Bayesian small area estimation for skewed business survey variables

Journal:	*Journal of the Royal Statistical Society*
Manuscript ID	JRSS-OA-SC-May-17-0010.R1
Manuscript Type:	Original Article - Series C
Date Submitted by the Author:	n/a
Complete List of Authors:	Fabrizi, Enrico; Università Cattolica del S. Cuore, DISES ferrante, Maria; University of Bologna, of Statistical Sciences Trivisano, Carlo; University of Bologna,
Keywords:	Log-normal distribution, Variance-Gamma distribution, robust estimation, local shrinkage priors, regional studies

Bayesian small area estimation for skewed business survey variables

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

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

Abstract

Regional and sectoral economic decision and policy rely on accurate business statistics on sub-national regions and business classes. Unfortunately, estimates based on business sample surveys can be imprecise due to the small sample sizes available for sub-populations. When this is the case, small area estimation methods can be helpful. 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 use a prior distribution for the random effects based on the idea of local shrinkage, which is particularly suitable when auxiliary variables with strong predictive power are available, another feature often displayed by business survey data. The proposed method is easily implemented using MCMC software. We discuss an application based on data from the Italian survey on Small and Medium Enterprises, where we estimate the total value added for subsets of a firm population obtained by cross-classifying industry, region and firm size. A simulation exercise explores the frequentist properties of the proposed estimator.

Keywords: robust estimation, Log-Normal distribution, local shrinkage priors, regional studies, Variance Gamma distribution.

1. Introduction

Regional economic decisions and policies rely on accurate business information regarding sub-national regions and business categories. The relevance of regional estimates of business aggregates and the interest in regional disparities in terms firm competitiveness and productivity is demonstrated by the growing number of scientific articles 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 coherently allocate funds (for examples of this, see OECD, 2013, Eurostat, 2011, Eurostat, 2015a). For instance, the (gross) value added, that is, the total value of new goods produced and services provided in a given time period, is routinely estimated at the national and sub-national levels. For the EU, Eurostat releases regional estimates of the value added at levels as detailed as the EU NUTS 3 regions (following the Nomenclature of Territorial Units for Statistics, Eurostat, 2015b) and industries (NACE Rev. 2, 1 digit, following the Statistical classification of economic activities in the European Community). Sub-national estimates of value added would be even more informative if they were disaggregated both in terms of industry and firm size for the purpose of 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 the poverty mapping (see Pfeffermann, 2014 and Pratesi, 2016 for a review) and few applications for business statistics. In the last few 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 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 models straightforwardly incorporate information on the sampling design and non-response re-weighting adjustments, leading to design-consistent estimators whenever direct estimators are design-consistent (Rao, 2003, p. 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 assumed model 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 in order 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. Literature on area level modelling on the log scale include Fay and Herriot (1979) and Slud and Maiti (2006) that 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 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 the 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 characterize 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).

This paper contributes 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 lead to predictions close to those we can obtain using standard priors for non-outlying areas, and to less biased predictors for the areas that can be labelled as outliers.

The strategy we propose may be applied to estimating business totals based on any positively skewed variables: value added, turnover, labor 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 sub-populations of Italian small- and medium-sized manufacturing firms classified according to sub-national region, industry and firm-size classes. We limit our attention to small and medium enterprises, that is, on firms with less than 100 employees because in general, as well as in Italy, larger firms are censused, and small area estimation is therefore not needed. We use data on the Small and Medium Enterprises sample survey (1-99 employees) conducted by the Italian National 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 EU Commission within the 7th Framework Programme.

The paper is organized as follows. Model specification is described in section 2. Specifically in section 2.1 closed formulas 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 motivate the recourse to 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 ($d = 1,...,m$; $N = \sum_{d=1}^{m} N_d$). A random sample of overall size n is taken 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 allow for reliable inference for most of the domains. We assume that individual weights w_{dj} $j = 1,...,n_d$ 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 $\hat{Y}_d = \sum_{j=1}^{n_d} w_{dj} y_{dj}$. 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:

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

where $[\bullet]$ is used to denote a parametrization in terms of mean and variance of the distribution. Exact or approximate design-unbiasedeness of totals' estimators is typical in survey sampling. The distributional assumption in [1] can be motivated directly assuming the log-normality of Y. Log-normal approximations of sums of independent log-normals are justified by several authors (e.g., Fenton, 1960; 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 is common in the small area literature (as in Fay and Herriot, 1979).

On the log-scale, a specification consistent with the *sampling model* [1] is given by:

$$\log\left(\hat{Y}_d\right) \mid \eta_d, \delta_d \sim N\left(\eta_d - \delta_d/2, \delta_d\right) \qquad [2]$$

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

In the small area literature, variances associated with direct estimators are usually treated as known constants. In practice, estimates obtained with methods such as linearization or bootstrap are smoothed using a model involving unknown parameters. In line with the literature on area-level models, we will 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 in order to improve the direct estimators:

$$\theta_d = \exp\left(\mathbf{x}_d^t \boldsymbol{\beta} + u_d\right) \qquad [3]$$

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

$$\theta_d \sim LN\left(\mathbf{x}_d^t \boldsymbol{\beta}, \psi_d\right) \qquad [4]$$

or, equivalently, $\eta_d \sim N\left(\mathbf{x}_d^t \boldsymbol{\beta}, \psi_d\right)$. We denote the model defined by *sampling model* [1] and *linking* [4] as the LN-LN model.

2.1 Analysis conditional on the variance components

To analyze the model defined by [1] and [4], note first that, assuming δ_d as known ($\delta_d = \hat{\delta}_d$) we can re-write [2] as $\hat{Z}_d \sim N\left(\eta_d, \hat{\delta}_d\right)$, 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:

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

where $\hat{\eta}_d^{B1} = \mathbf{x}_d^t \boldsymbol{\beta} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t \boldsymbol{\beta}\right)$, $g_{1,\mathrm{d}} = \gamma_d \hat{\delta}_d$ and $\gamma_d = \dfrac{\psi_d}{\psi_d + \hat{\delta}_d}$. 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 ($\hat{Z}_d$) and a synthetic component ($\mathbf{x}_d^t \boldsymbol{\beta}$), known as the linear composite estimator in the small area literature. If we assume quadratic loss and define $\hat{\theta}_d^{B1} = E\left(\theta_d \mid \boldsymbol{\beta}, \psi_d, data\right)$ as the point predictor for θ_d, we have that

$$\begin{aligned}\hat{\theta}_d^{B1} &= \exp\left\{\mathbf{x}_d^t \boldsymbol{\beta} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t \boldsymbol{\beta}\right) + \frac{1}{2}\gamma_d \hat{\delta}_d\right\} \\ &= \exp\left\{\mathbf{x}_d^t \boldsymbol{\beta} + \gamma_d\left(\log \hat{Y}_d - \mathbf{x}_d^t \boldsymbol{\beta}\right) + \gamma_d \hat{\delta}_d\right\}\end{aligned} \qquad [5]$$

This predictor is the product between $\exp\left(\hat{\eta}_d^{B1}\right)$ 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

$$E\left(\hat{\theta}_d^{B1}\right) = \exp\left\{\mathbf{x}_d^t \boldsymbol{\beta} + \frac{1}{2}\psi_d\right\} = E_M\left(\theta_d\right) \qquad [6]$$

where $E\left(\hat{\theta}_d^{B1}\right)$ is the expectation taken with respect to both linking and sampling models, while with $E_M(\bullet)$, we denote the expectation with respect to linking model [4]. This latter result means that $\hat{\theta}_d^{B1}$ is an unbiased predictor of θ_d in the same sense that BLUP 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 [6] can be found in Appendix 1.

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

$$\eta_d \mid \psi_d, data \sim N\left(\hat{\eta}_d^{B2}, g_{1,\mathrm{d}} + g_{2,\mathrm{d}}\right)$$

where $\hat{\eta}_d^{B2} = \mathbf{x}_d^t \hat{\boldsymbol{\beta}}_{gls} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t \hat{\boldsymbol{\beta}}_{gls}\right)$, $\hat{\boldsymbol{\beta}}_{gls} = \left(\sum_d \dfrac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d \mathbf{x}_d^t\right)^{-1} \sum_d \dfrac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d^t \log \hat{Y}_d$, $g_{2,\mathrm{d}} = \left(1-\gamma_d\right)^2 \mathbf{x}_d^t \left(\sum_d \dfrac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d \mathbf{x}_d^t\right)^{-1} \mathbf{x}_d$. As a consequence, the point predictor under quadratic loss will be given by

$$\hat{\theta}_d^{B2} = \exp\left\{ \mathbf{x}_d^t \hat{\boldsymbol{\beta}}_{gls} + \gamma_d \left(\hat{Z}_d - \mathbf{x}_d^t \hat{\boldsymbol{\beta}}_{gls} \right) + \frac{1}{2}\left(g_{1,\mathrm{d}} + g_{2,\mathrm{d}} \right) \right\} \qquad [7]$$

(see Appendix 1 for a proof). Unlike the empirical Bayes approach advocated by Slud and Maiti (2006), who plug estimates of unknown parameters into [7], a full Bayes approach accounts for the effect that the extra-variation 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 variance on the log scale.

To fully account 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 Markov Chain Monte Carlo considered in this paper.

