Small area estimation for business statistics: a study on Italian small and medium enterprises[1]

Luzi O., Solari F., Rocci F.[2]

Abstract

In 2013, Istat has developed a new statistical register, called Frame-SBS, to support the annual production of structural business statistics (SBS) based on the massive and integrated use of administrative data. Further developments foresee to achieve improved estimates also for those SBS which are not covered by administrative sources. In this context, small area estimates (SAE) can be used in order to obtain more accurate results. The SAE methods start from the direct sample survey estimates toward some regression estimates obtained by using additional auxiliary information (e.g. administrative data), to obtain more efficient estimates. This paper presents an experimental application of SAE in the Frame-SBS context and the advantages that can be ensured by using it, in terms of increased quality and reliability of economic variables. Different types of auxiliary information and approaches are used in order to identify the best in the estimation strategy.

.

Key Words: Structural Business Statistics, administrative data, Small Area Estimation.

1. Introduction

The availability of new and detailed quantitative data on Italian businesses is a key factor for assessing the competitiveness and the performance of our economic system, that plays a central role to set up or fine tune policy measures oriented to guarantee productivity and employment growth. High quality information at high level of detail is essential in order to allow business analysts and policy makers to better analyze the characteristics and behavior of sub-populations of firms.

In last years, Istat has proceeded with the deep revision of its estimation strategy in the area of economic statistics, moving from a production model essentially based on direct survey data complemented by secondary information, to a new approach where administrative data (AD hereafter) are extensively used and direct survey data are collected in order to complete the coverage of specific sub-populations or target variables.

In this context, a new statistical register for structural business statistics (SBS) has been developed in 2014 (Luzi *et al.,* 2014), based on the integrated use of AD. In this register, called *Frame-SBS*, hereafter Frame, a number of key SBS variables are available at firm level for the overall target population (~4,4 million of units).

---

[1] The views expressed in this paper are solely those of the authors and do not involve the responsibility of their Institutions.

[2] Orietta Luzi, Istat, Via C. Balbo 16, Rome, Istat, e-mail:luz@istat.it. Fabrizio Solari, Istat, Via C. Balbo 16, Rome, email:solari@istat.it., Fabiana Rocci, Istat, Via C. Balbo 16, Rome , email: rocci@istat.it

As for the SBS variables which are not covered by the administrative sources currently feeding the register, quality gains can be further achieved by estimating micro-aggregates representing detailed economic sectors, sub-populations, geographical areas, i.e. small domains or small areas. To this aim, direct sample surveys and/or suitable estimation strategies exploiting as much as possible the increased amount of available auxiliary information are to be designed, in order to complement the register information. In this framework, Small Area Estimation (SAE hereafter, see Rao, 2003) can play a central role in order to guarantee more accurate and efficient estimates of business statistics at high level of detail.

The paper discusses the main advantages that the use of SAE can ensure in this framework in terms of increased quality and reliability of both estimates and economic indicators. We also present the results of an experimental application of SAE , where auxiliary information is tested to be exploited for improving the estimates of variables, that are not directly available in administrative sources. In the study, two different types of auxiliary information are taken into account, one from the Italian Business Register (Asia) and the other one derived from the Frame itself, in order to assess also the potential information gathered by the new SBS system. Further, also the presence of a correlation among small domains has been considered, by means the definition of an area level linear mixed model with correlated area random effects.

The paper is structured as follows. In Section 2 the contents and the potentials of the Frame are delineated. Small area estimation methods are described in Section 3. Section 4 and 5 are devoted to the illustration of the case study and of the results, respectively. Finally, conclusions and future work are reported in Section 6.

2. Exploiting the potential of administrative data in the SBS area

The economic analysis of the factors affecting the competitiveness of the modern industrial systems increasingly requires complex statistical information, able to combine aggregated measurements with quantitative evidence on the degree of heterogeneity within the system of enterprises. The greater is the complexity and heterogeneity of the structure of an economy, the greater the information connected to an analysis based on very detailed aggregate data. This applies particularly to the analysis of the Italian production system, which is characterized by a large presence of small enterprises – the firms with less than 10 persons employed account about 50% of total employment - and of highly specialized sectors.

