WP.

LISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Paris, France, 28-30 April 2014)

Topic (ii): New and emerging methods

Imputation with multi-source data: the case of Italian Structural Business Statistics

Prepared by M.Di Zio, U.Guarnera, R.Varriale,
Istat, Italy

I. INTRODUCTION

1. In Istat, Structural Business Statistics (SBS) for small and medium enterprises (SME) are traditionally based on sample surveys. In the last years, the increasing availability of information from administrative sources made it possible to take into account the possibility of using administrative data to improve the quality of the produced statistics. Until now this information has been generally used as auxiliary information to treat non-response in survey data and to calibrate the estimates on known aggregates.

2. The level of maturity in the analysis of these kinds of data lead Istat to use administrative data as a primary source for information to produce SBS statistics. Data from administrative sources as *Financial Statement, Studi di settore, Tax Return* are used to build a microdata file composed of the main economic variables. The choice of producing a microdata file follows from the difficulty of providing coherent estimates at different level of aggregation, in this regard we remind that these data are also used by National Accounts to build national economic aggregates.

3. Since not all the variables are available in all the data sources, and the sources cover only subsets of the target population, the microdata file is a result of an imputation process. The imputation procedure is based on a combination of different techniques that are introduced to comply with requirements given by constraints, such as statistical relationships among main variables, balance edits, and presence of zero-inflated variables.

4. Given such a complexity, the assessment of the procedure is not an easy task. A comparison with official estimates based on the SME sample survey data is carried out. The differences are decomposed in terms of sampling and measurement errors. The analysis of the impact of the different error sources may be useful to validate the results and to improve the process of production of statistics in this context.

5. The paper is structured as follows. Section II describes the informative context of SME statistics based on administrative data. The imputation process is described in Section III, and some results about the evaluation of the estimation procedure are reported and discussed in Section IV.

II. INFORMATIVE CONTEXT

6. The administrative data sources are the Financial Statements, Studi di settore, and Tax Return data. The units of Financial statements (FS) are the companies, mainly corporate firms, liable to fill in the financial statement. The "Studi di settore" (SDS) is a Fiscal Authority survey that aims at evaluating the capacity of enterprises to produce income and at indirectly assessing whether they pay taxes correctly. The units compiling the SDS 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 return data are mainly based on *"Unico"* and, for a residual part of units represented by corporate firms, on *"Irap"* (the Italian regional tax on productive activities).

7. There are apparently many variables observed in the administrative data sources that can be used for the SBS estimates, but even after an harmonization step, only some of them are considered enough reliable both in terms of consistency of definitions with the ones described by the SBS regulation, and in terms of reported values compared to the SME observations. The list of the variables used in the the imputation process is reported in Table 1.

TABLE 1. Variable used in the imputation process

Section	Label	Variable
Revenues	Y_1	Income from sales and Services (Turnover)
	Y_2	Changes in stock of finished and semi-finished products
	Y_3	Changes in contract work in progress
	Y_4	Changes in internal work capitalized under fixed assets
	Y_5	Other income and earnings (neither financial, nor extraordinary)
Costs	Y_6	Purchases of goods
	Y_7	Purchases of services
	Y_8	Use of third party assets
	Y_9	Changes in stocks of raw materials and for resale
	Y_{10}	Other operating charges
	PC	Personnel Costs

It is worthwhile to remark that the variable *personnel costs* is always observed and it is used as auxiliary variable in the imputation procedure. In addition to PC, some derived variables are used as auxiliary variables in the imputation process. In fact in some cases they are considered more reliable than the variables used for the derivation, this is due to a kind compensation process that is not easy to model. The derived variables refer to the Cost section and are $CS = Y_2 - Y_9$ (Total Change in Stocks), $GS = Y_6 + Y_7$ (Purchases of Goods and Services) and $IC = GS + Y_8 + Y_{10}$ (Total Intermediate Costs).

8. We remark that some variables are observed in more than one data sources, this means that for each of this variable (generally) different values are available. In this application a hierarchical approach is chosen. It consists in assigning a hierarchy to the administrative sources and consequently values of the variables are chosen according to this raking. The hierarchy is established according to some quality criteria, and it assigns the first rank to FS, then to SDS, and finally to Tax Return data. The general coverage of the administrative sources for the 2011 is reported in the Table 2.
It is however worthwhile to remark that the coverage of each single variable depends on its availability in the different data sources.

