Estimating Structural Business Statistics in the Italian Public Sector from multiple administrative data sources

Orietta Luzi, Tiziana Pichiorri, Roberta Varriale
Italian National Statistical Institute (Istat)
National Accounts and Business statistics Department
{luzi,pichiorr,varriale}@istat.it

1. Abstract

At the Italian National Statistical Institute (Istat in the following), the direct use of administrative data for estimating business statistics has progressively increased, stimulated by the augmented availability and quality of secondary data on both private and public businesses. In this context, in 2014 a research project has started aiming at developing a statistical information system to support the estimation of the economic accounts of Public bodies, based on the integrated use of microdata from different administrative sources. The new system is expected to ensure higher quality and better consistency of Structural Business Statistics and National Accounts in the Italian Public Sector.

Given the peculiarities of the target population and the characteristics of the available sources, the development of the system implies the management of a number of challenging issues, such as the harmonization of concepts in the sources (target populations/units, target variables), the evaluation of their quality and usability (coverage, accuracy, etc.), the identification and treatment of integration and linkage errors, the analysis and treatment of measurement, coverage and response errors.

This paper focuses on the quality issues addressed and the methodological solutions adopted to deal with missing information for the sub-population of Italian Municipalities (about 8.100 units). The aim is to evaluate the potential biasing effect of estimating the main items of the economic accounts of municipalities based on predictions (imputations) of microdata values. The analyses have been conducted by running a MC simulation study.

The paper is organized as follows: section 2 and 3 describe the data and the simulation study, respectively. Section 4 ends the paper with some conclusions and suggestions for future research.

2. Data sources

The available administrative sources of information for the sub-population of Italian Municipalities are the Economic Account Certificate (EAC in the following), which is the primary source of information on municipalities' accounts, and the Information System on Public Bodies Operations (Siope in the following), which contains information related to the target economic variables. Additional information used in the estimation strategy comes from statistical sources such as: the 2011 Census of Industry and Services (providing information on structural characteristics of Public Institutions, including number of employees); the 2011 Population Census; the Istat annual survey on resident population of municipalities by gender, year of birth and marital status; the Italian Register of Public Institutions.

In 2012, the Municipalities in the Italian Register of Public Institutions are 8.092. Out of them, 7.387 units (91.3%) have information from the EAC, and all of them have information from Siope. It has to be remarked that, in our specific application context, the non-availability of information for

a given Administration in a given source can be essentially due to either genuine non-response, or to under-coverage, or to unit identification errors (e.g. due to units demography). Tables 1 and 2 show the distribution of missing values for geographical macro regions and population size: the highest non response rate is observed in the South and Islands, especially in Puglia, Calabria and Sicilia, and in the municipalities with less than 5.000 inhabitants.

Table 1. Distribution of Municipalities with missing information, for geographical macro regions and population size (in classes), year 2012

Population	Missing values		Total
	N	%	N
North-West	136	4,4	3059
North-East	82	5,5	1480
Center	80	8,0	996
South	267	14,9	1790
Islands	140	18,3	767
Total	705	8,7	8092

Population size (inhabitants)	Missing values		Total
	N	%	N
< 1500	284	9,9	2866
[1500,5000)	248	8,8	2832
[5000,10000)	96	8,1	1189
[10000,60000)	73	6,6	1104
[60000,100000)	3	5,5	55
> 100000	1	2,2	46
Total	705	8,7	8092

3. Simulation study

Based on the analysis of the observed missing data patterns, and on specific assumptions on the nature of the missing data mechanism (see below), a number of alternative parametric and non-parametric imputation methods have been considered, including longitudinal approaches which exploit the panel structure of the data.

The methods that have been applied are:

- *Nearest Neighbour Donor* (NND): the value that is imputed in unit *i* is the per-capita value of the response variable *Y* (*ratio hot-deck*, De Waal *et al.*, 2011), computed with respect to an auxiliary variable *X* that is known for all the population: $Y_{i,pc} = Y_i/X_i$. After identifying the NND *d* with respect to some matching variables statistically associated to the variable *Y*, the value Y_i^* to be imputed in municipality *i* is computed by the relationship: $Y_i^* = X_i \times Y_{d,pc} = X_i \times Y_d/X_d$ where Y_d and X_d are the values of the response variable *Y* and the auxiliary variable *X* of the donor municipality *d*.
- *Predictive Mean Matching* (PMM): the PMM is 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 with continuous target variables, a typical application of the PMM uses a regression model to compute the predictive mean of each unit (Di Zio and Guarnera, 2009). The selection of donors is based on the Mahalanobis distance defined in terms of the residual variance-covariance matrix in the regression model (Little, 1988).
- *Longitudinal NND* (LNND): this method is equal to the NND, except that the matching variables (M_1,..,M_k) include information on municipalities from 2011.
- *Longitudinal deterministic methods*: these methods start from the value of the response variable observed in 2011 for unit *i* (Y_i*_2011)* and updated it with an individual trend that is computed on an auxiliary variable (from Siope, Census, etc.) observed in 2011 and 2012 for that specific unit, or with a median trend.
- *Mixed methods*: in these approaches a longitudinal approach is combined with NND methods.

