Small domain estimation of business statistics by using multivariate skew normal models

SMALL DOMAIN ESTIMATION OF BUSINESS STATISTICS USING MULTIVARIATE SKEW NORMAL MODELS

Summary

Small domain business statistics are becoming important for better planning business policies. We focus on the estimation of the averages of value added and labour cost in small domains. To take into account the positive skewness in the distribution of outcomes and the correlation between them, we propose a bivariate skew-normal small area model. Estimates are obtained from real survey data. The performance of the proposed estimator is evaluated based on both survey data and a synthetic firm population. Results show that the model proposed increases the estimates reliability and that the estimates obtained make it possible to perform detailed regional economic studies.

Keywords: Firm sample surveys, Hierarchical Bayesian modeling, Regional economic studies, Skew-Normal distribution

1. Introduction and motivations

Small domain business statistics are becoming more and more important for better evaluating regional and sectoral firms' competitiveness. Unfortunately, data on economic aggregates and indicators are rarely available at a local level and/or for firms' categories. In fact, in this context estimates are usually obtained from sample survey data and then founded on design-based (direct) estimators that, due to small sample size, cannot provide reliable estimates for small domains (or, put it differently, they produce estimates with a large error). Obviously, the problem can be overcome by increasing survey sample size, but this solution is usually not pursued because it consumes time and budget resources. The Small Area Estimation (SAE) methods are devoted to producing reliable small area (or small domain) estimates based on the information available, by relying on model-based (indirect) estimators. More in detail, SAE methods use models to predict

estimates of interest for all the small areas, and reliable estimate for one small area is obtained by "borrowing strength" from sample survey data collected in other small areas. Predictors are auxiliary data available at a small area level and measured without (or with a small) error, such as administrative or census data. SAE models produce good estimates provided that good auxiliary variables are available and the model is correctly specified. For an up-to-date review on SAE methods, see Pfeffermann (2013) and Rao and Molina (2015).

In spite of the rapid growth in SAE literature over the past ten years, the small area estimation of social and economic parameters has so far mostly concerned the small area estimation of poverty or employment indicators, whereas it has seldom been used to estimate parameters related to firm activity and performance. Only recently has the literature on small area estimation methods focused on business survey data (Chandra and Chambers, 2011; Chandra et al., 2012; Burgard et al., 2014; Schmid et al., 2016). The reasons for this increased interest can be found in the needs expressed by the National Statistical Institutes for improving official local business statistics, and by economists and policy makers for better monitoring enterprise performance and promoting entrepreneurship at a regional level.

In this paper, we propose a model-based small area estimator of two important business aggregates: the averages of Value Added (VA) and Labour Cost (LC). We focus on VA and LC because they form the basis of some important economic competitiveness indicators that are useful for obtaining a mapping of firms' performance and important drivers of the changes in living standards (OECD, 2016): labour productivity (VA/number of employees), cost competitiveness (LC/number of employees), and gross profitability ((VA-LC)/revenue). Monitoring and promoting competitiveness growth and sharing related gains, through the creation of new technologies, investment in human capital and production innovation, is important to foster both growth and the reduction of inclusion gaps. Disparities in competitiveness and productivity arise among countries and within countries (OECD, 2016; Eurostat, 2016). Besides, industrial sectors differ from each other with respect to their competitiveness and productivity. In most countries, the economic growth in post-crisis period

has been much weaker than in the pre-crisis period, but this decline is of different intensity in different sectors, a disparity that is likely to have a negative impact on well-being and inequalities. A further element of competitive heterogeneity is due to firms' size. In most countries, the gaps between micro, small and large firms remain relatively high. Larger firms generally show higher levels of productivity and competitiveness compared to small and micro firms, and this gap increased in the manufacturing sectors from 2008 to 2013 (OECD, 2016).

In a situation of great heterogeneity within countries and firms' categories, the evaluation of both regional competitiveness and regional and sectoral economic disparity has become more and more important, also in order to detect the presence of competitive regions/firms' categories in less competitive countries and vice versa. One of the main aims of the Europe 2020 strategy, the plan for long-term recovery adopted by the European Union, is the reduction of regional disparities. In this context, accurate business statistics on sub-national regions and business categories could support regional and sectoral economic decisions.

As already mentioned, due to insufficient sample size, National Statistical Institutes are able to produce estimates only to a certain level of detail. For example, Eurostat produces (gross) value added estimates by EU NUTS3 regions (following the Nomenclature of Territorial Units for Statistics, Eurostat, 2015) and NACE Rev. 2, 1 digit sectors (following the Statistical classification of economic activities in the European Community). The availability of geographically disaggregated estimates, computed by firm sector and size, could help policy makers to implement better-targeted and more effective policies.

The above considerations motivate our interest in estimating economics aggregates by cross-classifying regions, firm size classes and economic sectors. It is worth noting that small areas or domains of interest can be defined with others criteria with respect to those we select: the small area problem arises whenever statistical data are gathered from a sample survey unable to support reliable estimates at a disaggregated level, because the domain sample size is too small.

We focus on Italian manufacturing industrial sectors and we perform a real data application based on information collected by the Italian National Statistical Institute (ISTAT) in the small and medium enterprises sample survey. To evaluate the properties of the estimators through a simulation study, we resort to a synthetic firm population generated by data on real firms (Kolb et al., 2013). This study has been included in the BLUE-Enterprise and Trade Statistics (BLUE-ETS) project, financially supported by the European Commission.

In specifying a small domain estimation model for business data, some particular issues that arise in business surveys (Cox et al., 1995; Rivière, 2002) must be taken into account. One of the most relevant is the fact that business data, due to the presence of a majority of small firms, are generally characterized by a positively skewed distribution. In addition, firm aggregates representing totals are generally highly related amongst themselves due to an underlying factor which is firm size.

We propose a small domain model that deals with both these issues, by operating in the "area level" model framework (Rao and Molina, 2015, p. 123). Since area level models are estimated starting from design-based estimates, they easily incorporate information on sampling design and on non-response adjustments. The area level model consists of i) a "sampling" model, specifying direct design-based estimates as measurements of an underlying area descriptive parameter, the variance of which is considered as known; the input for the model then consists of design-based estimates, called "direct estimates", and their associated estimated variances, ii) a "linking" model, relating the area parameters to auxiliary information accurately known at the area level and to area specific random effects.

Small area models often rely on the assumption of normality for direct estimators and area (random) effects, which are inadequate for asymmetric outcomes. Even in the presence of skewed data, in small area literature the assumption of normality at the sampling model level is often justified invoking the Central Limit Theorem. However, when dealing with small sample sizes, this assumption might be hardly sustainable. To take into account the asymmetry of data, we relax the normality assumption of the most popular so-called normal-normal model (Fay and Herriot, 1979)

by adopting a skew-normal distribution both in sampling and in linking models. The class of skew-normal distributions proves to be quite useful in modelling real data-sets and enjoys remarkable properties in terms of mathematical tractability (Azzalini, 1985; Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999, 2003). In particular, the skew-normal specification offers some advantages with respect to other non-symmetric distributions, because it includes the normal distribution as a special case and allows for modelling zero and negative values. In the context of the area level model-based estimation and of the Bayesian framework for inference, Ferraz and Moura (2011) tackle the problem of skewness by assuming a skew-normal distribution at the sampling model level and in a univariate context. In addition, Fabrizi and Trivisano (2010), in their study on the use of a robust linear mixed model for small area estimation, propose the assumption of skewed Exponential Power distribution at the linking model level. In the frequentist framework, Slud and Maiti (2006) propose a small area model based on log-normality assumption at the sampling model level. The skew normal distribution is considered by Diallo and Rao (2014), who derive empirical best estimators for unit level models where a skew normal distribution is assumed for both area-specific effects and random errors. Furthermore, Diallo (2014) proposes a replication based method for estimating MSE under SN small area models.

Furthermore, in order to take advantage of the relationship usually observed within business data, we propose a multivariate extension of such a skew-normal small area model, which considers the high correlation among direct estimators and/or the correlation among area-specific random effects. The multivariate specification of the small area model offers some advantages over the univariate one. Univariate small area models improve on the traditional estimates by "borrowing strength" from related small areas or relevant covariates which are available for the population. A further improvement in estimate reliability can be obtained in a multivariate approach by 'borrowing strength' from related dependent variables. This approach could provide better estimates, by taking into account the correlations between the response variables after conditioning on the auxiliary variables. A multivariate extension of the Fay–Herriot model is considered in Datta et al. (1991,

1996) where information on three- and five-person families is included in order to estimate the median income for four-person families for the 50 U.S. states and the District of Columbia. Fabrizi et al. (2008) propose a multivariate small area estimation approach to obtain reliable estimates for certain poverty parameters. All these studies focus on normality assumption. Only recently, some attention has been devoted to multivariate small area models relying on non-normal distributions. The multivariate beta regression with application to small area estimation is proposed by Souza and Moura (2012). Ferrante and Trivisano (2010) propose a multivariate small area estimation approach for count data based on the multivariate Poisson-log normal distribution. A multivariate logistic-normal model is adopted by Fabrizi et al. (2011) with the aim of estimating poverty rates based on different thresholds.

This paper is organized as follows. In Section 2 we provide a brief description of the multivariate skew normal distribution, while presenting the multivariate skew normal small area model. Section 3 contains a description of the strategy of business outcome estimation based on the proposed small area model. In Section 4 we evaluate the performance of the estimators proposed with reference to real survey data and compare it with some competitor estimators on the basis of certain performance criteria. In Section 5 the properties of the estimators proposed is evaluated by carrying out a simulation study. Results show that the consideration of the asymmetry of data and the correlation between outcomes greatly increases the reliability of the estimates, and the estimator proposed offers good randomization properties. In Section 6 we present an example of how the estimates we obtained could be important in interpreting the regional and industry disparities in labour productivity. Section 7 offers some conclusions.

2. Model and prediction strategy

A multivariate version of the skew-normal distribution is defined in Azzalini and Dalla Valle (1996). Vector **Y** has a K-multivariate skew normal distribution (k=1,…,K), $SN_K(\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\lambda})$, with

vector of location parameters $\boldsymbol{\xi}$, dispersion matrix $\boldsymbol{\Omega}$ and vector of shape parameters $\boldsymbol{\lambda}$, if its density function can be expressed by:

$$g(\mathbf{y};\boldsymbol{\xi},\boldsymbol{\Omega},\boldsymbol{\lambda}) = 2\phi_K(\mathbf{y}-\boldsymbol{\xi};\boldsymbol{\Omega})\Phi\left(\boldsymbol{\lambda}^T\boldsymbol{\omega}^{-1}(\mathbf{y}-\boldsymbol{\xi})\right) \tag{1}$$

where $\phi_K(\mathbf{x};\boldsymbol{\Omega})$ is the density of a multivariate normal distribution with zero mean, $N_K(\mathbf{0},\boldsymbol{\Omega})$, $\Phi(\cdot)$ is the cumulative function of the univariate standard normal distribution and $\boldsymbol{\omega} = Diag(\boldsymbol{\Omega})^{1/2}$. For $\lambda_k = 0$ the skew-normal distribution is the normal and for $\lambda_k \to \infty$ the skew-normal converges to the half-normal distribution.

