ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

Clustering in small area estimation with area level linear mixed models

Elaheh Torkashvand, Mohammad Jafari Jozani and Mahmoud Torabi

University of Manitoba, Winnipeg, Canada

[Received January 2016. Final revision June 2017]

Summary. Finding reliable estimates of parameters of subpopulations (areas) in small area estimation is an important problem especially when there are few or no samples in some areas. Clustering small areas on the basis of the Euclidean distance between their corresponding covariates is proposed to obtain smaller mean-squared prediction error (MSPE) for the predicted values of area means by using area level linear mixed models. We first propose a statistical test to investigate the homogeneity of variance components between clusters. Then, we obtain the empirical best linear unbiased predictor of small area means by taking into account the difference between variance components in different clusters. We study the performance of our proposed test as well as the effect of the clustering on the MSPE of small area means by using simulation studies. We also obtain a second-order approximation to the MSPE of small area means and derive a second-order unbiased estimator of the MSPE. The results show that the MSPE of small area means can be improved when the variance components are different. The improvement in the MSPE is significant when the difference between variance components is considerable. Finally, the methodology proposed is applied to a real data set.

Keywords: Combined clustering; Complete clustering; Empirical best linear unbiased predictor; Mean-squared prediction error; Simple clustering; Small area estimation

1. Introduction

Consider a small area estimation problem where the ultimate goal is to predict area means with a higher precision. Adding random effects to linear models reduces the bias of predictors of small area means whereas it increases their variability. It is a common practice in mixed effect models to assume homogeneous random effects. However, there are many applications where such an assumption is not valid. To provide more realistic predictions of small area means, one might decide to work with non-homogeneous random effects. Recently, there has been much research in this direction. For example, Maiti *et al.* (2011) studied the effect of poverty on the educational performance of students in different school districts. In this application, small areas (school districts) were clustered on the basis of their poverty status and different regression models were assumed for different clusters. Random effects of school districts belonging to the same cluster were considered to follow a normal distribution with a different variance component due to the effect of socio-economic status of families and areas on the students' education.

In other work, Datta *et al.* (2011) and Datta and Mandal (2015) argued whether or not the presence of the random effects in small area estimation is necessary, especially when the main concern is the prediction of small area means. Datta *et al.* (2011) gave an example that showed that including random effects might not always be useful. To test how influential the

Address for correspondence: Mohammad Jafari Jozani, Department of Statistics, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada.
E-mail: M_Jafari_Jozani@umanitoba.ca

 0964–1998/17/180000

presence of the random effects in an area level model is, they developed methodologies based on a frequentist approach. They also introduced a statistical test to decide whether the inclusion of the random effect in the small area models is necessary. They concluded that including random effects in the model decreases the rate of convergence to the true values of area parameters. The decrease is significant especially when the sample size in areas, n_i, are large. They also pointed out that dropping the random effects can lead to more accurate point and/or interval estimators, although the flexibility and adaptivity of the area level (also called Fay–Herriot) model might be lost. The disadvantage of this methodology is that the random effects will be eliminated from all areas whereas it might be necessary to keep it for some areas. To address this issue, Datta and Mandal (2015) took a Bayesian view for the presence of the random effects in the area level small area estimation. They implemented a spike-and-slab prior distribution for the random effect to investigate whether it is necessary to include random effects in each small area. In particular, they assumed that the random effects follow a non-degenerate and unique distribution with probability p in the small area whereas it is absent from the modelling with probability $1-p$. Their method addresses the disadvantage of the method that was proposed in Datta *et al.* (2011) and gives more flexibility to area level linear models.

Jiang and Nguyen (2012) considered a heteroscedastic nested error regression model in which they treated each small area as a cluster with unknown sampling variance, $\sigma^2_{e_i}$, and unknown $\gamma = \mathrm{var}(u_i)/\mathrm{var}(e_{ij})$, where u_i is the area-specific random effect, to give more flexibility to small area models. Subsequently, they proposed an optimal method for prediction. Heterogeneity of the model that was proposed in Jiang and Nguyen (2012) is due to the heterogeneity of the sampling variances and variance components in small areas whereas γ is assumed to be fixed.

Rigby and Stasinopoulos (2005) introduced generalized additive models for location, scale and shape to give more flexibility in modelling. Generalized additive models for location, scale and shape define different generalized linear mixed models for different model parameters and use a back-fitting algorithm to solve for a proper model. Rigby and Stasinopoulos considered the same variance components for the response variable. However, for the simplest case of the mean and the variance, generalized additive models for location, scale and shape do not take into account the difference between variability of the response variable in clusters of small areas.

In this work, we consider an area level model where we assume that there is no access to unit level data in areas (e.g. due to confidentiality). The sampling variance is assumed to be known whereas the variance of the random effects is unknown. As is often the case, in observational studies, there are different latent variables in different clusters that affect the magnitude of random effects and, consequently, the bias of synthetic estimators. Clustering small areas with similar covariates in terms of the Euclidean distance can be used to take into account the inherent differences between areas and probably increases the precision of the small area mean prediction. Note that this inherent difference comes back to the features of small areas and not of the covariates. There is also no solid mathematical formula for the relationship between variance components and covariates (either increase or decrease) in clusters.

We introduce clustering by a frequentist approach to give more flexibility to the Fay–Herriot model although we do not omit random effects from small area means. Clustering small areas by using the hierarchical clustering technique based on the Euclidean distance between their corresponding covariates is proposed to construct different groups of small areas where, inside each group, areas are homogeneous and areas from different groups or clusters are non-homogeneous. The idea comes from the fact that, in the regression analysis, when the covariates are sufficiently close in terms of the Euclidean distance, we expect associated means of the response variable to be close. However, the fluctuation around the regression line might be from different sources of error due to different latent variables in the clusters. The magnitude of these

fluctuations might also differ. As an example, consider evaluating the effect of body mass index on waste circumference for different groups based on age, sex, poverty, education and ethnicity. Here, it is reasonable to expect small areas that are similar in terms of the Euclidean distance between their covariates to show a similar pattern in terms of the random effects. Obviously, the trend of accumulating body fat for underweight and obese people is different. So, we can easily consider different variance components for different clusters. We start with the assumption that the random effect in each cluster follows a normal distribution with a different variance component to give more flexibility to the behaviour of the random effects. We introduce a test statistic to test the null hypothesis of equality of variance components of the normal distributions. If the null hypothesis is rejected, we implement a modified version of Tukey's method (Tukey, 1949) to combine some clusters. We assess the effect of different distributions of random effects on the precision of small area means predictors by using the mean-squared prediction error (MSPE) and study situations where the methodology proposed results in more reliable predictors of small area means.

The outline of the paper is as follows. In Section 2, we review the general area level model and study clustering in small area estimation. In Section 3, we introduce a test statistic to evaluate the assumption of homogeneity of variance components and prove some of its properties. Moreover, we show how a modification of Tukey's method can be used to combine some clusters. Using the new distributional form of the random effects, we find the empirical best linear unbiased predictor (EBLUP) of small area means in Section 4. We also obtain an approximation to the MSPE of the EBLUP and derive an unbiased estimator of the MSPE up to the second order of approximation. In Section 5, a real data set is analysed. Implementing the simulation studies, we evaluate the performance of our proposed test statistic under various scenarios in Section 6. The precision of our proposed approach in predicting small area means in terms of the MSPE is also assessed in this section. In addition, we study the relative bias (RB) of the estimator of the MSPE. Finally, we give some concluding remarks in Section 7.

2. Complete-clustering approach in area level small area model

Consider the area level regression model

$$y_i = \mathbf{X}_i\beta + u_i + e_i, \qquad i = 1, \ldots, m, \tag{2.1}$$

where y_i is the variable of interest, $\mathbf{X}_i = (1, X_{i1}, X_{i2}, \ldots, X_{ip})$ is the vector of covariates, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ is the vector of regression coefficients, m, u_is and e_is are the number of areas, the area level random effects and the random errors respectively. Also, assume that the u_is are independent and identically distributed (IID) from an $N(0, \sigma_u^2)$ distribution, and the e_is are independent with $e_i \sim N(0, D_i)$, where the D_is are all known and σ_u^2 is unknown. We assume that there is no sample selection bias and the sampling design is not informative (Fay and Herriot, 1979; Pfeffermann and Sverchkov, 2005, 2007).

Random effects u_i in model (2.1) explain the lack of information provided by covariates about small area means, $\theta_i = \mathbf{X}_i\beta + u_i$, for $i = 1, \ldots, m$. In model (2.1), it is usually assumed that the variance component σ_u^2 is the same for all areas. The magnitude of the random-effect part u_i depends on how good the covariates explain small area means. In many applications, however, we might expect random effects for areas with similar covariates in terms of the Euclidean distance, $\|\mathbf{X}_i - \mathbf{X}_j\|_2 = \sqrt{\{\Sigma_{t=1}^{p}(X_{it} - X_{jt})^2\}}$, to show similar behaviour compared with random effects that are associated with other areas. This motivates the use of clustering to form different groups that contain similar small areas. So, using the hierarchical clustering approach, we put small areas into different clusters, C_l, $l = 1, \ldots, k$, such that each cluster contains the most

similar small areas in terms of the Euclidean distance between the values of their corresponding covariates. In this paper, we also suggest the use of a different variance component, $\sigma^2_{u_l}$, for $l=1,\ldots,k$, in different clusters. In other words, we assume that $u_{j_l} \sim^{\text{IID}} N(0,\sigma^2_{u_l})$ for $j_l \in C_l$, the lth cluster, and $j_l=1,\ldots,n_{c_l}$, where n_{c_l} is the number of small areas in the lth cluster. Under this setting, model (2.1) can be represented as

$$y_{j_l}=\mathbf{X}_{j_l}\beta+u_{j_l}+e_{j_l}, \qquad j_l=1,\ldots,n_{c_l}, \quad l=1,\ldots,k, \tag{2.2}$$

where y_{j_l} and $\mathbf{X}_{j_l}$ are the response variable and the covariate vector that are associated with the jth element in the lth cluster respectively. Also, $e_{j_l} \sim N(0, D_{j_l})$ with known D_{j_l}s. Throughout the paper, we call this the complete-clustering approach where an optimal number of clusters, say k, is chosen such that there are no more significant changes in the distance between clusters for $k^* > k$. This makes the corresponding Fay–Herriot model more flexible to catch the true behaviour of the random effects.

Although complete clustering separates small areas into more homogeneous clusters, variance components might still be equal. If this happens, then model (2.2) reduces to model (2.1), and we refer to this as simple clustering. In Section 3, we test the assumption of equality of the variance components in different clusters by introducing a test statistic and showing some of its asymptotic properties. We might also have a situation where variance components are equal for some clusters, but not all. In this case, we combine clusters with equal variance components and the method will be referred to as the combined clustering approach. The MSPE of the simple, combined and complete methods will be calculated in Section 6. As shown in Section 6, the complete- and combined clustering approaches lead to more precise predictions of $\theta_{j_l}=\mathbf{X}_{j_l}\beta+u_{j_l}$, for $j_l=1,\ldots,n_{c_l}$ and $l=1,\ldots,k$ in terms of the MSPE compared with the prediction that is based on the usual area level small area model (2.1).

