

Bayesian Statistics for Small Area Estimation

V. Gómez-Rubio, N. Best, S. Richardson, G. Li

Department of Epidemiology and Public Health

Imperial College London

St. Mary's Campus, Norfolk Place

W2 1PG London - United Kingdom

Tel: +44 (0) 20 759 43302 - Fax: +44 (0) 20 740 22150

and Philip Clarke

Office for National Statistics, United Kingdom

Abstract. National statistical offices are often required to provide statistical information about characteristics of the population, such as mean income or unemployment rate, at several administrative or small area levels. Having good area level estimates is important because policies will often be based on this type of information.

In this paper we describe how Bayesian hierarchical models can help in the task of providing good quality small area estimates. Starting from direct estimates obtained from survey data, we describe a range of Bayesian hierarchical models that incorporate different types of random effects and show that these give improved estimates. Models that synthesise individual and aggregated information are considered as well. Finally, we highlight some additional applications that further exploit the estimates produced, such as the classification and ranking of areas and how to approach the problem of having no direct information in several areas.

Keywords: Bayesian Hierarchical Models; Missing Data; Policy Making; Small Area Estimation; Spatial Models

1. Introduction

Small Area Estimation (Rao, 2003) tackles the important statistical problem of providing reliable estimates of a target variable in a set of small geographical areas. The main difficulty is that it is nearly always impossible to measure the value of the target variable for all the individuals in the areas of interest and, hence, a survey is conducted to obtain a representative sample (Cochran, 1977). Surveys are often designed to include different sorts of data in order to

E-mail: v.gomezrubio@imperial.ac.uk

make the best use of them. The information collected is used to produce some direct estimate of the target variable that relies only on the survey design and the sampled data. Unfortunately, sampling from all areas can be expensive in resources and time. A more practical approach is to select a subset of areas where the survey is conducted; estimates for all areas are then produced using the sample and some additional auxiliary information which must be available for all small areas (Särndal et al., 1992).

Regression models are often used in this context to provide estimates for the non-sampled areas. The model is fitted on data from the sample and used together with additional information from auxiliary data to compute estimates for non-sampled areas. This method is often known as *synthetic* estimation (Gonzalez, 1973).

Regression estimators can be extended to include random effects, which can be estimated by Empirical Best Linear Unbiased Predictors (usually known as EBLUP estimators, see Robinson, 1991). Jiang and Lahiri (2006) provide a review on this topic for small area estimation. In addition to the usual covariates, mixed-effect models include random effects to model different types of individual variation and space and time interactions (Singh et al., 2005; Petrucci and Salvati, 2005). Different types of responses can be modelled with this framework, but the computations for non-Normal responses can be quite involved and techniques such as Penalised Quasi-Likelihood are required (Breslow and Clayton, 1993). EURAREA Consortium (2004) make a summary of direct and model-based likelihood-based small area estimators for several target variables for different national datasets from different countries.

The potential of spatial and spatio-temporal modelling in Small Area Estimation has been addressed by Jiang and Lahiri (2006) and the discussion therein. The two main benefits that they point out are the possibility of borrowing information from neighbouring areas when estimating spatially-correlated random effects and improving estimation in non-sampled areas. Petrucci and Salvati (2005) have also addressed the use of spatial random effects to produce improved Small Area estimates.

Bayesian alternatives of both the non-spatial and spatial mixed effects models for Small Area Estimation have been proposed (see, for example, Datta and Ghosh (1991), Ghosh et al. (1998), and Rao (2003) for a recent review). In particular, Bayesian small area spatial modelling has already been successful in other similar contexts, such as the estimation of the rate of disease in different geographic regions (Best et al., 2005). Complex mixed-effects and correlation between areas can be easily handled and modelled hierarchically in different layers of the model. For example, Besag et al. (1991) propose a spatial model in which the area variation not explained by the available covariates is split into two components: one which is unstructured (and independent) for each

area and another one which reflects likely correlation between neighbouring regions. Note that disease mapping applications are based on data available on disease status for all individuals in every area, whilst Small Area Estimation is usually based on a survey which provides access to a limited sample of the population under study and does not cover all areas, hence creating its own set of methodological issues.

Although implementation of complex Bayesian models requires computationally intensive Markov Chain Monte Carlo simulation algorithms (Gilks et al., 1995), there are a number of potential benefits of the Bayesian approach for small area estimation. It offers a coherent framework that can handle different types of target variable (e.g. continuous, dichotomous, categorical), different random effects structures (e.g. independent, spatially correlated), areas with no direct survey information, models to smooth the survey sample variance estimates, and so on, in a consistent way using the same computational methods and software whatever the model. Uncertainty about all model parameters is automatically captured by the posterior distribution of the small area estimates and any functions of these (such as their rank), and by the predictive distribution of estimates for small areas not included in the survey sample. Bayesian methods are particularly well suited to sparse data problems (for example, when the survey sample size per area is small) since Bayesian posterior inference is exact (modulo Monte Carlo simulation error associated with the estimation algorithms) and does not rely on asymptotic arguments. The posterior distribution obtained from a Bayesian model also provides a much richer output than the traditional point and interval estimates from a corresponding likelihood-based model. In particular, the ability to make direct probability statements about unknown quantities — for example, the probability that the target variable exceeds some specified threshold in each area — and to quantify all sources of uncertainty in the model, make Bayesian small area estimation well suited to informing and evaluating policy decisions.

In this paper, we aim to illustrate some of these points by considering a range of Bayesian hierarchical models for small area estimation that incorporate different types of spatial and non-spatial random effects structures. We compare the predictive accuracy of the small area estimates produced by each model, and focus in particular on two common problems faced by statistical bureaus when dealing with Small Area Estimation: (1) ranking and classification of areas, and (2) providing estimates in areas that have been left out of the survey. In a Bayesian framework, we will tackle the ranking problem by means of posterior ranks and posterior probabilities of being among a certain proportion of areas, whilst for the problem of missing direct information we will rely on observed area level covariates and the use of spatial random effects at different administrative levels to predict the missing data.

We consider the case in which the response variable is Normal, but all these techniques can be easily extended to the more general case. The use of the models and methods that we are proposing is illustrated through an example based on the average equivalised income per household in the 284 municipalities of Sweden. This data set has also been considered by EURAREA Consortium (2004), where many different types of likelihood-based estimators were computed and compared.

The paper is structured as follows. In Section 2 an introduction to typical survey design and synthetic estimators for Small Area Estimation is presented. Section 3 includes a general description of Bayesian methods and different types of models for Small Area Estimation. How to provide estimates for areas with no direct observations is considered in Section 4. The classification of areas for policy making is described in Section 5. An example with real data is shown in Section 6. Finally, some conclusions and remarks are presented in Section 7.

2. Direct and Linear Regression Estimators

2.1. Standard Regression Estimators

A common approach to the estimation of the mean value for an area involves the use of regression methods. The link between direct estimation and linear regression is as follows. We will consider the case of estimating area level averages, but the case of area totals can be worked out similarly. Assuming that an area level *direct* estimate $\hat{Y}_{D,i}$ (Rao, 2003) has been computed for area i , we can combine this estimate with linear regression by using the following Fay-Herriot model (Fay and Herriot, 1979):

$$\hat{Y}_{D,i} = \mu_i + e_i \quad (1)$$

where μ_i is the true area mean and e_i is a random term which reflects the variation of a direct estimator $\hat{Y}_{D,i}$ around the mean and which we will assume Normally distributed with zero mean and variance V_i^2 . In practice, V_i^2 is replaced by an estimate \hat{V}_i^2 . Here we have taken \hat{V}_i^2 equal to the variance of the direct estimator (often termed the *design variance*; Särndal et al., 1992). Alternatively, area level variances can be smoothed, for example using generalized variance functions (Jiang and Lahiri, 2006, see)[page 6], which should yield more stable estimates when the within-area sample size is small. Note that, rather than follow this 2-stage approach, in Section 3, we propose a Bayesian model with a hierarchical structure on both the small area means *and variances*, that smooths the sample variances as an integral part of the model fitting.

Standard regression techniques can then be employed to model the mean μ_i on area level covariates \bar{X}_i . These covariates are the area average of the

individual values x_{ij} over the population: $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$.

Hence, the mean is modelled as

$$\mu_i = \alpha^* + \bar{X}_i \beta^*$$

and the coefficients α^* and β^* can be estimated using typical model fitting algorithms.

Note that when the direct estimators are missing for some areas and only the values of \bar{X}_i are available, a *synthetic* estimator (Rao, 2003) can be computed as

$$\hat{\mu}_{S,i} = \hat{\alpha}^* + \bar{X}_i \hat{\beta}^* \tag{2}$$

where the value of the regression coefficients $\hat{\alpha}^*$ and $\hat{\beta}^*$ are computed using the data available from other areas.

Alternatively, a unit level version of this model can be fitted by regressing y_{ij} on the unit level covariates \mathbf{x}_{ij}

$$\begin{aligned} y_{ij} | \mu_{ij}, \sigma_i^2 &\sim N(\mu_{ij}, \sigma_i^2) \\ \mu_{ij} &= \alpha + \mathbf{x}_{ij} \beta \end{aligned}$$

Note that the coefficients for this model may be different from those estimated for the aggregate model, hence the different notation for the regression coefficients (see Section 2.2). The area level average can then be computed by averaging over all the values of μ_{ij} :

$$\mu_i = \sum_{j=1}^{N_i} \frac{\mu_{ij}}{N_i} = \alpha + \bar{X}_i \beta \tag{3}$$

Hence, a small area estimate that combines aggregated and individual information by means of the fitted values of α and β from the unit level model is

$$\hat{\mu}_{s,i} = \hat{\alpha} + \bar{X}_i \hat{\beta} \tag{4}$$

Synthetic estimators based on regression models described above have been widely used in statistical bureaus to provide small area estimates for areas not included in the sample (see, for example, Heady et al., 2003). This topic is further discussed in Section 4.

