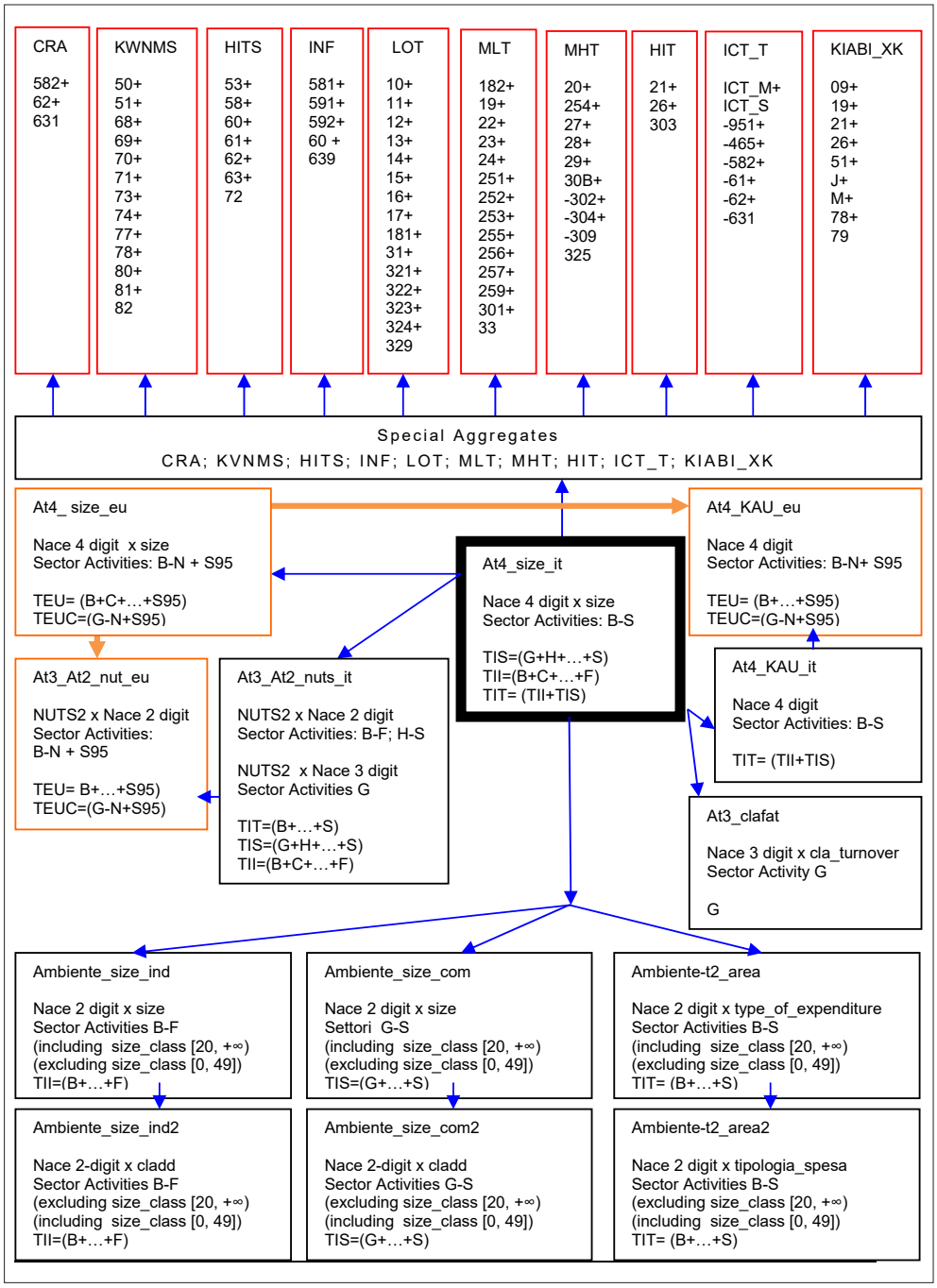
Figure 2 reports the diagram related to the aggregates TEU e TIT that Istat builds and releases.

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 $x_1 \geq x_2 \geq ... \geq x_N$ 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 $a,b,c,d, \geq 0$.

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-Argus10 (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.

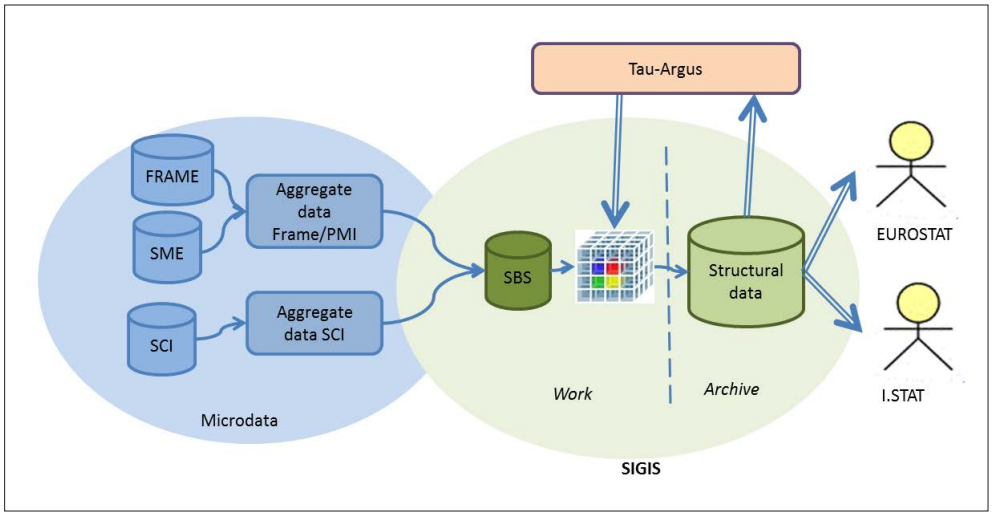
The result of the whole protection process by τ-Argus is a set of tables, exhaustive of the all domains to be published, whose cells contain confidentiality flags.