The request for a more coherent approach to the measurement of micro/macro aspects has stimulated Istat to move towards a massive use of AD to feed the need of standard economic variables for large populations of businesses. In this framework, SBS play a central role for the analysis of businesses productivity and competitiveness. In Italy, SBS has been traditionally estimated based on data collected through two annual surveys: the sample survey on *Small and Medium Enterprises* (SME hereafter) (about 100,000 sampled enterprises with less than 99 persons employed representing a population of about 4.3 million of units), and the total survey on Large Enterprises (LE hereafter) (about 11,000 enterprises with 100 or more persons employed). Both surveys estimate totals of profit-and-loss accounts variables, employment, investments etc. in the

industrial, construction, trade and non-financial services sectors, at different level of disaggregation by economic activity and business size (in terms of number of persons employed), as requested by the SBS Eurostat Regulation. A large number of secondary variables are also estimated, e.g. for National Accounts estimation purposes.

In the view of exploiting as much as possible the available micro-information from any AD sources on enterprises, in 2014 the combined use of information from *Financial Statements, Sector Studies, Unico Model, IRAP, Social Security Data* has made possible to achieve the statistical register Frame (Luzi *et al.*, Q2014; Curatolo *et al*., 2016). In the Frame, firm-level data for the *main* economic variables are directly acquired from the AD sources (covering about the 95% of the whole SBS target population) after a harmonization phase. The combination of the AD source almost fully covers the main profit-and-loss accounts variables (main SBS hereafter), like *production value, turnover, intermediate costs, value added, wages, labor cost*. As a consequence, the corresponding estimated totals can be obtained at any level of detail (economic sector, size in terms of number of persons employed, territorial) and for specific sub-populations of enterprises (e.g. exporters, sub-contractors, micro-enterprises, etc.) by summing-up variables micro-data[3]. The availability of census-like data for the main SBS has stimulated the implementation of a more comprehensive information system based on the integration of the Frame with other Istat statistical registers (e.g. the *Trade by Enterprise Characteristics* register, and the register on *employment in the Italian companies*). The aim is to allow for the joint analysis of economic performance, internationalization, employment and territorial structure of enterprises at high level of detail.

In the Frame context, the estimation of the economic account variables which are *components* of the *main* SBS (i.e. which are related to them through mathematical constraints) is performed by using a design based/model assisted approach known as projection estimator (Kim *et al.*, 2011). This approach exploits the randomization process of the SME sample selection under consistency constraints, at pre-defined levels of detail. The remaining key SBS (such as *capital stocks, investments, structure of intermediate costs for goods and services, structure of labor costs*) which are not covered by the utilized AD sources, are currently estimated based on the direct survey results. In this context, however, SAE can play a central role in order to obtain more efficient estimates starting from those direct estimates, for level of domain for which sample design methods can result inefficient. Indeed, further auxiliary information and the relationships among the estimates along different domains can be further exploited to improve the quality of the final estimates at a given level of aggregation.

3. Small Area Estimation for Business Statistics: the model

Model based small area estimation (SAE) techniques use explicit modelling for relating unit survey data or area direct estimates to a set of auxiliary variables. The most widely used class of models is linear mixed models, which include area random effects to account for between area variation beyond that explained by auxiliary information.

In the unit level model (Battese et al., 1988), individual survey data are required for both target and auxiliary information, while at population level totals or mean values of

[3] It has to be mentioned here that for these variables statistical imputation is adopted to compensate for the sources incompleteness and/or under-coverage w.r.t. specific SBS sub-populations.

auxiliary variables are needed for each small area. When unit level survey data are not available, an area level mixed model estimator can be implemented (Fay and Herriot, 1979). Area level models require strong auxiliary information at area level, which should be available for sampled and non-sampled areas. Moreover, direct survey estimates and their corresponding sampling variance need to be available for each sampled area.

By means of mixed models methodology (see, for instance, Jiang and Lahiri, 2006), a best linear unbiased predictor (BLUP) is used to estimate small area parameters. Since the variance components are usually unknown, the correspondent empirical best linear unbiased predictor (EBLUP) is used instead.

The basic model of this type is the Fay-Herriot model, introduced by Fay and Herriot (1979), to estimate per capita income for small places in the United States.

Let θ_d be the parameter to be estimated for each domain d. A linking model between θ_d and a set of covariates, whose values are known for each domain of interest, is assumed. Using matrix notation, we can write:

$$\boldsymbol{\theta} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{u}, \qquad (1)$$

where $\mathbf{X}$ is the covariate matrix and $\mathbf{u} = (u_1,...,u_D)$ is the vector of area effects, assumed to be independently distributed with mean zero and variance σ_u^2.