2.2. Area vs. unit level models

In principle, the estimators $\hat{\mu}_{S,i}$ and $\hat{\mu}_{s,i}$ look very similar, but they often provide different results given that the estimates of the parameters of the regression model are computed in a different way. The compatibility between the estimates of the coefficients of the covariates between area and unit level models is not always fulfilled (Greenland and Robins, 1994). We would expect to have similar fitted values of the coefficients in both cases, but when using aggregated data we may observe some bias in the estimates, often referred to as ecological bias. This bias may happen, for example, when there is confounding in the covariates at the individual level and the aggregated covariates do not explain it. Hence, we have to be careful not to interpret the coefficients causally as measuring individual level effects.

The choice of model will depend on what kind of data we have. Aggregated data is usually easier to obtain, as different statistical bureaus and institutions regularly produce volumes with area level statistical data.

3. Bayesian Hierarchical Models

3.1. Unit level models

First of all, we will consider the case in which individual level information on the target variable and covariates from the survey sample is available in all areas. Models with aggregated data are considered in the next section. For continuous variables, the response (possibly after appropriate transformation) can be modelled using a Normal distribution (Rao, 2003):

$$\text{MODEL 1} \quad y_{ij} | \mu_{ij}, \sigma_e^2 \sim N(\mu_{ij}, \sigma_e^2) \quad (5)$$

where μ_{ij} is the true value of the target variable of individual j in the sample from area i and σ_e^2 reflects the individual sampling variation which we will assume the same in all areas for the time being (i.e., $\sigma_i^2 = \sigma_e^2$).

Although covariates are typically introduced to model dependence between the mean and some explanatory factors, it is likely that some residual variance will remain unexplained by the covariates. This can be accounted for effectively by including random effects in the model. These random effects capture the unobserved patterns such as spatial dependence and between area variation. In particular, we consider the following random effects regression model for the unit level means:

$$\mu_{ij} = \alpha + \mathbf{x}_{ij}\beta + u_i + v_i \quad (6)$$

which leads to an area-level mean model:

$$\mu_i = \sum_j \frac{\mu_{ij}}{N_i} = \alpha + \bar{X}_i \beta + u_i + v_i, \quad (7)$$

Here α is the intercept of the model, β is the vector of coefficients of the covariates \mathbf{x}_{ij} , u_i is a random effect which accounts for area level variation and is distributed independently as

$$u_i | \sigma_u^2 \sim N(0, \sigma_u^2)$$

and v_i represents spatially correlated random effects. Initially, we consider the intrinsic Conditional Autoregressive (CAR) specification for v_i (Besag et al., 1991). Under this specification the conditional distribution of v_i given values v_{-i} in all the remaining areas only involves the neighbouring areas

$$v_i | v_{-i}, \sigma_v^2 \sim N\left(\sum_{j \in \delta_i} \frac{v_j}{|\delta_i|}, \frac{\sigma_v^2}{|\delta_i|}\right) \quad (8)$$

where δ_i is the set of neighbours of area i and $|\delta_i|$ the number of neighbours. In addition, we have added the constraint that the sum of the values of all the random effects v_i is zero to make the intercept and the random effects identifiable (see, Banerjee et al., 2004, pages 163–164).

As an alternative to this conditional specification, we can model the mean μ_{ij} by including spatial random effects w_i which are correlated according to the distance d_{kl} between two areas k and l (Diggle et al., 1998):

$$\mu_{ij} = \alpha + x_{ij} \beta + w_i \quad (9)$$

with w distributed as a Multivariate Normal

$$w | \Sigma \sim MVN(0, \Sigma); \Sigma_{kl} = \sigma_w^2 \exp\{-\phi d_{kl}\} \quad (10)$$

σ_w^2 is the variance at any given point and ϕ is a smoothing parameter that controls the scale of the correlation between areas.

Unlike model (7), we do not include a separate independent random effect u_i in model (9). The motivation for doing so in model (7) lies with the fact that the spatial dependence of the intrinsic CAR random effects (8) is pre-determined by the neighbourhood structure. Hence unstructured effects are also included to allow for Bayesian learning about the strength of spatial dependence in the data, via the relative contribution of the u_i and v_i to the posterior (Besag et al., 1991; Eberley and Carlin, 2000). In the case of model (9), Bayesian learning about the strength of spatial dependence of the w_i random effects takes place directly

via the posterior estimation of the correlation parameter ϕ in (10) ($\phi \rightarrow 0$ implies no spatial correlation). It is technically possible to include a separate independent random effect term in (9), but in practice this can result in poorly identified posterior distributions (Diggle et al., 2002).

For all these models, a sensible area level estimate is

$$\hat{\mu}_{b,i} = E_{\cdot|y}[\alpha + \bar{X}_i\beta + z_i] = \hat{\alpha} + \bar{X}_i\hat{\beta} + \hat{z}_i$$

$E_{\cdot|y}[\cdot]$ denotes posterior expectation and z_i denotes the random effects, which are specified as either $z_i = u_i + v_i$ (as in (6)) or $z_i = w_i$ (as in (9)). In this case, we compute the posterior means $\hat{\alpha}$, $\hat{\beta}$ and \hat{z}_i of α , β and z_i , respectively, assuming that area level averages of the covariates \bar{X}_i are available.

Model 1 is essentially the one proposed by Battese et al. (1988) modified to include different types of random effects. It assumes the same within-area variation (σ_e^2) for individuals in all areas, which is usually unrealistic because individual variation is likely to differ between areas. We therefore consider an extension of this model to the more general case in which we have a different variance σ_i^2 in each area:

$$\begin{array}{l} \text{MODEL 2} \\ y_{ij} | \mu_{ij}, \sigma_i^2 \sim N(\mu_{ij}, \sigma_i^2) \\ \sigma_i^2 \sim \text{vague prior} \end{array} \quad (11)$$

In this case, each area variance is estimated using the information only from the sample from area i . When the survey data within each area are sparse, this can lead to poor estimates of σ_i^2 . An alternative is to use a hierarchical structure on these variances to borrow information across areas to obtain more robust estimates. In particular, we can model the logarithm of the variances as follows:

$$\begin{array}{l} \text{MODEL 3} \\ y_{ij} \quad | \mu_{ij}, \sigma_i^2 \sim N(\mu_{ij}, \sigma_i^2) \\ \log(\sigma_i^2) \quad | \sigma^2 \sim N(0, \sigma^2) \\ \sigma^2 \sim \text{vague prior} \end{array} \quad (12)$$

This last model is similar in spirit to the use of generalised variance functions to smooth the area level variances, but is fully model-based, so that uncertainty about the variance estimates is reflected in the resulting posterior variance of the small area estimates.

Model 2 is essentially the same model proposed by Arora and Lahiri (1997) with random effects. They also proposed an area level model (see below) that allows for the area level variances to be estimated using a hierarchical model. Arora et al. (1997) use a similar unit level model with independent random effects and propose an empirical Bayes approach for the estimation of the random

effects. Kleffe and Rao (1992) approximate the MSE of this unit level model with different area level variances with common prior distribution.

3.2. Area level models

For area-level data, we can extend the model shown in equation (1) in order to include covariates and random effects. For example

$$\begin{aligned} \hat{Y}_{D,i} \mid \mu_i, \hat{V}_i^2 &\sim N(\mu_i, \hat{V}_i^2) \\ \mu_i &= \alpha^* + \bar{X}_i \beta^* + u_i^* + v_i^* \end{aligned} \tag{13}$$

where u_i^* and v_i^* are assigned independent normal and CAR distributions, respectively. Note that now area level variances of the area mean estimates are assumed to be known and, hence, represented by a square.

The estimate of the area mean is then provided by

$$\hat{\mu}_{B,i} = E_{\cdot|y}[\alpha^* + \bar{X}_i \beta^* + u_i^* + v_i^*] = \hat{\alpha}^* + \bar{X}_i \hat{\beta}^* + \hat{u}_i^* + \hat{v}_i^*$$

where, as before, the 'hat' notation denotes the posterior mean of the relevant parameter. For this model we have again written α^* and β^* instead of α and β (used for the unit level models) to highlight the fact that area level models can produce different estimates than unit level models. Similarly, the estimates of the random effects u_i^* and v_i^* are also likely to be different from those of u_i and v_i .

3.3. Prior distribution for the parameters in the model

For the intercept α and the coefficients of the covariates β we have employed improper flat priors but that induce proper posteriors. We use an inverted Gamma as a prior distribution for each one of the variances σ_u^2 , σ_v^2 , σ_w^2 , σ_i^2 (except in Model 3) and σ^2 . In order to give vague prior information and let the model learn from the data, we use small values for the parameters of the Gamma distribution. In particular, we have used 0.001 and 0.001 for the scale and shape parameters, respectively. Finally, the prior used for the parameter ϕ depends on the range and scale of measurement of the distances between small areas. In our application, distances range from 2.71 km to 1471.43 km, and we used a uniform between 0.01 and 5 to accommodate a reasonable range of values for the spatial correlation.

Gelman (2006) has recently shown that inverted gamma priors with small shape and scale parameters for random effects variances may induce spurious shrinkage, specially when the number of groups is small and there are few observations per group, and he proposes several alternatives. Following his suggestions, we have also used a half-Cauchy distribution on the standard deviation

of the random effects, but we have not found differences in the small area estimates.