3. Equality of variance components in different clusters

In this section, we introduce a test statistic to test the following hypothesis regarding the equality of variance components in model (2.2):

$$H_0: \sigma^2_{u_1}=\sigma^2_{u_2}=\ldots=\sigma^2_{u_k} \quad \textit{versus} \quad H_{\text{a}}: \sim H_0.$$

For this, because of the difference between variance components and sampling variances under model (2.2), Kolmogrov's strong law of large numbers is used to define the test statistic. Because of the complexity that the weighted least squares estimate of β introduces in the method-of-moments estimates of variance components, we use ordinary least square (OLS) to estimate β, where the consistency of $\hat{\beta}_{\text{OLS}}$ under model (2.2) is shown in lemma 1. Using $\hat{\beta}_{\text{OLS}}$, the modified method-of-moments (MMM) estimators of variance components are introduced in theorem 1. The asymptotic distributions of these estimators are also found, and the test statistic is constructed accordingly.

Lemma 1. In model (2.2), let $\mathbf{X}=(\mathbf{X}'_1,\ldots,\mathbf{X}'_m)'$ and assume that $\mathbf{X}'\mathbf{X}$ is full rank, the columns of $\mathbf{X}$ are independent and the covariate matrix of $\mathbf{X}$ is independent of u_{j_l}s and e_{j_l}s. Then, the OLS estimator of the regression coefficient, $\hat{\beta}_{\text{OLS}}$, is a consistent estimator of β.

For a proof of lemma 1, see Appendix A.

Theorem 1. Let

$$\hat{\sigma}^2_{u_l}=\frac{1}{n_{c_l}}\sum_{j_l=1}^{n_{c_l}}\{(y_{j_l}-\mathbf{X}_{j_l}\hat{\beta}_{\text{OLS}})^2-D_{j_l}\} \tag{3.1}$$

and assume that $n_{c_l} \to \infty$ as $m \to \infty$, for $l = 1, \ldots, k$. Under the assumptions of lemma 1 and the general model (2.2), the asymptotic distribution of $\hat{\sigma}^2_{u_l}$ as $m \to \infty$ is given by

$$\hat{\sigma}^2_{u_l} \sim N\left\{\sigma^2_{u_l}, \frac{2}{n^2_{c_l}} \sum_{j_l=1}^{n_{c_l}} (\sigma^2_{u_l} + D_{j_l})^2\right\}, \qquad l = 1, \ldots, k. \tag{3.2}$$

For a proof of theorem 1, see Appendix A.

Remark 1. It is worth noting that the estimator of σ^2_u in theorem 1 is asymptotically unbiased. Because of the positive nature of the variance component $\sigma^2_{u_l}$, for $l = 1, \ldots, k$, interest lies in obtaining positive estimates of the variance component. As the underlying distribution in expression (3.2) is a normal distribution, it is likely that $\hat{\sigma}^2_u$ becomes negative. However, one can easily show that for large n_{c_l} (where $m \to \infty$) the probability of observing a negative estimate is negligible, i.e.

$$\Phi\left[-\sigma^2_{u_l} n_{c_l} \Big/ \sqrt{\left\{2 \sum_{j_l=1}^{n_{c_l}} (\sigma^2_{u_l} + D_{j_l})^2\right\}}\right] \to 0, \tag{3.3}$$

provided that $\sigma^2_{u_l}$ and D_{j_l}, for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$, are bounded.

Remark 2. Estimators of $\sigma^2_{u_l}$, for $l = 1, \ldots, k$, proposed in distribution (3.2) are based on the method of moments. They are unbiased and consistent estimators. However, their variances depend highly on the number of small areas in each cluster. To guarantee a precise estimate of variance components, we suggest implementing this method when the number of small areas is sufficiently large in different clusters.

Let $\hat{\sigma}^2_u = (\hat{\sigma}^2_{u_1}, \ldots, \hat{\sigma}^2_{u_k})$. Under the null hypothesis, $H_0 : \sigma^2_{u_1} = \sigma^2_{u_2} = \ldots = \sigma^2_{u_k} = \sigma^2_u$, we have

$$\hat{\sigma}^2_u \sim N_k(\sigma^2_0, \Sigma_0), \tag{3.4}$$

where $\sigma^2_0 = (\sigma^2_u, \ldots, \sigma^2_u)$ and

$$\Sigma_0 = \operatorname{diag}\left\{\frac{2}{n^2_{c_1}} \sum_{j_1=1}^{n_{c_1}} (\sigma^2_u + D_{j_1})^2, \ldots, \frac{2}{n^2_{c_k}} \sum_{j_k=1}^{n_{c_k}} (\sigma^2_u + D_{j_k})^2\right\}.$$

Thus,

$$(\hat{\sigma}^2_u - \sigma^2_0)' \Sigma_0^{-1} (\hat{\sigma}^2_u - \sigma^2_0) \sim \chi^2_{k-p-1}, \tag{3.5}$$

and the p-value of the test is approximately equal to

$$\mathbb{P}(\chi^2_{k-p-1} > \chi^2_0), \tag{3.6}$$

where $\chi^2_0 = (\hat{\sigma}^2_u - \hat{\sigma}^2_0)' \hat{\Sigma}_0^{-1} (\hat{\sigma}^2_u - \hat{\sigma}^2_0)$, $\hat{\sigma}^2_0 = (\hat{\sigma}^2_{0u}, \ldots, \hat{\sigma}^2_{0u})$, $\hat{\sigma}^2_{0u} = (1/k) \Sigma_{l=1}^k \hat{\sigma}^2_{u_l}$ and

$$\hat{\Sigma}_0 = \operatorname{diag}\left\{\frac{2}{n^2_{c_1}} \sum_{j_1=1}^{n_{c_1}} (\hat{\sigma}^2_{0u} + D_{j_1})^2, \ldots, \frac{2}{n^2_{c_k}} \sum_{j_k=1}^{n_{c_k}} (\hat{\sigma}^2_{0u} + D_{j_k})^2\right\}.$$

In Section 6, we evaluate the performance of distribution (3.5) under various scenarios. The results indicate that the test has a large power. However, the simulated values of the type I error are larger than the level of significance. Note that distribution (3.2) depends highly on the number of small areas in each cluster. For larger clusters, it is expected to obtain more accurate estimates of variance components. However, because of theorem 1, estimates that are obtained

from larger clusters might be significantly different from estimates in smaller clusters. In other words, the difference between the information that is provided from clusters based on different small areas might result in the rejection of the null hypothesis even though it is correct, which explains inflated observed type I error in our simulation studies.

3.1. Combined clustering approach

If the null hypothesis of equality of variance components is rejected, there might be some clusters that have the same variance components. To estimate fewer numbers of parameters, one might consider combining such clusters. This is similar to what happens after rejecting the null hypothesis in the analysis-of-variance context. Tukey (1949) proposed a solution to this problem by using a combination of T- and F-tests. In this paper, we modify his approach to combine the clusters with the same variance components. For this, we take the following steps.

(a) We first sort the MMM estimates of the variance components.
(b) Considering distribution (3.2) and conducting a T-test under an unequal variance set-up and using the same significance level as in p-value (3.6), we make groups of clusters. For this, starting from the cluster with the smallest variance estimate, we compare it with the cluster with the second smallest variance estimates. If the null hypothesis of equality of the related variance components, using distribution (3.2), is not rejected, we make a new group consisting of corresponding clusters. Otherwise, we keep the cluster corresponding to the smallest variance component as a group with a single element. Then, the second smallest number is compared with the third smallest. Similarly, if the null hypothesis of equality is not rejected, we add the corresponding cluster to the group that the second cluster belongs to. Similarly to Tukey (1949), this process stops if all groups are constructed by one or two clusters. If not, we go to the next step.
(c) For a group with number of clusters larger than 2, we find the maximum and the average values of the MMM estimates of the variance components, $\hat{\sigma}^2_{\max}$ and $\hat{\sigma}^2_{\text{mean}}$ respectively. Following Tukey (1949) and depending on the number of clusters in each group, say k', we construct a test statistic W as follows:

$$W = \begin{cases} \frac{1}{3}\left(0.25 + \frac{1}{n_{c_{\max}}}\right)^{-1} \left\{ \frac{\hat{\sigma}^2_{\max} - \hat{\sigma}^2_{\text{mean}}}{\text{var}(\hat{\sigma}^2_{\max} - \hat{\sigma}^2_{\text{mean}})} - 1.2\log_{10}(k') \right\}, & k' > 3, \\ \frac{1}{3}\left(0.25 + \frac{1}{n_{c_{\max}}}\right)^{-1} \left\{ \frac{\hat{\sigma}^2_{\max} - \hat{\sigma}^2_{\text{mean}}}{\text{var}(\hat{\sigma}^2_{\max} - \hat{\sigma}^2_{\text{mean}})} - 0.5 \right\}, & k' = 3, \end{cases}$$

where $n_{c_{\max}}$ is the number of small areas inside the cluster corresponding to $\hat{\sigma}^2_{\max}$. Here

$$\text{var}(\hat{\sigma}^2_{\max} - \hat{\sigma}^2_{\text{mean}}) = \left(1 - \frac{1}{k'}\right)^2 \text{var}(\hat{\sigma}^2_{\max}) + \left(\frac{1}{k'}\right)^2 \sum_{j \neq j_{\max}} \text{var}(\hat{\sigma}^2_j)$$

where $j_{\max}$ is the index related to the $\hat{\sigma}^2_{\max}$ and $\text{var}(\hat{\sigma}^2_j)$s are obtained by using the estimated values of the variance in distribution (3.2).

The aim is to see whether or not we can split a group of clusters into smaller clusters. If W is larger than the critical value of the standard normal distribution for the two-sided test with the level of significance in p-value (3.6), we put the cluster corresponding to the maximum variance component into a new group. We repeat this step for the new maximum if $k' > 2$. If the new maximum should be separated as well, we put it in the same group as the old maximum.

(d) When the number of clusters in a group remains the same in the previous step, depending on the number of clusters inside a group, we test the assumption of equality of the variance components for the group with size larger than 2. If the size of the group is larger than $p+1$, we implement the test statistic that is proposed in this paper. Otherwise, simultaneous T-tests with the same significance level as in p-value (3.6) for a small p are conducted. If the null hypothesis is rejected, we split the group with an even number of clusters into subgroups of two clusters by starting from the smallest MMM estimate and moving forwards to the largest. For the odd number size, we let the last subgroup have three clusters and then test the assumption of equality of the variance components one more time. If the null hypothesis is rejected, we make two new subgroups of two and one clusters starting from the cluster corresponding to the smallest MMM estimate.

After implementing this algorithm, some of the clusters will be combined. Hence, to estimate the MSPE, denoted by mspe, we need to deal with fewer parameters. This results in less variability because of the estimation of the model parameters in the estimation of small area means and the corresponding MSPEs. In Section 6, it is shown that, when there are clusters with the same variance component, the combined cluster has the same performance as the complete cluster in estimating small area means, but significantly better than simple clustering. However, the results of Section 6 indicate that, if the difference between variance components is huge, complete clustering performs better in estimating small area means compared with combined and simple clustering approaches regardless of the number of parameters to be estimated.

In the next section the EBLUPs of small area means are given. Following Prasad and Rao (1990), we also provide an approximation to the MSPE of the EBLUP and derive an estimator of the MSPE of the EBLUP which is a second-order unbiased approximation.