Besides, let us specify the sampling model. A design unbiased direct estimators $\hat{\theta}_d$ is supposed to be available (but not necessarily for all the domains), that is

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \mathbf{e}, \qquad (2)$$

where $\mathbf{e} = (e_1,...,e_D)$ is the vector of sampling errors associated with the direct estimators, for which, for $d = 1,...,D$, $E(e_d \mid \theta_d) = 0$, i.e., the direct estimator is assumed to be unbiased, and $V(e_d \mid \theta_d) = \varphi_d$, where the variances φ_d are supposed to be known in order to avoid identifiability problems.

Combining equations (1) and (2) the following linear mixed model is obtained:

$$\hat{\boldsymbol{\theta}} = \mathbf{X}^T\boldsymbol{\beta} + \mathbf{u} + \mathbf{e}. \qquad (3)$$

On the basis of model (3), for each domain d the empirical best linear unbiased estimator (EBLUP) is

$$\hat{\theta}_d^{\text{EBLUP}} = \gamma_d\hat{\theta}_d + (1-\gamma_d)\mathbf{X}_{.d}^T\hat{\boldsymbol{\beta}}\ , \qquad (4)$$

where $\gamma_d = \hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \varphi_d)$ is the weight of the direct estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\hat{\boldsymbol{\theta}}$ is the generalized least square (GLS) estimator of the

regression coefficient vector, with $\hat{\mathbf{V}} = \hat{\sigma}_u^2(\mathbf{I}_D + \boldsymbol{\Phi})$ is the estimate of the model variance matrix of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\Phi} = \text{diag}(\phi_1,...,\phi_D)$. The estimation for the parameters σ_u^2 and $\boldsymbol{\beta}$ is attained iteratively by means, for instance, of ML or REML estimation, assuming normality of the random effects, or by the method of fitting constants. Computational details can be found in Rao (2003, pp. 115-120).

Nevertheless, if information at unit level is available, then, under the hypothesis of homoscedasticity, the variance φ_d can be estimated from a unit level model (see, for instance, Rao, 2003) or a generalized variance function. Anyway, this would affect the MSE of the predicted domain values (Bell, 2008).

For more details, methods for estimation of $\hat{\sigma}_u^2$ see Rao (2003, pp. 115-120). Details on the estimation of the MSE are given in Rao (2003, pp. 103, 128-130).

Under the classic model specification (3) the area specific random effects are assumed to be independent. This hypothesis means that no correlation structure of the data is considered. Instead, it is reasonable to assume the random effects between the neighbouring areas (defined, for example, by a contiguity criterion) being correlated and the correlation decaying to zero as distance increases. Petrucci *et al.* (2005) extended model (3) to allow for correlated area effects. In details, the uncorrelated vector of random effects u is substituted with a correlated vector of random effects v.

Let $\mathbf{v} = (v_1,...,v_D)$ follow a Simultaneously Autoregressive (SAR) process with proximity matrix W, unknown autoregression parameter ρ (see Anselin, 1992; Cressie, 1993), and let u be defined as before, i.e.,

$$\mathbf{v} = \rho\mathbf{W}\mathbf{v} + \mathbf{u}. \tag{5}$$

If the matrix $(\mathbf{I}_D - \mathbf{W})$ is assumed to be non-singular, then v can be expressed as

$$\mathbf{v} = (\mathbf{I}_D - \mathbf{W})^{-1}\mathbf{u}. \tag{6}$$

Equation (6) implies that v has mean vector 0 and covariance matrix G equal to

$$\mathbf{G} = \sigma_u^2[(\mathbf{I}_D - \rho\mathbf{W})^T(\mathbf{I}_D - \rho\mathbf{W})]^{-1}. \tag{7}$$

Then, model (3) becomes

$$\hat{\boldsymbol{\theta}} = \mathbf{X}^T\boldsymbol{\beta} + (\mathbf{I}_D - \rho\mathbf{W})^{-1}\mathbf{u} + \mathbf{e} \tag{8}$$

Under model (8), the EBLUP of the quantity of interest θ_d is

$$\hat{\theta}_d^{\text{SEBLUP}} = \mathbf{X}_{.d}^T\hat{\boldsymbol{\beta}} + \mathbf{b}_d^T\hat{\mathbf{G}}\hat{\mathbf{V}}^{-1}(\hat{\boldsymbol{\theta}} - \mathbf{X}^T\hat{\boldsymbol{\beta}}), \tag{9}$$