3.4. Assessing the quality of the estimates

Evaluating the quality of small area estimates that have been obtained with area and unit level models can be difficult in practice. We are usually interested in their variances or mean square prediction errors (MSPE), but obtaining good estimators of MSPE is typically difficult for frequentist SAE methods since closed form expressions that account for the variability caused by estimation of the model parameters do not exist. Various approximate formula have been proposed, as well as jackknife and parametric bootstrap estimators (see, e.g. Jiang and Lahiri (2006) and Rao (2003) for reviews). On the other hand, the natural Bayesian measure of accuracy — the posterior variance of the small area estimates — is obtained automatically from the posterior output, and fully accounts for uncertainty about all the model parameters.

A criterion that can be used for model comparison in Bayesian statistics is the *Deviance Information Criterion* (DIC, Spiegelhalter et al., 2002). It is based on the deviance of the model penalised for model complexity and its interpretation is similar to the AIC, with models having smaller DIC being preferred.

For simulation studies, where the true mean value \bar{Y}_i is known, we can calculate the Relative Bias (RB) and the Relative Root Mean Squares Error (RRMSE) to evaluate the accuracy of the small area estimates. They are defined as

$$RB_i = \frac{1}{K} \sum_{k=1}^K \frac{(\hat{Y}_i^{(k)} - \bar{Y}_i)}{\bar{Y}_i}, \quad RRMSE_i = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{Y}_i^{(k)} - \bar{Y}_i)^2}}{\bar{Y}_i}$$

where k indexes the survey samples.

As global measures, we can take the Mean Absolute Relative Bias (MARB) and the Mean Relative Root Mean Square Error (MRRMSE):

$$MARB = \frac{1}{m} \sum_{i=1}^m |RB_i|, \quad MRRMSE = \frac{1}{m} \sum_{i=1}^m RRMSE_i \quad (14)$$

where m is the total number of areas. In this way, we can assess if the estimators are biased and if they are variable. Either of these latter two criteria can be used to decide on the best model. Better estimators will be those which produce smaller values of the MARB and MRRMSE.

We can also calculate the Relative Bias (RBvar) and Relative Root Mean Square Error (RRMSEvar) of the *variance* estimates in a simulation study, to assess how well they estimate the true error of the small area estimator. They are defined as

$$RBvar_i = \frac{1}{K} \frac{\sum_{k=1}^K (\hat{\text{var}}(\bar{Y}_i^{(k)}) - EMSE_i)}{EMSE_i},$$

$$RRMSEvar_i = \frac{\sqrt{\frac{1}{K} \sum_{k=1}^K (\hat{\text{var}}(\bar{Y}_i^{(k)}) - EMSE_i)^2}}{EMSE_i}$$

where $EMSE_i$ is the true empirical error

$$EMSE_i = \frac{1}{K} \sum_{k=1}^K (\hat{Y}_i^{(k)} - \bar{Y}_i)^2.$$

The Mean Absolute Relative Bias (MARBvar) and Mean Relative Root Mean Square Error (MRRMSEvar) of the variance estimates are then defined analogously to (14).

4. Small Area Estimation in Absence of Direct Information

In order to reduce costs, surveys are often carried out in a subset of areas by taking a sample of the population which is representative of the whole study region. This means that direct estimates can only be provided for a few areas and that estimates for the out-of-sample areas must be obtained by other means. Hence, we can split the areas between in-sample and off-sample areas, according to whether or not they have been included in the survey.

In area level models we will therefore have missing the values of $\hat{Y}_{D,i}$ and $\hat{\sigma}_i^2$ for the off-sample areas, whilst in unit level models we will miss the values y_{ij} and \mathbf{x}_{ij} from off-sample areas. However, we will assume that the area level covariates \bar{X}_i are available for all areas since they are obtained from a different source to the survey. This is a key (but realistic) assumption in order to be able to provide small area estimates for all areas.

Note that here we only consider the problem of data that are missing by design of the survey (see section 6 for further particulars). Non-response in surveys, for example, is another common source of missing data in Small Area Estimation, but we do not address this issue here.

A simple approach to tackle the problem of not having direct observations in several areas is to employ a regression model (such as described in Section

2.1) on some covariates, which is fit using survey data (from in-sample areas). Estimates for the off-sample areas are *imputed* by relying on the estimated model and additional information (e.g. area level covariates). The main drawback of this method is that the imputed values do not account for the uncertainty in the estimation of the regression coefficients or spatial correlation between the target variable in different areas.

4.1. Spatially correlated random effects

If we consider any of the models discussed in sections 3.1 and 3.2, the spatially correlated random effects can also be taken into account in addition to the covariates when predicting the small area estimates in off-sample areas. This approach of borrowing information from neighbouring areas when there are missing direct estimates has been considered, for example, by Staubach et al. (2002); LeSage and Pace (2004) and Saei and Chambers (2005) in a non-Bayesian approach.

The way information for areas with no direct observations is borrowed from other areas is as follows. If we want to get an estimate in off-sample areas using the area level model shown in equation (13) (for unit level model the procedure is similar) we could write the model as

$$\begin{bmatrix} \hat{Y}_{D,s} \\ \mu_{\underline{s}} \end{bmatrix} = \alpha + \begin{bmatrix} \bar{X}_s \\ \bar{X}_{\underline{s}} \end{bmatrix} \beta + \begin{bmatrix} z_s \\ z_{\underline{s}} \end{bmatrix} + \begin{bmatrix} e_s \\ 0 \end{bmatrix} \quad (15)$$

where z represents spatially correlated random effects (which can have different specifications, as discussed below), the subindex s refers to the observed (in-sample) areas and \underline{s} to the unobserved (off-sample) areas. The value of $z_{\underline{s}}$ can be estimated by exploiting the (spatial) correlation with z_s .

4.1.1. Multivariate Normal specification

When the full vector of spatial random effects $z = w$ is $MVN(0, \Sigma)$, as shown in equation (10), the conditional distribution of $w_{\underline{s}}|w_s$ is

$$MVN(\Sigma_{\underline{s}s}\Sigma_{ss}^{-1}w_s, \Sigma_{\underline{s}\underline{s}} - \Sigma_{\underline{s}s}^T\Sigma_{ss}\Sigma_{\underline{s}s})$$

as explained in Diggle et al. (1998). The estimate of $\hat{w}_{\underline{s}}$ is then the posterior expectation of the mean of the above conditional MVN, i.e. $E_{\cdot|y}[\Sigma_{\underline{s}s}\Sigma_{ss}^{-1}w_s]$.

The final estimator for the set of areas in \underline{s} becomes

$$\hat{Y}_{\underline{s}} = \hat{\alpha} + \hat{\beta}\bar{X}_{\underline{s}} + \hat{w}_{\underline{s}}$$

This estimator can be regarded as an improved version of the synthetic estimator and it is likely that it will reduce the bias in the estimation.

4.1.2. CAR specification

Prediction of spatial random effects, $w_{\underline{s}}$, in off-sample areas using the multivariate normal specification (10) is unambiguous, since the joint distribution of $w = (w_s, w_{\underline{s}})$, and hence the conditional predictive distribution of $w_{\underline{s}}|w_s$, are uniquely defined. By contrast, prediction of CAR random effects, $v_{\underline{s}}$ is *ad hoc* since the joint distribution of the full vector $v = (v_s, v_{\underline{s}})$ is not defined (e.g. Banerjee et al., 2004, Section 3.3) Instead, prediction proceeds by directly specifying the conditional distributions $v_{\underline{s}}|v_s$; these are well-defined, but in principle, are not unique. This point is illustrated by Banerjee et al. (2004, p.82-83) for the case of a (proper) CAR model fitted to point level (rather than area) data. Prediction at a new location can be achieved either by constructing a CAR model for the set of observed locations and separately specifying the conditional distribution of the new location given the observed ones, or by constructing a CAR model for the full set of observed and new locations. Both approaches are valid, but lead to different predictive distributions. In the case of area level data, it does not make sense to consider the former approach, since it is not obvious how to specify the neighbourhood structure of just the observed (in-sample) areas ignoring the off-sample areas. Hence we specify a CAR model (8) for the full set of spatial random effects in the in-sample and off-sample areas, $v = (v_s, v_{\underline{s}})$, and simply treat the response data in the off-sample areas (i.e. the $y_{i,j}$'s for areas $i \in \underline{s}$) as missing. This leads to a modified set of full conditional distributions for the spatial random effects in off-sample areas in the MCMC scheme used to estimate the posterior distribution (see Appendix for details).

Although the above conditional specification for predicting off-sample random effects is valid, when we lack direct observations from several areas we often encounter the following practical difficulties:

- Fitting of v_i in in-sample areas may be problematic if most or all of the neighbouring areas are off-sample (i.e. have no observations).
- Similarly, when predicting v_i in off-sample areas we may find areas with few or no in-sample neighbours.

In principle, the model can still be fitted, even if there are areas with all neighbours missing (provided the area is not an island), but if too many of them are missing, the estimation procedure will be very unstable. We discuss some possible remedies to this problem in the next section.

Non-spatial random effects u can be included in the model as well, but all the elements of $u_{\underline{s}}$ must be set to zero to avoid problems of identifiability with $v_{\underline{s}}$. In a sense, these unstructured random effects measure the discrepancy between the response (observed data) and the predicted data (fixed and spatial random) effects in the model. If we lack the response, given that the random effects u are not correlated across areas, it is impossible to estimate the scale of the values in $u_{\underline{s}}$.

