XXXV CONFERENZA ITALIANA DI SCIENZE REGIONALI[1]

ESTIMATING BUSINESS STATISTICS FOR SMALL DOMAINS

Maria Rosaria FERRANTE[2], Silvia PACEI[3]

ABSTRACT

The improvement of small domains business statistics is becoming increasingly important in order to improve official statistics, to better monitor enterprise performance and to better plan policies promoting entrepreneurship at the local level. Using data on the Small and Medium Enterprises sample survey conducted by the Italian National Statistical Institute (ISTAT), we estimate two parameters, value added and labour cost, with reference to small domains defined by cross-classifying regions, industries and firm size. The survey considered is designed to provide reliable estimates for domains larger than those we are interested in, hence the number of firms sampled from many of our domains is too low to obtain reliable estimates using the "direct" estimation strategy currently employed by Istat. To improve estimates reliability we propose to use model-based small area estimators. One of the most relevant statistical issues which arise in business surveys is the skewness in the distribution of outcomes, due to the majority of small firms. To take this issue into account, we relax the normality assumption of the classic normal-normal small area model by using a model based on the skew-normal distribution. In addition we propose a bivariate extension of such skew-normal models, which enables us to take into account the high correlation between the target variables. Results highlight i) the importance of considering the asymmetry of data and the correlation between outcomes, ii) the relevance of obtained estimates for the regional economic studies.

---

[1] This paper has been completed within the BLUE-ETS project "Blue-Enterprises and Trade Statistics", a research project funded by the Seventh Framework Programme of the European Commission, FP7-COOPERATION-SSH.

[2] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, via Belle Arti 41, 40126 Bologna (Italy), maria.ferrante@unibo.it

[3] Department of Statistical Sciences "Paolo Fortunati", University of Bologna, via Belle Arti 41, 40126 Bologna (Italy), silvia.pacei@unibo.it (corresponding author)

1. Introduction

Small area estimation of social and economic parameters have so far concerned especially the small area estimation of poverty or employment indicators, while these models have seldom been used to estimate parameters related to firm activity and performance. Only recently the literature on small area estimation methods has focused on business survey data (Chandra and Chambers, 2011; Chandra et al., 2012; Zimmermann and Munnich R., 2013; Munnich et al., 2013). The reasons for this increased interest are not only to improve official business statistics, but also to better monitor enterprise performance and promote entrepreneurship at the local level.

Some peculiar issues arise in business surveys (Cox et al., 1995; Rivière, 2002) that have to be considered in the small area estimation framework. One of the most relevant is that business data, due to the presence of a majority of small firms, are characterized by a positively skewed distribution. Therefore model specifications are of major importance. Besides, firms' parameters representing totals are generally highly related to an underlying factor which is firm size.

We propose a small area model that deals with both these issues, by operating in the "area level" model framework (Rao, 2003). Area level models consist of i) a "sampling" model, specifying direct estimates as measurements of an underlying area descriptive parameter whose variance is known (the input for the model are direct estimates and their associated variances), ii) a "linking" model, modeling the domain parameters as a function of auxiliary information accurately known at the domain level. Small area models often rely on the assumption of normality for direct estimates and area parameters, inadequate for asymmetric outcomes.

To take into account the asymmetry of data we relax the normality assumption of the classic so called normal-normal model (Fay and Herriot, 1979) by assuming that both survey estimators and random effects are skew-normally distributed. Secondly, in order to take advantage of the relationship usually observed within business data, we propose a multivariate extension of such skew-normal small area model, that considers the high correlation among direct estimators and/or the correlation among area-specific random effects.