2.2 The distribution for the random effects and specification of hyperpriors

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

Datta et al. (2011) note 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 analyzing business data, it is quite likely that auxiliary variables with strong predictive power are available. When this is the case, 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:

$$u_d | \psi_d \overset{ind}{\sim} N(0, \psi_d) \quad [8]$$

$$\psi_d | a, \lambda \overset{ind}{\sim} Gamma(a, \lambda) \quad [9]$$

$$\lambda | b_0, c_0 \sim Gamma(b_0, c_0) \quad [10]$$

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) observe 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 behavior 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* 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.,

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

(see Bibby and Sørensen, 2003 for more details on this distribution). This marginal prior distribution is symmetric and has expectation $E(u_d) = 0$ and variance $V(u_d) = a / \lambda$. It belongs to the family of generalized hyperbolic distribution (Barndorff-Nielsen, 1977). The conjugate hierarchy in [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, while we specify a prior on the global parameter λ. As far as *a* is concerned, we focus on two choices, $a = 1$, $a = 0.5$.

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

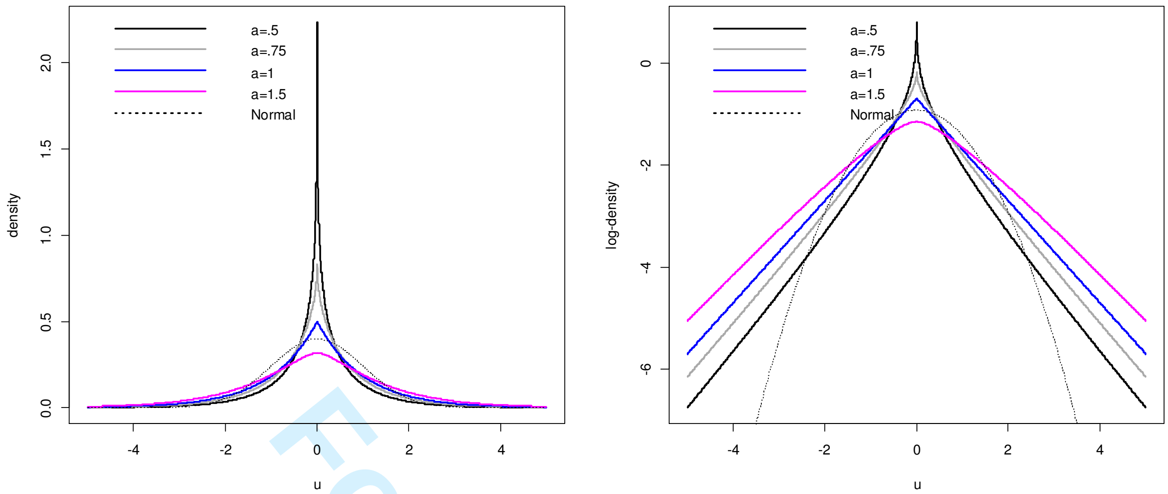

Figure 1: *density (left) and log-density (right) functions of the Variance Gamma distribution.* $VG\left(a,\sqrt{2\lambda},0,0\right)$ *for* $\lambda=1$ *and various values of a.*

The case $a = 0.5$ represents a more peaked prior distribution and encourages more shrinkage towards 0 of small random intercepts (see Figure 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 sizeable mass of prior probability close to 0. It can be shown that once $\boldsymbol{\psi}=\{\psi_d\}$ and $\boldsymbol{\tau}=\left\{\sqrt{\psi_d}\right\}$, $d=1,\ldots,m$, are defined, under prior [8]-[10] and $a = 0.5$,

$$\boldsymbol{\tau}\mid b_0,c_0 \sim Mht\left(0,\frac{2b_0}{c_0}\mathbf{I},2b_0\right). \qquad [11]$$

With $Mht(\bullet)$ 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 $\sqrt{\psi_d}\mid b_0,c_0$ is a (univariate) half-*t* and the priors for two different variance components are uncorrelated, i.e. $Cov\left(\psi_d,\psi_{d^*}\mid b_0,c_0\right)=0$ whenever $d\neq d^*$. See appendix 1 for a proof of [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 $E\left(\lambda^{-1}\right)=c_0$. This helps to interpret c_0 as a scaling constant for the random effects variance $V\left(u_d\right)=a/\lambda$. The choice of c_0 depends on the scale for the random effects in the problem being considered. According to

[11], the parameter b_0 can be interpreted in terms of degrees of freedom of the marginal prior $p(\boldsymbol{\tau})$, 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. An Application for Italian Small and Medium Enterprise Survey Data

In this section, we illustrate the methodology we discussed using real survey data. We use data on the Small and Medium Enterprises (SME) sample survey, wave 2008, conducted by the Italian National Statistical Institute (ISTAT). Specifically, we target the estimation of the total value added (VA) for small domains of the population of Italian small and medium manufacturing firms (less than 100 employees). The domains we focus on are smaller than those for which ISTAT provides reliable estimates. Specifically, our domains are defined by cross-classifying: the 20 Italian NUTS 2 administrative regions, the economic industrial sector (NACE Rev. 2, 2 digit, 22 industries) and firm size (4 classes: less than 10 employees, 10 to 19 employees, 20 to 49 employees, 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.

3.1 Direct estimators and sampling model

The SME survey sampling design is stratified and strata are defined by cross-classifying NACE Rev. 2 (4 digits) Italian administrative regions and company size in the four classes defined above. The domains 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 ($r = 1,\ldots,20$). 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 characterized by a sampling rate over 0.75.

The actual sample sizes for the domains we consider ranges from 2 to 184, with a median of 8, a mean of 13.5 and .75 and .9 quantiles equal to 16 and 30 respectively.

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 (the inverse of the inclusion probabilities) by factors adjusting for non-response and calibrating to known totals. Let the estimated total be denoted as $\hat{Y}_{ijr} = \sum_{k \in d_{ijr}} w_{ijr,k}\, y_{ijr,k}$, where $y_{ijr,k}$ is the value added of the *k*-th firm in sector *i*, size class *j*, region *r*.

We assume that $E\left(\hat{Y}_{ijr}\right)=\theta_{ijr}$ with $Var\left(\hat{Y}_{ijr}\right)=V_{ijr}$. We estimate V_{ijr} using linearization-based variance estimators and denote these estimates as $\hat{V}_{ijr}$.

In our sampling model we assume log-normality according to [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 we have in our analysis.

For all the domains with $n_{ijr} \geq 3$, we tested normality and log-normality using the Shapiro-Wilk test. Results are reported in table 1 below. In reporting the results we consider separately the smallest 90% of the domains $n_{ijr} \leq 30$ 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, while log-normality is accepted in more than 70% of the cases.

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

n_{ijr}	normality	log-normality
≤ 30	0.733	0.959
> 30	0.087	0.713
overall	0.672	0.943

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 illustrated in previous section. To check this, we consider a set of log-normal populations: $Y_i \sim LN\left(\tilde{\mu}_i, \tilde{\sigma}_i\right)$, $i=1,\ldots,L$ where $L=77$ is the number of domains with $n_{ijr} > 30$ for which log-normality was not rejected and $\tilde{\mu}_i, \tilde{\sigma}_i$ 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: $n_i = 5,10,15,20$. 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 presence of negligible departures from the null. 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.

Table 2: *Kolmogorov-Smirnov distances between the Monte Carlo ($R = 10,000$) distribution of the sample mean and two reference distributions, for various sample sizes.*

	sample size (n_i)			
Reference distribution	5	10	15	20
Log-normal	0.012	0.014	0.015	0.015
Normal	0.119	0.097	0.081	0.073

In order 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 introduced in [1], we have that

$$Var\left\{\log\left(\hat{Y}_{ijr}\right)\right\} = \log\left\{CV^2\left(\hat{Y}_{ijr}\right) + 1\right\}. \qquad [12]$$

Thus, the smoothing can be conducted on $CV^2\left(\hat{Y}_{ijr}\right) = \frac{\hat{V}_{ijr}}{\hat{Y}_{ijr}^2}$. After careful explorative analysis, we assume that $CV^2\left(\hat{Y}_{ijr}\right)$ 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}, $\hat{V}_{ijr}$:

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

with $E\left(\upsilon_{ijr}\right) = 0$, $V\left(\upsilon_{ijr}\right) = \kappa$ 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 for the size of the domain n_{ijr}. The domain sample size n_{ijr} in the denominator of [13] allows for the decrease in the coefficient of variation when the sample size increases. Smoothed squared estimated coefficient of variations are given by $CV_{smooth}^2\left(\hat{Y}_{ijr,k}\right) = \frac{\phi_j}{n_{ijr}}\left(1 - \frac{n_{ijr}}{N_{ijr}}\right)$; the first, second and third quartiles of

$CV^2_{smooth}\left(\hat{Y}_{ijh,k}\right)$ 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 [1] to our problem as

$$\log \hat{Y}_{ijr} \left| \eta_{ijr}, \hat{\delta}_{ijr} \sim N\left(\eta_{ijr} - \frac{1}{2}\hat{\delta}_{ijr}, \hat{\delta}_{ijr}\right)\right. \qquad [14]$$

where $\eta_{ijr} = \log\left(\theta_{ijr}\right)$ and $\hat{\delta}_{ijr} = \log\left\{CV^2_{smooth}\left(\hat{Y}_{ijr}\right)+1\right\}$ according to [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-ASIA 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

$$\eta_{ijr} \sim N\left(\beta_0 + x_{ijr}\beta_1, \psi_{ijr}\right) \qquad [15]$$

where x_{ijr} is the log-total turnover for the domain in question. Equivalently, we can write $\eta_{ijr} = \beta_0 + ltt_{ijr}\beta_1 + u_{ijr}$. The prior for the vector of domain-specific random intercepts u_{ijr} is specified according to [8]-[10]. As for the prior specifications not already discussed, we set $\beta_0 \sim N\left(0,10^5\right)$, $\beta_1 \sim N\left(0,10^5\right)$, $b_0 = 2$, $c_0 = 1$. We chose these values as they provide a reasonable scale for the random effects variance in our problem.