4. Empirical best linear unbiased predictor and its mean-squared prediction error estimation

In this section, the BLUP of small area means and, consequently, the EBLUP for the new distributional form of random effects are obtained. Following Henderson (1950), the BLUP of small area means is given by

$$\tilde{\theta}_{j_l} = \mathbf{X}_{j_l}\tilde{\beta} + \tilde{u}_{j_l}, \tag{4.1}$$

where $\tilde{u}_{j_l} = \mathbf{G}\mathbf{V}^{-1}(y_{j_l} - \mathbf{X}_{j_l}\tilde{\beta})$, $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$, $\mathbf{y} = (y_1, \ldots, y_m)$, $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_k)$, $\mathbf{V}_l = \text{diag}(\sigma_{u_l}^2 + D_{1_l}, \ldots, \sigma_{u_l}^2 + D_{n_{c_l}})$, $\mathbf{G} = \text{diag}(\mathbf{G}_1, \ldots, \mathbf{G}_k)$, $\mathbf{G}_l = \sigma_{u_l}^2 \mathbf{I}_{n_{c_l} \times n_{c_l}}$ and $\mathbf{I}$ is the identity matrix for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. The MSPE of $\tilde{\theta}_{j_l}$ can be written as

$$\begin{aligned} \text{MSPE}(\tilde{\theta}_{j_l}) &= E(\tilde{\theta}_{j_l} - \theta_{j_l})^2 \\ &= g_{1j_l}(\gamma) + g_{2j_l}(\gamma), \end{aligned}$$

where $\gamma = (\sigma_{u_1}^2, \ldots, \sigma_{u_k}^2)$, $g_{1j_l}(\gamma)$ is the jth element of the lth cluster on the diagonal of $\mathbf{G} - \mathbf{G}\mathbf{V}^{-1}\mathbf{G}$, $g_{2j_l}(\gamma) = \mathbf{d}_{j_l}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{d}_{j_l}$, $\mathbf{d}_{j_l}' = \mathbf{X}_{j_l} - \mathbf{b}_{j_l}'\mathbf{X}_{j_l}$ and $\mathbf{b}_{j_l}'$ is the jth row of the lth cluster of $\mathbf{G}\mathbf{V}^{-1}$.

Because of unknown γ, $\hat{\theta}_{j_l}$s, the EBLUPs of θ_{j_l}s, are obtained for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. For this, the MMM estimators of variance components are used in formula (4.1). The MSPE of the EBLUP can be decomposed as

$$\begin{aligned} \text{MSPE}(\hat{\theta}_{j_l}) &= E(\hat{\theta}_{j_l} - \theta_{j_l})^2 \\ &= E(\tilde{\theta}_{j_l} - \theta_{j_l})^2 + E(\hat{\theta}_{j_l} - \tilde{\theta}_{j_l})^2 + 2E(\hat{\theta}_{j_l} - \tilde{\theta}_{j_l})(\tilde{\theta}_{j_l} - \theta_{j_l}). \end{aligned}$$

Under the normality assumption for random effects as well as the sampling error, the cross-product term is 0 (Rao and Molina, 2015). Therefore

$$\mathrm{MSPE}(\hat{\theta}_{jl}) = g_{1jl}(\gamma) + g_{2jl}(\gamma) + g_{3jl}(\gamma),$$

where

$$g_{3jl}(\gamma) = \mathrm{tr}\left\{ \left(\frac{\partial \mathbf{b}'_{jl}}{\partial \gamma} \right) \mathbf{V} \left(\frac{\partial \mathbf{b}'_{jl}}{\partial \gamma} \right) \mathrm{var}(\gamma) \right\}, \tag{4.2}$$

$$\frac{\partial \mathbf{b}'_{jl}}{\partial \gamma} = \begin{pmatrix} \frac{\partial \mathbf{b}'_{jl1}}{\partial \sigma^2_{u_1}} & \cdots & \frac{\partial \mathbf{b}'_{jlm}}{\partial \sigma^2_{u_1}} \\ \frac{\partial \mathbf{b}'_{jl1}}{\partial \sigma^2_{u_2}} & \cdots & \frac{\partial \mathbf{b}'_{jlm}}{\partial \sigma^2_{u_2}} \\ & \vdots & \\ \frac{\partial \mathbf{b}'_{jl1}}{\partial \sigma^2_{u_k}} & \cdots & \frac{\partial \mathbf{b}'_{jlm}}{\partial \sigma^2_{u_k}} \end{pmatrix}_{k \times m},$$

and $\mathrm{var}(\gamma) = \Sigma_0$.

Prasad and Rao (1990) gave the second-order MSPE estimate of small area means as the measure of the variability of the EBLUP as

$$\mathrm{mspe}(\hat{\theta}_{jl}) \approx g_{1jl}(\hat{\gamma}) + g_{2jl}(\hat{\gamma}) + 2g_{3jl}(\hat{\gamma}), \tag{4.3}$$

where $\hat{\gamma}$ is the consistent estimate of γ and $g_{1jl}(\hat{\gamma})$, $g_{2jl}(\hat{\gamma})$, $g_{3jl}(\hat{\gamma})$, $\hat{\mathbf{d}}_{jl}$, $\widehat{\partial \mathbf{b}'_{jl}/\partial \gamma}$ and $\widehat{\mathrm{var}(\gamma)} = \hat{\Sigma}_0$ are obtained by substituting $\hat{\gamma}$ in their original definition. Let $\hat{\mathbf{V}}_l = \mathrm{diag}(\hat{\sigma}^2_{u_l} + D_{1_l}, \ldots, \hat{\sigma}^2_{u_l} + D_{n_{c_l}})$ and $\hat{\mathbf{G}}_l = \hat{\sigma}^2_{u_l} \mathbf{I}_{n_{c_l} \times n_{c_l}}$, for $l = 1, \ldots, k$. Accordingly, let $\hat{\mathbf{V}} = \mathrm{diag}(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_k)$ and $\hat{\mathbf{G}} = \mathrm{diag}(\hat{\mathbf{G}}_1, \ldots, \hat{\mathbf{G}}_k)$. Note that Prasad and Rao (1990) proposed approximation (4.3) for the consistent estimators of γ. As the MMM estimates of the variance components are consistent, we use them for data analysis and simulation studies.

It is worth mentioning that the magnitude of g_{3jl} depends on the number of variance components in the model. So, the more parameters we have in the model, the larger g_{3jl} will be. However, the magnitudes of g_{1jl} and g_{2jl} decrease significantly such that we gain improvement in terms of the overall MSPE. To evaluate the performance of the estimator of MSPE, we use the RB as the measure of the precision.

5. Real data analysis

In this section, we use the National Health and Nutrition Examination Survey for 2011–2012 as a unit level data set to predict waist circumference on the basis of the body mass index. Vague (1947) showed that there is a high association between waist circumference and cardiovascular disease, type 2 diabetes and also hypertension. He concluded that 'apple-shaped obesity' observed in men is a high-risk obesity whereas 'gynoid obesity', which is often found in women, has a lower risk.

The data are categorized in small domains by using the following variables:

(a) age groups as 20.00–33.00, 33.00–48.00, 48.00–48.56, 48.56–63.00 and 63.00–80.00 years;
(b) gender as male and female;
(c) education as the highest grade or level of education completed by adults 20 years old and older grouped as less than ninth-grade education, 9th–11th-grade education (includes

12th grade and no diploma), high school graduate or general education development, some college or associates degree, and college graduate or higher;
(d) ethnicity as Mexican, other Hispanic, non-Hispanic white, non-Hispanic black, non-Hispanic Asian, and other race—including multiracial;
(e) poverty groups as 0.00–0.97, 0.97–1.88, 1.88–2.41, 2.41–4.03 and 4.03–5.00.

We consider the mean of waist circumference of people belonging to a small area as the response variable y_i, and the mean of their body mass index as the covariate x_i. The units inside each small area are used to obtain the sampling variance D_i. For this, we propose the following method which is an extension of the method in Wang and Fuller (2003).

Suppose that we have enough sampled units n_i such that the regression line $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + e_{ij}$ is estimable, where e_{ij} is the sampling error, $n_i > 1$ is the number of sampled units in each small area and $j = 1, \ldots, n_i$. Let

$$y_i = \beta_0 + \beta_1 x_i + u_i + e_i,$$

where $e_i = (1/n_i)\Sigma_{j=1}^{n_i} e_{ij}$, $x_i = (1/n_i)\Sigma_{j=1}^{n_i} x_{ij}$ and $y_i = (1/n_i)\Sigma_{j=1}^{n_i} y_{ij}$ and $i = 1, \ldots, 1362$. First, the interest lies in the estimation of the sampling variance $\sigma_{e_i}^2$. An unbiased estimate of the sampling variance inside the small area, $\hat{\sigma}_{e_i}^2$, is given by

$$\frac{1}{n_i - 1} \sum_{j=1}^{n_i} \{(y_{ij} - y_i)^2 - \beta_1^2 (x_{ij} - x_i)^2\}. \tag{5.1}$$

Note that β_1 is unknown. To fit a regression line in each area, at least two sampled units are required. As there are small areas with only one sample unit, we use the overall regression line that is obtained from the complete data to estimate $\hat{\sigma}_e^2$. Consequently, the sampling variance for the mean of the response variable in each area is given by

$$D_i = \hat{\sigma}_{e_i}^2 / n_i. \tag{5.2}$$

The D_is range from 4.58×10^{-16} to 61.63. There are 519 small areas with $n_i = 1$ where the overall regression line is used to estimate $D_i = 61.63$.

There are a few approaches to cluster small areas. For example, k-means clustering is an approach that minimizes the Euclidean distance of the covariate from the mean of covariates belonging to a cluster (Hartigan and Wong, 1979). Another approach is to use hierarchical clustering (Ward, 1963) based on the Lance–Williams algorithm through the squared Euclidean distance between covariates. The silhouette method (Rousseeuw, 1987) also gives a suitable number of clusters based on the average of silhouettes. It is worth mentioning that we do not look for the optimal number of clusters (i.e. the smallest number of clusters) as the aim is to have small areas inside a cluster that are as similar as possible. Hence, we choose the number of clusters such that the distance between clusters does not change significantly after that. Even if the number of clusters is far from the optimum value, using Tukey's method, we can always combine clusters which are not significantly different. So, as Fig. 1 shows, we cluster small areas in seven groups by using hierarchical clustering and model data as

$$y_{j_l} = \beta_0 + \beta_1 x_{j_l} + u_{j_l},$$

where $j_l = 1, \ldots, n_{c_l}, l = 1, \ldots, 7$ and $n_c = (165, 521, 333, 64, 252, 9, 18)$. Fig. 2 shows a boxplot of the response variable in different clusters. As we observe, the variance of clusters is different. The MMM estimates of the variance components are $(11.40, 0.36, 2.87, 4.05, 1.76, 420.77, 36.17)$.

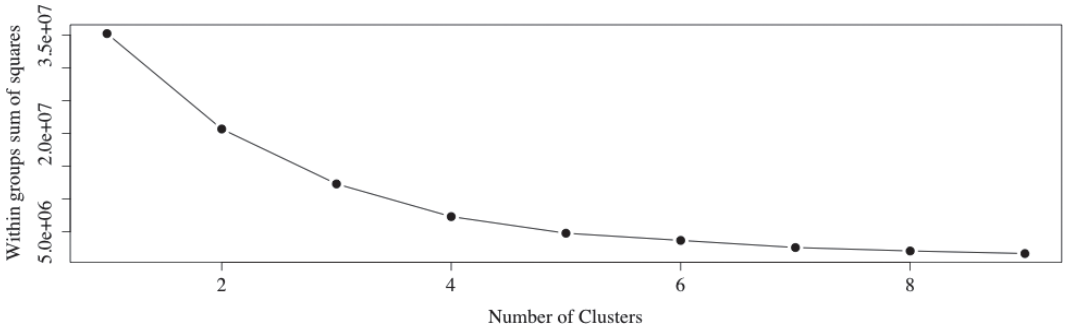

Fig. 1. Effect of number of clusters on within-groups sums of squares in the *k*-means clustering

Table 1. Comparison between complete, combined and simple clustering in terms of the quantiles of their corresponding mspe

Statistic	*Complete over simple*	*Combined over simple*	*Complete over combined*
Minimum	0.00	0.00	0.28
1-decile	0.86	1.00	0.90
2-decile	1.00	1.06	0.94
3-decile	1.49	1.75	0.98
4-decile	2.59	2.95	1.00
5-decile	3.06	3.14	1.00
6-decile	5.48	7.29	1.00
7-decile	10.95	11.08	1.00
8-decile	15.09	11.14	1.00
9-decile	62.24	62.22	1.17
Maximum	62.65	62.64	1.60
Mean	12.43	12.39	0.98

The assumption of equality of the variance components is rejected with $\chi^2_4 = 90.80$ with a corresponding p-value of 0. The residual is defined as

$$\epsilon_{j_l} = y_{j_l} - \hat{\theta}_{j_l}$$

where $\hat{\theta}_{j_l}$ is obtained by using equation (4.1) and substituting the unknown parameters from either the simple or complete-clustering methods for $j_l = 1, \ldots, n_{c_l}, l = 1, \ldots, 7$. Figs 3 and 4 show boxplots of the residual variable belonging to different clusters by using the simple and complete methods respectively. After implementing complete clustering, the behaviour of residuals in different clusters is more homogeneous in Fig. 4. By using Tukey's method, we have five combined clusters where the modified number of small areas in the clusters are $n_c = (165, 521, 649, 9, 18)$ with the variance components $(11.40, 0.36, 2.89, 420.77, 36.17)$.