4.2. *Borrowing information at higher administrative levels*

As noted previously, when the number of areas with no in-sample data is very high prediction of spatial random effects using a CAR model may be problematic because information is borrowed from neighbouring areas (and, indirectly, from second and higher order neighbours) and not enough in-sample neighbours may be available to estimate the spatial pattern. To overcome this problem, a CAR specification at a higher geographical level could be used to deal with the sparseness in the data.

All areas are assumed to be part of one of M higher administrative levels, each of which is made of several areas. We add a random effect $r_k, k = 1, \dots, M$ instead of the spatial effect v_i so that the mean in area level models is decomposed as follows:

$$\mu_i = \alpha + \beta \bar{X}_i + u_i + r_{k(i)} \quad (16)$$

where index $k(i)$ represents the higher administrative level index of area i . A similar formulation can be proposed for μ_{ij} to create a unit level model. $r_{k(i)}$ is given a spatial structure to capture the large scale spatial variation. A CAR prior similar to that assigned to v_i is assumed, but adjacency is defined according to the higher administrative level and it is assumed that at least one area in each region is sampled. If the spatial pattern at this higher level is too weak or non-existent, the random effects $r_{k(i)}$ may be assigned a non-spatial distribution, such a $N(0, \sigma_r^2)$.

5. **Classification of areas to inform policy**

Policy makers are often interested in targeting areas with particular needs in order to conduct specific actions. For example, areas with the highest unemployment rates can be selected to carry out training programmes to improve the possibilities of finding a job or becoming self-employed. In statistical terms, this is the same as selecting the areas which are in the (lower or upper) tails of the distribution of the area level target variable using some form of rank-based estimator.

We note the advice of Goldstein and Spiegelhalter (1996) and consider the ranking of the areas with care. As they argue, comparison of areas is difficult due to the (often considerable) uncertainty in estimating the ranks and the results must be taken more as a guidance than a definitive classification of the areas. Nevertheless, a number of different approaches to ranking a set of units (small areas, hospitals, schools etc.) have been proposed in the literature, and we consider how some of these can be applied in the present context.

The first approach consists of estimating the posterior distribution of each area's rank, R_i . This is straightforward to do using Bayesian sampling-based (MCMC) methods, and just involves ranking the values of the predicted target variable (μ_i) across areas for every posterior sample. The mean of the posterior distribution of each area's rank, $\hat{R}_i = E_{\cdot|y}[R_i]$, has been shown to be the optimal point estimate under squared error loss (Shen and Louis, 1998). Bell (2004) offers some insight into this approach and its applications to poverty mapping in the U.S. in an unpublished work. Note that the posterior mean of the ranks will not necessarily be the same as the ranks based on the posterior mean of the target variable itself, since the rank is a non-linear transformation of the latter. This inconsistency may be seen as an undesirable feature if both the posterior means of the target variable and posterior means of the ranks are to be reported.

To address this problem, Shen and Louis (1998) considered the problem of producing a set of area-specific estimates that satisfy the 'triple goals' of being good estimates of (1) the true histogram (distribution) of parameters across areas; (2) the true ranks; (3) the true parameter values. They note that no single set of estimates can simultaneously optimise all three goals, but that producing a single set of estimates with good performance on each criterion is important in many policy settings. Their proposed estimator produces point estimates of the area-specific target parameter that optimise the first two criteria (in particular, the rank of these point estimates is equivalent to the posterior mean ranks), and that produce estimates of the individual area-specific parameters that, although not as good as the posterior means of the target parameter, generally produce acceptable estimates.

Other authors who have worked on trying to produce an ensemble of small area estimates whose empirical distribution (histogram) is not overshrunk, in order to provide good rankings, include Louis (1984); Spjøtvoll and Thomsen (1987); Lahiri (1990) and Ghosh (1992). In all these cases, the estimates are based on some form of 'constrained' hierarchical or empirical Bayes estimation. The point estimates are optimised under a particular loss function (usually squared error loss) subject to constraints on the mean and variance of the ensemble of estimators — usually that they match the posterior expectation of the mean and variance, respectively, of the true ensemble distribution. However,

Shen and Louis (1998, 2000) have shown that, although constrained estimators yield improved estimates of the histogram of small area target values, they always produce rankings equivalent to the ranking of the posterior means of the target variables, which are not optimal. On the other hand, triple goal estimates have similar or improved performance in terms of estimating both the ranks and the shape of the empirical distribution. Relatively little work has been done on evaluating posterior mean ranks or constrained or triple goal estimators in the context of *spatial* hierarchical models, although work by Conlon and Louis (1999) suggests that performance should be similar to that found for non-spatial settings. On a cautionary note, a number of authors have found constrained and triple goal estimates to be sensitive to model mis-specification (Shen and Louis, 2000; Paddock et al., 2006).

Ranking is relative to the other areas and, for example, will not determine whether a particular area reaches the desired level of income or wealth. For this purpose, a different approach can be followed by defining a specific threshold, T , and estimating the probability that the target variable in each area is above (or below) it. For example, when estimating the average income per household, we can set this threshold to the poverty line (see section 6.2 for details). This ‘*exceedence probability*’ is simply the tail-area probability of the posterior distribution of the *target variable* in each area that is above (or below) the threshold, i.e.

$$p_i^{\text{ex}}(T) = \text{pr}_{\cdot|y}(\mu_i \geq T) \quad (17)$$

These exceedence probabilities can either be used directly, or can be used to rank areas. Morris and Christiansen (1996) and Normand et al. (1997) propose a similar approach to ranking the performance of different hospitals and the identification of those that might be under-performing.

A difficulty with the previous approach is that it can be problematic to set a suitable threshold because it may have to be chosen subjectively. An alternative approach which combines the ideas of ranks and exceedence probabilities is to estimate the probability of being ranked in the top (or bottom) $Q\%$ of areas. Different values of Q can be used at the same time if required. These ‘*percentile probabilities*’ are computed by first converting the ranks to percentiles as follows

$$P_i = R_i/(m + 1)$$

(where m is the total number of small areas) and then calculating the tail area probability of the posterior distribution of P_i that is above (or below) the threshold Q , i.e.

$$p_i^{\text{pc}}(Q) = \text{pr}_{\cdot|y}(P_i \geq Q) \quad (18)$$

Again, these percentile probabilities can either be used directly, or used to compute a final ranking of the areas.

Lin et al. (2006) consider the theoretical and empirical performance of various ranking criteria, including \hat{R}_i (or its percentile equivalent, $\hat{P}_i = \hat{R}_i/(m+1)$), $\hat{P}_i^{\text{ex}}(T) = \text{rank}(p_i^{\text{ex}}(T))/(m+1)$ and $\hat{P}_i^{\text{pc}}(Q) = \text{rank}(p_i^{\text{pc}}(Q))/(m+1)$, in the context of normal-normal and Poisson-gamma hierarchical models for comparing mortality estimates in around 4000 renal dialysis centres in the US. They conclude that the optimal estimator will depend on the specific purpose for which the ranking or classification of units is to be used, but that in most applications, the above three criteria perform well. They also note that all the ranking estimates can perform poorly if the underlying model parameters (equivalent to our small area estimates μ_i) are imprecisely estimated. We illustrate and compare these various criteria in the context of small area estimation in the example in Section 6.

6. Example: Average Income per Household in Sweden

The LOUISE Population register in Sweden contains different socio-economic variables at the individual and household level for all municipalities in the country. Among these variables, we have selected the *equivalised* income per household as the target variable and our aim is to obtain area level estimates of the average equivalised income per household for each municipality. The equivalised income is based on the net income divided by the number of people in the household, but considering different weights for adults and children under 16. In particular, it is defined as

$$\text{EqInc} = \text{NetIncome} / (1 + 0.5 * (\# \text{ of adults} - 1) + 0.3 * (\# \text{ of children under 16}))$$

As auxiliary data, we have considered several covariates that measure different characteristics of the household and the head of household. In particular, the number of people living in the household and the number of employed people were recorded and, for the head of household, we have the age, gender and whether he/she finished tertiary education. The time period is restricted to year 1992. This data set has been analysed by EURAREA Consortium (2004) using a wide range of likelihood-based small area estimators. These reports can be downloaded from <http://www.statistics.gov.uk/eurarea/>.

Our aim is to assess the performance of our methods against a gold standard: the complete survey. To achieve this, we simulate a “mock” survey. The survey design, same as in the EURAREA Reports, includes all of the 284 municipalities of Sweden, with a sample size in each area equal to the 1% of the total number of households, which have been sampled without replacement. The sample sizes range from 7 to 2910, with an average of 110. Variables recorded

for each household include the equivalised income plus the five covariates already described. We have simulated 100 replicated surveys from the complete population survey in order to evaluate the sample-to-sample variability of the small area estimates. A discussion of the impact of the sample size used in the survey is given in Section 6.3.

In addition to the different survey data, the area means (or proportions) of the five covariates are available for every municipality. They have been computed using the whole of the records from the LOUISE register, but other sources of information could have been used as well.

The adjacency of the areas has been computed considering that two regions are neighbours if they have a common boundary. Furthermore, the centroids of the areas (not population-weighted but computed from the boundaries) are also available. This information will be used when considering the correlation structure between the spatial random effects as in (10).

6.1. *Small Area Estimation of Average Income*

We have computed the direct estimator, area level synthetic estimator and all the estimators from area and unit level Bayesian models proposed in Section 3 for each of the 100 survey samples that showed appropriate convergence.

For each Bayesian model we have considered four versions depending on what random effects are included in the model (u_i , v_i , $u_i + v_i$ or w_i). Note that the 5 covariates described before are always included.