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

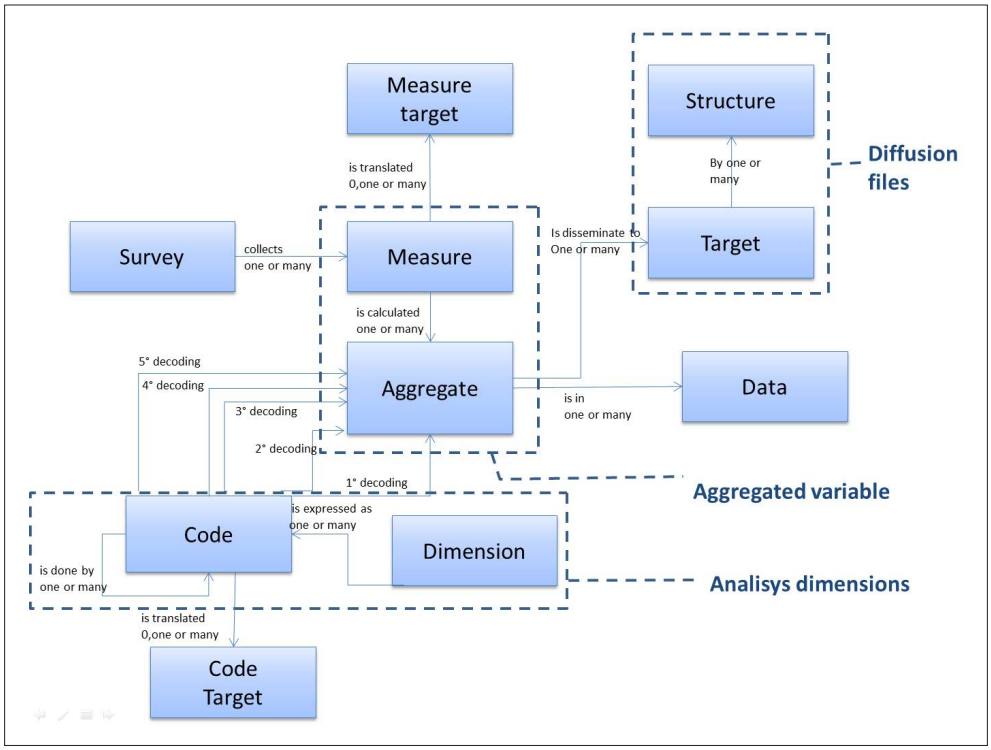
The management information system of the structural indicators, Sigis, consists of two environments:

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

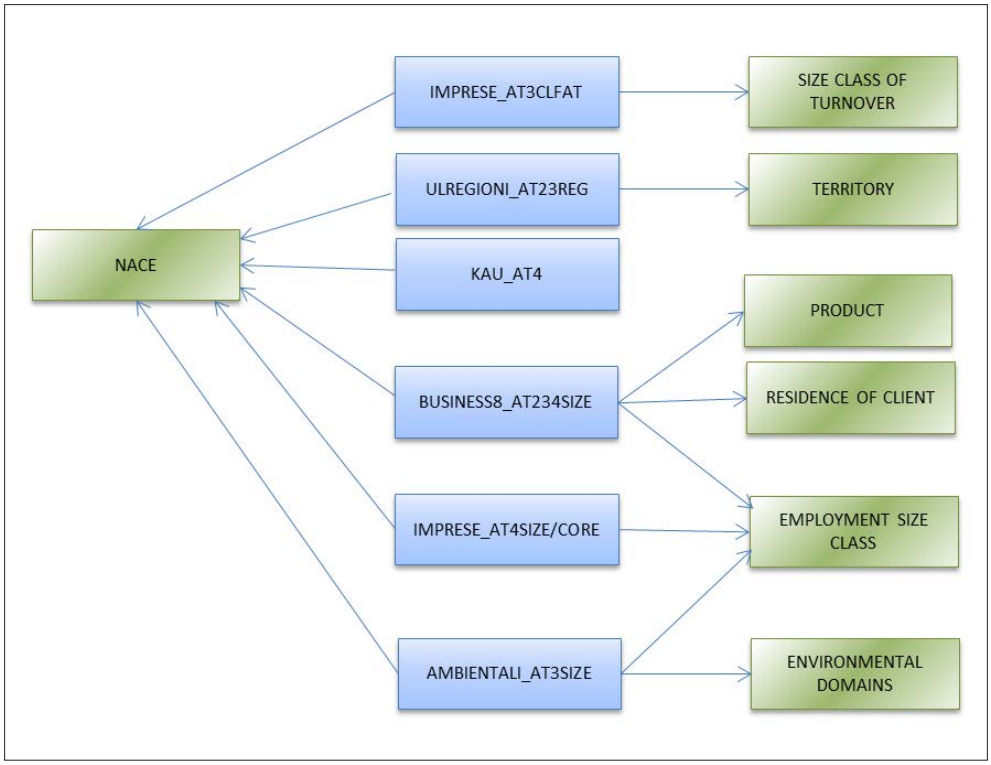
is carried out; any information that would be appropriate can be jot down, in particular if it is confidential or not.

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.

