Bayesian small area estimation methods for business survey statistics

Enrico Fabrizi
Università Cattolica del S. Cuore, Piacenza, Italia – enrico.fabrizi@unicatt.it

Maria Rosaria Ferrante*, Carlo Trivisano
Università di Bologna, Bologna, Italy – maria.ferrante@unibo.it, carlo.trivisano@unibo.it

Abstract

Reliable statistical information on business aggregates by geographical region, industry sector and firm size is an important tool for regional economic studies. Data on small and medium enterprises are typically obtained by means of sample surveys that produce reliable estimates only for relatively large domains. For more detailed estimates small area estimation models can be helpful. In this paper we propose a small area techniques for totals of skewed target variables, a situation that is typical of business data. Specifically, we focus on the value added since it is the basis for calculating important economic aggregates and indicators. We adopt a Bayesian approach to inference, that can be particularly suitable in a context where data transformations are used. The estimates we obtain are noticeably more reliable than those obtained by using standard survey weighted estimators.

Keywords: skewed data, Log-Normal distribution, value added, regional studies.

1. Introduction

The relevance of regional estimates of business aggregates and indicators for researchers in economics emerges from the growing number of scientific articles (Breinlich et al., 2014). Regional statistics are produced by the National Statistical Institutes and used by governments for coherently allocating funds. Moreover, regional economic decision and policies rely on accurate business information on sub-national regions and business categories. More specifically, to measure how large a national (and sectorial) economy is, we look at the Gross Domestic Products. At sub-national level, the total value of new goods produced and services provided in a given time period, is measured by the (Gross) Value Added. Sub-national estimates of value added would be even more informative if they were disaggregated both in terms of industry and firms' size since in order to measure the relative contribution of an industry and/or of certain firms size classes to the regional economy. For example Eurostat releases regional estimates of the (gross) value added up to the detail level of EU NUTS 3 regions (following the Nomenclature of territorial units for statistics, Eurostat, 2011) and branch (NACE 1 digit, Rev. 2, following the Statistical classification of economic activities in the European Community, as in http://ec.europa.eu/eurostat/statistics-explained/index.php). No estimates for regions or business size classes finer than those described in these documents are published because sample sizes of official business surveys are too small for the standard design-based estimators to be adequately precise. This limitation can be overcome by model-based small area estimation methods. The small area estimation literature has until very recently focused mostly on the analysis of surveys, with estimation goals such as the poverty mapping (Pfeffermann, 2014) with few applications on business statistics. In the last few years the attention to this field of application is growing (Militino et al., 2015; Burgard et al. 2014, Chandra, 2009).

In this paper we propose a new small area estimation method aimed at the estimation of business statistics. We focus on sub-populations of Italian small and medium sized manufacturing firms classified according to sub-national region, industry and firm size classes, and on the estimation of the value added since it is the basis for calculating important economic aggregates and indicators as labor productivity. We limit our attention to small and medium enterprises, that is on firms with less than 100 employees, since generally, and so in Italy, larger firms are censused and small area estimation is therefore not needed. 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, financially supported by the EU Commission within the 7th Framework Programme.

The domains we are interested in are smaller than those for which ISTAT provide reliable estimates. Specifically, our domains are defined as: the 20 Italian NUTS 2 administrative regions, the economic industrial sector (NACE Rev. 2, 2 digit) and the firms' size (less than 10 employees, from 10 to 19 employees, from 20 to 49 employees, from 50 to 99 employees). As anticipated, for domains as small as those that we target, standard design-based estimators are characterized by unacceptably large variances. For this reason we use small area estimation methods and, specifically, area-level models (Rao, 2003).

In the context of our problem, specifying area-level small area estimation models we should at first take into account some peculiarities of business data (Rivière, 2002) and namely the typical positive skeweness of data. When population is skewed and sample sizes are small, the normality distribution assumed for estimators of means and totals, typical of small area estimation models and based on Central Limit Theorem, can fail.