9. In Table 3 the frequency of missing data per variables conditionally on the different data sources is illustrated. The symbols 'X' and '?' stand for observed and missing data respectively. In this table,

TABLE 2. Coverage of administrative sources for the 2011 year

Source	Frequency	Relative frequency (%)
FS	714885	16.1
SDS	2836100	64.0
Unico	714894	16.1
Irap	4201	0.1
NA	162848	3.7
Total	4432928	100

SDS-F, SDS-G and Unico1-Unico8 refer to the different kinds of SDS and Unico that enterprises have to fill in depending on their legal status. In particular, the units compiling SDS-G and Unico5 are represented by professionals and "minimum taxpayers", respectively.

TABLE 3. Coverage of administrative data per variables related to Revenues and Costs section, and sources for the 2011 year

Source	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	PC	CS	GS	IC	Coverage rate (%)
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	50.08
SDS-G	X	X	X	X	X	?	?	?	X	X	X	X	?	X	13.90
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	10.86
Unico6	X	?	?	?	?	?	?	?	?	?	X	?	?	?	0.16
Unico7	X	?	?	?	?	?	?	?	?	?	X	?	?	?	0.31
Unico8	X	?	?	?	?	?	?	?	?	?	X	?	?	?	0.49
Irap	X	X	X	X	X	X	X	X	X	X	X	X	X	X	0.09
NA	?	?	?	?	?	?	?	?	?	?	X	?	?	?	3.67

The rate of missing data per variable is reported in Table 4. We notice a very high rate of missing data for the variables Y_2 and Y_9, however it is worthwhile to remember that they have a deterministic relation with the variable $CS = Y_2 - Y_9$ and this mitigates the impact of such a high level of missing information.

TABLE 4. Percentage of missing data per variable

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9	Y_{10}	CS	GS	IC
2.7	58.3	8.2	4.7	8.2	19.2	19.2	19.8	58.3	19.8	6.7	15.6	15.6

The economic data described in this paragraph are used both by the SBS sector and by National Accounts, requiring many domains of estimation. In order to avoid consistency problems, missing data are imputed to obtain a microdata file.

III. THE IMPUTATION PROCESS

10. The imputation procedure is based on a combination of different techniques.
The entire imputation process is composed by 4 sequential steps:

(1) deterministic imputation based on the guidelines of subject matter experts;
(2) imputation of the variables Y_1, Y_6, Y_7 and CS, through Predictive Mean Matching (PMM);
(3) imputation of the variables Y_3, Y_4, Y_8, Y_{10} and Y_5, through Nearest Neighbor Donor (NND);
(4) imputation of the variables Y_9, Y_2 through a two-step procedure composed by a logistic and a linear regression model.

In this paper we focus on the description and evaluation of the imputation process related to the last three steps, in particular we remark that the pattern of missing data depicted in Table 3 is the one obtained after the deterministic imputation in step 1. The steps from 2 to 4 have been carried out inside strata based on the economic divisions Nace2 cross-classified with the two subsets of observations characterised by having or not personnel costs. For the PMM, also the item GS has been used to define strata, the distinction is made between units either with or without purchases of goods and services.

11. The enterprises with information coming from SDS-G and Unico5 are represented by professionals and "minimum taxpayers", and they are considered to behave quite differently from the rest. Since the imputation process tends to reproduce in the non-observed part of the population the behaviour of the observed units, for this subset of population the imputation is made by resorting to the SME survey, details will be given later in the paper.

12. The choice of each imputation method for different groups of variables is due to: the percentage of missing values, the variable distribution characteristics (only positive, zero-inflated, etc.), the presence of a (weak/strong) relationship between variables and the presence of balance edits. All these characteristics influence the choice of a statistical model in the imputation process.

13. The PMM [Little, 1988] can be considered as a NND imputation technique based on a distance function where matching variables are weighted through their predictive power with respect to the variables that have to be imputed. In a multivariate context, the PMM is typically applied to match each recipient to the donor having the closest predictive mean with respect to a regression model of the target variables on a set of covariates. Selection of donors is based on the Mahalanobis distance defined in terms of the residual covariance matrix from the regression model. Intuitively, Mahalanobis matrix gives largest weights to the variables with the smallest prediction error.

14. The NND method is a common hot deck method, in which a donor is selected from the complete cases in order to minimize some similarity measure, such as the Euclidean distance. In this context, the matching variables to compute the Euclidean distance are Y_1, Y_6, Y_7 and PC (if present), and the variables to be imputed are the ratios of Y_3, Y_4, Y_5, Y_8 and Y_{10} on Y_1. The final imputed value is obtained by multiplying the imputed ratios by the size variable Y_1 of the recipient unit, this technique is also known as ratio hot deck [de Waal et al., 2011]. This method is preferable because it ensures that the values of the variables to be imputed are coherent with respect to the value of the reference variable. The reason why Y_1 has been treated as a size variable, instead of the commonly used *Number of employees*, is that it has both the lowest rate of missing data and the highest quality from a content point of view. When Y_1 is zero the ratio cannot be computed, in this case the standard NND is used.