The skew-normal specification offers some advantages with respect to other non-symmetric distributions, because it includes the normal distribution as special case and allows for modelling zero and negative values. In the normal-normal model, Gaussianity is assumed for both sampling direct estimators in the sampling model and unobservable parameters in the linking model. However, Normality can be a strong assumption for data sets arising from different areas of application, included economics. Genton (2004) shows, for instance, that symmetric distributions, like normal or t-distributions, do not model many real data-sets

sufficiently well. Data are often skewed in business surveys and economic theory suggests that regression relationships are typically multiplicative, i.e. linear in the log scale. In these cases the Normal distribution, which is particularly easy to handle, could not be adequate. Even in presence of skewed data, in the small area literature the assumption of Gaussianity at the sampling model level is often justified invoking the Central Limit Theorem. However, as we deal with small sample sizes, this assumption could be hardly sustainable. The approach is motivated by the recent increasing interest in the literature in classes of parametric distributions that are built by multiplying a symmetric density function by a function that introduces skewness in the resulting probability distribution function (Castro et al., 2008). Besides, the skew-normal class of distributions enjoys remarkable properties in terms of mathematical tractability and it proves to be quite useful in modelling real data-sets (Azzalini, 1985; Azzalini and Dalla Valle, 1996; Azzalini and Capitanio, 1999, 2003). In the context of the small area estimation, only Ferraz and Moura (2011) have recently faced the problem of skewness by assuming a skew-normal distribution for direct estimates in a univariate context. Besides, Fabrizi and Trivisano (2010), in their study on the use of robust linear mixed model for small area estimation, propose the assumption of skewed Exponential Power distribution only at the linking model level.

The multivariate specification of the small area model offers some advantages with respect to the univariate one. In the framework of "area level" models, univariate small area models improve on the traditional estimates by realizing the "borrowing strength" from related small areas or relevant covariates available for population. A further improvement in estimate reliability can be obtained by 'borrowing strength' from related dependent variables in a multivariate approach. 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) and in Datta et al. (1996) who in order to estimate median income for four-person families for the 50 U.S. states and the District of Columbia, included in the model information on three- and five-person families. Fabrizi et al. (2008) proposed a multivariate small area estimation approach in order to obtain reliable estimates for some 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. Multivariate beta regression with application to small area estimation was proposed by Souza and Moura (2012). Ferrante and Trivisano (2010) proposed a multivariate small area estimation approach for count data based on the Multivariate Poisson-Log Normal distribution. A multivariate logistic-normal model has been adopted in Fabrizi et al. (2011) with the aim of estimating poverty rates based on different thresholds.

Using data on the Small and Medium Enterprises, sample survey conducted by the Italian National Statistical Institute, we focus on two basic outcomes, the averages of *Value Added*

(VA) and *Labour Cost* (LC) and we refer to the Italian manufacturing economic sectors. We aim to obtain reliable estimates of relevant economic aggregates for smaller domains, defined by cross-classifying geographical areas, economic activity and size. Notice that by using estimates of VA and LC totals, the Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) can be estimated too, since this is defined as the difference between VA and LC.

We selected the mentioned outcomes because they are the basis of some important economic competitiveness indicators: labour productivity (VA/number of employees), cost competitiveness (LC/number of employees) and also gross profitability (EBITDA/revenue). The domain estimates obtained could then be useful to obtain a firm performance mapping. The relevance of the mentioned indicators is in fact highlighted by the increasing divergences in economic competitiveness among regions within the different EU member states observed in recent years (European Commission, 2010). In a situation of large heterogeneity within countries, the evaluation of the regional competitiveness and of the economic regional disparity have become more and more important, also in order to detect the presence of competitive regions in less competitive countries. The main aim of the Europe 2020 strategy, the new plan for long-term recovery adopted by the European Union, is the reduction of disparities. The evaluation of the EU regional cohesion could be interesting in countries like Italy, where the spatial concentration of economic activities is a feature of the economic system (Becattini et al., 2009). Besides, the availability of geographically disaggregated estimates, computed by firm sector and size, could help policy makers to implement better targeted and effective policies.

