**The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data**

Orietta LUZI, Ugo GUARNERA and Paolo RIGHI

ISTAT - Italian National Statistical Institute

**Abstract**

In Istat, Structural Business Statistics for small and medium enterprises are traditionally based on sample surveys. The increasing availability of stable and timely information from administrative sources has made it possible to use them as a primary source for information to produce SBS, overcoming their use as essentially auxiliary information to treat non-response in survey data and to calibrate the estimates. The paper illustrates the main characteristics and methodological features of the new statistical information system (called frame) based on the integrated use of administrative archives recently produced at Istat. In the paper the overall production process of the frame, and the methodological solutions adopted in the new information context are illustrated.

**Key words:** administrative data, structural business statistics, imputation, projection estimator

1. **Introduction**

Increasing the efficiency and consistency of statistical production processes is an issue in the area of Structural Business Statistics (hereafter SBS) mainly due to the facts that the Italian economic system is characterized by a large amount enterprises (about 4.5 million of units), and a high level of detail is required by the current SBS Regulations. Actually, Istat has been working from several years to the reduction of SBS production costs and burden on enterprises by exploiting administrative and fiscal data sources (hereafter AD), especially on small and medium enterprises (see for example [1]; [2]; [5]). In this specific context, in 2013 the Istat Department for National Accounts and Economic Statistics has achieved a strategic result, consisting in the completion of a firm-level integrated Statistical Data Warehouse (“frame” in the following) for small and medium enterprises based on AD as primary sources of information, integrated with direct survey data to estimate information not available in AD sources.

Based on the frame, starting from the 2011 reference year onwards, estimates for key SBS can be calculated at an extremely refined level of detail, overcoming some limitations of the current statistical production strategy, which is essentially based on direct surveys integrated with AD for the non-respondents. Improvements are expected in terms of efficiency, accuracy of cross-sectional estimates within and across statistical domains, coherence of estimates over time. The dissemination of larger, more detailed and better focused data to end users will be also facilitated: actually, the availability of micro-data values allows providing coherent estimates at different levels of aggregation, taking into account that these data are also used by National Accounts (NA).

This is result is in line with the requirements emerged at European level, where European projects like BLUE-ETS[[1]](#footnote-1), the ESSnet Admin Data[[2]](#footnote-2), the ESSnet on Data Integration[[3]](#footnote-3) have provided a methodological framework for quality evaluation and estimation when using AD in the area of economic statistics. High levels of consistency, efficiency, information detail will be also required by the new FRIBS (Framework Regulation In Business Statistics) to be achieved by 2015.

This paper provides an overview of the solutions adopted for producing the *frame*, with specific attention to concepts and definitions harmonization, to the assessment of quality and usability of AD w.r.t. the SBS estimation purposes, to the estimation strategy (see for example [8] and [9] for a general discussion of these issues).

The paper is structured as follows. Section 1 illustrates the preliminary data validation and harmonization of the sources’ information contents Section 2 contains an overview of the activities carried out to ensure the usability of the AD sources, while in Section 3 the quality evaluation strategy adopted to ensure the statistical usability of the sources is illustrated. In Section 4 the estimation strategy is described. Section 5 contains some concluding remarks.

1. **The sources and the preliminary data treatment**

Currently, SBS on SMEs (enterprises with less than 99 persons employed, about 4.4 million of units) are obtained based on an annual sampling survey which collects information on profit and loss accounts, as well as on employment, investments etc. on about 100,000 enterprises in the industrial, construction, trade and non-financial services sectors. A large number of secondary variables are also estimated, also for NA estimation purposes. The target population is identified based on the Italian Business Register (hereafter BR), counting about 4.5 million enterprises. Recently, a new version of the BR has been produced, where information on enterprises has been extended with particular reference to structural variables on employment. Very detailed estimation domains are required from the SBS Regulations[[4]](#footnote-4). The survey response rate is generally close to 40% (varying according to size classes and economic activity sectors) in terms of reliable replies.