Using the complete-clustering approach, explained in model (2.2) where we derived estimates of variance components by using equation (3.1), we observe more than 49% improvements over the simple approach (the Fay–Herriot model) in mspe in 70% of small areas. In 10.00% of small areas, mspe of the simple clustering approach is up to 62.24 times more than mspe of the complete-clustering approach. On average we obtain an mspe of the simple clustering approach up to 12.43 times more than the complete-clustering approach. After implementing Tukey's method (Section 3.1), there is over 75% improvement in 70% of small areas. The complete- and

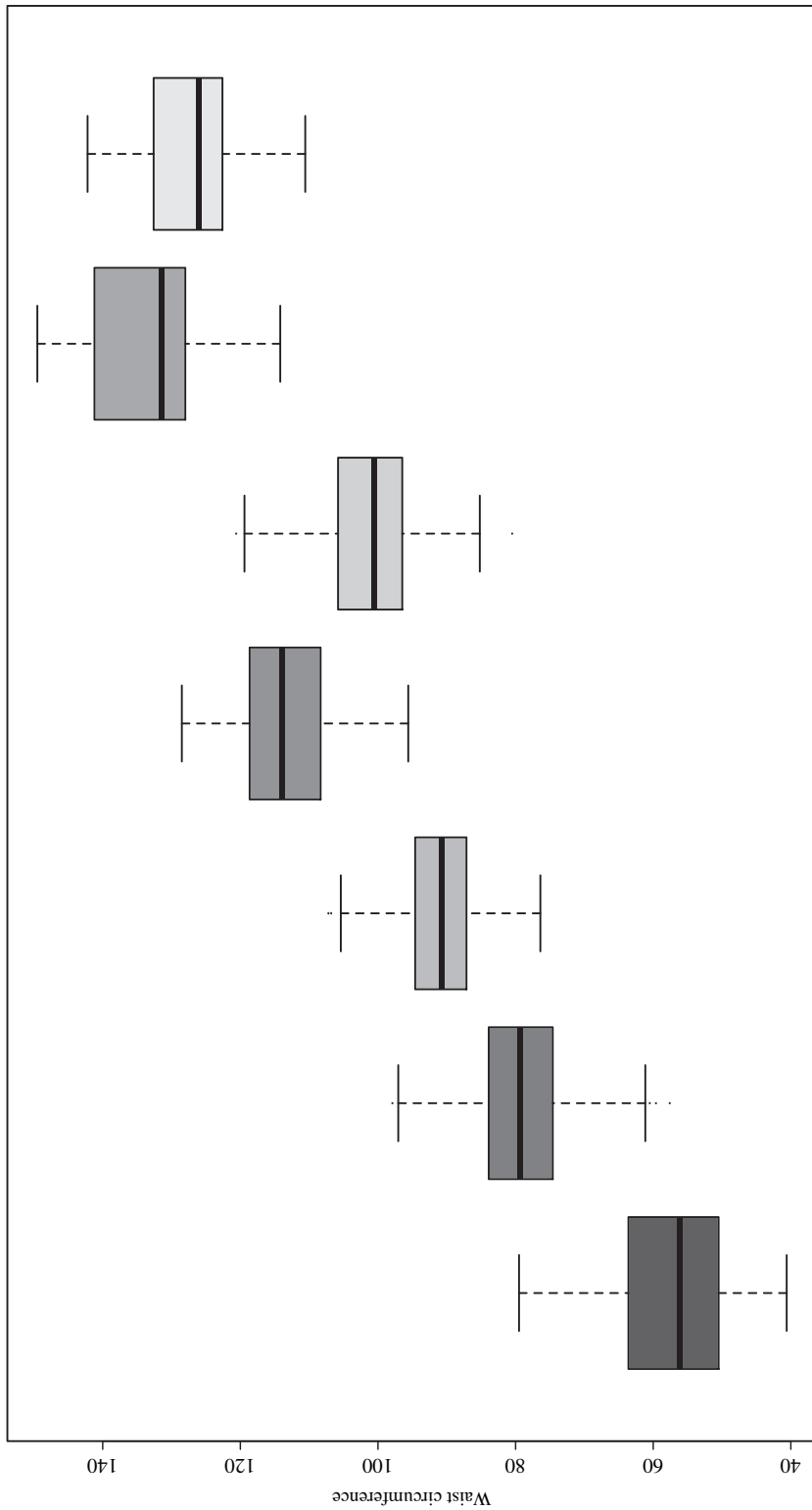

Fig. 2. Boxplot of the response variable (waist circumference) in various clusters after implementing the hierarchical method

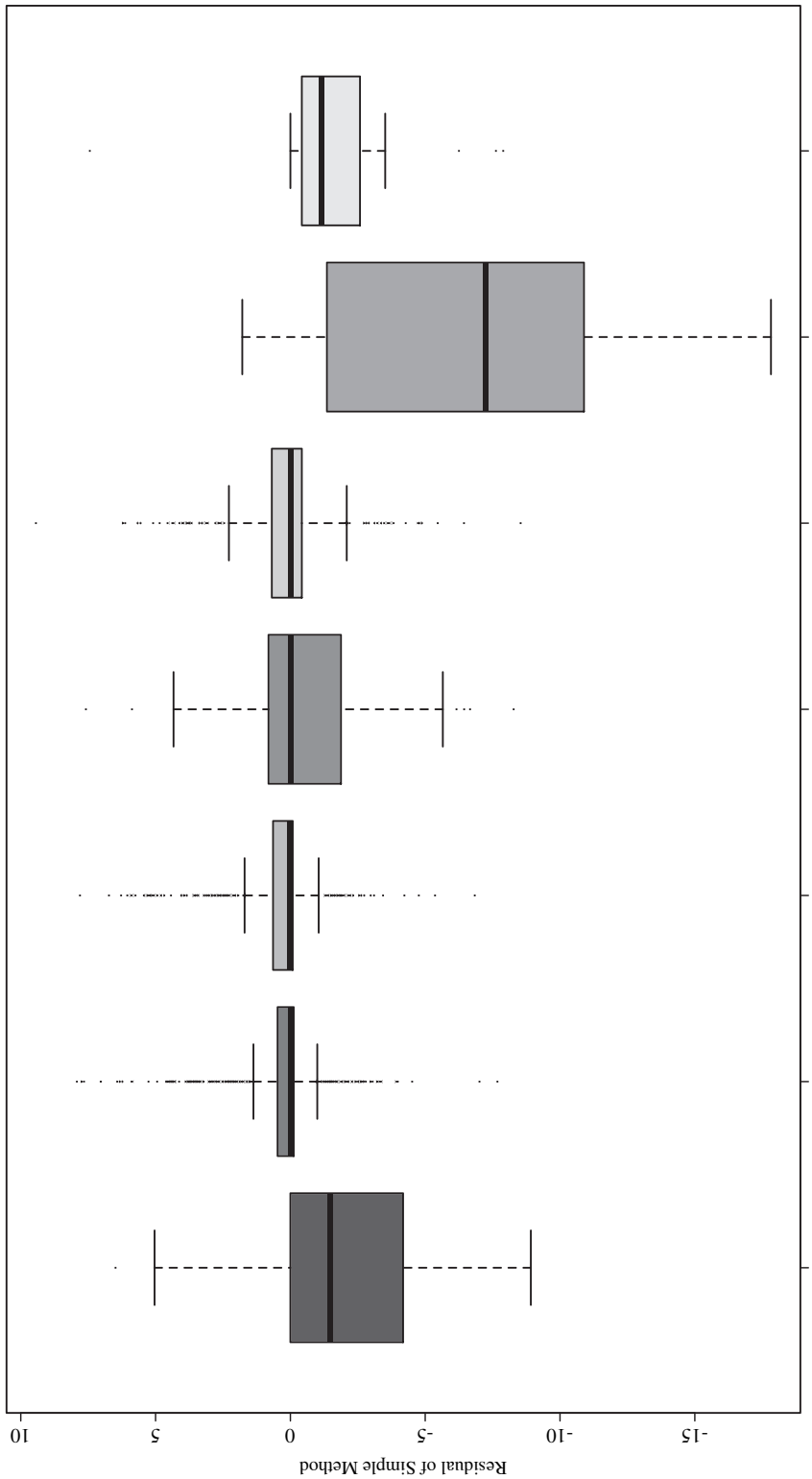

Fig. 3. Boxplot of the residual variable belonging to various clusters after implementing the simple method (the Fay–Herriot model)