At the end, we aim at suggesting an approach that could be extended to the estimation of other business statistics and that could be used with data collected by other European countries, given that the statistics on value added and on labour cost are provided by 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).

The paper is organised as follows. in Section 2 we provide a brief description of the multivariate skew normal distribution, Section 3 presents the multivariate skew normal small area model, Section 4 contains a description of the strategy of business outcome estimation based on the proposed small area model, in Section 5 we compare this model with competitors and we evaluate the performance of the estimators associated to the small area model. In Section 6 we try to present how the estimates we obtained could be relevant in interpreting the regional and the industry disparities.

2. The Multivariate Skew-Normal distribution

Here we introduce the Multivariate Skew-Normal distribution and some of its features which are useful in this context.

A multivariate version of the skew-normal distribution has been defined in Azzalini and Dalla Valle (1996). Vector $\mathbf{Y}$ 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:

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

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:

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

3. The Multivariate Skew Normal small area model

Among the many small area methods proposed in the literature we suggest using model-based small area estimators relying on area-level models (Rao, 2003), where the term “area” refers to domain. Note that the input of these models is given by direct estimates and their associated variances; that is, direct estimates are considered as measurements of an underlying area descriptive parameter, subject to a zero-mean measurement error, whose variance is known. This part of the area model is named “sampling” model. In the second part of the small area model (the “linking” model) the domain parameters are then modelled as a function of auxiliary information accurately known at the domain level.

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 sampling model the vector $\hat{\boldsymbol{\theta}}_i$ of direct estimators are supposed to follow a multivariate skew-normal distribution:

$$\hat{\boldsymbol{\theta}}_i | \boldsymbol{\theta}_i, \boldsymbol{\lambda}, n_i, \boldsymbol{\Omega}_i \sim SN_K(\boldsymbol{\theta}_i, \boldsymbol{\Omega}_i, \boldsymbol{\lambda}_i)$$

$$\lambda_{ik} = \lambda_k / \sqrt{n_i}$$

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 that assumption. As is customary, we assume that the element of matrix $\boldsymbol{\Omega}_i$ are known, and substitute them with their respective estimates.

We try to increase the flexibility of the model by proposing the specification of a multivariate skew-normal distribution also for the linking model:

$$\boldsymbol{\theta}_i | \boldsymbol{\mu}_i, \boldsymbol{\lambda}_V, \boldsymbol{\Omega}_V \sim SN_K(\boldsymbol{\mu}_i, \boldsymbol{\Omega}_V, \boldsymbol{\lambda}_V)$$

where the location parameters are linear function of some auxiliary area level variable $\mu_i = \mathbf{x}_{ik}^T \boldsymbol{\beta}_k$.

Our parameters of interest are the expectations of the marginal distributions of $\hat{\boldsymbol{\theta}}_i$ under the described model:

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

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

Note that to obtain a continuous transition from non-normality to normality, we model the location parameter rather than modelling the mean of the skew normal distribution, as done in Ferraz and Moura (2011), while considering a univariate skew-normal model.

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

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 SN 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 explore both these solutions by adopting the second one, because the first one does not work as well from an MCMC point of view, by performing with extremely slow convergence and bad mixing of chains associated with hyperparameters.

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

4.1. Data

We rely on data collected by the Small and Medium Enterprises (SME) sample survey (1-99 employees), carried out by the Italian National Statistical Institute (ISTAT) that provided us with this information in 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, Italian administrative regions (NUTS2) and company size. A detailed description of the SME survey can be found in Faramondi et al. (2010). We consider data collected in 2008 which refer to 25,925 firms in the 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. In more detail, 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: less 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, the 50th and the 75th percentiles of the domain size are respectively equal to 20, 43, 80. Hence the number of units sampled from many of our domains is too low to obtain reliable direct estimates and a small area estimation method is advisable.
We observe that both outcomes have a distribution generally characterized by an extraordinary heterogeneity and asymmetry. A preliminary analysis reveals that they are considerably positively skewed and correlated: the Fisher skewness coefficients are approximately 6.6 for VA and 2.5 for LC, while the coefficient of correlation is 0.82.