At the same time, Istat annually gets a number of AD sources containing information on enterprises’ profit and loss accounts. They have different degrees of coverage of the SMEs population and different characteristics concerning variables:

* the Financial Statements (hereafter FS), from the Chamber of Commerce, which annually provides profit and loss statements of limited liability companies (about 750,000 units);
* the Sector Studies survey (hereafter SS), which is a fiscal survey including each year about 3.5 million SMEs with the *Turnover* in the interval [30,000-7,5 million] euros[[5]](#footnote-5);
* the Tax returns form data (hereafter Unico), from the Ministry of Economy and Finance, which is based on a unified model of tax declarations by legal form, containing economic information for different legal forms for about 4.5 million of units each year;
* the Social Security Data (hereafter SSD), from the National Security Institute, which includes firm data and individual (employees) data on occupation and labour cost for all enterprises.

Assessing the feasibility of a new SBS estimation process based on the (integrated) use of the above sources has implied a preliminary data analysis and validation consisting, for each archive, in the following activities, taking into account the specific SBS estimation purposes:

* evaluation of the source usability in terms of timeliness, stability, coverage;
* editing the source data (identification of errors, duplicated units, not coherent data, etc.);
* harmonization of variables definitions;
* evaluation of the *statistical* usability (quality) of the source.

In terms of timeliness and punctuality, each source is regular and stable over time: the possibility of delays in the data provision, and the possible changes in the sources’ contents have been considered of limited effects on the overall SBS production process especially once an appropriate process and data monitoring service is established with the external Authorities.

In terms of units coverage, each source covers (partially overlapping) sub-sets of the SME population: overall, more than 95% of SMEs is covered by the considered sources. Concerning coverage of variables, not all the SBS variables can be directly derived from each source, while some sources provide information on common key SBS items.

The preliminary editing of sources’ data was necessary in order to detect measurement errors, duplication of units, and incoherent intra-unit data. It has to be remarked that in each source, the statistical units (the enterprises) are identified based on a complex procedure performed at the BR construction stage, together with their structural characteristics (e.g. economic activity, size, localization). However, despite the unique ID code assigned to each enterprise in each source, different amounts of duplicated units in each archive have been identified. Automatic corrections were also performed, whenever possible, in case of not coherent firm data. A number of units were discarded in case of too complex data inconsistencies.

A critical phase consisted in the first harmonization of the administrative contents w.r.t. the SBS definitions described by the SBS regulation. It has to be noted that it was not always possible to “reconcile” administrative and statistical definitions. As a consequence, some source information was either discarded or used as “auxiliary information” at the data estimation stage. A specific case was represented by variable *Personnel Cost*, which is available from the SSD: this variable was used as auxiliary information to “adjust” information on *Personnel Cost* drawn from the other sources, to better approximate SBS definitions.

For the reference year 2011, the information framework determined by the units and variables coverage of (partially overlapping) sources is illustrated in Figure 1. As it can be seen, both BR and SSD provide information on the overall population: the BR contains enterprises’ structural information, like *Economic Activity* (*Ateco*), *Number of Employees* (*NEmp*), and a proxy of *Turnover* (*TBR*), while SSD provides data on *NEmp*, *Personnel Cost* (*PC*), *Wages and Salaries* (*WS*), *Worked Hours* (WH) and *Social Contributions* (*SC*) for all the enterprises with at least one employee: as said above, *PC* serves as auxiliary information in the harmonization phase, while *WS*, *WH* and *SC* are directly used in the estimation process (see section 4).

1. **Quality evaluation and variables selection**

Let Y*p\** be the subset of target variable which can be potentially estimated by directly using AD. All the *p* variables could be estimated based on the survey data: let YjS (*j*=1,..,*p*) indicate the survey variables. On the contrary, the three sources contain useful information only for (different) sub-sets of Y*p\**: let (Y1i,…,Yki) indicate the variables which could be estimated by using source *i*, where Y*ki*⊂ Y*p\** for each *i*=1,2,3. However, the three sub-sets of variables are partly overlapping, as a number of key items could be estimated from more than one source.

For the above reasons, it was necessary to evaluate the statistical adequacy of each Y*ji* for estimating Y*j\**, thus identifying the “best source” for each item and “prioritizing” the sources in case of overlaps. The evaluation was based on the comparison between each Y*ji* and the corresponding YjS on the common set of survey respondents, assuming that survey values represent the best available measures of Y*j\** as they are collected based on the correct SBS definitions .