The marginal distribution of Y_k is the scalar skew-normal $SN\left(\xi_j, \omega_j^2, \tilde{\lambda}_j\right)$, where $\tilde{\lambda}_j = \delta_j \Big/ \sqrt{1-\delta_j^{\,2}}$, and vector $\boldsymbol{\delta}$ may be obtained from the parameters of the density function as follows: $\boldsymbol{\delta} = \frac{1}{\sqrt{1+\boldsymbol{\lambda}'\overline{\boldsymbol{\Omega}}\boldsymbol{\lambda}}}\overline{\boldsymbol{\Omega}}\boldsymbol{\lambda}$ and $\overline{\boldsymbol{\Omega}} = \boldsymbol{\omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\omega}^{-1}$ (Frühwirth-Schnatter and Pyne, 2009). The expected value of the marginal distribution is:

$$E(\mathbf{Y}) = \boldsymbol{\xi} + \boldsymbol{\omega}\boldsymbol{\delta}\sqrt{\frac{2}{\pi}}\,. \tag{2}$$

2.1.The Multivariate Skew Normal small area model

Based on the skew normality assumption, we propose the following small area model. Let the $\hat{\theta}_{ik}$ be the direct estimator of the outcome parameter θ_{ik} in the i-th domain (i=1,…,m), referred to the k-th outcome (k=1,…,K). In the sampling model the vector $\hat{\boldsymbol{\theta}}_i$ of direct estimators is supposed to follow a multivariate skew-normal distribution:

$$\hat{\boldsymbol{\theta}}_i \Big| \boldsymbol{\theta}_i^*, \boldsymbol{\lambda}, n_i, \boldsymbol{\Omega}_i \;\sim\; SN_K\left(\boldsymbol{\theta}_i^*, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i\right) \tag{3}$$

$$\lambda_{ik} = \lambda_k \Big/ \sqrt{n_i} \tag{4}$$

In this model, each shape parameter is set equal to a common parameter divided by the square root of the sample size, so that when the sample size increases, the shape parameter tends to zero and the

skew normal tends to the normal distribution. Gupta and Kollo (2003) give a formal justification for this assumption. As is customary, we assume that the elements of matrix $\mathbf{\Omega}_i$ are known, and substitute them with their respective estimates.

We propose the specification of a multivariate skew-normal distribution for the linking model also. The assumption of normality for the small domain parameters is indeed difficult to justify, and by allowing a non-symmetric distribution for the random effects also, we may increase the flexibility of the model at the expense of an additional complexity in the model.

Hence in the linking model:

$$\boldsymbol{\theta}_i^* | \boldsymbol{\mu}_i, \boldsymbol{\lambda}_V, \mathbf{\Omega}_V \sim SN_K(\boldsymbol{\mu}_i, \mathbf{\Omega}_V, \boldsymbol{\lambda}_V) \tag{5}$$

where the location parameters are a linear function of some auxiliary area level variables:

$$\mu_i = \mathrm{x}_{ik}^T \beta_k . \tag{6}$$

Our parameters of interest are the expectations of the marginal distributions of $\hat{\boldsymbol{\theta}}_i$ under the skew normal model described which, according to eq. (2), is given by:

$$\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^* + \boldsymbol{\omega}_i \boldsymbol{\delta}_i \sqrt{\frac{2}{\pi}} \tag{7}$$

where $\boldsymbol{\omega}_i = Diag(\mathbf{\Omega}_i)^{\frac{1}{2}}$ and $\boldsymbol{\delta}_i = \frac{1}{\sqrt{1 + \boldsymbol{\lambda}_i' \overline{\mathbf{\Omega}}_i \boldsymbol{\lambda}_i}} \overline{\mathbf{\Omega}}_i \boldsymbol{\lambda}_i$ and $\overline{\mathbf{\Omega}}_i = \boldsymbol{\omega}_i^{-1} \mathbf{\Omega}_i \boldsymbol{\omega}_i^{-1}$.

Note that, as done by Ferraz and Moura (2011), to obtain a continuous transition from non-normality to normality, in equation (5) we model the location parameter $\boldsymbol{\theta}_i^*$ of the skew normal distribution of the sampling model, rather than modelling the mean of the skew normal distribution $\boldsymbol{\theta}_i$. So that when the sample size increases, the shape parameter, $\lambda_{ik} = \lambda_k / \sqrt{n_i}$, tends to zero, the mean $\boldsymbol{\theta}_i$ tends to the location parameter $\boldsymbol{\theta}_i^*$ (see equation (7)), and the sample distribution of $\hat{\boldsymbol{\theta}}_i$ converges to the normal distribution, as must be to satisfy the Central Limit Theorem.

Since we assume multivariate skew normality both at the sampling and at the linking model, we named this model the Multivariate Skew Normal-Skew Normal model.

2.2 Prediction strategy

In this study, as far as estimation is concerned, we have adopted a Hierarchical Bayesian approach implemented by means of a Monte Carlo Markov Chain (MCMC) computational method. This approach to inference has a number of potential benefits for small area estimation (for an up-to-date review, see Rao and Molina, 2015, section 10): it makes it possible to easily manage distributional assumptions other than the normal one, to obtain straightforward estimates for areas with no sample information, to capture the uncertainty about all parameters through posterior distribution, and to deal with multivariate models with correlations among design-based estimators and among random effects.

Under the Hierarchical Bayesian framework, we assume a quadratic loss and define $\hat{\boldsymbol{\theta}}_i^{HB} = E(\boldsymbol{\theta}_i \mid data)$ as the point predictor for $\boldsymbol{\theta}_i$ and the posterior variance $Var(\boldsymbol{\theta}_i \mid data)$ as a measure of the precision of the estimator. The posterior variance, that is, the variance of the posterior density, describes the uncertainty of the parameter, which is a random variable in the Bayesian paradigm. For complex problems, the evaluation of the posterior variance requires the evaluation of high dimensional integrals and MCMC methods are used for this purpose. These methods generate samples from the posterior distribution and then use the simulated samples to approximate the posterior quantities of interest. To implement MCMC calculations, we use the OpenBugs open source software (Thomas et al., 2006; Spiegelhalter et al., 2002), widely adopted in the analysis of a large class of Bayesian models, particularly in the estimation of applied small area hierarchical models (Rao and Molina, 2015, p.339). OpenBugs has the further advantage that it can be easily run from R, the software we use for the simulation study.

OpenBugs Program does not take the skew normal distribution into consideration. There are two possible solutions to this problem: a) hierarchically generating samples of the skew normal density

by using the stochastic representation (Henze, 1986; Ferraz and Moura, 2011), b) explicitly writing the skew normal density formula into the BUGS code, which can be done by using what is known as "the trick for specifying new distributions" (Spiegelhalter et al., 2002). We explored both these solutions and adopt the latter, since the former does not work as well from an MCMC standpoint, by performing with extremely slow convergence and bad mixing of chains associated with hyper-parameters.

3. The estimation of business outcomes based on the multivariate skew normal-skew normal small area model

3.1. Data

We rely on the firms' official data collected in the small and medium enterprises (SME) sample survey (1-99 employees), conducted by the Italian National Statistical Institute (ISTAT). Data are provided to us within the framework of the BLUE-ETS project. The survey sampling design is stratified and strata are defined by cross-classifying NACE 4 Rev. 2, 2 digits sectors, Italian administrative regions (NUTS2), and firm size. A detailed description of the SME survey can be found in Faramondi et al. (2010). We consider the data collected in 2008, which refer to 25,925 firms in manufacturing sectors.

With reference to the outcomes we focus on, i.e. VA and LC, ISTAT provides reliable estimates for domains defined alternatively by: i) cross-classification of administrative region and economic activity (NACE Rev. 2, 2 digits), ii) cross-classification of size (in classes) and economic activity (NACE Rev. 2, 3 digits), iii) economic activity (NACE Rev. 2, 4 digits). Hence the SME survey is designed to provide reliable estimates for domains that are larger than those we target. The domains we are interested in are obtained by cross-classifying the following variables: macro-regions where firms are located (north-west, north-east, centre, south, islands), firm economic activity (NACE Rev. 2, 2 digits), firm size (four classes: fewer than 10 employees, from 10 to 19 employees, from 20 to 49 employees, from 50 to 99 employees). We obtain 426 domains, and the number of firms in

each domain ranges from a minimum of 2 to a maximum of 335. The 25th, 50th, and 75th percentiles of the domain size are respectively equal to 20, 43, and 80. Hence the number of units sampled from many of our domains is too low to obtain reliable direct estimates, so a small area estimation method is advisable. We observe that both outcomes have a distribution generally characterized by an extraordinary heterogeneity and positive skewness. A preliminary analysis reveals that they are considerably positively skewed and correlated: the Fisher skewness coefficient is approximately 2.60 for VA and 2.15 for LC, while the coefficient of correlation between VA and LC is 0.82.

3.2. Direct estimators and the estimation of their variance

Direct estimates and the estimates of their standard errors form the input information for area level models. As the domains of interest are a collection of strata, we easily obtain direct estimates by using a Horvitz-Thompson estimator. ISTAT final weights are obtained by multiplying base sampling weights (the inverse of inclusion probabilities) by two factors adjusting for i) non-response and ii) calibration with respect to known totals. As far as the estimation of standard errors of direct estimates is concerned, we are not able to replicate the ISTAT procedure due to the unavailability of some information, as the weighting cells used for calibration and the different components of final weights mentioned above, necessary to estimate the standard error of a calibration estimator according to the methodology described in the ISTAT manual (ISTAT, 2007).

In order to obtain design-based variances, we test two different approximation strategies: the linearization method and the bootstrap technique. To implement the bootstrap we use the technique for finite populations proposed by Särndal et al. (1992, page 442). We decide to adopt the bootstrap strategy which, in addition to estimating the standard errors of direct estimates, enables us to also estimate the covariance between direct estimators in a simple way, which is necessary when a multivariate sampling model is specified. The robustness of the strategy adopted is confirmed by the great coherence between the estimates obtained through the two techniques (the correlation between estimates is 0.96).

In the end, estimators of sampling variances are smoothed by using the "generalised variance functions" method. We use a log-log function to link the estimated variances to the correspondent direct estimates (Wolter, 1985, Section 5). Smoothed estimates of the sampling variances are considered as the true sampling variance in the model. To obtain smoothed estimates of covariances, necessary for bivariate models, which were coherent with correspondent variances, we multiply the square root of the smoothed estimates of the two variances and the bootstrap estimate of the correspondent correlation coefficient.

Referring to the smoothed bootstrap estimates, it appears that the first, second, and third quartiles of the coefficient of variation estimated for direct estimates obtained for the VA are 11%, 12% and 14% respectively, while its maximum value is 24%. For the LC, the first, second, third quartile, and maximum of the coefficient of variation estimated for direct estimates are respectively 9%, 10%, 12%, and 22%. These results further confirm the need to improve direct estimates by adopting a small area model approach.

3.3 Auxiliary variables and priors

As auxiliary variable in the linking model, we use for both outcomes the number of employees in small areas, data available from the ISTAT statistical archives of active enterprises (ASIA); these data are updated annually through a process of integration with various administrative archives, and provide a source of official data on the structure of firm population. The coefficient of correlations between the auxiliary variable and the direct estimates of value added and labor cost are 0.87 and 0.93 respectively.

As regards the prior specification needed to complete the Bayesian specification of the model, we assume non-informative priors. This reflects the lack of prior information on model parameters, which is the usual scenario in real Hierarchical Bayesian application on small areas and in area level models (Rao and Molina, 2015, section 10.2). We accordingly specify a bivariate normal distribution for the regression parameters of eq. (6) with dispersion matrix $\mathbf{B}^{-1}$, and we specify a

Wishart distribution (with scale matrix given by the identity matrix I_2 and 2 degrees of freedom) for both $\mathbf{B}^{-1}$ and the dispersion matrixes of the linking models $\mathbf{\Omega}_V^{-1}$.

$$\mathbf{\Omega}_V^{-1} \sim Wishart(I_2,2),\ \boldsymbol{\beta} \sim N_2(0,\mathbf{B}),\ \mathbf{B}^{-1} \sim Wishart(I_2,2), \tag{7}$$

We adopt a different approach for the shape parameter. As it has been discussed in the literature (Ferraz and Moura, 2011; Liseo and Loperfido, 2006), the estimation of the shape parameter poses some difficulties, since small differences in the shape parameters correspond to SN models not very different from each other. That problem can be tacked by using an informative prior for it wherever available. In the specific case we study, outcome variables have a positively skewed distribution, hence a positive shape parameter. We thus specify a normal distribution truncated at zero for it, with precision parameter *D* (which is the inverse of the variance parameter), both in the linking and in the sampling models:

$$\lambda_k \sim TN_{[0,\infty]}(0,D),\ (k = VA, LC) \tag{8}$$

Furthermore, we are interested in the possible effects caused by the choice of such priors on λ_k. At this aim we evaluate the sensitivity of the posterior means to the choice of the dispersion parameter *D*. We set *D* at three different values, *i.e.*, 0.01, 0.001 and 0.0001, and then compare the posterior means and the posterior standard deviations obtained using those different priors. Figure 1 compares the posterior means through three scatter plots for the VA. It is clear from Figure 1 that the small area domain estimates are very stable, for the points representing the small domains appear aligned. Similar graphics are obtained for CL. Besides, the coefficient of variations calculated as the ratio between the posterior standard deviations and the posterior means do not show significant differences between them, as they are for the three decreasing values of *D* equal to 0.0863, 0.0863, 0.0864 and 0.0762, 0.0763, 0.0762 respectively for VA and CL. Hence, we opt to use the most informative prior for λ_k, which correspond to the smallest value considered for *D* (0.01).

