Clustering in Small Area Estimation with Area-Level Linear Mixed Models

Elaheh Torkashvand[a], Mohammad Jafari Jozani[a,1], and Mahmoud Torabi[b]

[a] *Department of Statistics, University of Manitoba, Winnipeg, MB, CANADA, R3T 2N2*

[b] *Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, CANADA, R3E 0W3*

Summary

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

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 into linear models reduces the bias of predictors of small

[1] Corresponding author: M_Jafari_Jozani@umanitoba.ca

area means while it increases their variability. It is a common practice in mixed-effect models to assume homogenous random effects. However, there are many applications where such an assumption is not valid, as the lurking variables affecting the response variable in different areas could be different. In order to provide more realistic predictions of small area means, one might decide to work with non-homogenous random effects. Recently, there has been several research in this direction. For example, Maiti et al. (2011) considered the idea of clustering based on consequences of differences between the covariate where they studied the effect of the poverty on the educational performance of students in different school districts. In this application, small areas (school districts) were clustered based on their poverty status and different regression models were assumed for different clusters. Random effects of school districts belonging to the same cluster 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 works, Datta et al. (2011) and Datta and Mandal (2015) argued whether or not the presence of the random effects in the 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 including random effects might not always be useful. In order to test how influential the presence of the random effects in the area-level small area model is, they developed methodologies based on 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 the random effects in the model decreases the rate of convergence to the true value of the parameter in each area. The decrease is significant specially when the sample size in areas, n_i's, 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 while it might be necessary to keep it for some areas. To address this issue, Datta and Mandal (2015) had Bayesian view towards the presence of the random effects in the area-level small area estimation. They implemented the 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 while it is absent from the modelling with probability $1 - p$. Their method addresses the disadvantage of the method proposed in Datta et al. (2011) and gives more flexibility to the area-level linear models.

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

Rigby and Stasinopoulos (2005) introduced generalized additive model for location, scale, and shape (GAMLSS) in order to give more flexibility into modeling. GAMLSS defines different generalized linear mixed models for different model parameters and uses the back-fitting algorithm to solve for a proper model. Rigby and Stasinopoulos (2005) considered the same variance components for the response variable. However, for the simplest case of the mean and the variance, GAMLSS does 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 there is no access to unit level data in areas (e.g., due to the confidentiality, etc.). The sampling variance is assumed to be known while the variance of the random effects is unknown. As it is often the case, in observational studies, there are different lurking 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 most likely increase the precision of the small area mean prediction. Note that this inherent difference comes back to the features of small area and not 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 as a frequentist approach to give more flexibility to the Fay-Herriot model while we do not omit random effects from small area means. Clustering small areas 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/clusters are non-homogenous. The idea comes from the fact that in the regression analysis, when the covariates are close enough 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 coming from different error sources due to different lurking variables in clusters. The magnitude of these fluctuations might also differ. As an example, consider evaluating the effect of the body mass index (BMI) on the 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 show similar pattern in terms of the random effects. Obviously, the trend of accumulating body fat for underweight and obese people is different. So, one can easily consider different variance components for different clusters. We start with an 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 the 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 using the mean squared prediction error (MSPE) and study situations where the proposed methodology 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 in order 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 a second order of approximation. In Section 5, a real dataset is analyzed. Implementing the simulation studies, we evaluate the performance of our proposed test statistic under different 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 MSPE. Finally, we give some concluding remarks in Section 7.