4.2. The direct estimators and the estimation of their variances

Let $\hat{\theta}_{ik}$ be the direct estimator of the outcome parameter θ_{ik} in the i-th domain (i = 1,…, m), referring to the k-th outcome ($k = VA, LC$).

Firstly we focus on direct (design-based) estimates and on the estimation of their standard errors, both input information of the area level models. As the domains of interest are a collection of strata, direct estimates can be easily obtained by using a Horvitz-Thompson estimator. As far as estimation of standard errors is concerned, ISTAT adopts calibration estimators (for more details on the sample design and estimation strategy see ISTAT, 2007). Unfortunately, due to the unavailability of some design information, we are not able to replicate the ISTAT procedure. ISTAT calibrates sampling basic weights given by the inverse of the inclusion probabilities, hence the final weights are unequal within the strata. Besides, analytic variance is based on linear approximations and is not straightforward to calculate.

In order to obtain design-based variances, we test two different approximation strategies: the linearization method and the bootstrap technique. In order to implement the bootstrap we follow the technique for finite populations proposed by Särndal et al. (1992, page 442). We decide to adopt the bootstrap strategy that, besides the estimates of standard errors of direct estimates, enables us to estimate also the covariance between direct estimators in a simple way, which is necessary when a multivariate sampling model is specified. The robustness of the strategy used is confirmed by the very large coherence between estimates obtained through the two techniques (the correlation between estimates is 0.96). By referring to the bootstrap estimates, it arises that the first, second and third quartiles of the estimated coefficient of variation of direct estimates are 8%, 12% and 18% respectively, while its maximum is 123%. These results further confirm the necessity to improve the direct estimates by adopting a small area model approach.

At the end, estimators of sampling variances are smoothed using the 'generalised variance functions' method, and the smoothed estimators are then treated as the true sampling variances; see Wolter (1985, Ch. 5) for details of the method. In particular, we use a log-log function to link the estimated variances to the correspondent direct estimates. Smoothed estimates of the sampling variances are then considered as known in the model.

4.3. Auxiliary variables and priors

As auxiliary variable in the linking model, we adopt the number of employees in the small areas for both outcomes, which are available from the ISTAT statistical archives of active enterprises (ASIA), updated annually through a process of integration of various administrative archives and representing a source of official data on the structure of firm population.

As regards the prior specification needed to complete the Bayesian specification of the model, we assume non informative priors, apart from the shape parameter. It is well known that difficulties arise in the estimation of the shape parameter that can be overcome by using an informative prior for it. In the specific case we are researching, we presume a positively skewed distribution for our outcome variables and, hence, a positive shape parameter. Therefore, we specify a normal distribution truncated at zero:

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

$$\lambda_k \sim TN_{[0,\infty]}(0,D),\ (D=0.01)\ (k=VA,LC).$$

To implement MCMC calculations we used the OpenBugs software (Thomas et al., 2006; Spiegelhalter et al., 2002) which is very widely used in applied hierarchical modelling. For the MCMC simulation we run three parallel chains of length 250,000, discard the first 100,000 and thin the chain by taking every 50th sample value.

5. Performance evaluation of the proposed small area model

Here we compare the proposed bivariate SN-SN model with models where: i) the correlation between sampling estimators and that between random effects are assumed to be zero, ii) the shape parameter has been assumed equal to zero at the linking level and/or at both sampling and linking level. The joint use of these two restrictions defines the following models both univariate and bivariate: the univariate normal-normal model (univN-univN), the univariate skew normal-normal model (univSN-univN), the univariate skew normal-skewnormal model (univSN-univSN), and the correspondent bivariate ones (bivN-bivN, bivSN-bivN, bivSN-bivSN). The comparison among univariate models allows 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 if the “borrowing strength” among the correlation between outcomes could further improve the performance of estimators. Obviously, to reach a conclusive result on the model choice, a simulation exercise has to be performed.

