Appl. Statist. (2018)
67, *Part* 4, *pp.* 1–19

Bayesian small area estimation for skewed business survey variables

Enrico Fabrizi

Università Cattolica del S. Cuore, Piacenza, Italy

and Maria Rosaria Ferrante and Carlo Trivisano

Università di Bologna, Italy

[Received October 2016. Final revision October 2017]

Summary. In business surveys, estimates of means and totals for subnational regions, industries and business classes can be too imprecise because of the small sample sizes that are available for subpopulations. We propose a small area technique for the estimation of totals for skewed target variables, which are typical of business data. We adopt a Bayesian approach to inference. We specify a prior distribution for the random effects based on the idea of local shrinkage, which is suitable when auxiliary variables with strong predictive power are available: another feature that is often displayed by business survey data. This flexible modelling of random effects leads to predictions in agreement with those based on global shrinkage for most of the areas, but enables us to obtain less shrunken and thereby less biased estimates for areas characterized by large model residuals. We discuss an application based on data from the Italian survey on small and medium enterprises. By means of a simulation exercise, we explore the frequentist properties of the estimators proposed. They are good, and differently from methods based on global shrinkage remain so also for areas characterized by large model residuals.

Keywords: Local shrinkage priors; Log-normal distribution; Regional studies; Robust estimation; Variance gamma distribution

1. Introduction

Regional economic decisions and policies rely on accurate business information regarding subnational regions and business categories. The relevance of regional estimates of business aggregates and the interest in regional disparities in terms of firm competitiveness and productivity is demonstrated by the growing number of scientific papers in this field (see Breinlich *et al*. (2014) for a review).

Regional statistics are produced by the national statistical institutes, and governments use them to allocate funds coherently (for examples of this, see Organisation for Economic Co-operation and Development (2013) and Eurostat (2011, 2015a)). For instance, the (gross) value added, i.e. the total value of new goods produced and services provided in a given time period, is routinely estimated at the national and subnational levels. For the European Union, Eurostat releases regional estimates of the value added at levels as detailed as the European Union '*Nomenclature des unités territoriales statistiques*' (NUTS) 3 (Eurostat, 2015b), and industries ('*Nomenclature statistique des activités dans la Communauté européenne*' (called 'NACE'), revision 2, one digit, following the 'Statistical classification of economic activities in the European

Address for correspondence: Enrico Fabrizi, Dipartimento di Scienze Economichee Sociale, Università Cattolica del S. Cuore, Via Emilia Parmense 84, Piacenza, Emilia-Romagna, Italy.
E-mail: enrico.fabrizi@unicatt.it

 0035–9254/18/67000

Community'). Subnational estimates of value added would be even more informative if they were disaggregated in terms of both industry and firm size for measuring the relative contribution of an industry and of certain firm size classes to the regional economy. Unfortunately, sample sizes of official business surveys are too small for the standard design-based estimators (known as 'direct estimators') to be sufficiently precise in small domains.

This limitation can be overcome by model-based small area estimation methods. The small area estimation literature has until very recently focused largely on the analysis of social surveys, with estimation goals such as poverty mapping (see Pfeffermann (2014) and Pratesi (2016) for a review) and few applications for business statistics. In recent years, awareness of this field of application has grown (Burgard *et al*., 2014; Ferrante and Trivisano, 2010; Militino *et al*., 2015), as well as the availability of reliable administrative archives for firms that can be used to obtain auxiliary information.

Small area models may be broadly classified into *area level* and *unit level*. In area level models, survey-weighted (direct) estimates that are obtained for each domain are related to auxiliary information at the same level of population disaggregation. In unit level models, the target variables and unit level information on auxiliary variables are related at this microlevel. Area level models straightforwardly incorporate information on survey weights, leading to design consistent estimators whenever direct estimators are design consistent (Rao (2003), page 117). 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 model assumed completely fails. Area level modelling is less demanding in terms of data disclosure and overcomes potential problems of record linkage between the survey sample and the administrative archive. For these reasons, area level models will be considered in this paper.

Many business survey variables are positive and positively skewed (Rivière, 2002), so normality is not a tenable assumption in most of the cases. Log-transformation can then be introduced to apply normal linear mixed models on the log-scale. Predictions on the original data scale require back-transformation that is a potential source of bias. Positive skewness of survey variables may cause estimators of means and totals to have non-normal (positively skewed) sampling distributions, when calculated on small samples (see Fay and Herriot (1979) and Karlberg (2000)). Literature on area level modelling on the log-scale includes Fay and Herriot (1979) and Slud and Maiti (2006) which both consider an empirical Bayes approach to inference. In this paper we propose a full Bayes approach, accounting for all sources of uncertainty, effectively dealing with back-transformation bias and implementable with widely available Markov chain Monte Carlo (MCMC) software.

When predicting means or totals for business survey variables, strong covariates from administrative archives are often available. For instance, in our application, aimed at predicting gross value added at the domain level, we can exploit knowledge of turnover for each firm in the population. Area level totals of turnover are strongly correlated with those of value added. Nonetheless, a minority of the areas will typically deviate from the relationship that characterizes most of the others. If we think of modelling in terms of mixed models, we have that random effects would be needed for a subset of the areas (Datta, 2011) or alternatively that there are subsets of random effects characterized by different variances. The specification of spike-and-slab priors can be useful in this case (Datta and Mandal, 2015).

We contribute to the small area literature by proposing an approach based on local shrinkage priors for the random effects (Frühwirth-Schnatter and Wagner, 2010), where spike-and-slab priors are replaced by continuous gamma scale mixture-of-normal distributions (Griffin and Brown, 2010) that lead to marginal variance gamma distributions for the random effects. This flexible modelling of random effects leads to predictions that are close to those that we can

obtain by using standard priors for non-outlying areas, and to less biased predictors for the areas that can be labelled as outliers.

The paper is organized as follows. Model specification is described in Section 2. Specifically in Section 2.1 closed formulae for posterior means conditionally on variance components are illustrated as posterior means are proposed as point predictors. In Section 3, we apply our methodology on real survey data. We use data on the small and medium enterprise (SME) sample survey (1–99 employees) conducted by the Italian Statistical Institute (ISTAT), which provided us with this information within the framework of the BLUE-ETS project; this project has been financially supported by the European Union Commission within the Seventh framework programme. For these data we motivate the recourse to a log-normal likelihood for the direct estimators. In Section 4, we introduce a simulation exercise to explore the frequentist properties of the proposed predictor in comparison with some alternatives, including the estimator of Slud and Maiti (2006). Section 5 presents the study's conclusions.

2. Small area estimation model

Let Y be the target variable, which we assume positive with a positively skewed distribution. Assume that Y is defined on a population U of N units, partitioned into a set of m non-overlapping domains of size $N_d (d = 1, \ldots, m; N = \Sigma_{d=1}^{m} N_d)$. A random sample of overall size n is taken by using a possibly complex design: samples of sizes n_d are drawn from each domain. The small area nature of the problem lies in n_d being too small to enable reliable inference for most of the domains. We assume that individual weights $w_{dj}, j = 1, \ldots, n_d$, are attached to responses y_{dj} to account for unequal selection probabilities and possibly other selection adjustments.

The normal distribution is not suitable to describe either the distribution of Y in the population nor the sampling distribution of the domain totals' direct estimators $\hat{Y}_d = \Sigma_{j=1}^{n_d} w_{dj} y_{dj}$. Although these are linear combinations of individual observations and can be assumed to be approximately normally distributed in large samples, in samples of small size, the sum of a few positively skewed variables remains positively skewed. We assume that the total direct estimators are log-normally distributed:

$$\hat{Y}_d \big| \theta_d, V_d \sim \mathrm{LN}\left([\theta_d], [V_d]\right) \tag{1}$$

where $[\cdot]$ is used to denote a parameterization in terms of the mean and variance of the distribution. Exact or approximate design unbiasedeness of totals' estimators is typical in survey sampling. The distributional assumption in expression (1) can be motivated directly assuming the log-normality of Y. Log-normal approximations of sums of independent log-normals have been justified by several researchers (e.g. Fenton (1960) and Cobb *et al.* (2012)). Moreover, Mazmanyan *et al.* (2009) proposed a log-normal central limit theorem for the approximation of the sum of positively skewed random variables, although not necessarily log-normal. Eventually, the assumption of normality on the log-scale when dealing with mean or total estimators of skewed variables became common in the small area literature (as in Fay and Herriot (1979)).

