Fabrizi, Enrico; Ferrante, Maria; Trivisano, Carlo

Conference Paper

Small area estimation of labor productivity for the Italian manufacturing SME cross-classified by region, industry and size

53rd Congress of the European Regional Science Association: "Regional Integration: Europe, the Mediterranean and the World Economy", 27-31 August 2013, Palermo, Italy

Provided in Cooperation with:
European Regional Science Association (ERSA)

Suggested Citation: Fabrizi, Enrico; Ferrante, Maria; Trivisano, Carlo (2013) : Small area estimation of labor productivity for the Italian manufacturing SME cross-classified by region, industry and size, 53rd Congress of the European Regional Science Association: "Regional Integration: Europe, the Mediterranean and the World Economy", 27-31 August 2013, Palermo, Italy, European Regional Science Association (ERSA), Louvain-la-Neuve

This Version is available at:
https://hdl.handle.net/10419/124053

Small area estimation of labor productivity for the Italian manufacturing SME cross-classified by region, industry and size

M.R. Ferrante, C. Trivisano,
Dipartimento di Scienze Statistiche "P. Fortunati"
Università di Bologna, Italia
E. Fabrizi
DISES, Università Cattolica del S. Cuore, Italia

Abstract: In this paper we propose a new small area estimation methodology aimed at the estimation of Value Added, Labor Cost and related competitiveness indicators for subsets of the population of Italian small and medium sized manufacturing firms classified according to geographical region, industrial sector and firms size. This disaggregation is needed in regional comparisons in order to avoid the confounding effect of sectorial and firm size composition of a region's manufacturing industry. We use data on the Small and Medium Enterprises sample survey conducted by the Italian National Statistical Institute (year 2009) that provided us this information in the framework of the BLUE-ETS project. The estimates obtained with our method are more reliable than those that would have been obtained using standard survey weighted estimators, and offer therefore the basis for more sound economic analysis. The small area methods that we propose are model based and take into account the peculiarities of business such as the skewness of target variables' distributions. For this reason the model we propose is based on the log-normal distribution. We consider a multivariate model in which two different variables (Value Added and Labor Cost) and jointly modeled in order to exploit their correlation. We adopt a Bayesian approach to inference. The problem of prior specification is considered and two alternative solutions compared. Since we produce estimates for several variables and hundreds of subset of the target population results are difficult to summarize. A general conclusion may be that, for Italy, the North-South divide in productivity levels is more apparent in capital and knowledge intensive sectors, especially when industrial districts are present. The productivity gap tends to grow for larger firms, but there exists several exception to this rule. Many industries traditionally associated to the Italian productive system (furniture, clothing, textile) are characterized by low labor productivity levels: in these cases the productivity gap between Northern and Southern regions is less pronounced or absent. As the paper is mostly about the methodology needed to obtain the estimates, it is relevant not only for those interested in Italian economy. The same ideas may be applied to data from other countries. The relevance of the mentioned indicators is highlighted by the increasing divergences in economic competitiveness among regions within the different EU member states observed in these last years.

Key words: competitiveness, value added, labor cost, sample survey, Bayesian inference.

JEL: R12, C11, C13, J24.

1. Introduction

Slow productivity growth is the root of competiveness losses recorded by the Italian economy since the late 90s (IMF, 2011). The country needs to enhance its competitiveness and in a long term perspective to re-orient its productive structure towards more innovative, high productivity sectors (Montanari, 2011). Italy is a country of wide, persistent regional disparities: the per-capita GDP in the South is only 60% of that of the Center-North. Roughly one third of this gap may be accounted by disparities in labor productivity (Bardone and Reitano, 2009). Italy's manufacturing industry is also characterized by a predominance of small and medium-sized enterprises, which are unable to exploit economies of scale.

These facts about Italian economy motivates our interest in estimating labor productivity measures separately for Italian regions, firm size classes and economic sectors. We use data on the Small and Medium Enterprises (SME) sample survey (1-99 employees) conducted by the Italian National Statistical Institute

(ISTAT) that provided us this information in the framework of the BLUE-ETS project. We consider the 2009 wave of the SME survey.

We illustrate an estimation method to obtain reliable estimates of two relevant economic aggregates, value added (VA) and Labor cost (LC) for small subsets (domains) of the population of Italian small and medium manufacturing enterprises defined by cross-classifying by region, sector, and size. With small and medium enterprises we mean those with less than 100 employees. We focus on VA and LC since they are the basis for calculating important aggregates and indicators: i) Earnings Before Interest, Taxes, Depreciation and Amortization (EBITDA) that is obtained as the difference between VA and LC; ii) labor productivity (VA/number of employees); iii) cost competitiveness (LC/number of dependent employees); iv) gross profitability (EBITDA/revenue). Labor productivity will be given the most attention when discussing results. In this paper we consider only firms with 99 employees or less. The reason is twofold: first, larger firms are censused and small area estimation are therefore not needed; second, the study of productivity of small and medium enterprises is specific interest in economics.

An atlas of productivity / competitiveness measures by detailed sectors, region and size classes provides a valuable information basis for economic analysis and policy decision as it will discussed more in detail in section 2.1.

With reference to the mentioned variables, ISTAT provides reliable estimates for domains defined alternatively by: i) cross-classification of administrative region and economic activity (NACE Rev. 2, 2 digit), ii) cross-classification of size (in classes) and economic activity (NACE Rev. 2, 3 digit), iii) simple classification of economic activity (NACE Rev. 2, 4 digit). These estimates are characterized by sizeable sampling errors (ISTAT, 2010b). The domains we are interested are smaller, obtained cross classifying the population according to three criteria. Specifically, they are: regions where firms are located (North-West, North-East, Centre, South, Islands), economic activity (NACE Rev. 2, 2 digit), size (in four classes: less than 10 employees, from 10 to 19 employees, from 20 to 49 employees, from 50 to 99 employees).

