Bayesian Analysis (2017) **12**, Number 3, pp. 729–752

A Generalised Semiparametric Bayesian Fay–Herriot Model for Small Area Estimation Shrinking Both Means and Variances

Silvia Polettini[*]

Abstract. In survey sampling, interest often lies in unplanned domains (or small areas), whose sample sizes may be too small to allow for accurate design-based inference. To improve the direct estimates by borrowing strength from similar domains, most small area methods rely on mixed effects regression models.

This contribution extends the well known Fay–Herriot model (Fay and Herriot, 1979) within a Bayesian approach in two directions. First, the default normality assumption for the random effects is replaced by a nonparametric specification using a Dirichlet process. Second, uncertainty on variances is explicitly introduced, recognizing the fact that they are actually estimated from survey data. The proposed approach shrinks variances as well as means, and accounts for all sources of uncertainty. Adopting a flexible model for the random effects allows to accommodate outliers and vary the borrowing of strength by identifying *local* neighbourhoods where the exchangeability assumption holds. Through application to real and simulated data, we investigate the performance of the proposed model in predicting the domain means under different distributional assumptions. We also focus on the construction of credible intervals for the area means, a topic that has received less attention in the literature. Frequentist properties such as mean squared prediction error (MSPE), coverage and interval length are investigated. The experiments performed seem to indicate that inferences under the proposed model are characterised by smaller mean squared error than competing approaches; frequentist coverage of the credible intervals is close to nominal.

Keywords: Dirichlet process prior, Fay–Herriot, Hierarchical models, mixed effects regression models, small area, smoothing of sampling variances.

1 Introduction and Motivation: The Fay–Herriot Model

The information content of a sample survey is clearly not limited to the planned domains, and researchers are often interested in obtaining estimates for a whole variety of subpopulations, that we refer to as small areas. The reason for this terminology is that often the sample sizes for such domains are too small to provide sufficiently accurate design-based estimates of the domain parameters.

To improve such estimates, indirect estimators introduce linking models that rely on auxiliary information to connect small areas, thus borrowing strength and increasing the effective sample size. Modern methods for small area estimation (SAE hereafter)

[*]Dipartimento MEMOTEF, Sapienza Università di Roma, Via del Castro Laurenziano 9, 00161 Roma, Italy, silvia.polettini@uniroma1.it

DOI: 10.1214/16-BA1019

heavily rely on mixed effects modelling. The book by Rao and Molina (2015) contains a thorough analysis of the model based approach to small area estimation and SAE methods in general.

In this contribution we focus on area level models. Area-level models rely on aggregated, area-specific, quantities and are often the only available model under microdata confidentiality protocols. Indeed, due to disclosure limitation procedures, data aggregated to the area level are currently more readily available to users than are unit level data, for both the variable of interest as well as for the auxiliary information. Another advantage of area level modelling is that it allows one to account for the sampling design in a straightforward way, by introducing the direct survey estimators of the target small-area parameters in the sampled areas, and their corresponding (design-based) variance estimates. Typically, in small area estimation problems the direct estimators may present unacceptably high variances due to small sample size in some or all of the small areas.

In the literature, the first and most popular area-level model is the one proposed by Fay and Herriot (1979). The Fay–Herriot model prescribes a sampling model for the direct survey estimates, supplemented by a linking model for the small area parameters of interest. Let m be the number of sampled small areas. Under the sampling model, design unbiased, direct estimators $\hat{\theta}_i$ of the small area parameters θ_i, $i = 1, \ldots, m$, are assumed to be available, whose sampling error is ϵ_i. Usually the ϵ_i's are assumed to be independent, normally distributed random variables with known variance ψ_i, $\epsilon_i \sim N(0, \psi_i)$, so that

$$\hat{\theta}_i \sim N(\theta_i, \psi_i), \quad i = 1, \ldots, m. \tag{1}$$

To achieve the desired borrowing of strength across areas, a linking model for θ_i is introduced, namely $\theta_i = x_i'\beta + \nu_i$, where $x_i = (x_{i1}, \ldots, x_{ip})'$ is a vector of auxiliary variables, β is a vector of regression coefficients, and finally the v_i's are area-specific random effects accounting for heterogeneity and lack of fit. Normality of the random effects is usually assumed: $\nu_i \overset{i.i.d.}{\sim} N(0, \sigma_\nu^2)$, so that

$$\theta_i | \beta, \sigma_\nu^2 \sim N(x_i'\beta, \sigma_\nu^2), \quad i = 1, \ldots, m. \tag{2}$$

Combining the previous equations together, one obtains a mixed effects linear regression model with normal random components, $\hat{\theta}_i = x_i'\beta + \epsilon_i + \nu_i$. Since areas of interest may not be all sampled in practice, it is assumed that the combined area-level model above also holds for the non-sampled areas. This amounts to assuming no selection bias for areas. If the parameters β, σ_ν^2 were known, the small area estimator would be the Best Linear Unbiased Predictor (BLUP)

$$\tilde{\tilde{\theta}}_i = {x_i}'\beta + \gamma_i(\hat{\theta}_i - {x_i}'\beta), \quad \gamma_i = \frac{\sigma_\nu^2}{\sigma_\nu^2 + \psi_i}, \quad i = 1 \ldots, m;$$

because the above parameters are unknown, plug-in estimators are implemented, that are referred to as the Empirical Best Linear Unbiased Predictor (EBLUP) or empirical Bayes (EB) estimator. Many papers in the literature discuss procedures for estimating model parameters and measures of uncertainty for the EBLUP $\tilde{\theta}_i$, usually the mean squared prediction error, defined as $\text{MSPE}_i = E(\tilde{\theta}_i - \theta_i)^2$. See e.g. Prasad and Rao (1990); Datta and Lahiri (2000); Rao and Molina (2015); Datta and Ghosh (2012) and references therein.

In the Fay–Herriot setup, accuracy and precision of small area predictors depend on the validity of the model. In this respect, we focus specifically on two major underlying assumptions, namely the assumption of known sampling variances in (1) and the normality of random effects (2). As to the first point, an identifiability issue arises for the Fay–Herriot model, unless the sampling variances ψ_i are assumed to be known. Most of the literature has focused on reflecting uncertainty in estimating σ_ν^2 while assuming known ψ_i (Prasad and Rao, 1990; Datta and Lahiri, 2000). In practice however the latter quantities are estimated from the sample; in some applications, smoothed estimators, usually obtained by means of the generalised variance function (GVF, see Dick, 1995) approach, are introduced, and then treated as known. Quoting Hawala and Lahiri (2010) "the small area data, especially the design-based variance estimates, could be very unreliable and noisy making it difficult to identify a reasonable GVF model with small area specific random effects". Wang and Fuller (2003) and Rivest and Vandal (2003) consider estimation of the MSE when the ψ_is are estimated. Bell (2008) investigated sensitivity of inferences to uncertainty about sampling variances, reporting larger true mean squared errors (MSE) and biases in reported MSE when the survey variances are poorly estimated, the major problem being identified in underestimation of survey variances when the ratio ψ_i/σ_v^2 is large, which is typical for small counties. Rao and Molina (2015) report the relative increase in MSE of the BLUP obtained assuming ψ_i known and equal to the estimated values, and the relative bias of the reported MSE estimator.

As to the second point, whereas for area-level models the distributional assumption on the sampling errors ϵ_i is usually justified by the properties of the direct estimators $\hat{\theta}_i$, the normality assumption for the random effects ν_i has no justification other than computational convenience. Moreover, this assumption is difficult to detect in practice, since it involves unobservable quantities. The problem affects both frequentist and Bayesian analyses, although availability of Markov Chain Monte Carlo (MCMC) techniques makes computational convenience less relevant in the latter framework. Datta and Lahiri (1995) and Chakraborty et al. (2016) discuss the impact of outliers on the shrinkage property of the standard Fay–Herriot model. They observe that even a single substantively outlying observation may corrupt the shrinkage property of the small area model due to overestimation of the model variance σ_ν^2. To overcome the problem and accommodate for outliers Datta and Lahiri (1995) model the small area means by means of heavy-tailed distributions, obtained as scale mixture of normals with given mixing distribution. More recently, to avoid specification of the mixing distribution above, Chakraborty et al. (2016) propose a two-component normal mixture, analogous to the frequentist outlier-robust model of Sinha and Rao (2009). The second element of the mixture describes the model error for the outlying units and has an inflated variance: $\theta_i = x_i'\beta + (1-\delta_i)\nu_{1i} + \delta_i\nu_{2i}$, the δ_is, $i = 1, \ldots, m$ being $Bernoulli(p)$ random variables characterising the outlying areas, and $\nu_1 \sim N(0, V_1), \nu_2 \sim N(0, V_2)$. Using noninformative priors, Chakraborty et al. (2016) explore the shrinkage effect under the mixture model above and assess flexibility and robustness of their proposal.