INSERT Figure 1 Here

Regarding the MCMC simulation, we run three parallel chains of a 250,000 length, discard the first 100,000, and thin the chain by taking every 50th sample value. The CPU time necessary to process 426 domains is about two hours.

4. Performance evaluation of the proposed small area model

We compare the proposed bivariate skew normal-skew normal model with models where: i) the correlation between sampling estimators and between random effects is assumed to be zero, ii) the shape parameter is assumed to be equal to zero at the linking level, at the sampling level and at both linking and sampling levels. The joint use of these two restrictions defines the following models, where we denote the distribution of the linking model before and after that of the sampling model: the univariate normal-normal model (univN-univN), the univariate normal-skew normal model (univN-univSN), the univariate skew normal-normal model (univSN-univN), the univariate skew normal-skew normal model (univSN-univSN), and the corresponding bivariate ones (bivN-bivN, bivN-bivSN, bivSN-bivN, bivSN-bivSN). The comparison among univariate models makes it possible for us to evaluate the improvement provided by the specification of the skew normal distribution. The comparison between univariate and bivariate models allows us to appreciate whether the "borrowing strength" from the correlation between outcomes could further improve the performance of estimators.

In the univariate models, multivariate priors are substituted with univariate ones. In particular, the Gamma prior is used for precision instead of the Wishart distribution, which is used for the inverse of covariance matrix.

We use the Deviance Information Criterion (DIC) to compare model specifications in terms of the fit of data (Table 1) and the logarithm of the pseudo-marginal likelihood (LPML, Ibrahim et al., 2001). The DIC measure is calculated with the posterior mean of deviance penalized by the effective number of parameters under the Bayesian framework (Spiegelhalter et al., 2002); it then

balances the fit of a model to the data with its complexity. The model with the smallest DIC should be the one best able to predict a replicate dataset with the same structure as the one currently observed. The LPML corresponds to a Bayesian leave-one-out cross-validation measure, and it evaluates the accuracy of prediction based on a summary statistic of the conditional predictive ordinate criterion (Gelfand et al., 1992). Models with larger LPML indicate a better fit of competing models. Table 1 reports the DIC and the LPLM results for the whole set of small area models estimated.

INSERT Table 1 Here

As expected in both the univariate and bivariate cases, the skew normal distributional assumption reduces the DIC value and increases the LPML value: the SAE models where the skew normal distribution is assumed at least at the sampling level (univSN-univSN, univN-univSN, bivSN-bivSN, bivN-bivSN) show the best fit. The skew normal distribution does not lead to a particular improvement in the fit if used at the linking level only (univSN-univN, bivSN-bivN). The normal-normal models, both in the univariate (Fay-Herriot model) and in the bivariate cases, have the worst performance. Furthermore, we observe that all bivariate models fit the data better than their correspondent univariate counterparts: consideration of the correlation, both between direct estimates and between random effects, greatly improves the fit. At the end, the most suitable models for our data set are the bivSN-bivSN and bivN-bivSN for both DIC and LPML, even though the two tools yield slightly different rankings of the models.

We furthermore evaluate the performance of model-based estimates through the percentage Coefficient of Variation Reduction, defined for each domain as:

$$CVR_i^{HB} = 100\left(1 - CV_i^{HB} / CV_i^{dir}\right), \qquad i = 1,...m \tag{9}$$

where CV_i^{HB} is the coefficient of variation referred to all the Hierarchical Bayes (*HB)* estimators considered and obtained from the posterior variance. The coefficient of variation of the direct estimator, CV_i^{dir} , is obtained from the smoothed bootstrap variance calculated as described in

Section 3.2. Hence, the *CVR* measures the gain in efficiency provided by each model-based estimator compared to the direct design-based one (Table 2). Before presenting the results obtained for the CVR, it might be useful to give some warnings on the use of this indicator. The comparison between CVs of model-based and design-based estimators might be spurious and inconclusive for selecting among different models, mainly because the model-based CV could be design-biased even where the model is correct. Nevertheless, in the literature on the Hierarchical Bayesian approach the CVs of model-based and design-based estimators have been frequently compared (You and Zhou, 2011; Molina et al., 2015; Rao and Molina, 2015; Fabrizi and Trivisano, 2016; Fabrizi et al., 2011). Therefore, we consider also of the information provided by this comparison in our application to sample data, jointly with the two already discussed model selection tools. In the next Section, we further deepen the properties of the estimators by carrying out a simulation study on a synthetic population.

Focusing on the best-performing models, the bivSN-bivSN and the bivN-bivSN models, we notice that they also ensure a relevant gain in efficiency compared to the direct estimator. The coefficients of variation reduction are, for both models, more than 30% and 25%, for the VA and the LC respectively, on median and on average compared to the direct estimator. For 10% of the domains, these reductions reach up to 43% for both variables. The gain in efficiency is a bit higher for the VA than for the CL, because the direct estimates obtained for the VA are a little more unreliable than those obtained for the CL.

INSERT Table 2 Here

To sum up, these results highlight that: i) it is important to take into account the skewness of data, mainly with reference to the sampling errors; ii) "borrowing strength" also from the correlation between outcomes further improves the model fitting; iii) the use of a non-symmetric distribution for the random term does not seem essential, when non-symmetric distribution is already specified for the sampling error. We think this result is strictly linked to the strong explanatory power of our auxiliary variable (the number of employees), and the use of a non-symmetric distribution for the

random term could ameliorate the fit of the model in those applications where the predictive power of covariates is weak.

Based on these last considerations, we analyse in greater depth the performance of both the bivSN-bivSN and the bivN-univSN model-based estimators, through a graphic comparison. Figures 2a and 2b depict the estimates obtained by the bivSN-bivSN and bivN-bivSN small area models versus direct estimates of the VA and LC averages. For both models, the points lie along the $y = x$ line and the correlations between the two sets of estimates are 0.96 and 0.99 for VA and LC respectively; this suggests that model-based estimates are approximately design-unbiased, even if there is a slight shrinkage of the model Bayesian predictor in the right-upper part of the two figures, let us say when the model-based prediction is more than 2,000. The points that show a large difference between the y and x coordinates refer mainly to those domains belonging to the highest firm size class (50-99 employees), where the sample size is particularly small and the variability is particularly high and, consequently, direct estimates tend to be particularly unreliable. However, regarding the bias, an in-depth analysis will be carried out in the simulation study (Section 5).

INSERT Figures 2a and 2b Here

Improvement in the reliability of estimates obtained by the bivSN-bivSN and bivN-bivSN small area models in each domain can be visualized in Figures 3a and 3b, which shows the values of the coefficient of variation of model-based estimates versus the coefficient of variation of direct estimates. Again, for both models (and for both VA and LC) the coefficients of variation of model-based estimates are smaller than those for direct estimates, and are markedly smaller for most of the domains. The only domain where the direct estimate is more reliable than the model-based estimate has a high sample rate (50%).

INSERT Figures 3a and 3b Here

To evaluate the precision gain of estimates obtained by adopting bivSN-bivSN or bivN-bivSN small area models instead of the most popular normal-normal model (univN-univN), we consider the

coefficient of variation reduction of the estimates. The two sets of estimates (bivSN-bivSN or bivN-bivSN compared to univN-univN), corresponding to black and grey points respectively, are plotted against domain sample size (Figure 4a and 4b). The plots show, for both VA and LC, that both the bivSN-bivSN and the bivN-bivSN models generally lead to a greater gain in efficiency than the univN-univN model. For this last model, the coefficient of variation reduction is smaller than zero in some domains, thus indicating that the estimates based on the univN-univN model are less reliable than direct estimates. With regard to this result, by analyzing more thoroughly the relation between the domain sample size and the CVs of direct and univN-univN estimates, we find that the CV of the univN-univN estimates is higher than the CV of direct estimates in about 10% of the smallest domains (sample size less than 50 units) for both the outcome variables, and that the differences between those CVs are often very small (the percentage coefficient of variation reduction (eq. 9) is less than -1 in 60% and 70% of the cases for the value added and the labor cost respectively). This result is probably due to the estimation of the standard error of direct estimates, which can be very unstable when the sample size is small.

From figures 4a and 4b it also emerges that the differences between the coefficients of variation reduction in the two sets of estimates decrease when the sample size increases.

INSERT Figure 4a and 4b Here

5. Simulation Study

In our design-based simulation study, we take advantage of the availability of a fully synthetic data set, TRItalia, which was produced within the BLUE-ETS project (Kolb et al., 2013). TRItalia data were generated, starting from the already-mentioned ASIA archive, in order to reproduce, as closely as possible, the structure of the used sample regarding dependencies and similarities among variables. TRItalia data are treated as the real population in our simulation. We prefer to base our study on the TRItalia dataset, rather than use data generated under some distribution model, because that synthetic population may provide a more realistic view of small area estimation problems that

occur in real life situations. Furthermore, we prefer to use that synthetic population rather than draw repeated samples from the ISTAT sample considered in the previous sections, in order to work with a large population with a high variability of data. Lastly, the TRItalia dataset has been already used to evaluate and compare the performance of small area estimators: see, for example, Burgard et al. (2014) and Schmid et al. (2016). In the TRItalia dataset, the distributions of our target variables, although not identical to those observed in the real sample, are still asymmetric and correlated.

Only a few firm characteristics are recorded in the ASIA archive, such as, for example, the sector of activity, the municipality where the firms are located, the number of employees, and the turnover in classes. Therefore, in the TRItalia dataset other important variables, such as VA and LC, are imputed from the ISTAT small and medium enterprise sample survey according to statistical models. For further details on the construction of the TRItalia dataset, see Kolb et al. (2013).

The main purpose of our simulation experiment is to assess whether the estimator proposed offers good randomization properties in most of the domains and whether it meets essential requirements such as design consistency and asymptotic unbiasedness. Having found that the bivariate models work best in our case for the high correlation between the outcome variables, in this simulation we focus on the bivariate specifications described in Section 4.

To reduce the problem of computational time, we limit analysis to five sectors, chosen from among the most relevant ones in Italy: “manufacture of food products”, “manufacture of textiles”, “manufacture of wearing apparel”, “manufacture of fabricated metal products, except machinery & equipment” and “manufacture of machinery & equipment”. That population consists of 279,501 enterprises. Combining the selected sectors with the five macro-regions and four classes of employees introduced in Section 3.1, we obtain a total of 100 domains. We discard 5 domains because of the low number of units. We repeatedly select 1,000 stratified samples, where strata correspond to the domains. We repeat the simulation study by considering three different percentage sampling rates: 5%, 3.5%, and 2.5%. Only in the smallest domains is the sampling rate eventually increased upwards, in order to have at least two units per domain. The 3.5% sample is

selected from the 5% sample, and the 2.5% sample is selected from the 3.5% sample, for purposes of attenuating the effect of sampling variability on the results obtained for the three different sampling fractions considered. The use of different sampling rates enables us to evaluate the possible improvement in efficiency gains provided by the estimators proposed compared to the direct one, also where the number of units sampled from the domains is reduced. In this simulation settings, the direct estimates are particularly unreliable in certain domains. For the 2.5% sample, for example, the coefficient of variation of direct estimates ranges from 7.3% to 135.9% with an average value of 46.3% and from 6.5% to 92.6% with an average value of 33.1%, respectively for the VA and the CL.