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

$$u_{ijr} \left| \sigma^2 \sim N\left(0,\sigma^2\right),\ \sigma^2 \sim InverseGamma\left(c,d\right)\right. \qquad [16]$$

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

3.3 Markov Chain Monte Carlo computational issues

Parameter estimates are obtained by summarizing the posterior distributions approximated by the output of Markov Chain Monte Carlo (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 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 autocorrelation diagrams and by means of the potential scale reduction known as the Gelman-Rubin statistic (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 can be downloaded for free on the internet and it is open source.

3.4 Comparing alternative models

In order to choose among competing models, we compute the Deviance Information Criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPLM, 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 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 using LPLM is consistent with that obtained using the DIC. The adoption of the Variance Gamma for the random intercepts u_{ijr} leads to a further reduction in DIC with respect to the more common specification [16].

Table 3: *Comparison of alternative assumptions on the distributions of the random effects*

Shrinkage	Prior on random intercepts u_{ijr}	a	DIC	LPLM	median CVR
global	[16]	-	15340	-7846	0.391
local	[10]-[12]	1	15230	-7808	0.421
local	[10]-[12]	0.5	15220	-7798	0.455

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 $CVR_k^h = 1 - CV_k^h / CV_k^{DIR}$, where the 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, while CV_k^{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 suggested in the context of small area estimation byYou and Rao (2002) and considered also in Fabrizi and Trivisano (2016):

$$\text{dis}_{ijr} = P\left(\hat{Y}_{ijr} < Y_{ijr}^*\right)$$

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 centered around the direct estimate. Values of dis_{ijr} far from 0 and 1 would provide evidence of systematic under or over estimation. For the log-normal model endowed with priors [10]-[12] and $a = 0.5$ (i.e. the best model in terms of DIC and LPLM), 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 considered in table 3.

Results on $median(CVR_k^h)$, reported in table 3, highlight that the whole set of considered small area estimators 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 impact. Prior specifications mimicking the spike and slab behavior allow for a further gain in efficiency with respect to priors ordinarily used in this type of analysis.

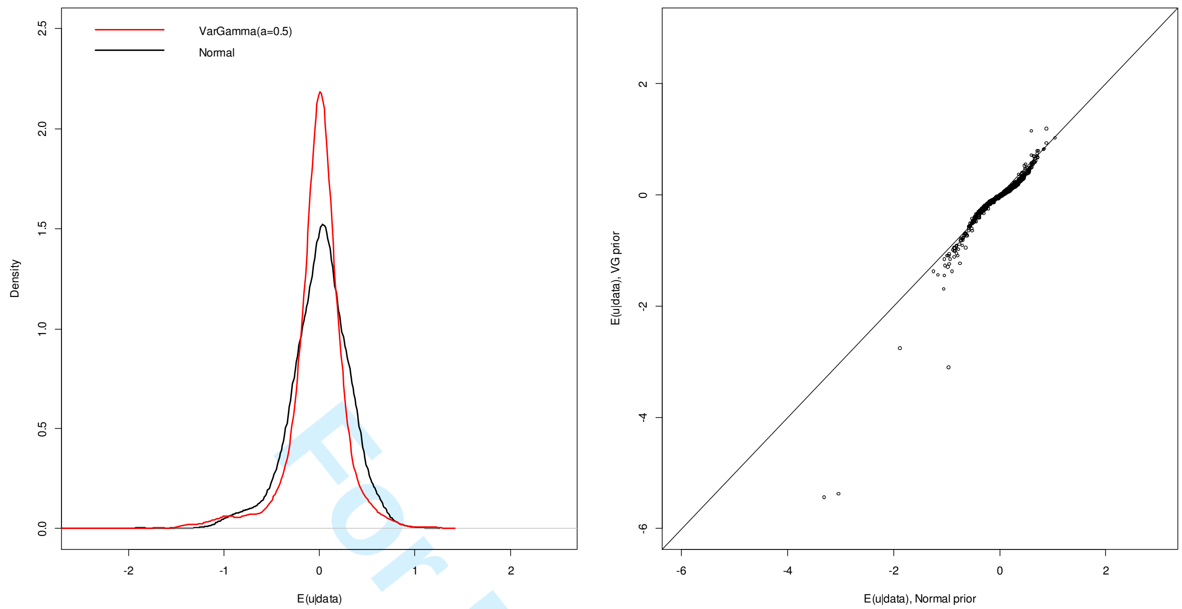

Figure 2*: Left panel: kernel densities of* $E\left(u_{ijr} \mid data\right)$ *over the set of small areas under Normal and VG priors. Right panel:* $E\left(u_{ijr} \mid data\right)$ *under Normal prior vs* $E\left(u_{ijr} \mid data\right)$ *under VG prior. For VG priors,* $a = 0.5$ *is assumed.*

To evaluate the improvements allowed 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 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, while for the model based ones this number grows to 70%. Although the uncertainty around the small area estimates remain sizeable and not all estimates would be publishable, the application of the proposed method endows most sub-population with a publishable estimate in spite of the small sample sizes.

Figure 2 displays the impact of alternative prior specifications on the ensemble of the random intercepts' posterior means $E\left(u_{ijr} \mid data\right)$ under the Normal prior [16] and Variance Gamma priors $u_{ijr} \sim VG\left(0.5, \sqrt{2\lambda}, 0, 0\right)$. Based on the left panel, it is clear that the Variance Gamma leads to a more peaked distribution of estimated random intercepts, as predicted by theory. As tail behavior is difficult to read from density estimates, in the right panel of Figure 2, we plotted the point posterior expectation under the Normal prior vs those obtained under the Variance Gamma. The peak around 0 is still apparent from the inflection of the points' cloud approximately at 0; heavier tails under Variance Gamma can also be appreciated: under the Normal prior, $E\left(u_{ijr} \mid data\right)$ lies within the interval $(-4, 2)$, while they do not under the Variance Gamma specification.

As the purpose of small area estimators is to complement the direct estimates obtained using ordinary survey weighted methods, robustness with respect to modelling assumptions is a major concern. As anticipated, 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) introduced in section 4. 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 discussed in previous section. We drop from the 2007 wave those firms sampled also in the 2008. We obtain a population of size $N = 30451$. Domains are defined by cross-classifying the population by firm size and industry sector; with respect to the data analysis of previous section we collapse the regions. By reducing the number of domains we create sub-populations large enough to be sampled using reasonable sampling rates. Collapsing by region has a milder impact on sub-population skewness with respect to firm size or industry sector. Thereby our synthetic population is divided into $m = 88$ domains whose sizes N_d ($d = 1,\dots,m$) range from 14 to 1339 with an average of 346. The same target parameters (total VA) and auxiliary variable (turnover) studied in section 4 are considered.

We keep this synthetic population as fixed and we repeatedly 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 using analytic formulas. The point estimators we compare in the simulation are:

i) the posterior mean associated to the Fay-Herriot model specified on the un-transformed scale (UFH). The unstrasformed Fay-Herriot model can be described as follows: $\hat{Y}_d \sim N(\theta_d, V_d)$, where V_d is the variance of the direct estimator $\hat{Y}_d$, $\theta_d \sim N(\beta_0+\beta_1 x_d, \sigma_v^2)$, $\sigma_v \sim Unif(0,A)$, $A=1000$, $\beta_i \sim N(0,10^5)$ $i=0,1$. We consider the UFH model as it is probably the most "basic" Bayesian model a practitioner would think of for analyzing these data.

ii) the predictor proposed by Slud and Maiti (2006), we denote as SM;

iii) the posterior mean obtained from the log-normal model [1]-[3] endowed with the global shrinkage prior [16]. We denote the estimator as LNGS ("Log-Normal with Global Shrinkage")

iv) the posterior mean associated to the model [1]-[3] but endowed with local shrinkage prior described in [8]-[10]; we label the estimator as LNLS ("Log-Normal with Local Shrinkage").

Let's denote with est_{ds} the generic estimator calculated for the domain d in the 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 (w.r.t. the coverage probability $1-\alpha$) and set $\alpha=0.05$. Comparison tools are defined as follows:

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

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

$$COV95=\frac{1}{m}\sum_{d=1}^{m}\frac{1}{S}\sum_{s=1}^{S}1\left\{Y_d\in\left(p_{\theta_{ds}|data}(0.025),p_{\theta_{ds}|data}(0.975)\right)\right\}$$

As we are interested in the frequentist coverage of Bayes estimators, $COV95$ is calculated only for *i*, *iii*, *iv*. Results are summarized in table 4.

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

est	$RABIAS$	$RRMSE$	$COV95$
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

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 parametrized model. Actually, they are based on similar ideas and models, only the priors or the way hyper-parameters 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 you compromise between the efficiency of a biased synthetic predictor and the unbiasedeness of large variance direct estimators; to some extent estimators associated to areas with very small area-specific sample sizes are naturally biased.

The naïve Fay-Herriot model, specified on the un-transformed 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 performances reflect 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 already noted that this assumption is not completely unrealistic, so misspecification of the sampling model can be held as 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, that is those characterized by a deviation from the synthetic component much larger than most of the remaining areas. We investigate performances separataly for the areas 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 $n_d = 34$, well above the average sample size of the simulation.

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

est	*RABIAS*	*RRMSE*	*COV*95
SM	0.4763	0.4790	--
LNGS	0.4414	0.4453	0.505
LNLS	0.0144	0.2515	0.947

The LNGS and the SM predictor are based on a similar global shrinkage idea. Results in table 5 show how the common variance parameter 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. Local shrinkage prior associated to 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, a feature often displayed in business survey data.