*Figure 1: Unit and variables coverage by source. Year 2011*



Let *du*=(Y*uji* – Y*uj*S)/Y*uj*S be the relative difference between variables’ values Y*ji* and Y*j*S in unit *u* (*u=1,…, nir*; *nir*=number of survey respondents in source *i*). Quality indicators include the rate of units having *du* in the interval ±5%, the mean, the median and the interquartile value (IQR *du*) of the *du* distribution; the coefficient of variation (Cv *du*), the *Kolmogorov-Smirnov* indicator (KS) . In Table 1 an example of indicators’ values when comparing survey data with SS data for some target variables (25.109 linked units) are reported. Similar comparisons were performed for each Y*ji* in each source *i*,by variable and data domain (e.g. *economic activity* and *size class*) (see [7]).

Analyses highlighted that, despite the similarity of definitions ensured by the harmonization step, some of the Y*ji* systematically diverged from the corresponding Y*j\** (mainly due to the different behavior of enterprises when providing information for fiscal purposes, or to the fact that some items could be of lower interest for the data owner so maybe they are not checked at all). Actually, from the evaluation study it resulted that:

* some Y*j\** had a good level of coverage (*main economic aggregates*), then could be estimated by directly using Y*ji* information from the different sources on different SME sub-populations, with the following priority: *FS* (which resulted the best harmonized source w.r.t. SBS Regulation definitions, supplying the largest amount of good quality information), SDS and finally Unico. These variables, however, contain “partially” missing values, due to the lack of information in some sources: these values which will be considered as *item non responses*.
* some Y*j\** could not be directly estimated based on AD source, as the level of coverage was considered as not adequate (*components of the main economic variables*);

*Table 1: Quality indicators by variable (SS vs Sample Survey)*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| VARIABLES | KS | *d*u± 5% (% units) | Mean *d*u (000€) | Median *d*u (000€) | IQR *d*u (000€) | Cv *d*u |
| **Income from sales and Services (Turnover)** | **1,0** | **90,8** |  | **-0,1** | **0,0** | **0,0** | **-1704,4** |
| **Changes in internal work capitalized under fixed assets**  | **0,5** | **98,9** |  | **-0,9** | **0,0** | **0,0** | **-84,2** |
| **Changes in contract work in progress** | **1,0** | **95,0** |  | **-1,7** | **0,0** | **0,0** | **-104,4** |
| Changes in stocks of finished and semi-finished products | 1,8 | 82,1 |  | -0,8 | 0,0 | 0,0 | -236,5 |
| Other income and earnings (neither financial, nor extraordinary) | 12,1 | 56,5 |  | -1,3 | 0,0 | 0,0 | -108,4 |
| Purchases of Goods | 12,3 | 58,7 |  | -3,7 | 0,0 | 5,1 | -63,7 |
| Purchases of Services | 4,2 | 29,0 |  | 0,2 | 0,0 | 12,8 | 1001,2 |
| **Purchases goods/services** | **1,0** | **58,0** |  | **-3,5** | **-0,6** | **9,6** | **-51,7** |
| Use of third party assets | 4,6 | 77,5 |  | 0,3 | 0,0 | 0,0 | 115,7 |
| Other operating charges | 13,9 | 11,9 |  | -1,2 | 0,0 | 6,9 | -84,8 |
| **Labor Cost** | **4,6** | **82,8** |  | **1,4** | **0,0** | **0,0** | **35,8** |
| Amortization | 3,0 | 64,8 |  | -3,8 | 0,0 | 0,7 | -12,4 |
| **Value Added**  | **1,7** | **46,9** |  | **-0,3** | **1,3** | **12,3** | **-610,0** |
| Gross Operating Margin | 3,0 | 35,2 |  | -1,7 | 1,1 | 11,0 | -102,1 |
| Net Operating Margin | 4,7 | 31,1 |  | 3,7 | 1,3 | 13,2 | 52,1 |

Table 2 contains the SME coverage by source of the resulting integrated database, number of units, number of employees, revenues, and value added. As it can be seen, about 96% of enterprises and of employees are covered: the SS is the most relevant administrative source in terms of population units coverage (64%), while FS covers more than 66% of total population revenues. The *not covered units* will be considered as *unit non-responses*. The presence of both unit and item non responses has driven the design of the estimation strategy described in section 4.