Where $\hat{\mathbf{G}}$ is obtained from (7) replacing the variance components σ_u^2 and ρ with their estimates $\hat{\sigma}_u^2$ and $\hat{\rho}$, $\hat{\mathbf{V}} = \hat{\mathbf{G}} + \mathbf{\Phi}$, and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \hat{\boldsymbol{\theta}}$ is the GLS estimator of the regression parameter $\boldsymbol{\beta}$, and $\mathbf{b}_d$ is D-dimensional vector $(0, ..., 0, 1, .0, .., 0)$ with 1 in the d-th position. The vector of regression coefficient $\boldsymbol{\beta}$ and the variance components σ_u^2 and ρ can be estimated either by ML or REML methods. Details for the estimation of the model parameters and the MSE can be found in Petrucci *et al.* (2005).

4. The empirical study

Among the remaining key SBS not covered by the used AD sources, the variable taken into account for the empirical study is *Total depreciation of fixed assets*. The final aim is to assess if it is possible to achieve more efficient estimates by exploiting the relationship among the released estimates of the given variables across different domain.
For SBS purposes, estimates at different level of details are required: 4-digits Nace-code, 3-digits Nace-code by 7 size classes, 2-digits Nace code by Regions (NUTS-2). This first study takes into account the estimates released at 4-digits Nace-code.

As auxiliary information, two different variables have been studied, that are derived from different sources. The comparison between the results, according to the different auxiliary information, makes it possible to assess the different potential information that can be gained.

The two variables used as auxiliary information are:

1. Proxy of the Turnover, delivered by the Italian BR (Asia);
2. Value Added, delivered by the Frame.

The target variable is correlated with both the Turnover and the Value Added, nevertheless the different origin of the auxiliary variables can play an important role in gathering information and hence to achieve a significant gain in efficiency. In the first case, the variable is derived from the BR, in the second one it is one of the main variables elaborated through the Frame process.

In this view, the results are analyzed by comparing, for each 4-digit domain, the MSE of the direct estimates to the MSE of the SAE estimates. Four different types of area level SAE strategies have been tested according to the presence of correlation among the economic activities, and the source of information used. When it is assumed to be no correlation among the economic activities, the adopted model specification is the Fay-Herriot model, while when a correlation structure is assumed for the economic activities, we use the model specification given in Petrucci *et al.* (2005). In this case, economic activity having the same first three digits are considered as neighbor in the proximity matrix.

Then, all the SAE strategies are resumed in the following table.

Table 1: SAE strategies used in the case study

		Source of Information	
		Asia	Frame
SAE method	Fay Herriot	FH_ASIA	FH_FRAME
	Spatial Fay Herriot	SFH_ASIA	SFH_FRAME

The results of the experimental study are reported in the following section.

5. Results

The results of the empirical study have been compared in terms of estimates variability, to assess that using auxiliary information can help in achieving more efficient estimates.

The analysis of the differences is run in two steps: at first, the comparison between the direct and the SAE estimates is done, in order to ascertain whether the SAE method can really improve the results efficiency. Afterwards, once the SAE method is assessed to gather a gain in efficiency, the comparison is between the two kind of SAE estimates, that vary according to the correlation scheme and the auxiliary information. This allows to go further deeply in the analysis to understand the potential of the several informative contexts taken into consideration.

All the comparison between the methods are performed in terms of efficiency, i.e. ratio of the corresponding MSEs. Obviously, when the efficiency is greater than 1 then the method corresponding to the denominator of the efficiency indicator is more efficient than the other method.

We report in Figure 1 the efficiency of FH and SFH models with respect to the direct estimates, when using either Asia or Frame as source of auxiliary information. Figure 2 displays the efficiency of SFH model with respect to FH. Also in this case, both plots for Asia and Frame are reported.

The Figure 1a and 1b b show that efficiency is always greater than 1, showing the better efficiency of the SAE estimates with respect to the direct estimates. Furthermore, SFH estimates results more efficient than FH estimates, implying that there is evidence for a correlation between the economic activities. In particular, Figure 2 displays that the gain in efficiency deriving from SFH is higher when Asia is used, that is the less informative source in terms of correlation with the target variable. Therefore, using a very informative source allows to obtain good estimates even by using the simplest model specification.