We propose a small area estimation model for positively skewed survey data based on the log-normality assumption. We adopt a full Bayesian approach to inference, that allows us to easily handle a model based on lognormal distributional assumption to propose posterior means as point estimators and to use posterior variances as associated measures of uncertainty. Bayesian inference involve the specification of priors: we discuss a careful choice of these priors. We evaluate the performance of the proposed estimators through the usual indicators adopted in the small area literature (Rao, 2003).

2. Review of the relevant small area literature and motivation of the proposed approach

Small area models may be broadly classified into “area level” and “unit level”. In area level models survey weighted (direct) estimates obtained for each domain are related with auxiliary information at the same level of population disaggregation. In “unit level” models the target variables and auxiliary variables are related at the statistical unit level. “Area level” model straightforwardly incorporate information on the sampling design and non-response re-weighting adjustments, leading to design-consistent estimators whenever 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 Fay–Herriot model (Fay and Herriot, 1979), where normality is assumed for both the sampling distribution of direct estimators and in the model for unobservable parameters. Normality can be a strong assumption for business survey variables, that are typically positively skewed. Moreover, economic theory suggests that regression relationships involving production 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. Logarithmic transformations is usually applied to the dependent variables in regression models in order to remove its skewness and then to use linear regression models. Unfortunately predictions require back-transformation (the exponential one) and naïve back-transformation would lead to negatively biased estimators, as is it the case for small area estimators based on exponentiating expected value of data. More specifically, for the “area level” models we deal with, in a frequentist framework to inference, Slud and Maiti (2006) present a bias correction term for small area estimators based on log-transformation. They adopt a mixed effect linear model where the small area model response variable is the log-transformation of the observed small area sampled data. Consequently the small area parameter estimated by this model is the log-transformed outcome parameter. They assume that unbiased design based estimators are available for the small area parameters in the log scale. After obtaining the small area estimates of log transformed responses they achieve good estimates of the small area responses by applying the inverse (exponential) transformation to the first ones corrected for the bias due to the area random effect. The effectiveness of the correction they adopt for the bias relies on a second-order Taylor expansion. The estimator they obtain for the outcome parameter is asymptotically unbiased and the approximation is, in other words, due to the presence in the correction term of the estimated random effect variance on the log-scale.

Also Fay and Herriot (1979), who estimate per-capita income in small areas, adopt a linear mixed model where the outcome variable is the log-transformation of the small area direct estimators. They obtained the small area EBLUP estimator of the parameter in the log-scale and then transformed it back to the original scale. To correct the back-transformation bias, Fay and Herriot benchmark the back-transformed small area estimates to ensure that: i) the total estimated income for small areas in a state with the corresponding direct estimates at the state level, ii) the total estimated income for small areas in a county with the corresponding direct estimates at the county level.

Currently, the Small Area Income and Poverty Estimates (SAIPE) program at the U.S. Census Bureau estimates the number of poor children aged 5-17 in U.S. counties through a small area Fay-Herriot model based on the log of the observed number of the related poor children in the areas of interest (U.S. Census Bureau, 2010). After obtaining the estimates in the log scale, differently from Slud and Maiti, they back transform them in the original scale by using a first-order correction factor and adjust through benchmarking the estimates in the original scale.

In this article we consider models where normality is replaced by the log-normality assumption in this way avoiding the back-transformation of estimates and then the need of correcting for the bias of estimates. More specifically, with respect to the SAE literature where the log-transformation is adopted, we explicitly assume log-normality of both the sampling distribution of the direct estimators and in the model for the underlying population parameters. We specify the SAE model based on design-unbiased direct estimates of the outcome parameter, that is of the parameter in the original scale, differently from the literature just mentioned that assume the design-unbiasedness of direct estimates in the log-scale of the parameter in the log scale.