The proposed methodology can be easily implemented using widely available MCMC software. Openbugs codes, as well as formulas for the full conditional distributions needed for an independent implementation of the algorithm, are available upon request from the authors.

In summary, we showed that the predictor based on local shrinkage prior has overall acceptable frequentist properties, comparable to the alternatives we consider in the exercise. We introduce this prior specification to deal with situations where powerful auxiliary information is available, most of the areas are well fit by the assumed model and only a minority are outlying, characterized by larger model residuals. For these areas, local shrinkage priors can lead to estimators with reduced bias and thereby more efficient.

References

Barndorff-Nielsen, O.E. (1977). Exponentially decreasing distributions for the logarithm of particle size, Proceedings of the Royal Statistical Society, series A, 353, 401-419.

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.

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.

Datta, G. S., Hall, P. and Mandal, A. (2011). Model selection by testing for the presence of small-area effects in area-level data. Journal of the American Statistical Association, 106, 362-374.

Datta, G.S., Mandal A. (2015) Small Area Estimation with Uncertain Random Effects, Journal of the American Statistical Association, in press.

Eurostat (2011), Key figures on European business - with a special feature on SMEs, Eurostat Pocketbooks.

Eurostat (2015a). Eurostat regional yearbook 2015. Statistical books, General and regional statistics.

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

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

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

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.

Fenton, L. (1960) The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems. *IRE Transactions on Communication Systems* 8(1), 57-67.

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.

Gelman A., Meng X., Stern H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6, 733-807.

Gelman A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, 3, 515–533.

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

Horrace W.C. (2005) Some results on the multivariate truncated normal distribution, Journal of Multivariate Statistics, 94, 209-221.

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

Jiang J. & Lahiri P. (2005) Mixed model prediction and small area estimation (with discussion), TEST, 15, 1-96

Mazmanyan, L., Ohanyan, V., Trietsch, D. (2009) The lognormal central limit theorem for positive random variables. Reproduced as Appendix in Baker K.R, Trietsch D. (2009), Principles of sequencing and scheduling, John Wiley and Sons, New York.

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, 178, 1051-1067.

Park, Casella G. (2008). The Bayesian lasso, Journal of the American Statistical Association, 103, 681–686.

Pratesi M. ed. (2016). Analysis of Poverty Data by Small Area Estimation, John Wiley, New York.

OECD (2013), OECD Regions at a Glance 2013: OECD Publishing. http://dx.doi.org/10.1787/reg_glance-2013-en

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

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

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

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.

Statistics Canada (2007) 2005 Survey of Financial Security. Public Use Microdata File User Guide. Ottawa: Statistics Canada. (Available from http://www.statcan.gc.ca/pub/13f0026m/13f0026m2007001-eng.htm.)

Wolter K.M. (1985). Introduction to variance estimator, New York, Springer-Verlag.

You, Y. & Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. Canadian Journal of Statistics, 30, 3–15.

Appendix 1

Proof of [6]

To start with we note that

$$\hat{\theta}_d^{B1} = \exp\left\{\mathbf{x}_d^t\boldsymbol{\beta} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t\boldsymbol{\beta}\right) + \frac{1}{2}\gamma_d\hat{\delta}_d\right\} = \exp\left\{\gamma_d\hat{Z}_d + (1-\gamma_d)\mathbf{x}_d^t\boldsymbol{\beta} + \frac{1}{2}\gamma_d\hat{\delta}_d\right\} \text{ so}$$

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

We note that $E\left(\hat{Z}_d\right) = E_M\left\{E\left(\hat{Z}_d \mid \eta_d\right)\right\} = E_M\left\{\eta_d\right\} = \mathbf{x}_d^t\boldsymbol{\beta}$; analogously $V\left(\hat{Z}_d\right) = V_M\left\{E\left(\hat{Z}_d \mid \eta_d\right)\right\} + E_M\left\{V\left(\hat{Z}_d \mid \eta_d\right)\right\} = \psi_d + \hat{\delta}_d$. As a consequence $E\left(\gamma_d\hat{Z}_d\right) = \gamma_d\mathbf{x}_d^t\boldsymbol{\beta}$, $V\left(\gamma_d\hat{Z}_d\right) = \gamma_d^2\left(\psi_d + \hat{\delta}_d\right) = \psi_d\gamma_d$ and $\mathrm{E}\left[\exp\left\{\gamma_d\hat{Z}_d\right\}\right] = \exp\left\{\gamma_d\mathbf{x}_d^t\boldsymbol{\beta} + \frac{1}{2}\psi_d\gamma_d\right\}$. This leads to

$E\left(\hat{\theta}_d^{B1}\right) = \exp\left\{\mathbf{x}_d^t\boldsymbol{\beta} + \frac{1}{2}\gamma_d\left(\psi_d + \hat{\delta}_d\right)\right\} = \exp\left\{\mathbf{x}_d^t\boldsymbol{\beta} + \frac{1}{2}\psi_d\right\}$ that coincides with $E_M\left(\theta_d\right)$.

Proof of [7]

Let's introduce some matrix notation. Let $\mathbf{z} = vec\left(\hat{Z}_d\right)$ is the vector containing, $\boldsymbol{\Psi} = diag\left(\psi_d\right)$, $\boldsymbol{\Delta} = diag\left(\hat{\delta}_d\right)$ the diagonal matrixes containing the variance components; let $\mathbf{X}$ be the matrix with rows $\mathbf{x}_d^t$, $d = 1,\ldots,m$.

Standard Bayesian analysis of normal linear mixed model lead to $\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\psi} \sim N\left(\hat{\boldsymbol{\beta}}_{gls}, \mathbf{V}(\boldsymbol{\Psi})\right)$ where $\hat{\boldsymbol{\beta}}_{gls} = \left[\mathbf{X}^t\left(\boldsymbol{\Psi}+\boldsymbol{\Delta}\right)^{-1}\mathbf{X}\right]^{-1}\mathbf{X}^t\left(\boldsymbol{\Psi}+\boldsymbol{\Delta}\right)^{-1}\mathbf{z}$, $\mathbf{V}(\boldsymbol{\Psi}) = \left[\mathbf{X}^t\left(\boldsymbol{\Psi}+\boldsymbol{\Delta}\right)^{-1}\mathbf{X}\right]^{-1}$. We can calculate $E\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}\right) = E_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}}E\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi}\right)$ where $\boldsymbol{\eta} = vec\left(\eta_d\right)$. $E_{\beta\mid z,\Psi}E\left(\eta \mid z, \beta, \Psi\right) = E_{\beta\mid z,\Psi}\left(X\beta\right) = X\hat{\boldsymbol{\beta}}_{gls}$. Analogously $V\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}\right) = V_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}}E\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi}\right) + E_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}}V\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi}\right)$. If we denote the vector of small area predictors (on the log scale) conditional on $\boldsymbol{\beta}$ as $\boldsymbol{\eta}^{B1} = E\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi}\right) = \boldsymbol{\Gamma}\mathbf{z} + \left(\mathbf{I} - \boldsymbol{\Gamma}\right)\mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\Gamma} = \boldsymbol{\Psi}\left(\boldsymbol{\Psi}+\boldsymbol{\Delta}\right)^{-1}$ we have that $V\left(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}\right) = V_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}}\boldsymbol{\Gamma}\mathbf{z} + \left(\mathbf{I} - \boldsymbol{\Gamma}\right)\mathbf{X}\boldsymbol{\beta} + E_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}}\mathbf{G}_1$ with $\mathbf{G}_1 = \boldsymbol{\Psi}\left(\boldsymbol{\Psi}+\boldsymbol{\Delta}\right)^{-1}\boldsymbol{\Delta}$. Taking expectation with respect to $p\left(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Psi}\right)$ we get

$V(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}) = (\mathbf{I}-\boldsymbol{\Gamma})\mathbf{X}'\boldsymbol{\beta}\mathbf{V}(\boldsymbol{\Psi})\mathbf{X}(\mathbf{I}-\boldsymbol{\Gamma}) + \mathbf{G}_1 = \mathbf{G}_2 + \mathbf{G}_1$. We note that $p(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi})$ is a multivariate normal distribution. If we consider an individual η_d have that $E(\eta_{ijr} \mid \mathbf{z}, \boldsymbol{\Psi}) = \mathbf{x}_d^t \hat{\boldsymbol{\beta}}_{gls}$,

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

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

Proof of [11] and subsequent statements

From [9] we have that $p(\boldsymbol{\psi} \mid \lambda) = \prod_{d=1}^{m} \frac{\lambda^a}{\Gamma(a)} \psi_d^{a-1} e^{-\lambda \sum_{d=1}^{m} \psi_d} \propto \lambda^{ma} e^{-\lambda \sum_{d=1}^{m} \psi_d} \prod_{d=1}^{m} \psi_d^{a-1}$

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

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

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

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

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

$\tau_d > 0$, $\forall d$. 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 is a multivariate t for which we apply the absolute value transformation on each component. We can then write formula [13].

To prove that univariate priors $p(\tau_d)$ let's start from $p(\tau_d)=\int\ldots\int p(\boldsymbol{\tau})\,d\boldsymbol{\tau}_{-d}$. We can represent as $p(\boldsymbol{\tau})$ as the result of applying the absolute value transformation on a multivariate t distribution, i.e $\boldsymbol{\tau}=\left|\boldsymbol{\tau}^*\right|$ with $p(\boldsymbol{\tau}^*)=\int p(\boldsymbol{\tau}^*\mid\xi)p(\xi)\,d\xi$ where $p(\boldsymbol{\tau}^*\mid\xi)\sim MVN\left(\mathbf{0},\frac{2b_0}{c_0}\xi\mathbf{I}_m\right)$ and $p(\xi)\sim IGamma(b_0,b_0)$. 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.