On the log-scale, a specification that is consistent with the *sampling model* (1) is given by

$$\log(\hat{Y}_d) | \eta_d, \delta_d \sim N(\eta_d - \delta_d/2, \delta_d) \tag{2}$$

where $\eta_d = \ln(\theta_d)$ and $\delta_d = \mathrm{var}\{\log(\hat{Y}_d)\}$. $E\{\log(\hat{Y}_d)\} = \eta_d - \delta_d/2$ is in line with assuming the availability of an unbiased estimator on the original scale of the data: if $E(\hat{Y}_d) = \theta_d$, then $E\{\log(\hat{Y}_d)\} < \log(\theta_d)$. Note that $V_d = \exp(2\theta_d + \delta_d)\{\exp(\delta_d) - 1\}$ will depend on both parameters of the log-normal distribution.

In the small area literature, variances that are associated with direct estimators are usually treated as known constants. In practice, estimates that are obtained with methods such as linearization or the bootstrap are smoothed by using a model involving unknown parameters. In line with the literature on area level models, we shall assume that variances on the log-scale are known and denote them as $\hat{\delta}_d$.

We assume a multiplicative *linking model* for θ_d that links the outcome parameter to the auxiliary information to improve the direct estimators:

$$\theta_d = \exp(\mathbf{x}_d^t \boldsymbol{\beta} + u_d). \tag{3}$$

The p-row vector $\mathbf{x}_d^t$ contains the covariates that are known for domain d from external sources, and u_d is a random intercept that is associated with θ_d. Let us assume that $u_d \sim N(0, \psi_d)$, which implies that

$$\theta_d \sim \mathrm{LN}(\mathbf{x}_d^t \boldsymbol{\beta}, \psi_d) \tag{4}$$

or, equivalently, $\eta_d \sim N(\mathbf{x}_d^t \boldsymbol{\beta}, \psi_d)$. We denote the model that is defined by sampling model (1) and *linking* model (4) as the log-normal–log-normal model.

2.1. Analysis conditional on the variance components

To analyse the model that is defined by expressions (1) and (4), note first that, assuming δ_d as known ($\delta_d = \hat{\delta}_d$) we can rewrite expression (2) as $\hat{Z}_d \sim N(\eta_d, \hat{\delta}_d)$, where $\hat{Z}_d = \log(\hat{Y}_d) + \frac{1}{2}\hat{\delta}_d$. We can use standard results from the analysis of linear mixed models (see Rao (2003), chapter 5) to prove that, conditionally on the regression coefficients $\boldsymbol{\beta}$ and the variances ψ_d,

$$\eta_d | \boldsymbol{\beta}, \psi_d, \mathrm{data} \sim N(\hat{\eta}_d^{B1}, g_{1,d})$$

where $\hat{\eta}_d^{B1} = \mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \gamma_d(\hat{Z}_d - \mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta})$, $g_{1,d} = \gamma_d \hat{\delta}_d$ and $\gamma_d = \psi_d / (\psi_d + \hat{\delta}_d)$. Note that, as a function of the shifted direct estimates $\hat{Z}_d$, $\hat{\eta}_d^{B1}$ is a convex linear combination of a direct component $\hat{Z}_d$ and a synthetic component $\mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta}$, known as the linear composite estimator in the small area literature. If we assume quadratic loss and define $\hat{\theta}_d^{B1} = E(\theta_d | \boldsymbol{\beta}, \psi_d, \mathrm{data})$ as the point predictor for θ_d, we have that

$$\begin{aligned} \hat{\theta}_d^{B1} &= \exp\{\mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \gamma_d(\hat{Z}_d - \mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta}) + \tfrac{1}{2}\gamma_d \hat{\delta}_d\} \\ &= \exp[\mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \gamma_d\{\log(\hat{Y}_d) - \mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta}\} + \gamma_d \hat{\delta}_d]. \end{aligned} \tag{5}$$

This predictor is the product between $\exp(\hat{\eta}_d^{B1})$ and a factor that corrects for the main bias term in the back-transformation; it is in line with formula (4) of Slud and Maiti (2006).

It can also be shown that

$$E(\hat{\theta}_d^{B1}) = \exp(\mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \tfrac{1}{2}\psi_d) = E_M(\theta_d) \tag{6}$$

where $E(\hat{\theta}_d^{B1})$ is the expectation taken with respect to both linking and sampling models, whereas with $E_M(\cdot)$ we denote the expectation with respect to linking model (4). This result means that $\hat{\theta}_d^{B1}$ is an unbiased predictor of θ_d in the same sense that best linear unbiased predictors are unbiased: the unconditional frequentist expectation of the estimator and the expectation of the estimand under the linking model are the same. A proof of equation (6) can be found in Appendix A.

If we remove the conditioning on $\boldsymbol{\beta}$ and assume a non-informative flat prior on $\boldsymbol{\beta}$, i.e. $p(\boldsymbol{\beta}) \propto \mathbf{1}$, then we have that

$$\eta_d|\psi_d, \text{data} \sim N(\hat{\eta}_d^{B2}, g_{1,d} + g_{2,d})$$

where

$$\hat{\eta}_d^{B2} = \mathbf{x}_d^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{gls}} + \gamma_d(\hat{Z}_d - \mathbf{x}_d^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{gls}}),$$

$$\hat{\boldsymbol{\beta}}_{\mathrm{gls}} = \left(\sum_d \frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d\mathbf{x}_d^{\mathrm{T}}\right)^{-1} \sum_d \frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d^{\mathrm{T}} \log(\hat{Y}_d)$$

and

$$g_{2,d} = (1-\gamma_d)^2\, \mathbf{x}_d^{\mathrm{T}} \left(\sum_d \frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d\mathbf{x}_d^{\mathrm{T}}\right)^{-1} \mathbf{x}_d.$$

As a consequence, the point predictor under quadratic loss will be given by

$$\hat{\theta}_d^{B2} = \exp\{\mathbf{x}_d^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{gls}} + \gamma_d(\hat{Z}_d - \mathbf{x}_d^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{gls}}) + \tfrac{1}{2}(g_{1,d} + g_{2,d})\} \tag{7}$$

(see Appendix A for a proof). Unlike the empirical Bayes approach that was advocated by Slud and Maiti (2006), who plugged estimates of unknown parameters into equation (7), a full Bayes approach accounts for the effect that the extra variation that is implied by the estimation of $\boldsymbol{\beta}$ has on the point predictor; in fact, the expectation of a log-normal variable depends on both the expectation and the variance on the log-scale.

To account fully for all sources of uncertainty, we should remove the conditioning on the variance components ψ_d; unfortunately, for sensible choices of the prior, this leads to posterior distributions for θ_d that cannot be written in closed form and should therefore be explored by means of computational algorithms such as the MCMC algorithm that is considered in this paper.

2.2. The distribution for the random effects and specification of hyperpriors

The main difference between model (4) and the linking model that has been adopted by most of the small area literature on Fay–Herriot-type models (Jiang and Lahiri, 2005; Pfeffermann, 2014) is that the variances that is associated with random intercepts are in model (4) domain specific, implying local shrinkage instead of the ordinary global shrinkage that we would have had assuming that $\psi_d = \psi$, $\forall\ d$. In a different context, the specification of a distribution for random intercepts based on local shrinkage is discussed in Frühwirth-Schnatter and Wagner (2010).

Datta *et al*. (2011) noted that, in the presence of good covariates, the variability of the small area parameters may be accounted for by a synthetic estimator, and the inclusion of a random-effect term may be unnecessary. When random effects are needed for a subset of the areas, the specification of spike-and-slab priors can be useful (see Datta and Mandal (2015)). Spike-and-slab priors amount to assuming that random intercepts are sampled from a mixture of two normal distributions.

When analysing business data, it is quite likely that auxiliary variables with strong predictive power are available. When this is so, the bulk of the direct estimates will be well fitted by the synthetic model (without random intercepts), so the associated ψ_d are likely to be small, with a minority of areas that require larger area-specific intercepts (and thereby larger ψ_d).

Our specification for the distribution of $u_d, d = 1, \ldots, m$, is based on infinite mixtures of normal distributions. Following Griffin and Brown (2010), our specification uses *gamma* mixtures of normal distributions. Specifically, we assume that

$$u_d|\psi_d \overset{\text{ind}}{\sim} N(0, \psi_d), \tag{8}$$

$$\psi_d|a, \lambda \overset{\text{ind}}{\sim} \text{gamma}(a, \lambda), \tag{9}$$

$$\lambda|b_0, c_0 \sim \text{gamma}(b_0, c_0). \tag{10}$$

This leads to spiked priors for the random effects that at the same time have tails that are heavier than those of the normal distribution. Griffin and Brown (2010) observed that, for small values of the shape parameter a, the prior specification (8)–(10) leads to a marginal prior distribution for u_d that mimics the behaviour of spike-and-slab priors based on finite mixtures. This infinite mixture specification is computationally easier to deal with.