For domains as small as those that we target, ordinary design weighted estimators do not lead to precise enough estimators. For this reason we use small area estimation methods (see Rao, 2003, Hidiroglou and Smith, 2005 and Chandra, 2009). The idea is that of combining survey data with auxiliary information accurately known for each unit in the population. Specifically, the source of our auxiliary information is the ASIA administrative archive. In small area estimation survey and administrative data are usually linked using a statistical model.

Technically, we use area-level small area estimation models. In specifying these model we should take into account some peculiarities of business data (Cox et al., 1995; Rivière, 2002) and namely the typical positive skeweness of data. In these cases, models relying on the normality assumption , which are typical in small area estimation, can be inadequate. In this article we explore the use of various alternative models based on the log-normality assumption. Moreover we adopt a full Bayesian approach to inference, that allows us to handle the estimation of various related target parameters (such as VA and LC totals along with related productivity indicators). Bayesian inference involve the specification of priors: we discuss a careful choice of these priors in order to obtain Bayes estimators with good frequentist properties.

From a subject matter point of view, the results we obtain are difficult to summarize since estimates are produced for several variables and hundreds of domains. A general conclusion may be that the North-South divide in productivity levels is more apparent in capital and knowledge intensive sectors, especially when industrial districts are present. The productivity gap tends to grow for larger firms, but there exists several exception to this rule. Many industries traditionally associated to the Italian productive system (furniture, clothing, textile) are characterized by low labor productivity levels: in these cases the productivity gap between Northern and Southern regions is less pronounced or absent.

As the paper is mostly about the methodology needed to obtain the estimates, it is relevant not only for those interested in Italian economy. The same ideas may be applied to data from other countries. The relevance of the mentioned indicators is highlighted by the increasing divergences in economic competitiveness among regions within the different EU member states observed in these last years (European Commission, 2010).

The paper is organized as follows. In the next section we first present some data that may be helpful to illustrate the need for small area estimation in the analysis of labor productivity and we review the small area literature that we consider relevant for this paper. Section 3 presents the survey data we analyze, section 4 discusses the adopted small area models. A statistical analysis of the results may be found in section 5, while section 6 contains some examples of the use that can be done of the results.

2. Stylized facts and review of relevant small area literature

2.1 Stylized facts motivating small area estimation

In 2008, the Italian manufacturing industry employs 4.4 million workers in 460 thousand firms (average number of employees per firm is 9.6) with an overall value of 213 million euros. It accounts for 10.4% of firms, 25.5% of employees and 29.8% of the value added created by the private sector (ISTAT 2010b). Most manufacturing firms are very small (only 2.4% of them crosses the threshold of 50 employees); a feature that distinguished Italy and that has been held responsible for the staggering growth of the country's economy and especially its low levels of labor productivity that is equal for manufacturing industry to 48.4 thousand euros in 2008.

Labor productivity in the manufacturing industry grows with firm size as it ranges from 26.1 thousand euros for micro firms (less than 10 employees) to 70.2 of firms with 250 employees or more. Labor productivity differs widely across sectors, as it may in part expected in view of different capital intensity of the various industries. The oil refinement and the pharmaceutical sectors are characterized by labor productivity more than three times larger than the textile and clothing sectors. Of course there may an interplay between size and sector, so a comparison of productivity across groups simply defined in terms of size class or sector may be misleading.

The same is true if we consider geographical comparisons of labor productivity. Italy is characterized by large regional disparities. Bardone and Reitano (2009) note that about one third of the observed disparities in the per-capita GDP between the affluent Northern regions and those of the South and Islands may be attributed to the gap in labor productivity. In this line we may note that, in 2008, labor productivity in Lombardia (North) is twice that of Calabria in the South (ISTAT, 2010b). Nonetheless composition by sector and size of the regional industries should be taken into careful consideration. For instance, we have that labor productivity is higher in Sicilia with respect to Lombardia in 2007, a very counterintuitive result. Lombardia is an industrial region while Sicilia is not and this reflects in a value added 6.4 times larger than that of Sicilia for the manufacturing industry with a population that is only twice as big. Nonetheless the capital intensive oil refinement sector, although employs only 3% of the manufacturing industry, produces 65% of the value added of rather under-developed industry in Sicilia, leading to very high productivity figures.

To avoid these type of problems when analyzing labor productivity data we consider the estimation of productivity levels for domains created cross-classifying the population according to three criteria: firm size (4 classes), geography (5 regions) and economic activity (23 sectors). As anticipated, the domains defined in this way are too small for the standard survey weighted estimators to lead to reliable enough estimators. Small area estimation may then be useful. In small area estimation survey and administrative data are usually linked using a statistical model.

2.2 Review of relevant small area literature

Small area estimation models have seldom been used to estimate parameters related to firms' activity and performance. A couple of empirical studies carried out on business surveys are given in Hidiroglou and Smith (2005) and Chandra (2009), but they did not exploit Bayesian inference. A reason is that National Statistical Institutes, which are highly interested in small area estimation for business surveys, are obliged to as far as possible published official statistics that are based on estimators with negligible bias. Fear of model misspecification has been a hindrance to wide application of small area estimation (Hedlin, 2008) to business surveys.

Small area models may be broadly classified into "area level" and "unit level". In area level models survey weighted (direct) estimates obtained for each of the domains being studied are related with auxiliary information at the same level of population disaggregation. In "unit level" models the target and auxiliary variables are related at the individual observation level. "Area level" model that straightforwardly lead to design-consistent estimators provided that the direct estimators are design-consistent. Design consistency is a general purpose form of protection against model failures, as it guarantees that, at least for large domains, estimates make sense even if the assumed model completely fails. For this reason "area level" models will be considered in this paper.