Average properties over all the domains are measured by the Average Absolute Relative Bias (AARB), the Average Mean Squared Error (AMSE), and the Average Relative Efficiency (AEFF), which compare the mean squared error of the small area estimators to that of the direct one:

$$AARB = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{1000}\sum_{r=1}^{1000}\left(\hat{\theta}_{ir}^{HB}/\theta_i - 1\right)\right|$$

$$AMSE\left(est\right) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{1000}\sum_{r=1}^{1000}\left(est_{ir} - \theta_i\right)^2 \qquad (10)$$

$$AEFF = \sqrt{AMSE\left(\hat{\theta}\right)/AMSE\left(\hat{\theta}^{HB}\right)}$$

In (10) $\hat{\theta}_{ir}^{HB}$ denotes the value of the small area estimate obtained under the small area model *HB* for the *r*.th simulated sample and the *i*-th domain (*i*=1,…*m*), and est_{ir} represents the value of an estimator (alternatively $\hat{\theta}_{ir}^{HB}$ or the direct one, $\hat{\theta}_i$) for the *r*.th simulated sample.

Table 3 shows the percentage values of indicators in (10) obtained with reference to $\hat{\theta}_{ir}^{HB}$ and to $\hat{\theta}_i$ under the three sampling rates.

INSERT Table 3 Here

These summary measures show that all the small area estimators perform significantly better than the direct estimator in terms of AEFF for both outcomes. Among them, the bivSN-bivSN and the bivN-bivSN estimators perform better than the other two for both outcomes and show very similar values for the measures in (10), with values slightly better for bivN-bivSN. The bivSN-bivSN and bivN-bivSN estimators show an AMSE which is always much lower than the direct estimator, and decreases as the sample rate increases. However the AMSE of the direct estimator decreases more than that of the small area model estimators as the sample rate increases and, consequently, the gain in efficiency provided by the small area model estimators decreases as the sample rate increases. This result holds for both outcomes. The gain in efficiency is more evident for the VA. In particular, for the VA the AEFF value ranges from 253 to 325% for the bivSN-bivSN, and from 269 to 333% for the bivN-bivSN; for the CL the AEFF ranges from 200 to 237% for the bivSN-bivSN, and from 211 to 242% for the bivN-bivSN.

The bias of the bivSN-bivSN and the bivN-bivSN estimators, measured by AARB, is found to be slightly higher for the VA than for the LC, and it decreases with the increasing sampling rate for both outcomes, in particular for the CL. AARB reaches its maximum values for the 2.5% sample (10% and 9%, respectively, for the VA and the LC).

The other two models, bivSN-bivN and bivN-bivN, even though they also perform significantly better than the direct estimator in terms of AEFF, appear to be always worse than bivSN-bivSN and bivN-bivSN both in terms of overall gain in efficiency and bias. In fact, AARB is found to be a bit higher for these estimators, reaching in the 2.5% sample 14% and 12% for VA and LC respectively.

These findings confirm those obtained for the real sample data: to consider a shape parameter for area random effects is found not to improve the estimates, whereas taking into account a shape parameter for the direct estimates, according to the domain sample size (because $\lambda_{ik} = \lambda_k / \sqrt{n_i}$), enables a significant improvement in the performance of the small area estimator.

To better understand the distribution of the bias and of the accuracy of the best performing estimators, bivSN-bivSN and bivN-bivSN, in individual domains, we carry out a graphical analysis. We focus on the results obtained for the 2.5% sampling rate. We notice that in general the two estimators show very similar results. The relationship between the absolute relative bias and the domain sample size is set out in Figures 5a and 5b. The absolute relative bias is given by:

$$ARB\left(\hat{\theta}_i^{HB}\right)=\left|\frac{1}{1000}\sum_{r=1}^{1000}\left(\frac{\hat{\theta}_{ir}^{HB}}{\theta_i}-1\right)\right|$$

Figures 5a and 5b show that the absolute bias rapidly decreases as the domain-specific sample size becomes larger, for both outcomes. This confirms our claim that the suggested small domain model-based estimators are design-consistent and as a consequence asymptotically design-unbiased. For domains with a small n_i, small domain estimators are biased by construction; indeed, they aim at reducing overall MSE using the principle of "borrowing strength" from a model assumption. This will markedly reduce variance at the expense of some bias inflation. Hence, a moderate bias when n_i is small is expected.

INSERT Figures 5a and 5b Here

We further observe the MSE reduction provided by the bivSN-bivSN and bivN-bivSN estimators compared to the direct estimator in individual small domains by plotting, separately for each outcome, the $\sqrt{MSE(\hat{\theta}_i^{HB})}$ versus $\sqrt{MSE(\hat{\theta}_i)}$ (Figures 6a and 6b). The bivSN-bivSN and bivN-bivSN estimators provide more reliable estimates than the direct estimator in almost all domains.

INSERT Figures 6a and 6b Here

To analyze the relation between the gain in efficiency provided by the suggested estimators with the domain sample size, the square root of the ratio between $MSE(\hat{\theta}_i)$ and $MSE(\hat{\theta}_i^{HB})$ is plotted versus the domain sample size (Figures 7a and 7b). The gain in efficiency appears strictly linked to the domain sample size, reaching high values when the sample size is particularly small (less than 30). We notice also that for some domains the ratio is a little less than 1. For example for the VA and

bivSN-bivSN model it happens in 7 domains where the CV of direct estimates is particularly low (from a minimum value of 7.3% to a maximum value of 19.0%, and a mean value of 13.9%).

Finally, we try to verify whether the use of our model based estimators could introduce some bias on those domain estimates that would be reliable when obtained from the direct estimator. To this purpose we focus on those domains for which we obtain the smallest CVs for direct estimates, less than 15% (12 domains), and we check the magnitude of the bias introduced by the bivSN-bivSN model in those domains. We find that, while the average Absolute Relative Bias (ARB) calculated for the whole set of domains is equal to 10%, it is only 4% when calculated for the 12 domains with smallest CVs of direct estimates. Similar results are obtained for the labour cost. Hence, we may expect that even when starting from direct estimates more reliable than those considered in our simulation, the strategy proposed would not introduce bias of worrying magnitude.

6. The small area estimates of labour productivity

This section provides a quick overview of some results obtained from the ISTAT SME sample survey, in order to show the potential of the analysis offered by the small area estimates obtained. As illustration, we consider the labour productivity (LP) of the food industry (Ateco 2002, code 10), a sector representing one of the most dynamic specialization models for the Italian manufacturing system, and a parachute for Italian manufacturing in terms of exports. The model-based estimates obtained from the bivSN-bivSN model applied to ISTAT sample data are highly reliable, as their coefficient of variation ranges from a minimum of 6% to a maximum of 8% (very similar results are obtained for bivN-bivSN model).

In Figure 8, we report estimated LP for i) manufacturing industry (deep grey bar) and food industry (light grey bar), by macro-regions, ii) food industry by regions and firm size classes (green bars). As expected, the results shown in Figure 8 clearly highlight the well-known north-south productivity divide in Italy, since the LP for the whole manufacturing industry decreases smoothly from north to south. However, when estimates are obtained for a higher level of detail, the picture is

not so clear or interpretable through the categories of the economic analysis traditionally used to evaluate territorial disparities in Italy. In fact, if we consider the food industry in macro-regions, distinguishing also among firm size classes, heterogeneity increases.

INSERT Figure 8 Here

Firstly, the north-south gap in LP is less evident in food industry with respect to the whole manufacturing sector. In fact the difference in LP between the North-West region (where the LP is at its maximum) and the Islands region (minimum LP) is larger for the manufacturing industry (about 17%) than for the food industry (about 10%). Secondly, if we take a look at the differences across LP in the food industry by regions but also by size classes, we notice that, in general, larger firms tend to be more productive, as expected, than smaller ones. Still, some interesting heterogeneity aspects arise among size classes if we focus on the north-south divide. The LP in the South for the classes from 10 to 99 employees is close to that observed in the productive North-East region, and the LP in the Center for the 50-99 employees class is not far from that of the North-West region. Besides, in general, LP for larger firms that operate in the food industry is larger than that observed for the whole food industry, in all the regions. Great differences arise among size classes within the group of firms located in the north: the LP ranges from a value of about 20 for the micro firms located in the North-East, to about 55 for the firms located in the North-West and in the larger size class considered. To sum up, the usual reading key based on the category of the north-south divide is not so evident for the food industry. Therefore, the availability of this type of results poses a new challenge to economists in search of ways to explain the heterogeneity in this sector and to formulate adequate policies to foster firms' productivity.

7. Conclusions

We propose a small domain strategy based on the Skew-Normal distribution for the simultaneous estimation of business parameters that takes into account both the asymmetry and the correlation that typically characterize the target variable distributions. The results obtained from the application

to real data highlight how considering an adequate model in the presence of asymmetry and correlation may improve the fit of the data and the reliability of estimates. The simulation study reassures us that the estimator meets some important properties as design consistency and asymptotic unbiasedness. Furthermore, an example of the economic analysis that can be drawn up from the obtained estimates further highlights the usefulness of these results for users of business small domain estimates, and in general for regional and industrial economists interested in explaining territorial, industrial, and dimensional disparities.

Nevertheless the strategy proposed may be further improved by taking into account of other issues that may be relevant when producing small domain estimates. One of the most important issues regards the benchmarking techniques, which may allow to reduce estimates bias, making the small domain estimates comparable with those obtained for larger domains. This is done by modifying the small area estimators to satisfy constraints. This practice may however reduce the efficiency of small domain estimates (Bell et al. 2013; Pfeffermann and Tiller, 2006).

Lastly, the approach we suggest can easily be extended to the estimation of other skewed business statistics and to different domains, and may also be used with data collected for other European countries, given that the statistics on VA and on LC are provided within the framework of the EU Council Regulation on structural business statistics of industry and services (58/97), which guarantees the quality of data products and their international comparability (ISTAT, 2007).

Acknowledgments

The research is embedded within the BLUE-ETS project, which is financially supported by the European Commission within the 7th Framework Programme (cf. http://www.blue-ets.eu). The authors would like to thank the Italian National Institute of Statistics (ISTAT) for kindly providing the data sets on which this study is based. The release of data is prohibited by a data access agreement.

References

Azzalini A. (1985), A Class of Distributions Which Includes the Normal Ones, *Scandinavian Journal of Statistics*, 12: 171-178.

Azzalini A., Capitanio A. (1999), Statistical Application of the Multivariate Skew Normal Distribution, *Journal of the Royal Statistical Society. Series B*, 61: 579-602.

Azzalini A., Capitanio A. (2003), Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution, *Journal of the Royal Statistical Society, Series B,* 65: 367-389.

Azzalini A., Dalla Valle A. (1996), The Multivariate Skew Normal Distribution, *Biometrika*, 83: 715-726.

Bell W.R., Datta G.S., Ghosh M. (2013), Benchmarking Small Area Estimators, *Biometrika*, 100, 1: 189-202.

Burgard, J. P. ,Munnich, R. T. , Zimmermann, T. (2014): The impact of sampling designs on small area estimates for business data, Journal of Official Statistics, 30, 4: 749-771.

Chandra H., Chambers R. (2011), Small area Estimation under Transformation to linearity, *Survey Methodology*, 31, 1: 39-51.