Other choices for the mixing distribution such as the popular *inverse gamma* distribution would lead to platikurtic distributions with heavy tails, such as those in the t-family; this contrasts with the idea of severe shrinkage for most of the areas, which is consistent with a large probability mass close to 0.

Specifically, prior specification (8)–(10) implies that $u_d|a, \lambda$ follows a Variance Gamma distribution, i.e.

$$u_d \sim \text{VG}\{a, \sqrt{(2\lambda)}, 0, 0\}$$

(see Bibby and Sørensen (2003) for more details on this distribution). This marginal prior distributions is symmetric and has expectation $E(u_d) = 0$ and variance $V(u_d) = a/\lambda$. It belongs to the family of generalized hyperbolic distributions (Barndorff-Nielsen, 1977). The conjugate hierarchy in specification (8)–(10) also facilitates MCMC sampling.

In line with Frühwirth-Schnatter and Wagner (2010), we set the shape parameter a to a fixed (small) value, whereas we specify a prior on the global parameter λ. As far as a is concerned, we focus on two choices: $a = 1$ and $a = 0.5$.

The choice $a = 1$ implies that u_d is *a priori* distributed as a double-exponential distribution or Laplace, which, combined with the normal prior conditional on ψ_d, recalls the Bayesian lasso of Park and Casella (2008).

The case $a = 0.5$ represents a more peaked prior distribution and encourages more shrinkage towards 0 of small random intercepts (Fig. 1). Moreover, it leads to a *half-t* marginal prior on $\sqrt{\psi_d}$. The half-t-prior for standard deviations is discussed in Gelman (2006) and recommended whenever it makes sense to put a sizable mass of prior probability close to 0. It can be shown that once $\boldsymbol{\psi} = \{\psi_d\}$ and $\boldsymbol{\tau} = \{\sqrt{\psi_d}\}$, $d = 1, \ldots, m$, have been defined, under prior (8)–(10) and $a = 0.5$,

$$\boldsymbol{\tau}|b_0, c_0 \sim \text{Mht}\left(0, \frac{2b_0}{c_0}\mathbf{I}, 2b_0\right). \tag{11}$$

With $\text{Mht}(\cdot)$ we denote the multivariate distribution (with support $\mathbb{R}^{+n}$) that is obtained from a multivariate t-distribution applying the absolute value transformation on each component of the random variable. We can also prove that each prior $\sqrt{\psi_d}|b_0, c_0$ is a (univariate) half-t-priors and the priors for two different variance components are uncorrelated, i.e. $\text{cov}(\psi_d, \psi_{d*}|b_0, c_0) = 0$ whenever $d \neq d^*$. See Appendix A for a proof of result (11) and the other statements.

As for the prior specification of the remaining parameters, diffuse independent normal priors can be specified for the components of $\boldsymbol{\beta}$. We can set $b_0 = 2$, which implies that $E(\lambda^{-1}) = c_0$. This helps to interpret c_0 as a scaling constant for the random-effects variance $V(u_d) = a/\lambda$. The choice of c_0 depends on the scale for the random effects in the problem being considered. According to result (11), the parameter b_0 can be interpreted in terms of degrees of freedom of

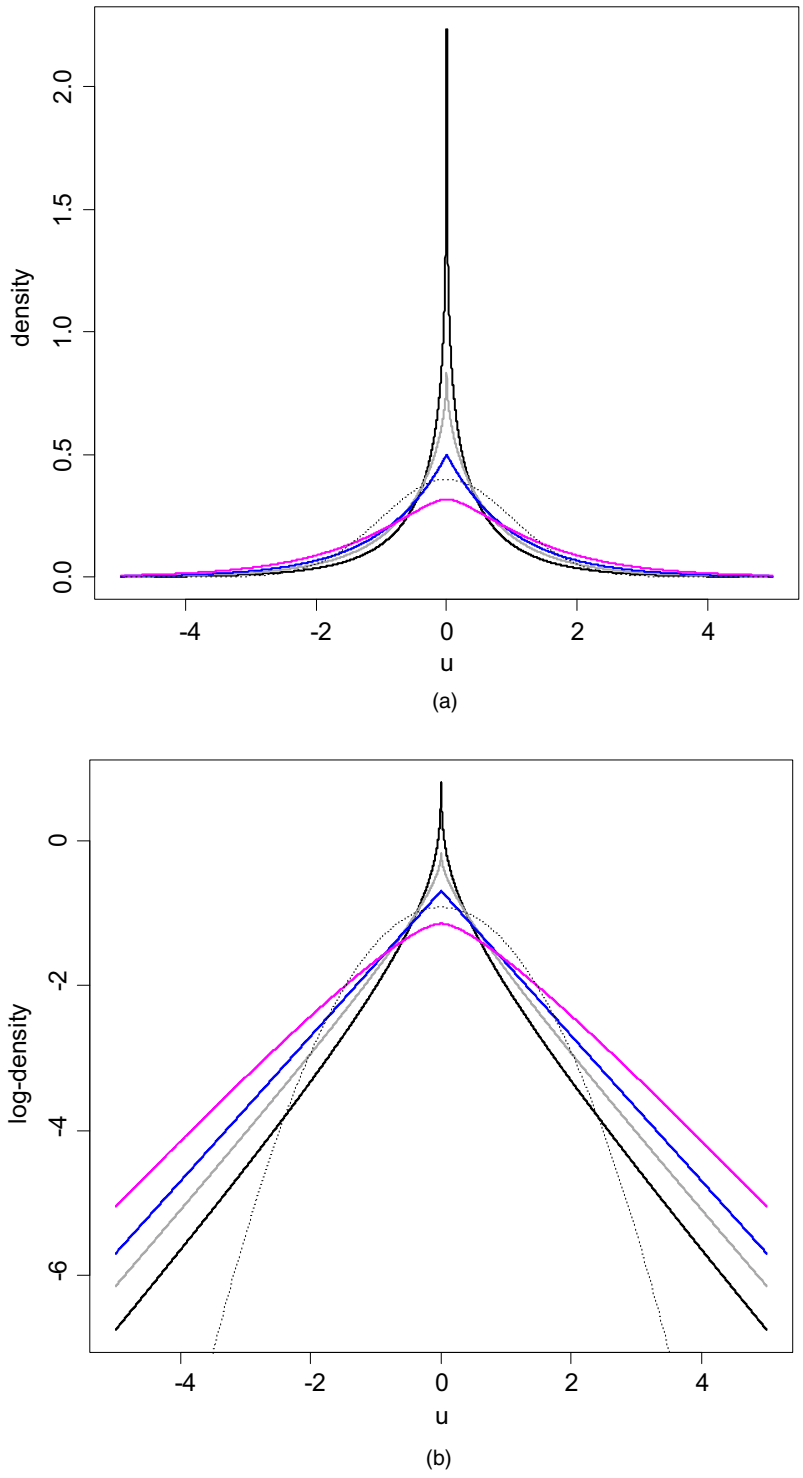

Fig. 1. (a) Density and (b) log-density functions of the variance gamma distribution $VG(a, \sqrt{(2\lambda)}, 0, 0)$ for $\lambda = 1$ and various values of a: ——, $a = 0.5$; ——, $a = 0.75$; ——, $a = 1$; ——, $a = 1.5$; · · · · · ·, normal

the marginal prior $p(\tau)$, so the choice of $b_0 = 2$ is in line with selecting half-t-priors with a very small number of degrees of freedom (Gelman, 2006).

3. Estimation for Italian small and medium enterprise survey data: an application

In this section, we illustrate the methodology that we discussed by using real survey data. We use data on the SME sample survey, wave 2008, conducted by ISTAT. Specifically, we target the estimation of the total value added for small domains of the population of Italian small and medium manufacturing firms (fewer than 100 employees). The domains that we focus on are smaller than those for which ISTAT provides reliable estimates. Specifically, our domains are defined by cross-classifying the 20 Italian *Nomenclature des unités territoriales statistiques* level 2 administrative regions, the economic industrial sector (NACE revision 2, two digit, 22 industries) and firm size (four classes: fewer than 10 employees; 10–19 employees; 20–49 employees; 50–99 employees). As expected, for domains as small as those that we target, standard design-based estimators are characterized by unacceptably large variances.

3.1. Direct estimators and sampling model

The SME survey uses a stratified sampling design and strata are defined by cross-classifying NACE revision 2 (four digits) Italian administrative regions and company size in the four classes that were defined above. The domains that we are interested in are planned because they are unions of sampling strata.