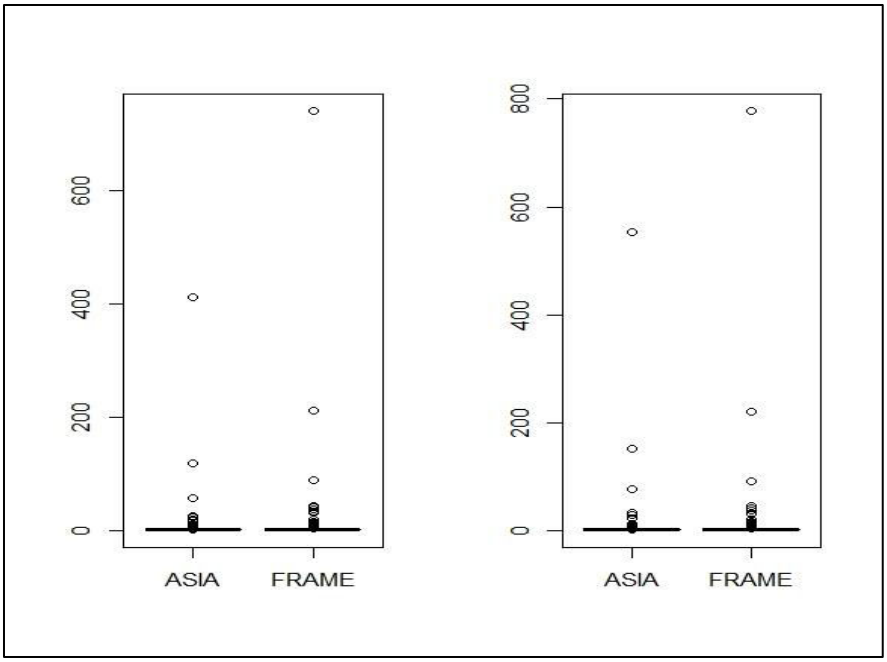

Figure 1a: Boxplot of the efficiency of FH (left) and SFH (right) estimates with respect to the direct estimates, using Asia and Frame.

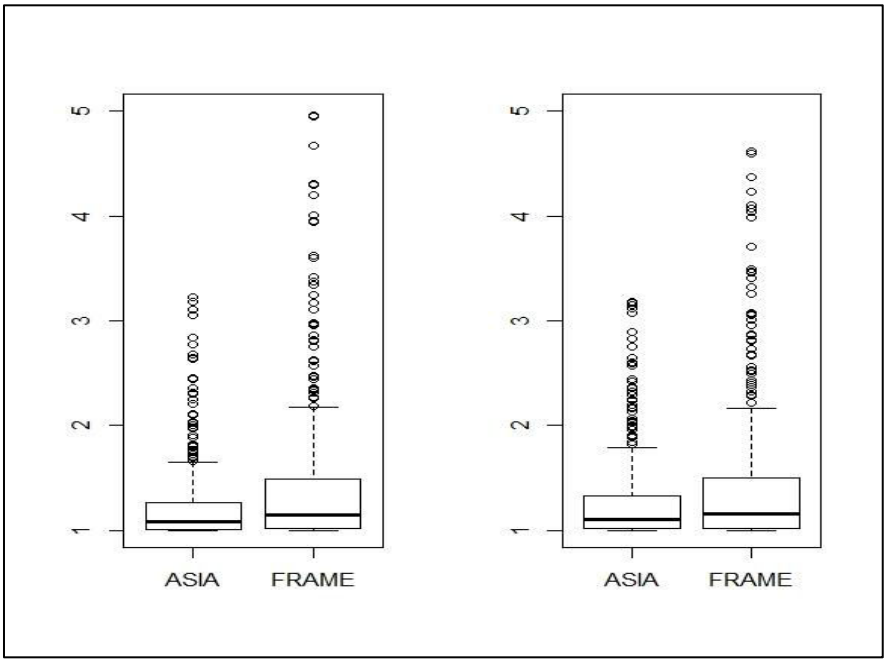

Figure 1b: Boxplot of the efficiency of FH (left) and SFH (right) estimates with respect to the direct estimates, using Asia and Frame. Only the domains for which the CV of direct estimates is less than 5% are reported.