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

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

$$p(\tau_d)=\int\ldots\int\left[\int\prod_{d=1}^{m}p(\tau_d\mid\xi)\,p(\xi)\,d\xi\right]d\boldsymbol{\tau}_{-d}$$

where each $p(\tau_d\mid\xi)$ is distributed as an half-Normal distribution.

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

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

$$V(\boldsymbol{\tau})=E_\xi\left\{V(\boldsymbol{\tau}\mid\xi)\right\}+V_\xi\left\{E(\boldsymbol{\tau}\mid\xi)\right\}$$

and note that $E(\boldsymbol{\tau}\mid\xi)=\mathbf{0}$ while $E_\xi\left\{V(\boldsymbol{\tau}\mid\xi)\right\}=E(\xi)\frac{2b_0}{c_0}\mathbf{I}_m$, which is of course diagonal.

Bayesian small area estimation for skewed business survey variables

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

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

Abstract

In business surveys, estimates of means and totals for sub-national regions, industries and business classes can be too imprecise due to the small sample sizes available for sub-populations. 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 often displayed by business survey data. This flexible modelling of random effects lead to predictions in agreement with those based on global shrinkage for most of the areas, but allows 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 proposed estimators. They are good, and with difference to methods based on global shrinkage remain so also for areas characterized by large model residuals.

Keywords: robust estimation, Log-Normal distribution, local shrinkage priors, regional studies, Variance Gamma distribution.

1. Introduction

Regional economic decisions and policies rely on accurate business information regarding sub-national regions and business categories. The relevance of regional estimates of business aggregates and the interest in regional disparities in terms firm competitiveness and productivity is demonstrated by the growing number of scientific articles 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 coherently allocate funds (for examples of this, see OECD, 2013, Eurostat, 2011, Eurostat, 2015a). For instance, the (gross) value added, that is, the total value of new goods produced and services provided in a given time period, is routinely estimated at the national and sub-national levels. For the EU, Eurostat releases regional estimates of the value added at levels as detailed as the EU NUTS 3 regions (following the Nomenclature of Territorial Units for Statistics, Eurostat, 2015b) and industries (NACE Rev. 2, 1 digit, following the Statistical classification of economic activities in the European Community). Sub-national estimates of value added would be even more informative if they were disaggregated both in terms of industry and firm size for the purpose of 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 the poverty mapping (see Pfeffermann, 2014 and Pratesi, 2016 for a review) and few applications for business statistics. In the last few 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 obtained for each domain are related with 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 micro level. Area level models straightforwardly incorporate information on survey weights, leading to design-consistent estimators whenever direct estimators are design-consistent (Rao, 2003, p. 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 assumed model 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 in order 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; Karlberg, 2000). Literature on area level modelling on the log scale include Fay and Herriot (1979) and Slud and Maiti (2006) that 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 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 the 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 characterize 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 lead to predictions close to those we can obtain 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 formulas 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 Enterprises sample survey (1-99 employees) conducted by the Italian National 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 EU Commission within the 7$^{\text{th}}$ Framework Programme. For these data we motivate the recourse to 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 ($d = 1,...,m$; $N = \sum_{d=1}^{m} N_d$). A random sample of overall size n is taken 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 allow for reliable inference for most of the domains. We assume that individual weights w_{dj} $j = 1,...,n_d$ 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 $\hat{Y}_d = \sum_{j=1}^{n_d} w_{dj} y_{dj}$. 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:

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

where $[\bullet]$ is used to denote a parametrization in terms of mean and variance of the distribution. Exact or approximate design-unbiasedeness of totals' estimators is typical in survey sampling. The distributional assumption in [1] can be motivated directly assuming the log-normality of Y. Log-normal approximations of sums of independent log-normals are justified by several authors (e.g., Fenton, 1960; 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 is common in the small area literature (as in Fay and Herriot, 1979).

On the log-scale, a specification consistent with the *sampling model* [1] is given by:

$$\log\left(\hat{Y}_d \right) | \eta_d, \delta_d \sim N\left(\eta_d - \delta_d / 2, \delta_d \right) \qquad [2]$$

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

In the small area literature, variances associated with direct estimators are usually treated as known constants. In practice, estimates obtained with methods such as linearization or bootstrap are smoothed using a model involving unknown parameters. In line with the literature on area-level models, we will 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 in order to improve the direct estimators:

$$\theta_d = \exp\left(\mathbf{x}_d^t \boldsymbol{\beta} + u_d\right) \qquad [3]$$

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

$$\theta_d \sim LN\left(\mathbf{x}_d^t \boldsymbol{\beta}, \psi_d\right) \qquad [4]$$

or, equivalently, $\eta_d \sim N\left(\mathbf{x}_d^t \boldsymbol{\beta}, \psi_d\right)$. We denote the model defined by *sampling model* [1] and *linking* [4] as the LN-LN model.

2.1 Analysis conditional on the variance components

To analyze the model defined by [1] and [4], note first that, assuming δ_d as known ($\delta_d = \hat{\delta}_d$) we can re-write [2] as $\hat{Z}_d \sim N\left(\eta_d, \hat{\delta}_d\right)$, 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:

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

where $\hat{\eta}_d^{B1} = \mathbf{x}_d^t \boldsymbol{\beta} + \gamma_d \left(\hat{Z}_d - \mathbf{x}_d^t \boldsymbol{\beta}\right)$, $g_{1,\mathrm{d}} = \gamma_d \hat{\delta}_d$ and $\gamma_d = \dfrac{\psi_d}{\psi_d + \hat{\delta}_d}$. 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 ($\hat{Z}_d$) and a synthetic component ($\mathbf{x}_d^t \boldsymbol{\beta}$), known as the linear composite estimator in the small area literature. If we assume quadratic loss and define $\hat{\theta}_d^{B1} = E\left(\theta_d \mid \boldsymbol{\beta}, \psi_d, data\right)$ as the point predictor for θ_d, we have that

$$\begin{aligned}\hat{\theta}_d^{B1} &= \exp\left\{\mathbf{x}_d^t\boldsymbol{\beta} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t\boldsymbol{\beta}\right) + \frac{1}{2}\gamma_d\hat{\delta}_d\right\} \\ &= \exp\left\{\mathbf{x}_d^t\boldsymbol{\beta} + \gamma_d\left(\log\hat{Y}_d - \mathbf{x}_d^t\boldsymbol{\beta}\right) + \gamma_d\hat{\delta}_d\right\}\end{aligned} \qquad [5]$$

This predictor is the product between $\exp\left(\hat{\eta}_d^{B1}\right)$ 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

$$E\left(\hat{\theta}_d^{B1}\right) = \exp\left\{\mathbf{x}_d^t\boldsymbol{\beta} + \frac{1}{2}\psi_d\right\} = E_M\left(\theta_d\right) \qquad [6]$$

where $E\left(\hat{\theta}_d^{B1}\right)$ is the expectation taken with respect to both linking and sampling models, while with $E_M(\bullet)$, we denote the expectation with respect to linking model [4]. This latter result means that $\hat{\theta}_d^{B1}$ is an unbiased predictor of θ_d in the same sense that BLUP 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 [6] can be found in Appendix 1.

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

$$\eta_d \mid \psi_d, data \sim N\left(\hat{\eta}_d^{B2}, g_{1,\mathrm{d}} + g_{2,\mathrm{d}}\right)$$

where $\hat{\eta}_d^{B2} = \mathbf{x}_d^t\hat{\boldsymbol{\beta}}_{gls} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t\hat{\boldsymbol{\beta}}_{gls}\right)$, $\hat{\boldsymbol{\beta}}_{gls} = \left(\sum_d \frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d\mathbf{x}_d^t\right)^{-1}\sum_d \frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d^t\log\hat{Y}_d$, $g_{2,\mathrm{d}} = \left(1-\gamma_d\right)^2\mathbf{x}_d^t\left(\sum_d \frac{1}{\psi_d + \hat{\delta}_d}\mathbf{x}_d\mathbf{x}_d^t\right)^{-1}\mathbf{x}_d$. As a consequence, the point predictor under quadratic loss will be given by

$$\hat{\theta}_d^{B2} = \exp\left\{\mathbf{x}_d^t\hat{\boldsymbol{\beta}}_{gls} + \gamma_d\left(\hat{Z}_d - \mathbf{x}_d^t\hat{\boldsymbol{\beta}}_{gls}\right) + \frac{1}{2}\left(g_{1,\mathrm{d}} + g_{2,\mathrm{d}}\right)\right\} \qquad [7]$$

(see Appendix 1 for a proof). Unlike the empirical Bayes approach advocated by Slud and Maiti (2006), who plug estimates of unknown parameters into [7], a full Bayes approach accounts for the effect that the extra-variation 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 variance on the log scale.

To fully account 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 Markov Chain Monte Carlo considered in this paper.