Addressing the above mentioned issues, this contribution extends the Fay–Herriot model within a Bayesian approach in two directions:

1. First, a hierarchical Bayesian (HB) formulation of the Fay–Herriot model is proposed, in which the default normality assumption for random effects is replaced by a nonparametric specification, namely using a Dirichlet process (DP hereafter, see Ferguson, 1973; Antoniak, 1974). Using the HB representation of the standard Fay–Herriot model, it is easy to see that it borrows strength from all areas by shrinking small area parameters to a common mean; usually shrinkage is higher for areas where the direct estimates have more variability. Borrowing strength is achieved by applying an exchangeability assumption across small areas, with the amount of shrinkage being only adjusted, not the sets of areas where borrowing can take place. Several definitions of closeness, besides spatial closeness, may be implemented to select such sets; also, it may be difficult to identify which units belong to each set. In addition, some small areas may be outliers and should not be used in drawing inference about other areas and, vice versa, other areas should not be used in drawing inference about them. In our formulation, the aim of achieving different levels of shrinkage and accommodating outliers is pursued by choosing the Dirichlet process to model the random effects. This amounts to define a probability model over all possible partitions of small areas. Due to the clustering structure induced by the DP (see Ferguson, 1973), the resulting estimates selectively use data from only subsets of the observed small areas. Indeed the posterior predictive distribution of the area means θ_i will support common random effects for areas that are "close", that is, belong to the same cluster. The uncertainty about which clusters are valid is automatically accounted for. This model is useful when, a priori, one suspects that small areas are not equally similar but there is no information about where the dissimilarities are (Malec and Müller, 2008).

2. Second, following e.g. You and Chapman (2006), uncertainty on variances is introduced in the model, so as to reflect the fact that they are actually estimated from survey data. To allow for unknown variances, the model proposed in the next Section adds a level in the hierarchy, that permits to smooth variances without introducing an exogenous model for that. Denoting by n_i the sample size observed in area i, You and Chapman (2006) model the sampling within-area variances as

$$(n_i - 1)S_i^2 \sim \psi_i \chi^2_{(n_i-1)}, \quad i = 1, \dots, m$$

independently of $\hat{\theta}_i$. You and Chapman (2006) assign the parameters ψ_i separate priors with known hyperparameters, therefore their model does not enjoy the shrinkage property of variances. The previous model holds e.g. when data for area i, $i = 1, \dots, m$ are i.i.d. normal so that $\hat{\theta}_i$ and S_i^2 are independent and $(n_i - 1)$ are the degrees of freedom in estimating ψ_i. Wang and Fuller (2003, p. 721) suggest that when the normality assumption does not hold, the chi-squared distribution

> may still represent a good approximating model for the sampling variances once a careful determination of the "equivalent" degrees of freedom δ_i is performed.
>
> In the proposed approach shrinkage is obtained by assuming a common distribution generating the variance parameters as in Dass et al. (2012). As highlighted in Maiti et al. (2014) an important feature of the model is the dual shrinkage property. In particular, modelling of variances is performed within the model, and inference on means also reflects uncertainty about sampling variances. We refer to Maiti et al. (2014) for a discussion of the different approaches taken in the literature concerning the issue of unknown sampling variances.

While MSE estimation under several types of model misspecification has received considerable attention, the construction of confidence intervals for the area quantities has not been investigated thoroughly in the literature. Recently Diao et al. (2014) focus on deriving accurate CIs for small area means using the EBLUPs and estimators of mean square prediction error of EBLUPs based on various methods of estimation of model parameters, under the standard Fay–Herriot model. They also investigate robustness to misspecifications of the random component distribution via simulation. In this paper we consider both point and interval prediction and investigate the performance of the proposed model in both these problems. In Maiti (2003) a unit-level Bayesian hierarchical model is proposed, where the random intercept is modelled by a Gaussian mixture with an unknown number of components; a simulation study illustrates that in particular interval estimation may be misleading when the distribution of the random effects is misspecified. Malec and Müller (2008) discuss use of DP in the context of small area estimation. Their model is formulated as a unit-level setup in which the county-specific random effects are described by a mixture of Dirichlet processes. Dass et al. (2012) and Maiti et al. (2014) propose an extension of the Fay–Herriot model shrinking both means and variances, but do not include a flexible representation of the random effects. To our knowledge, the formulation proposed in this paper has not previously appeared in the literature.

The paper is structured as follows: Section 2 describes the proposed model; viability of the approach and the effect of introducing a flexible specification of the random effects and an explicit shrinkage of sampling variances are investigated through application to simulated and real data in Sections 3 and 4. Section 5 contains final remarks.

2 Proposed Approach

As mentioned in the Introduction, this contribution focuses on two major assumptions underlying the Fay–Herriot model, namely the normality of random effects and the assumption of known sampling variances. The proposed model aims to address both aspects within a Bayesian formulation. We assume that (independent) information is available about sampling variances and choose to include this as an additional stage in the proposed hierarchical model, so that estimation of area means and sampling variances is performed at the same time and smoothing of variances is performed within the same model. Inference on small area means incorporates the uncertainty on sampling variances: a further consequence is that the proposed approach allows to produce a formal quantification of sampling variances. Adopting a flexible specification of the random effects would allow greater flexibility, and robustness against model misspecifications. Although McCulloch and Neuhaus (2011) conclude in favour of a substantial insensitivity of linear random effects models to assumptions on the random effects distribution, they point out an impact on the prediction of the random effects, which is reportedly modest as far as the mean squared error of prediction is concerned, and high as far as the shape of the distribution of the best predicted values is concerned. Interval estimation may be affected by this problem. Accurate prediction of the random effects is crucial for predicting small area quantities, and the presence of outliers may be problematic. Also, estimation of nonlinear functionals may suffer from misrepresentation of the law of the random effects, as remarked in Fabrizi and Trivisano (2010). The previous authors develop two robustified versions of the Fay–Herriot model, assuming either exponential power (EP) or skewed-EP distributed random effects, and investigate robustness of such models under deviations from normality. Their aim is to understand whether estimates of linear and, especially, nonlinear functionals such as ranks are sensitive to deviations from normality of the random effects. Although the models proposed by Fabrizi and Trivisano (2010) are based on distributions that generalise, and contain, the normal, yet these parametric models may fail to adequately describe the distribution of the random effects, and again the problem of checking the adequacy of these models arises. Many proposals in the literature try to extend the model for the small area means. As mentioned in Section 1, Datta and Lahiri (1995) generalise the Fay–Herriot model by introducing an heavy-tailed distribution, defined as a scale mixture of normal distributions, to properly account for outliers; in the same context, Sinha and Rao (2009) and Chakraborty et al. (2016) propose a two-components mixture of normals. Maiti (2003) proposes a Bayesian hierarchical unit-level model where the random intercept is modelled by a Gaussian mixture with an unknown number of components. Here we consider a different though related extension of the Fay–Herriot model based on Dirichlet process (DP) priors, where the assumption of normality producing the linking model (2) is replaced by

$$\nu_i|\sigma_\nu^2, M \overset{i.i.d.}{\sim} G(\cdot), \quad i = 1, \ldots, m, \quad G \sim DP(M, N(0, \sigma_\nu^2)), \tag{3}$$

where $DP(M, G_0)$ stands for the Dirichlet process (DP) with precision parameter M and base measure G_0. In the context of a generalisation of the Fay–Herriot model, it is natural to assume G_0 to be a normal distribution.

The representation above not only relaxes the normality assumption, but also provides an enlarged model for describing the random effects, with specific advantages. As mentioned, a flexible model for describing the random effects having the capacity of borrowing strength from subsets of units would be key to improve the model, induce robustness against outliers and reduce the prediction error. In particular, (3) defines a probability model over all possible partitions of small areas. As a consequence, small areas are partitioned into clusters sharing the same random effect, with areas within the same cluster being independent, but not independent between clusters. Given m observations from a Dirichlet process, the probability of observing a partition $m_1, \ldots, m_k$

such that $\sum_{j=1}^{k} m_j = m, m_j > 0, j = 1, \ldots, k$ is

$$\pi(m_1, \ldots, m_k | M, m, k) = \frac{\Gamma(M)}{\Gamma(M+m)} M^k \prod_{j=1}^{k} \Gamma(m_j). \tag{4}$$

Predictions are thus obtained by combining local information, e.g. by focusing on clusters of the observed small areas, the uncertainty about the clustering structure being accounted for through (4). From the previous equation it is clear that choice of M, the precision parameter of the DP, is crucial as it drives the clustering property of the DP. Following the literature on the subject (e.g. Escobar and West, 1994), we introduce a further layer in the hierarchy, modelling this unknown parameter according to a Gamma distribution, see (12). This prior is usually justified due to its conditional conjugacy property. However the problem of selecting the Gamma parameters is nontrivial. Dorazio (2009) shows that the posterior mass for the number of clusters tends to concentrate on $k = 1$ as the shape of the Gamma distribution goes to zero, therefore the usual noninformative argument is not appropriate. Dorazio (2009) suggests to choose the parameters of the Gamma distribution in such a way that the resulting posterior for the total number of clusters is closest to the uniform in terms of the Kullback–Leibler divergence. See also Murugiah and Sweeting (2012), who propose an alternative elicitation of the Gamma prior for M.