Let $\hat{Y}_{ijr}$ be the direct estimator of the parameter θ_{ijr}, where i indexes the economic activity ($i = 1, \ldots, 22$), j the size classes ($j = 1, \ldots, 4$) and r the regions ($r = 1, \ldots, 20$). Given this peculiar structure of the data the domain index d that was used in Section 2 is now replaced by the triplet ijr. The potential number of 1760 domains falls to 1165, as some of the populations obtained by cross-classification are empty and some very small. We excluded all the domains that were characterized by a sampling rate over 0.75.

The actual sample sizes for the domains that we consider ranges from 2 to 184, with a median of 8, a mean of 13.5 and 0.75- and 0.9-quantiles equal to 16 and 30 respectively.

Direct estimates can be obtained by 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 (the inverse of the inclusion probabilities) by factors adjusting for non-response and calibrating to known totals. Let the estimated total be denoted as $\hat{Y}_{ijr} = \Sigma_{k \in d_{ijr}} w_{ijr,k} y_{ijr,k}$, where $y_{ijr,k}$ is the value added (VA) of the kth firm in sector i, size class j and region r. We assume that $E(\hat{Y}_{ijr}) = \theta_{ijr}$ with $\mathrm{var}(\hat{Y}_{ijr}) = V_{ijr}$. We estimate V_{ijr} by using linearization-based variance estimators and denote these estimates as $\hat{V}_{ijr}$.

In our sampling model we assume log-normality according to model (1). To justify this assumption for our data we proceed in two steps: first, we check whether log-normality is a sensible assumption for domain-specific sample data; then we use a simple simulation exercise to assess whether log-normality is to be preferred to normality as the sampling distribution of total estimators given the sample sizes that we have in our analysis.

For all the domains with $n_{ijr} \geqslant 3$, we tested normality and log-normality by using the Shapiro–Wilk test. Results are reported in Table 1. In reporting the results we consider separately the smallest 90% of the domains $n_{ijr} \leqslant 30$ and the largest 10%. In the smaller domains, for which the test is relatively less powerful, both normality and log-normality tend to be not rejected, but normality fails clearly more often. In larger domains, when the test has more power, normality is rejected in the large majority of cases, whereas log-normality is accepted in more than 70% of the cases.

Table 1. Checking normality and log-normality within domain-specific samples by using the Shapiro–Wilk test: percentage of non-rejections at the 0.01 significance level

n_{ijr}	*Normality (%)*	*Log-normality (%)*
$\leqslant 30$	0.733	0.959
> 30	0.087	0.713
Overall	0.672	0.943

Table 2. Kolmogorov–Smirnov distances between the Monte Carlo ($R = 10000$) distribution of the sample mean and two reference distributions, for various sample sizes

Reference distribution	*Results for the following sample sizes* n_i*:*			
	5	*10*	*15*	*20*
Log-normal	0.012	0.014	0.015	0.015
Normal	0.119	0.097	0.081	0.073

From Table 1 we conclude that log-normality is a sensible assumption for the distribution of VA within domains. We actually assume that direct estimators are log-normally distributed according to the arguments that were illustrated in previous section. To check this, we consider a set of log-normal populations $Y_d \sim \mathrm{LN}(\tilde{\mu}_d, \tilde{\sigma}_d)$, $d = 1, \ldots, L$, where $L = 77$ is the number of domains with $n_{ijr} > 30$ for which log-normality was not rejected and $\tilde{\mu}_d$ and $\tilde{\sigma}_d$ are the parameters according to maximum likelihood for these domains. For each of these populations we generated simple random samples of sizes $R = 10\,000$ for each of the following sample sizes: $n_d = 5, 10, 15, 20$. Note that 20 represent the 0.8-quantile of the sample size distribution in our application.

We evaluate how far is the empirical sampling distribution of the sample mean from the normal and the log-normal distributions in terms of Kolmogorov–Smirnov distance averaging over the L populations. In fact, formal hypothesis testing of distributional assumptions with a sample of replicates as large as 10 000 would lead to rejections even in the presence of negligible departures from the null hypothesis. Results, summarized in Table 2, show how log-normality is to be preferred to normality for all sample sizes. We can also note that, as the sample size grows larger, the difference between the two distances decreases.

To obtain more stable direct variance estimates, we smooth them through the generalized variance function approach (Wolter, 1986). To begin with, we consider that, under the log-normality assumption that was introduced in model (1), we have that

$$\mathrm{var}\{\log(\hat{Y}_{ijr})\} = \log\{\mathrm{CV}^2(\hat{Y}_{ijr}) + 1\}. \tag{12}$$

Thus, the smoothing can be conducted on $\mathrm{CV}^2(\hat{Y}_{ijr}) = \hat{V}_{ijr}/\hat{Y}^2_{ijr}$. After careful exploratory analysis, we assume that $\mathrm{CV}^2(\hat{Y}_{ijr})$ varies with the size class j but not with economic activity i or with regions r. This leads to the following smoothing equation for the direct estimate of V_{ijr}, $\hat{V}_{ijr}$:

$$\hat{V}_{ijr} = \phi_j \frac{\hat{Y}^2_{ijr}}{n_{ijr}} \left(1 - \frac{n_{ijr}}{N_{ijr}}\right) + \upsilon_{ijr} \tag{13}$$

with $E(\upsilon_{ijr}) = 0$ and $V(\upsilon_{ijr}) = \kappa$ and where a finite population correction factor is also considered to account for varying and occasionally non-negligible sample rates. The parameter ϕ_j can be interpreted as the smoothed squared coefficient of variation multiplied by the size of the domain n_{ijr}. The domain sample size n_{ijr} in the denominator of equation (13) allows for the decrease in the coefficient of variation when the sample size increases. Smoothed squared estimated coefficients of variation are given by

$$\mathrm{CV}^2_{\mathrm{smooth}}\left(\hat{Y}_{ijr,k}\right) = \frac{\phi_j}{n_{ijr}} \left(1 - \frac{n_{ijr}}{N_{ijr}}\right);$$

the first, second and third quartiles of $\mathrm{CV}^2_{\mathrm{smooth}}(\hat{Y}_{ijh,k})$ estimated on our data set are 31%, 45% and 65% respectively. These results confirm the need to adopt a small area model approach.

We can then adapt the sampling model (2) to our problem, changing the index from d to ijr and δ_{ijr} with $\hat{\delta}_{ijr} = \log\{\mathrm{CV}^2_{\mathrm{smooth}}(\hat{Y}_{ijr}) + 1\}$ defined according to equation (12).

3.2. Auxiliary information and linking model

As an auxiliary variable, the log-total-turnover in each domain is available. This auxiliary information refers to the Italian firms' population and it is provided by the Italian statistical register of active enterprises archive. The predictive power of this covariate is quite strong: the squared correlation coefficient is equal to 0.87 when calculated on variables on their original scale, and it is equal to 0.79 for the log-transformations. In the original scale the high correlation level is influenced by few observations with a larger scale with respect to most of the others.

We assume the multiplicative linking model (4) for θ_{ijr} to link the outcome parameter to the auxiliary information given by the log-total-turnover for the domain in question. With reference to the log-scale, we can write $\eta_{ijr} = \beta_0 + \mathrm{ltt}_{ijr}\beta_1 + u_{ijr}$. The prior for the vector of domain-specific random intercepts u_{ijr} is specified according to expressions (8)–(10). As for the prior specifications not already discussed, we set $\beta_0 \sim N(0, 10^5)$, $\beta_1 \sim N(0, 10^5)$, $b_0 = 2$ and $c_0 = 1$. We chose these values as they provide a reasonable scale for the random-effects variance in our problem.

We also consider the log-normal–log-normal model with an alternative choice for the prior distribution on u_{ijr}:

$$u_{ijr}|\sigma^2 \sim N(0, \sigma^2), \qquad \sigma^2 \sim \mathrm{InverseGamma}\,(c, d). \tag{14}$$

This prior specification, which implements global shrinkage, can be considered as a benchmark for evaluating the effects of prior specification approximating the spike-and-slab prior that was introduced in the previous section, and it represents a routine choice in many applications. We set $c = 0.01$ and $d = 0.01$.

3.3. Markov chain Monte Carlo computational issues

Parameter estimates are obtained by summarizing the posterior distributions approximated by the output of 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 a measure of uncertainty. To assess the convergence carefully, we ran three parallel chains of 25 000 runs each, the starting point being drawn from an overdispersed

distribution. The convergence of the Gibbs sampler was monitored by visual inspection of the chains' plots and auto-correlation diagrams and by means of the potential scale reduction known as the Gelman–Rubin statistic (Carlin and Louis (2000), chapter 5). Both models displayed fast convergence; we discarded the first 5000 iterations from each chain. To obtain estimates, we used the OpenBugs software package, which can be downloaded for free on the Internet and is open source.