2.2 The distribution for the random effects and specification of hyperpriors

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

Datta et al. (2011) note 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 analyzing business data, it is quite likely that auxiliary variables with strong predictive power are available. When this is the case, 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 , $d = 1,...,m$ is based on infinite mixtures of normal distributions. Following Griffin and Brown (2010), our specification uses *Gamma* mixtures of normal distributions. Specifically, we assume:

$$u_d \mid \psi_d \overset{ind}{\sim} N\left(0, \psi_d\right) \qquad [8]$$

$$\psi_d \mid a, \lambda \overset{ind}{\sim} Gamma\left(a, \lambda\right) \qquad [9]$$

$$\lambda \mid b_0, c_0 \sim Gamma\left(b_0, c_0\right) \qquad [10]$$

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) observe 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 behavior 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* 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.,

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

(see Bibby and Sørensen, 2003 for more details on this distribution). This marginal prior distribution is symmetric and has expectation $E\left(u_d\right) = 0$ and variance $V\left(u_d\right) = a / \lambda$. It belongs to the family of generalized hyperbolic distribution (Barndorff-Nielsen, 1977). The conjugate hierarchy in [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, while we specify a prior on the global parameter λ. As far as *a* is concerned, we focus on two choices, $a = 1$, $a = 0.5$.

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

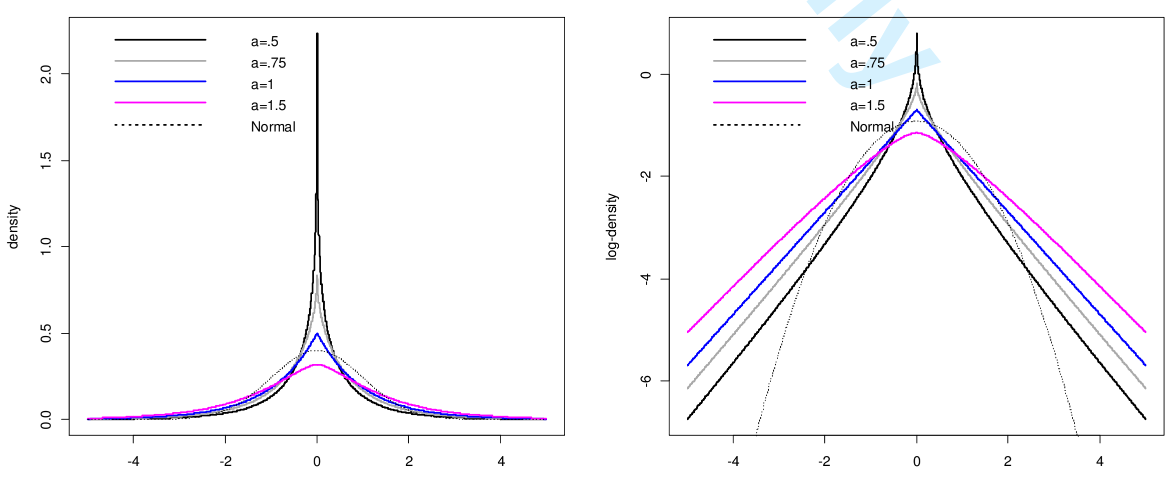

Figure 1: *density (left) and log-density (right) functions of the Variance Gamma distribution.* $VG\left(a,\sqrt{2\lambda},0,0\right)$ *for* $\lambda=1$ *and various values of a.*

The case $a = 0.5$ represents a more peaked prior distribution and encourages more shrinkage towards 0 of small random intercepts (see Figure 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 sizeable mass of prior probability close to 0. It can be shown that once $\boldsymbol{\psi}=\{\psi_d\}$ and $\boldsymbol{\tau}=\left\{\sqrt{\psi_d}\right\}$, $d=1,\ldots,m$, are defined, under prior [8]-[10] and $a = 0.5$,

$$\boldsymbol{\tau}\mid b_0,c_0 \sim Mht\left(0,\frac{2b_0}{c_0}\mathbf{I},2b_0\right). \qquad [11]$$

With $Mht(\bullet)$ 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 $\sqrt{\psi_d}\mid b_0,c_0$ is a (univariate) half-*t* and the priors for two different variance components are uncorrelated, i.e. $Cov\left(\psi_d,\psi_{d^*}\mid b_0,c_0\right)=0$ whenever $d \neq d^*$. See appendix 1 for a proof of [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 $E\left(\lambda^{-1}\right)=c_0$. This helps to interpret c_0 as a scaling constant for the random effects variance $V\left(u_d\right)=a/\lambda$. The choice of c_0 depends on the scale for the random effects in the problem being considered. According to [11], the parameter b_0 can be interpreted in terms of degrees of freedom of the marginal prior $p(\boldsymbol{\tau})$, 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 we discussed using real survey data. We use data on the Small and Medium Enterprises (SME) sample survey, wave 2008, conducted by the Italian National Statistical Institute (ISTAT). Specifically, we target the estimation of the total value added (VA) for small domains of the population of Italian small and medium manufacturing firms (less than 100 employees). The domains we focus on are smaller than those for which ISTAT provides reliable estimates. Specifically, our domains are defined by cross-classifying: the 20 Italian

NUTS 2 administrative regions, the economic industrial sector (NACE Rev. 2, 2 digit, 22 industries) and firm size (4 classes: less than 10 employees, 10 to 19 employees, 20 to 49 employees, 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.

3.1 Direct estimators and sampling model

The SME survey uses stratified sampling design and strata are defined by cross-classifying NACE Rev. 2 (4 digits) Italian administrative regions and company size in the four classes defined above. The domains 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 $(i=1,\dots,22)$, j the size classes ($j=1,\dots,4$) and r the regions ($r=1,\dots,20$). Given this peculiar structure of the data the domain index d 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 characterized by a sampling rate over 0.75.

The actual sample sizes for the domains we consider ranges from 2 to 184, with a median of 8, a mean of 13.5 and .75 and .9 quantiles equal to 16 and 30 respectively.

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

In our sampling model we assume log-normality according to [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 we have in our analysis.

For all the domains with $n_{ijr}\geq 3$, we tested normality and log-normality using the Shapiro-Wilk test. Results are reported in table 1 below. In reporting the results we consider separately the smallest 90% of the domains $n_{ijr} \leq 30$ 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, while log-normality is accepted in more than 70% of the cases.

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

n_{ijr}	normality	log-normality
≤ 30	0.733	0.959
> 30	0.087	0.713
overall	0.672	0.943

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 illustrated in previous section. To check this, we consider a set of log-normal populations: $Y_d \sim LN\left(\tilde{\mu}_d, \tilde{\sigma}_d\right)$, $d = 1, ..., L$ where $L = 77$ is the number of domains with $n_{ijr} > 30$ for which log-normality was not rejected and $\tilde{\mu}_d, \tilde{\sigma}_d$ 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 presence of negligible departures from the null. 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.

Table 2: *Kolmogorov-Smirnov distances between the Monte Carlo ($R = 10{,}000$) distribution of the sample mean and two reference distributions, for various sample sizes.*

	sample size (n_i)			
Reference distribution	5	10	15	20
Log-normal	0.012	0.014	0.015	0.015
Normal	0.119	0.097	0.081	0.073

In order 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 introduced in [1], we have that

$$Var\left\{\log\left(\hat{Y}_{ijr}\right)\right\} = \log\left\{CV^2\left(\hat{Y}_{ijr}\right)+1\right\}. \qquad [12]$$

Thus, the smoothing can be conducted on $CV^2\left(\hat{Y}_{ijr}\right) = \frac{\hat{V}_{ijr}}{\hat{Y}_{ijr}^2}$. After careful explorative analysis, we assume that $CV^2\left(\hat{Y}_{ijr}\right)$ 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}, $\hat{V}_{ijr}$:

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

with $E\left(\upsilon_{ijr}\right)=0$, $V\left(\upsilon_{ijr}\right)=\kappa$ 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 for the size of the domain n_{ijr}. The domain sample size n_{ijr} in the denominator of [13] allows for the decrease in the coefficient of variation when the sample size increases. Smoothed squared estimated coefficient of variations are given by $CV_{smooth}^2\left(\hat{Y}_{ijr,k}\right) = \frac{\phi_j}{n_{ijr}}\left(1-\frac{n_{ijr}}{N_{ijr}}\right)$; the first, second and third quartiles of $CV^2{}_{smooth}\left(\hat{Y}_{ijh,k}\right)$ 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 $\hat{\delta}_{ijr} = \log\left\{CV_{smooth}^2\left(\hat{Y}_{ijr}\right)+1\right\}$ defined according to [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-ASIA 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 log-scale, we can write $\eta_{ijr} = \beta_0 + ltt_{ijr}\beta_1 + u_{ijr}$. The prior for the vector of domain-specific random intercepts u_{ijr} is specified according to [8]-[10]. As for the prior specifications not already discussed, we set $\beta_0 \sim N\left(0,10^5\right)$, $\beta_1 \sim N\left(0,10^5\right)$, $b_0 = 2$, $c_0 = 1$. We chose these values as they provide a reasonable scale for the random effects variance in our problem.

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

$$u_{ijr} \left| \sigma^2 \sim N\left(0,\sigma^2\right),\ \sigma^2 \sim InverseGamma\left(c,d\right) \right. \qquad [14]$$

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

3.3 Markov Chain Monte Carlo computational issues

Parameter estimates are obtained by summarizing the posterior distributions approximated by the output of Markov Chain Monte Carlo (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 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 autocorrelation diagrams and by means of the potential scale reduction known as the Gelman-Rubin statistic (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 can be downloaded for free on the internet and it is open source.

3.4 Comparing alternative models

In order to choose among competing models, we compute the Deviance Information Criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPLM, 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 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 using LPLM is consistent with that obtained using the DIC. The adoption of the Variance Gamma for the random intercepts u_{ijr} leads to a further reduction in DIC with respect to the more common specification [14].