All methods have been applied within imputation classes (domain, *D*), i.e. homogeneous cells defined using some auxiliary variables that are considered explicative of the missing mechanism and are known for all the population (taken from the available statistical sources).

The "best" imputation method has been identified by means of a comparative evaluation study based on a MonteCarlo (MC) simulation, which allowed us to assess the quality of each method in terms of accuracy of results at both aggregate and microdata level. The simulation has been structured in the following steps:

1. starting from complete data, simulation of a rate of missing values on the response variable, following a *Missing Completely At Random* (MCAR) mechanism. The simulated non-response rate is equal to the observed percentage of non-response of the target variables in 2012;
2. application of different imputation methods to predict missing values;
3. computation of distance measures between imputed and observed values, both at aggregate and unit level;
4. iteration of steps 1-3 for *k=1.000* times;
5. computation of quality indicators based on the measures computed at step 3.

The indicators used to compare the imputation methods are (Luzi *et al.*, 2007):

- *Relative1 Bias (RB)* - or *Relative estimation error due to imputation* – in the domain *D*:

$$RB_Y^D = \frac{1}{K}\sum_{k=1}^{K}\frac{\left(\hat{T}_{Y,true}^D - \hat{T}_{Y,imp}^D(k)\right)}{\hat{T}_{Y,true}^D} \times 100$$

  where $\hat{T}_{Y,true}^D$ and $\hat{T}_{Y,imp}^D(k)$ are the total estimates of the response variable *Y*, computed respectively on the observed true values and on the imputed values (for each iteration *k*, *k=1,…,1000)* in the domain *D*.

- *Relative Root Mean Squared Error (RMSE)*

$$RMSE_Y^D = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\left(\hat{T}_{Y,true}^D - \hat{T}_{Y,imp}^D(k)\right)^2}{\hat{T}_{Y,true}^D}} \times 100$$

- *Relative Imputation Error (RIE)*

$$RIE_Y = \frac{1}{K}\sum_{k=1}^{K}\sqrt{\frac{\sum_{i=1}^{n^*}\left(y_{true,i} - y_{imp,i}\right)^2}{\sum_{i=1}^{n^*} y_{true,i}}}$$

  where $y_{true,k}$ and $y_{imp,k}$ are, respectively, the original and imputed values of the response variable *Y* for unit *i*, and *n** is the number of respondent units with simulated missing values.

The target variables of the present work are *Compensation of employees* and *Intermediate consumption.* Variable *Y* indicates each time the target variable under investigation that is directly measured in the EAC, on legal accrual bases. With '*S'* we indicate the variable from the source Siope (corresponding to *Y*), measured on cash bases, which is used as auxiliary information in the imputation process of *Y*. Some explorative analyses have shown very strong statistical correlation between *Y* and *S*.

Other auxiliary variables that have been used in the imputation process are: number of employees in 2011 and 2012 *(Nempl)*; surface of the municipality (*Surface*); population *(Pop)* in 2011 and 2012,

both in size classes (see Table 1) and absolute terms; geographical macro region of the municipality and *Geographical characteristics* (plain/not plain) of the municipality territory.

Table 3 shows the non-response pattern in the EAC and Siope data sources in 2011 and 2012 for the two target variables: values 1 and 0 indicate the presence and absence of information, respectively. Although the number of municipalities that need to be imputed in 2012 (705, corresponding to the grey cells in the Table) is equal for both variables, the number of units with no missing values for all sources is equal to 7.121 and 7.132, respectively.