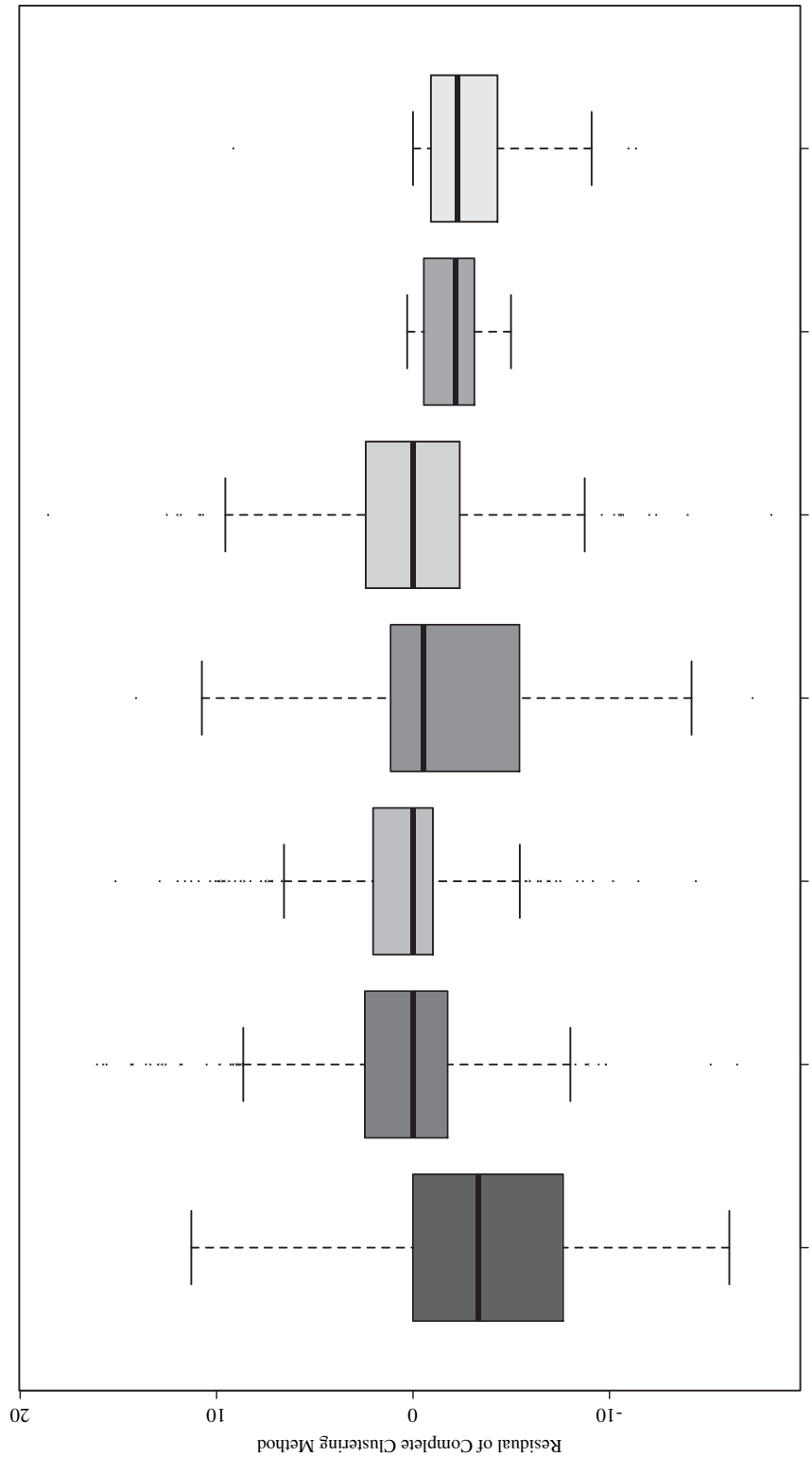

Fig. 4. Boxplot of the residual variable belonging to various clusters after implementing the complete method

combined clustering approaches have the same performance in over 80% of small areas in terms of mspe. To compare the performance of the proposed methods with the simple clustering approach, we first calculate mspe of the complete-clustering, mspe_c, combined clustering, mspe_{cb}, and the simple approach, mspe_s. Then, $\text{mspe}_s/\text{mspe}_c$, $\text{mspe}_s/\text{mspe}_{cb}$ and $\text{mspe}_{cb}/\text{mspe}_c$ are calculated. An overall comparison of three methods is given in Table 1.

As the MMM estimates of the variance components are significantly different, use of either the complete- or combined clustering schemes is justifiable. However, the MMM estimates of the variance components for three clusters, 2.87, 4.05 and 1.76, are very close. Because of this similarity, the combined clustering scheme merges the corresponding clusters.

Figs 5 and 6 present the mspe- and predicted values of the small area means respectively.

6. Simulation studies

In this section, we consider various scenarios of the sampling variances, D_{jl}s, and the variance components, $\sigma^2_{u_l}$, for $j_l = 1, \ldots, n_{cl}$ and $l = 1, \ldots, k$ to evaluate the performance of the test statistic. We also design simulation studies to see the performance of the proposed method in the reduction of the MSPE. For this, the empirical MSPE (EMSPE) of small area means by using different clustering schemes is calculated. Further, we evaluate the performance of the complete-

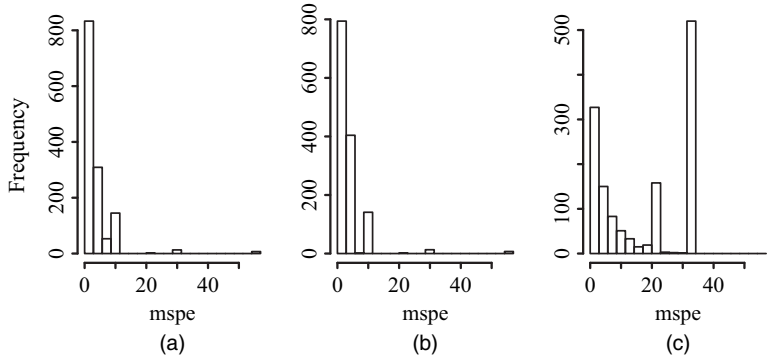

Fig. 5. Histogram of mspes: (a) complete-clustering method; (b) combined clustering method; (c) simple method

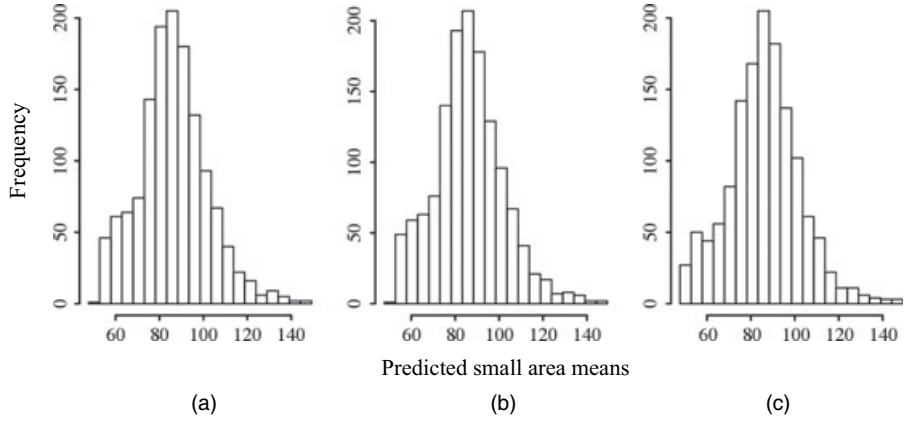

Fig. 6. Histogram of predicted small area means: (a) complete-clustering method; (b) combined clustering method; (c) simple method

and combined clustering approaches on the estimation of the MSPE, mspe, using the RB. In Section 5, an unbiased estimator of the sampling variance inside the small area was proposed. We assess the performance of this estimator by using simulation studies.

6.1. Evaluating the test statistic and the effect of complete and combined clustering on the empirical mean-squared prediction error and mspe

For the simulation studies, we use an area level data set regarding the prescription costs from the Union Régionale des Caisses d'Assurance Maladie of the Midi-Pyrénées Region in the south-west of France, during the period January–December 1999. The data set consists of $m = 268$ cantons (a type of administrative division of a country) that are considered as small areas and the goal is to predict the average prescription cost in each area. In general, cantons are relatively small in terms of population size when compared with counties, departments or provinces. Because of the confidentiality issue and the privacy concerns, the data set is only available in the area level format. This data set has been used by Cressie *et al.* (2005, 2006) to assess the performance of their proposed estimators in the context of the spatial model. Kang *et al.* (2009) also considered this data set in a spatial analysis to predict the average prescription amount in each canton.

In this work, we consider small area estimation to conduct simulation studies using this data set. The area level small area model provides an appropriate link between different small areas to aggregate the information from all small areas to predict the small area mean. The idea is different from Kang *et al.* (2009) as, in the spatial model, cantons are considered to be dependent whereas in our set-up we assume that they are independent. Following Kang *et al.* (2009), we use the percentage of patients over 70 years as the covariate in each canton. We consider the average prescription amount in each canton as the response variable.

Cantons can be clustered on the basis of the percentage of patients over 70 years. We expect patients in similar age groups to have reasonably similar prescription costs as they need similar numbers of follow-up visits, etc. Another important factor is the type of medication that is taken by the patient. Apparently, people in similar age groups need similar supplementary treatments (Speros, 2009).

Similarly to the approach in Section 5, we initially obtain $k = 10$ clusters. However, using Tukey's method, at the end of the analysis, we obtain an updated number of clusters.

Consider $n_c = (33, 37, 32, 14, 34, 65, 4, 13, 18, 18)$. Let x_{j_l} denote the jth covariate in the lth cluster for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. Throughout this section, different scenarios for D_{j_l}s and $\sigma^2_{u_l}$s are considered. Generally, D_{j_l}s are generated from a uniform distribution with different ranges, $U(0.25, 0.5)$, $U(0.25, 1.5)$, $U(0.25, 2.5)$ and $U(0.25, 3.5)$, for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. The response random variable is simulated following two steps. First, let $\theta_{j_l} \sim N(\mathbf{X}_{j_l}\beta, \sigma^2_{u_l})$ where $\mathbf{X}_{j_l} = (1, x_{j_l})$ and $\beta = (2.8093, 0.0059)$ that is obtained by regressing the prescription cost on the percentage of patients over 70 years old. We also choose different and arbitrary sets of $\sigma^2_{u_l}$ to evaluate the performance of the MMM estimates, the hypothesis testing procedure and three methods of clustering, complete, combined and simple, on the EMSPE and mspe. Then, $y_{j_l} \sim N(\theta_{j_l}, D_{j_l})$, for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. We generate $R = 5000$ simulations of the response random variables.

In theorem 1, the MMM estimate of the variance components is introduced. Throughout this chapter, we work with the MMM estimates of the variance components rather than their restricted maximum likelihood estimates to predict small area means and also estimate the MSPE. We performed simulation studies to compare our proposed estimates of the variance components with the restricted maximum likelihood estimates of the variance components (Cressie, 1992; Rao and Molina, 2015) in terms of mean-squared error (MSE). The aim is to

show that, although MMM estimators of the variance components are based on the MMM, their performance is comparable with the restricted maximum likelihood estimates of the variance components. Tables 2 and 3 show that the two methods have almost the same performance by using the MSE criterion. As Table 3 shows, the MSEs for the same value of the variance component are different since we have different cluster sizes even for two clusters with the same variance components. According to Tables 2 and 3, we expect to obtain a smaller MSE of variance components for larger cluster sizes.

To calculate the power of the test proposed, the value of the test statistic is calculated for each simulated response variable. After finding the corresponding p-value, we reject the null hypothesis of equality of variance components if the p-value is less than the predetermined level of significance ($\alpha = 0.05$). The average of the number of times that the null hypothesis is rejected when the alternative hypothesis is correct is used to evaluate the power of the test. Table 4 gives the power of the test under various set-ups.