To complete the specification of the proposed model, we explicitly model the sample variances S_i^2 as suggested in Dass et al. (2012). This accounts for uncertainty on the sampling variances ψ_i (the true variances of $\hat{\theta}_i$) in predicting small area quantities.

The proposed model reads as follows:

$$\begin{aligned}
\hat{\theta}_i &= \theta_i + e_i, \quad e_i \sim N(0, \psi_i), \quad \text{independently}, \quad i = 1, \ldots, m && (5)\\
\theta_i &= x_i'\beta + \nu_i, \quad \nu_i \sim G(\cdot), \quad \text{independently}, \quad i = 1, \ldots, m && (6)\\
G &\sim DP(M, N(0, \sigma_\nu^2)) && (7)\\
\delta_i S_i^2 &\sim \psi_i \chi^2_{\delta_i}, \quad \text{independently and independent on } \hat{\theta}_i, \quad i = 1, \ldots, m && (8)\\
\psi_i^{-1} &\sim Ga(a_0, b_0), \quad \text{independently}, \quad i = 1, \ldots, m && (9)\\
\sigma_\nu^{-2} &\sim Ga(a_1, b_1) && (10)\\
\beta &\sim N(0, d\mathrm{I}) \quad \text{where I is the identity matrix} && (11)\\
M &\sim Ga(a_2, b_2), && (12)
\end{aligned}$$

where δ_i represents the degrees of freedom for estimating the sampling variance, $Ga(a, b)$ denotes the Gamma distribution with shape a and rate b, and $a_0, b_0, a_1, b_1, d, a_2, b_2$ are known constants. The hyperparameters in (10)–(12) are fixed; for comparability with the EBLUP, the hyperprior on the β vector is assumed to be normal with large variance.

The modelling assumption in (8) is fully justified in a normal setting, with simple random sampling; in our case, it can be taken as an attempt to achieve a more comprehensive quantification of uncertainty in estimating domain means that also accounts for variability of sampling variances. Arora and Lahiri (1997) and You and Chapman

(2006) use $\delta_i = n_i - 1$, n_i being the number of sampled units belonging to each small area. Selecting an appropriate number of degrees of freedom allows one to approximate the distribution reasonably well, as suggested in Wang and Fuller (2003), even in the non-normal case; an application of this principle can be found in Maples et al. (2009). Thorough investigation of the quality of the approximation in complex sample surveys has not been undertaken. Some comments can be found in Maiti (2003).

Unlike You and Chapman (2006), assuming a single prior for the ψ_is allows shrinking the sample variances, to an extent that depends on $n_i, i = 1 \ldots, m$ (see also Dass et al., 2012). The proposed hierarchical model differs from the one analysed in Maiti et al. (2014) for the nonparametric modelling of the random effects.

Setting $\Psi = \{diag(\psi_i)\}$, under the proposed model the likelihood factorizes as $L(\beta, \Psi|\hat{\boldsymbol{\theta}}) \prod_i L(\psi_i|S_i^2)$ where, following Lo (1984), Liu (1996) the first component is

$$L(\beta, \Psi|\hat{\boldsymbol{\theta}}) = \sum_{c=1}^{m} \sum_{C:|C|=c} \frac{\Gamma(M)}{\Gamma(M+m)} M^c \prod_{j=1}^{c} \Gamma(m_j) \int p(\hat{\theta}_{(j)}|\beta, \Psi, \nu_j) dG_0(\nu_j),$$

where C is a partition of areas $\{1, \ldots, m\}$ into c groups (or clusters), m_j is the number of small areas in the j-th cluster, $1 \leq m_j \leq m$, $\hat{\theta}_{(j)}$ is the vector of the direct estimates belonging to cluster j and finally

$$p(\hat{\theta}_{(j)}|\beta, \Psi, \nu_j) = \prod_{k \in cluster\ j} \frac{1}{\sqrt{2\pi\psi_k}} \exp\{-\frac{1}{2\psi_k}(\hat{\theta}_k - x_k'\beta - \nu_j)^2\},$$

and the dependence on Ψ of the latter equation is only through the elements involved in the j-th cluster. As evident from the previous equation, all areas belonging to a given cluster are assigned the same random effect; furthermore, the number of clusters in each partition is unknown. Data are assumed exchangeable only within the same cluster and the posterior predictive distribution of the area means θ_i will support common random effects for areas that are "close", that is, belong to the same cluster. In this formulation all possible partitions are explored and the uncertainty about which clusters are valid is automatically accounted for. In a small area context, such feature may prove particularly useful because, a priori, one can expect that areas are not equally similar, and that some outlying areas should be singled out, but often there is no information about how to form clusters and where the dissimilarities are. Under the assumption of normal random effects, one can expect both "undue influence of larger outliers, and undue impact on smaller outlying areas as they are shrunk towards the overall mean" (Ohlssen et al., 2007): under DP-distributed random effects, shrinkage can be higher for some non-outlying areas and lower for outlying areas, and this is performed in a data-driven way.

It is of interest here to understand the performance of the model in predicting small area quantities under the extended Fay–Herriot model above, primarily the domain means θ_i in (6). We consider measures of prediction error, for which a natural quantification of uncertainty under the proposed approach is the posterior variance. We focus on frequentist properties of the small area predictors and assess the MSPE through simulation under various data generating processes. The measures above reflect in-sample performance and do not have a formal Bayesian justification: to assess the model's predictive performance from a Bayesian viewpoint, measures of prediction accuracy were also evaluated, specifically DIC (Spiegelhalter et al., 2002), WAIC (Watanabe, 2009, 2010) and log pseudo marginal likelihood (LPML, see Geisser and Eddy, 1979). For missing data and mixture models, DIC can be defined and estimated as discussed in Celeux et al. (2006). WAIC and LPML are cross-validated criteria, indicated in the literature as better suited to assess a model's predictive performance, in that a correction is made for using the data twice: in estimating the model and in assessing the model's fit. Besides this, an aspect that has received less attention in the literature is the construction of confidence intervals for the area means. Recently Diao et al. (2014) focus on deriving accurate CIs for small area means using the EBLUPs and estimators of MSPE of EBLUPs based on various methods of estimation of model parameters. Under the Bayesian model proposed in this paper, a natural approach to the former issue is to produce posterior credible intervals. Dass et al. (2012) also develop confidence intervals for the small area means based on a decision theoretic approach. They obtain Bayes confidence intervals by minimizing the expected loss, specified through a function that takes into account both the coverage probability and the length of the interval and that depends on a tuning parameter that allows to vary the weight attached to the interval length compared to the coverage probability. The actual coverage properties of the intervals proposed in Dass et al. (2012) are only investigated under the assumption of normality. We address the performance of the credible intervals obtained under the proposed model and compare it to existing solutions through simulation.

The variable shrinkage property allows us to reduce prediction error and although the proposed model accounts for all sources of uncertainty, applications show that the resulting credible intervals tend to be shorter than other frequentist solutions. The frequentist coverage of such intervals is also analysed. In Section 3 we assess the predictive performance of the model under known data generating processes and describe the findings of a simulation study aimed at investigating the frequentist properties of the proposed approach.

2.1 Posterior Computations

The posterior distribution of the small area quantities θ_i is analytically intractable, so we adopt MCMC techniques to perform inference. We use a Gibbs sampler, repeatedly sampling one set of parameters at a time, specifically $\beta|$rest, $\psi|$rest, $\nu|$rest, $M|$rest, $\sigma_\nu^2|$rest. Given model specification and the previous equations, sampling the fixed effects parameters given the cluster configuration proceeds as in standard normal hierarchical models; updating $\sigma_\nu^2|$rest is also standard, whereas given the structure of the model, one can exploit conjugacy in sampling the random effects; the approach proposed in West et al. (1994) is used. To avoid use of Metropolis steps, $M|$rest is updated using the scheme proposed in Escobar and West (1994). In synthesis, the model is augmented with an extra variable η whose distribution, given M and the number of clusters k, is Beta$(M+1, m)$, such that the conditional distribution of $M|\eta, k$ reduces to a mixture of two gamma densities (see formula (13) in Escobar and West, 1994), which allows straightforward implementation within a Gibbs sampling scheme.