2. Complete Clustering Approach in Area-Level Small Area Model

Consider the following area-level regression model

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

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_i's, and e_i's are the number of areas, the area-level random effects, and the random errors, respectively. Also, assume that u_i's are independent and identically distributed (i.i.d) from a $N(0, \sigma_u^2)$-distribution, and e_i's are independent with $e_i \sim N(0, D_i)$, where D_i's 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's 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 (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, one might expect random effects for areas with similar covariates in terms of the Euclidian distance, $\|\mathbf{X}_i - \mathbf{X}_j\|_2 = \sqrt{\sum_{t=1}^{p}(X_{it} - X_{jt})^2}$, to show similar behaviour compared with random effects associated with other areas. This motivates to use 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 Euclidian distance between the values of their corresponding covariates. In this paper, we also suggest to use a different variance component, $\sigma_{u_l}^2$, for $l = 1, \ldots, k$, in different clusters. In other words, we assume that $u_{j_l} \overset{i.i.d}{\sim} N(0, \sigma_{u_l}^2)$ for $j_l \in C_l$, the l'th cluster, and $j_l = 1, \ldots, n_{c_l}$, where n_{c_l} is the number of small areas in the l'th cluster. Under this setting, model (2.1) can be represented as follows

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

where y_{j_l} and $\mathbf{X}_{j_l}$ are the response variable and the covariate vector associated with the j'th element in the l'th 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 is no more significant changes in the distance between clusters for $k^* > k$. This makes the corresponding Fay-Herriot model more flexible in order to catch the true behaviour of the random effects.

Although complete clustering separates small areas into more homogeneous clusters, still variance components might be equal. If this happens, then (2.2) reduces to (2.1), and we refer to this as the 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 one 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)

$$H_0 : \sigma^2_{u_1} = \sigma^2_{u_2} = \cdots = \sigma^2_{u_k}, \quad vs. \quad H_a : \sim H_0.$$

To this end, because of the difference between variance components and sampling variances under the model (2.2), the Kolmogrov's strong law of large numbers is used in order to define the test statistic. Due to the complexity that the weighted least square estimate of β introduces into the method of moments estimates of variance components, we use the ordinary least square (OLS) to estimate β, where the consistency of $\hat{\beta}_{OLS}$ under the model (2.2) is shown in Lemma 1. Using $\hat{\beta}_{OLS}$, the modified method of moments estimators of variance components (MMM) 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, $\hat{\beta}_{OLS}$, is a consistent estimator of β.*

Proof. See the Appendix for the proof.

Theorem 1. *Let*

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

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

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

Proof. See the Appendix for the proof.

Remark 1. *It is worth noting that the estimator of σ^2_u in Theorem 1 is asymptotically unbiased. Due to the positive nature of the variance component, $\sigma^2_{u_l}$, for $l = 1, \dots, k$, the interest lies in getting positive estimates of the variance component. As the underlying distribution in (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 $m \longrightarrow \infty$) the probability of observing a negative estimate is negligible, i.e.*

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

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

Remark 2. *Estimators of $\sigma^2_{u_l}$, for $l = 1, \dots, k$, proposed in (3.2) are based on the method of moments. They are unbiased and consistent estimators. However, their variances highly depend on the number of small areas in each clusters. In order to guarantee a precise estimation of variance components, we suggest to implement this method when the number of small areas is large enough in different clusters.*

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

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

where $\sigma^2_0 = (\sigma^2_u, \dots, \sigma^2_u)$ and $\Sigma_0 = \text{diag}\left(\frac{2}{n^2_{c_1}} \sum_{j_1=1}^{n_{c_1}} (\sigma^2_u + D_{j_1})^2, \dots, \frac{2}{n^2_{c_k}} \sum_{j_k=1}^{n_{c_k}} (\sigma^2_u + D_{j_k})^2 \right)$. 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

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

where $\chi_0^2 = (\hat{\sigma}_u^2 - \hat{\sigma}_0^2)'\hat{\Sigma}_0^{-1}(\hat{\sigma}_u^2 - \hat{\sigma}_0^2)$, $\hat{\sigma}_0^2 = (\hat{\sigma}_{0u}^2, \dots, \hat{\sigma}_{0u}^2)$, $\hat{\sigma}_{0u}^2 = \frac{1}{k}\sum_{l=1}^{k} \hat{\sigma}_{u_l}^2$, and