All models were run in WinBUGS[†], using two different chains starting at different sets of initial values. Convergence was checked visually, and was excellent in most cases, partly because the response variable is Normal and partly because we used known 'tricks' to improve mixing, such as standardising the covariates in the model. There were a small number of the 100 survey samples for which some of the unit level models did not converge, and these were excluded from the results. The number of replicates used to compute the results reported are shown under column n . Sensitivity to the priors was checked by running some of the models (usually, the one that provided the best estimates) using the alternative priors discussed in Section 3.3 and we found negligible differences in the small area estimates.

The left half of Table 1 summarises the MARB and MRRMSE of the various small area estimators over the n survey samples. The direct estimator has the smallest MARB (probably because it is unbiased by design) but the highest MRRMSE. All the Bayesian estimators have lower bias and mean square error than the corresponding synthetic estimator.

[†]Code available at <http://www.bias-project.org.uk/research.htm>

Comparing the different Bayesian estimators, the biggest impact comes from the way the sampling variances are treated. Unit level model 1, which assumes a common variance for all areas, has higher bias and mean square error than the other models which all assume area-specific variances. When the sample size per area is sufficiently large, as in the present case, the design variance is a good estimate and so there is little to choose between treating the area-specific variances as fixed (as in our area-level model) or modelling them independently or hierarchically (unit level models 2 and 3 respectively). A different picture emerges in the sparse data case (see Section 6.3).

Inclusion of spatial random effects with a CAR structure (v_i), in combination with unstructured random effects, tends to reduce bias and mean square error of the small area estimates compared to models with only unstructured or only spatial random effects. Models including spatial random effects based on the distance between areas (w_i , equation 10) do not perform well compared to the models with spatial random effects based on the CAR specification (results not shown). In practice, the performance is very similar to the model with unstructured random effects u_i . We believe that this happens because the stationarity assumption underlying the former specification does not hold. Another drawback of using this model is that fitting takes much longer than with other spatial models.

Table 1 also shows that, broadly speaking, the ranking of models by DIC is similar to that based on MARB or MRRMSE, indicating that DIC can be useful for model selection in a real application setting (although note that the latter cannot be used to compare models based on different data, i.e. area versus unit level models).

The right half of Table 1 summarises the bias and mean square error of the variance estimates from the different models. For all models, the variance estimates have much higher relative bias and mean square error than do the small area estimates themselves. Nevertheless, most of the Bayesian variance estimates (except unit level model 1 and the unit level model 2 with only spatial random effects) are substantially better than the corresponding synthetic variance estimates, and comparable or better than the variance of the direct estimator. In contrast to the findings for the small area point estimates, however, the Bayesian models with spatial random effects only tend to perform poorly relative to models including unstructured random effects (either alone or in combination with spatial effects). The variance estimates from the former models tend to be systematically smaller than those from the latter (data not shown), suggesting that while borrowing of information from neighbouring areas can improve the small area estimates, it tends to over-estimate their precision. Thus models with both unstructured and spatial random effects may represent the best compromise in terms of producing accurate small area estimates which

also have reasonable variance estimates.

As an alternative way of assessing the variability of the small area estimates we have also computed frequentist coverage rates for all these models. Rao (2003, Chapter 10) discusses these issues and points out that posterior variances tend to underestimate the frequentist MSE when the number of areas and/or the between-area variance is small, which may lead to too narrow credible intervals. However, our results in Table 1 show that the coverage rates for the various Bayesian models considered here are only slightly lower than the nominal 95% in most cases. The exceptions are the Bayesian models with only spatial random effects, where there is modest under-coverage, for the reasons already discussed. In contrast, there are severe problems with the coverage of the synthetic estimator, probably due to the fact that it has high bias and high variability.

6.2. Classification of areas

We have selected the area level model and unit level Model 3 among unit level models (on the grounds of giving good results and having a better convergence) to rank the areas, using some of the different classification criteria discussed in Section 5.

For the ranking method based on the posterior mean ranks, we have computed the mean root MSE (which seems a more reasonable criterion for ranks than the relative measures, MARB and MRRMSE) to assess which model is the best in terms of ranking the areas. In general, each type of model produces a very similar ranking, but models with unstructured random effects perform better in this regard. Furthermore, unit level models produce a slightly better ranking than area level models. In particular, mean root MSE for unit level model 3 with independent random effects is 101.76 whilst for the area level model with the same random effects it is 103.63. Thus the models that achieve the best small area estimates do not necessarily produce the most accurate rankings. As discussed earlier, Shen and Louis (1998) propose the use of triple-goal estimates to produce good small area estimates that also produce a good ranking of the areas.

Figure 1 shows an example of the posterior ranking obtained with area level model with unstructured and spatial random effects. Although only the results for one survey sample are shown here, the results for the other samples are similar. Figure 1(a) displays posterior mean ranks and 95% credible intervals, with an ordering based on the true area ranks. It is clear that it is difficult to separate the low ranked areas from the medium ranked ones as many intervals overlap. Similarly, Figure 1(b) shows the posterior small area estimates of the average income per household.

Table 1. Summary of the performance of the small area estimates and their variance estimates for different models fitted to the 1% survey samples. The values are averaged over the n replicates.

Model	MARB ×100	MRRMSE ×100	DIC*	MARBvar [†] ×100	MRRMSEvar [†] ×100	Cov. rate [‡]	n
Direct	0.5	6.5	—	68.7	68.8	0.94	100
<i>Area level</i>							
Synthetic	3.5	3.6	—	89.3	89.3	0.17	100
Bayesian							
u_i	2.0	3.1	3246	58.8	69.1	0.93	100
v_i	2.1	2.8	3279	80.2	89.7	0.91	100
$u_i + v_i$	1.8	2.9	3232	60.9	72.0	0.93	96
<i>Unit level</i>							
Synthetic	6.0	6.0	—	94.6	99.4	0.06	100
Bayesian							
Model 1							
u_i	3.0	4.3	495714	122.8	185.4	0.94	98
v_i	3.7	4.3	495825	109.4	144.6	0.85	98
$u_i + v_i$	2.8	3.6	490784	109.6	130.0	0.93	72
Bayesian							
Model 2							
u_i	2.0	3.5	474461	42.4	61.5	0.93	100
v_i	2.8	3.3	474682	102.5	124.6	0.88	100
$u_i + v_i$	1.9	3.2	474122	49.4	69.6	0.91	75
Bayesian							
Model 3							
u_i	2.0	3.5	474118	43.9	64.0	0.93	94
v_i	2.0	3.0	474077	62.3	72.6	0.90	94
$u_i + v_i$	1.9	3.2	474363	56.2	76.5	0.94	76

*Area and unit DICs are not comparable because they are computed using different data.

[†]Variance estimate for direct estimator is the design variance; for synthetic estimator it is $Var(\bar{X}_i\hat{\beta})$; for Bayesian estimators it is the posterior variance of $\hat{\mu}_i$.

[‡]95% intervals for direct and synthetic estimators are the 95% confidence intervals assuming normality; for the Bayesian estimators they are the central 95% posterior credible interval for $\hat{\mu}_i$.

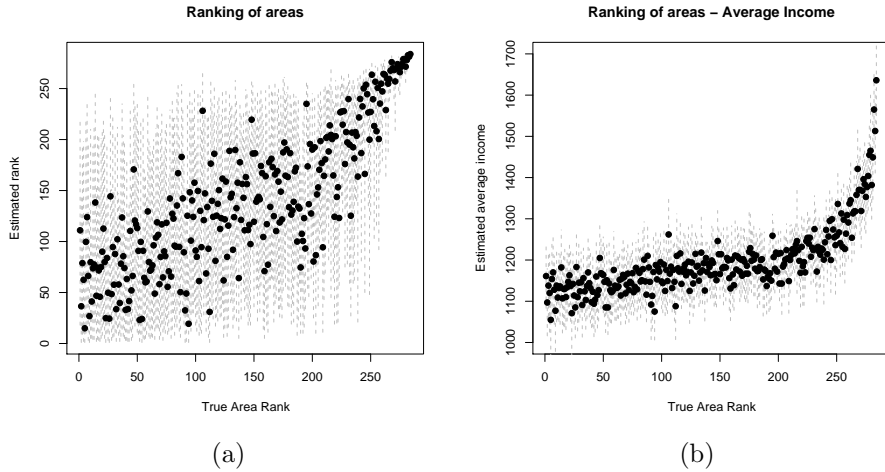


Figure 1. Posterior estimates of (a) the area rank, and (b) the target variable (income) for a randomly chosen dataset, estimated using the area level model with unstructured and CAR spatial random effects. 95% credible intervals (in grey) show the high uncertainty about the estimates, particularly for the ranks.

For the ranking method based on a threshold level (exceedence probabilities), we have used the *poverty line* as the cut-off of interest. The poverty line can be defined in several ways. A common definition (used, for example, by the Organisation for Economic Co-operation and Development) is 60% of the national median equivalised income per household (Hills, 2004, page 40). However, it turns out that none of the municipalities in Sweden fall below such a threshold. For illustration, we therefore set the cut-off point by considering the 0.05 quantile of the empirical distribution of the direct estimators. In all 100 survey data sets these quantiles were very close to 1050, which was set as the cut-off point.

We have also considered ranking of areas based on percentile probabilities — specifically, the probability of being among the 10% and 20% of areas with the lowest equivalised income per household (28 and 57 municipalities, respectively) and of being the poorest area.

The plots in Figure 2 show the results of applying these various probability-based ranking criteria to the Swedish income example, for the area level model with independent and spatial random effects (the results for other models are similar). For each criterion we have included the following three plots:

- Posterior probabilities against true area rank for a single randomly chosen data set
- Mean of posterior probabilities across datasets against true area rank, with 95% sampling intervals (see below).
- Mean of the ranks of the posterior probabilities across datasets, with 95% sampling intervals (see below).