*Table 2: Final units coverage of the frame by source (% values)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **Number Units** | **Number Employees** | **Revenues** | **Value Added** |
| *FS* | 16.1 | 38.2 | 66.2 | 54.1 |
| *SS* | 64.0 | 49.2 | 24.5 | 36.4 |
| *Unico* | 16.2 | 8.3 | 5.5 | 6.1 |
| ***Total covered*** | **96.3** | **95.7** | **96.2** | **96.6** |
| ***Not covered*** | **3.7** | **4.3** | **3.7** | **3.4** |
| ***Total*** |  ***100.0*** |  ***100.0*** |  ***100.0*** |  ***100.0*** |

Before estimation, the integrated database was edited in order to identify possible influential errors: based on a model-based selective editing approach ([4]), we identified the influential units on the estimates of the population total (*Ty*) of *Value added* and *Intermediate costs* by economic activity (98 branches of economic activity). Auxiliary information used in models were *Turnover, Number of employees*, *Personnel cost* (for enterprises with at least one employee). Fixed at level 0.05 the estimated error remaining in data after selective editing, 1,685 critical units were selected (about 0.04% of the total edited units) for interactive analysis: the manual inspection has determined in some cases the manual data correction, in other cases the acceptance of the values, in other cases the cancellation of the values and their automatic imputation.

1. **The estimation strategy**

In order to obtain the SBS estimates, a hybrid procedure has been adopted: a model based (predictive) approach for estimating the *main economic aggregates*, and a design based (model assisted) approach for the *components of the main economic aggregates*.

* 1. *Imputation of the main economic aggregates*

For the main target variables, the predictive approach naturally leads to build a microdata file where the not available information is predicted (imputed) based on the observed data. When some item is available from more than one data source, the corresponding value is chosen from the “more liable” source, according to the hierarchy illustrated in the previous section. In Table 3 the variables treated in the first phase are listed together with the corresponding labels. It is worthwhile to remind that the variable *PC* (*Personnel Costs*) is always available, thus it is used as an auxiliary variable in the imputation procedure. Moreover, in some cases, linear combinations of the variables in Table 3, instead of their own values, are observed.

*Table 3 – Variables treated in the first estimation phase*

|  |  |  |
| --- | --- | --- |
| **Section** | **Label** | **Variable** |
| Revenues | *Y1* | Income from sales and Services (Turnover) |
| *Y2* | Changes in stocks of finished and semi-finished products |
| *Y3* | Changes in contract work in progress |
| *Y4* | Changes in internal work capitalized under fixed assets |
| *Y5* | Other income and earnings (neither financial, nor extraordinary) |
|  |  |  |
| Costs | *Y6* | Purchases of goods |
| *Y7* | Purchases of services |
| *Y8* | Use of third party assets |
| *Y9* | Changes in stocks of raw materials and for resale |
| *Y10* | Other operating charges |
| ***PC*** | Personnel Costs |

Specifically, for some sources we have considered the values of the “derived variable” CS=*Y2*-*Y9* (*Total Change in Stocks*), GS=*Y6*+*Y7* (*Purchases of Goods and Services*) and IC=GS+*Y8*+*Y10* (*Total Intermediate Costs*). The resulting informative framework is reported in Table 4. For each source (row) and each variable (column), observed and missing values are represented by the symbols “X” and “?” respectively. Note that for SS and Unico more than one raw is indicated, corresponding each to a different legal form. The coverage of each source is reported in the last column.

*Table 4: Coverage of the main economic aggregates by administrative source*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Source*** | ***Y1*** | ***Y2*** | ***Y3*** | ***Y4*** | ***Y5*** | ***Y6*** | ***Y7*** | ***Y8*** | ***Y9*** | ***Y10*** | ***PC*** | ***GS*** | ***CS*** | ***IC*** | ***Coverage (%)*** |
| *FS* | X | X | X | X | X | X | X | X | X | X | X | X | X | X | 16.13 |
| *SDS-F* | X | ? | X | X | X | X | X | X | ? | X | X | X | X | X | 51.08 |
| *SDS-G* | X | ? | X | X | X | ? | ? | ? | X | X | X | ? | X | X | 13.00 |
| *Unico1* | X | X | X | X | X | ? | ? | ? | X | ? | X | X | X | ? | 0.78 |
| *Unico2* | X | X | X | X | X | ? | ? | X | X | ? | X | X | X | ? | 0.04 |
| *Unico3* | X | ? | X | X | X | ? | ? | ? | ? | ? | X | X | X | ? | 2.73 |
| *Unico4* | X | ? | X | X | X | X | X | ? | ? | ? | X | X | X | ? | 0.76 |
| *Unico5* | X | X | X | X | X | ? | ? | ? | X | X | X | X | ? | X | 10.86 |
| *Unico6* | X | ? | ? | ? | ? | ? | ? | ? | ? | ? | X | ? | ? | ? | 0.16 |
| *Unico7* | X | ? | ? | ? | ? | ? | ? | ? | ? | ? | X | ? | ? | ? | 0.31 |
| *Unico8* | X | ? | ? | ? | ? | ? | ? | ? | ? | ? | X | ? | ? | ? | 0.49 |
| *Unico9* | X | X | X | X | X | X | X | X | X | X | X | X | X | X | 0.09 |
| *NA* | ? | ? | ? | ? | ? | ? | ? | ? | ? | ? | X | ? | ? | ? | 3.67 |