As is shown in Table 4, when the difference between D_{jl}s and $\sigma^2_{u_l}$s becomes larger, the test becomes more powerful. Moreover, when the similarities between the $\sigma^2_{u_l}$s, for $l = 1, \ldots, k$, increase, the power of the test decreases. Generally, removing clusters from the analysis may

Table 2. MSE of the restricted maximum likelihood and MMM estimates for different variance components

σ^2_u	n_c	*MSE, restricted maximum likelihood*	*MSE, MMM*
1	33	0.07	0.09
8	37	2.38	2.37
25	32	27.00	26.57
5	14	2.67	2.73
64	34	159.23	155.73
10	65	2.03	2.03
49	4	767.33	702.52
36	13	123.32	119.69
40	18	118.05	112.03
13	18	12.91	12.80

Table 3. MSE of the restricted maximum likelihood and MMM estimates for similar variance components in some clusters

σ^2_u	n_c	*MSE, restricted maximum likelihood*	*MSE, MMM*
1	33	0.07	0.10
8	37	2.39	2.40
25	32	27.01	26.60
25	14	58.25	55.83
64	34	159.25	155.78
25	65	12.15	12.08
49	4	772.09	704.98
40	13	152.07	147.54
40	18	118.75	112.08
13	18	13.02	12.83

Table 4. Power of the test statistic for various values of σ_u^2 and D_{j_l}

σ_u^2	D_{j_l}			
	U(0.25, 0.5)	*U(0.25, 1.5)*	*U(0.25, 2.5)*	*U(0.25, 3.5)*
(1, 8, 25, 5, 64, 10, 49, 36, 40, 13)	1	1	1	1
(1, 8, 25, 25, 64, 25, 49, 40, 40, 13)	1	1	1	1
(1, 1, 2, 4, 2.5, 4, 3, 3, 5, 6)	1	0.97	0.90	0.86
(1.5, 1.5, 2.4, 6, 4, 4, 5.9, 4, 6, 4)	0.96	0.89	0.79	0.74
(0.1, 0.5, 0.2, 0.5, 0.25, 0.4, 0.49, 0.3, 0.4, 0.13)	0.50	0.26	0.19	0.22
(0.1, 0.1, 0.2, 0.15, 0.25, 0.3, 0.12, 0.3, 0.09, 0.13)	0.36	0.21	0.17	0.20

Table 5. Effect of removing clusters on the power of the test statistic for the case $\sigma_u^2 = (1, 1, 2, 4, 2.5, 4, 3, 3, 5, 6)$

	D_{j_l}			
	U(0.25, 0.5)	*U(0.25, 1.5)*	*U(0.25, 2.5)*	*U(0.25, 3.5)*
Removing the largest cluster	0.99	0.95	0.87	0.82
Removing the smallest cluster	1	0.98	0.92	0.87
Removing two largest clusters	0.94	0.87	0.75	0.70
Removing three largest clusters	0.95	0.89	0.78	0.72
Removing four largest clusters	0.69	0.60	0.53	0.48

increase or decrease the power of the test (Table 5). For instance, looking at the second row of Table 5, removing the smallest cluster with four small areas increases the power. As this cluster has the same variance component as the eighth, we expect to have higher power because of the more distinct values of the variance components for the remaining clusters. Obviously, removing a large cluster decreases the power of the test more significantly. The simulation study indicates that in this scenario the test rejects the null hypothesis more than the predetermined level of significance $\alpha = 0.05$ (Table 6). As is explained in Section 2, this is because the precision of the estimate of the $\sigma_{u_l}^2$s, for $l = 1, \ldots, k$, depends highly on the number of small areas in each cluster. In our set-up, the seventh cluster contains four small areas whereas the sixth cluster contains 65 small areas. So, even though the variance components are the same in both, the null hypothesis of equality might be rejected because of the difference between n_{c_6} and n_{c_7}.

We also perform simulation studies to evaluate the performance of our proposed method in terms of the EMSPE. For this, we calculate the EBLUP of small area means, $\hat{\theta}_{j_l}^{(r)}$, by finding the MMM estimates of the variance components and substituting them in expression (4.1). The EMSPE is

$$\text{EMSPE}(\theta_{j_l}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{j_l}^{(r)} - \theta_{j_l}^{(r)})^2, \qquad \text{for } j_l = 1, \ldots, n_{c_l}, \quad l = 1, \ldots, k,$$

where $\theta_{j_l}^{(r)}$ is the small area mean in the rth iteration. Figs 7 and 8 show the EMSPE that is obtained by using different methods for different set-ups.

Table 6. Percentage of times the null hypothesis is rejected by mistake

σ_u^2	D_{jl}			
	U(0.25, 0.5)	*U(0.25, 1.5)*	*U(0.25, 2.5)*	*U(0.25, 3.5)*
(0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1)	0.14	0.15	0.13	0.13
(1, 1, 1, 1, 1, 1, 1, 1, 1, 1)	0.15	0.16	0.15	0.18
(1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2, 1.2)	0.15	0.16	0.15	0.18
(2, 2, 2, 2, 2, 2, 2, 2, 2, 2)	0.15	0.15	0.14	0.15
(3, 3, 3, 3, 3, 3, 3, 3, 3, 3)	0.15	0.16	0.15	0.18
(20, 20, 20, 20, 20, 20, 20, 20, 20, 20)	0.15	0.16	0.15	0.16

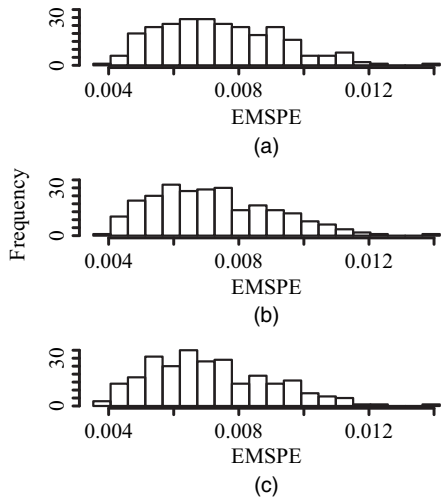

Fig. 7. Histogram of the EMSPE for $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$: (a) simple method; (b) combined method; (c) complete method

We compare the EMSPE of the complete and combined clustering, EMSPE_c and EMSPE_cb, with the simple method, EMSPE_s, by finding for example the following ratio for each small area:

$$\frac{\text{EMSPE}_\text{s}}{\text{EMSPE}_\text{c}}.$$

Values larger than 1 indicate that the complete clustering reduces the true MSPE for small areas. We find the ratios of $\text{EMSPE}_\text{s}/\text{EMSPE}_\text{cb}$ and $\text{EMSPE}_\text{cb}/\text{EMSPE}_\text{c}$ to compare the EMSPE of the combined version of the clustering, EMSPE_cb, after implementing Tukey's method, with the simple method and complete clustering.

We observe that, when σ_l^2s are highly different, the improvement in either EMSPE_c or EMSPE_cb over the simple clustering becomes larger. Dealing with $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$, the improvement for the complete clustering reaches up to 67% with the minimum of -8% compared with the simple approach. We obtain larger improvements for areas belonging to clusters with variance components far from their overall average (25.1) whereas the negative improvement shows a scattered pattern. For combined clusters, we gain up to 49% improve-

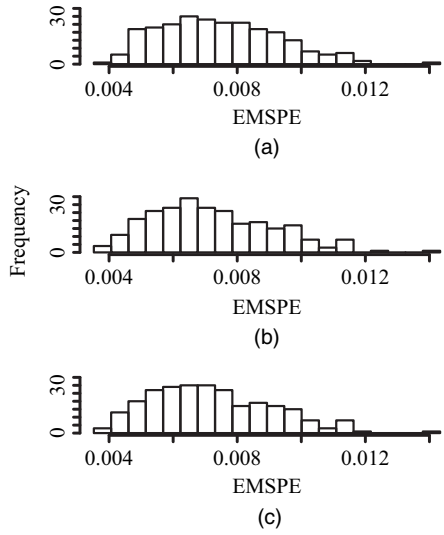

Fig. 8. Histogram of the EMSPE for $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$: (a) simple method; (b) combined method; (c) complete method

Table 7. Comparison of the EMSPE of predictors of small area means by using different approaches based on their deciles where $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$ and $D_{j_l} \sim U(0.25, 0.5)$

Statistic	$EMSPE_s/EMSPE_c$	$EMSPE_s/EMSPE_{cb}$	$EMSPE_{cb}/EMSPE_c$
Minimum	0.92	0.90	0.97
1-decile	0.99	0.98	0.99
2-decile	0.99	0.99	1.00
3-decile	1.00	0.99	1.00
4-decile	1.00	1.00	1.00
5-decile	1.01	1.01	1.00
6-decile	1.02	1.01	1.00
7-decile	1.03	1.03	1.01
8-decile	1.04	1.04	1.02
9-decile	1.19	1.14	1.04
Maximum	1.67	1.49	1.18
Mean	1.05	1.03	0.92

ment with a minimum of -10% compared with the simple approach. Also, complete clustering performs better than combined clustering for up to 18% and a minimum of -2%. Similarly to the complete-clustering scheme, in the case of combined clustering, the larger improvement happens for areas belonging to clusters with variance components that are far from the overall average. Table 7 summarizes the results for $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$.

To have a numerical evaluation of the performance of expression (4.3), we use the RB defined by

$$\mathrm{RB}_{j_l} = \frac{E(\mathrm{mspe}_{j_l})}{\mathrm{EMSPE}_{j_l}} - 1 \qquad \text{for } j_l = 1, \ldots, n_{c_l}, \quad l = 1, \ldots, k. \tag{6.1}$$

where $E(\mathrm{mspe}_{j_l})$, for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$, is the average of obtained values from expression (4.3) over $R = 5000$ iterations. Tables 8 and 9 give the summary statistics of the RB

Table 8. Summary statistics for the RB of the estimator of the MSPE of small area means by using different approaches for $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$

Statistic	*Complete-clustering approach*	*Combined clustering approach*	*Simple approach*
Minimum	−0.34	−0.34	−0.34
1-decile	−0.16	−0.17	−0.16
2-decile	−0.12	−0.12	−0.11
3-decile	−0.07	−0.08	−0.08
4-decile	−0.04	−0.05	−0.03
5-decile	0.00	−0.01	0.00
6-decile	0.02	0.02	0.04
7-decile	0.08	0.08	0.07
8-decile	0.12	0.12	0.12
9-decile	0.20	0.20	0.20
Maximum	0.55	0.56	0.57
Mean	0.01	0.01	0.01

Table 9. Comparison of mspes of predictors of small area means by using different approaches based on deciles of their variance where $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$ and $D_{jl} \sim U(0.25, 0.5)$

Statistic	$mspe_{\mathrm{c}}$	$mspe_{\mathrm{cb}}$	$mspe_{\mathrm{s}}$
Minimum	0.0008	0.0010	0.0012
1-decile	0.0013	0.0014	0.0014
2-decile	0.0016	0.0016	0.0016
3-decile	0.0018	0.0019	0.0019
4-decile	0.0021	0.0021	0.0023
5-decile	0.0023	0.0025	0.0026
6-decile	0.0027	0.0028	0.0030
7-decile	0.0031	0.0031	0.0035
8-decile	0.0037	0.0037	0.0039
9-decile	0.0040	0.0040	0.0043
Maximum	0.0048	0.0048	0.0047
Mean	0.0026	0.0026	0.0028

and variance of the EMSPE for various approaches when $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$ and $D_{jl} \sim U(0.25, 0.5)$. Our findings indicate that we achieve small RB for the three approaches. In particular, the value of |RB| is less than 0.55, 0.57 and 0.57 for the complete-, combined and simple clustering methods. All three methods have almost the same performance in terms of the variance of mspe. Our analysis shows that the simple, combined and complete-clustering methods have the same performance in terms of the coefficient of variation of mspe.