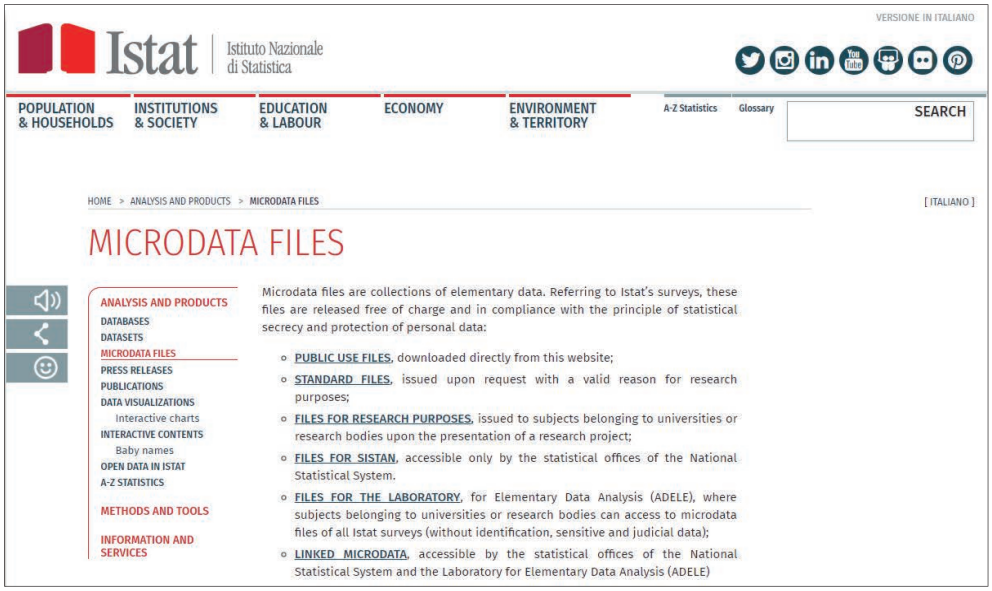
Frame-SBS file contains the twenty following variables:

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.



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.

A quality evaluation framework for the statistical register *Frame-SBS*

Orietta Luzi, Fabiana Rocci, Roberto Sanzo, Roberta Varriale [1]

Abstract

In 2013, Istat implemented the new statistical register "Frame-SBS" for the annual production of economic accounts statistics based on the integrated use of administrative and survey data, overcoming most of the limits of the traditional survey-based estimation strategy. The transition to a production strategy essentially based on the use of administrative data required the development of innovative methodological approaches, and determined the need of new tools for quality evaluation of both the data and the statistical process. In this paper we propose a first scheme of indicators for measuring and documenting the quality of the Frame-SBS. The final goal is to implement a quality control system to regularly monitor the register, by the identification of possible process and data weaknesses, and supporting quality improvements.

Keywords: Statistical Register, Administrative data, Quality.

1 Orietta Luzi (luzi@istat.it); Fabiana Rocci (rocci@istat.it); Roberto Sanzo (sanzo@istat.it); Roberta Varriale (varriale@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.

1. Introduction

In the last years, Istat has strongly increased the amount of administrative (hereafter *admin*) archives that are centrally acquired and used for statistical production purposes. Such an increase calls for a tailoring of the current approaches for quality measurement and assessment, in order to build a wider framework based on: the measurement of the quality of input sources, that are centrally acquired by Istat (Ambroselli *et al.*, 2014); designing proper tools to extend quality auditing to the statistical processes using admin data (Brancato *et al.*, 2014); measuring, monitoring and assessing the quality of any statistical process and product derived by using admin data, which is the main aim of this paper. One of the most common output based on an intensive use of admin data is the establishment of statistical registers, representing the transformation of the admin information for statistical purposes, according to the statistical definition of the target population and variable (Wallgren *et al.*, 2007).

This paper deals with the quality assessment of the statistical register *Frame-SBS*, (Luzi *et al.*, 2016; Luzi *et al.*, 2014) which is currently used at Istat for the annual estimation of Structural Business Statistics (hereafter *SBS*). Its implementation has been guaranteed by the use of a number of admin sources integrated in an appropriate strategy to survey data. Hence, the availability of stable, timely and reliable admin sources providing high quality and detailed information on enterprises' profit and loss accounts, has allowed since 2013 Istat to use a new estimation strategy. The *Frame-SBS* contains microdata for the main economic variables for all the enterprises in industry and services (excluding financial companies and insurance) with less than 100 persons employed which are active for more than six months in the reference year (about 4.4 million of units), for every SBS domain required by the European Regulation.

Therefore, based on the *Frame-SBS*, estimates for the main SBS can be computed at an extremely refined level of detail, overcoming some limitations of the previous estimation strategy. As a consequence, improvements have been achieved in terms of both accuracy of cross-sectional estimates and consistency of estimates over time and among related statistical domains, with particular reference to National Accounts. Concerning accuracy, however, it has been underlined that even if sampling error components have been essentially removed, additional sources of non-sampling error need to

be assessed due to the admin data characteristics and coverage and to the features of the integration process.

The present work focusses on the definition and the implementation of the quality framework to assess the *Frame-SBS* production process, starting from the framework proposed by Zhang (2012). First considerations concerning the suitability of the Zhang proposal with respect to the Istat experience are also reported. In particular, in the paper a first application of the proposed quality framework is reported, with the introduction of an additional step to better deal with the admin sources integration phase. The focus is on the main register variables, that are those which can be directly derived by the admin sources with a high level of quality and coverage (see Curatolo *et al.*, 2016).

The paper is structured as follows. Section 2 contains a description of the main characteristics of the *Frame-SBS* register. In Section 3, the proposed quality framework associated to the *Frame-SBS* production process is illustrated, and the corresponding list of quality indicators is proposed. Some results from the established framework system are also provided, to show how the monitoring is currently assessed year by year. In section 4 some concluding remarks are provided and the directions for further developments are delineated.

2. The *Frame-SBS*

In this section we describe the main features and the production process of the *Frame-SBS*.

The target population of the register consists of all the Italian small and medium enterprise (enterprises with less than 100 persons employed) in the industrial, construction, trade and non-financial services sectors (about 4.3 million of units) which are active for more than six months in the reference year, for every SBS domain required by the European Regulation. This population is completely identified by the Italian Business Register (BR) ASIA[2] (Istat, 2016) which contains structural and classification information on the Italian active enterprises. Actually, ASIA allows to identify all the potential theoretical sub-populations (e.g. by legal form) which could be investigated for statistical purposes.

The *main* target variables of the register are the profit and loss account variables as identified by the E.U. regulation:

- Revenues
  - Income from sales and services (Turnover)
  - Changes in stock of finished and semi-finished products
  - Changes in contract work in progress
  - Changes in internal work capitalised under fixed assets
  - Other income and earnings (neither financial, nor extraordinary)
- Costs
  - Purchases of goods
  - Purchases of services
  - Use of third party assets
  - Changes in stocks of raw materials and for resale

2 The Italian Business Register represents the official source on the structure of the business population and demography that identifies the Italian enterprises, and their statistical variables. Asia has the role of the frame list for all Istat business survey. It is also a reference to update structural information on enterprises (economic activity, persons employed, employees, etc.) and allows linking all the available administrative sources through the fiscal code.