An area level model frequently employed in the small area estimation framework is the so called Fay–Herriot model (Fay and Herriot, 1979), where normality is assumed for both the sampling distribution of direct estimators and in the model on unobservable parameters. However, Normality can be a strong assumption for data sets arising from business surveys, that are typically positively skewed and economic

theory suggests that regression relationships are typically multiplicative, that is linear in the log scale. The log-normal distribution is frequently assumed by economists when estimating the production function (where the outcome is the value added): in this literature the log-normal distribution is also named the Cobb-Douglas distribution from the Cobb-Douglas production function (Johnson et al., 1994).

In this article we consider models where normality is replaced by the log-normality assumption. In the literature on small area the use of the log normality for skewed data has been explored by Chandra and Chambers (2011) for the "unit-level" models and Maiti (2004) and Slud and Maiti (2006) for "area-level" models. The approach we adopt in this paper is new as we assume log-normality of both the sampling distribution of the direct estimator and in the model for the underlying population parameter. The approximate log-normality of mean estimators calculated on samples from log-normal populations is reviewed in Cobb et al. (2012).

We adopt a full Bayesian approach to inference, which is particularly appropriate in a context where functional transformations need to be taken into account because it allows to easily compute posterior distributions of transformations of the target parameters. Specifically, once posterior for VA and LC can be simulated from, it is straightforward to make inference on firms' competitiveness.

3. Survey data and direct estimation

We use data on the Small and Medium Enterprises (SME) sample survey (1-99 employees) conducted by the Italian National Statistical Institute (ISTAT) that provided us this information in the framework of the BLUE-ETS project. The sampling design is stratified and strata are defined cross classifying NACE 4 Rev. 2, 2 digits, Italian administrative regions (NUTS2) and company size. A detailed description of the SME survey can be find in Faramondi et al. (2010). With reference to the mentioned outcomes, ISTAT provides reliable estimates for domains defined alternatively by: i) cross-classification of administrative region and economic activity (NACE Rev. 2, 2 digit), ii) cross-classification of size (in classes) and economic activity (NACE Rev. 2, 3 digit), iii) cross-classification of economic activity (NACE Rev. 2, 4 digit).

3.1 The direct estimators and the estimation of their variances

The domains we are interested are obtained cross classifying the population according to three criteria. Specifically, they are: regions where firms are located (North-West, North-East, Centre, South, Islands), economic activity (NACE Rev. 2, 2 digit), size (in four classes: less than 10 employees, from 10 to 19 employees, from 20 to 49 employees, from 50 to 99 employees).

Let the $\hat{\theta}_{ijh,k}$ be the direct estimator of the outcome parameter $\theta_{ijh,k}$ in the jh-th domain, where i indexes the regions ($i=1,...,5$), j the economic activity ($j=1,...,23$), h the classes of employees ($h=1,...,4$). The k index refers to the two outcomes, that are value added (VA) and labour cost (LC), then $k=VA,LC$.

As the domains of interest are union of strata, direct estimates can be obtained using the calibration estimator ISTAT adopts for the SME survey. Calibration estimators (Deville and Särndal, 1992) can be written as weighted sums. ISTAT's published weights are obtained by multiplying base weights (inverse of the inclusion probabilities) by factors adjusting for non-response and calibrating to known totals. More details on the construction of weights for the SME survey can be found in ISTAT (2007).

In order to obtain design-based variances we use a bootstrap procedure. Specifically, we consider the technique for finite populations proposed by Särndal et al. (1992, page 442). With respect to possible alternatives, such as the linearization method, the bootstrap strategy enables us to explore the sampling distribution of sampling estimates, calculating confidence intervals. Moreover the procedure can be extended to the calculation of the variance for non-linear parameters. We also calculated the variances using the linearization method and we found that this procedure leads to estimates that are consistent with those obtained using the bootstrap procedure (the correlation between the two estimates' series is 0.96).

Looking at the bootstrap estimates, we note that the first, second and third quartiles of the estimated coefficient of variation of direct estimates are respectively 8%, 12% and 18% while its maximum is 123%. These results confirm the necessity to adopt a small area model approach.

4. The considered small area estimation models

From the analyses of bootstrap replications we observed that the sampling distribution of the direct estimators are positively skewed in most cases, a consequence of the large positive skewness of VA and LC population distributions. The log transformation of direct estimators can be suitably adopted in order to get more symmetric distribution.

Let $\boldsymbol{\theta}_{ijh} = \left(\theta_{ijh,k}\right)_{k=VA,LC}$ the bivariate vector of the population parameters we are interested in, for the $ijh-$th domain, $\hat{\boldsymbol{\theta}}_{ijh}$ the corresponding vector of direct estimators and $\hat{\boldsymbol{\eta}}_{ijh} = \left(\hat{\eta}_{ijh,k}\right)_{k=VA,LC}$ the vector of the log transformed estimators, that is $\hat{\eta}_{ijh,k} = \ln \hat{\theta}_{ijh,k}$, $\hat{\eta}_{ijh,k} \in (0,+\infty)$.

Let's assume that the log transformed direct estimator $\hat{\eta}_{ijh,k}$ is normally distributed. Consequently the direct estimator $\hat{\theta}_{ijh,k}$ has a log-normal distribution. This assumption may be justified using the arguments in Cobb et al. (2012), namely the approximate log-normality of VA and LC in the population.