We follow Rao (2003, p. 133) that suggests that for handling non-linear cases the Empirical Bayes and the Hierarchical Bayes approaches are better suited and we adopt a full Bayesian approach to inference. In the framework where we operate, this approach offers various advantages. A first it is particularly appropriate in a context where non-linear transformations need to be taken into account because it allows to easily compute posterior distributions of transformations of the target parameters. Secondly, MSE is used as a measure of variability under other approaches to SAE, while in the HB approach, that uses the posterior variance as a measure of variability (assuming a prior distribution on the model parameters), the estimation is straightforward in the sense that, the posterior distributions, once computed, can be used for all inferential purposes. A third advantage of the HB approach is its ability to easily manage models assuming that random effects follow a class of distributions instead of relying on the normal distribution (Datta and Lahiri, 1995; Fabrizi et al., 2011; Ferrante and Trivisano, 2010). Fourth, the HB approach is flexible enough to take into account the uncertainty associated to estimates of direct estimators’ variances by incorporating smoothing models in the estimation process, while usually in SAE context the direct sampling variances are assumed known even though they are often estimated (Fabrizi et al., 2011).

3. Survey data, direct estimation and estimation of sampling variances

The SME survey sampling design is stratified and strata are defined cross classifying NACE 4 Rev. 2, 2 digits, Italian administrative regions (NUTS2) and company size. The domains we are interested are obtained cross classifying the firms’ population according to the following variables: Regions where firms are located (NUTS 2), 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). Domains are planned since they are aggregations of sampling strata. We consider 4 repetitions of the survey (2004-2007); we note that samples are independent from occasion to occasion.

Let the $\hat{Y}_{ijrt}$ be the direct estimator of the parameter θ_{ijrt}, where i indexes the economic activity ($i=1,...,22$), j the size classes ($j=1,...,4$), r the NUTS 2 regions ($r=1,...,20$) and t the survey occasions

($t = 1,...,4$). The number of domains is then 1165 that repeated for 4 survey occasion defines 4660 direct estimates input of the SAE model.

As the domains of interest are union of strata, direct estimates can be obtained using the calibration estimator that ISTAT adopts for the SME survey. Calibration estimators 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. Let the estimated total be denoted as $\hat{Y}_{ijrt} = \sum_{k \in d_{ijrt}} w_{ijrt,k} y_{ijrt,k}$ where $y_{ijrt,k}$ is the value added of the *k*-th firm in sector *i*, size class *j*, region *r*, year *t*. For simplicity from now on we denote *ijrt* as "domain". We assume that $E\left(\hat{Y}_{ijrt}\right) = \theta_{ijrt}$ and that $Var\left(\hat{Y}_{ijrt}\right) = V_{ijrt}$. We estimate V_{ijrt} using linearization-based variance estimators.

4. Small area estimation model

Let's indicate with *Y* the outcome variable, that is the firms' value added. Since we observe that the empirical distribution of the *Y* in the whole sample of firms is positively skewed, we assume that this variable is log-normally distributed. As we deal with small sample sizes, the normality assumption is not tenable also for direct estimators that are linear combinations of log-normal observations. A log-normal populations in this case is justified by several authors (e.g. Cobb et al., 2012). Note also that in the econometric context the log transformation of the outcome variable is frequently adopted in order to get more symmetric distribution in the estimation of linear models.

Let $\eta_{ijrt} = \ln\left(\theta_{ijrt}\right)$. We assume that the total' direct estimator $\hat{Y}_{ijrt}$ is log-normally distributed:

$$\hat{Y}_{ijrt} \left| \theta_{ijrt}, V_{ijrt} \sim LN\left(\left[\theta_{ijrt}\right], \left[V_{ijrt}\right]\right)\right. \qquad [1]$$

where $[\bullet]$ is used to denote expectation and variance of the distribution. A specification on the log-scale consistent with the [1] is given by:

$$\log\left(\hat{Y}_{ijrt}\right) \sim N\left(\eta_{ijrt} - \delta_{ijrt}/2, \delta_{ijrt}\right) \qquad [2]$$

where $\delta_{ijrt} = Var\left(\log\left(\hat{Y}_{ijrt}\right)\right)$. In order to get more stable direct variance estimates we smooth them considering that under the log-normality assumption of $\hat{Y}_{ijrt}$,

$$Var\left(\log\left(\hat{Y}_{ijrt}\right)\right) = \log\left[CV^2\left(\hat{Y}_{ijrt}\right) + 1\right]. \qquad [3]$$