We now consider the simulation studies for $D_{jl} \sim U(0.25, 0.5)$ and $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$ to evaluate to what extent the difference between the variance components affects the estimator of the MSPE. We gain improvement in terms of the EMSPE by using clustering based on the covariate (Table 10). A maximum improvement of 66% for complete clustering and a minimum of −4% are obtained compared with the simple approach. For the combined version,

Table 10. Comparison of the EMSPE of predictors of small area means by using different approaches based on their deciles where $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$ and $D_{j_l} \sim U(0.25, 0.5)$

Statistic	$EMSPE_s/EMSPE_c$	$EMSPE_s/EMSPE_{cb}$	$EMSPE_{cb}/EMSPE_c$
Minimum	0.96	0.96	0.96
1-decile	0.99	0.99	0.99
2-decile	0.99	0.99	1.00
3-decile	1.00	1.00	1.00
4-decile	1.00	1.00	1.00
5-decile	1.00	1.00	1.00
6-decile	1.01	1.01	1.00
7-decile	1.01	1.01	1.00
8-decile	1.03	1.02	1.01
9-decile	1.20	1.17	1.01
Maximum	1.66	1.62	1.10
Mean	1.04	1.04	0.96

Table 11. Summary statistics for the RB of the estimator of the MSPE of small area means by using different approaches for $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$

Statistic	*Complete-clustering approach*	*Combined clustering approach*	*Simple approach*
Minimum	−0.34	−0.34	−0.34
1-decile	−0.16	−0.16	−0.16
2-decile	−0.12	−0.11	−0.12
3-decile	−0.07	−0.07	−0.08
4-decile	−0.04	−0.04	−0.04
5-decile	0.00	−0.01	0.01
6-decile	0.02	0.03	0.04
7-decile	0.08	0.08	0.07
8-decile	0.13	0.12	0.12
9-decile	0.20	0.20	0.20
Maximum	0.56	0.56	0.57
Mean	0.01	0.01	0.01

the maximum improvement of 62% and the minimum of −4% are obtained compared with the simple approach. Also, complete clustering performs better than combined clustering for up to 10% and a minimum of −4%. Similarly to Table 7, we cannot determine a specific trend for the negative improvement. The greatest amount of improvement happens for small areas with the variance component far from the overall average ($\sigma_{u_1}^2 = 1$). Tables 11 and 12 give the summary statistics for the RB and variance of the estimation of the MSPE by using different approaches when $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$. Complete and combined clustering have similar performance with $|\mathrm{RB}| \leqslant 0.56$ whereas the simple approach results in $|\mathrm{RB}| \leqslant 0.57$. All three methods have almost the same performance in terms of the variance of mspe. Our analysis shows that the simple, combined and complete-clustering methods have the same performance in terms of the coefficient of variation of mspe.

Table 12. Comparison of mspe of predictors of small area means by using different approaches based on deciles of their variance where $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$ and $D_{j_l} \sim U(0.25, 0.5)$

Statistic	$mspe_{\mathrm{c}}$	$mspe_{\mathrm{cb}}$	$mspe_{\mathrm{s}}$
Minimum	0.0009	0.0009	0.0012
1-decile	0.0013	0.0013	0.0014
2-decile	0.0016	0.0016	0.0017
3-decile	0.0019	0.0019	0.0020
4-decile	0.0021	0.0021	0.0023
5-decile	0.0024	0.0024	0.0026
6-decile	0.0027	0.0027	0.0030
7-decile	0.0032	0.0032	0.0035
8-decile	0.0037	0.0037	0.0039
9-decile	0.0042	0.0042	0.0043
Maximum	0.0048	0.0048	0.0047
Mean	0.0026	0.0026	0.0028

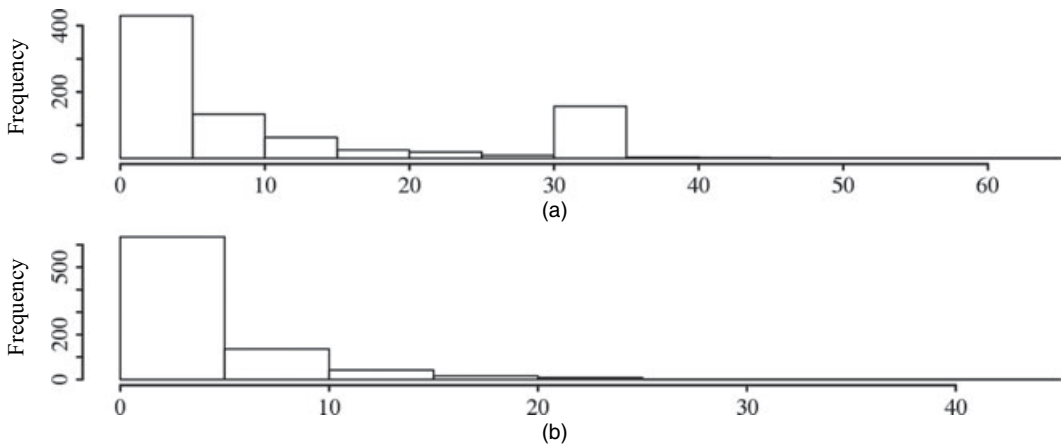

Fig. 9. (a) Histogram of the true D_is and (b) histogram of the estimates of D_is

6.2. Assessing the performance of the proposed estimator of D_i

In formula (5.2), we proposed an estimator of D_is for the unit level data. To evaluate the performance of this estimator, we carry out simulation studies. The estimated values of the parameters from Section 5 and also its covariate matrix are used to generate the response variable (waist circumference). We have 1362 small areas with 519 of them having one sample unit. We consider the MMM estimates of the variance components, $(0.36, 2.87, 4.05, 1.76, 11.40, 420.77, 36.17)$, as well as the estimated D_is ranging from 4.58×10^{-16} to 61.63 to generate the data. In each iteration, the sampling variance of the mean of the response variable in each small area is calculated. Fig. 9 shows the true and estimated values of D_i by using formula (5.2). Fig. 9 displays that formula (5.2) underestimates the sampling variance, but, in general, it gives reliable estimates. As we explained in Section 5, for small areas where the regression line cannot be defined, we use the overall regression line that is obtained from all sampled units. Mathematically, expression (5.1) is an unbiased estimator of σ_e^2. Expanding expression (5.1) results in

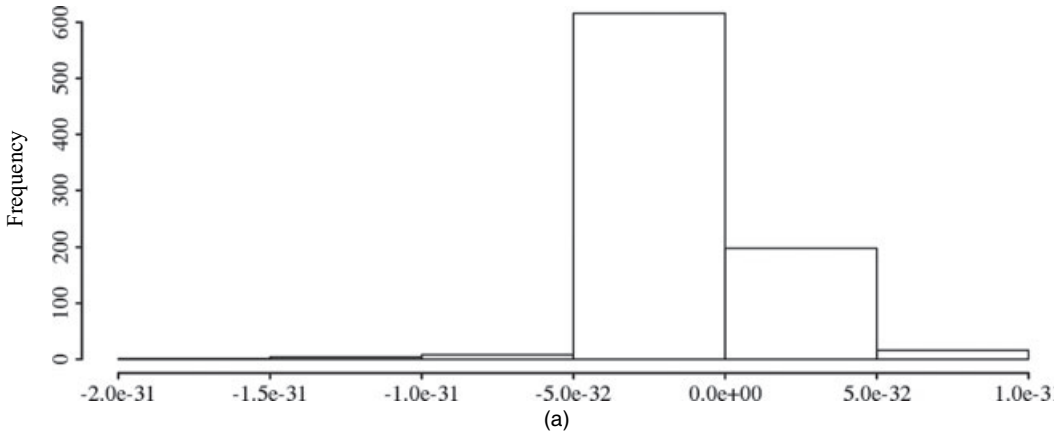

Fig. 10. Histogram of the neglected terms in estimation of $\sigma^2_{e_i}$s given in equation (6.2)

$$\frac{1}{n_i - 1}\left\{2\hat{\beta}_1 \sum_{j=1}^{n_i}(x_{ij} - x_i)(e_{ij} - e_i) + \sum_{j=1}^{n_i}(e_{ij} - e_i)^2\right\}. \tag{6.2}$$

We perform simulation studies to check the magnitude of the first term in expression (6.2),

$$\frac{1}{n_i - 1}\left\{2\beta_1 \sum_{j=1}^{n_i}(x_{ij} - x_i)(e_{ij} - e_i)\right\}.$$

Fig. 10 shows the histogram of this term for all small areas, which is almost zero. This indicates that the error of neglecting the first term in expression (6.2) is negligible.

7. Concluding remarks

In small area estimation, the ultimate goal is to find reliable estimates of parameters of small areas although only a few or no sampled units are available in some areas. Using a model-based approach, a link between different small areas is made to take into account the information from other small areas for prediction. In this paper, the main interest lies in predicting small area means whereas the precision of the predictor is quantified by using mspe.

Clustering small areas by using the Euclidean distance between covariates is proposed. The goal is to obtain more accurate predictions of small area means. For this, a hypothesis test is conducted after implementing hierarchical clustering of covariates to check the assumption of the equality of variance components in different clusters. Our results indicate that the test has a high power with inflated type I error. Following Tukey (1949), we combine some clusters with similar variance components. Small area means are predicted by either taking into account the difference between variance components in clusters, complete- or combined clustering schemes, or using the simple method (Fay–Herriot model) of equality of variance components in all clusters. To compare the performance of the new predictors with the simple predictor of small area means, the EMSPEs of three methods are calculated by using simulations. The results show improvement in terms of the EMSPE especially when the difference between variance components is significant. The simulation studies (which are not shown here) indicate the superiority of the complete- and combined clustering methods not only over the usual Fay–Herriot model, but also over the direct estimator of small area means.

A real data set is also analysed corresponding to the unit level model. To make it the area level model and to implement the methodologies that were developed here, we obtain the mean of the

response variable and the covariate. The mspe that is obtained by using the simple method is on average 19.40 times greater than the mspe obtained by using the complete-clustering approach. The estimated values of variance components for this data set are significantly different. Tukey's method is implemented and similar clusters in terms of the variance components are merged. We propose to consider combined clustering for this data set because of the reduction in the number of clusters after implementing Tukey's method.

This paper uses clustering in small area estimation based on similarity of covariates in small areas to account for the inherent differences between areas better and most likely to increase the precision of the small area mean prediction. We developed our methodology on the basis of a linear mixed model. Extending the results of this paper to generalized linear mixed models is of great importance and will be studied in our future work. In addition, as theorem 1 shows, it is possible to obtain a negative estimate for the variance component by using our proposed estimator. Addressing this limitation is of importance and will make the methodology proposed more general. Another future work of interest is extending the results of this paper to linear mixed models with discrete covariates whereas clustering is done by using other similarity measures such as the Gower distance.

Acknowledgements

We thank two referees and the Associate Editor for their constructive comments on an earlier version of the paper. Mohammad Jafari Jozani and Mahmoud Torabi gratefully acknowledge partial support from the Natural Sciences and Engineering Research Council of Canada.

Appendix A

A.1. Proof of lemma 1

Note that

$$\hat{\beta}_{\mathrm{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

where $\mathbf{y} = (y_1, \ldots, y_m)$. First, we rewrite $\hat{\beta}_{\mathrm{OLS}}$ as

$$\begin{aligned}\hat{\beta}_{\mathrm{OLS}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta,\end{aligned}$$

where $\delta = (u_1 + e_1, \ldots, u_m + e_m)'$. Now, $\hat{\beta}_{\mathrm{OLS}}$ is a consistent estimator of β, if $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta \to^{\mathrm{p}} 0$ as $m \to \infty$. To show this, note that

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta &= \left(\frac{1}{m}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{m}\mathbf{X}'\delta\right) \\ &= \left(\frac{1}{m}\mathbf{X}'\mathbf{X}\right)^{-1}\begin{pmatrix} \frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}}\delta_{j_l} \\ \frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}}X_{j_l1}\delta_{j_l} \\ \vdots \\ \frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}}X_{j_lp}\delta_{j_l}\end{pmatrix}.\end{aligned}$$

Let $S_{j_l i} = X_{j_l i}\delta_{j_l}$ and $i = 1, \ldots, p$, and assume that $X_{j_l i} \sim X_i$, for $j_l = 1, \ldots, n_{c_l}$, $l = 1, \ldots, k$, where $E(X_i) = \mu_i$ and $\text{var}(X_i) = \sigma_i^2$. Since $\mathbf{X}$ and δ are independent, we have

$$E(S_{j_l i}) = E(X_{j_l i})E(\delta_{j_l}) = 0,$$

and

$$\begin{aligned}\text{var}(S_{j_l i}) &= \text{var}\{E(S_{j_l i}|X_{j_l i})\} + E\{\text{var}(S_{j_l i}|X_{j_l i})\} \\ &= 0 + E\{X_{j_l i}^2(\sigma_{u_l}^2 + D_{j_l})\} \\ &= (\mu_i^2 + \sigma_i^2)(\sigma_{u_l}^2 + D_{j_l}).\end{aligned}$$

In contrast, $\exists M > 0$ such that $\sigma_{u_l}^2 + D_j \leqslant M$ for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. One can easily show that

$$\frac{1}{m^2}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}} \text{var}(S_{j_l i}) = \frac{1}{m^2}(\mu_i^2 + \sigma_i^2)\left(\sum_{l=1}^{k} n_{c_l}\sigma_{u_l}^2 + \sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}} D_{j_l}\right) < \infty,$$

Using Kolmogrov's strong law of large numbers, we have

$$\frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}} S_{j_l i} \to 0 \qquad \text{almost surely.}$$

This implies that $(1/m)\Sigma_{l=1}^{k}\Sigma_{j_l=1}^{n_{c_l}} X_{j_l i}\delta_{j_l} \to^{\text{p}} 0$ as $m \to \infty$. Using similar arguments, it is easy to show that $(\mathbf{X}'\mathbf{X})^{-1} \to^{\text{p}}$ constant as $m \to \infty$, which completes the proof.