The first two rows displayed in Figure 2 show the probabilities of being among the 20% and 10% of the most deprived areas, respectively. These posterior probabilities are not simple monotonic functions of the ranks and some additional variability is observed. Dotted lines representing the uncertainty about the replication variability have also been displayed. These sampling intervals correspond to the variability between the different survey samples, because in this case the Bayesian approach does not provide them for an individual data set. As with the ranks, there is considerable uncertainty about these probabilities, highlighting the difficulty of finding a reliable way to classify or discriminate a number of areas from the rest.

As alternatives to reduce sample-to-sample variance of these probabilities, we consider more extreme criteria, and compute the posterior probability of being the most deprived area and the posterior probability of the average income being below the threshold of 1050. In both of these plots, the sampling variability of the probability is greatly reduced, and we can be quite confident that the areas with non-zero probabilities are poorer than those with zero probabilities. However, it is still difficult to discriminate between the areas with non-zero probabilities due to overlapping intervals.

In order to measure the performance of the ranking methods based on computing different percentile probabilities, we have computed the Operating Characteristic (OC) proposed in Lin et al. (2009, see page 28 for details), which are shown in Table 2. The Operating Characteristic is similar to computing the probability of wrongly classifying an area as below (or above) the specified quantile. The OC's are then averaged over all samples and 95% sampling intervals have been computed from the OC values for all the data sets. The results indicate the OC's for a particular quantile are quite consistent across data sets and across models, although models that include a spatial structured random effect generally have slightly lower OC values than those with only unstructured random effects, showing better classification performance. There are substantial differences in the OC's for different quantiles however. In this particular example, all the models perform better at discriminating between the richest areas and the rest (for example, the OC is around 0.3 for classifying areas as above or below the 80th percentile), and less good at correctly identifying the

Table 2. Operating characteristic (with 95% sampling intervals). Low values of the OC indicate good performance in ranking the areas.

MODEL	Percentile (r)				
	80%	50%	20%	10%	(1/284)%
<i>Area level</i>					
u_i	0.28 (0.24,0.33)	0.39 (0.34,0.41)	0.47 (0.43,0.51)	0.53 (0.47,0.58)	0.81 (0.54,0.91)
v_i	0.24 (0.20,0.28)	0.36 (0.32,0.39)	0.45 (0.41,0.49)	0.50 (0.45,0.55)	0.71 (0.47,0.86)
$u_i + v_i$	0.25 (0.20,0.31)	0.37 (0.33,0.40)	0.45 (0.40,0.50)	0.51 (0.46,0.57)	0.74 (0.43,0.90)
<i>Unit level</i>					
<i>Model 2</i>					
u_i	0.36 (0.33,0.39)	0.42 (0.39,0.45)	0.48 (0.44,0.52)	0.54 (0.49,0.59)	0.82(0.61,0.93)
v_i	0.33 (0.30,0.36)	0.44 (0.39,0.49)	0.50 (0.44,0.56)	0.52 (0.45,0.61)	0.61 (0.39,0.82)
$u_i + v_i$	0.30 (0.24,0.39)	0.41 (0.38,0.44)	0.47 (0.42,0.51)	0.52 (0.45,0.58)	0.72 (0.39,0.90)
<i>Model 3</i>					
u_i	0.36 (0.33,0.40)	0.42 (0.39,0.46)	0.49 (0.45,0.53)	0.55 (0.50,0.60)	0.82 (0.61,0.93)
v_i	0.28 (0.24,0.32)	0.41 (0.38,0.45)	0.47 (0.42,0.52)	0.52 (0.45,0.59)	0.69 (0.37,0.88)
$u_i + v_i$	0.33 (0.30,0.37)	0.43 (0.40,0.46)	0.50 (0.46,0.54)	0.55 (0.50,0.60)	0.77 (0.51,0.89)

poorest areas (for example, the OC is around 0.5 for classifying areas as above or below the 10th or 20th percentile). This is reflected in the plots in Figure 2, which show much less uncertainty about the probabilities and ranks of the richer areas.

6.3. Effect of the sample size

We have based our primary results on a sample size of 31,144 households which is the same as was used in the analysis of this data by the EURAREA Project, and typical for many major surveys. For example, the Family Resources Survey (FRS) carried out yearly in Great Britain by the Office for National Statistics had an initial sample size of 49,800 households in 2005-2006 (Barton et al., 2007) (out of approximately 25.29 millions in 2006). The response rate was 62% or 28,029 households. Although this is similar in absolute sample size to our example, in percentage terms, our Swedish sample corresponds to around 1% of households, whereas the FRS sample corresponds to around 0.11% of the total households in Great Britain.

For our Swedish example, we have therefore also considered a 0.1% sample of the households in the area (or a sample size of 5, whatever is higher), which resulted in a total sample size of 3,358 households. Although these numbers seem too small for a national survey, this example will illustrate how the performance of the various SAE models is affected by variation of the sample size.

For this reduced sample, we have computed the MARB and MRRMSE for both the small area estimates and their variance estimates for the area level models and unit level models 3. These are shown in Table 3 together with the DICs and coverage rates of the 95% intervals. As expected, bias and mean

Table 3. Summary of the performance of the small area estimates and their variance estimates for different models fitted to the 0.1% survey samples. The values are averaged over the n replicates.

Model	MARB ×100	MRRMSE ×100	DIC*	MARBvar [†] ×100	MRRMSEvar [†] ×100	Cov. rate [#]	n
Direct	1.5	17.6	—	95.4	95.4	0.46	100
<i>Area level</i>							
Synthetic	4.3	5.1	—	98.5	98.5	0.12	100
Bayesian							
u_i	1.4	11.9	3288	84.1	85.6	0.50	100
v_i	2.4	8.1	5334	37.4	81.9	0.70	100
$u_i + v_i$	1.4	11.9	3290	84.0	85.5	0.50	100
<i>Unit level</i>							
Synthetic	5.8	6.3	—	79.8	155.5	0.19	100
Bayesian							
Model 3							
u_i	2.9	6.0	50648	46.5	70.2	0.93	100
v_i	3.0	4.4	50565	82.5	99.7	0.90	100
$u_i + v_i$	3.0	4.6	50565	96.5	120.5	0.93	100

*Area and unit DICs are not comparable because they are computed using different data.

[†]Variance estimate for direct estimator is the design variance; for synthetic estimator it is $Var(\bar{X}_i\hat{\beta})$; for Bayesian estimators it is the posterior variance of $\hat{\mu}_i$.

[#]95% intervals for direct and synthetic estimators are the 95% confidence intervals assuming normality; for the Bayesian estimators they are the central 95% posterior credible interval for $\hat{\mu}_i$.

square error of both the estimates and their variances are generally higher than for the 1% sample size. For the unit level model 3, the overall picture for the 0.1% and 1% sample sizes is similar, however, with the Bayesian estimates outperforming the synthetic estimates on most criteria, and inclusion of spatial random effects leading to improved small area estimates but over-precise variance estimates. For the Bayesian area level model, the reduction in sample size has a particularly detrimental impact on the mean square error and coverage rates of the small area estimates. This probably reflects the fact that the area level model considered here uses the design variance as the (fixed) estimate of the sampling variance in each area, and this is likely to be a poor estimate when the sample size in each area is small. Specifying a hierarchical model to smooth the variances, as in unit level model 3, could be used to address this problem.

Regarding the convergence of the models, it seems to be better for unit level model 3 than the other Bayesian models, probably because of the hierarchical structure on the area level variances. In all cases, the model with both random effects u_i and v_i showed that it is difficult to disentangle spatial and non-spatial variation when data are sparse because of poor convergence and possible confounding between the two effects; a phenomenon documented in disease mapping (Best et al., 2005).

6.4. *Estimation in the absence of direct information*

We have extended our analysis to consider the case where there are only a few areas in the sample. This is usually done in practice to reduce the survey costs. To be precise, we have considered a mock survey with only 100 municipalities, with a mean sample size of 166 (range 17 to 2910) for the 1% sample, and mean sample sizes of 17 (5 to 291) for the 0.1% sample. For the selection of the regions, we have created several socio-economic strata according to the area average number of employed people per household and the proportion of head of household with tertiary studies. Each variable was divided in three intervals, which led to 9 strata. A sample size was assigned at each strata proportionally to the number of municipalities that it contained, with the only constraint that at least one municipality must be considered at each strata. Hence, the data is now made of the previous 100 survey data sets for each of the 1% and 0.1% samples but deleting the areas not included in the sampled regions. This means that the samples taken within each area are the same as before, so that MARB and MRRMSE can be directly compared.

When dealing with areas with missing observations it is important to consider carefully the missingness mechanism, because ignoring it can have an impact on the outcome (Little and Rubin, 2002). In our case, the probability of being missing (i.e., not appearing in the sample) depends on a set of covariates used to assign each area to a stratum which are assumed to be known. Hence, the missing data are *Missing At Random* (MAR) and, as discussed in Gelman et al. (1995, page 205), the missingness mechanism can be ignored and the results will not be affected because the covariates are included to estimate the area level means.

Similar models to those used in the previous sections have been fitted to the samples from the reduced set of areas. The results for area level model and unit level model 3 for both 1% and 0.1% samples are summarised in Table 4.

For in-sample areas, the bias and mean square error of the new small area estimates are similar to those based on samples for the full set of areas, whilst, as expected, estimates for off-sample areas have higher bias and mean square error. The notable exception is again the area-level model fitted to the 0.1% sample

Table 4. Summary of the performance of the small area estimates for selected models, in the absence of direct information for some areas.