By denoting with $\mu_{ijh,k}$ and $\sigma_{ijh,k}$ respectively the mean and the standard deviation of the $\hat{\eta}_{ijh,k}$ variable, it is well known that the expected value and the variance of the log-normally distributed variables $\hat{\theta}_{ijh,k}$ are respectively:

$$E\left(\hat{\theta}_{ijh,k}\right) = \exp\left(\mu_{ijh,k} + 0.5\sigma^2{}_{ijh,k}\right) = \theta_{ijh,k} \quad [1]$$

$$VAR\left(\hat{\theta}_{ijh,k}\right) = \left[\exp\left(\sigma^2{}_{ijh,k}\right) - 1\right]\exp\left[2\mu_{ijh,k} + \sigma^2{}_{ijh,k}\right] \quad [2]$$

Equivalently the parameters $\mu_{ijh,k}$ and $\sigma^2{}_{ijh,k}$ can be obtained depending on the expected value and variance of the $\hat{\theta}_{ijh,k}$ variable:

$$\mu_{ijh,k} = \ln\left[E\left(\hat{\theta}_{ijh,k}\right)\right] - 0.5\ln\left[1 + \frac{VAR\left(\hat{\theta}_{ijh,k}\right)}{E\left(\hat{\theta}_{ijh,k}\right)^2}\right] = \ln\theta_{ijh,k} - 0.5\ln\left[1 + CV\left(\hat{\theta}_{ijh,k}\right)\right] \quad [3]$$

$$\sigma^2_{ijh,k} = \ln\left[1 + \frac{VAR\left(\hat{\theta}_{ijh,k}\right)}{E\left(\hat{\theta}_{ijh,k}\right)^2}\right] = \ln\left[1 + CV\left(\hat{\theta}_{ijh,k}\right)\right] \quad [4]$$

In order to get more stable variance estimates than the direct ones, we smooth the estimated variances using a Generalized Variance function approach (Wolter, 1985). Let's indicate with $VAR\left(\hat{\theta}_{ijh,k}\right)$ the variance of direct estimator and with $\text{var}\left(\hat{\theta}_{ijh,k}\right)$ its estimate obtained through the bootstrap algorithm.

In order to specify the smoothing equation, we consider that, as in [4], the variance of the Normally distributed estimator $\hat{\eta}_{ijh,k}$ depends on the coefficient of variation of the loglormally distributed estimator $\hat{\theta}_{ijh,k}$. Consequently we propose a smoothing strategy for this coefficient of variation. Based on the identity $CV^2\left(\hat{\theta}_{ijh,k}\right) = VAR\left(\hat{\theta}_{ijh,k}\right) \Big/ \hat{\theta}^2_{ijh,k}$ and in the spirit of Ferrante and Trivisano (2010) and of Fabrizi et al. (2011), we assume that $CV^2\left(\hat{\theta}_{ijh,k}\right)$ varies with the size class (h) and with the parameter (k) but neither with the economic activity (j), nor with the regions (i). This leads us to specify the following smoothing equation: $\text{var}\left(\hat{\theta}_{ijh,k}\right) = \delta_{h,k}\,\hat{\theta}^2_{ijh,k} \big/ n_{ijh} + \upsilon_{ijh,k}$, with $E\left(\upsilon_{ijh,k}\right) = 0$. The parameter $\delta_{h,k}$ can then be interpreted as the smoothed squared coefficient of variation multiplied for the size of the domain n_{ijh}, thus allowing the decrease of the coefficient of variation decreases when the sample size increases.

The fit of the smoothing model for both *VA* and *LC* is satisfactory. Smoothed estimated coefficient of variations can be obtained as $cv_{smooth}\left(\hat{\theta}_{ijh,k}\right)=\hat{\delta}_{h,k}^{1/2}/n_{ijh}$. Based on [4], the $cv_{smooth}^2\left(\hat{\theta}_{ijh,k}\right)$ are the inputs of the sampling models that will shortly be presented.

Based on the assumptions highlighted in the section 3.4.7 and on results obtained in section 3.4.6, we specify the following small area sampling model:

$$\hat{\boldsymbol{\theta}}_{ijh}\left|\boldsymbol{\mu}_{ijh},\boldsymbol{\Sigma}_{ijh}\right.\overset{ind}{\sim}LogN_2\left(\boldsymbol{\mu}_{ijh},\boldsymbol{\Sigma}_{ijh}\right) \quad [5]$$

The diagonal elements of the 2×2 matrix $\boldsymbol{\Sigma}_{ijh}=\left(\sigma_{ijh,kk'}\right)_{k,k'=VA,LC}$ are considered as known and, based on [4], we assume that $\sigma_{ijh,kk}^2=\ln\left[1+cv_{smooth}^2\left(\hat{\theta}_{ijh,k}\right)\right]$. We assume the independence of direct estimators, then the off-diagonal elements are equals to zero, that is $\sigma_{ijh,kk'}=0$. The role of the sampling direct estimators correlation could be evaluated by relaxing this last assumption.

As we aim at estimating the expected value of the $\hat{\boldsymbol{\theta}}_{ijh}$ estimator, based on [3] we specify the following bivariate linking model for $\ln\boldsymbol{\theta}_{ijh}$:

$$\ln\boldsymbol{\theta}_{ijh}\left|\boldsymbol{\gamma}_{ijh},\boldsymbol{\Sigma}_{v}\right.\overset{ind}{\sim}N_2\left(\boldsymbol{\gamma}_{ijh},\boldsymbol{\Sigma}_{v}\right) \quad [6]$$

where $\boldsymbol{\gamma}_{ijk}=\left(\gamma_{ijh,k}\right)_{k=VA,LC}$. $\boldsymbol{\Sigma}_{v}$ is a positive definite 2×2 matrix where diagonal elements are defined as $\sigma_{v,k}^2$ and the off-diagonal elements are given by $\sigma_{v,kk'}$ ($k,k'=VA,LC$). We denote this as LN-LN model.

