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

Orietta Luzi, Tiziana Pichiorri, Roberta Varriale
National Accounts and Business statistics Department
Italian National Statistical Institute (Istat)

08 September 2015

Outline

1. Introduction
2. Data sources and informative context
3. Preliminary analysis
4. MC simulation study
    - steps
    - imputation methods
    - indicators
    - results
5. Conclusions

Introduction

Context:
At the Italian National Statistical Institute, 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.

Research project:
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.

Challenging issues:
- harmonization of concepts in the sources (target populations/units, target variables),
- evaluation of quality and usability of administrative data (coverage, accuracy, etc.),
- identification and treatment of integration and linkage errors,
- analysis and treatment of measurement, coverage and response errors.

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

Informative context (1)

Target population:
Italian Municipalities, about 8.100 units

Reference year:
2012

The administrative sources of information:
- Economic Account Certificate (EAC),
- Information System on Public Bodies Operations (Siope)

Additional information, from statistical sources:
- 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.

Informative context (2)

In 2012 the Municipalities in the Italian Register of Public Institutions are 8092, 7387 (91.3%) have information from the EAC, and all of them are covered by the Siope

Non response = non-availability of information for a given Administration in a source can be essentially due to:

- genuine non-response,
- under-coverage,
- unit identification errors.

Distribution of missing values for geographical macro regions and population size:

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	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

Target variables

Compensation of employees and Intermediate Costs

- *Variable Y:* indicates the target variable under investigation that is directly measured in the EAC, on *legal accrual bases*.

- *Variable S:* indicates 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*.

Other auxiliary variables used in the imputation process:

- *number of employees* in 2011 and 2012 (Nempl);
- *surface* of the municipality (Surface);
- *population* (Pop) in 2011 and 2012, both in size classes and absolute terms;
- *geographical macro region* of the municipality;
- *geographical characteristics* (plain/not plain) of the municipality territory.

Preliminary analysis - Compensation of employees (1)

Aims:

- investigate the variable characteristics
- detect and remove outliers or anomalies from the dataset

Methods:

- Explorative analyses
- A (robust) regression model

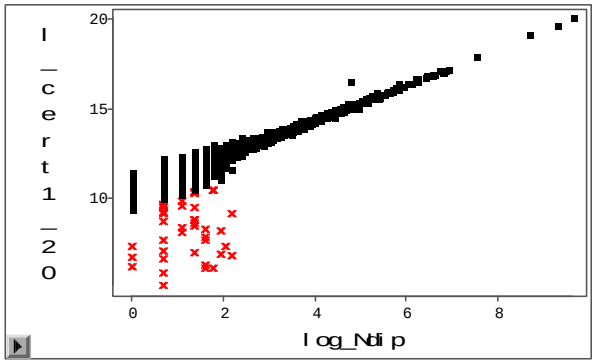
Results:

- strong statistical correlation between *Y* and *S*
- outliers have been interactively treated

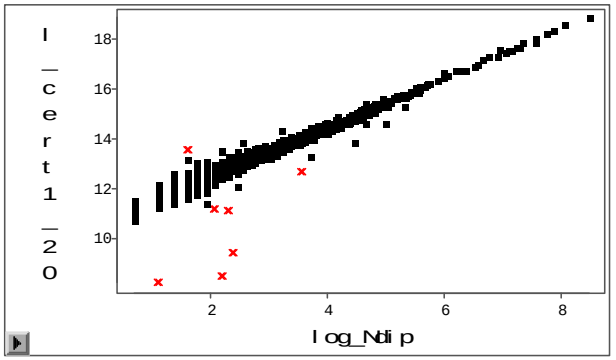
Scatter-plot of municipalities (logarithmic scale):
Compensation of employees and Number of employees

outliers in red

North-West

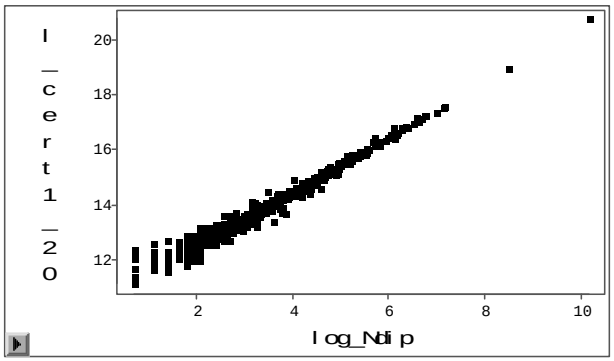


North-East

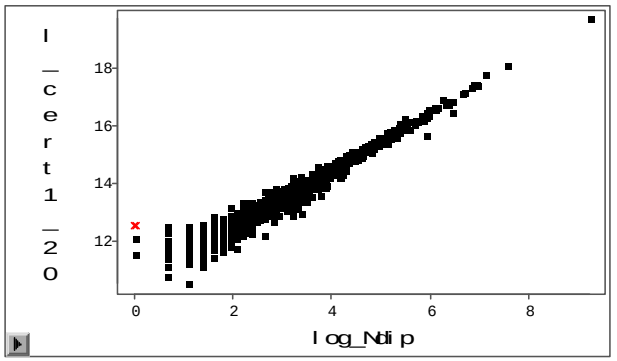


Preliminary analysis - Compensation of employees (2)