- Other operating charges
- Personnel Costs.

In order to estimate the target variables[3], in *Frame-SBS* micro-data from different admin data sources available in the Italian information system are properly integrated. Such sources are currently acquired by Istat through a unique entry point ensuring a standardised and consistent management of the relationships with data owners. The sources are (Curatolo *et al.*, 2016):

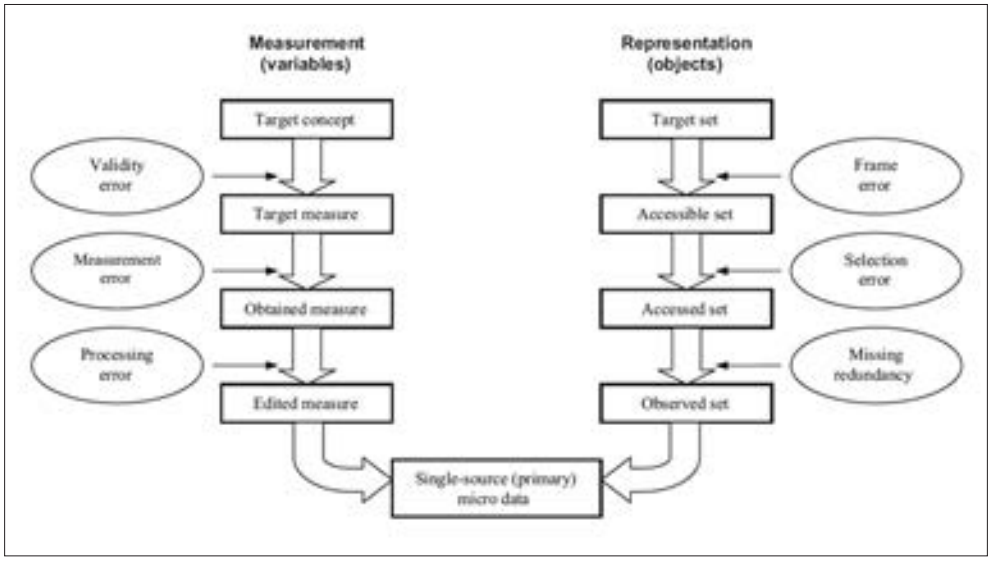
- Financial Statements (hereafter *FS*). FS are registered by the Italian Chambers of Commerce. Profit and loss account items of the financial statements are annually provided for limited liability companies (about 750,000 units). Variable definitions in FS, which are designed to check the balance sheet of corporate companies, are the closest to those required by SBS regulations. For this reason, this source plays a central role in the integration process described in the following;
- Sector Studies survey (hereafter *SS*). SS is a Fiscal Authority survey, including each year about 3.5 million of units, that aims at evaluating the capacity of enterprises to produce income and at indirectly assessing whether they pay taxes correctly. The units compiling the SS form, composed of detailed information on costs and income, are the enterprises with a turnover less than 7,500,000 Euros belonging to many activity sectors;
- Tax returns (hereafter *Modello Unico*). The Modello Unico data is provided by the Ministry of Economy and Finance, is based on a unified model of tax declarations by legal form and containins economic information for different legal forms for about 4.5 million of units each year;
- Regional Tax on Productive Activities (hereafter *Irap*). The Irap form is used to declare the regional tax on productive activities carried out by enterprises. It is filled regardless of the accounting system adopted and is composed of several sub-forms in accordance with the different type of the enterprises.

3 The variable personnel costs is always observed and it is used as auxiliary variable in the estimation process.

3. The proposed quality evaluation framework, first considerations based on *Frame-SBS* study

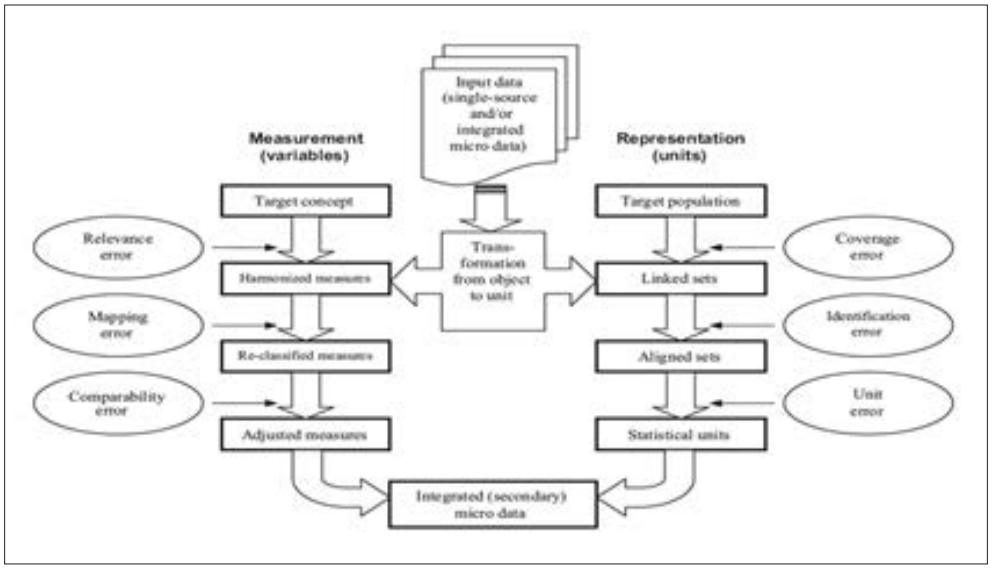
In order to assess the quality of the *Frame-SBS* data and, indirectly, of the statistics produced based on its data, we adapted the framework proposed by Zhang (2012), where a well-defined data processing scheme with the associated list of errors for the production of statistics based on the combination of various admin and statistical datasets is presented. The framework consists of two main phases, represented in the lifecycle diagrams reported in Figure 3.1 and Figure 3.2. The first phase, dealing with each single source, categorizes errors arising with respect to the original source's target population and concepts, in order to support the assessment of the quality of the source itself.