We propose to obtain the HB estimator of $\boldsymbol{\theta}_{ijh}$ as follows:

$$_{LN-LN}\hat{\theta}_{ijh,k}^{HB}=E\left[\exp\left(\mu_{ijh,k}^{HB}+0.5\ \ln\left(1+cv_{smooth}^2\left(\hat{\theta}_{ijh,k}\right)\right)\right)\middle|data\right] \quad [7]$$

We also estimate a normal-log-normal (N-LN) model in order to evaluate the impact of assuming log-normality for the sampling distribution of the direct estimators. We consider the following alternative model:

$$\hat{\boldsymbol{\theta}}_{ijh}\left|\boldsymbol{\theta}_{ijh},\boldsymbol{\Psi}_{ijh}\right.\overset{ind}{\sim}N_2\left(\boldsymbol{\theta}_{ijh},\boldsymbol{\Psi}_{ijh}\right) \quad [8]$$

where in the 2×2 matrix $\boldsymbol{\Psi}_{ijh}=\left(\Psi_{ijh,kk'}\right)_{k,k'=VA,LC}$ the diagonal elements are defined as $\Psi_{ijh,kk}=\theta_{ijh,k}^2cv_{smooth}^2\left(\hat{\theta}_{ijh,k}\right)$ and the off-diagonal ones are assumed equals to zero, that is $\Psi_{ijh,kk'}=0$.

For the linking model we assume the following bivariate distribution:

$$\ln\boldsymbol{\theta}_{ijh}\left|\boldsymbol{\gamma}_{ijh},\boldsymbol{\Sigma}_{v}\right.\overset{ind}{\sim}N_2\left(\boldsymbol{\gamma}_{ijh},\boldsymbol{\Sigma}_{v}\right)$$

where $\boldsymbol{\gamma}_{ijk}=\left(\gamma_{ijh,k}\right)_{k=VA,LC}$. $\boldsymbol{\Sigma}_{v}$ is a positive definite 2×2 matrix where diagonal elements are defined as $\sigma_{v,k}^2$ and the off-diagonal elements as $\sigma_{v,kk'}$. In this case the HB estimator of $\boldsymbol{\theta}_{ijh}$ is obtained as:

$$_{N-LN}\hat{\theta}_{ijh,k}^{HB}=E\left(\exp\left(\theta_{ijh,k}\right)\middle|data\right)$$

As usual, in order to improve the direct estimates we use auxiliary information by assuming for the LN-LN and N-LN models by setting $\gamma_{ijh,k}=\alpha_k+x_{ijh}\beta_k$. α_k and β_k are respectively intercept and slope specific for *VA* and *LC*.; x_{ijh} is the log total number of firms' employees in each domain. This auxiliary information refers to the Italian firms' population and is provided by the ISTAT-ASIA administrative archive.

With regard to the prior distribution, we assume $\boldsymbol{\alpha}\sim N_2\left(\mathbf{0},\mathbf{A}\right)$ and $\boldsymbol{\beta}\sim N_2\left(\mathbf{0},\mathbf{B}\right)$ where $\boldsymbol{\alpha}=\left(\alpha_k\right)_{k=VA,LC}$ and $\boldsymbol{\beta}=\left(\beta_k\right)_{k=VA,LC}$. The priors variances are set to very large values with respect to the order of magnitude

of the parameters. We assume $\boldsymbol{\Sigma}_v^{-1} \sim Wishart(\mathbf{I}_2, 2)$, a popular reference prior for precision matrices. These assumptions are to reflect the lack of prior information about model parameters, thus defining diffuse but proper specification of priors.

In order to evaluate the role of the prior variance we consider an alternative prior for the $\boldsymbol{\Sigma}_v^{-1}$ matrix with reference to LN-LN model. Fabrizi and Trivisano (2012) suggest the use of a Generalized Inverse Gaussian (GIG) prior for the estimation of the mean of a log-normal variable. These authors show that, for a careful choice of the prior parameters, assuming GIG on σ^2 parameter of a log-normal may lead to Bayes estimators with optimal frequentist properties. In this line, we adopt independent GIG priors for $\sigma_{v,k}^2$ while a uniform prior over the range $(-1,1)$ is assumed for $\rho_{v,kk'}$; the implicit prior for the off-diagonal elements is based on the relationship $\sigma_{v,kk'} = \rho_{v,kk'}\sigma_{v,k}\sigma_{v,k'}$ $(k, k' = VA, LC)$. The GIG distribution is characterized by three parameters: λ, δ, γ. We set them to $\lambda = -0.5$, $\delta = 0.1$, $\gamma = 2.1$ according to the suggestions in Fabrizi and Trivisano (2012).

For simplicity we will denote the LN-LN models as LN-LN-WISH and LN-LN-GIG according to the variance chosen for the precision matrix; the associated point estimators are denoted as ${}^{WISH}_{LN-LN}\hat{\theta}^{HB}_{ijh,k}$ ${}^{GIG}_{LN-LN}\hat{\theta}^{HB}_{ijh,k}$ respectively. For consistency, the N-LN model endowed with the Wishart prior on $\boldsymbol{\Sigma}_v^{-1}$ will be referred to as N-LN-WISH and the associated point estimator ${}^{WISH}_{N-LN}\hat{\theta}^{HB}_{ijh,k}$

Parameters estimates are obtained by summarizing the posterior distributions approximated by the output of Monte Carlo Markov Chain (MCMC) integration via the Gibbs sampling algorithm. For a brief introduction to this numerical method applied to Bayesian inference, see, for instance, Carlin and Louis (2000, ch. 5). By assuming a quadratic loss, the posterior means are adopted as estimates of the area specific parameters. Posterior variances are used as measure of uncertainty. In order to carefully assess the convergence, we run three parallel chains of 25,000 runs each, the starting point being drawn from an over-dispersed distribution. The convergence of the Gibbs sampler was monitored by visual inspection of the chains' plots and of autocorrelation diagrams, and by means of the potential scale reduction factor proposed by Gelman and Rubin (1992). Both models displayed fast convergence, we discarded the first 5,000 iterations from each chain. To obtain estimates we used the OpenBugs software package (Thomas et al., 2006).