Based on the identity $CV^2\left(\hat{Y}_{ijrt}\right) = \frac{V_{ijrt}}{\hat{Y}^2_{ijrt}}$ and after some explorative analysis, we assume that $CV^2\left(\hat{Y}_{ijrt}\right)$ varies with the size class (*j*) and with time (*t*) but neither with the economic activity (*i*), nor with the regions (*r*). This leads to the following smoothing equation for the direct estimate of the V_{ijrt}, $\mathrm{var}\left(\hat{Y}_{ijrt}\right)$:

$$\mathrm{var}\left(\hat{Y}_{ijrt}\right) = \phi_{jt} \frac{\hat{Y}^2_{ijrt}}{n_{ijrt}} \left(1 - \frac{n_{ijrt}}{N_{ijrt}}\right) + \upsilon_{ijrt} \qquad [4]$$

with $E\left(\upsilon_{ijrt}\right) = 0$ and where a finite population correction factor is also considered to account for varying and occasionally non-negligible sample rates. The parameter ϕ_{jt} can be interpreted as the smoothed squared coefficient of variation multiplied for the size of the domain n_{ijrt}, thus allowing the decrease of the coefficient of variation when the sample size increases. The fit of the smoothing model is satisfactory. Smoothed squared estimated coefficient of variations can be obtained as $cv^2{}_{smooth}\left(\hat{Y}_{ijh,k}\right) = \frac{\phi_{jt}}{n_{ijrt}}\left(1 - \frac{n_{ijrt}}{N_{ijrt}}\right)$ and they are the inputs of the sampling models that will be presented. In other words we assume that $\delta_{ijrt} = \log\left[cv_{smooth}{}^2\left(\hat{Y}_{ijrt}\right) + 1\right]$. The first, second and third quartiles of

$cv^2{}_{smooth}\left(\hat{Y}_{ijh,k}\right)$ are respectively 31%, 45% and 65%. These results confirm the need to adopt a small area model approach.

We assume a multiplicative model for θ_{ijrt} that links the outcome parameter to the auxiliary information in order to improve the direct estimates:

$$\theta_{ijrt} = \exp\left(\alpha + x_{ijrt}\beta + u_{ijrt}\right) \qquad [5]$$

where u_{ijrt} is a generic random effect associated to θ_{ijrt} which may be specified in different ways.

We denote the model defined by [2] and [5] as LN-LN model. As auxiliary variable we use the log total turnover in each domain. This auxiliary information refers to the Italian firms' population and is provided by the ASIA archive.

In order to evaluate the role of the log-normality assumption of the direct estimates, we also estimate a Normal-Log-Normal (N-LN) model where the [1] is replaced by:

$$\hat{Y}_{ijrt}\left|\theta_{ijrt}, V_{ijrt} \sim N\left(\theta_{ijrt}, cv^2_{smooth}\theta^2_{ijrt}\right)\right. \qquad [6]$$

and the linking model is as defined in [5].

We consider two different specifications for the prior distribution on u_{ijrt} with the aim of using the one that leads to best fit and small area estimator performances.

i) $\mathbf{u}_{\bullet t}\left|\sigma^2_t \sim MVN\left(0, I\sigma^2_t\right)\right.$, $\sigma^2_t \sim IG\left(a,b\right)$

ii) $\mathbf{u}_{\bullet t}\left|\psi_{ijrt} \sim MVN\left(0, \mathrm{diag}\left(\psi_{ijrt}\right)\right)\right.$, $\psi_{ijrt}\left|a_t, \lambda_t \sim Gamma\left(a_t, \lambda_t/2\right)\right.$, $\lambda_t\left|c_{0t}, b_{0t}, a_t \sim Gamma\left(a_{0t}, b_{0t}/2a_t\right)\right.$

Specification i) represent the standard solution in the applied Bayesian literature. As regards specification ii), we note that it implies that $u_{ijrt}\left|a_t, \lambda_t\right.$ follows a Variance Gamma distribution, i.e.