MODEL	1% SURVEY SAMPLE			0.1% SURVEY SAMPLE		
	MARB ×100 (in/off)	MRRMSE ×100 (in/off)	Coverage rate [#] (in/off)	MARB ×100 (in/off)	MRRMSE ×100 (in/off)	Coverage rate [#] (in/off)
<i>Area level</i>						
Synthetic	3.4/3.9	3.6/4.1	0.29/0.25	4.1/4.1	5.5/5.6	0.17/0.19
Bayesian						
u_i	1.8/3.7	3.0/3.9	0.92/0.44	1.2/3.5	11.3/5.1	0.46/0.71
v_i	1.9/3.4	2.9/3.9	0.90/0.86	1.2/3.3	11.4/7.8	0.48/0.99
$u_i + v_i$	1.8/3.4	2.9/3.9	0.91/0.85	1.2/3.5	11.4/5.4	0.46/0.85
REGIONAL	1.8/3.4	2.9/3.9	0.90/0.64	1.2/3.5	11.3/5.1	0.46/0.76
<i>Unit level</i>						
Synthetic	5.7/6.8	5.7/6.8	0.09/0.04	5.8/6.9	6.4/7.4	0.24/0.16
Bayesian Model 3						
u_i	2.1/4.9	3.5/5.1	0.91/0.15	2.1/3.5	3.5/5.1	0.91/0.15
v_i	2.0/3.8	3.0/4.1	0.86/0.86	2.0/3.0	3.8/4.1	0.89/0.86
$u_i + v_i$	2.0/3.8	3.0/4.2	0.85/0.84	2.0/3.0	3.8/4.1	0.89/0.85
REGIONAL	1.9/3.9	3.1/4.5	0.64/0.56	1.9/3.1	3.9/4.6	0.90/0.56

[#]95% intervals for synthetic estimators are the 95% confidence intervals assuming normality; for the Bayesian estimators they are the central 95% posterior credible interval for $\hat{\mu}_i$.

size. Somewhat counter-intuitively, the mean square error in the off-sample areas in this scenario is about *half* that of the in-sample areas. The explanation lies with the problem noted earlier, that treating the sample (design) variance in in-sample areas as fixed in the area level model is leading to poor estimates when data are sparse. The estimates for off-sample areas do not suffer from this problem, since there is no data in these areas and so their estimates are simply predicted from the fitted model and do not depend on the observed design variance.

The average coverage rates for in-sample and off-sample areas have also been included in the Table 4. In general, coverage rates for the in-sample areas are similar to, or slightly lower than, the full data case (again, the exception to this is the area level model with 0.1% sample size, which has much lower than expected coverage). In off-sample areas, coverage is similar to that in in-sample areas only for the models that include spatial random effects, and is lower for other models. This is because the unstructured random effects are set to zero in these areas for identification, and so do not help to improve the small area

estimates in those areas.

6.5. *Regional model*

Area and unit level models based on the structure for μ_i in equation (16) have been computed to assess the value of borrowing information at a higher regional level when there are areas with missing observations. In general, these models perform similarly to the models with area (municipality) level spatial random effects, and better than the models with only unstructured random effects. This suggests that whilst inclusion of spatial random effects is important for improving the small area estimates, the precise structure assumed for these random effects is less critical.

7. Discussion

In this paper we have described different Bayesian area and unit level models for the estimation of variables in small areas by combining information from survey data and other sources. We have studied the importance of taking into account non-spatial and spatially correlated area level variation, and we have found that by including random effects to model both these sources of variation the small area estimates can be improved. These improvements are somewhat offset by a tendency for spatial models to under-estimate the variances of the estimates, however. Despite this, we have shown that it becomes particularly important to model spatial dependence when some areas have no direct survey estimates, as information from nearby areas can be used to improve prediction in the off-sample areas.

When comparing area versus unit level models, the models performed similarly when using a 1% survey sample. However, when a 0.1% sample was employed, area level models had a smaller bias but were worse than unit level models in terms of MRRMSE. We believe that this is due to the fact that the direct estimators on which area level models rely are unbiased by design and have a wide design variance in this case and are not very reliable. Smoothing the within-area variances, as we did for the Bayesian unit level models, would help to address this problem.

Regarding unit level models, we considered three ways of modelling the within-area variance. Clearly, allowing for a different within-area variance for each area improves the fitting. When there are sufficient data to estimate the within area variance with accuracy there is not much difference between Models 2 and 3. However, the hierarchical structure on the area level variances as in Model 3 leads to better convergence of the MCMC simulations. In a more general framework, a hierarchical structure based on covariates (for example,

linear regression) could be employed to model the different area level variances (see, for example, Gelman, 2006). Hence, we have shown the importance of borrowing strength across areas when data are sparse not only to estimate the area level means but also the area level variances.

In this work we have only considered models with a Normal response. Generalised Linear Mixed Models can be used to deal with non-Normal variables and the ideas presented here can still be applied. However, combining individual and aggregated data is not so straightforward because in these models the response and the explanatory covariates are not linked linearly any more and so care is required to specify an appropriate aggregate form of the individual-level model. For example, as shown by Jackson et al. (2006), it is possible to synthesise different sources of data, with different levels of aggregation when an appropriately specified model is used. Otherwise, a bias in the estimation of the coefficients of the covariates is introduced, which may bias the small area estimates of the target variable as well. We intend to explore these ideas in the future and tackle their application to the estimation of area level counts and rates, such as the number of persons per household and rate of unemployment.

Areas can be classified to help inform policy issues by exploiting the results provided by Bayesian inference. We have considered different approaches to the ranking of areas. Accounting for the uncertainty of the estimates is crucial because when areas tend to be similar it will be difficult to separate low-ranked areas from the rest. Alternatively, the probability of being among the $q\%$ lowest ranked areas can be used instead, for some suitable quantile, q . We have shown that choosing more extreme quantiles (e.g. the lowest ranked area rather than the bottom 10% or 20%) reduces uncertainty about the ranks due to sampling variation. However, it is still difficult to confidently identify all but the most extreme ranked areas.

When some areas are not included in the survey, it is still possible to provide estimates for those areas by relying on the fitted Bayesian models (using in-sample areas) and their spatial correlation to off-sample areas. Area level covariates are still required in all areas to compute the small area estimates. As expected, these estimates are less accurate than in the case with survey data in all areas but, despite the loss of performance, the results are still reasonable and have lower bias and better coverage than traditional synthetic estimates. When data are very sparse, spatial random effects can be incorporated at a regional level, so that larger-scale spatial patterns are modelled. This can help to cope with large amounts of areas with no direct observations and provide reliable results.

8. Acknowledgements

We would like to thank ONS for providing access to the Swedish data, which were originally provided by Statistics Sweden. This work was carried out as part of the Imperial College BIAS node of the ESRC National Centre for Research Methods (grant numbers RES-576-25-5003 and RES-576-25-0015). We would also want to thank the associate editor and two anonymous referees whose comments have helped to improve the quality of this paper.

References

- Arora, V. and P. Lahiri (1997). On the superiority of the Bayesian method over the BLUP in Small Area estimation problems. *Statistica Sinica* 7, 1053–1063.
- Arora, V., P. Lahiri, and K. Mukherjee (1997). Empirical Bayes estimation of finite population means from complex surveys. *Journal of the American Statistical Association* 92(440), 1555–1562.
- Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, Florida.
- Barton, J., R. Chung, D. Donaldson, J. Shome, J. Snow, E. Whiting, and M. Willitts (Eds.) (2007). *Family Resources Survey. United Kingdom 2005-2006*. National Statistics, Department for Work and Pensions (UK).
- Battese, G. E., R. M. Harter, and W. A. Fuller (1988). County crop areas using survey data and satellite data. *Journal of the American Statistical Association* 83(401), 28–36.
- Bell, W. R. (2004). *A Bayesian approach to recognizing Statistical Uncertainty in ranked Survey Estimates*. U.S. Census Bureau, Census Advisory Committee of Professional Association Meetings. April 22-23, 2004.
- Besag, J., J. C. York, and A. Mollié (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 43, 1–59.
- Best, N., S. Richardson, and A. Thomson (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research* 14(1), 35–59.
- Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear models. *Journal of the American Statistical Association* 88(421), 9–25.

- Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). John Wiley & Sons, Inc., New York.
- Conlon, E. M. and T. A. Louis (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, and R. Bertollini (Eds.), *Disease Mapping and Risk Assessment for Public Health*, pp. 31–47. Wiley.
- Datta, G. S. and M. Ghosh (1991). Bayesian prediction in linear models: Applications to small area estimation. *The Annals of Statistics* 19(4), 1748–1770.
- Diggle, P., R. Moyeed, B. Rowlingson, and M. Thomson (2002). Childhood malaria in the Gambia: a case-study in model-based geostatistics. *Applied Statistics* 51, 493–506.
- Diggle, P., J. Tawn, and R. Moyeed (1998). Model-based geostatistics. *Applied Statistics* 47, 299–350.
- Eberley, L. E. and B. P. Carlin (2000). Identifiability and convergence issues for Markov Chain Monte Carlo fitting of spatial models. *Statistics in Medicine* 19, 2279–2294.
- EURAREA Consortium (2004). Project reference volume. Technical report, EURAREA Consortium.
- Fay, R. E. and R. A. Herriot (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74(366), 269–277.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (Comment on article by Browne and Draper). *Bayesian Analysis* 31, 515 – 533.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, Florida.
- Ghosh, M. (1992). Constrained Bayes estimation with applications. *Journal of the American Statistical Association* 87(418), 533–540.
- Ghosh, M., K. Natarajan, T. W. F. Stroud, and B. P. Carlin (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association* 93(441), 273–282.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton, Florida.