5. Model comparison and checking. Statistical evaluation of small area estimators

In order to choose among competing models, we compute the Deviance Information Criterion (DIC). The DIC is a model selection criterion according to which a model's performance is evaluated as the sum of a measure of fit (the posterior mean of the deviance D) and a measure of model complexity obtained as the difference between D and the deviance evaluated at the parameters' posterior mean. In this way, a model is preferred if it displays a lower DIC value (Spiegelhalter et al., 2002). Table 1 reports the DIC results for the whole set of small area models estimated. DIC values show that the two models in which the lognormality distribution is assumed both in the sampling and linking models (the LN-LN-WISH model and the LN-LN-GIG) perform better in terms of DIC with respect to the remaining models.

Model	DIC
LN-LN-WISH	3576
LN-LN-GIG	3538
N-LN-WISH	3607

Table 1*: Model comparison using DIC statistic*

Secondly we base the comparison on the percentage reduction of the coefficient of variation of small area estimators versus the direct ones, defined as $CVR_k = 100\left(1 - CVR_k^B \,/\, CVR_k^{DIR}\right)$, where the CVR_k^{HB} and CVR_k^{DIR} are respectively the coefficient of variation of Bayesian estimators and of direct estimators. In Table 2 and Table 3 summaries for the CVR_k are reported respectively for VA and LC.

Results reported in Table 2 and Table 3 highlight that for both VA and LC the whole set of considered small area estimators reduce the variability of direct estimators. The median CVR reduction associated with the estimators based on the best performing model as evaluated by DIC (i.e. the LN-LN-GIG model) is satisfying: it is 26% for VA and 17% for LC. For the 25% of the domains the CVR referred to VA is greater than 26% for the estimators based on LN-LN-WISH model and greater than 38% for the estimator based on LN-LN-GIG. For LC the same summaries are respectively equals to 25% for the LN-LN-WISH model and to 34% for the LN-LN-GIG.

	Estimators		
Summaries	$\hat{\theta}^{HB}_{ijh,VA}$ (WISH, LN−LN)	$\hat{\theta}^{HB}_{ijh,VA}$ (GIG, LN−LN)	$\hat{\theta}^{HB}_{ijh,VA}$ (WISH, N−LN)
perc. 0.1	10.90	13.52	7.43
perc. 0.25	15.10	18.00	11.30
median	19.61	23.00	16.54
average	21.54	26.73	17.98
perc. 0.75	26.39	38.11	22.37
perc. 0.90	33.68	43.94	30.89

Table 2: *Summaries for the Coefficient of Variation Reduction of the HB estimators versus the direct one Outcome: total Value Added*

	Estimators		
Summaries	$\hat{\theta}^{HB}_{ijh,LC}$ (WISH, LN−LN)	$\hat{\theta}^{HB}_{ijh,LC}$ (GIG, LN−LN)	$\hat{\theta}^{HB}_{ijh,LC}$ (WISH, N−LN)
perc. 0.1	6.71	8.1	4.37
perc. 0.25	9.79	11.1	8.10
median	14.78	17.3	12.63
average	18.30	22.3	14.42
perc. 0.75	24.70	34.4	17.81
perc. 0.90	35.04	43.0	26.90

Table 3:*Summaries for the Coefficient of Variation Reduction of the HB estimators versus the direct one Outcome: total Labour Cost*

In summary, we may conclude that the LN-LN models are more performing than that based on the assumption of normality for the distribution of the direct estimators. This latter assumption is often justified invoking central limit theorem effects that are unlikely to operate in small samples from skewed populations. As regards the specification of the prior on the variance components we have that alternative priors lead to different results both in terms of fit (as measured by DIC) and efficiency of the associated small area estimators. In our application the use of GIG priors lead to better results.

6. A look to the results

In this section we present some of the results in order to illustrate that can be helpful in the analysis of Italian manufacturing industry. For simplicity we focus on the analysis of labor productivity defined as

$$LP = \frac{\text{Value added}}{\text{Number of employees}}$$

which will be estimated by dividing the our model based estimates of the Value Added with the number of employees that is known from the ISTAT-ASIA administrative archive. The model based estimates

considered in this section are based on [7] and the adoption of the GIG prior for the variance components, i.e. we consider the predictor ${}^{GIG}_{LN-LN}\hat{\theta}^{HB}_{ijh,LC}$.

Let's first consider some industries that traditionally characterize Italy's manufacturing sector. In Figures 1 through 3 we report estimates for LP for the Food, Furniture, Machinery and equipment (n.e.c) sectors.

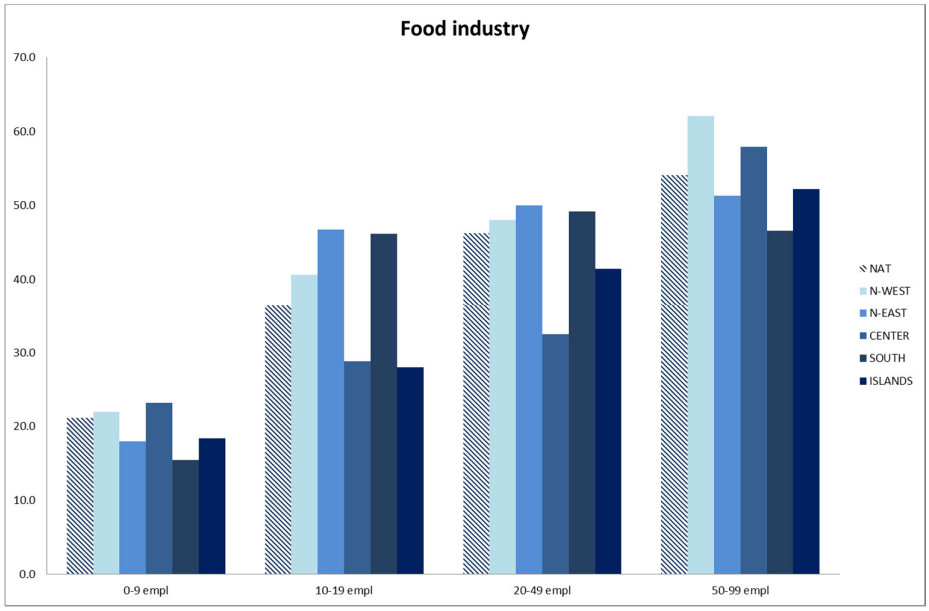

Figure 1: *Labour productivity estimates for the Food industry by region and firm class size.*