We also note that the values of all the target variables are available for the respondents of the SME sample. Starting from the situation represented in Table 4, the missing values have been imputed in order to obtain a complete data set corresponding to the whole SME population. The imputation process results from a combination of different techniques that have been applied to separate groups of variables. Basically, the whole procedure can be thought of as composed of 4 sequential steps (see [3] for more details):

1. deterministic imputation based on the guidelines of subject matter experts;
2. imputation of variables *Y1*, *Y6 Y7*, and CS through Predictive Mean Matching (PMM);
3. imputation of variables *Y3*, *Y4 Y8*, *Y10*, and *Y5*, through Nearest Neighbor Donor (NND);
4. imputation of variables *Y9 , Y2* via a two-step procedure based on logistic and linear regression.

The deterministic step is essentially limited to obvious imputations such as setting to zero some items for some specific categories of enterprises. As far as the genuine “statistical phase” is concerned, the differences in the distributional characteristics of the target variables imply the need of using different methodologies to manage missing values. Specifically, PMM has been used for variables not affected by “zero inflation”, i.e. variables that can be (at least approximately) considered as genuinely continuous. For these variables, we used a multivariate version of PMM composed of the following steps:

1. estimating via the Expectation Maximization algorithm the parameters of a multivariate normal model;
2. computing predictive means corresponding to missing items based on estimates obtained in a);
3. imputing missing values via NND using the predictive means as matching variables.

For the variables characterized by a “semi-continuous” distribution (i.e., affected by zero inflation), NND imputation based on ordinary Euclidean distance seems to be more appropriate. More precisely, for each incomplete record (recipient) whit positive value of *Y1* (Turnover), ratios between the variables in step 3) and *Y1* have been imputed from the closest donor, and the values of missing variables have been obtained as product of the imputed ratios by the value of *Y1*  in the recipient. For zero-turnover units, standard NND has been used. Finally, a fully parametric approach has been used for Y2 and Y9 (the two components of changes in stocks). Specifically, a logistic regression has been used to decide if the values of the two variables are zero or not, and in cases where nonzero values were to be imputed, an ordinary linear regression was used.

It is worthwhile noting that the imputation steps 1)-3) have been performed within suitably chosen imputation strata mainly based on *sector of economic activity*.

Due to the complexity of the imputation process, the evaluation of the whole procedure is a difficult task. However, we tried to assess the impact of the new estimation strategy, by comparing the estimates obtained from AD with the traditional ones based on the sample survey. In detail, for a given variable *Y*, with corresponding population total *Ty*, we have computed the relative difference , where is the estimate of *Ty* obtained with the sample data through the calibration estimator currently used for the SME survey and  is the estimate computed on the entire archive by summing up all the values. We also tried to decompose the difference above into a “sampling component” and a “measurement component”, by considering also the additional estimate  obtained by using the SME estimator on the sample units, but replacing the survey data with the corresponding data in the frame. We obtain an approximate measure of the SME sampling error and of the measurement effect respectively, through the following two indicators:

, .

Thus, the equality  provides the desired decomposition. In Table 5, the above indicators are reported for variables *T* (*Turnover*), *IC* (*Intermediate Costs*), and the variable *Value Added* (VA) defined as: , together with the population size (N) and the SME sample size (*n*). As it can be seen, actually the total difference between the survey and *frame-*based estimates is mainly due to the “sampling component”. This evidence indicates that the *frame* may ensure final estimates of the main target variables which are free of the “traditional” levels of sampling errors, deriving from the necessity of estimating target parameters for a very large population based on a relatively small survey sample size.