Table 2. Missing values in the EAC and Siope sources, years 2011 and 2012

Non-response pattern								
Type	Information source				Compensation of employees		Intermediate consumption	
	EAC 2012	EAC 2011	Siope 2012	Siope 2011	N	%	N	%
1	1	1	1	1	7.121	88,00	7.132	88,14
2	1	1	1	0	3	0,04	0	0,00
3	1	1	0	1	2	0,02	0	0,00
4	1	1	0	0	85	1,05	79	0,98
5	1	0	1	1	165	2,04	165	2,04
6	1	0	0	0	11	0,14	11	0,14
7	0	1	1	1	534	6,60	534	6,60
8	0	1	0	0	3	0,04	3	0,04
9	0	0	1	1	167	2,06	167	2,06
10	0	0	0	0	1	0,01	1	0,01

Before running the simulation study, an exploratory data analysis has been conducted and a (robust) regression model has been applied to both investigate the variable characteristics and detect and remove outliers or anomalies from the dataset. The outlier observations have been treated interactively.

Imputation methods for variable *Compensation of employees* are implemented as follows:

- **NND**: for each municipality, the per-capita value of the target variable is computed with respect to the number of employees ($Nempl$): $Y_{i,pc} = Y_i / Nempl_i$. The matching variables are *S_2012/ Nempl*, *Surface*, *Pop_2012* and the variable used to form the imputation cells is *Geographical macro region.*

- **PMM**: as a first step of the PMM method, a multivariate linear regression model has been estimated within each imputation cell (defined using *Geographical macro region*):
  $Y_i = \alpha + \beta_1 S_i\_2012/Nempl_i + \beta_2 Pop_i\_2012 + \beta_3 Surface_i + e_i$
  Subsequently, the target variable has been imputed with a minimum distance donor with respect to the value Y_i^P predicted using the regression model.

- **LNND**: in this method, the matching variables used to identify the minimum distance donor for each unit are *Y_2011/Nempl*, *S_2011/Nempl*, *S_2012/S_2011*, *Population 2012*, *Surface* and the classification variable is *Geographical macro region.*

- **Longitudinal deterministic methods**: the imputation cells of this class of methods are build using a cross classification of the variables *Geographical macro region* and *Pop_2012* (in classes). In the formulas, $med_D(.)$ represents the median value in cell *D* of the distribution of variable (.). Types from 7 to 10 refer to the different non-response pattern represented in Table 2.

a. **Long EAC Sio**
   Types 7 and 8: $Y_i\_2012 = Y_i\_2011 * med_D(Y_i\_2012/Y_i\_2011)$
   Type 9: $Y_i\_2012=Y_i\_2011^P * (S_i\_2012/S_i\_2011)$
   where $Y_i\_2011^P$ is the value of Y_i_2011 predicted by the robust regression model for each *Geographical macro region*: $Y_i\_2011 = \alpha + \beta_1 S_i\_2011 + e_i$.
   Type 10: $Y_i\_2012= med_D(Y_i\_2012)$

b. Long EAC Pop
   Types 7 and 8: $Y_i\_2012 = Y_i\_2011 * med_D(Y_i\_2012/Y_i\_2011)$
   In all other cases: $Y_i\_2012= med_D(Y_i\_2011) * (Pop_i\_2012/Pop_i\_2011)$

c. Long Pop
   Type 7 and 8: Y_i is imputed by $Yi\_2012 = Yi\_2011 * (Pop_i\_2012/Pop_i\_2011)$
   In all other cases: $Y_i\_2012 = med_D(Y_i\_2011) *(Pop_i\_2012/Pop_i\_2011)$

d. Long Sio
   Type 7: $Y_i\_2012= Y_i\_2011 * (S_i\_2012/S_i\_2011)$
   Type 8: $Y_i\_2012= Y_i\_2011 * med_D(S_i\_2012/S_i\_2011)$
   Type 9: $Y_i\_2012= Y_i\_2011^P * (S_i\_2012/S_i\_2011)$
   Type 10: $Y_i\_2012= med_D(Y_i\_2012)$

- **Mixed methods**: in the first phase (Type 7 and 8), the process has a deterministic step where the target variable (Y_i_2012) is imputed using longitudinal information, within classes defined by *Geographical macro region* and *Population size*. In the second step (Type 9 and 10), the procedure is completed by a non-deterministic step using a LNND method within classes defined by *Geographical macro region*. The mixed methods use different auxiliary information in the deterministic step (*Pop, Y, S*):

  a. NND Long Mixed Pop: $Y_i\_2012= Y_i\_2011 * (Pop_i\_2012/Pop_i\_2011)$
  b. NND Long Mixed EAC: $Y_i\_2012 = Y_i\_2011 * med_D(Y_i\_2012/Y_i\_2011)$
  c. NND Long Mixed Sio: $Y_i\_2012 = Y_i\_2011 * (S_i\_2012/S_i\_2011).$

Imputation methods for *Intermediate consumption* are implemented as follows:

- **NND**: the per-capita value of the target variable is computed with respect to the variable *Population* for each municipality: $Y_{i,pc} = Y_i / Pop_i\_2012$. The matching variables are *S_2012/Pop_2012, Nempl, Surface, Geographical characteristics* and the variable used to classify units in class of imputation is *Geographical macro regions*.