Figure 3.1 - Sources of error in phase one of Zhang's framework



Source: Zhang, 2012

Figure 3.2 - Sources of error in phase two of Zhang's framework



Source: Zhang, 2012

The second phase focusses on errors arising when data from several sources are combined to produce a statistical output. In this case, the aim is to measure the quality of the transformation process which is needed to adapt the data from their original purpose to the statistical one. Indeed, in this phase the targets correspond to the statistical population and to the statistical concepts to be measured. For more details see Zhang (2012) and Zabala (2013).

3.1 The *Frame-SBS* case study

In this paragraph the *Frame-SBS* production process is described. We start from the Zhang's framework, which is actually useful in order to clearly analyze the design of any mixed-source statistical process. The final aim is to understand the error sources potentially affecting the register's output data, that may result from the characteristics of each admin archive and/or from the design choices underlying the statistical production process.

Nevertheless, we propose to represent the process in three-phases: the first phase can be assimilated to the Zhang's phase one, while the Zhang's second phase has been split into two sub-phases in order to better distinguish the

specific steps of the transformation process the original data have to go through: in phase two the admin data are evaluated according to the SBS targets (both units and variables), however we define a first sub-phase (phase 2a), where each admin source is evaluated separately in order to determine the criteria according which to select the data and to combine them, and a second phase (phase 2b), where the integrated dataset is created and is further elaborated to attain the final register data.

Phase 1. Pre-treatment of admin sources. The first phase of the *Frame-SBS* production process consists of the pre-treatment of each admin source data. This phase is carried out separately for every source, covering each a different population and characterised by a peculiar structure and specific contents. Firstly, only the subset of items which are useful for deriving the target SBS are selected. On the objects side, for each admin source, the following actions are performed: verify if there are substantial changes over time in the population coverage and in the time of the source supply, identify and eliminate duplicated units or unacceptable information. On the measurements side, an initial assessment of formal data inconsistencies is carried out, based on the use of accounting rules (edits). At this stage, a proportion of FS units containing errors that cannot be resolved are discarded. The remaining errors are resolved by adopting a deterministic data imputation approach.

Phase 2a. Treatment of the admin sources, taking into account the SBS purposes. During phase 2a, the units belonging to the SBS population are selected from each source. Note that the statistical units in each source are identified at the archive acquisition stage from the external supplier, therefore units identification errors are not expected in the *Frame-SBS* production process. The admin (original) items of each source are harmonised w.r.t. the target SBS variables. The harmonisation process is a result of accurate preliminary analyses of admin data and their associated metadata, with the aim of comparing the economic contents derived from the admin items with the corresponding SBS definitions, as described by the SBS European regulation (Curatolo *et al.*, 2016). As it is not always possible to directly "reconcile" the admin and the statistical definitions, the admin information is used to obtain the harmonised variables, however a certain amount of information is

discarded and this causes a given amount of "item non response". Finally, the source coverage w.r.t. the target population is evaluated and the information content of the entries in the various admin sources are assessed. As a direct consequence of this assessment, different degrees of reliability are associated to the different admin sources, and a pre-defined priority is associated to each archive so that the best source is used for each target (sub)population in case of overlaps.

Phase 2b. Integration of the sources. In this phase, the final list of the units belonging to the target population is identified (based on the BR identification code) and a specific admin source is associated to each of them, following the predefined priority in case of concurrent (overlapping) sources. For each statistical unit all information from a single source (when available) is derived, to preserve the internal data consistency at unit level. There are some "exceptions to the priority", according to which the most reliable source is discarded and the source with next priority is used. For example, in case of inconsistencies resulting from the pre-treatment of each source (phase 1) that cannot be resolved. Another exception is based on the analysis of the *per capita* (per employee) labor cost of the enterprises, that when not coherent with auxiliary information available from the Istat Employee Wage Register (RACLI), may determine the selection of the units from the source with next priority. Once the above process is completed, an integrated dataset of target units and variables is determined. However, a certain amount of both under-coverage w.r.t. the SBS target population, and incompleteness w.r.t. SBS target variables remain, to be properly recovered. Therefore, after an editing activity aiming at identifying and treating possible outliers and influential errors, an imputation process to predict unit and item non-responses on the integrated data is performed (Di Zio *et al.*, 2016). A macro-editing strategy is used for the final cross-sectional and longitudinal validation of the final SBS estimates at the level of detail required by the Eurostat regulation.

For each phase of the *Frame-SBS* production process, in the Tables 3.1, 3.2 and 3.3 we propose a set of quality indicators consisting of both new measures and some adaptation of the indicators proposed by Zabala (2013). For each process phase, the indicators are presented by subject (variables, objects and units), process step and error type (as reported in Figure 3.1 and Figure 3.2).