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

We use the Deviance Information Criterion (DIC) to compare the specifications in terms of fit of the data (Table 1), and the percentage coefficient of variation reduction (CVR) to measure the gain in efficiency provided by the small area estimators used with respect to the direct

estimator (Tables 2). The model with the smallest DIC should be the one that would best predict a replicate dataset which has the same structure as that currently observed.

Table 1 - Models comparison

	DIC				
	Univariate models				Bivariate models
	Value Added	*Labour Cost*	*tot*		*Value Added & Labour cost*
univN-univN	1541	930	2471	bivN-bivN	2152
univSN-univN	1502	923	2425	bivSN-bivN	1725
univSN-univSN	1495	917	2412	bivSN-bivSN	1706

First we may observe that the bivariate models fit the data better that their correspondent univariate ones: the consideration of correlation both between direct estimates and between random effects greatly improves the fit.

Secondly, in both the univariate and bivariate cases, the skew normal distributional assumption reduces the DIC value: the skew normal-skew normal models show the best fit, followed by the skew normal-normal models, whereas the normal-normal models, mainly in the bivariate context, perform worse.

In addition, the importance of specifying a skew-normal distribution also for the random coefficients arises both for univariate and multivariate models.

Focusing on the best performing model in terms of DIC, the bivSN-bivSN model, we notice that it also provides a relevant gain in efficiency with respect to the direct estimator. The reduction of the coefficient of variation is about 30% and 25% respectively for the value added and the labour cost on median with respect to the direct estimator, and for 10% of domains this reduction reaches 43% for both variables.

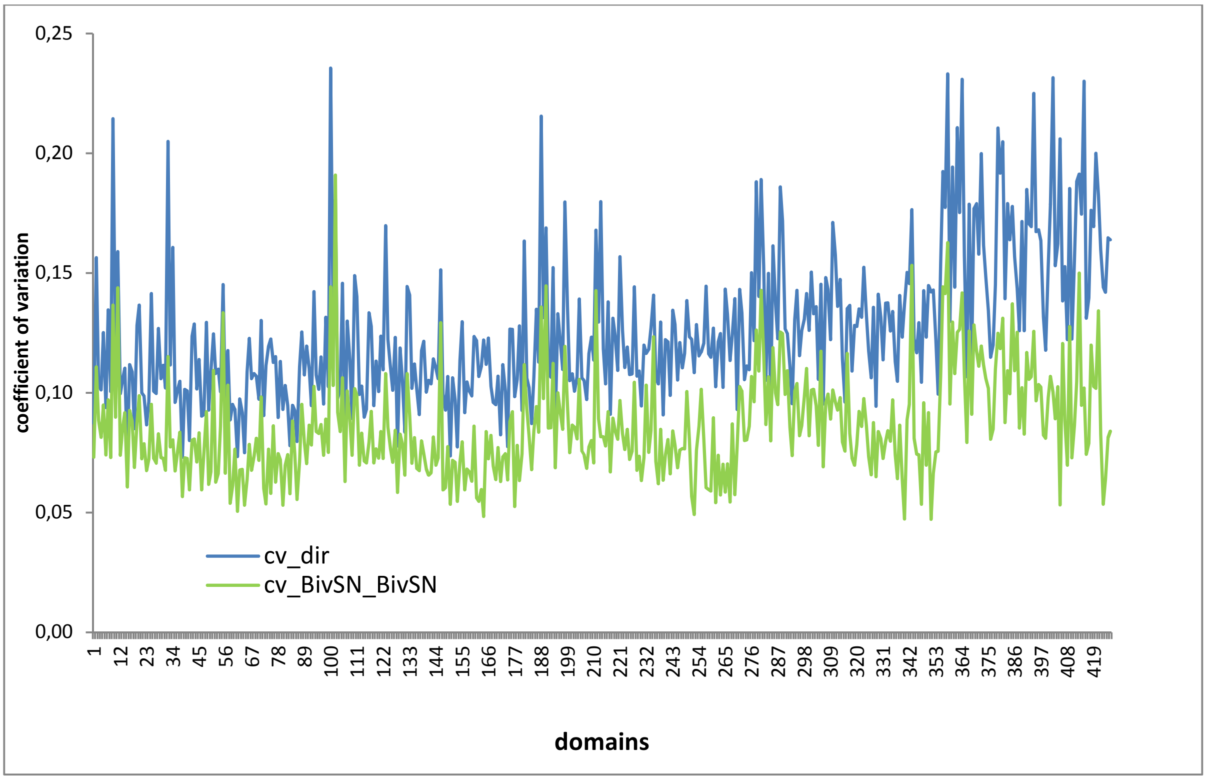
Table 2 - Summaries for the Coefficient of Variation Reduction of the HB estimators associated to the bivSN-bivSN versus the direct one.