3.4. Comparing alternative models

To choose between competing models, we compute the deviance information criterion (DIC) and the logarithm of the pseudo-marginal likelihood LPML (Ibrahim *et al.*, 2001). A model is preferred if it displays a lower DIC value. Table 3 reports the DIC results for the whole set of small area models that were estimated. DIC values show that, in line with expectations, the log-normality assumption at the sampling level performs better in terms of DIC with respect to the model assuming normality. The ordering of alternative models by using LPML is consistent with that obtained by using the DIC. The adoption of the variance gamma model for the random intercepts u_{ijr} leads to a further reduction in DIC value with respect to the more common specification (14).

We also compare the median reduction of the coefficient of variation of estimators with respect to the direct ones, defined as median(CVR_k^h). CVR_k^h is defined as $\mathrm{CVR}_k^h = 1 - \mathrm{CV}_k^h/\mathrm{CV}_k^{\mathrm{DIR}}$, where CV_k^h is the coefficient of variation calculated on the posterior of θ_k (k being a generic index for the areas) under model h, whereas $\mathrm{CV}_k^{\mathrm{DIR}}$ is the coefficient of variation of the direct estimators calculated from the randomization distribution.

The posterior predictive approach can be used to assess the fit of a model (Gelman *et al.*, 1996). We consider a discrepancy measure that was suggested in the context of small area estimation by You and Rao (2002) and considered also in Fabrizi and Trivisano (2016):

$$\mathrm{dis}_{ijr} = P(\hat{Y}_{ijr} < Y^*_{ijr})$$

where Y^*_{ijr} is generated from the posterior predictive distribution. The discrepancy measure is aimed at checking whether, for each area, the posterior predictive distribution is centred near the direct estimate. Values of dis_{ijr} far from 0 and 1 would provide evidence of systematic underestimation or overestimation. For the log-normal model endowed with priors (10)–(12) and $a = 0.5$ (i.e. the best model in terms of DIC and LPML), we have that the average of the discrepancy measure over the set of areas is 0.499 with 0.25- and 0.75-quantiles equal 0.32 and 0.68 respectively, which means an adequate fit. Less than 10% of the areas shows dis_{ijr} out of the range (0.2, 0.8). Similar summaries can be obtained for the other models that are considered in Table 3.

Table 3. Comparison of alternative assumptions on the distributions of the random effects

Shrinkage	*Prior on random intercepts* u_{ijr}	a	*DIC*	*LPML*	*Median CVR*
Global	(14)	–	15340	−7846	0.391
Local	(10)–(12)	1	15230	−7808	0.421
Local	(10)–(12)	0.5	15220	−7798	0.455

Results on median(CVR_k^h), reported in Table 3, highlight that the whole set of small area estimators considered considerably reduce the variability of direct estimators, which is consistent with the availability of a strongly predictive auxiliary variable. Nonetheless, even if exploiting the same auxiliary information, the models perform differently, as the prior specification has a non-negligible effect. Prior specifications mimicking the spike-and-slab behaviour enable a further gain in efficiency with respect to priors that are ordinarily used in this type of analysis.

To evaluate the improvements that are enabled by the model-based proposed predictor we can compare the number of small areas with values of the coefficient of variation CV less than 16.6%, between 16.6% and 33.3% and over 33.3% for the direct and the model-based predictor. These thresholds for CV were suggested by Statistics Canada (2007) to provide quality level guidelines for publishing small area estimates; those with a coefficient of variation that is less than 16.6% are considered reliable for general use. Estimates with a coefficient of variation between 16.6% and 33.3% should be accompanied by a warning to users. Estimates with coefficients of variation larger than 33.3% are deemed to be unreliable. Less than 25% of the direct estimates have associated CV below the 33.3% threshold, whereas for the model based estimates this number grows to 70%. Although the uncertainty around the small area estimates remains sizable and not all estimates would be publishable, the application of the method proposed endows most subpopulation with a publishable estimate in spite of the small sample sizes.

Fig. 2 displays the effect of alternative prior specifications on the ensemble of the random intercepts' posterior means $E(u_{ijr}|\text{data})$ under the normal prior (14) and variance gamma priors $u_{ijr} \sim \mathrm{VG}\{0.5, \sqrt{(2\lambda)}, 0, 0\}$. On the basis of Fig. 2(a), it is clear that the variance gamma prior leads to a more peaked distribution of estimated random intercepts, as predicted by theory. As tail behaviour is difficult to read from density estimates, in Fig. 2(b), we plot the point posterior expectation under the normal prior *versus* those obtained under the variance gamma prior. The peak around 0 is still apparent from the inflection of the points' cloud approximately at 0; heavier tails under the variance gamma prior can also be appreciated: under the normal prior, $E(u_{ijr}|\text{data})$ lies within the interval $(-4, 2)$, whereas they do not under the variance gamma specification.

As the purpose of small area estimators is to complement the direct estimates that are obtained by using ordinary survey-weighted methods, robustness with respect to modelling assumptions is a major concern. As expected, the recourse to area level model-based methods entails design consistency. When the area-specific sample sizes are large (and occasionally they are) the small area estimate will be close to that obtained under standard design-based methods. This offers, at least for the larger domains, protection against model misspecifications; moreover, it automatically guarantees that, in the case of large domains, model-based and design-based estimates are automatically in agreement.

4. A simulation assessment of the frequentist properties of the proposed point estimators

In this section we introduce a simulation exercise with the aim of investigating the frequentist properties of point estimators (i.e. posterior means) that were introduced in Section 2 and applied to the analysis of the SME survey in Section 3. We study bias, mean-square error and frequentist coverage of posterior probability intervals.

The simulation is design based and we do not assume any parametric distribution when generating the data. We create a synthetic population merging the samples of the 2007 and 2008 SME survey that was discussed in the previous section. We drop from the 2007 wave those firms that were sampled also in the 2008 wave. We obtain a population of size $N = 30451$. Domains

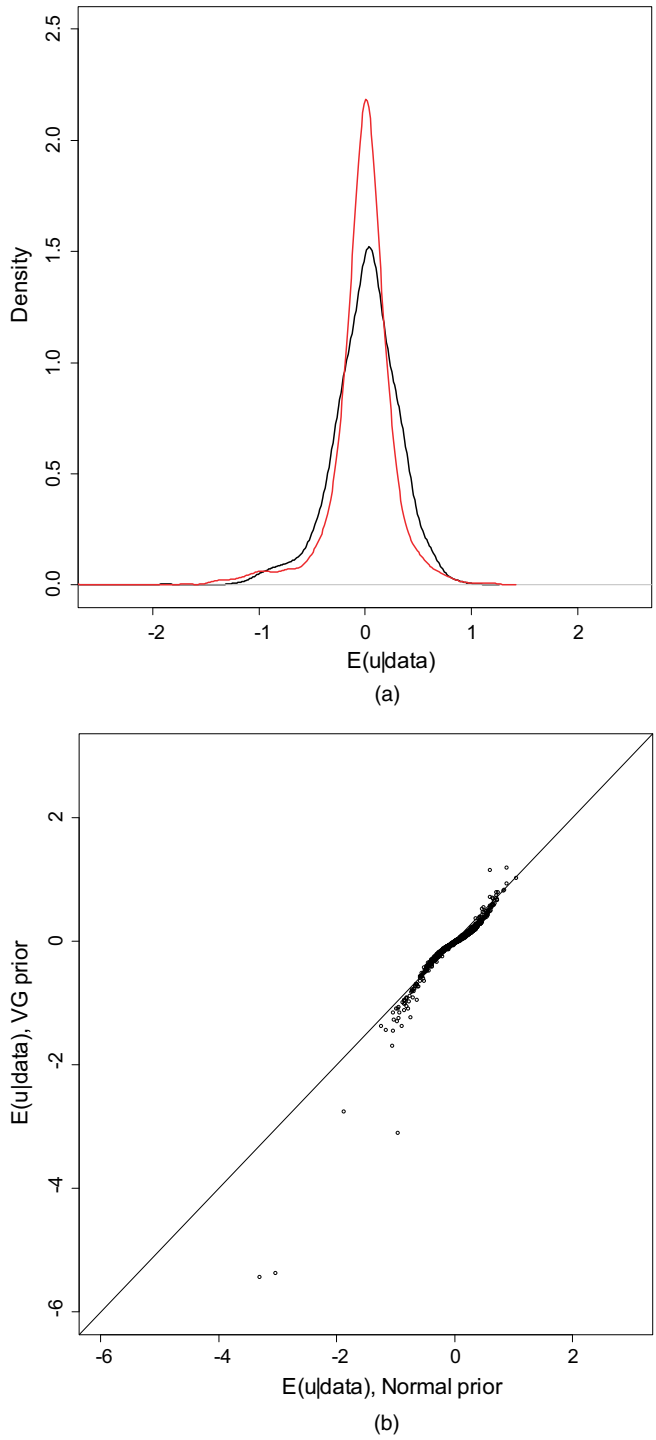

Fig. 2. (a) kernel densities of $E(u_{ijr}|\text{data})$ over the set of small areas under normal (———) and variance gamma (———) priors and (b) $E(u_{ijr}|\text{data})$ under a normal prior *versus* $E(u_{ijr}|\text{data})$ under a variance gamma prior: for variance gamma priors, $a = 0.5$ is assumed