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

In Section 6, we evaluate the performance of (3.5) under different scenarios. The results indicate that the test has a large power. However, the simulated values of the type I error are larger than the significance level. Note that the distribution of (3.2) highly depends on the number of small areas in each cluster. For larger clusters, it is expected to get more accurate estimates of variance components. However, due to Theorem 1, estimates obtained from larger clusters might be significantly different from estimates in smaller ones. In other words, the difference between the information 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 the equality of variance components is rejected, there might exist some clusters that have the same variance components. In order to estimate fewer number of parameters, one might consider combining such clusters. This is similar to what happens after rejecting the null hypothesis in the ANOVA context. Tukey (1949) proposed a solution to this problem by using combination of T- and F-tests. In this paper, we modify his approach to combine the clusters with the same variance components. To this end, we take the following steps:

- We first sort the MMM estimates of the variance components.
- Considering (3.2) and conducting the T-test under the unequal variance set-up and using the same significance level as in (3.6), we make groups of clusters. To this end, starting from the cluster with the smallest variance estimate, we compare it with the one with the second smallest variance estimates. If the null hypothesis of the equality of the related variance components, using (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 one. Similarly, if the null hypothesis of the equality is not rejected, we add

the corresponding cluster to the group that the second cluster belongs to. Similar to Tukey (1949), this process stops if all groups are constructed by one or two clusters. If not, we go to the next step.

- For a group with the number of clusters larger than two, we find the maximum and the average values of the MMM estimates of the variance components, $\hat{\sigma}^2_{max}$ and $\hat{\sigma}^2_{mean}$, respectively. Following Tukey (1949) and depending on the number of clusters in each group, say k', we construct a test statistic W as follows
$$W = \frac{1}{3}\left(0.25 + \frac{1}{n_{c_{max}}}\right)^{-1}\left(\frac{\hat{\sigma}^2_{max} - \hat{\sigma}^2_{mean}}{var(\hat{\sigma}^2_{max} - \hat{\sigma}^2_{mean})} - 1.2\log_{10} k'\right), \quad k' > 3,$$
$$W = \frac{1}{3}\left(0.25 + \frac{1}{n_{c_{max}}}\right)^{-1}\left(\frac{\hat{\sigma}^2_{max} - \hat{\sigma}^2_{mean}}{var(\hat{\sigma}^2_{max} - \hat{\sigma}^2_{mean})} - 0.5\right), \quad k' = 3,$$
where $n_{c_{max}}$ is the number of small areas inside the cluster corresponding to $\hat{\sigma}^2_{max}$. Here
$$var(\hat{\sigma}^2_{max} - \hat{\sigma}^2_{mean}) = (1 - \frac{1}{k'})^2 var(\hat{\sigma}^2_{max}) + (\frac{1}{k'})^2 \sum_{j \neq j_{max}} var(\hat{\sigma}^2_j)$$
where j_{max} is the index related to the $\hat{\sigma}^2_{max}$ and $var(\hat{\sigma}^2_j)$'s are obtained using the estimated values of the variance in (3.2).

  The aim is to see whether or not we can split a group of clusters into smaller ones. If W is larger than the critical value of the standard Normal distribution for the two-sided test with the level of significance in (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 one.

- 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 the equality of the variance components for the group with the size larger than two. If the size of the group is larger than $p+1$, we implement the test statistic proposed in this paper. Otherwise, simultaneous T-tests with the same significance level as in (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 forward to the largest one. For the odd number size, we let the last subgroup have three clusters and then test the assumption of the 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 the above algorithm, some of the clusters will be combined. Hence, in order to estimate the MSPE, denoted by mspe, one needs to deal with fewer number of parameters. This results in less variability due to the estimation of the model parameters into 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 clustering has the same performance as the complete one in estimating small area means, but, significantly better than the simple clustering. However, the results of Section 6 indicate that if the difference between variance components is huge, the 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 EBLUP 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 MSPE of the EBLUP which is a second order unbiased approximation.