Chandra H., Chambers R., Salvati N. (2012), Small area estimation of proportions in business surveys, *Journal of Statistical Computation and Simulation*, 82, 6: 783-795.

Cox B.G., Binder D.A., Chinnappa N., Christianson A., Colledge M.J., Kott P.S. (eds.) (1995), *Business Survey Methods*. New York: Wiley.

Datta G.S., Fay R.E., Ghosh, M. (1991), Hierarchical and empirical multivariate Bayes analysis in small area estimation, *Proceedings of the Seventh Annual Research Conference of the Bureau of the Census:* 63-79.

Datta G.S., Ghosh M., Nangia N., Natarajan K. (1996), Estimation of median income of four-person families: A Bayesian approach, *Bayesian Analysis in Statistics and Econometrics*, (Eds. D.A. Berry, K.M. Chaloner and J.K. Geweke), 129-140, Wiley.

Diallo M.S. (2014), *Small Area Estimation Under Skew-Normal Nested Errors Models*, PHD Thesis, Ottawa-Carleton Institute for Mathematics and Statistics.

Diallo M.S., Rao J.N.K. (2014), Small Area Estimation of Complex Parameters Under Unit-level Models with Skew-Normal Errors, *JSM 2014 – Survey Research Methods Section*.

European Commission (2010), Fifth Report on Economic and Social Cohesion, http://ec.europa.eu/regional_policy/sources/docoffic/official/reports/cohesion5/index_en.cfm

Eurostat (2015b), Regions in the European Union - Nomenclature of territorial units for statistics - NUTS 2013/EU-28.

Fabrizi E., Ferrante M.R., Pacei S. (2008), Measuring Sub-National Income Poverty by Using a Small Area Multivariate Approach, *The Review Of Income And Wealth*, 4: 597-615.

Fabrizi E., Ferrante M.R., Pacei S., Trivisano C. (2011), Hierarchical Bayes multivariate estimation of poverty rates based on increasing thresholds for small domains, *Computational Statistics & Data Analysis*, 55: 1736-1747.

Fabrizi E., Ferrante M. R., Trivisano C. (2016), Hierarchical Beta regression models for the estimation of poverty and inequality parameters in small areas. In: *Analysis of poverty data by small area methods*. Pratesi, M. (Ed.), John Wiley & Sons, pp. 299-314.

Fabrizi E., Trivisano, C. (2010), Robust Linear Mixed Models for Small Area Estimation, *Journal of Statistical Planning and Inference*, 140: 433-443.

Fabrizi E., C. Trivisano (2016) Small area estimation of the Gini concentration coefficient, *Computational Statistics & Data Analysis*, 99, pp. 223 - 234

Faramondi A., Baldassarini A., Battellini F., Ciaccia D., Veroli N. D., Dol P., Donnarumma I., Forte A., Greca G., Lancioni G., Maresca S., Marotta M., Milani A., Nardone T., Pascarella C., Puggioni A., Riccioni S., Sacco G., Tartamella F. (2010), Regional Gva Inventory ITALY. Research Project Report, *Metodi e Norme*, 44, Inventory on the implementation of regional gross value added in Italy.

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

Ferrante M.R., Trivisano C. (2010), Small area estimation of the number of firms' recruits by using multivariate models for count data, *Survey Methodology*, 36, 2: 171–180.

Ferraz V.R.S., Moura F.A.S. (2011), Small area estimation using skew normal models, *Computational Statistics and Data Analysis,* 56, 10: 2864-2874.

Frühwirth-Schnatter S., Pyne S. (2009), Bayesian inference for finite Mixture of univariate and multivariate skew-normal and skew-t distributions*, Biostatistics*, 11, 2: 317-336.

Gelfand, A. E., Dey, D. K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling based methods (with discussion). in: Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (Eds.), Bayesian Statistics 4, Oxford University Press, Oxford, 147–167.

Gupta A.K., Kollo T. (2003), Density Expansion Based on the Multivariate Skew Normal Distribution, *Sankhya*, 66:821-835.

Henze N. (1986), A Probabilistic Representation of the 'Skew-Normal' Distribution, *Scandinavian Journal of Statistics*, 13, 4: 271-275.

Ibrahim J., Chen M., Sinha D. (2001). *Bayesian Survival Analysis*. Springer-Verlag, New York.

ISTAT (2007), Conti economici delle imprese - Anno 2003, *Informazioni*, 8 (in Italian).

Kolb, J.-P., R. Münnich, F. Volk, and T. Zimmermann. 2013. "TRItalia dataset." In BLUE-ETS Deliverable D6.2: Best practice recommendations on variance estimation and small area estimation in business surveys, edited by R. Bernardini Papalia, C. Bruch, T. Enderle, S. Falorsi, A. Fasulo, E. Fernandez-Vazquez, M. Ferrante, J.P. Kolb, R. Münnich, S. Pacei, R. Priam, P. Righi, T. Schmid, N. Shlomo, F. Volk, and T. Zimmermann, 168–188. Available at: http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.2.pdf (accessed October 16, 2014).

Liseo B., Loperfido N. (2006), A Note On Reference Priors For a Scalar Skew-Normal Distribution. *Journal of Statistical Planning and Inference*, 136, 373-389.

Molina I., B. Nandram, J.N.K. Rao (2015), Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach, *Annals of Applied Statistics*, 8, 2, pp.852-885.

Pfeffermann D., Tiller R. (2006), Small-Area Estimation With State-Space Models Subject to Benchmark Constraints, *Journal of the American Statistical Association*, 101, 476: 1387-1397.

Pfeffermann D. (2013), New Important Developments in Small Area Estimation, *Statistical Science*, 28, 1: 40-68.

Rao J.N.K., Molina I. (2015), Small Area Estimation – Second Edition. New Jersey: John Wiley & Sons.

Rao, J. N. K. and Molina, I. (2016) Empirical Bayes and Hierarchical Bayes Estimation of Poverty Measures for Small Areas, in *Analysis of Poverty Data by Small Area Estimation* (ed M. Pratesi), John Wiley & Sons.

Rivière P. (2002), What Makes Business Statistics Special?, *International Statistical Review*, 70, 1: 145-159.

Särndal C.E., Swensson B., Wretman J. (1992), *Model assisted survey sampling*. Springer series in Statistics, Berlin: Springer-Verlag, New York: Heidelberg.

Schmid T., Tzavidis N., Munnich R., Chambers R. (2016), Outlier robust small area estimation under spatial correlation, Scandinavian Journal of Statistics, DOI: 10.1111/sjos.12205, forthcoming.

Slud, E. V. & Maiti, T. (2006). Mean-squared error estimation in transformed Fay–Herriot models. *Journal of the Royal Statistical Society, Series B,* 68, 2: 239–257.

Souza D.F., Moura F.A.S. (2012), Multivariate beta regression with application to small area estimation. Technical Report. http://www.dme.im.ufrj.br/arquivos/publicacoes/arquivo246.pdf.

Spiegelhalter D.J., Best N., Carlin B.P., Van der Linde, A. (2002), Bayesian measures of model complexity and fit (with discussion)*, Journal of the Royal Statistical Society, Series B*, 64: 583-639.

Thomas A., O'Hara B., Ligges U., Sturz, S. (2006), Making BUGS open, *R News*, 6: 12–17.

You Y, Zhou Q.M. (2011), Hierarchical Bayes Small Area Estimation under a Spatial Model with Application to Health Survey Data, *Survey Methodology*, 37, pp. 25-37.

Wolter K.M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.

Table 1. Models comparison: DIC and LPML values (ISTAT SME sample survey)

DIC					
	Univariate models			Bivariate models	
	VA	*LC*	*tot*		*VA&LC*
univN-univN	1541	930	2471	bivN-bivN	2152
univSN-univN	1537	928	2465	bivSN-bivN	2121
univN-univSN	1502	923	2425	bivN-bivSN	1725
univSN-univSN	1495	917	2412	bivSN-bivSN	1706
LPML					
	Univariate models			Bivariate models	
	VA	*LC*	*tot*		*VA&LC*
univN-univN	-799.5	-503.2	-1302.7	bivN-bivN	-1232
univSN-univN	-795.3	-497.7	-1293.0	bivSN-bivN	-1227
univN-univSN	-780.3	-466.7	-1247.0	bivN-bivSN	-976
univSN-univSN	-780.9	-470.6	-1251.5	bivSN-bivSN	-1003

Table 2. Summaries for the Coefficient of Variation Reduction (CVR) of the HB estimators versus the direct one (ISTAT SME sample survey).

CVR%								
univariate models								
	univN-univN		univSN-univN		univN-univSN		univSN-univSN	
Summaries	VA	LC	VA	LC	VA	LC	VA	LC
perc. 0.10	-1.0	-1.3	-1.2	-1.0	1.1	0.6	0.8	0.3
perc. 0.25	0.7	0.7	0.7	0.6	3.5	2.4	3.0	2.6
Median	4.4	4.8	5.0	5.3	8.1	7.7	8.8	8.2
Average	9.9	10.8	10.4	11.4	13.8	13.8	14.1	13.9
perc. 0.75	13.5	17.2	16.2	16.5	19.5	18.8	19.8	19.0
perc. 0.90	33.1	36.8	35.1	38.4	35.2	37.8	36.6	39.2
bivariate models								
	bivN-bivN		bivSN-bivN		bivN-bivSN		bivSN-bivSN	
Summaries	VA	LC	VA	LC	VA	LC	VA	LC
perc. 0.10	-0.2	-0.7	-0.4	-1.0	20.3	15.1	19.3	13.2
perc. 0.25	1.4	0.8	1.6	0.9	26.6	21.3	25.0	19.1
Median	6.8	6.4	5.9	5.2	32.3	27.4	30.7	25.6
Average	11.6	11.3	10.0	9.8	33.6	29.1	32.2	27.3
perc. 0.75	17.4	18.3	14.4	15.2	38.8	34.9	37.0	32.8
perc. 0.90	33.4	33.4	28.2	28.5	44.9	44.1	43.4	43.5

Table 3. Summary of performance measurements based on the simulation study carried out on the synthetic population.

Value Added					
2.5% sample					
	BivSN-BivSN	BivN-BivSN	BivSN-BivN	BivN-BivN	Dir
AARB%	10.54	9.86	14.26	13.73	0.01
AMSE	9.93	9.32	16.55	15.11	103.75
AEFF%	324.55	333.65	250.38	262.04	
3.5% sample					
	BivSN-BivSN	BivN-BivSN	BivSN-BivN	BivN-BivN	Dir
AARB%	10.42	9.62	13.98	12.91	0.01
AMSE	9.86	9.11	16.00	14.15	87.81
AEFF%	298.42	310.47	234.27	249.11	
5% sample					
	BivSN-BivSN	BivN-BivSN	BivSN-BivN	BivN-BivN	Dir
AARB%	10.09	8.96	13.60	12.62	0.01
AMSE	9.85	8.75	15.46	12.47	63.49
AEFF%	252.86	269.37	202.65	225.64	
Labour Cost					
2.5% sample					
	BivSN-BivSN	BivN-BivSN	BivSN-BivN	BivN-BivN	Dir
AARB%	8.85	8.27	12.46	11.92	0.01
AMSE	5.66	5.41	9.18	8.64	31.79
AEFF%	236.99	242.41	186.09	191.82	
3.5% sample					
	BivSN-BivSN	BivN-BivSN	BivSN-BivN	BivN-BivN	Dir
AARB%	8.37	7.75	12.01	11.07	0.01
AMSE	5.40	5.08	8.56	7.89	27.04
AEFF%	223.77	230.71	177.73	185.12	
5% sample					
	BivSN-BivSN	BivN-BivSN	BivSN-BivN	BivN-BivN	Dir
AARB%	7.26	6.98	11.31	10.91	0.01
AMSE	5.17	4.68	7.87	6.79	20.85
AEFF%	200.82	211.07	162.77	175.23	

Figure 1: Comparison of estimates obtained from bivSN-bivSN model using different values for *D* in the prior for the shape parameters (ISTAT SME sample survey).