*Table 5: Total difference (dt) and measurement component (dm) in comparing estimates based on administrative data and sample data. Year 2011*

|  |  |  |
| --- | --- | --- |
|  | *dt* | *dm* |
| ***N*** | ***n*** | ***T*** | ***IC*** | ***VA*** | ***T*** | ***IC*** | ***VA*** |
| 4,432,928 | 74,112 | -6.5 | -8.8 | 0.2 | -0.9 | -0.5 | -0.9 |

* 1. *Estimation of the components of the main economic aggregates*

For the components of each economic aggregate, for which the only informative source is essentially the SME sample survey, we chose a design based/ model assisted approach to the inference exploiting the randomization process of the sample selection. To define the estimation process we have been driven by some requirements of the final output. The estimated total of the sum of the elementary components belonging to a given economic aggregate must be coherent to the (estimated) total of the aggregate at domain level according to the current SBS Regulations. Furthermore, the output should be a flexible statistical tool and it can be used for other aims. In particular, NA will base its procedure on the *frame*, so the coherence of the estimates must be fulfilled for the NA domains that are generally highly detailed.

The estimation procedure of the components have been performed using the *projection estimator* ([6]). The procedure takes into account the sample of respondents of the SME survey and, after the adjustment of the sample weights for nonresponse, sets a regression model for each variable of interest (component) being the parameters estimated using the sampling weights. The component values of the unobserved enterprises of the population are imputed or projected, using the estimated regression parameters. The final estimates are achieved as the sum of the values projected by the models. The implementation of the estimation procedure uses the economic aggregate as covariates. The underlying assumption is that they are really observed and not obtained by means of the imputation procedure (section 4.1). The reliability of the subsequent inference will depend on the goodness of the previous imputation step. The estimator has two important properties: 1) it allows the use of ad hoc regression model for each component; 2) micro level estimates are obtained.

As far as the first point is concerned, we point out that the traditional regression estimator is assisted by only one regression model, therefore the *projection estimator* improves the precision of the estimates. Eventually, the availability of the micro level estimates does not mean that we may use estimates for each type of subpopulations (especially highly disaggregate) since the estimator could produce biased values. In particular the *projection estimator* is unbiased for a given domain (or at more aggregate level) when the domain intercept is included in the model. In case of hetheroscedastic model, where the variance is a square root function of a covariate, the regression parameter of such covariate must be estimated at domain level as well.

In the practical implementation of the procedure we have taken into account of the trade-off between precision and bias; models defined at highly detailed domain level allows to compute unbiased detailed estimates, but variance increases and then we cannot use too specific regression models. Hence, the estimation procedure first considered the coverage of about 33,600 respondents of the whole population (SME survey, year 2011). The analysis led to consider models estimated for domain defined according to the Nace Rev. 2 three digit economic activity by the size class of employed persons (0-5, 6-19 , 20-99 ). Under this level of detail the estimates could be biased.

We obtained about 600 domains (and regression models) with generally a minimum number of sampled units equal to 25 and an average number of about 45 of sampled units. In some cases we obtained smaller size domains but with a high sample rate (Table 6).

*Table 6 Rules for not collapsing the domain*

|  |  |
| --- | --- |
| (If) Number of respondents | (than) Sample rate(respondents/population size) |
| 1-2 | 1.00 |
| 4-5 | 0.50 |
| 6-8 | 0.10 |
| 9-14 | 0.02 |

Finally, it was necessary to collapse some economic activities / or classes of employed persons to gather an enough number of respondents for estimating the models. On the same domains we adjusted the sampling weights with the inverse of the domain unit non response rate, being the sample size of 2011 SME survey of about 97,000 enterprises. We remark that the decision to use the same domains for the nonresponse adjustment and for defining the projection estimator has been obtain after some studies carried out to evaluate the best way to deal with the nonresponse. In particular, the logistics model and different nonresponse classes for nonresponse adjustment has been compared. The results have been not significantly different from the ones achieved by the correction with nonresponse rate computed in the domains of the projection estimator. Nevertheless, the use of logistic model for estimating the response probability can be cumbersome if the process has been performed in each survey occasion.