3 Simulation Study

To assess model's performance, we conducted a simulation study, that enables us to benchmark the fitted values to the true underlying values. Following the scheme adopted in Wang and Fuller (2003) and Maiti et al. (2014), unit-level data were generated from the model

$$Y_{ij} = \beta + \nu_i + \epsilon_{ij}, \qquad j = 1, \ldots, n_i,\ i = 1, \ldots, m,$$

with $\beta = 10$, and $\epsilon_{ij} \sim N(0, n_i\psi_i)$. Unequal sampling variances were ensured by choosing three levels of ψ_i, namely $1, 4, 16$, each assigned to one third of the areas. The corresponding area level model is

$$Y_i = \beta + \nu_i + \epsilon_i,$$

with $Y_i = \bar{Y}_i = \sum_{i=1}^{n_i} Y_{ij}/n_i$, $\epsilon_i = \sum_{j=1}^{n_i} \epsilon_{ij}/n_i$.

Therefore $\epsilon_i \sim N(0, \psi_i)$ and $Y_i|\theta_i, \psi_i \sim N(\theta_i, \psi_i)$ with $\theta_i = \beta + \nu_i$. As in Wang and Fuller (2003), we set $n = 36$ areas, with $n_i = 9$ units each, $\beta = 10$ and $\sigma^2_\nu = 1$.

Within the simulation scheme just described, three different assumptions on the random effects ν_i are investigated: we first generate the random effects from a normal distribution, which is the standard setup examined in Wang and Fuller (2003) and Maiti et al. (2014), and then relax such an assumption by introducing a skew-t distribution and a mixture of normals to investigate the model's robustness against departures from normality in the regression component of the model.

The details of the simulation setup are specified below:

M.1) First, the normal case is considered, in which $\nu_i \sim N(0, \sigma^2_\nu)$, as in Wang and Fuller (2003) and Maiti et al. (2014); although Maiti et al. (2014) select three different levels for σ^2_ν, we only consider $\sigma^2_\nu = 1$.

M.2) Second, a skew-t distribution (Azzalini and Capitanio, 2003) for the random effects ν is considered. The skew-t is an extremely flexible distribution, allowing to introduce skewness and heavy tails, and therefore its use induces the presence of outliers. This scheme also allows us to test the effect of misspecifying the model for the sampling variances: indeed the independence between means and variances does not hold in this case and the chi-square distribution is no longer the correct model in (8). The skew-t distribution has four parameters, namely location (ξ), scale (ω), slant (α), and degrees of freedom (v). In the simulation the parameters were set to $\xi = 0, \omega = 1.404, \alpha = 12, v = 10$. Setting $\delta = \alpha/\sqrt{1+\alpha^2}$, the mean of the distribution can be expressed as (Azzalini and Capitanio, 2003)

$$\mu = \xi + \omega\delta\sqrt{\frac{v}{\pi}} * \frac{\Gamma(0.5(v-1))}{\Gamma(0.5v)};$$

the random effects were then centred using the above expression to ensure zero mean. The variance of the skew-t distribution is

$$\sigma^2 = \omega^2\left[\frac{v}{v-2} - \left(\delta\frac{v}{\pi}\frac{\Gamma(0.5(v-1))}{\Gamma(0.5v)}\right)^2\right];$$

Therefore setting $\omega = 1.404$ ensures unit variance of the random effects as in the normal scenario, whereas setting $\omega = 1$ implies a variance of about 0.5, which makes the regression model for the θ_is comparatively more accurate than the direct estimators. This latter choice was also investigated for comparison.

M.3) Finally, a contaminated distribution, analogous to the scheme adopted in Sinha and Rao (2009) was considered for modelling the random effects: $\nu_i \sim (1-\gamma)\,N(0,\sigma_\nu^2) + \gamma\,N(0,\sigma_\nu^{*2})$, with $\sigma_\nu^2 = 1$ and $\sigma_\nu^{*2} = 25$. This gives a rather strong contamination scenario, potentially inducing highly outlying units. In order to assess the effect of increasing the proportion of outliers, two choices of γ were considered, namely, $\gamma = 0.11$ and $\gamma = 0.25$.

Denoting by $\tilde{\theta}_i$ the small area predictions under a given model, we consider as measures of prediction error the following quantities: total bias: $\mathrm{B} = \frac{1}{m}\sum_{i=1}^m E(\tilde{\theta}_i - \theta_i)$ and total mean squared error of prediction: $\mathrm{MSPE} = \frac{1}{m}\sum_{i=1}^m E(\tilde{\theta}_i - \theta_i)^2$, and their relative counterparts:

$$\mathrm{RB} = \frac{1}{m}\sum_{i=1}^{m} E\left(\frac{\tilde{\theta}_i - \theta_i}{\theta_i}\right)$$

$$\mathrm{RMSPE} = \frac{1}{m}\sum_{i=1}^{m} E\left(\frac{\tilde{\theta}_i - \theta_i}{\theta_i}\right)^2.$$

In order to check the model's predictive ability under each of the data generating processes, Bayesian predictive criteria were evaluated for data simulated according to the schemes just described. Results are reported in Table 1: with the exception of the normal setup, in which the parametric model is preferred, in the other cases the semiparametric model is slightly superior in terms of estimated predictive criteria.

3.1 Simulation Results

A total of 3000 simulations from each of the models described above was considered. The following hyperparameters were selected: $a_0 = 0.5, b_0 = .1, a_1 = 1, b_1 = 1, d = 5000, a_2 = 1, b_2 = 0.04$. The choice of the parameters a_1, b_1 was based on the fact that as the base measure of the DP becomes diffuse, the probability of adding new clusters decreases, thus implicitly favouring models with a small number of components (see Rossi, 2014). Indeed, using the Blackwell and MacQueen (1973) representation, the posterior probability of selecting a new atom for the random effect ν_i given the data and the current cluster configuration, conditional on the remaining $m-1$ random effects, is proportional to $Mh_i(\hat{\theta}_i)$, with

$$h_i(\hat{\theta}_i) = \int p(\hat{\theta}_i|\beta, \Psi, \nu_i)dG_0(\nu_i)$$

(see West et al., 1994). Diffuse specifications for G_0 are therefore informative and not desirable.

		WAIC	DIC	LPML
M.1	*Normal*			
	Proposed model	352.68	340.78	-195.75
	You–Chapman model	349.49	339.02	-196.57
M.2	*Skew-t*			
	Proposed model	351.99	343.57	-198.87
	You–Chapman model	354.09	344.23	-199.42
	Skew-t, unit variance			
	Proposed model	335.25	331.95	-188.58
	You–Chapman model	338.44	334.59	-188.43
M.3	*Mixture of normals, 10%*			
	Proposed model	204.84	196.01	-131.78
	You–Chapman model	210.69	197.68	-137.03
	Mixture of normals, 25%			
	Proposed model	229.13	212.30	-136.85
	You–Chapman model	250.89	216.59	-161.97

Table 1: Models' predictive measures under the data generation models adopted in the simulation study: comparison between the proposed semiparametric model and the model of You and Chapman (2006).

For each simulation, the model based point predictions were computed and empirical measures of total bias and MSPE were obtained, averaging over areas within homogeneous groups. For schemes M.1 and M.2 the groups above comprise areas having the same true sampling variance; for the scheme M.3 there are two groups, one for each level of the variance selected for the random effects; the overall figures are also reported.

M.1) The results for the normal setup are shown in Table 2. The figures for the absolute and relative measures of bias and mean squared prediction error indicate a certain stability over different levels of sampling error, with some larger discrepancies when $\psi = 1$. For comparison, in Table 3 we report an extract from Table 2 in Maiti et al. (2014), where the performance of the proposed model is assessed by simulation and contrasted to the method of Wang and Fuller (2003). Although with a different number of replications, there is a remarkable difference in the MSPEs under the three different approaches, showing that the method proposed in this paper outperforms the parametric approach. Compared to the bias-corrected method of Maiti et al. (2014), the proposed model carries a higher, downwards, bias, but the MSPEs indicate a clear advantage in terms of prediction error. The frequentist coverage of the prediction intervals, not reported in the cited paper, seems for the proposed method to vary with the level of the sampling variance, and is lower than the nominal level for the 95% intervals, but is still around the nominal value, except for $\psi = 1$. In light of the simulation results, it would be advisable to refer to the 99% credible intervals.