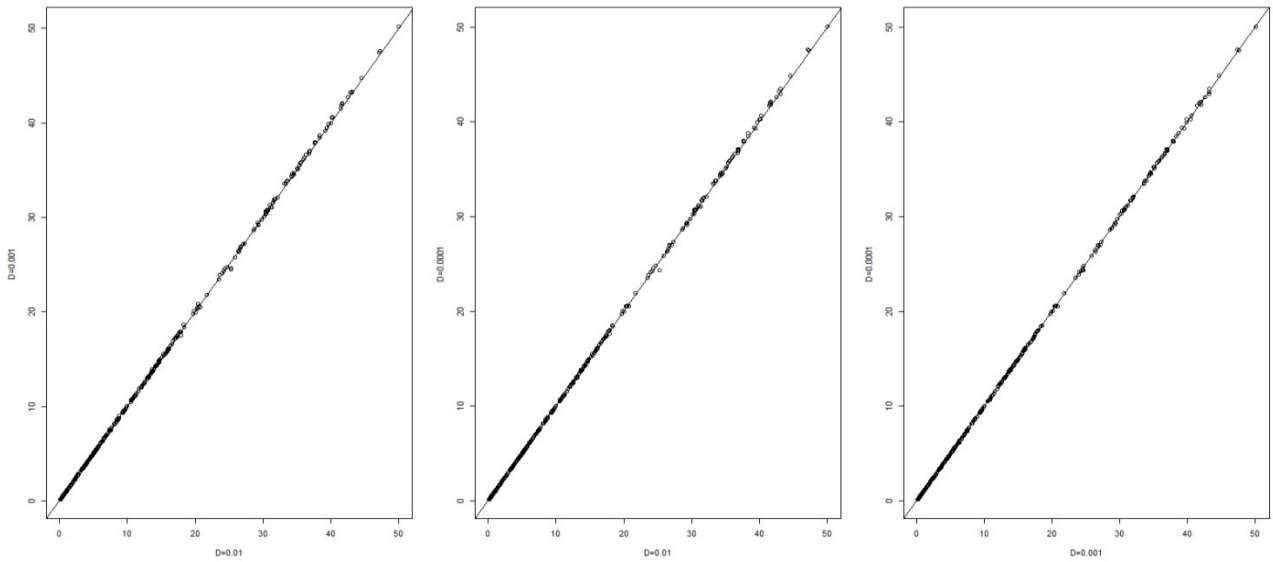

Figure 2a. Direct estimates versus Model based bivSN-bivSN estimates (ISTAT SME sample survey). Value Added (left) and Labour Cost (right).

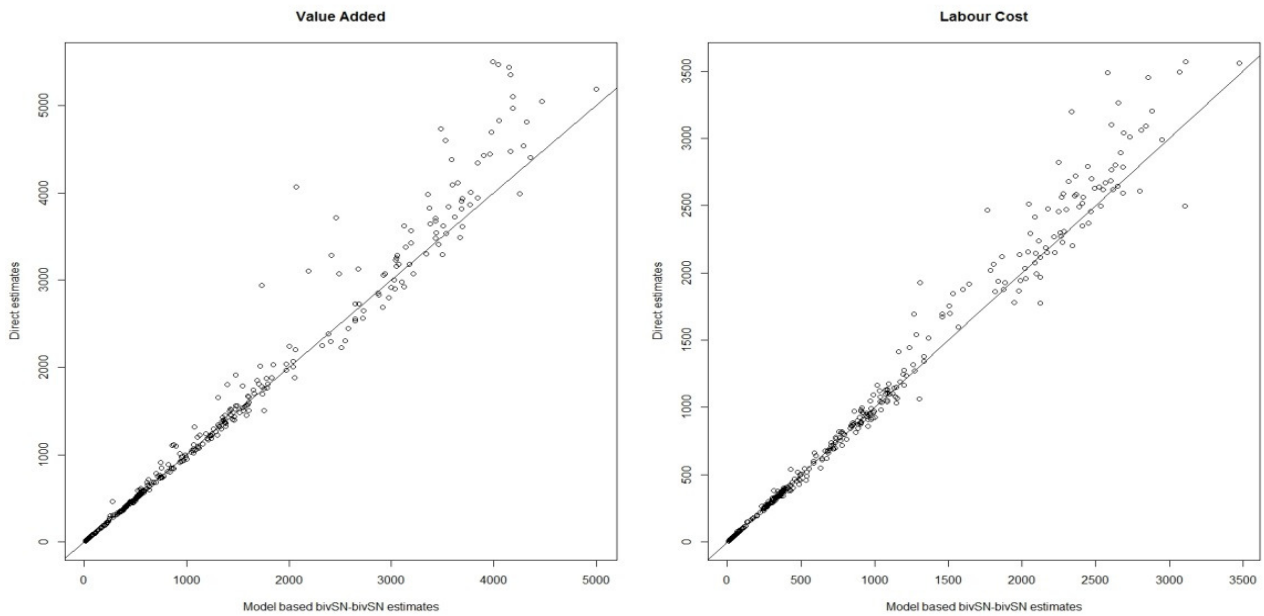

Figure 2b. Direct estimates versus Model based bivN-bivSN estimates (ISTAT SME sample survey). Value Added (left) and Labour Cost (right).

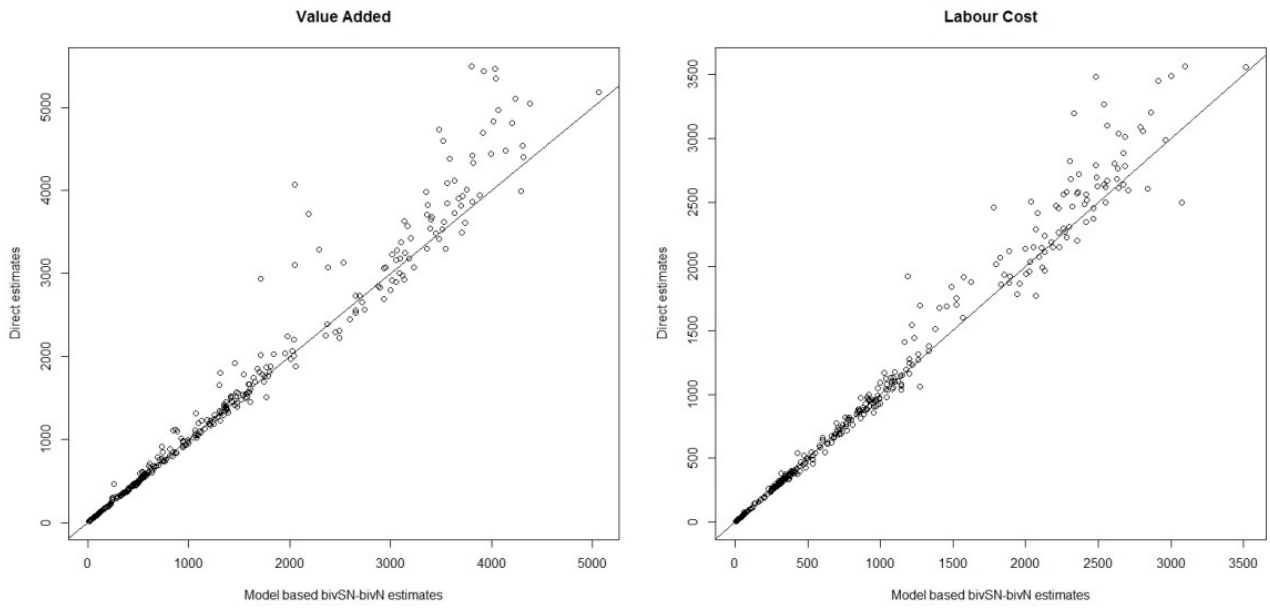

Figure 3a. Coefficient of variation (%) of model based bivSN-bivSN estimates versus coefficient of variation of direct estimates (ISTAT SME sample survey). Value Added (left) and Labour Cost (right).

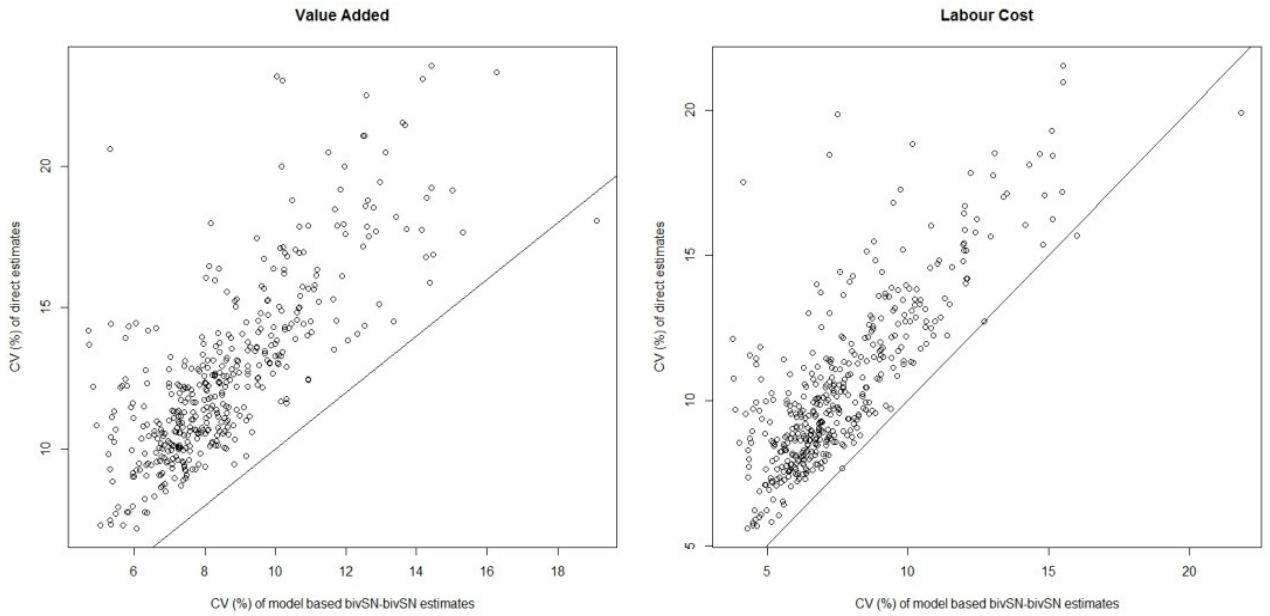

Figure 3b. Coefficient of variation (%) of model based bivN-bivSN estimates versus coefficient of variation of direct estimates (ISTAT SME sample survey). Value Added (left) and Labour Cost (right).