- **PMM**: as a first step of the PMM method, a multivariate linear regression model has been estimated for each imputation cell (defined using *Geographical macro regions*):
  $Y_i = \alpha + \beta_1 S_i\_2012/ Pop_i\_2012 + \beta_2 Surface_i + \beta_3 Nempl_i + e_i$
  Subsequently, the target variable has been imputed with a minimum distance donor with respect to the predicted value Y_i^P.

- **LNND**: the matching variables are *Y_2011/Pop_2011, S_2011/Pop_2011, S_2012/S_2011, Nempl, Surface, Geographical characteristics,* and the classification variable is *Geographical macro region.*

- **Longitudinal deterministic methods**: these methods are similar to those used for variable *Compensation of employees* except for the classification variables used to determine the imputation cells, which are: *Geographical macro region, Pop_2012 (in classes)* and *Geographical characteristics*.

- **Mixed methods**: these methods are similar to those used for variable *Compensation of employees* except for the classification variables used to determine the imputation cells in the deterministic step, which are *Geographical macro region, Pop_2012 (in classes) and Geographical characteristics.* In the non-determinist step the classification variable is *Geographical macro region.*

Table 3 and 4 show the results obtained from the MC simulation study for the variable *Compensation of employees* and *Intermediate consumption*, respectively. The imputation methods ensuring higher levels of accuracy in terms of RMSE are those exploiting the longitudinal information of units with missing data in the reference year 2012. Taking into account all the indicators, the preferred methods result to be *Long Sio* and *NND long Mixed Sio*, which use also the auxiliary information from Siope. These results are confirmed also at regional level (not reported here).

Table 3. *Compensation of employees*: quality indicators by imputation method - Italian national level, year 2012

Indic.	Methods									
	NND	PMM	LNND	Long EAC Sio	Long EAC Pop	Long Pop	Long Sio	NND long Mixed Pop	NND long Mixed EAC	NND long Mixed Sio
RB	0,045	0,366	0,044	-0,004	0,023	-0,320	-0,021	-0,354	-0,012	-0,028
RMSE	0,362	0,981	0,361	0,182	0,192	0,355	0,103	0,383	0,182	0,105
RIE	0,240	0,433	0,237	0,108	0,122	0,130	0,075	0,118	0,108	0,075

Table 4. *Intermediate consumption*: quality indicators by imputation method - Italian national level, year 2012

Indic.	Methods									
	NND	PMM	LNND	Long EAC Sio	Long EAC Pop	Long Pop	Long Sio	NND long Mixed Pop	NND long Mixed EAC	NND long Mixed Sio
RB	0,584	-0,418	0,601	0,284	0,298	0,359	0,002	0,321	0,260	-0,018
RMSE	1,975	1,680	1,968	1,159	1,163	0,491	0,338	0,464	1,153	0,340
RIE	1,238	0,816	1,227	0,171	0,176	0,219	0,275	0,216	0,172	0,277

4. Conclusions

The analyses presented in the paper show good performances of some of the considered imputation procedures for the Italian Municipalities' economic accounts in terms of result accuracy at both aggregate and microdata level, especially when longitudinal information and auxiliary data are used in imputation models.

Future research is still needed: the multivariate nature of variables should be considered, and estimation methods for different and more complex key variables in economic accounts are to be assessed.

From a content point of view, the future work will be addressed on a deeper analysis of the informative context by subject matter experts in order to further exploit the informative power of all the available auxiliary information. From a methodological point of view, additional studies will be carried out in order to verify if a *Missing At Random* assumption for the non-response mechanism is more appropriate in this specific application context.

References

1. de Waal T., Pannekoek J., Scholtus S. (2011). *Handbook of Statistical Data Editing and Imputation.* Wiley
2. Di Zio, M., Guarnera, U. (2009). Semiparametric predictive mean matching. *AStA - Advances in Statistical Analysis*. 93, 175-186
3. R.J.A. Little (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6, 3, pp. 287-296
4. Luzi O., Di Zio M., Guarnera U., Manzari A., De Waal T., Pannekoek J., Hoogland J., Tempelman C., Hulliger B., Kilchmann D. (2007): *Recommended Practices for Editing and Imputation in Cross sectional Business Surveys*, EDIMBUS project report