M.2) The results obtained under the model with skew-t-distributed random effects (see Table 4) are encouraging, giving estimates of the sampling variances that are close

ψ	bias	Rbias	MSPE	RMSPE	$\hat{\psi}$	C_{95}	C_{99}
1	-0.0067	-0.0023	0.3158	0.0546	1.0338	0.9233	0.9805
4	-0.0138	-0.0035	0.3937	0.0603	4.1331	0.9402	0.9880
16	-0.0133	-0.0038	0.4035	0.0603	16.5175	0.9499	0.9918

Table 2: Results from the simulation scheme of Wang and Fuller (2003) and Maiti et al. (2014), with $n = 36$ and $n_i = 9, \quad i = 1, \ldots, n$. The selected sampling variances are shown in the first column. The last two columns report frequentist coverage of the 95 and 99% credible intervals.

ψ	bias_M	bias_{WF}	MSPE_M	MSPE_{WF}
1	0.002	0.001	0.564	0.623
4	0.003	0.002	0.937	1.131
16	0.002	-0.001	1.101	1.359

Table 3: Results from Table 2 in Maiti et al. (2014); the selected sampling variances are shown in the first column. The suffix WF refers to the method in Wang and Fuller (2003) while M refers to the method of Maiti et al. (2014).

ψ	bias	Rbias	MSPE	RMSPE	$\hat{\psi}$	C_{95}	C_{99}
1	0.0096	-0.0026	0.5885	0.0736	1.0295	0.9258	0.9765
4	0.0121	-0.0052	1.0349	0.0950	4.1043	0.9417	0.9821
16	0.0083	-0.0072	1.2147	0.1007	16.4930	0.9544	0.9865

Table 4: Results from the simulation scheme M.2: errors generated from a skew t distribution with unit variance.

ψ	bias	Rbias	MSPE	RMSPE	$\hat{\psi}$	C_{95}	C_{99}
1	0.0058	-0.0015	0.4493	0.0642	1.0302	0.9399	0.9819
4	0.0040	-0.0031	0.6892	0.0775	4.1184	0.9592	0.9889
16	0.0041	-0.0039	0.7434	0.0786	16.4522	0.9728	0.9938

Table 5: Results from the simulation scheme M.2: errors generated from a skew t distribution with $\omega = 1$ and variance about 0.5.

to the target values, and moderate amount of relative bias, comparable to what obtained under the normal setup. As expected, the prediction error increases, with RMSPE as high as 1.7 times the corresponding figures for the normal case; yet the maximum RMSPE amounts to about 10%. Compared to the normal case, the coverage corrupts slightly, especially when $\psi = 1$, with values below the nominal level, but still around 98% for the 99% credible intervals. To assess the coverage of the credible intervals, data were also drawn for comparison from a skew-t distribution with variance 0.5. Results, reported in Table 5, indicate an improvement in bias, MSPE and coverage, showing that the model performance improves when the variance of the random effects distribution is lower than the sampling variance, a result analogous to the normal case. This is a case when the model is relatively more informative than the direct estimator and an advantage can be

σ^2_ν	bias	Rbias	MSPE	RMSPE	$\hat{\psi}$	C_{95}	C_{99}
25	-0.0074	-0.0155	1.1605	0.3088	1.0380	0.8707	0.9524
1	0.0030	-0.0032	0.6203	0.0796	1.0319	0.9254	0.9764
overall	-0.0033	-0.0067	0.7568	0.1842	1.0342	0.9117	0.9704

Table 6: Results for the simulation scheme in M.3: errors generated from a mixture of normals as in Sinha and Rao, with an 11% of outliers. The last line reports the assessment for the whole simulated sample. The first column reports the variance of the random components in each subgroup.

σ^2_ν	bias	Rbias	MSPE	RMSPE	$\hat{\psi}$	C_{95}	C_{99}
25	0.0028	-0.0198	1.3126	0.5710	1.0581	0.8804	0.9578
1	-0.0044	-0.0031	0.6680	0.0824	1.0335	0.9288	0.9770
overall	0.0046	-0.0107	0.8376	0.2294	1.0361	0.9167	0.9722

Table 7: Results for the simulation scheme in M.3. Errors generated from a mixture of normals as in Sinha and Rao, with a 25% of outliers. The last line reports the assessment for the whole simulated sample.

expected from using a model based approach for small area estimation. Consequently, the proposed nonparametric model may be useful when it is expected that the model is reliable. Even though there is evidence of some undercoverage, the MSPE indicate that the true area means are not far off the interval bounds.

M.3) The third setup considered for modelling the random effects is a mixture of normals with mean zero and different variances (1 and 25, respectively). This setup is analogous to the one analysed in Sinha and Rao (2009) and represents a situation in which the model is weak and the direct estimator is comparatively more reliable. The assessment was performed for the whole simulated sample, and separately for the data from each submodel. The overall behaviour is similar, although with lower performances, to the skew-t setup (first line of Table 4), which however refers to a balanced case where the sampling and error variances have equal weight. Table 6 presents the results for the normal mixture with $\gamma = 0.11$, whereas Table 7 contains the figures for the same setup, but with $\gamma = 0.25$. When the proportion of outlying areas is $\gamma = 0.11$ the performance of the submodel with unit sampling variance is similar to the setup M.2, with analogous undercoverage and slightly higher MSPE. When $\gamma = 0.25$ this effect is widened and the whole model performance is slightly worse. As may be expected, the subset of units characterised by large sampling variance has high prediction error and low coverage. For these units, the performance is poor in both configurations, with considerable increase in the MSPE with respect to both the normal and the skew-t setup, and worsening behaviour when the proportion of outliers increases. This can also be ascribed to the fact that this is a situation in which the model is very weak and the direct estimator is comparatively more, and increasingly, reliable. The overestimation of the true ψ seems to increase with the proportion of outliers, and so do the bias and prediction error. While on a smaller scale, the same increase can be noticed also for the non-outlying areas. In fact, for the contamination pattern considered, the presence of outliers affects the whole model, to an extent that depends on the proportion of outliers, but clearly carries the strongest consequences on the outlying units.

4 Application

Next, we consider for illustration two simple applications to real data, that have been extensively analysed in the literature. In particular they are used in You and Chapman (2006) and Dass et al. (2012), to which we compare our results.

We first consider the milk data set studied in Arora and Lahiri (1997). Table 8 contains a comparison with results presented in Table 3 of You and Chapman (2006). Although the sample sizes are all large, and the CV not exceedingly high for this data set, this example is interesting for testing the proposed model because of the auxiliary information introduced, which amounts to a classification of areas into a number of major areas (four in our application, as in the analysis of You and Chapman, 2006). Indeed, as discussed in Section 1, one of the features of the proposed approach is to explore possible configurations of the sampled areas into clusters, thus producing alternative aggregations of “similar” areas, useful for the purpose of borrowing strength in small area prediction, while accounting for uncertainty about sampling variances. To understand the role of the nonparametric formulation of the random effects, the model without covariates is also estimated; results are shown in the last column of Table 8.

For comparability with the model fitted by You and Chapman (2006), a flat Gamma prior ($a_0 = 0.0001, b_0 = 0.0001$) was used for the sampling variance parameters ψ_i. To assist the prior elicitation on M, the distribution of the number of clusters (Antoniak, 1974), integrated with respect to the Gamma prior, was considered (see Dorazio, 2009):

$$\pi(k|m, a_2, b_2) = \frac{{b_2}^{a_2}|S_{m,k}|}{\Gamma(a_2)} \int_0^\infty \frac{\Gamma(M)}{\Gamma(M+m)} M^{k+a_2-1} e^{-b_2 M} dM \tag{13}$$

where $|S_{m,k}|$ is the unsigned Stirling number of the first kind.

In Figure 1 the graph of such distribution under various choices of (a_2, b_2) is reported, for $m = 43$ as with the milk dataset; numerical integration of (13) was performed. The induced distribution over K is not completely flexible. As suggested by Dorazio (2009), practical choice of the pair (a_2, b_2) may be performed with the aim of spreading the distribution (13) as much as possible. For this example, a Gamma prior with shape 0.1 and rate 0.004 was selected for both models. Such a choice seems to be a good compromise in terms of prior mean and spread of the distribution, see Table 9. Although this prior gives a larger weight to $K = 1$, the number of clusters concentrates on larger values, as reported in Figure 2.