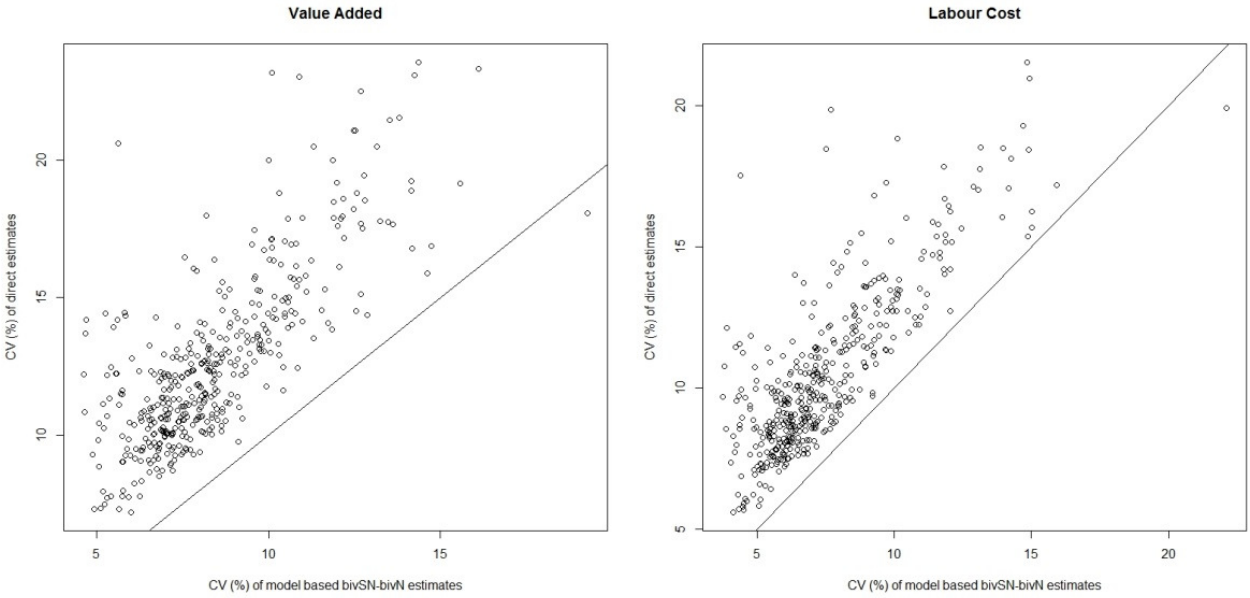

Figure 4a. Coefficient of Variation Reduction (%) of model based bivSN-bivSN estimates (black) and of univN-univN (grey) versus domain sample size (ISTAT SME sample survey). Value Added (left) and Labour Cost (right).

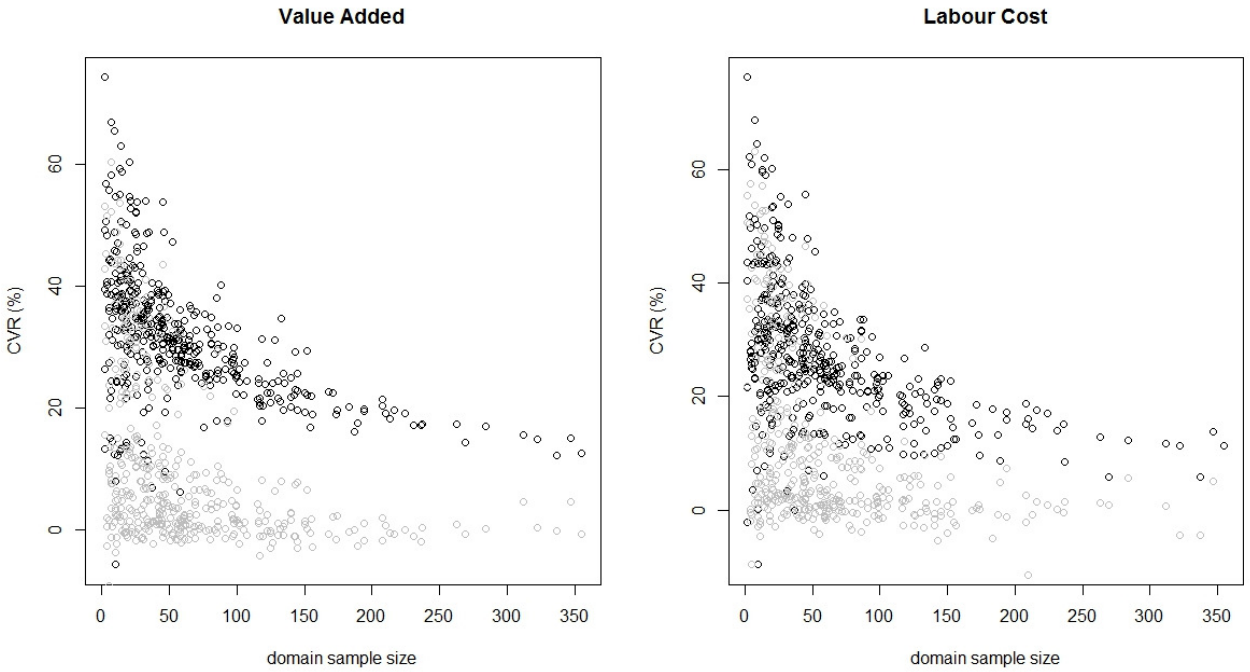

Figure 4b. Coefficient of Variation Reduction (%) of model based bivN-bivSN estimates (black) and of univN-univN (grey) versus domain sample size (ISTAT SME sample survey). Value Added (left) and Labour Cost (right).

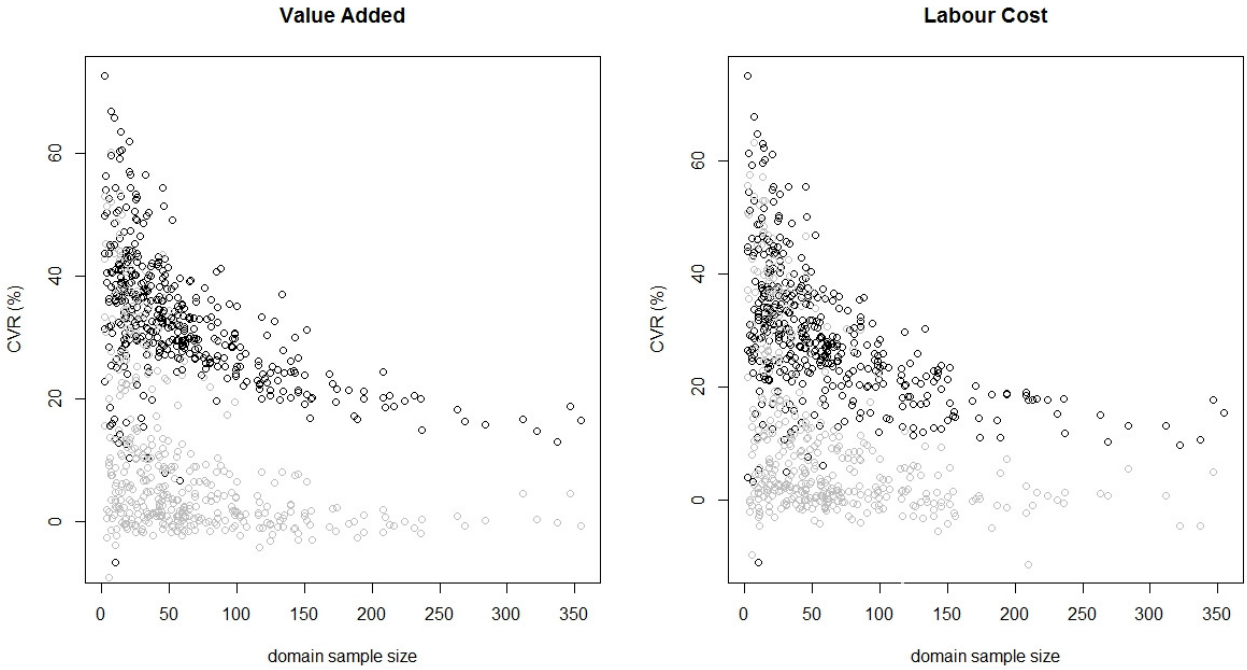

Figure 5a. Absolute relative bias of model based bivSN-bivSN estimates plotted against the domain sample size (results from the simulation study carried out on the synthetic population). Value Added (left) and Labour Cost (right).

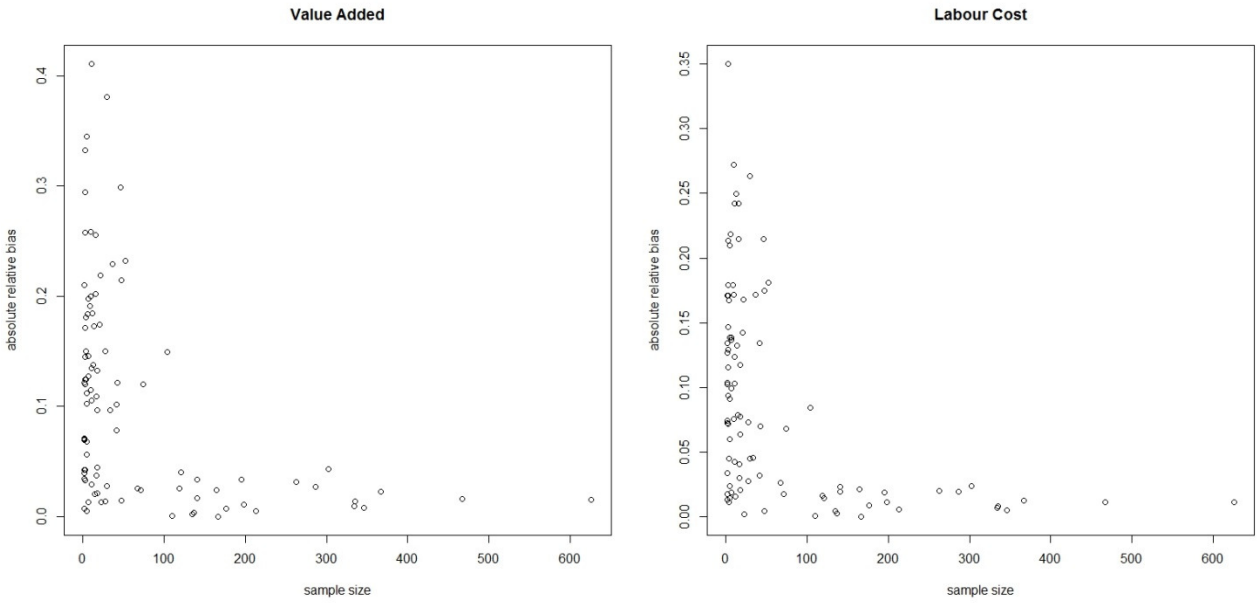

Figure 5b. Absolute relative bias of model based bivN-bivSN estimates plotted against the domain sample size (results from the simulation study carried out on the synthetic population). Value Added (left) and Labour Cost (right).

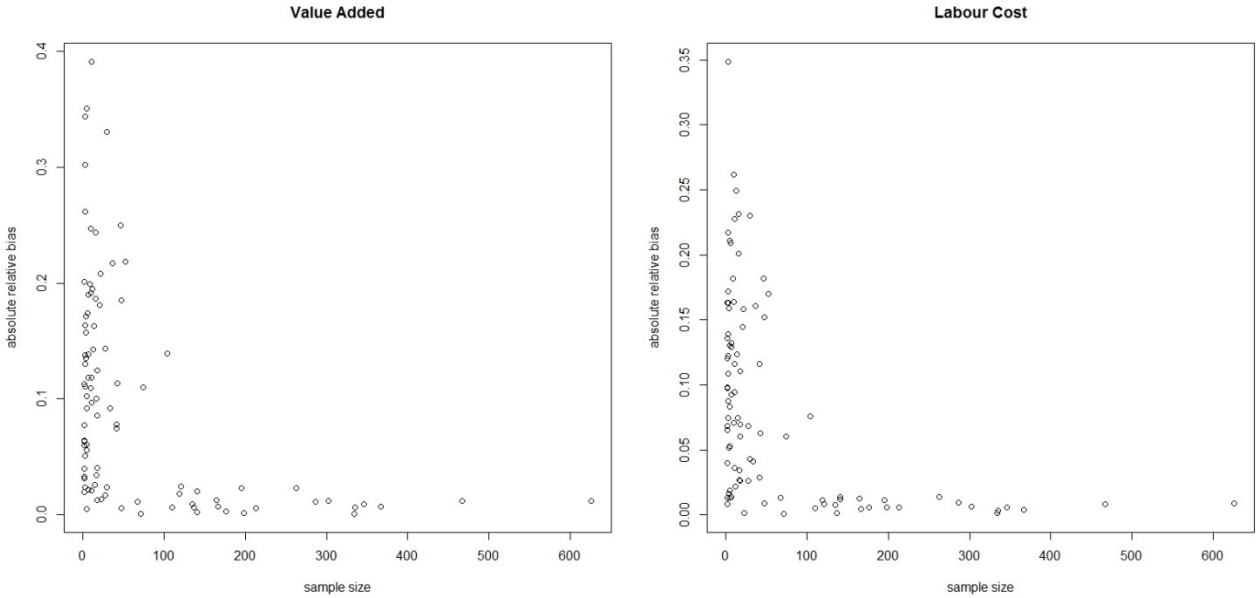

Figure 6a. Comparison between the $\sqrt{MSE\left(\hat{\theta}_i^{SN.SN}\right)}$ versus $\sqrt{MSE\left(\hat{\theta}_i\right)}$ (results from the simulation study carried out on the synthetic population). Value Added (left) and Labour Cost (right).