$$p\left(u_{ijrt}\left|a_t, \lambda_t\right.\right) = \frac{\lambda_t^{0.5a_t+0.25}}{\sqrt{\pi}\,2^{a_t-1/2}\,\Gamma\left(a_t\right)}\left|u_{ijrt}\right|^{a_t-1/2} K_{a_t-1/2}\left(\left|u_{ijrt}\right|\lambda_t^{0.5}\right)$$

where $K(\ )$ denotes the modified Bessel function of the third kind. This marginal prior distribution is either peaked around 0 and has heavier than normal tails belongs to the family of continuous shrinkage priors that can be considered valid alternatives to "spike and slab" priors (Griffin and Brown, 2010). The conjugate hierarchy in ii) also facilitates MCMC sampling. Hyperprior specification: $\alpha \sim N\left(0,.001\right)$, $\beta \sim N\left(0,.001\right)$, $a = .01$, $b = .01$. Prior distributions for a_t and λ_t are chosen in line with Fruhwirth-Schnatter and Wagner (2010) and Griffin and Brown (2010): $a_t \sim exp(1)$, $c_{0t} = 2$, $b_{0t} = 2$.

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. 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 known as Gelman-Rubin statistic (see Carlin and Louis, 2000, ch. 5). Both models displayed fast convergence, we discarded the first 5,000 iterations from each chain. To obtain estimates we used the OpenBugs software package which is downloadable for free on the internet and is open source.

5. Model comparison

In order to choose among competing models, we compute the Deviance Information Criterion (DIC). A model is preferred if it displays a lower DIC value. Table 1 reports the DIC results for the whole set of small area models estimated. DIC values show that the log-normality assumption at the sampling level the perform better in terms of DIC with respect to the model assuming normality. Besides, the adoption of the prior (ii) leads to a remarkable further reduction of DIC with respect to the standard solution (i) in the applied Bayesian literature

Table 1: *Model comparison*

model	prior	DIC	median CVR
N-LN	i	59810	36.02%
LN-LN	i	59500	37.71%
LN-LN	ii	58850	46.11%

We also base the comparison on the median percentage reduction of the coefficient of variation of 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. Results on median of the CVR_k, reported in Table 1, highlight that the whole set of considered small area estimators reduce the variability of direct estimators. The median *CVR* is sizeable for all the estimated models and that associated with the estimators based on the best performing LN-LN model with the (ii) prior (as evaluated by DIC) is really satisfying and equals to 46%.

References

Breinlich, H., Ottaviano, G.I.P., & Temple, J.R.W. (2014). Regional Growth and Regional Decline. in: Handbook of Economic Growth, edition 1, vol. 2, ch. 4, 683-779 Elsevier.

Burgard, J.P., Munnich R., & Zimmermann T. (2014). The Impact of Sampling Designs on Small Area Estimates for Business Data. Journal of Official Statistics, 30, 4, 749–771.

Carlin, B.P., & Louis, T.A. (2000). Bayes and empirical Bayes data analysis. New York, Chapman and Hall.

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

Cobb B.R., Rumì R. & Salmeron A. (2012). Approximating the distribution of a sum of log-normal random variables. In: Proc. of the VI European Workshop on Probabilistic Graphical Models, Granada.

Datta, G.S. & Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in presence of covariates and outliers. Journal of Multivariate Analysis, 54, 2, 310-328.

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.

Fay, R. & 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., & 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.

Fruhwirth-Schnatter, S., & Wagner, H. (2010). Bayesian variable selection for random intercept modelling of Gaussian and non-Gaussian data. In J. Bernardo, M. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, M. West (eds.), Bayesian Statistics, 9, 165–200, Oxford Univ. Press.

Griffin, J.E., & Brown, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. Bayesian Analysis, 5, 171- 188.

Militino, A.F., Ugarte, M.D. & Goicoa T. (2015). Deriving small area estimates from information technology business surveys. Journal of the Royal Statistical Society A, doi: 10.1111/rssa.12105.

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

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.

U.S. Census Bureau (2010), "2006 - 2009 County-Level Estimation Details" http://www.census.gov/did/www/saipe/methods/statecounty/20062009county.html

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

Pfeffermann D. (2014). Small Area Estimation. In International Encyclopedia of Statistical Science, M. Lovric (eds.), 1346-1349, Springer-Verlag.