A.2. Proof of theorem 1

Note that

$$y_{j_l} \sim N(\mathbf{X}_{j_l}\beta, \sigma_{u_l}^2 + D_{j_l}), \tag{A.1}$$

for $j_l = 1, \ldots, n_{c_l}$, $l = 1, \ldots, k$. Considering small areas that belong to the lth cluster, we have

$$E(y_{j_l} - \mathbf{X}_{j_l}\beta)^2 = \sigma_{u_l}^2 + D_{j_l},$$

for $j_l = 1, \ldots, n_{c_l}, l = 1, \ldots, k$, which leads to the following estimator of $\sigma_{u_l}^2$:

$$\hat{\sigma}_{u_l}^2 = \frac{1}{n_{c_l}}\sum_{j_l=1}^{n_{c_l}}\{(y_{j_l} - \mathbf{X}_{j_l}\beta)^2 - D_{j_l}\}.$$

Let $Z_{j_l} = (y_{j_l} - \mathbf{X}_{j_l}\beta)^2 - D_{j_l}$ for $j_l = 1, \ldots, n_{c_l}, l = 1, \ldots, k$. It is easy to show that

$$\begin{aligned}E(Z_{j_l}) &= \sigma_{u_l}^2, \\ \text{var}(Z_{j_l}) &= 2(\sigma_{u_l}^2 + D_{j_l})^2.\end{aligned}$$

As Z_{j_l}s are not identically distributed, to find the asymptotic distribution of $\hat{\sigma}_{u_l}^2$, we check Lindeberg's condition (Billingsley, 2008). Let $s_{n_{c_l}}^2 = \Sigma_{j_l=1}^{n_{c_l}}\text{var}(Z_{j_l})$. The interest is to show that

$$\lim_{n_{c_l}\to\infty}\frac{1}{s_{n_{c_l}}^2}\sum_{j_l=1}^{n_{c_l}} E\{(Z_{j_l} - \sigma_{u_l}^2)^2\mathbf{1}_{|Z_{j_l} - \sigma_{u_l}^2| > \epsilon s_{n_{c_l}}}\} = 0, \tag{A.2}$$

where $\epsilon > 0$ and $\mathbf{1}$ is the indicator function. For this, we first expand the expectation term. Let $t = (y_{j_l} - \mathbf{X}_{j_l}\beta)$. Based on distribution (A.1), we have

$$
\begin{aligned}
E\{(Z_{j_l}-\sigma_{u_l}^2)^2\mathbf{1}_{|Z_{j_l}-\sigma_{u_l}^2|>\epsilon s_{n_{c_l}}}\} &= 2\int\limits_{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}}^{\infty} \frac{(t^2-D_{j_l}-\sigma_{u_l}^2)^2}{\sqrt{\{2\pi(\sigma_{u_l}^2+D_{j_l})\}}}\exp\left\{-\frac{t^2}{2(\sigma_{u_l}^2+D_{j_l})}\right\}\mathrm{d}t \\
&= 2\int\limits_{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}}^{\infty} t^4\frac{1}{\sqrt{\{2\pi(\sigma_{u_l}^2+D_{j_l})\}}}\exp\left\{-\frac{t^2}{2(\sigma_{u_l}^2+D_{j_l})}\right\}\mathrm{d}t \\
&\quad -4(\sigma_{u_l}^2+D_{j_l})\int\limits_{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}}^{\infty} t^2\frac{1}{\sqrt{\{2\pi(\sigma_{u_l}^2+D_{j_l})\}}}\exp\left\{-\frac{t^2}{2(\sigma_{u_l}^2+D_{j_l})}\right\}\mathrm{d}t \\
&\quad +(\sigma_{u_l}^2+D_{j_l})^2\int\limits_{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}}^{\infty} \frac{1}{\sqrt{\{2\pi(\sigma_{u_l}^2+D_{j_l})\}}}\exp\left\{-\frac{t^2}{2(\sigma_{u_l}^2+D_{j_l})}\right\}\mathrm{d}t \\
&= 2\left[-t^3\sqrt{\left(\frac{\sigma_{u_l}^2+D_{j_l}}{2\pi}\right)}\exp\left\{-\frac{t^2}{2(\sigma_{u_l}^2+D_{j_l})}\right\}\right]_{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}}^{\infty} \\
&\quad +2(\sigma_{u_l}^2+D_{j_l})\left[-t\sqrt{\left(\frac{\sigma_{u_l}^2+D_{j_l}}{2\pi}\right)}\exp\left\{-\frac{t^2}{2(\sigma_{u_l}^2+D_{j_l})}\right\}\right]_{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}}^{\infty} \\
&\quad +3(\sigma_{u_l}^2+D_{j_l})^2[1-\Phi\{\sqrt{(\epsilon s_{n_{c_l}})+\sigma_{u_l}^2}-\mathbf{X}_{j_l}\boldsymbol{\beta}\}]. \qquad \text{(A.3)}
\end{aligned}
$$

As mentioned, D_{j_l}s and $\sigma_{u_l}^2$ are bounded in the small area estimation for $j_l=1,\ldots,n_{c_l}$ and $l=1,\ldots,k$. Let $M_0=\max\{D_{j_l}$ and $\sigma_{u_l}^2; j_l=1,\ldots,n_{c_l}, l=1,\ldots,k\}$. So, expression (A.2) is less than

$$
\begin{aligned}
&\lim_{n_{c_l}\longrightarrow\infty}\frac{1}{s_{n_{c_l}}^2}\left\{\exp\left(-\frac{1}{4M_0}\epsilon s_{n_{c_l}}\right)\left(\sum_{j_l=1}^{n_{c_l}}\exp\{(\sigma_{u_l}^2-\mathbf{X}_{j_l}\boldsymbol{\beta})^2+2\sqrt{(\epsilon s_{n_{c_l}})}(\sigma_{u_l}^2-\mathbf{X}_{j_l}\beta)\}\right.\right. \\
&\qquad\left.\left.\times\left[\sqrt{\left(\frac{M_0}{\pi}\right)}\{\sqrt{(\epsilon s_{n_{c_l}})}+\sigma_{u_l}^2-\mathbf{X}_{j_l}\beta\}^3+4\sqrt{\left(\frac{M_0^3}{2\pi}\right)}\{\sqrt{(\epsilon s_{n_{c_l}})}+\sigma_{u_l}^2-\mathbf{X}_{j_l}\beta\}\right]\right)\right\} \\
&\qquad+\lim_{n_{c_l}\longrightarrow\infty}12\frac{1}{s_{n_{c_l}}^2}M_0^2\sum_{j_l=1}^{n_{c_l}}[1-\Phi\{\sqrt{(\epsilon s_{n_{c_l}})}+\sigma_{u_l}^2-\mathbf{X}_{j_l}\beta\}]. \qquad \text{(A.4)}
\end{aligned}
$$

Noting that $s_{n_{c_l}}\to\infty$ as $n_{c_l}\to\infty$, expression (A.4) goes to 0. Thus, equation (A.2) holds. Since $\hat{\sigma}_{u_l}^2=(1/n_{c_l})\Sigma_{j_l=1}^{n_{c_l}}Z_{j_l}$, the asymptotic distribution of $\hat{\sigma}_{u_l}^2$ is easily obtained as follows:

$$
\hat{\sigma}_{u_l}^2\sim N\left\{\sigma_{u_l}^2,\frac{2}{n_{c_l}^2}\sum_{j_l=1}^{n_{c_l}}(\sigma_{u_l}^2+D_{j_l})^2\right\},
$$

as $n_{c_l}\to\infty$.

References

Billingsley, P. (2008) *Probability and Measure*. New York: Wiley.

Cressie, N. A. (1992) REML estimation in empirical bayes smoothing of census undercount. *Surv. Methodol.*, **18**, 75–94.

Cressie, N., Perrin, O. and Thomas-Agnan, C. (2005) Likelihood-based estimation for gaussian MRFs. *Statist. Methodol.*, **2**, 1–16.

Cressie, N. A., Perrin, O. and Thomas-Agnan, C. (2006) Doctors prescribing patterns in the Midi-Pyrénées region of France: point-process aggregation. In *Case Studies in Spatial Point Process Modeling* (eds A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan), pp. 183–195. New York: Springer.

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

Datta, G. S. and Mandal, A. (2015) Small area estimation with uncertain random effects. *J. Am. Statist. Ass.*, **110**, 1735–1744.
Fay, R. E. and Herriot, R. A. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. *J. Am. Statist. Ass.*, **74**, 269–277.
Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A *K*-means clustering algorithm. *Appl. Statist.*, **28**, 100–108.
Henderson, C. R. (1950) Estimation of genetic parameters. *Biometrics*, **6**, 186–187.
Jiang, J. and Nguyen, T. (2012) Small area estimation via heteroscedastic nested-error regression. *Can. J. Statist.*, **40**, 588–603.
Kang, E. L., Liu, D. and Cressie, N. (2009) Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computnl Statist. Data Anal.*, **53**, 3016–3032.
Maiti, T., Ren, H., Dass, S. C., Lim, C. and Maier, K. S. (2011) Clustering-based small area estimation: an application to MEAP data. *Calc. Statist. Ass. Bull.*, **66**, 73–93.
Pfeffermann, D. and Sverchkov, M. (2005) Small area estimation under informative sampling. *Statist. Transn*, **7**, 675–684.
Pfeffermann, D. and Sverchkov, M. (2007) Small-area estimation under informative probability sampling of areas and within the selected areas. *J. Am. Statist. Ass.*, **102**, 1427–1439.
Prasad, N. and Rao, J. (1990) The estimation of the mean squared error of small-area estimators. *J. Am. Statist. Ass.*, **85**, 163–171.
Rao, J. N. and Molina, I. (2015) *Small Area Estimation*, 2nd edn. Hoboken: Wiley.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Computnl Appl. Math.*, **20**, 53–65.
Speros, C. (2009) More than words: promoting health literacy in older adults. *Online J. Iss. Nursng*, **14**, no. 3, article 5.
Tukey, J. W. (1949) Comparing individual means in the analysis of variance. *Biometrics*, **5**, 99–114.
Vague, J. (1947) Sexual differentiation, a factor affecting the forms of obesity. *Presse Med.*, **30**, 339–340.
Wang, J. and Fuller, W. A. (2003) The mean squared error of small area predictors constructed with estimated area variances. *J. Am. Statist. Ass.*, **98**, 716–723.
Ward, Jr, J. H. (1963) Hierarchical grouping to optimize an objective function. *J. Am. Statist. Ass.*, **58**, 236–244.