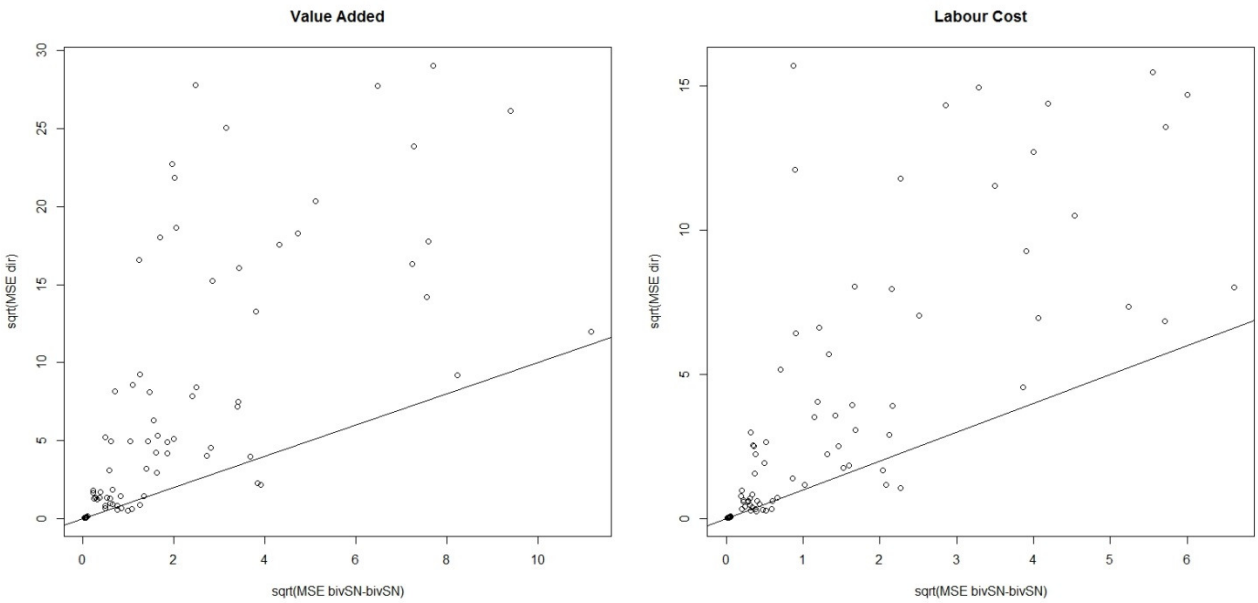

Figure 6b. Comparison between the $\sqrt{MSE\left(\hat{\theta}_i^{N.SN}\right)}$ versus $\sqrt{MSE\left(\hat{\theta}_i\right)}$ (results from the simulation study carried out on the synthetic population). Value Added (left) and Labour Cost (right).

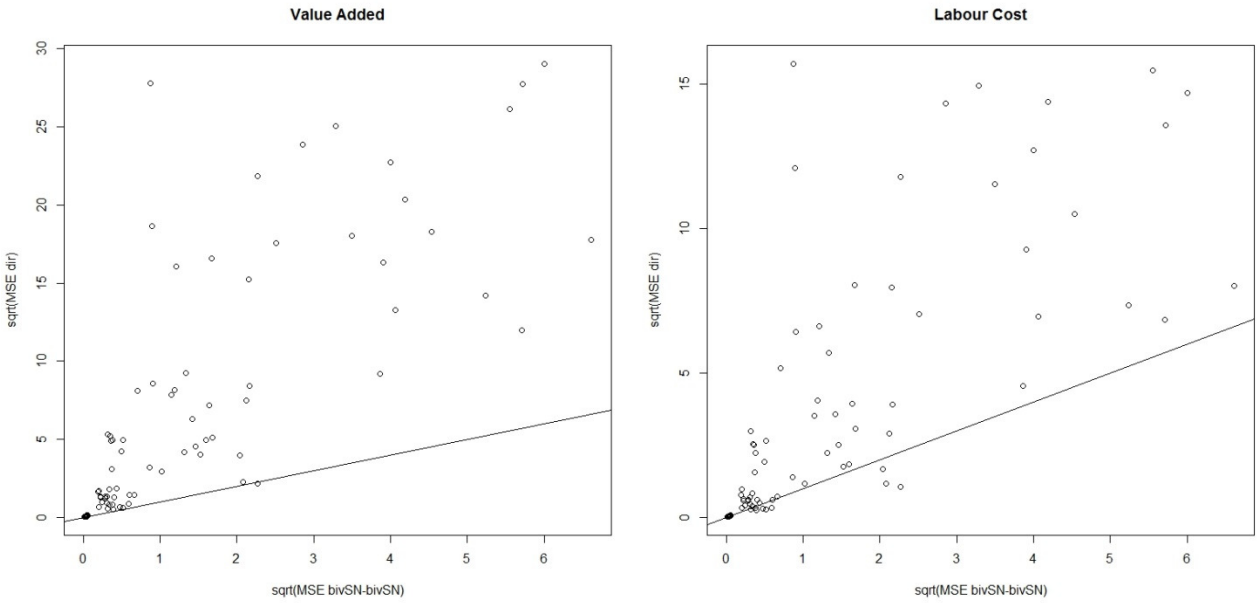

Figure 7a. Comparison between the $\sqrt{MSE\left(\hat{\theta}_i\right)/MSE\left(\hat{\theta}_i^{SN.SN}\right)}$ versus domain sample size (results from the simulation study carried out on the synthetic population). Value Added (left) and Labour Cost (right).

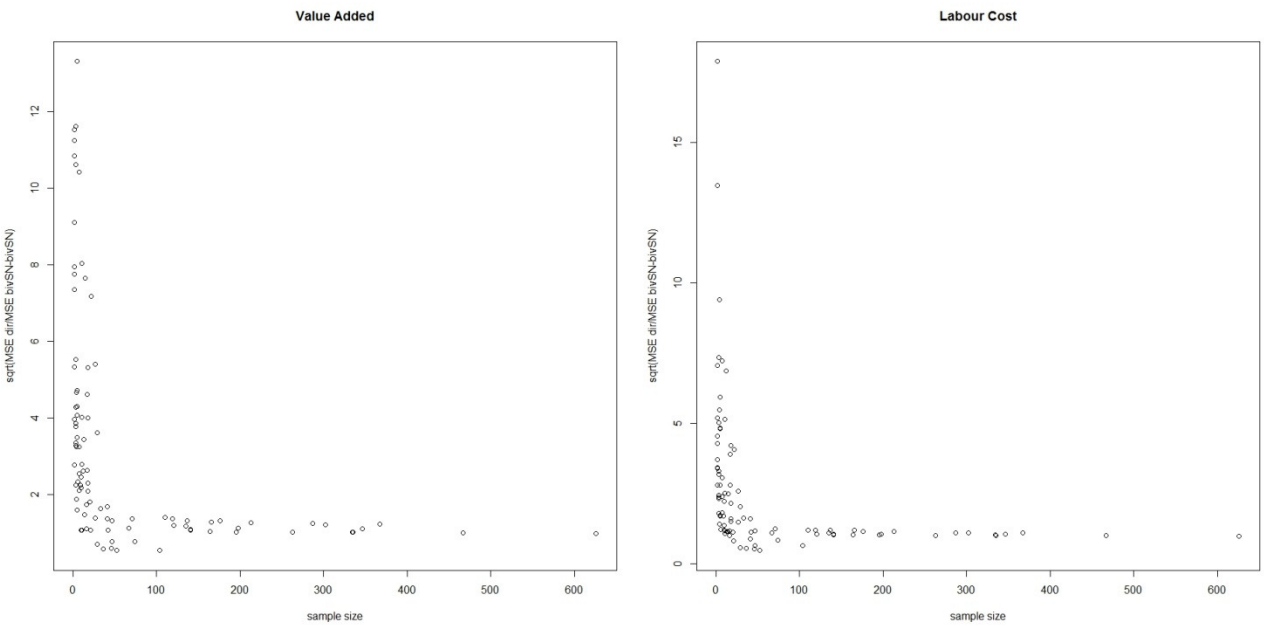

Figure 7b. Comparison between the $\sqrt{MSE\left(\hat{\theta}_i\right) / MSE\left(\hat{\theta}_i^{N.SN}\right)}$ versus domain sample size (results from the simulation study carried out on the synthetic population). Value Added (left) and Labour Cost (right).

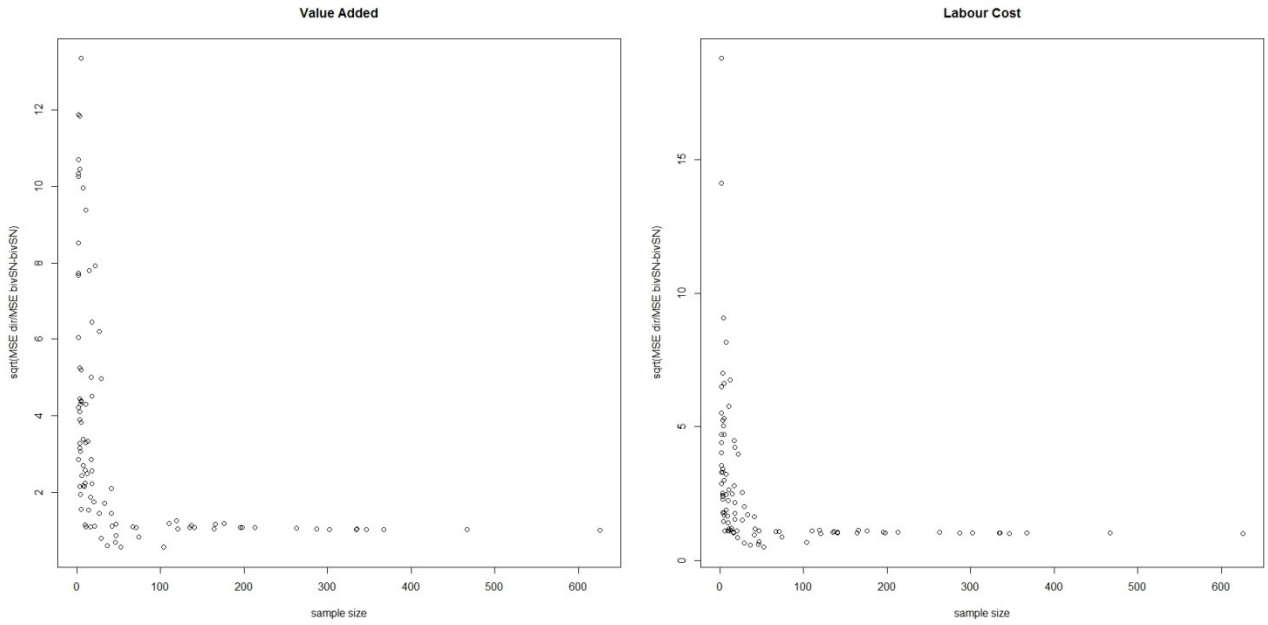

Figure 8. Labour productivity estimates for the Food industry by region and firm size class resulting from bivSN-bivSN model (ISTAT SME sample survey).