	CVR	
Summaries	Value Added	Labour Cost
perc. 0.10	19.3	13.2
perc. 0.25	25.0	19.1
Median	30.7	25.6
Average	32.2	27.3
perc. 0.75	37.0	32.8
perc. 0.90	43.4	43.5

The results are also confirmed by the graph reported in Figure 1, where the green line reports the coefficients of variation obtained for the small area model based estimators referred to the value added in each domain. The coefficient of variation of direct estimates is shown in blue.

The graph obtained for the labour cost is very similar. We may observe that the coefficient of variation of estimators associated to the bivariate skew normal-skew normal model is significantly lower than the one concerning direct estimates.
Summarizing, these results highlight the importance: i) of taking into account the skewness of data; ii) of specifying a non symmetric distribution for the random term even if a non-symmetric distribution is already specified for the sampling error, iii) of "borrowing strength" also from the correlation between outcomes.

Figure 1 - Coefficient of variation for the whole set of domains. In green those referring to the bivariate SN-SN estimator, in blue those referring to the direct estimator. Outcome variable: Value Added



6. The small area estimates of labour productivity

A quick overview of some results is provided in this section in order to show the potential of the analysis offered by the small area estimates obtained. As an illustrative example we consider the labour productivity (LP) of the Food industry (Ateco 2002, code 10), a sector representing one of the most dynamic specialization model for the Italian manufacturing system and a parachute for Italian manufacturing in terms of export. The estimates we refer to are highly reliable as their coefficient of variation goes from a minimum of 6% to a maximum of 8%.