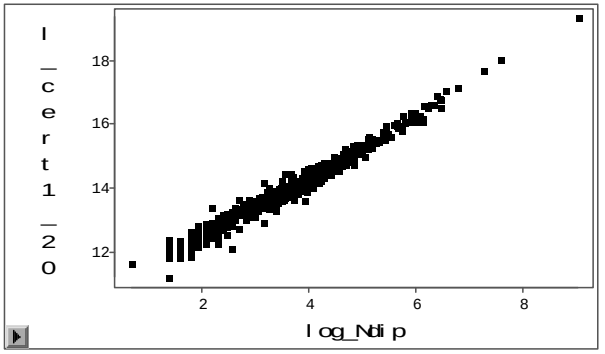
Center



South



Islands



Scatter-plot of municipalities (logarithmic scale):
Compensation of employees
and Number of employees

outliers in red

Non-response pattern

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.

- The number of municipalities that need to be imputed in 2012 (705) is equal for both variables
- The number of units with no missing values for both sources is equal to 7.121 and 7.132

Non-response pattern

Type	Information source				Compensation of employees		Intermediate Costs	
	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

Simulation study: steps

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.

Simulation study: imputation methods (1)

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), computed w.r.t. an auxiliary variable *X* (known for all the population):
$Y_{i,pc} = Y_i / X_i$.

a. Identify the NND donor *d* w.r.t. the matching variables statistically associated to the variable *Y*
b. Impute the value $Y_i^* = X_i \times Y_{d,pc} = X_i \times Y_d / X_d$
   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 w.r.t. 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.

Simulation study: imputation methods (2)

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:
a longitudinal approach is combined with NND methods.

Simulation study: indicators

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

Relative Bias (RB) - 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 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)

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

where $y_{true,k}$ and $y_{imp,k}$ are 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.

Imputation methods, Compensation of employees

NND:

for each municipality, the per-capita value of the target variable is computed w.r.t. the number of employees (*ratio hot-deck*):

$$Y_{i,pc} = Y_i / Nempl_i$$

Matching variables: S_2012/ Nempl, Surface, Pop_2012

Variable for imputation cells: Geographical macro region.

PMM:

a. multivariate linear regression model within each imputation cell:

$$Y_i = \alpha + \beta_1 S_i\_2012/Nempl_i + \beta_2 Pop_i\_2012 + \beta_3 Surface_i + e_i$$

b. minimum distance donor with respect to the value Y_i^P (predicted using the regression model)

Variable for imputation cells: Geographical macro region.

LNND:

Matching variables: Y_2011/Nempl, S_2011/Nempl, S_2012/S_2011, Population 2012, Surface

Variable for imputation cells: Geographical macro region

Imputation methods, Compensation of employees

Longitudinal deterministic methods:

Long EAC Sio
Long EAC Pop
Long Pop
Long Sio

update the value Y_i_2011 with an individual trend that is computed on an auxiliary variable (from Siope, Census, etc.) or with a median trend.
Variable for imputation cells: Geographical macro region * Population.

Mixed methods:

NND Long Mixed Pop
NND Long Mixed EAC
NND Long Mixed Sio

a. deterministic step: update the value Y_i_2011 with an individual trend computed on an auxiliary variable (Pop, Y, S)
   Variable for the imputation cells: Geographical macro region * Population.
b. non-deterministic step using a LNND method
   Variable for imputation cells: Geographical macro region

Imputation methods, Intermediate consumption

Same methods used for compensation of employees, *differences*:

NND:
the per-capita value of the target variable for each municipality is computed w.r.t. the variable Population:

$$Y_{i,pc} = Y_i / Pop_i\_2012$$

Matching variables: S_2012/Pop_2012, Nempl, Surface, Geographical characteristics

PMM:
a. multivariate linear regression model within each imputation cell:
$Y_i = \alpha + \beta_1 S_i\_2012/ Pop_i\_2012 + \beta_2 Surface_i + \beta_3 Nempl_i + e_i$

LNND:
Matching variables: Y_2011/Pop_2011, S_2011/Pop_2011, S_2012/S_2011, Nempl, Surface, Geographical characteristics

Longitudinal deterministic methods:
Variable for imputation cells: Geographical macro region, Pop_2012 (in classes) and Geographical characteristics.

Mixed methods:
Variable for imputation cells in the deterministic step: Geographical macro region, Pop_2012 and Geographical characteristics.

Results

- the imputation methods ensuring higher levels of accuracy in terms of RMSE are those exploiting the longitudinal information of units with missing data
- among them, 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
- results are confirmed also at regional level

Compensation of employees

Indicator	NND	PMM	NND long	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**

Intermediate consumption

Indicator	NND	PMM	NND long	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	\
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**

Conclusions

Results:

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:

- 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

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

Di Zio, M., Guarnera, U. (2009). Semiparametric predictive mean matching. *AStA - Advances in Statistical Analysis*. 93, 175-186

R.J.A. Little (1988). Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics*, 6, 3, pp. 287-296

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*

Thank you for your attention