- Goldstein, H. and D. J. Spiegelhalter (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, Series A* 159(3), 385–443.
- Gonzalez, M. E. (1973). Use and evaluation of synthetic estimators. In *Proceedings of the Social Statistics Section*, pp. 33–36. American Statistical Association, Washington, D.C.
- Greenland, S. and J. Robins (1994). Ecologic studies—biases, misconceptions, and counterexamples. *American Journal of Epidemiology* 139(8), 747–760.
- Heady, P., P. Clarke, and others (2003). Small Area Estimation Project Report. Technical report, Office for National Statistics, United Kingdom.
- Hills, J. (2004). *Inequality and the State*. Oxford University Press.
- Jackson, C., N. Best, and S. Richardson (2006). Improving ecological inference using individual-level data. *Statistics in Medicine* 25(12), 2136–2159.
- Jiang, J. and P. Lahiri (2006). Mixed model prediction and small area estimation. *Test* 15(1), 1–96.
- Kleffe, J. and J. N. K. Rao (1992). Estimation of mean square error of empirical bests linear unbiased predictors under a random variance linear model. *Journal of Multivariate Analysis* 43, 1–15.
- Lahiri, P. (1990). “Adjusted” Bayes and Empirical Bayes estimation in finite population sampling. *Sankhyā, Series B* 52(1), 50–66.
- LeSage, J. P. and R. K. Pace (2004). Models for spatially dependent missing data. *Journal of Real Estate Finance and Economics* 29(2), 233–254.
- Lin, R., T. A. Louis, S. M. Paddock, and G. Ridgeway (2006). Loss function based ranking in two-stage hierarchical models. *Bayesian Analysis* 1(4), 915–946.
- Lin, R., T. A. Louis, S. M. Paddock, and G. Ridgeway (2009). Ranking US-RDS provider specific SMRs from 1998–2001. *Health Services and Outcomes Research Methodology* 9, 22–38.
- Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc., Hoboken, New Jersey.

- Louis, T. A. (1984). Estimating a population of parameter values using Bayes and Empirical Bayes methods. *Journal of the American Statistical Association* 79(386), 393–398.
- Morris, C. N. and C. L. Christiansen (1996). Hierarchical model for ranking and for identifying extremes, with applications (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 5*, pp. 277–296. Oxford University Press.
- Normand, S. L., M. E. Glickman, and C. Gatsonis (1997). Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association* 92, 803–814.
- Paddock, S. M., G. Ridgeway, R. Lin, and T. A. Louis (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational Statistics and Data Analysis* 50, 3243–3262.
- Petrucci, A. and N. Salvati (2005). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological & Environmental Statistics* 11(2), 169–182.
- Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science* 6, 15–51.
- Saei, A. and R. Chambers (2005). Working paper M05/03: Empirical Best Linear Unbiased Prediction for out of sample areas. Technical report, Southampton Statistical Sciences Research Institute, University of Southampton.
- Särndal, C.-E., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer-Verlag Inc., New York.
- Shen, W. and T. A. Louis (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B* 60(2), 455–471.
- Shen, W. and T. A. Louis (2000). Triple-goal estimates for disease mapping. *Statistics in Medicine* 19, 2295–2308.
- Singh, B. B., G. K. Shukla, and D. Kundu (2005). Spatio-temporal models in small area estimation. *Survey Methodology* 31(2), 183–196.
- Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc. B* 64, 583–639.

- Spjøtvoll, E. and I. Thomsen (1987). Applications of some Empirical Bayes methods to Small Area Estimation. *Bulletin of the International Statistical Institute* 46(2), 435–449.
- Staubach, C., V. Schmid, L. Knorr-Held, and M. Ziller (2002). A Bayesian model for spatial wildlife disease prevalence data. *Preventive Veterinary Medicine* 56(1), 75–87.

A. Full conditional distributions for CAR specification with missing observations

Let us assume that we have an area level model and that we have data from the first l areas, i.e., we have the values of direct estimates and their sampling variances for these areas. In this context, $s = \{1, \dots, l\}$ and $\underline{s} = \{l + 1, \dots, m\}$.

The full conditionals to be used in Gibbs Sampling for this model are:

$$\mathbf{\Pi}(\alpha, \beta | \dots) = N \left((\bar{X}^T V^{-1} \bar{X})^{-1} \bar{X}^T V^{-1} (\hat{Y} - v), (\bar{X}^T V^{-1} \bar{X})^{-1} \right)$$

where $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_l)^T$, $v = (v_1, \dots, v_l)^T$, $V = \text{diag}(\hat{V}_1^2, \dots, \hat{V}_l^2)$ and

$$\bar{X}^T = \begin{pmatrix} 1 & \dots & 1 \\ \bar{X}_1 & \dots & \bar{X}_l \end{pmatrix}$$

$$\mathbf{\Pi}(v_i | \dots) \propto \exp \left\{ \frac{-1}{\hat{\sigma}_i^2} (\hat{Y}_i - \alpha - \beta \bar{X}_i - v_i)^2 + \frac{-1}{\sigma_v^2} n_i (v_i - \bar{v}_i)^2 \right\}, \quad i \in s$$

$$\mathbf{\Pi}(v_j | \dots) \propto \exp \left\{ \frac{-1}{\sigma_v^2} n_j (v_j - \bar{v}_j)^2 \right\}, \quad j \in \underline{s}$$

where n_i and n_j represent the number of neighbours of areas i and j , respectively.

$$\mathbf{\Pi}(\sigma_v^2 | \dots) = Ga^{-1} \left(\varepsilon + \frac{n}{2}, \varepsilon + \frac{1}{2} \sum_{k \sim l} (v_k - v_l)^2 \right)$$

Note that the full conditional for each v_i depends directly on observed data and its neighbours, whilst v_j only depends on its neighbours through \bar{v}_j and not directly on the observed data. Furthermore, α and β are only informed by the data from the areas included in the survey, as we would expect.

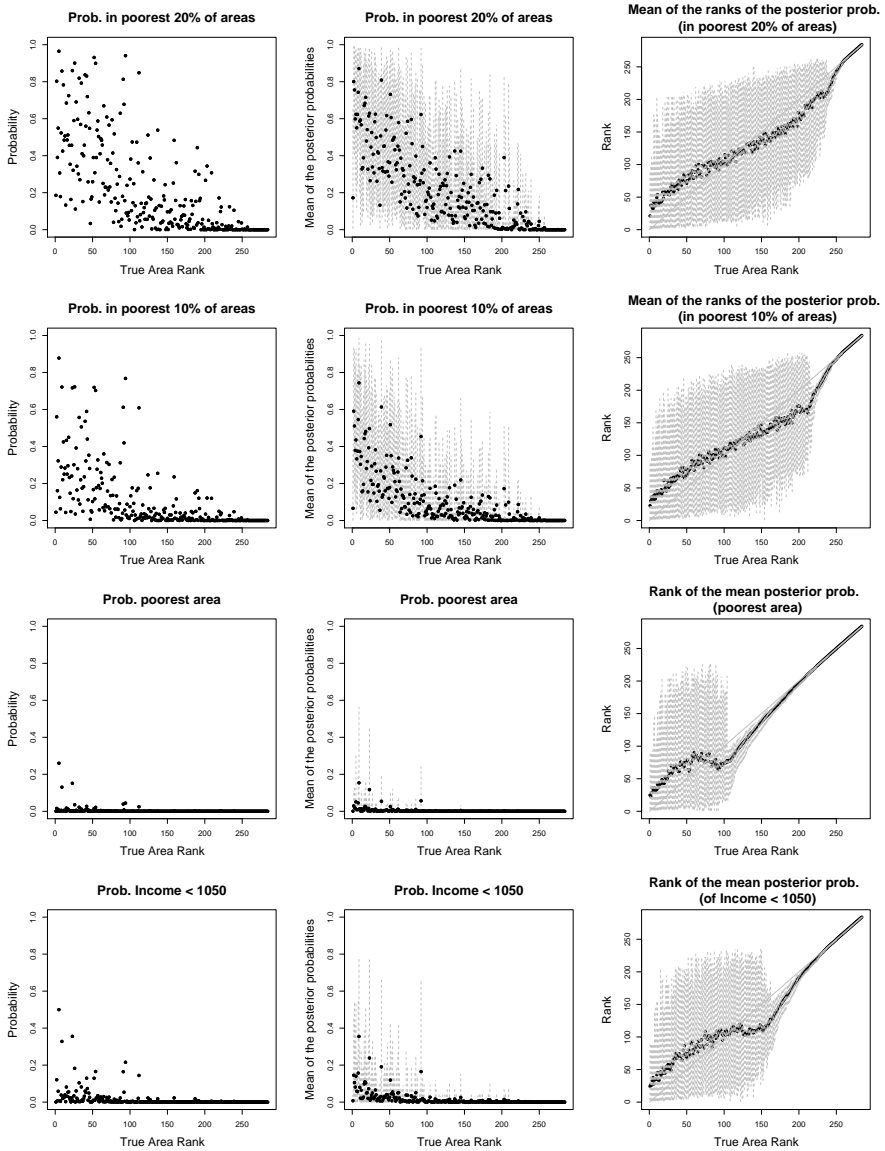


Figure 2. Ranking of the areas according to four different criteria. Left column shows the exceedence or percentile probabilities for each area for a randomly selected data set; middle column shows the mean of these probabilities across all datasets; right column shows mean of the rank of these probabilities across datasets. Grey dashed lines are 95% 'sampling intervals' and show the uncertainty about the replication variability for the probabilities and ranks across the survey samples.