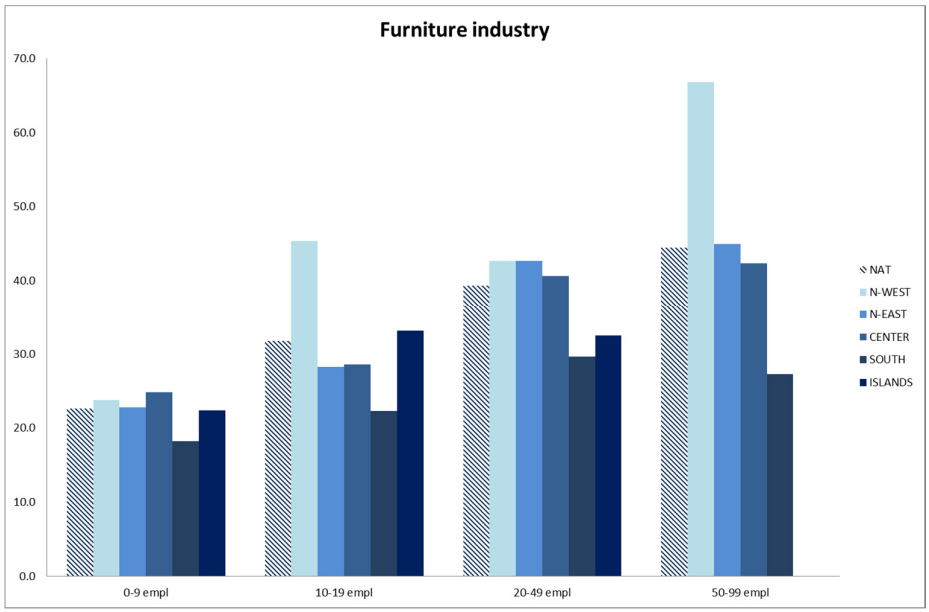

Figure 2: *Labour productivity estimates for the Furniture industry by region and firm class size.*

From Figure 1 (food industry) we may note that the national average of *LP* grows smoothly with firm size. At the regional level this remains approximately true, even if the patterns are different. If we consider the North-East region we have a large increase from the 0-9 to the next size class, but the remaining three attains similar values. For the Center region, the first three (up to 49 employees) have close productivities while it is the fourth to be clearly more productive. The North-South divide is not evident for this industry.

In the furniture industry (Figure 2) we may observe that micro-firms have close *LP* levels in all regions, while for larger firms the North-West attains higher levels, in particular for the size classes 10-19

and 50-99, a result that may be attributed to the presence of several furniture districts in the region. The productivity gap of Southern regions is detectable from the plot, but only for firms with 20 employees or more.

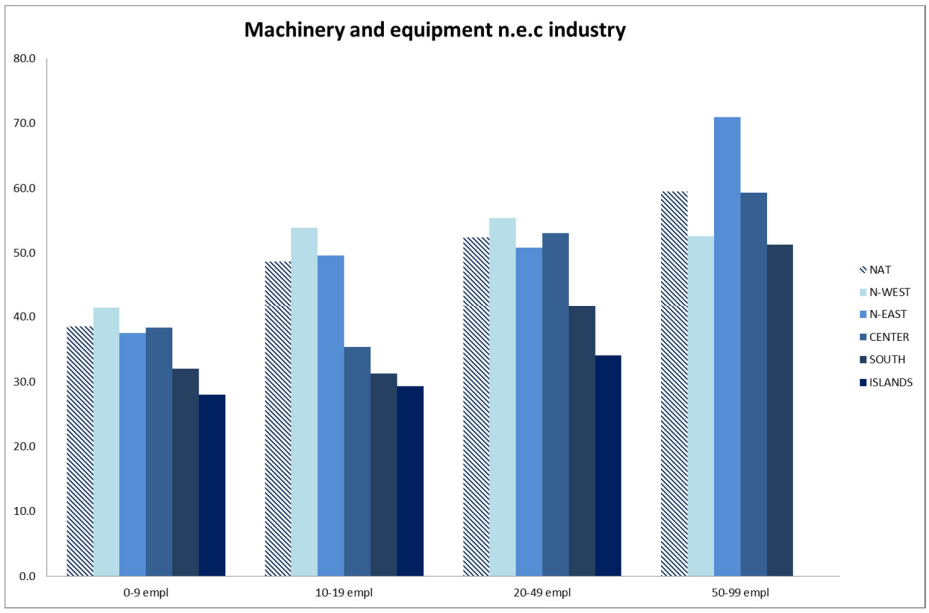

Figure 3: *Labour productivity estimates for the Machinery and equipment industry by region and firm class size.*

From Figure 3 we may observe that in a capital, knowledge intensive sector such as the manufacturing of machinery and equipment, the regional disparities in the *LP* are more evident, with South and Islands lagging behind especially in firms of smaller size. This may depend on the presence of several districts related to this sector in North-East and North-West regions.