Under the model with covariates, point estimates are robust to different choices of the hyperparameters for M and σ^2_ν. The number of clusters may vary according to different prior choices; nonetheless, point estimates are stable and, as a consequence, the

$\hat{\theta}_i^{HB}$	SD	CV	$\hat{\theta}_i^{YC}$	SD	CV	$\tilde{\theta}_i$	SD	CV	$\tilde{\theta}_i^{wo}$	SD	CV
1.020	0.113	0.111	1.020	0.112	0.110	1.047	0.086	0.082	1.122	0.124	0.110
1.045	0.072	0.069	1.045	0.072	0.069	1.059	0.061	0.057	1.147	0.054	0.047
1.065	0.073	0.069	1.063	0.074	0.069	1.066	0.061	0.057	1.152	0.049	0.042
0.767	0.095	0.124	0.770	0.096	0.125	0.717	0.111	0.155	0.698	0.058	0.082
0.849	0.096	0.113	0.851	0.097	0.114	0.834	0.135	0.162	0.723	0.066	0.092
0.975	0.103	0.106	0.974	0.103	0.105	1.022	0.095	0.093	1.024	0.190	0.186
1.058	0.125	0.118	1.057	0.123	0.117	1.062	0.089	0.084	1.150	0.102	0.089
1.097	0.099	0.090	1.096	0.099	0.090	1.150	0.105	0.091	1.144	0.079	0.069
1.219	0.121	0.099	1.213	0.121	0.100	1.218	0.100	0.082	1.183	0.081	0.068
1.192	0.122	0.102	1.189	0.122	0.103	1.206	0.101	0.084	1.174	0.077	0.066
0.793	0.094	0.119	0.800	0.097	0.122	0.755	0.109	0.144	0.694	0.059	0.085
1.213	0.131	0.108	1.208	0.130	0.108	1.215	0.106	0.087	1.183	0.094	0.079
1.206	0.112	0.093	1.204	0.112	0.093	1.214	0.096	0.079	1.178	0.071	0.060
0.984	0.107	0.109	0.986	0.107	0.108	0.975	0.153	0.157	0.800	0.161	0.202
1.187	0.105	0.088	1.186	0.104	0.087	1.195	0.088	0.074	1.156	0.074	0.064
1.156	0.104	0.090	1.157	0.103	0.089	1.173	0.097	0.083	1.140	0.095	0.083
1.225	0.101	0.083	1.225	0.099	0.080	1.216	0.078	0.064	1.168	0.060	0.051
1.284	0.115	0.089	1.281	0.114	0.089	1.237	0.087	0.071	1.187	0.087	0.073
1.234	0.101	0.082	1.233	0.100	0.081	1.220	0.078	0.064	1.171	0.062	0.053
1.233	0.110	0.089	1.233	0.109	0.089	1.217	0.083	0.068	1.169	0.071	0.060
1.092	0.097	0.089	1.094	0.097	0.089	1.106	0.118	0.107	1.074	0.157	0.147
1.192	0.128	0.107	1.192	0.127	0.107	1.188	0.111	0.094	1.095	0.167	0.153
1.122	0.103	0.092	1.124	0.101	0.090	1.142	0.110	0.096	1.105	0.135	0.122
1.221	0.113	0.092	1.221	0.110	0.090	1.211	0.086	0.071	1.164	0.074	0.063
1.193	0.086	0.072	1.193	0.086	0.072	1.203	0.072	0.060	1.163	0.050	0.043
0.761	0.091	0.120	0.760	0.091	0.120	0.757	0.067	0.089	0.734	0.083	0.112
0.763	0.092	0.120	0.762	0.092	0.120	0.758	0.067	0.089	0.737	0.086	0.117
0.734	0.125	0.170	0.732	0.123	0.167	0.732	0.103	0.141	0.823	0.196	0.238
0.768	0.085	0.110	0.766	0.083	0.109	0.760	0.062	0.082	0.732	0.071	0.097
0.615	0.076	0.124	0.618	0.075	0.122	0.630	0.101	0.160	0.675	0.067	0.099
0.769	0.122	0.158	0.769	0.120	0.156	0.753	0.090	0.119	0.894	0.215	0.240
0.795	0.119	0.150	0.791	0.116	0.147	0.765	0.085	0.111	0.961	0.217	0.225
0.771	0.091	0.118	0.768	0.091	0.118	0.761	0.066	0.086	0.740	0.088	0.119
0.612	0.060	0.099	0.614	0.061	0.100	0.624	0.088	0.141	0.671	0.064	0.095
0.701	0.085	0.121	0.701	0.084	0.120	0.725	0.077	0.106	0.710	0.054	0.076
0.757	0.094	0.123	0.757	0.093	0.123	0.756	0.070	0.093	0.737	0.090	0.122
0.534	0.080	0.150	0.538	0.081	0.150	0.500	0.113	0.226	0.633	0.103	0.162
0.744	0.096	0.129	0.741	0.096	0.129	0.748	0.073	0.098	0.727	0.078	0.107
0.754	0.082	0.108	0.753	0.081	0.108	0.756	0.062	0.082	0.724	0.057	0.079
0.768	0.088	0.115	0.768	0.088	0.114	0.761	0.065	0.085	0.733	0.075	0.103
0.747	0.071	0.095	0.746	0.071	0.096	0.754	0.057	0.076	0.722	0.051	0.070
0.801	0.093	0.116	0.800	0.092	0.115	0.773	0.067	0.087	0.789	0.147	0.186
0.682	0.094	0.139	0.683	0.093	0.137	0.706	0.094	0.134	0.704	0.061	0.087

Table 8: Comparison of results from You and Chapman (2006) with those obtained under the proposed method. The first portion of the table refers to the HB version of the Fay–Herriot model (known sampling variances and normal components); the second portion refers to the parametric model proposed by You and Chapman (2006) with unknown variances; the second half of the table refers to our semiparametric model with uncertainty on sampling variances, with and without covariates. SD refers to the posterior standard deviation and CV is obtained by the ratio of posterior standard deviation to posterior mean.

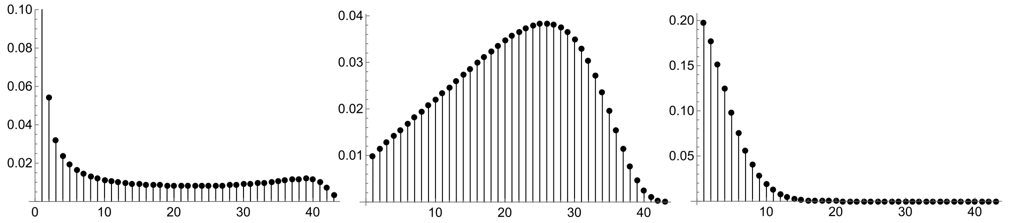

Figure 1: Prior distribution for K, integrated over the Gamma prior on M, for the milk data example. Left panel: $a_2 = 0.1, b_2 = 0.004$; centre: $a_2 = 1, b_2 = 0.04$; right: $a_2 = 1, b_2 = 1$.

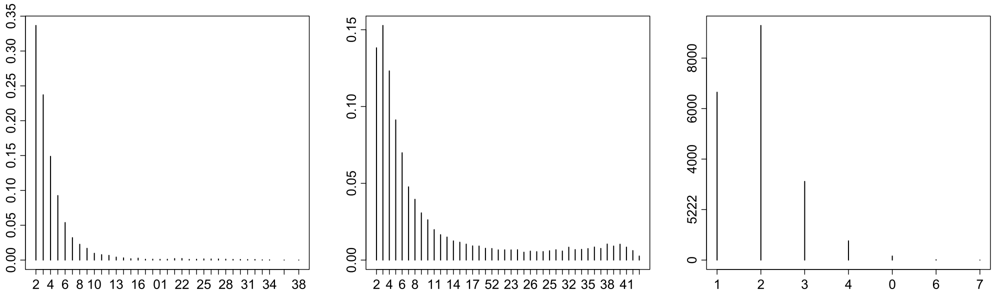

Figure 2: Posterior distribution of the number of clusters for the milk and corn data examples. Left panel: milk data, model without covariates; centre: milk data, model with covariates; right panel: corn data.

Gamma parameters	mean	sd
$a_2 = 0.1, b_2 = 0.004$	9.7	12.8
$a_2 = 1, b_2 = 0.04$	21.1	9.3
$a_2 = 1, b_2 = 1$	4.0	2.8

Table 9: Prior mean and variances for the number of clusters, K, under a Gamma prior for M for n=43 as for the milk data.

predictive criteria do not vary largely; the corresponding predictive scores are reported in Table 10. The estimated variances are also very stable.