are defined by cross-classifying the population by firm size and industry sector; with respect to the data analysis of the previous section we collapse the regions. By reducing the number of domains we create subpopulations that are sufficiently large to be sampled by using reasonable sampling rates. Collapsing by region has a milder effect on subpopulation skewness with respect to firm size or industry sector. Thereby our synthetic population is divided into $m = 88$ domains whose sizes N_d $(d = 1, \ldots, m)$ range from 14 to 1339 with an average of 346. The same target parameters (total value added) and auxiliary variable (turnover) that were studied in Section 4 are considered.

We keep this synthetic population as fixed and we repeatedly draw stratified samples with proportional allocation and a 4% sampling rate. Resulting domain-specific sample sizes n_d are adjusted so that $\min(n_d) = 3$. The resulting average domain-specific sample sizes is 14.05, with a maximum of 54. The mean sample size is very close to that of the application.

The Monte Carlo exercise is based on $S = 2000$ repeated samples. Direct estimates and their variances are calculated by using analytic formulae. The point estimators that we compare in the simulation are as follows:

(a) the posterior mean associated with the Fay–Herriot model specified on the untransformed scale, UFH (the untransformed Fay–Herriot model can be described as $\hat{Y}_d \sim N(\theta_d, V_d)$, where V_d is the variance of the direct estimator $\hat{Y}_d$, $\theta_d \sim N(\beta_0 + \beta_1 x_d, \sigma_v^2)$, $\sigma_v \sim \text{Unif}(0, A)$, $A = 1000$ and $\beta_i \sim N(0, 10^5)$, $i = 0, 1$; we consider model UFH as it is probably the most 'basic' Bayesian model that a practitioner would think of for analysing these data);
(b) the predictor that was proposed by Slud and Maiti (2006), SM;
(c) the posterior mean obtained from the log-normal model (1)–(3) endowed with the global shrinkage prior (14) we denote the estimator as LNGS ('log-normal with global shrinkage');
(d) the posterior mean that is associated with model (1)–(3) but endowed with the local shrinkage prior described in specification (8)–(10) and $a = 0.5$ (we label the estimator LNLS ('log-normal with local shrinkage').

In all cases we set $\beta_0 \sim N(0, 10^5)$, $\beta_1 \sim N(0, 10^5)$, $b_0 = 2$ and $c_0 = 1$, as in Section 3.

Denote by est_{ds} the generic estimator calculated for domain d in replication s. We compare alternative estimators in terms of relative bias, relative root-mean-square error and frequentist coverage of probability intervals based on the posterior distribution. Specifically we consider the frequentist coverage defined by the $\alpha/2$ and $1 - \alpha/2$ quantiles' posterior distribution of the target parameter (with respect to the coverage probability $1 - \alpha$) and set $\alpha = 0.05$. Comparison tools are defined as follows:

$$\text{RABIAS} = \frac{1}{m} \sum_{d=1}^{m} \frac{\left| \frac{1}{S} \sum_{s=1}^{S} \text{est}_{ds} - Y_d \right|}{Y_d},$$

$$\text{RRMSE} = \frac{1}{m} \sum_{d=1}^{m} \sqrt{\left\{ \frac{\frac{1}{S} \sum_{s=1}^{S} (\text{est}_{ds} - Y_d)^2}{Y_d^2} \right\}},$$

$$\text{COV95} = \frac{1}{m} \sum_{d=1}^{m} \frac{1}{S} \sum_{s=1}^{S} \mathbf{1}[Y_d \in \{ p_{\theta_{ds}|\text{data}}(0.025), p_{\theta_{ds}|\text{data}}(0.975) \}].$$

As we are interested in the frequentist coverage of Bayes estimators, COV95 is calculated only for estimators (a), (c) and (d). The results are summarized in Table 4.

Results from Table 4 show how the posterior means based on the log-normal model with either local or global shrinkage priors and the predictor of Slud and Maiti (2006) perform very closely in terms of mean-square error. In terms of bias LNLS is better, but its variance is somewhat bigger, as we can expect from a more flexible, richly parameterized model. Actually, they are based on similar ideas and models; only the priors or the way that hyperparameters are dealt with are different, so the results are in line with expectation. We did not expect the bias to be close to 0: in small area estimation there is a compromise between the efficiency of a biased synthetic predictor and the unbiasedeness of large variance direct estimators; to some extent estimators that are associated with areas with very small area-specific sample sizes are naturally biased.

The naive Fay–Herriot model, specified on the untransformed scale, performs worse in terms of both bias and mean-square error; the frequentist coverage of the posterior intervals is well below the 0.95 nominal level. This relatively poor performance reflects the misspecifcation of the model, based on the assumption of normality of the direct estimators. It also assumes a linear relationship between direct estimators and the auxiliary variable on the original scale of the data (instead of a linear relationship on the log-scale); we have already noted that this assumption is not completely unrealistic, so misspecification of the sampling model can be held responsible for the not completely satisfying results.

The two hierarchical models lead to close performances also in terms of frequentist coverage of posterior intervals. The advantage of using local shrinkage priors can be appreciated if we consider the performance for outlying areas, i.e. those characterized by a deviation from the synthetic component that is much larger than most of the remaining areas. We investigate performances separataly for the areas that are characterized by the larger (on average) model residual. Results related to the 'worst case' area are presented in Table 5. We note that this area-specific sample is $n_d = 34$, which is well above the average sample size of the simulation.

Predictors LNGS and SM are based on a similar global shrinkage idea. Results in Table 5

Table 4. Comparison of alternative predictors based on the Monte Carlo experiment

Estimator	*RABIAS*	*RRMSE*	*COV95*
UFH	0.1573	0.2026	0.708
SM	0.1263	0.1733	—
LNGS	0.1240	0.1719	0.929
LNLS	0.1113	0.1723	0.941

Table 5. Comparison of alternative predictors for area 3, characterized by the largest model residual

Estimator	*RABIAS*	*RRMSE*	*COV95*
SM	0.4763	0.4790	—
LNGS	0.4414	0.4453	0.505
LNLS	0.0144	0.2515	0.947

show how the common variance parameter that is assumed for the random effects cannot accommodate the 'outlier'; the associated model-based estimators are severely shrunken towards the synthetic component: this implies large bias and poor frequentist coverage of the posterior intervals. The local shrinkage prior that is associated with LNLS is more flexible and leads to an almost unbiased predictor and good coverage.

5. Conclusions

We introduced a Bayesian methodology that is useful for small area estimation of means and totals of variables that are positively skewed. This type of variable is often encountered in business surveys. We devote special attention to the specification of a prior distribution for the random effects; our proposal, based on the idea of local shrinkage, is well suited when auxiliary variables with strong predictive power is available, which is a feature that is often displayed in business survey data.

The methodology proposed can be easily implemented by using widely available MCMC software. Openbugs code, as well as formulae for the full conditional distributions that is needed for an independent implementation of the algorithm, is available on request from the authors.

In summary, we have shown that the predictor based on a local shrinkage prior has overall acceptable frequentist properties, comparable with the alternatives that we consider in the exercise. If most of the areas are well fitted by the model assumed and only a minority are outlying, characterized by larger model residuals, we have that, for these areas, local shrinkage priors can lead to estimators with reduced bias and thereby more efficient.

The strategy that we propose may be applied to estimating business totals based on any positively skewed variables: value added, turnover, labour cost and income from sales and services as well as the components of these main aggregates. We discuss the proposed model with reference to real survey data and, more specifically, to the estimation of the total value added, giving consideration to the fact that the value added is the input for calculating important economic aggregates and performance indicators. We address the subpopulations of Italian small and medium sized manufacturing firms classified according to subnational region, industry and firm size classes. We limit our attention to SMEs, i.e. on firms with fewer than 100 employees because in general, as well as in Italy, larger firms are censused, and small area estimation is therefore not needed.