15. In this context, both the PMM and NND approaches have the advantage to recover live values from donors. Since the PMM technique relies on a multivariate normal model, it has been used to treat variables having a genuine continuous distribution. On the contrary, the NND method has been used to treat variables with distribution characterized by 0 inflation and a non-linear relation.

16. Finally, the imputation of Y_2 and Y_9, representing the two components of CS has been carried out through a two-step process, composed by a logistic and a linear regression model. This approach has been compared with a NND approach through a simulation study, resulting in a better efficiency (both in terms of time consuming and accuracy of the estimates) of the two-step approach. The difficulty in the imputation of these variables is in both their nature and in the nature of the total CS that is semi-continuous and not positive. This means that the value CS could be generated by any linear combination of the two components. As an hypothesis, when the variable CS is equal to 0, the

two components have been imputed to be equal to 0. In the first step, we applied a stepwise logistic regression using as covariates Y_1, Y_3, Y_6, Y_7, CS, a modified version of the economic divisions Nace2 and PC (if present) in order to assign each enterprise to one of the 3 subpopulations characterized by the presence or absence of the two components (yes/yes, yes/no, no/yes). In particular, the assignment is based on random drawing from a multinomial distribution with parameters corresponding to the probabilities estimated through the logistic model. In the second step, for the enterprises which have been assigned to the subpopulation with only one component, the total value of CS has been imputed to such component. For the other enterprises, we estimated the value of the two components through a linear regression model with the same covariates used in the logistic model.

17. For enterprises with information coming from SDS-G and Unico5, all the information on revenues is complete. Costs are imputed through random ratio hot-deck within suitable defined imputation cells. The donors are chosen from the SME survey, this is the same as drawing a vector of ratios from the estimated distribution of the ratios in SME. In particular, we imputed the composition of the costs using as size variable the total costs and transferring the compositional information from the survey data.

IV. EVALUATION

18. The complexity of the imputation procedure and the particular nature of administrative data make the evaluation of the accuracy of the estimates a difficult task. A first overall evaluation has been obtained by comparing the estimates based on administrative data with the ones resulting from the classical procedure obtained by means of the SME sample survey data. The comparison has been made by using two different sets of estimation domains: the first is Nace2 (aggregation of economic sectors), and the second corresponds to the different administrative sources where information is taken from. While in the former case the domains are aggregations of planned survey domains - thus they are composed of sampling strata - in the latter case the domains - that have not been planned in the survey design phase - are used to analyze possible different levels of discrepancies between administrative data and survey data across the available sources.

19. For each typology of domain and each analysed variable, relative differences between total estimates based on administrative and sample data are considered. In detail, for a given variable Y with corresponding population total T_y, we have computed the indicator:

$$d_y^t = \frac{\hat{T}_y^s - \hat{T}_y^{ad}}{\hat{T}_y^{ad}} \times 100,$$

where $\hat{T}_y^s$ is the estimate of T_y obtained with the sample data through the calibration estimator currently used for SME survey, and $\hat{T}_y^{ad}$ is the estimate computed on the entire archive by summing up all the values. In order to distinguish the source of discrepancies due to the sampling and the measurement error, we have also considered, for each domain, the additional estimate $\hat{T}_y^{ad,s}$, that results from using the SME survey estimator on the sampled units, with the replacement of the survey data with the administrative data. As approximate measures of the measurement effect and sample error respectively, we introduce the following two indicators:

$$d_y^m = \frac{\hat{T}_y^s - \hat{T}_y^{ad,s}}{\hat{T}_y^{ad}} \times 100, \qquad d_y^s = \frac{\hat{T}_y^{ad,s} - \hat{T}_y^{ad}}{\hat{T}_y^{ad}} \times 100.$$

Thus, the total difference is decomposed into the sum of two differences associated with the two mechanisms:

$$d_y^t = d_y^m + d_y^s. \quad (1)$$

Note, however, that the indicator d_y^m that evaluates the "measurement effect", being based on the comparison of different measures only on the sample units, is also affected by sampling error. In particular, a few gross errors may have an high impact on the indicator.

20. Table 5 reports the indicators d^t and d^m for the three variables Y_1, IC, and the Value Added computed as $VA = \sum_{i=1}^{5} Y_i - IC - Y_9$ by source. In Table 6 results are shown for the following economic divisions Nace2: *Manufacture of textiles* (Nace2 =13), *Construction of buildings* (Nace2=41), *Wholesale and retail trade and repair of motor vehicles and motorcycles* (Nace2=45) and *Architectural and engineering activities; technical testing and analysis* (Nace2=71). In the tables, the size of domains in the population (N) and in the sample (n) are also reported.