As far as the model without covariates is concerned, the role of hyperparameters has a larger impact, as the number of clusters has a direct role in determining the small area predictions. Having selected a Gamma$(0.1, 0.004)$ prior for M, predictive measures were used to identify a suitable model, that is, suitable prior hyperparameters for σ_ν^2 within the range of values judged reasonable for the problem. Based on the predictive criteria, the pair $a_1 = 1, b_1 = 1$ was selected, still resulting in a unit prior mean for σ_ν^2.

The results for the semi- and the nonparametric model indicate lower CV if compared with the standard HB and the parametric model proposed in You and Chapman (2006)

	DIC	WAIC	LPML
Model by You and Chapman	-413.60	-425.91	187.28
Proposed model, with covariates	-418.67	-431.38	189.94
Proposed model, without covariates	-411.16	-426.44	180.01

Table 10: Milk data: estimated predictive criteria for comparison of the model by You and Chapman (2006) with the proposed method, with and without covariates.

except for a few areas. In this application, the estimated predictive criteria indicate that the proposed extended Fay–Herriot model is preferred over its parametric counterpart. It is interesting to note the good behaviour of the (fully nonparametric) model without covariates, whose predictive performance is not far from that of the models that explicitly introduce the categorical covariate: Table 10 confirms that the sole introduction of DP random effects makes the no-covariate model close to the parametric model with covariates estimated by You and Chapman (2006). Also, the application reveals that the posterior mean of the number of classes is 10.2 in the model with covariates and 4.3 for the model without covariates, showing the role of the clustering mechanism.

We also apply the proposed model to the Corn data set first analysed in Battese et al. (1988). For this data set, the very small area sizes make the direct estimates and the sampling variances highly unreliable and to be supplemented by a suitable statistical model. Based on these data, Dass et al. (2012) provide results that indicate strong superiority of their method compared to other proposals, including Wang and Fuller (2003), showing a dramatic reduction in the interval widths. Except for the nonparametric specification of the random effects, the proposed method and the one in Dass et al. (2012) share the model formulation to a large extent. However Dass et al. (2012) estimate the model parameters following an empirical Bayesian approach, whereas here the hyperparameters were elicited subjectively. A Gamma$(1, 0.25)$ prior was chosen for M; the implied prior mean for K is 4.1, with standard deviation 1.89. The posterior concentrates on a small number of clusters, see the right panel in Figure 2 (posterior mean: 1.9). Small area predictions are not sensitive to specification of M, but tend to vary with the hyperprior on ψ_i, in light of the very small area sizes for this data set. The extremely small area size makes use of a flat prior for ψ_i not advisable. The specification $a_0 = 1, b_0 = 0.001$ was used, implying a prior mean for the standard deviations of 31.6, which is compatible with the observed data range (see Table 11), and with the parameter estimates obtained by Dass et al. (2012) for the same data set.

The small area predictions provided by the method proposed in this paper are compatible with the results reported in Table 7 of Dass et al. (2012), with intervals of length comparable to those obtained under method I in the cited table; note however that compared to Dass et al. (2012) there is a greater variation in the interval widths, that do not depend so closely on area size. For some areas the proposed intervals are larger than the ones in Dass et al. (2012), which is not surprising since they minimize an expected loss function defined in terms of the interval length. Besides that, a larger interval length can be ascribed to the fact that the proposed intervals incorporate all sources of variation and rely on a nonparametric assumption. The variable borrowing of strength also impacts on the proposed intervals, that tends to be larger for those

	Semiparametric method				Method by Dass et al.					
	$\tilde{\theta}_i$	2.5%	97.5%	length	$\hat{\theta}_i^D$	2.5%	97.5%	length	$\sqrt{S_i^2}$	n
Franklin	145.97	109.89	165.46	55.57	131.81	104.09	159.37	55.29	5.70	3
Pocahontas	105.43	65.15	144.08	78.93	108.73	80.90	136.44	55.54	43.41	3
Winnebago	118.05	93.15	147.41	54.27	109.06	81.43	136.65	55.22	30.55	3
Wright	126.33	102.18	163.71	61.53	131.61	103.74	159.56	55.83	54.00	3
Webster	108.98	72.08	141.42	69.33	113.15	92.81	133.35	40.54	21.30	4
Hancock	125.12	102.16	153.01	50.85	129.43	111.78	147.19	35.41	15.66	5
Kossuth	118.79	99.60	142.01	42.41	121.01	103.45	138.63	35.18	12.11	5
Hardin	133.05	96.49	169.05	72.56	130.26	112.37	148.11	35.74	36.81	5

Table 11: Comparison between credible intervals obtained from the proposed method and confidence intervals obtained in Dass et al. (2012) for the modified Crop Area level data from Battese et al. (1988).

areas, that are not shrunk towards a common mean. At any rate, in light of the very small number of areas, and units within areas, all conclusions must be taken with care, especially for the proposed semiparametric method.

5 Final Remarks

This contribution investigates a semiparametric generalised Fay–Herriot model that allows for unknown sampling variance and specifies the random effects nonparametrically through a DP prior. The model formulation allows to relax parametric assumptions on the random effects by relying on a Dirichlet process prior, thus representing a means to account for outliers and overcome the problem of model misspecification. Thanks to the clustering property of the DP, the model is capable to uncover structure in the data that allows potential gains in estimation efficiency without incurring in a consistency–efficiency tradeoff. Indeed the clustering property allows to explore in a data driven way subgroups of areas that may be characterised by different features: see Articus and Burgard (2014) for a practical example in a real data setting. In light of such property, the proposed model may also provide an improvement over the direct estimator even in the absence of covariates.

Alternative approaches, specifically allowing for outliers, are finite mixtures of normal models and scale mixtures of normals, investigated e.g. in Maiti (2003) and Datta and Lahiri (1995); compared to classical mixture models, the number of clusters here is not fixed, and fixed effects parameter are estimated by pooling information from all possible clusters of areas.

Moreover the hierarchical formulation of the model permits to include as a further stage in the hierarchy an explicit specification of the sampling variances, usually assumed known, or modelled through a separate and exogenous step. This choice allows shrinkage of variances as well as of area means. As a consequence, under the proposed approach inference on small area quantities incorporates all sources of variation, thus providing a more comprehensive account of uncertainty. As a by-product, the method also produces smoothed estimates of the sampling variances. More elaborate models for smoothing the sampling variances, possibly depending on specific covariates, may be accommodated for in the hierarchical model. Although the distribution of sampling variances is not χ^2-distributed unless simple random sampling and normal population are assumed, simulation results under non-normal schemes indicate that the proposed Bayesian model is still reliable even under such model misspecification.

The simulation experiments performed seem to indicate that the proposed nonparametric model may be useful when it is expected that the model is reliable. Even though there is evidence of a little undercoverage, the MSPE indicate that the true area means are not far off the interval bounds and that the method makes efficient use of sample information. Alternative measures of coverage like the one proposed in Rossi (2014, p. 134) might be used to take into account both coverage and the distance from the true small area means.

Whenever availability of analogous published results made comparisons possible, the prediction error appears to be lower than that attained by other procedures. The poor coverage performances registered under the mixture of normals model may be ascribed to the fact that this is an instance where the model itself is highly unreliable compared to the direct estimator.

As commented, when the focus is point prediction of small area values, the parametric model is robust against departures from the normality assumption. Therefore one can expect that the predictive measures would be similar for the parametric and the non parametric model. Yet, under all the simulation schemes, predictive information measures favour the nonparametric model over its fully parametric counterpart. At the same time, analysis of the frequentist properties of the predictor obtained under the model with nonparametric random effects highlights a sensible reduction in MSPEs. In the applications, the proposal seems to produce typically smaller CVs than the analogous procedure in You and Chapman (2006); in comparison to the bias corrected procedure of Maiti et al. (2014) we noticed in simulations substantially lower MSPEs.

From the above findings we can conclude that the model effectively achieves flexibility in modelling the random effects and a more realistic representation of the uncertainty, without increasing the dimensionality of the problem.

Despite the focus in the application has been on prediction of small area means, other quantities might be of interest, such as ranks or CDF. In this respect, the robustness found in McCulloch and Neuhaus (2011) may not apply.

Finally, the same framework can also be extended to cover unit level and nonlinear models.