This research can be extended and complemented in many directions: one important problem that was not considered here is that of benchmarking of small area estimates to known totals for more aggregate domains. A second aspect to address is the longitudinal extension of the model specification to borrow strength not only from covariates but also information repeated over time. This also makes it possible to produce estimates at different time points.

Appendix A

A.1. Proof of equation (6)

To start with we note that

$$\hat{\theta}_d^{B1} = \exp\{\mathbf{x}_d^{\mathrm{T}}\boldsymbol{\beta} + \gamma_d(\hat{Z}_d - \mathbf{x}_d^{\mathrm{T}}\boldsymbol{\beta}) + \tfrac{1}{2}\gamma_d\hat{\delta}_d\} = \exp\{\gamma_d\hat{Z}_d + (1-\gamma_d)\mathbf{x}_d^{\mathrm{T}}\boldsymbol{\beta} + \tfrac{1}{2}\gamma_d\hat{\delta}_d\}$$

so

$$E[\exp\{\gamma_d\hat{Z}_d + (1-\gamma_d)\mathbf{x}_d^{\mathrm{T}}\boldsymbol{\beta} + \tfrac{1}{2}\gamma_d\hat{\delta}_d\}] = \exp\{(1-\gamma_d)\mathbf{x}_d^{\mathrm{T}}\boldsymbol{\beta} + \tfrac{1}{2}\gamma_d\hat{\delta}_d\}E\{\exp(\gamma_d\hat{Z}_d)\}.$$

We note that $E(\hat{Z}_d) = E_M\{E(\hat{Z}_d|\eta_d)\} = E_M\{\eta_d\} = \mathbf{x}_d^{\mathrm{T}}\boldsymbol{\beta}$; analogously

$$V(\hat{Z}_d) = V_M\{E(\hat{Z}_d|\eta_d)\} + E_M\{V(\hat{Z}_d|\eta_d)\} = \psi_d + \hat{\delta}_d.$$

As a consequence

$$E(\gamma_d \hat{Z}_d) = \gamma_d \mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta},$$

$$V(\gamma_d \hat{Z}_d) = \gamma_d^2(\psi_d + \hat{\delta}_d) = \psi_d \gamma_d$$

and

$$E\{\exp(\gamma_d \hat{Z}_d)\} = \exp\left(\gamma_d \mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \frac{1}{2}\psi_d \gamma_d\right).$$

This leads to

$$E(\hat{\theta}_d^{B1}) = \exp\left\{\mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \frac{1}{2}\gamma_d(\psi_d + \hat{\delta}_d)\right\} = \exp\left\{\mathbf{x}_d^{\mathrm{T}} \boldsymbol{\beta} + \frac{1}{2}\psi_d\right\}$$

which coincides with $E_M(\theta_d)$.

A.2. Proof of equation (7)

We introduce some matrix notation. Let $\mathbf{z} = \mathrm{vec}(\hat{Z}_d)$ be the vector containing, $\Psi = \mathrm{diag}(\psi_d)$ and $\boldsymbol{\Delta} = \mathrm{diag}(\hat{\delta}_d)$ the diagonal matrices containing the variance components; let $\mathbf{X}$ be the matrix with rows $\mathbf{x}_d^{\mathrm{T}}$, $d = 1, \ldots, m$.

Standard Bayesian analysis of normal linear mixed models leads to $\boldsymbol{\beta}|\mathbf{z}, \boldsymbol{\psi} \sim N\{\hat{\boldsymbol{\beta}}_{\mathrm{gls}}, \mathbf{V}(\Psi)\}$ where $\hat{\boldsymbol{\beta}}_{\mathrm{gls}} = \{\mathbf{X}^{\mathrm{T}}(\Psi + \boldsymbol{\Delta})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^{\mathrm{T}}(\Psi + \boldsymbol{\Delta})^{-1}\mathbf{z}$ and $\mathbf{V}(\Psi) = \{\mathbf{X}^{\mathrm{T}}(\Psi + \boldsymbol{\Delta})^{-1}\mathbf{X}\}^{-1}$. We can calculate $E(\boldsymbol{\eta}|\mathbf{z}, \Psi) = E_{\boldsymbol{\beta}|\mathbf{z},\Psi}\{E(\boldsymbol{\eta}|\mathbf{z}, \boldsymbol{\beta}, \Psi)\}$ where $\boldsymbol{\eta} = \mathrm{vec}(\eta_d)$. $E_{\boldsymbol{\beta}|z,\Psi}\{E(\eta|z, \beta, \Psi)\} = E_{\boldsymbol{\beta}|z,\Psi}(X\beta) = X\hat{\boldsymbol{\beta}}_{\mathrm{gls}}$. Analogously $V(\boldsymbol{\eta}|\mathbf{z}, \Psi) = V_{\boldsymbol{\beta}|\mathbf{z},\Psi}E(\boldsymbol{\eta}|\mathbf{z}, \boldsymbol{\beta}, \Psi) + E_{\boldsymbol{\beta}|\mathbf{z},\Psi}\{V(\boldsymbol{\eta}|\mathbf{z}, \boldsymbol{\beta}, \Psi)\}$. If we denote the vector of small area predictors (on the log-scale) conditionally on $\boldsymbol{\beta}$ as $\boldsymbol{\eta}^{B1} = E(\boldsymbol{\eta}|\mathbf{z}, \boldsymbol{\beta}, \Psi) = \boldsymbol{\Gamma}\mathbf{z} + (\mathbf{I} - \boldsymbol{\Gamma})\mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\Gamma} = \Psi(\Psi + \boldsymbol{\Delta})^{-1}$ we have that $V(\boldsymbol{\eta}|\mathbf{z}, \Psi) = V_{\boldsymbol{\beta}|\mathbf{z},\Psi}\boldsymbol{\Gamma}\mathbf{z} + (\mathbf{I} - \boldsymbol{\Gamma})\mathbf{X}\boldsymbol{\beta} + E_{\boldsymbol{\beta}|\mathbf{z},\Psi}\mathbf{G}_1$ with $\mathbf{G}_1 = \Psi(\Psi + \boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}$. Taking expectation with respect to $p(\boldsymbol{\beta}|\mathbf{z}, \Psi)$ we obtain $V(\boldsymbol{\eta}|\mathbf{z}, \Psi) = (\mathbf{I} - \boldsymbol{\Gamma})\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}\mathbf{V}(\Psi)\mathbf{X}(\mathbf{I} - \boldsymbol{\Gamma}) + \mathbf{G}_1 = \mathbf{G}_2 + \mathbf{G}_1$. We note that $p(\boldsymbol{\eta}|\mathbf{z}, \Psi)$ is a multivariate normal distribution. If we consider an individual η_d we have that $E(\eta_{ijr}|\mathbf{z}, \Psi) = \mathbf{x}_d^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{gls}}$,

$$V(\eta_d|\mathbf{z}, \Psi) = \gamma_d\hat{\delta}_d + (1 - \gamma_d)^2\mathbf{x}_d^{\mathrm{T}}\left(\sum_{d=1}^{m}\frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d\mathbf{x}_d^{\mathrm{T}}\right)^{-1}\mathbf{x}_d = g_{1,d} + g_{2,d}.$$

As $\theta_d = \exp(\eta_d)$ formula (7) follows.

A.3. Proof of expression (11) and subsequent statements

From assumption (9) we have that

$$p(\boldsymbol{\psi}|\lambda) = \prod_{d=1}^{m}\frac{\lambda^a}{\Gamma(a)}\psi_d^{a-1}\exp\left(-\lambda\sum_{d=1}^{m}\psi_d\right) \propto \lambda^{ma}\exp\left(-\lambda\sum_{d=1}^{m}\psi_d\right)\prod_{d=1}^{m}\psi_d^{a-1}.$$

Conditioning on a is omitted as it is treated as a known constant; m is a short-cut notation for the overall number of areas.

We can obtain the marginal prior $p(\psi)$ by using the integral $p(\psi) = \int_{\mathbb{R}^+} p(\psi|\lambda)\,p(\lambda)\mathrm{d}\lambda$. As $\lambda \sim \mathrm{Gamma}\,(b_0, c_0)$ we straightforwardly obtain

$$p(\psi) \propto \Gamma(ma + b_0) \Big/ \left(\sum_{d=1}^{m}\psi_d + c_0\right)^{ma+b_0}\prod_{d=1}^{m}\psi_d^{a-1}.$$

Applying the transformation $\tau_d = \sqrt{\psi_d}$ on each component of $\boldsymbol{\psi}$ we obtain

$$p(\boldsymbol{\tau}) \propto \left(\sum_{d=1}^{m}\tau_{ijr}^2 + c_0\right)^{-(ma+b_0)}\prod_{d=1}^{m}\tau_d^{2a-1}.$$

For the special case $a=\frac{1}{2}$ the density of $p(\boldsymbol{\tau})$ simplifies to $p(\boldsymbol{\tau}) \propto (\Sigma_{d=1}^{m}\tau_d^2 + c_0)^{-(ma+b_0)}$ or equivalently to

$$p(\boldsymbol{\tau}) \propto \left(1 + 1/c_0 \sum_{d=1}^{m} \tau_d^2\right)^{-(m/2+b_0)},$$

$\tau_d > 0, \forall\, d$. This expression can be recognized as the kernel of the density of a multivariate half-t-distribution with mean vector 0 and diagonal scale matrix. A multivariate half-t distribution is a multivariate t distribution for which we apply the absolute value transformation on each component. We can then write formula (13).