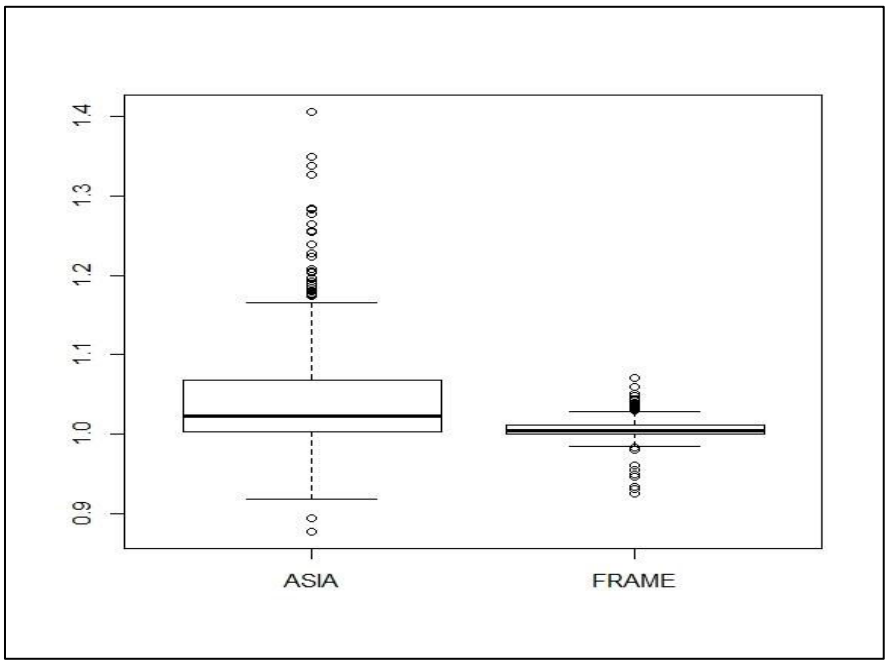

Figure 2: Boxplot of the efficiency of SFH with respect to FH estimates using Asia and Frame.

6. Conclusions and future work

The analysis and results presented in this paper show that new methodological development in exploiting administrative data in the context of the statistical register Frame can contribute to further improve and release a reliable Structural Business Statistics system.

In this view, SAE represents a very efficient method to produce very detailed and accurate economic statistics by jointly exploiting surveys estimates and admin data as auxiliary information.

Hence, it can be worthwhile to apply SAE methods to the direct SBS estimates in order to produce more reliable aggregated estimates.

The further steps of the research in this context would be to assess how to design the estimation methodology in order to assure the compliance of the SBS European regulation. In this perspective, the future activities will be focused on the following areas:

1. assessing the available information to perform the best SAE model: the Frame variables seem to gather good perspective from this point of view;
2. exploiting the opportunity for innovative applications – taking into account the peculiarities of economic variables and the characteristics of the Istat surveys on enterprises:
   - *benchmarking*: ensuring for large domains coherence between direct estimates and aggregated SAE estimates related to the small domains

included into the large one;
- design *multi-domain sampling* designs in Istat surveys on enterprise;
- development of new strategies to manage consistency and confidentiality constraints in the new information context.

References

Anselin, L. (1992). Spatial Econometrics. Method and Models, Boston: Kluwer Academic Publishers.

Banerjee, S., Carlin, B.P., and Gelfand, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data, New York: Chapman and Hall.

Battese, G.E., Harter, R.M., and Fuller, W.A. (1988). An error component model for prediction of

county crop areas using survey and satelite data. Journal of the American Statistical Association, 83(401): 28–36.

Bell, W.R. (2008). Examining sensitivity of small area inferences to uncertainty about sampling error variances. In Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, 327–334.

Cressie, N. (1991). Small-area prediction of undercount using the general linear model. In Proceedings of Statistics Symposium 90: Measurement and Improvement of Data Quality, Ottawa: Statistics Canada, 93–105.

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

Fay, R.E., and Herriot, R.A. (1979)., Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74(366): 269–277.

Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation (with discussion). Test, 15: 1–96.

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.

Luzi, O., Guarnera, U., Righi, P. 2014. The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. European Conference on Quality in Official Statistics (Q2014). Vienna, 3-5 June.

McCulloch, C.E. and Searle, S.R. (2001). Generalized, Linear, and Mixed Models. Chichester: John Wiley & Sons.

Petrucci, A., Pratesi, M., and Salvati, N. (2005). Geographic information in small area estimation: small area models and spatially correlated random area effects. Statistics in Transition, 7(3): 609–623.

Rao, J.N.K. (2003). Small Area Estimation, London: Wiley.