Table 3.1 - Phase 1 quality indicators

Objects. Accessible Set -> Accessed Set; Selection error	
Proportion of *units* in FS w.r.t. the FS theoretical population in the BR	**[*No. units in the source/ Total no. units in the FS theoretical population in BR*] x 100**
Proportion of *units* in the source w.r.t. the BR population, by source (SS, Unico, Irap)	**[*No. units in the source/ Total No. units in BR*] x 100**
Adherence to reporting period, for FS	***[No. units that do not adhere to the reporting period/Total No. units]* x 100**
Qualitative indicators, by source (SS, Unico, Irap)	*Changes in population coverage (Does coverage change over time?)* *Updating of reporting units (How are changes recorded and actioned? Is it proactive or reactive?)*
Objects. Accessed Set -> Observed Set; Missing/Redundancy error	
Percentage of multiple records, by source	***[No. units in Source S with multiple id code / No. of unique identification codes] x 100***
Qualitative indicators	*Detecting duplicate records (Describe how duplicate reporting units are identified)* *Methods of treating duplicate records (Describe how duplicate reporting units are handled)*
Variables. Process step: Target Measure -> Obtained Measure; Type of error: Measurement error	
Punctuality, by source	***[Date of receipt - date agreed]***
Lagged time between reference period and receipt of data	***[Date of receipt by Istat-Date of the end of the reference period over which the data provider reports]***
Qualitative indicators, by source	*Changes in administrative forms*
Variables. Obtained Measure -> Edited Measure; Processing error	
Proportion of *units* failing edit checks, by source	***[No. units failing edit checks/ Total no. of units checked]* x 100**
Proportion of *units* with all implausible values, by source	***[No. units whose values are all missing, or all values are equal to 0, or all values are equal to 1 / Total no. of units checked]* x 100**
Proportion of *units* with all missing values, by source	***[No. units with all values missing/ Total n. of units checked]* x 100**
Proportion of edit rules failed at least once, by source	***[No. of failed edit rules for source S/ Total no. of edit rules for source S]* x 100**
Proportion of imputed *values*, by source	***[Total no. of imputed values in source S / Total no. of values in source S]* x 100**
Composition of the proportion of imputed *values*, by source	***Modification rate: [Total no. of values changed from a code to another code in source S / Total no. of imputed values in source S] x 100*** ***Net imputation rate: [Total no. of values changed from missing or 0 to a code in source S / Total no. of imputed values in source S] x 100*** ***Cancellation rate: [Total no. of values changed from a code to 0 in source S / Total no. of imputed values in source S] x 100***

The proposed indicators include both quantitative and qualitative measures. Actually, for some types of errors (e.g. *Measurement errors* in phase 1, Relevance errors and *Mapping errors* in phase 2a), the description of the conceptual schemes developed provides key information for the assessment of the quality of the production process. The indicators proposed for phases 1 and 2a are typical of all statistical processes based on the integrated use of admin data. The most part of indicators proposed for variables in phase 2b, on the other hand, are similar to measures which are typically used to assess the quality of data collected by direct surveys.

Table 3.2 - Phase 2a quality indicators

Units. Target Population -> Linked Sets; Coverage error	
Proportion of *units* in the FS source w.r.t. the SBS sub-population of corporate companies	***[No. corporate companies of SBS pop. in source FS/ No. of corporate companies of the SBS pop.]* x 100**
Proportion of *units* in the source w.r.t. the SBS population, by source (SS, Unico, Irap)	***[No. units of SBS population in source S / No. of units of SBS population]* x 100**
Variables. Target Concept -> Harmonised Measures; Relevance error	
Qualitative indicators, by source	*Changes in definitions of all variables in each source and changes in definitions of SBS variables (Does definitions change over time?)* *Conceptual scheme representing the re-classification of administrative concepts needed to produce the SBS variable definitions*
Variables. Harmonised Measures -> Re-classified Measures; Mapping error	
Quantitative indicators, by source	*Comparison of each harmonised variable with SBS benchmark variable (histograms, univariate statistics, statistical tests, etc.), to be repeated when variable definitions change*
Proportion of target variables which not require reclassification or mapping, by source	***[No. variables captured directly from source S / Tot. no. variables]* x 100**
Proportion of target variables which can be derived through reclassification or mapping, by source	***[No. variables derived from source S after reclassification/ Tot. no. variables]* x 100**

Table 3.3 - Phase 2b quality indicators

Units. Target Population -> Linked Sets; Coverage error	
Proportion of *units* of the SBS population in the integrated dataset (coverage). Also in longitudinal perspective.	***[No. of units of SBS pop. in the integrated dataset / No. of units of SBS pop.]* x 100**
Proportion of *units* of the SBS population in the integrated dataset, by source S.	***[No. of units of SBS pop. in the integrated dataset from source S/ No. of units of SBS pop.]* x 100**
Proportion of *units* of the SBS population in the integrated dataset with information present in only one source	***[No. of units of SBS pop. in only one source / No. of units of SBS pop. in at least one source]* x 100**
Proportion of *units* of the SBS population in the integrated dataset with information available in more than one source	***[No. units of SBS pop. in more than one source / No. of units of SBS pop. in at least one source]* x 100**
Variables. Re-classified Measures -> Adjusted Measure; Comparability error	
Proportion of *units* with influential values, by variable	***[No. of units with influential errors / Total no. of units]* x 100**
Proportion of outliers, by variable	***[No. of outliers / Total no. of units]* x 100**
Proportion of *units* with at least one imputed value	***[No. of units with at least one imputed value / Total no. of units]* x 100**
Proportion of *units* failing at least one edit rule	***[No. of units failing edit checks / Total no. of units checked]* x 100**
Proportion of *variable values* imputed, by variable	***[No. of units with imputed values for variable Y / Total no. of unit]* x 100**
Composition of the proportion of imputed *variable values*, by variable	***Modification rate: [Total no. of values of the variable Y changed from a code to another code in source S / Total no. of imputed values of variable Y]* x 100** ***Net imputation rate: [Total no. of values of the variable Y changed from missing or 0 to a code / Total no. imputed values of variable Y] x 100*** ***Cancellation rate: [Total no. values of the variable Y changed from a code to 0 /Total no. of imputed values of variable Y] x 100***
Impact of data editing and imputation on microdata, by variable	***Simple and quadratic distance between pre-edited (Y) and post-edited (Y*) values of variable Y*** $DL_1(Y_i, Y_i^*) = S^N_{i=1} \lvert Y_i - Y_i^* \rvert$ */Total no. of units N;* $DL_2(Y_i, Y_i^*) = \ddot{O}S^N_{i=1} (Y_i - Y_i^*)^2$ */Total no. of units* N_i
Impact of data editing and imputation on distributions, by variable	*Kolmogorov-Smirnov distance on pre-edited and post-edited distributions* *Comparison of variable distributions (univariate statistics, etc.) pre- and post- editing and imputation*
Impact of data editing and imputation on statistical relations	*Pearson correlation index, Covariance matrix between variables*
Impact of data editing and imputation on aggregates, by variable	***[Variable total before editing and imputation / Variable total after editing and imputation]* x 100**

3.2 Selected results

In this section, we provide an example of how the quality indicators included in the proposed evaluation framework can be used for the analysis of the *Frame-SBS* inputs, data processing and outputs. It is straightforward to mention that the availability of the indicators values for subsequent years allow longitudinal analyses in order to monitor the changes of the quality of both input and output data.