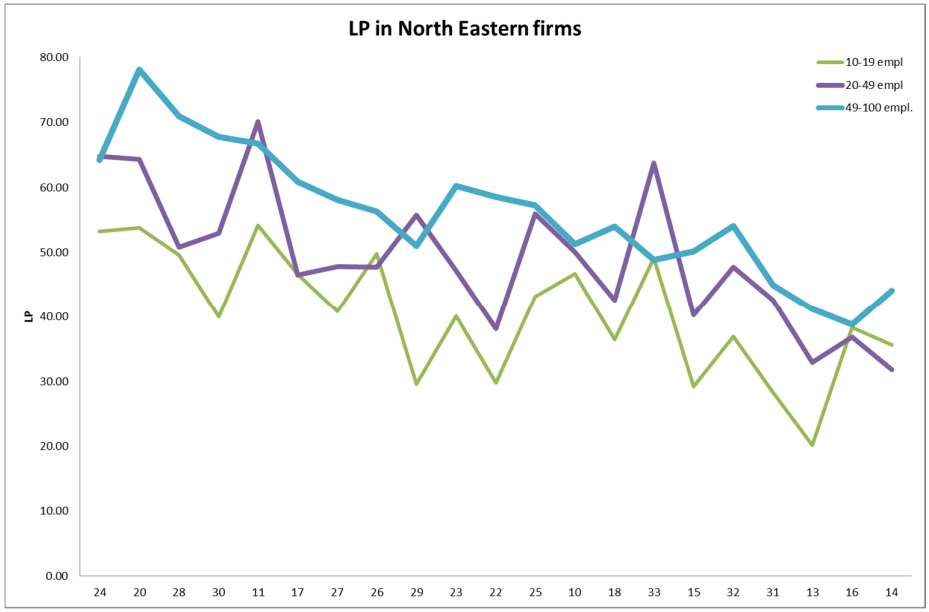

Figure 4: *Labour productivity estimates for the all manufacturing sectors and thee firm size classes in the North East region. Sectors on the x axis are sorted according to National average LP (decreasing). Thicker lines are associated to classes of larger firms.*

In Figure 4 we compare *LP* for 10-19, 20-49, 50-99 size classes and all the sectors for the region North-East (0-9 size class is omitted for simplicity's sake). Manufacturing sectors are presented, from right to left according to decreasing levels of average national labor productivity. We may note that, in general, larger

firms are characterized by higher values of *LP,* but labor productivity grows with size quite differently across sectors and there are several exception to this rule. For sector 33 (Repair and installation of machinery and equipment) firms in the class 20-49 are characterized by clearly higher productivity levels. The productivity gap of small firms is larger in high-tech, knowledge and capital intensive sectors such as 28 (Manufacture of machinery and equipment n.e.c) and 29 (Manufacture of motor vehicles, trailers and semi-trailers) while is smaller for labor intensive sectors. Similar patterns are observed also in the remaining regions.

Acknowledgments

This paper has been completed within the BLU-ETS project "Blue-Enterprises and Trade Statistics", a small or medium-scale focused research project funded by the Seventh Framework Programme of the European Commission, FP7-COOPERATION-SSH (Cooperation Work Programme: Socio-Economic Sciences and the Humanities).

References

Bardone, L. and Reitano, V.E. (2009) "Italy in the Euro Area: the Adjustment Challenge", in M. Buti (ed.) Italy in EMU: The Challenges of Adjustment and Growth, Basingstoke: Palgrave Macmillan.

Chandra, H. (2009): Small Area Estimation for Business Surveys, Proceedings of the Section on Survey Research Methods, American Statistical Association, p. 2803-2809.

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

Carlin, B. P. and Louis T. A. (2000): Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall: New York.

Cobb B.R., Rumì R., Salmeron A. (2012) Approximating the distribution of a sum of log-normal random variables. In: Proceedings of the VI European Workshop on Probabilistic Graphical Models, Granada. Document downloadable at: http://leo.ugr.es/pgm2012/submissions/pgm2012_submission_6.pdf.

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

Deville, J.C., and Särndal C.E. (1992). Calibration Estimators in Survey Sampling. In: Journal of the American Statistical Association, 87, 376-382.

Eberlein, E. and von Hammerstein E.A. (2004). Hyperbolic processes in finance. In R. Dalang, M. Dozzi, and R. F. (Eds.), Seminar on Stochastic Analysis, Random Fields and Applications IV, Progress in Probability, Berlin, DE, pp. 221-264. Birkhäuser Verlag.

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

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

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

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

Fabrizi, E. and Trivisano, C. (2012): Bayesian estimation of log-normal means with finite quadratic expected loss. In: Bayesian Analysis, 7, 975-996.

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. and Tartamella, F. (2010): Regional Gva Inventory ITALY. Research Project Report Metodi e Norme n. 44, 41100.2005.006-2007.252 Inventory on the implementation of regional gross value added in Italy.

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

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

Gelman, A., and Rubin, D.B. (1992): Inference from iterative simulation using multiple sequences. In: Statistical Science, 7, p. 457-511.

Hedlin D. (2008). Small Area Estimation: A Practitioner's Appraisal. In: Rivista Internazionale di Scienze Sociali, 4, p. 407-417.

Hidiroglou M.A. and Smith P.A. (2005): Developing Small Area Estimates For Business Surveys At The ONS. In: Statistics in Transition, 7 (3), p. 527-539.

International Monetary Fund (2011) Italy: Selected Issues, IMF Country Report No. 11/176

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

ISTAT (2010a) Struttura e competitività del sistema delle imprese industriali e dei servizi. Document downloadable at: http://www.istat.it/it/archivio/3972. (in Italian).

ISTAT (2010b) Conti economici delle imprese. Note Metodologiche. Document downloadable at: http://www.istat.it/it/archivio/40287. (in Italian).

Maiti, T. (2004): Applying Jackknife method of mean squared prediction error estimation in SAIPE. In: Statistics in Transition, 6, p. 685–695.

Montanari M. (2011) Italy's regional divide: does it matter also for exports?, Economic analysis from the European Commission's Directorate-General for Economic and Financial Affairs, volume 8, n.2.

Rao J.N.K. (2003) Small area estimation, John Wiley and Sons. New York.

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

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

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

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

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1995): BUGS: Bayesian Inference Using Gibbs Sampling. Version 0.50. Medical Research Council Biostatistics Unit, Cambridge.

Sudgen, R., Smith, T. and Jones R. (2000): Cochran's Rule For Simple Random Sampling. In: Journal of the Royal Statistical Society. Series B, 62, p. 787-793.

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