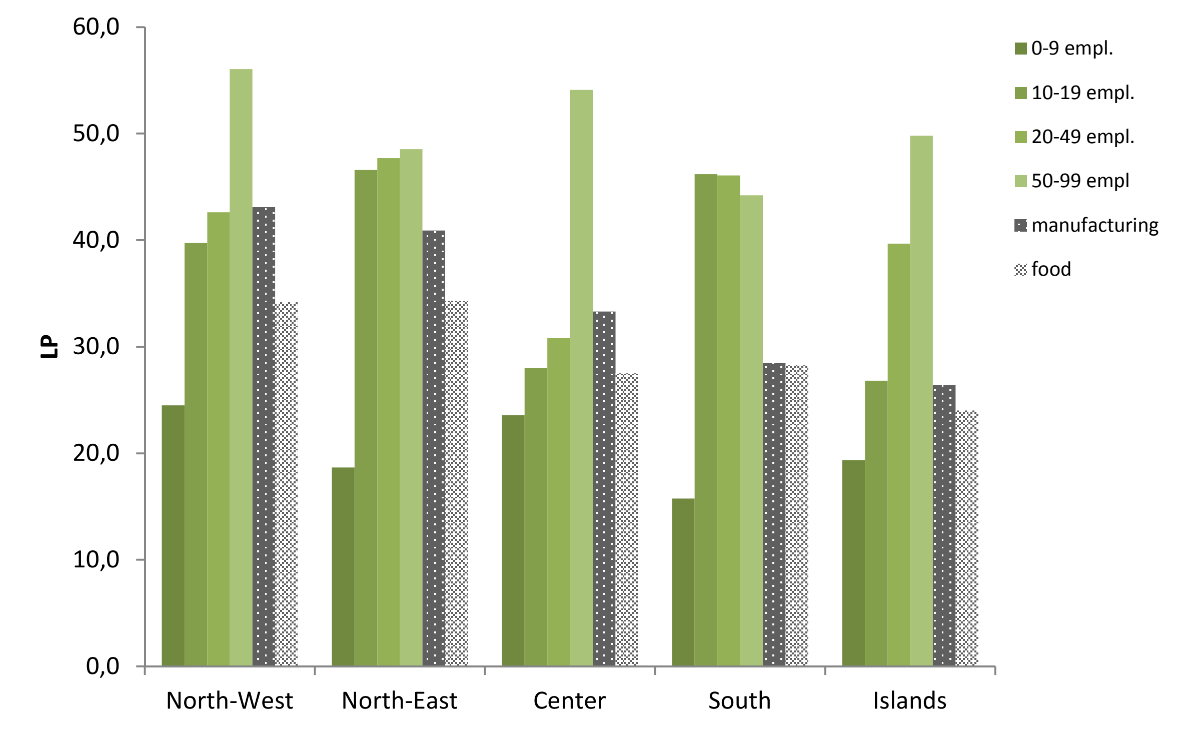
As expected, the results depicted in Figure 1 clearly show the well-known productivity North-South divide in Italy as the LP for the manufacturing industry decreases smoothly from the North to the South. However, when estimated are obtained for a higher level of detail, the picture is not so clear or interpretable with the categories of the economic analysis traditionally used to evaluate territorial disparities in Italy. If we consider the Food industry in macro-regions, by distinguishing also among firm size classes, heterogeneity increases.

Firstly, the North-South gap in LP is less evident when we focus on the Food industry: with respect to the whole manufacturing, the class of firms located in the northern regions and belonging to this sector has an LP closer to that of the southern regions.

Secondly, if we take a look at the differences among size classes, we notice that larger firms are more productive, as expected, than smaller ones. Moreover, some other interesting heterogeneity elements arise. Micro-firms have close LP levels in all regions and also the medium size firms (in the 50-99 employees class) located in the South are characterised by an LP level absolutely comparable to that of the similar firms located in the North. Besides, great differences arise among size classes inside the group of firms located in the North: the LP goes 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.

In summary, the North-South divide for the Food industry is not evident and these results could pose a new challenge to the economists in search of ways to explain heterogeneity in this sector.

Figure 2 - Labour productivity estimates for the Food industry by region and firm class size.



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.

Becattini G., Bellandi M., De Propris (2009), *A Handbook of Industrial District*. Edward Elgar Publishing Ltd.

Castro L.M., Loschi R.H., Arellano-Valle R.B. (2008), *Bayesian Inference for the Skew-Normal Shape Parameter: An Application to Change Point Problems*. Submitted.

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

Chandra H., Chambers R., N. Salvati (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.

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

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

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.

Genton M.G. (2004), *Skew-Elliptical Distributions and Their Applications: a Journey Beyond Normality*, Chapman & Hall, Boca Raton.

Gupta A.K., Kollo T. (2004), 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.

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

Munnich R., Schmid T., Zimmermann T. (2013), *Spatial robust small area estimation applied on business data*, working paper.

Rao J.N.K. (2003), *Small Area Estimation*. New Jersey: 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.

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.

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

Zimmermann T., Munnich R. (2013), *Coherent small area estimates for skewed business data*, working paper.