Table 3: *Comparison of alternative assumptions on the distributions of the random effects*

Shrinkage	Prior on random intercepts u_{ijr}	a	DIC	LPLM	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

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 $CVR_k^h = 1 - CV_k^h / CV_k^{DIR}$, where the 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, while CV_k^{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 suggested in the context of small area estimation byYou and Rao (2002) and considered also in Fabrizi and Trivisano (2016):

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

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 centered around the direct estimate. Values of dis_{ijr} far from 0 and 1 would provide evidence of systematic under or over estimation. For the log-normal model endowed with priors [10]-[12] and $a = 0.5$ (i.e. the best model in terms of DIC and LPLM), 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 considered in table 3.

Results on $median(CVR^h_k)$, reported in table 3, highlight that the whole set of considered small area estimators 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 impact. Prior specifications mimicking the spike and slab behavior allow for a further gain in efficiency with respect to priors ordinarily used in this type of analysis.

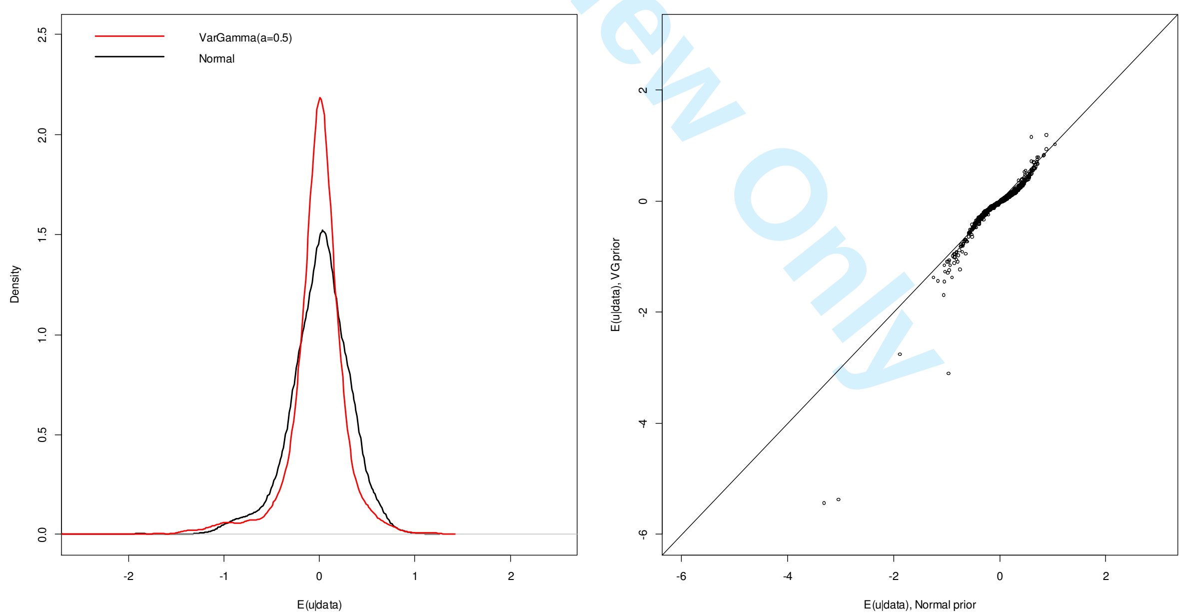

Figure 2: *Left panel: kernel densities of* $E\left(u_{ijr} \mid data\right)$ *over the set of small areas under Normal and VG priors. Right panel:* $E\left(u_{ijr} \mid data\right)$ *under Normal prior vs* $E\left(u_{ijr} \mid data\right)$ *under VG prior. For VG priors,* $a = 0.5$ *is assumed.*

To evaluate the improvements allowed 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 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, while for the model based ones this number grows to 70%. Although the uncertainty around the small area estimates remain sizeable and not all estimates would be publishable, the application of the proposed method endows most sub-population with a publishable estimate in spite of the small sample sizes.

Figure 2 displays the impact of alternative prior specifications on the ensemble of the random intercepts' posterior means $E\left(u_{ijr} \mid data\right)$ under the Normal prior [14] and Variance Gamma priors $u_{ijr} \sim VG\left(0.5, \sqrt{2\lambda}, 0, 0\right)$. Based on the left panel, it is clear that the Variance Gamma leads to a more peaked distribution of estimated random intercepts, as predicted by theory. As tail behavior is difficult to read from density estimates, in the right panel of Figure 2, we plotted the point posterior expectation under the Normal prior vs those obtained under the Variance Gamma. The peak around 0 is still apparent from the inflection of the points' cloud approximately at 0; heavier tails under Variance Gamma can also be appreciated: under the Normal prior, $E\left(u_{ijr} \mid data\right)$ lies within the interval $(-4, 2)$, while they do not under the Variance Gamma specification.

As the purpose of small area estimators is to complement the direct estimates obtained using ordinary survey weighted methods, robustness with respect to modelling assumptions is a major concern. As anticipated, 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) introduced in section 2 and applied to the analysis of 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 discussed in previous section. We drop from the 2007 wave those firms sampled also in the 2008. We obtain a population of size $N = 30451$. Domains are defined by cross-classifying the population by firm size and industry sector; with respect to the data analysis of previous section we collapse the regions. By reducing the number of domains we create sub-populations large enough to be sampled using reasonable sampling rates. Collapsing by region has a milder impact on sub-population skewness with respect to firm size or industry sector. Thereby our synthetic population is divided into $m = 88$ domains whose sizes N_d ($d = 1,...,m$) range from 14 to 1339 with an average of 346. The same target parameters (total VA) and auxiliary variable (turnover) studied in section 4 are considered.

We keep this synthetic population as fixed and we repeatedly 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 using analytic formulas. The point estimators we compare in the simulation are:

i) the posterior mean associated to the Fay-Herriot model specified on the un-transformed scale (UFH). The unstrasformed Fay-Herriot model can be described as follows: $\hat{Y}_d \sim N(\theta_d, V_d)$, where V_d is the variance of the direct estimator $\hat{Y}_d$, $\theta_d \sim N(\beta_0 + \beta_1 x_d, \sigma_v^2)$, $\sigma_v \sim Unif(0, A)$, $A = 1000$, $\beta_i \sim N(0, 10^5)$ $i = 0,1$. We consider the UFH model as it is probably the most "basic" Bayesian model a practitioner would think of for analyzing these data.

ii) the predictor proposed by Slud and Maiti (2006), we denote as SM;

iii) 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") *iv)* the posterior mean associated to the model [1]-[3] but endowed with local shrinkage prior described in [8]-[10] and $a = 0.5$; we label the estimator as LNLS ("Log-Normal with Local Shrinkage").

In all cases we set $\beta_0 \sim N\left(0,10^5\right)$, $\beta_1 \sim N\left(0,10^5\right)$, $b_0 = 2$, $c_0 = 1$, as in the application section.

Let's denote with est_{ds} the generic estimator calculated for the domain d in the 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 (w.r.t. the coverage probability $1-\alpha$) and set $\alpha = 0.05$. Comparison tools are defined as follows:

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

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

$$COV95 = \frac{1}{m}\sum_{d=1}^{m}\frac{1}{S}\sum_{s=1}^{S} 1\left\{Y_d \in \left(p_{\theta_{ds}|data}(0.025), p_{\theta_{ds}|data}(0.975)\right)\right\}$$

As we are interested in the frequentist coverage of Bayes estimators, $COV95$ is calculated only for *i*, *iii*, *iv*. Results are summarized in table 4.

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

est	*RABIAS*	*RRMSE*	*COV*95
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

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 parametrized model. Actually, they are based on similar ideas and models, only the priors or the way hyper-parameters 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 you compromise between the efficiency of a biased synthetic predictor and the unbiasedeness of large variance direct estimators; to some extent estimators associated to areas with very small area-specific sample sizes are naturally biased.

The naïve Fay-Herriot model, specified on the un-transformed 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 performances reflect 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 already noted that this assumption is not completely unrealistic, so misspecification of the sampling model can be held as 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, that is those characterized by a deviation from the synthetic component much larger than most of the remaining areas. We investigate performances separataly for the areas 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 $n_d = 34$, well above the average sample size of the simulation.

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

est	*RABIAS*	*RRMSE*	*COV*95
SM	0.4763	0.4790	--
LNGS	0.4414	0.4453	0.505
LNLS	0.0144	0.2515	0.947

The LNGS and the SM predictor are based on a similar global shrinkage idea. Results in table 5 show how the common variance parameter 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. Local shrinkage prior associated to 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, a feature often displayed in business survey data.

The proposed methodology can be easily implemented using widely available MCMC software. Openbugs codes, as well as formulas for the full conditional distributions needed for an independent implementation of the algorithm, are available upon request from the authors.

In summary, we showed that the predictor based on local shrinkage prior has overall acceptable frequentist properties, comparable to the alternatives we consider in the exercise. If most of the areas are well fit by the assumed model 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 we propose may be applied to estimating business totals based on any positively skewed variables: value added, turnover, labor 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 sub-populations of Italian small- and medium-sized manufacturing firms classified according to sub-national region, industry and firm-size classes. We limit our attention to small and medium enterprises, that is, on firms with less 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 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 in order 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.

References

Barndorff-Nielsen, O.E. (1977). Exponentially decreasing distributions for the logarithm of particle size, Proceedings of the Royal Statistical Society, series A, 353, 401-419.

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.

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.

Datta, G. S., Hall, P. and Mandal, A. (2011). Model selection by testing for the presence of small-area effects in area-level data. Journal of the American Statistical Association, 106, 362-374.

Datta, G.S., Mandal A. (2015) Small Area Estimation with Uncertain Random Effects, Journal of the American Statistical Association, in press.

Eurostat (2011), Key figures on European business - with a special feature on SMEs, Eurostat Pocketbooks.

Eurostat (2015a). Eurostat regional yearbook 2015. Statistical books, General and regional statistics.

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

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

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

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.