In Tables 3.4, 3.5 and 3.6 the values of a selected set of qualitative measures are reported for three reference years (2012, 2013 and 2014), for the designed phases of the *Frame-SBS* production process.

As it can be seen in Table 3.4, referring to *Objects: Selection error*, Unico is the archive with the lowest under-coverage rate w.r.t. its corresponding theoretical population. In particular, the under-coverage of FS w.r.t. its theoretical population (the Italian *corporate companies*) is essentially due to delays in the delivery of information to the Italian Chamber of Commerce by some of the enterprises, and to the fact that some deadlines for enterprises to supply their data are not compatible with the production of the register.

Concerning *Variables*, the proposed indicators relate to validation rules which identify within-records data inconsistencies with respect to the specific admin data coherence requirements. Note that a different number of rules has been defined to check the formal accuracy of data in the used sources[4]. From Table 3.4 it can be viewed that FS and SS have the highest quality in terms of proportions of units with all missing or implausible values. However, FS is the archive with the highest rate of edit rules failed at least once, while SS is the source with highest quality w.r.t. formal accounting rules. Very low proportions of imputed values result for all the sources involved in the production process, as imputation is performed on data after the elimination of the admin units containing unusable information (units with all missing values and units with all implausible values – missing, zero and 1).

4 FS: 29 edit rules defined (23 failed rules at least once in 2013); SS: 108 edit rules defined (40 rules failed at least once in 2013); Unico: 178 edit rules defined (52 rules failed at least once in 2013); Irap: 124 edit rules defined (26 rules failed at least once in 2013).

Table 3.4 - Phase 1, quality indicators by subject and error type. Years 2012, 2013 and 2014

INDICATOR	Year		
	2012	2013	2014
Objects. Selection error			
Proportion of units not in the source w.r.t. the theoretical population, by source			
FS	8.43	10.55	11.39
SS	12.80	12.55	10.60
Unico	4.48	5.52	6.39
Irap	22.62	22.17	26.00
Objects. Missing/Redundancy error			
Percentage of multiple records, by source			
FS	0.01	0.01	0.11
SS	0.00	0.00	0.00
Unico	2.24	2.13	2.03
Irap	2.23	1.21	0.95
Variables. Processing errors			
Proportion of units failing edit checks, by source			
FS	6.41	6.30	4.35
SS	0.01	0.00	0.00
Unico	18.46	0.68	0.59
Irap	0.01	10.88	10.56
Proportion of units with all missing values, by source			
FS	0.00	0.00	0.00
SS	0.00	0.00	0.00
Unico	1.42	16.57	0.01
Irap	1.19	12.55	0.03
Proportion of units with all implausible values, by source			
FS	0.01	0.01	0.01
SS	0.19	0.26	0.28
Unico	0.10	0.38	0.38
Irap	0.00	0.52	0.47
Proportion of edit rules failed at least once, by source			
FS	79.31	79.31	79.31
SS	-	37.04	40.74
Unico	-	29.21	28.73
Irap	0.01	20.97	15.45
Proportion of imputed values, by source			
FS	0.15	0.14	0.33
SS	0.25	0.00	0.00
Unico	0.41	0.01	0.01
Irap	0.00	0.40	0.39

Concerning phase 2a (Table 3.5), it results a high coverage rate of FS w.r.t. the SBS sub-population of corporate companies, as well as high coverage rates of the other sources SS, Unico and Irap w.r.t. the SBS target population.

Relating to variables, the *Variables: Mapping error* indicator is provided w.r.t. 20 key SBS, and taking into account that some of the sources are structured in multiple forms[5] corresponding to different classes of enterprises (e.g. different business legal forms) and providing different information accordingly (see Curatolo *et al.*, 2016, for more details). It is evident from the *mapping error* indicator that FS is the best harmonised source in terms of variables definitions w.r.t. the SBS estimation purposes. As expected, this indicator does not vary in the considered period: the variables definitions adopted for admin purposes did not change.

Table 3.5 - Phase 2a quality indicators by subject and error type. Years 2012, 2013 and 2014

INDICATOR	Year		
	2012	2013	2014
Units. Coverage error			
Proportion of units in the FS source w.r.t. the SBS sub-population of corporate companies	90.84	90.57	89.81
Proportion units in the source w.r.t. the SBS population, by source			
SS	79.99	80.84	80.11
Unico	77.57	78.30	74.54
Irap	95.42	94.76	93.91
Variables. Mapping errors			
Proportion of target variables which not require reclassification or mapping, by source			
FS	100.0	100.0	100.0
SS	86.0-90.0	86.0-90.0	86.0-90.0
Unico	6.0-73.0	6.0-73.0	6.0-73.0
Irap	25.0-80.0	25.0-80.0	25.0-80.0

Concerning phase 2b, in Table 3.6 the values of indicators on *coverage* in the integrated dataset are provided. Overall, the coverage of the target SBS population is about 97%, with the most part of information for each unit available from more than one source (about 93%). It has to be reminded that in the *Frame-SBS* the sources are used with a pre-defined priority, based on a preliminary assessment of the different quality levels of their information (see Curatolo *et al.*, 2016 for more details).

As it can be seen, SS is the source with the highest contribution in terms of proportion of units in the integrated dataset.