References

Antoniak, C. E. (1974). “Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems.” *Annals of Statistics*, 2: 1152–1174. MR0365969. 732, 743

Arora, V. and Lahiri, P. (1997). "On the superiority of the Bayesian method over the BLUP in small area estimation problems." *Statistica Sinica*, 7: 1053–1064. MR1488659. 735, 743

Articus, C. and Burgard, J. P. (2014). "A Finite Mixture Fay Herriot-type model for estimating regional rental prices in Germany." Technical report, University of Trier, Department of Economics. 747

Azzalini, A. and Capitanio, A. (2003). "Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 367–389. MR1983753. doi: http://dx.doi.org/10.1111/1467-9868.00391. 738

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). "An error-components model for prediction of county crop areas using survey and satellite data." *Journal of the American Statistical Association*, 83: 28–36. 746, 747

Bell, W. R. (2008). "Examining Sensitivity of Small Area Inferences to Uncertainty About Sampling Error Variances." In *ASA Proceedings of the Section on Survey Research*, 871–876. John Wiley. 731

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson distributions via Pólya urn schemes." *Annals of Statistics*, 1: 353–355. MR0362614. 739

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis*, 1(4): 651–673. MR2282197. doi: http://dx.doi.org/10.1214/06-BA122. 737

Chakraborty, A., Datta, G. S., and Mandal, A. (2016). "A two-component normal mixture alternative to the Fay–Herriot model." *Statistics in Transition new series*, 17(1): 67–90. 731, 732, 734

Dass, S. C., Maiti, T., Ren, H., and Sinha, S. (2012). "Confidence interval estimation of small area parameters shrinking both means and variances." *Survey Methodology*, 38: 173–187. 733, 735, 736, 737, 743, 746, 747

Datta, G. and Ghosh, M. (2012). "Small area shrinkage estimation." *Statistical Science*, 27(1): 95–114. MR2953498. doi: http://dx.doi.org/10.1214/11-STS374. 731

Datta, G. S. and Lahiri, P. (1995). "Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers." *Journal of Multivariate Analysis*, 54(2): 310–328. MR1345542. doi: http://dx.doi.org/10.1006/jmva.1995.1059. 731, 734, 747

Datta, G. S. and Lahiri, P. (2000). "A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems." *Statistica Sinica*, 10(2): 613–627. MR1769758. 731

Diao, L., Smith, D. D., Datta, G. S., Maiti, T., and Opsomer, J. D. (2014). "Accurate confidence interval estimation of small area parameters under the Fay–Herriot model." *Scandinavian Journal of Statistics*, 41(2): 497–515. MR3207183. doi: http://dx.doi.org/10.1111/sjos.12045. 733, 737

Dick, P. (1995). "Modelling net undercoverage in the 1991 Canadian Census." *Survey Methodology*, 21: 45–54. 731

Dorazio, R. M. (2009). "On selecting a prior for the precision parameter of Dirichlet process mixture models." *Journal of Statistical Planning and Inference*, 139(9): 3384–3390. MR2538090. doi: http://dx.doi.org/10.1016/j.jspi.2009.03.009. 735, 743

Escobar, M. D. and West, M. (1994). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90: 577–588. MR1340510. 735, 737

Fabrizi, E. and Trivisano, C. (2010). "Robust linear mixed models for Small Area Estimation." *Journal of Statistical Planning and Inference*, 140: 433–443. MR2558375. doi: http://dx.doi.org/10.1016/j.jspi.2009.07.022. 734

Fay, R. and Herriot, R. (1979). "Estimates of income for small places: an application of James–Stein procedures to Census Data." *Journal of the American Statistical Association*, 74: 269–277. MR0548019. 729, 730

Ferguson, T. S. (1973). "A Bayesian Analysis of some nonparametric problems." *Annals of Statistics*, 1(2): 209–230. MR0350949. 732

Geisser, S. and Eddy, W. F. (1979). "A predictive approach to model selection." *Journal of the American Statistical Association*, 74(365): 153–160. MR0529531. 737

Hawala, S. and Lahiri, P. (2010). "Variance modeling in the U.S. small area income and poverty estimates program for the American community survey." In *Proceedings of the American Statistical Association, Survey Methods Section, Denver, Colorado*. Alexandria, VA: American Statistical Association. 731

Liu, J. S. (1996). "Nonparametric hierarchical Bayes via sequential imputations." *Annals of Statistics*, 24(3): 911–930. MR1401830. doi: http://dx.doi.org/10.1214/aos/1032526949. 736

Lo, A. Y. (1984). "On a class of Bayesian nonparametric estimates. I. Density estimates." *Annals of Statistics*, 12(1): 351–357. MR0733519. doi: http://dx.doi.org/10.1214/aos/1176346412. 736

Maiti, T. (2003). "Modelling small area effects using mixture of Gaussians." *Sankhyā: The Indian Journal of Statistics*, 65(3): pp. 612–625. MR2060610. 733, 734, 736, 747

Maiti, T., Ren, H., and Sinha, S. (2014). "Prediction error of small area predictors shrinking both means and variances." *Scandinavian Journal of Statistics*, 41(3): 775–790. MR3249428. doi: http://dx.doi.org/10.1111/sjos.12061. 733, 736, 738, 740, 741, 748

Malec, D. and Müller, P. (2008). *A Bayesian semi-parametric model for small area estimation*, volume 3 of *Collections*, 223–236. Beachwood, Ohio, USA: Institute of Mathematical Statistics. MR2459227. doi: http://dx.doi.org/10.1214/074921708000000165. 732, 733

Maples, J., Bell, W., and Huang, E. (2009). "Small area variance modeling with application to county poverty estimates from the American community survey." In *Proceedings of the American Statistical Association, Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, 5056–5067. 736

McCulloch, C. E. and Neuhaus, J. M. (2011). "Misspecifying the shape of a random effects distribution: why getting it wrong may not matter." *Statistical Science*, 26(3): 388–402. MR2917962. doi: http://dx.doi.org/10.1214/11-STS361. 734, 748

Murugiah, S. and Sweeting, T. (2012). "Selecting the precision parameter prior in Dirichlet process mixture models." *Journal of Statistical Planning and Inference*, 142(7): 1947–1959. MR2903404. doi: http://dx.doi.org/10.1016/j.jspi.2012.02.013. 735

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. (2007). "Flexible random-effects models using Bayesian semi-parametric models: Applications to institutional comparisons." *Statistic in Medicine*, 26: 2088–2112. MR2364293. doi: http://dx.doi.org/10.1002/sim.2666. 736

Prasad, N. G. N. and Rao, J. N. K. (1990). "The estimation of the mean squared error of small-area estimators." *Journal of the American Statistical Association*, 85(409): 163–171. MR1137362. 731

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. Wiley Series in Survey Methodology. John Wiley & Sons, Inc., Hoboken, NJ, second edition. With a foreword by Graham Kalton. MR3380626. doi: http://dx.doi.org/10.1002/9781118735855. 730, 731

Rivest, L.-P. and Vandal, N. (2003). "Mean squared error estimation for small areas when the small area variances are estimated." In Roberts, G. and Bellhouse, D. (eds.), *Proceedings of the International Conference on Recent Advances in Survey Sampling, ICRASS 2002, Ottawa, Canada*. Statistics Canada. 731

Rossi, P. (2014). *Bayesian Non-and Semi-parametric Methods and Applications*. Princeton University Press. MR3288097. doi: http://dx.doi.org/10.1515/9781400850303. 739, 748

Sinha, S. K. and Rao, J. N. K. (2009). "Robust small area estimation." *Canadian Journal of Statistics*, 37(3): 381–399. MR2547205. doi: http://dx.doi.org/10.1002/cjs.10029. 731, 734, 739, 742

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian measures of model complexity and fit." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 583–639. MR1979380. doi: http://dx.doi.org/10.1111/1467-9868.00353. 737

Wang, J. and Fuller, W. A. (2003). "The mean squared error of small area predictors constructed with estimated area variances." *Journal of the American Statistical Association*, 98(463): 716–723. MR2011685. doi: http://dx.doi.org/10.1198/016214503000000620. 731, 732, 736, 738, 740, 741, 746

Watanabe, S. (2009). *Algebraic geometry and statistical learning theory*, volume 25 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge. MR2554932. doi: http://dx.doi.org/10.1017/CBO9780511800474. 737

Watanabe, S. (2010). "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." *Journal of Machine Learning Research*, 11: 3571–3594. MR2756194. 737

West, M., Müller, P., and Escobar, M. D. (1994). "Hierarchical priors and mixture models, with application in regression and density estimation." In Freeman, P. R. and Smith, A. F. M. (eds.), *Aspects of Uncertainty. A Tribute to D. V. Lindley*, 363–386. John Wiley & Sons. MR1309702. 737, 739

You, Y. and Chapman, B. (2006). "Small area estimation using area level models and estimated sampling variances." *Survey Methodology*, 32: 97–103. 732, 735, 736, 740, 743, 744, 745, 746, 748