The regression models assisting the *projection estimator* have been defined taking into account the space of the possible values. For all the component variables the constraint is to obtain non negative projected values, but the components of the change economic aggregates. So, for the former type of variables the heteroscedastic ratio models have been used, where each component has as covariate the economic aggregate to which it belongs to. We point out that with this model the sum of the components of a given aggregate is equal to economic aggregate at enterprises level. The regression model for the components of the change economic aggregate uses standard heteroscedatic model where the heroscedasticy terms is the square root of the number of employees.

Table 7 shows the coefficients of variation (CV) at national level for the components of some economic aggregates. It is worthwhile to note that the CV are generally extremely low. One component of the *Turnover*, the C11106 has a CV larger than 10% while for the *Purchases of services* the component C12207 has a CV equal to 11.64%. The analysis of the CV at disaggregate level is more controversial. Many domains have very large CV, but they are small domains and the estimates are referred to components with many zero (rare phenomena). Figure 2 depicts the CVs distribution by domains (3 digits Nace Rev.2 by size class) and components. In particular the first, second and third quartiles of the CV distribution are shown.

*Table 7 CV (%) of the components (label of the SME survey) of some main economic aggregates*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *Income from sales and Services (Turnover)* | **C11101** | **C11102** | **C11103** | **C11104** | **C11105** | **C11106** | **C11107** |  |
| CV | 1.67% | 1.05% | 7.22% | 9.81% | 6.75% | 13.62% | 1.69% |  |
| *Purchases of goods* | **C12101** | **C12102** | **C12103** |  |  |  |  |  |
|  | 1.81% | 9.36% | 1.10% |  |  |  |  |  |
| *Purchases of services (1)*  | **C12201** | **C12202** | **C12203** | **C12205** | **C12206** | **C12207** | **C12208** | **C12209** |
| CV |  3.7% |  4.7% |  5.7% |  5.3% |  5.1% |  11.6% |  3.5% |  6.1% |
| *Purchases of services (2)*  | **C12210** | **C12211** | **C12212** | **C12213** | **C12214** | **C12245** | **C12246** | **C12247** |
| CV | 3.00% | 9.55% | 5.82% | 2.45% | 3.06% | 2.98% | 9.44% | 2.15% |
| *Use of third party assets* | **C12301** | **C12302** | **C12304** |  |  |  |  |  |
| CV | 1.20% | 2.44% | 2.90% |  |  |  |  |  |
| *Other operating charges* | **C12903** | **C12905** |  |  |  |  |  |  |
| CV | 1.15% | 6.13% |  |  |  |  |  |  |

The impact of the *frame* on the main SBS for year 2011 has been evaluated based on the comparison between the target parameters estimates obtained by the current surveys and the corresponding ones based on the *frame*, at different levels of detail. Let  be the estimate of *Ty* based on the SME survey and on the total survey on Large Enterprise economic accounts (LE, about 11,000 enterprises with more than 100 employees), and let  be the corresponding estimate based on the combination of the *frame* and the LE micro-data. In Table 8 the relative distance  are reported for some main variables by size class.

*Figure 2. Distribution of the coefficient of variation (%) at domain level by components of some economic aggregates (blue: first quartile, red: median, green: third quartile)*



*Table 8.* *values between frame-based and survey-based estimates by variables and size class*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ***Size class*** | ***Turnover*** | ***Value Added*** | ***Labor Cost*** | ***Gross Operating Margin*** |
| *0-9* | 6.4 | -2.2 | 3.7 | -4.9 |
| *10-19* | 9.9 | 4.1 | 1.4 | 9.4 |
| *20-49* | 10.2 | 4.9 | 0.8 | 14.5 |
| *50-99* | 6.5 | -0.6 | -1.6 | 1.7 |
| *100-249* | 0 | 0 | 0 | 0 |
| *250+* | 0 | 0 | 0 | 0 |
| ***Total*** | **4.5** | **0.2** | **0.8** | **-0.5** |

As it can be seen, the use of the *frame* determines an increase of the estimates for all the reported variables, except than for *Gross Operating Margin*. The estimate of *Value Added* increases of about 0.2%, even if it decreases (-2.2%) in the class [0,9]. Concerning *Turnover*, the *frame* determines a significant increase of the total estimates for all the size classes.