5 SS involves 2 different forms; Unico involves 8 different forms; Irap involves 8 different forms.

Table 3.6 - Phase 2b quality indicators by subject and error type. Years 2012, 2013 and 2014

INDICATOR	Year		
	2012	2013	2014
Units. Target Population -> Linked Sets; Coverage error			
Proportion of missing units of the SBS population in the integrated dataset (under-coverage)	2.50	2.63	3.76
Proportion of units of the SBS population in the integrated dataset, by source			
FS	16.17	16.87	16.88
SS	67.26	67.67	67.05
Unico	12.26	10.80	11.07
Irap	1.80	2.03	1.23
Variables. Re-classified Measures -> Adjusted Measure; Comparability error			
Proportion of units with at least one imputed value	19.95	19.05	24.59
Proportion of variable values imputed, by variable			
Revenues	2.78	2.74	7.84
Purchases goods&services	13.44	12.88	16.44
Value Added	10.96	10.56	9.68
Modification rate, by variable			
Revenues	0.00	0.00	3.95
Purchases goods&services	5.25	6.01	6.01
Value Added	8.20	7.72	5.72
Net imputation rate, by variable			
Revenues	2.78	2.74	3.89
Purchases goods&services	8.19	6.87	10.37
Value Added	2.75	2.84	3.97
Cancellation rate, by variable			
Revenues	0.00	0.00	0.00
Purchases goods&services	0.00	0.00	0.00
Value Added	0.00	0.00	0.00
DL_1 (Impact of data editing and imputation on microdata), by variable			
Revenues	10,377	8,781	16,339
Purchases goods&services	8,402	7,954	13,194
Value Added	4,236	4,063	5,432
DL_2 (Impact of data editing and imputation on microdata), by variable			
Revenues	592,973	482,945	2,497,389
Purchases goods&services	449,541	431,047	1,652,552
Value Added	294,411	299,485	550,086
Kolmogorov-Smirnov Index (Impact of data editing and imputation on distributions), by variable			
Revenues	0.03	0.03	0.04
Purchases goods&services	0.08	0.07	0.10
Value Added	0.03	0.03	0.04
Impact of data editing and imputation on aggregates, by variable			
Revenues	102.70	102.30	104.30
Purchases goods&services	102.60	102.50	104.50
Value Added	102.30	102.40	103.90

4. Conclusions and future work

In this paper a comprehensive framework for the quality assessment of the statistical register *Frame-SBS* on enterprises accounts is proposed. In the definition of the framework, an effort has been made to adapt the proposals from Zhang (2012) and Zabala (2013) to the peculiarities of the register production process, in order to identify the actual sources of errors by using appropriate quality measures on both the variables and the objects/units sides. In fact, the identification of the error sources represents the basis for the systematic and continuous improvement of the production process quality through their elimination (or at least the reduction) in the subsequent replications of the production process of the register. Furthermore, the availability of such indicators for different reference years allow the analysis of both data and process quality in a longitudinal perspective. In addition, based on the proposed framework, a complete quality report could be developed for documentation and dissemination purposes.

Concerning future work, it has to be remarked that this proposal has to be considered as an initial step of a complex project. An in depth analysis of the proposed set of indicators is necessary in order to fine tune it, in order to eliminate the possible redundancies and potentially add new indicators. Concerning the latter aspect, additional quality measures could be defined as a consequence of the possible extension of the admin sources used and the detection of further sources of error. Furthermore, as imputation models are used in phase 2b to compensate for not available information, an evaluation of their impact on final estimates should be provided, e.g. by adopting iterative procedures (based e.g. on bootstrapping or on multiple imputation) to measure the additional uncertainty due to the imputation process, under appropriate assumptions on the missing data mechanisms.

It has to be underlined that the proposed framework needs to be harmonised with the tools under development at Istat for the quality documentation of all the admin sources acquired from external suppliers: in particular, the indicators on "input data" included in the so called *source quality card* associated to each admin archive acquired by Istat, have to be properly incorporated in the quality evaluation framework proposed in this paper.

Finally, a relevant development relates to the need of identifying appropriate combinations of the proposed quality indicators (e.g. by using composite indicators) in order to have a complete representation of the overall quality of the register data and of its production process.

References

Ambroselli, S., and G. Di Bella. 2014. "Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat". *European Conference on Quality in Official Statistics (Q2014)*. Wien, 2-5 June 2014.

Brancato, G., A. Boggia, F. Barbalace, and C. Busetti. 2014. "Quality Guidelines for statistical processes using administrative data". *European Conference on Quality in Official Statistics (Q2014)*. Wien, 2-5 June 2014.

Curatolo, S., V. De Giorgi, F. Oropallo, A. Puggioni, and G. Siesto. 2016. "Quality analysis and harmonisation issues in the context of the *Frame-SBS"*. *Rivista di Statistica Ufficiale*, N. 1/2016.

Di Zio, M., U. Guarnera, and R. Varriale. 2016. "Estimation of the main variables of the economic account of small and medium enterprises based on administrative sources". *Rivista di Statistica Ufficiale*, N. 1/2016.

Istituto Nazionale di Statistica - Istat. 2016. "Il Registro statistico Asia-Imprese". *Nota metodologica*. Roma: Istat. https://www.istat.it/it/files//2016/06/Nota-metodologica-1.pdf

Luzi, O., and R. Monducci. 2016. "The new statistical register *Frame-SBS*: overview and perspectives". *Rivista di Statistica Ufficiale*, N. 1/2016.

Luzi, O., U. Guarnera, and P. Righi. 2014. "The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data". *European Conference on Quality in Official Statistics (Q2014)*. Wien, 2-5 June 2014.

Wallgren, A., and B. Wallgren. 2007. *Register-based Statistics: statistical methods for administrative data*. Chichester, U.K.: John Wiley & Sons Ltd.

Zabala, F., G. Reid, J. Gudgeon, and M. Feyen. 2013. "Quality Measures for Statistical Outputs using Administrative Data". *Statistical Methods*. Statistics New Zealand.

Zhang, L.-C. 2012. "Topics of statistical theory for register-based statistics and data integration". *Statistica Neerlandica*. Volume 66, Issue 1: 41-63. Hoboken, NJ, U.S.: John Wiley & Sons, Inc.

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici e ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti per il perseguimento degli obiettivi della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca Istat". Nel 1999 la collana viene affidata a un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna a essere editore in proprio della pubblicazione.

1B012016003000000

ISSN 1828-1982

€ 15,00