To prove that univariate priors $p(\tau_d)$ we start from $p(\tau_d) = \int \ldots \int p(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}_{-d}$. We can represent as $p(\boldsymbol{\tau})$ the result of applying the absolute value transformation on a multivariate t-distribution, i.e. $\boldsymbol{\tau} = |\boldsymbol{\tau}^*|$ with $p(\boldsymbol{\tau}^*) = \int p(\boldsymbol{\tau}^*|\xi)p(\xi)\mathrm{d}\xi$ where

$$p(\boldsymbol{\tau}^*|\xi) \sim \mathrm{MVN}\left(\mathbf{0}, \frac{2b_0}{c_0}\xi \mathbf{I}_m\right)$$

and $p(\xi) \sim$ Inverse Gamma (b_0, b_0). We use the fact that a random vector distributed according to a multivariate t-distribution can be represented as an inverse gamma mixture of a multivariate normal distribution.

As the variance covariance matrix of $\boldsymbol{\tau}^*$ is diagonal $p(\boldsymbol{\tau}^*) = \int \Pi_{d=1}^{m} p(\tau_d^*|\xi)p(\xi)\mathrm{d}\xi$.

Horrace (2005) studied truncated multivariate normal distributions and showed that univariate marginal distributions from a multivariate half-normal distribution (obtained applying the absolute value transformation on each component) are univariate half-normal distributions if and only the variance–covariance matrix of the parent multivariate normal distribution is diagonal. As a consequence

$$p(\tau_d) = \int \ldots \int \left\{ \int \prod_{d=1}^{m} p(\tau_d|\xi)p(\xi)\mathrm{d}\xi \right\} \mathrm{d}\boldsymbol{\tau}_{-d}$$

where each $p(\tau_d|\xi)$ is distributed as a half-normal distribution.

If we change the order of integration and use conditional independence of $p(\tau_d|\xi)$ we obtain that $p(\tau_d)$ are marginally half-t distributed.

To prove that τ_d are linearly independent of each other we write

$$V(\boldsymbol{\tau}) = E_\xi\{V(\boldsymbol{\tau}|\xi)\} + V_\xi\{E(\boldsymbol{\tau}|\xi)\}$$

and note that $E(\boldsymbol{\tau}|\xi) = \mathbf{0}$ whereas

$$E_\xi\{V(\boldsymbol{\tau}|\xi)\} = E(\xi)\frac{2b_0}{c_0}\mathbf{I}_m,$$

which is of course diagonal.

References

Barndorff-Nielsen, O. E. (1977) Exponentially decreasing distributions for the logarithm of particle size. *Proc. R. Soc.* A, **353**, 401–419.

Breinlich, H., Ottaviano, G. I. P. and Temple, J. R. W. (2014) Regional growth and regional decline. In *Handbook of Economic Growth*, 1st edn, vol. 2, ch. 4, pp. 683–779. Amsterdam: Elsevier.

Burgard, J. P., Munnich R. and Zimmermann, T. (2014) The impact of sampling designs on small area estimates for business data. *J. Off. Statist.*, **30**, 749–771.

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

Cobb, B. R., Rumì, R. and Salmeron, A. (2012) Approximating the distribution of a sum of log-normal random variables. In Proc. 6th *Eur. Wrkshp Probabilistic Graphical Models, Granada.*

Datta, G. S., Hall, P. and Mandal, A. (2011) Model selection by testing for the presence of small-area effects in area-level data. *J. Am. Statist. Ass.*, **106**, 362–374.

Datta, G. S. and Lahiri, P. (1995) Robust hierarchical Bayes estimation of small area characteristics in presence of covariates and outliers. *J. Multiv. Anal.*, **54**, 310–328.

Datta, G. S. and Mandal, A. (2015) Small area estimation with uncertain random effects. *J. Am. Statist. Ass.*, to be published.

Eurostat (2011) Key figures on European business—with a special feature on SMEs. *Eurostat Pocketbook*. Luxembourg: Eurostat.

Eurostat (2015a) *Eurostat Regional Yearbook 2015: General and Regional Statistics*. Luxembourg: Eurostat.

Eurostat (2015b) *Regions in the European Union—Nomenclature of Territorial Units for Statistics—NUTS 2013/EU-28*. Luxembourg: Eurostat.
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. *Computnl. Statist. Data Anal.*, **55**, 1736–1747.
Fabrizi, E. and Trivisano, C. (2016) Small area estimation of the Gini concentration coefficient. *Computnl Statist. Data Anal.*, **99**, 223–234.
Fay, R. and Herriot, R. (1979) Estimates of income for small places: an application of James–Stein procedures to census data. *J. Am. Statist. Ass.*, **74**, 269–277.
Fenton, L. (1960) The sum of log-normal probability distributions in scatter transmission systems. *IRE Trans. Commun. Syst.*, **8**, 57–67.
Ferrante, M. R. and Trivisano, C. (2010) Small area estimation of the number of firms' recruits by using multivariate models for count data. *Surv. Methodol.*, **36**, 171–180.
Fruhwirth-Schnatter, S. and Wagner, H. (2010) Bayesian variable selection for random intercept modelling of Gaussian and non-Gaussian data. In *Bayesian Statistics 9* (eds J. Bernardo, M. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), pp. 165–200. Oxford: Oxford University Press.
Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models. *Baysn Anal.*, **1**, 515–533.
Gelman, A., Meng, X. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statist. Sin.*, **6**, 733–807.
Griffin, J. E. and Brown, P. J. (2010) Inference with normal-gamma prior distributions in regression problems. *Baysn Anal.*, **5**, 171–188.
Horrace, W. C. (2005) Some results on the multivariate truncated normal distribution. *J. Multiv. Statist.*, **94**, 209–221.
Ibrahim, J., Chen, M. and Sinha, D. (2001) *Bayesian Survival Analysis*. New York: Springer.
Jiang, J. and Lahiri, P. (2005) Mixed model prediction and small area estimation (with discussion). *Test*, **15**, 1–96.
Karlberg, F. (2000) Survey estimation for highly skewed populations in the presence of zeroes. *J. Off. Statist.*, **16**, 229–241.
Mazmanyan, L., Ohanyan, V. and Trietsch, D. (2009) The lognormal central limit theorem for positive random variables. In *Principles of Sequencing and Scheduling* (eds. K. R. Baker and D. Trietsch). New York: Wiley.
Militino, A. F., Ugarte, M. D. and Goicoa, T. (2015) Deriving small area estimates from information technology business surveys. *J. R. Statist. Soc.* A, **178**, 1051–1067.
Organisation for Economic Co-operation and Development (2013) *OECD Regions at a Glance 2013*. Geneva: Organisation for Economic Co-operation and Development Publishing.
Park, and Casella, G. (2008) The Bayesian lasso. *J. Am. Statist. Ass.*, **103**, 681–686.
Pratesi, M. (ed.) (2016) *Analysis of Poverty Data by Small Area Estimation*. New York: Wiley.
Pfeffermann, D. (2014) Small area estimation. In *International Encyclopedia of Statistical Science* (ed. M. Lovric), pp. 1346–1349. New York: Springer.
Rao, J. N. K. (2003) *Small Area Estimation*. New York: Wiley.
Rivière, P. (2002) What makes business statistics special? *Int. Statist. Rev.*, **70**, 145–159.
Slud, E. V. and Maiti, T. (2006) Mean-squared error estimation in transformed Fay–Herriot models. *J. R. Statist. Soc.* B, **68**, 239–257.
Statistics Canada (2007) *2005 Survey of Financial Security: Public Use Microdata File User Guide*. Ottawa: Statistics Canada.
Wolter, K. M. (1985) *Introduction to Variance Estimator*. New York: Springer.
You, Y. and Rao, J. N. K. (2002) Small area estimation using unmatched sampling and linking models. *Can. J. Statist.*, **30**, 3–15.