1. **Conclusions and future work**

In this paper the main features and methodological solutions adopted to produce the new statistical information system for SBS estimation (called *frame*) based on the use of AD are illustrated. The overall production strategy of the *frame* is depicted in Figure 3. As it can be seen, the *frame* consists of two main “products”: a database of elementary data (*frame SBS micro*) which allows the estimation of the *main economic aggregates* at any level of detail, and a database of estimates (*frame SBS macro*) for the *components of the main economic aggregates*, at a detail level which is in line with the SBS Regulation. Both databases represent inputs and “constraints” for NA.

*Figure 3: The overall production process of the frame*



Therefore, in the short term, the *frame* will serve as an intermediate output ensuring high levels of consistency between annual business statistics and NA, as early as the 2011 NA Benchmark. In the medium-long term, the *frame* is expected to have a strong impact on the overall system of the economic statistics: actually, the *frame* will be the basis of reference for the other economic surveys on enterprises, i.e. for a more integrated design of the main current structural economic surveys.

Future activities will focus mainly on improving the statistical use of AD sources, especially in relation to inferential aspects, as the use of AD in Official Statistics requires a new framework for evaluating the accuracy of parameters estimates. Furthermore, the need of and integrated use of the *frame* in the overall economic statistics system is stimulating the research and development of solutions for the efficient and consistent estimation of structural variables from different sources.

**References**

[1] Casciano, C., De Giorgi, V., Luzi, O., Oropallo, F., Seri, G., Siesto, G. (2011). *Combining administrative and survey data: potential benefits and impact on editing and imputation for a structural business survey*. UNECE Work Session on Statistical Data Editing, Ljubljana, Slovenia, 9-11 May.

[2] De Giorgi, V., Guarnera, U., Luzi, O., Rinaldi, M., Seri, G. (2011). *Estimating structural business statistics based on administrative data: the case of the Italian small and medium enterprises*, European Establishment Statistics Workshop (EESW11), Neuchatel, 12-14 September.

[3] Di Zio, M., Guarnera, U., Varriale R. (2014). *Imputation with multi-source data: the case of Italian SBS.* Paper presented at United Nations Economic Commission for Europe, conference of European statisticians

[4] Di Zio, M., and Guarnera, U. (2013). *A Contamination Model for Selective editing*. Journal of Official Statistics, Vol. 29, No. 4, pp. 539-555.

[5] Guarnera, U., Luzi, O., Silvestri, F., Buglielli, T., Nurra, A., Siesto, G. (2012). *Multivariate selective editing via mixture models: first applications to Italian structural business surveys*. UN/ECE Work Session on Statistical Data Editing, Oslo, 24-26 September.

[6] Kim, J. K. K., Rao, J. N. K. (2011). *Combining data from two independent surveys: a model-assisted approach*. Biometrika. No.8, pp. 1–16.

[7] Luzi, O., Di Zio, M., Oropallo, F., Puggioni, A., Sanzo R. (2013). *Integrating administrative and survey data in the new Italian system for SBS: quality issues*. Paper presented at the 3rd European Establishment Statistics Workshop, Nuremberg, Germany.

[8] Wallgren, A., Wallgren, B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*, John Wiley & Sons.

[9] Zhang, L.C. (2012). *Topics of statistical theory for register-based statistics and data integration*. Statistica Neerlandica, Vol. 66, No. 1, pp. 41–63.

1. BLUE – Enterprise and Trade Statistics: WP 4 “Improve the use of administrative sources”. http://www.cros-portal.eu/content/blue-ets [↑](#footnote-ref-1)
2. Use of administrative data and account data for business statistics: WP2 “Checklist for the quality of administrative data inputs”. http://www.cros-portal.eu/content/use-administrative-and-accounts-data-business-statistics. [↑](#footnote-ref-2)
3. Data Integration: WP2 “Development of Methods”. http://www.cros-portal.eu/content/data-integration-1. [↑](#footnote-ref-3)
4. 4-digits Nace-code, 3-digits Nace-code by 7 size classes, 2-digits Nace code by Regions (NUTS-2). [↑](#footnote-ref-4)
5. Two main SS models are available, depending on the unit’s legal form (individuals and private partnerships). [↑](#footnote-ref-5)