TABLE 5. Discrepancies between sample estimates and estimates based on administrative data for different administrative data sources: total differences (d^t) and measurement component (d^m)

			d^t			d^m		
Source	N	n	Y_1	TC	VA	Y_1	TC	VA
Tot	4432928	74112	-6.5	-8.8	0.2	-0.9	-0.5	-0.9
FS	714885	34284	-2.6	-4.6	3.1	-0.9	-0.7	-2.3
SDS-F	2220050	31732	11.4	11.3	12.7	-0.5	-1.2	1.6
SDS-G	616050	3844	23.1	26.5	26.7	-0.2	13.5	-0.3
Unico1	34570	296	-38.9	-69	-26.7	-1	-3.3	0.3
Unico2	1746	32	-41	-40.6	-40	-1.7	-2.5	0.9
Unico3	120876	756	-70.7	-78.5	-55.7	-0.7	-5.3	8.4
Unico4	33676	356	-69	-81.1	-34.6	-0.5	-5	14.6
Unico5	481517	1434	-58.2	-51.5	-59.1	-1.4	3.3	-1.3
Unico6	7371	151	-70.1	-76.9	-59.5	0.1	-2.5	-6.5
Unico7	13553	361	-68.5	-68.5	-59.4	-0.4	-2.5	13.8
Unico8	21585	381	-63.7	-55.3	-58.2	-2.4	8.7	-1.4
Irap	4201	89	-55.8	-63.1	-61	-0.1	-0.9	4.1
NA	162848	396	-91	-90.9	-89.8	-2.5	-1.3	-4.6

TABLE 6. Discrepancies between sample estimates and estimates based on administrative data for some economic divisions (Nace2): total differences (d^t) and measurement component (d^m)

			d^t			d^m		
Nace2	N	n	Y_1	TC	VA	Y_1	TC	VA
13	15669	1275	8.4	10.1	3.8	0	1.1	-3.5
41	150417	2625	-18	-21.7	-9.2	-0.8	2	1.5
45	118985	2649	0.4	0.3	1.7	1	0.9	-0.7
71	212880	1009	-9.5	-15.3	-4.6	-2.6	1.5	-0.5

Results in Tables (5) and (6) show that the largest component in the decomposition (1) is the one associated with the sampling error. This result is encouraging because it implies that the transition from designed based inference to an estimation approach based on administrative sources would result in a significant improvement of the estimation accuracy.

21. An important issue in the evaluation of an estimation procedure is the assessment of the estimate accuracy. According to the estimation approach so far used in Istat, the SBS estimates for SME are based on a sample survey, hence the assessment of their accuracy relies on designed-based inference. As already mentioned, massive use of administrative information requires a change of paradigm. In fact, differently from the context of sample survey, the availability of administrative information is not under control of the researcher, so that some model assumptions are necessary. In particular, one has to think of data as *iid* realizations from a statistical (possibly not explicitly specified) model. This is generally referred to as super-population model. In this framework, the

inferential approach is predictive, i.e., the missing values are imputed (predicted) on the basis of the available information. Thus, the uncertainty of the resulting estimates are essentially due to the prediction error associated with the imputation procedure.
If predictions were based on some parametric regression model and the missing patterns were enough simple, standard analytic techniques could be used to evaluate the estimate of the estimator variance Valliant et al. [2000]. In cases where missing patterns are arbitrary, but imputations are obtained from a unique multivariate normal, the Rubin multiple imputation approach can be (relatively) simply applied to assess the precision of the estimate of any finite population quantity Rubin [1987]. In the present case, however, the imputation procedure is complex and is composed of many different techniques. This complexity makes it difficult to use standard procedures for the assessment of the uncertainty in the final output. In particular, the assumed super-population model is not explicitly specified and it is only implicitly defined through the imputation procedures that have been used. Because of this characteristic, a replication approach seems to be more appropriate than an analytic approach. However, common univariate techniques for the variance estimation such as jacknife and bootstrap Wolter [2007] are difficult to extend to our context and further research is needed.

References

T. de Waal, J. Pannekoek, and S. Scholtus. *Handbook of Statistical Data Editing and Imputation*. Wiley, 2011.

R.J.A. Little. Missing-data adjustments in large surveys. *Journal of Business & Economic Statistic*, 6 (3):287–296, 1988.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, 1987.

R. Valliant, A.H. Dorfman, and R.M. Royal. *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, 2000.

K.M. Wolter. *Introduction to Variance Estimation*. Springer, 2007.