Fenton, L. (1960) The Sum of Log-Normal Probability Distributions in Scatter Transmission Systems. *IRE Transactions on Communication Systems* 8(1), 57-67.

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.

Gelman A., Meng X., Stern H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6, 733-807.

Gelman A. (2006). Prior distributions for variance parameters in hierarchical models. Bayesian Analysis, 1, 3, 515–533.

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

Horrace W.C. (2005) Some results on the multivariate truncated normal distribution, Journal of Multivariate Statistics, 94, 209-221.

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

Jiang J. & 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. Journal of Official Statistics, 16, 229-241.

Mazmanyan, L., Ohanyan, V., Trietsch, D. (2009) The lognormal central limit theorem for positive random variables. Reproduced as Appendix in Baker K.R, Trietsch D. (2009), Principles of sequencing and scheduling, John Wiley and Sons, New York.

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, 178, 1051-1067.

Park, Casella G. (2008). The Bayesian lasso, Journal of the American Statistical Association, 103, 681–686.

Pratesi M. ed. (2016). Analysis of Poverty Data by Small Area Estimation, John Wiley, New York.

OECD (2013), OECD Regions at a Glance 2013: OECD Publishing. http://dx.doi.org/10.1787/reg_glance-2013-en

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

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

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

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.

Statistics Canada (2007) 2005 Survey of Financial Security. Public Use Microdata File User Guide. Ottawa: Statistics Canada. (Available from http://www.statcan.gc.ca/pub/13f0026m/13f0026m2007001-eng.htm.)

Wolter K.M. (1985). Introduction to variance estimator, New York, Springer-Verlag.

You, Y. & Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. Canadian Journal of Statistics, 30, 3–15.

Appendix 1

Proof of [6]

To start with we note that

$$\hat{\theta}_d^{B1} = \exp\left\{ \mathbf{x}_d^t \boldsymbol{\beta} + \gamma_d \left(\hat{Z}_d - \mathbf{x}_d^t \boldsymbol{\beta} \right) + \frac{1}{2} \gamma_d \hat{\delta}_d \right\} = \exp\left\{ \gamma_d \hat{Z}_d + (1-\gamma_d) \mathbf{x}_d^t \boldsymbol{\beta} + \frac{1}{2} \gamma_d \hat{\delta}_d \right\} \text{ so}$$

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

We note that $E\left(\hat{Z}_d\right) = E_M\left\{ E\left(\hat{Z}_d \mid \eta_d\right)\right\} = E_M\{\eta_d\} = \mathbf{x}_d^t \boldsymbol{\beta}$; analogously $V\left(\hat{Z}_d\right) = V_M\left\{ E\left(\hat{Z}_d \mid \eta_d\right)\right\} + E_M\left\{ V\left(\hat{Z}_d \mid \eta_d\right)\right\} = \psi_d + \hat{\delta}_d$. As a consequence $E\left(\gamma_d \hat{Z}_d\right) = \gamma_d \mathbf{x}_d^t \boldsymbol{\beta}$, $V\left(\gamma_d \hat{Z}_d\right) = \gamma_d^2\left(\psi_d + \hat{\delta}_d\right) = \psi_d \gamma_d$ and $\mathrm{E}\left[\exp\left\{\gamma_d \hat{Z}_d\right\}\right] = \exp\left\{ \gamma_d \mathbf{x}_d^t \boldsymbol{\beta} + \frac{1}{2}\psi_d \gamma_d \right\}$. This leads to

$$E\left(\hat{\theta}_d^{B1}\right) = \exp\left\{ \mathbf{x}_d^t \boldsymbol{\beta} + \frac{1}{2} \gamma_d \left(\psi_d + \hat{\delta}_d\right) \right\} = \exp\left\{ \mathbf{x}_d^t \boldsymbol{\beta} + \frac{1}{2} \psi_d \right\}$$ that coincides with $E_M(\theta_d)$.

Proof of [7]

Let's introduce some matrix notation. Let $\mathbf{z} = vec\left(\hat{Z}_d\right)$ is the vector containing, $\boldsymbol{\Psi} = diag(\psi_d)$, $\boldsymbol{\Delta} = diag\left(\hat{\delta}_d\right)$ the diagonal matrixes containing the variance components; let $\mathbf{X}$ be the matrix with rows $\mathbf{x}_d^t$, $d = 1,\ldots,m$.

Standard Bayesian analysis of normal linear mixed model lead to $\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\psi} \sim N\left(\hat{\boldsymbol{\beta}}_{gls}, \mathbf{V}(\boldsymbol{\Psi})\right)$ where $\hat{\boldsymbol{\beta}}_{gls} = \left[\mathbf{X}^t (\boldsymbol{\Psi}+\boldsymbol{\Delta})^{-1} \mathbf{X}\right]^{-1} \mathbf{X}^t (\boldsymbol{\Psi}+\boldsymbol{\Delta})^{-1} \mathbf{z}$, $\mathbf{V}(\boldsymbol{\Psi}) = \left[\mathbf{X}^t (\boldsymbol{\Psi}+\boldsymbol{\Delta})^{-1} \mathbf{X}\right]^{-1}$. We can calculate $E(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}) = E_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}} E(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi})$ where $\boldsymbol{\eta} = vec(\eta_d)$. $E_{\beta\mid z,\Psi} E(\eta \mid z, \beta, \Psi) = E_{\beta\mid z,\Psi}(X\beta) = X\hat{\boldsymbol{\beta}}_{gls}$.

Analogously $V(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}) = V_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}} E(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi}) + E_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}} V(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi})$. If we denote the vector of small area predictors (on the log scale) conditional on $\boldsymbol{\beta}$ as $\boldsymbol{\eta}^{B1} = E(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\Psi}) = \boldsymbol{\Gamma}\mathbf{z} + (\mathbf{I}-\boldsymbol{\Gamma})\mathbf{X}\boldsymbol{\beta}$ with $\boldsymbol{\Gamma} = \boldsymbol{\Psi}(\boldsymbol{\Psi}+\boldsymbol{\Delta})^{-1}$ we have that $V(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}) = V_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}} \boldsymbol{\Gamma}\mathbf{z} + (\mathbf{I}-\boldsymbol{\Gamma})\mathbf{X}\boldsymbol{\beta} + E_{\boldsymbol{\beta}\mid\mathbf{z},\boldsymbol{\Psi}} \mathbf{G}_1$ with $\mathbf{G}_1 = \boldsymbol{\Psi}(\boldsymbol{\Psi}+\boldsymbol{\Delta})^{-1}\boldsymbol{\Delta}$. Taking expectation with respect to $p(\boldsymbol{\beta} \mid \mathbf{z}, \boldsymbol{\Psi})$ we get

$V(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi}) = (\mathbf{I}-\boldsymbol{\Gamma})\mathbf{X}^t\boldsymbol{\beta}\mathbf{V}(\boldsymbol{\Psi})\mathbf{X}(\mathbf{I}-\boldsymbol{\Gamma}) + \mathbf{G}_1 = \mathbf{G}_2 + \mathbf{G}_1$. We note that $p(\boldsymbol{\eta} \mid \mathbf{z}, \boldsymbol{\Psi})$ is a multivariate normal distribution. If we consider an individual η_d have that $E(\eta_{ijr} \mid \mathbf{z}, \boldsymbol{\Psi}) = \mathbf{x}_d^t \hat{\boldsymbol{\beta}}_{gls}$,

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

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

Proof of [11] and subsequent statements

From [9] we have that $p(\boldsymbol{\psi} \mid \lambda) = \prod_{d=1}^{m} \frac{\lambda^a}{\Gamma(a)} \psi_d^{a-1} e^{-\lambda \sum_{d=1}^{m} \psi_d} \propto \lambda^{ma} e^{-\lambda \sum_{d=1}^{m} \psi_d} \prod_{d=1}^{m} \psi_d^{a-1}$

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

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

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

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

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

$\tau_d > 0$, $\forall d$. 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 is a multivariate t for which we apply the absolute value transformation on each component. We can then write formula [13].

To prove that univariate priors $p(\tau_d)$ let's start from $p(\tau_d)=\int\ldots\int p(\boldsymbol{\tau})\,d\boldsymbol{\tau}_{-d}$. We can represent as $p(\boldsymbol{\tau})$ as the result of applying the absolute value transformation on a multivariate t distribution, i.e $\boldsymbol{\tau}=\left|\boldsymbol{\tau}^*\right|$ with $p(\boldsymbol{\tau}^*)=\int p(\boldsymbol{\tau}^*\mid\xi)p(\xi)\,d\xi$ where $p(\boldsymbol{\tau}^*\mid\xi)\sim MVN\left(\mathbf{0},\frac{2b_0}{c_0}\xi\mathbf{I}_m\right)$ and $p(\xi)\sim IGamma(b_0,b_0)$. 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.

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

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

$$p(\tau_d)=\int\ldots\int\left[\int\prod_{d=1}^{m}p(\tau_d\mid\xi)\,p(\xi)\,d\xi\right]d\boldsymbol{\tau}_{-d}$$

where each $p(\tau_d\mid\xi)$ is distributed as an half-Normal distribution.

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

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

$$V(\boldsymbol{\tau})=E_\xi\left\{V(\boldsymbol{\tau}\mid\xi)\right\}+V_\xi\left\{E(\boldsymbol{\tau}\mid\xi)\right\}$$

and note that $E(\boldsymbol{\tau}\mid\xi)=\mathbf{0}$ while $E_\xi\left\{V(\boldsymbol{\tau}\mid\xi)\right\}=E(\xi)\frac{2b_0}{c_0}\mathbf{I}_m$, which is of course diagonal.