4. EBLUP and its MSPE 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 $\tilde{u}_{j_l} = \mathbf{G}\mathbf{V}^{-1}(y_{j_l} - \mathbf{X}_{j_l}\tilde{\beta})$, $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$, $\mathbf{y} = (y_1, \ldots, y_m)$, $\mathbf{V} = \text{diag}(\mathbf{V}_1, \ldots, \mathbf{V}_k)$, $\mathbf{V}_l = \text{diag}(\sigma^2_{u_l} + D_{1_l}, \ldots, \sigma^2_{u_l} + D_{n_{c_l}})$, $\mathbf{G} = \text{diag}(\mathbf{G}_1, \ldots, \mathbf{G}_k)$, $\mathbf{G}_l = \sigma^2_{u_l}\mathbf{I}_{n_{c_l}\times n_{c_l}}$, 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 follows

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

where $\gamma = (\sigma^2_{u_1}, \ldots, \sigma^2_{u_k})$, $g_{1j_l}(\gamma)$ is the j'th element of the l'th 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 j'th row of the l'th cluster of $\mathbf{G}\mathbf{V}^{-1}$.

Due to unknown γ, $\hat{\theta}_{j_l}$'s, the EBLUP of θ_{j_l}'s, are obtained for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. To this end, the MMM estimators of variance components are used in formula (4.1). The MSPE of the

EBLUP can be decomposed as

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

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

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

where

$$g_{3j_l}(\gamma) = \text{trace}\left((\frac{\partial \mathbf{b}'_{\mathbf{j_l}}}{\partial \gamma}) \mathbf{V} (\frac{\partial \mathbf{b}'_{\mathbf{j_l}}}{\partial \gamma})' \mathbf{var}(\gamma) \right), \tag{4.2}$$

$$\frac{\partial \mathbf{b}'_{j_l}}{\partial \gamma} = \begin{bmatrix} \frac{\partial \mathbf{b}'_{j_l 1}}{\partial \sigma^2_{u_1}} & \cdots & \frac{\partial \mathbf{b}'_{j_l m}}{\partial \sigma^2_{u_1}} \\ \frac{\partial \mathbf{b}'_{j_l 1}}{\partial \sigma^2_{u_2}} & \cdots & \frac{\partial \mathbf{b}'_{j_l m}}{\partial \sigma^2_{u_2}} \\ & \vdots & \\ \frac{\partial \mathbf{b}'_{j_l 1}}{\partial \sigma^2_{u_k}} & \cdots & \frac{\partial \mathbf{b}'_{j_l m}}{\partial \sigma^2_{u_k}} \end{bmatrix}_{k \times m},$$

and $var(\gamma) = \Sigma_0$.

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

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

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

It is worth mentioning that the magnitude of g_{3j_l} depends on the number of variance components in the model. So, the more parameters we have in the model, the larger g_{3j_l} will be. However, the magnitude of g_{1j_l} and g_{2j_l} 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 relative bias (RB) as the measure of the precision.

5. Real Data Analysis

In this section, we use the National Health and Nutrition Examination Survey (NHANES) for 2011-2012 as a unit-level dataset to predict the waist circumference based on the Body Mass Index (BMI). Vague (1947) showed that there is a high association between the 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 while the "gynoid obesity", often found in women, has a lower risk.

The data is categorized to small domains using the following variables:

- Age groups as 20.00 – 33.00, 33.00 – 48.00, 48.00 – 48.56, 48.56 – 63.00, 63.00 – 80.00.
- Gender as male and female.
- Education as the highest grade or level of education completed by adults 20 years and older grouped as less than 9th grade education, 9 – 11th grade education (includes 12th grade and no diploma), High school graduate/GED, some college or associates (AA) degree, and college graduate or higher.
- Ethnicity as Mexican, other hispanic, non-hispanic white, non-hispanic black, non-hispanic Asian, and Other race -including multi racial.
- Poverty groups as 0.00 – 0.97, 0.97 – 1.88, 1.88 – 2.41, 2.41 – 4.03, 4.03 – 5.00.

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

Suppose 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 $j = 1, \dots, n_i$. Let

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

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

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

Note that β_1 is unknown. In order 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 obtained from the complete data to estimate $\hat{\sigma}^2_e$. Consequently, the sampling variance for the mean of the response variable in each area is given by

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

D_i's ranges 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 $D_i = 61.63$.

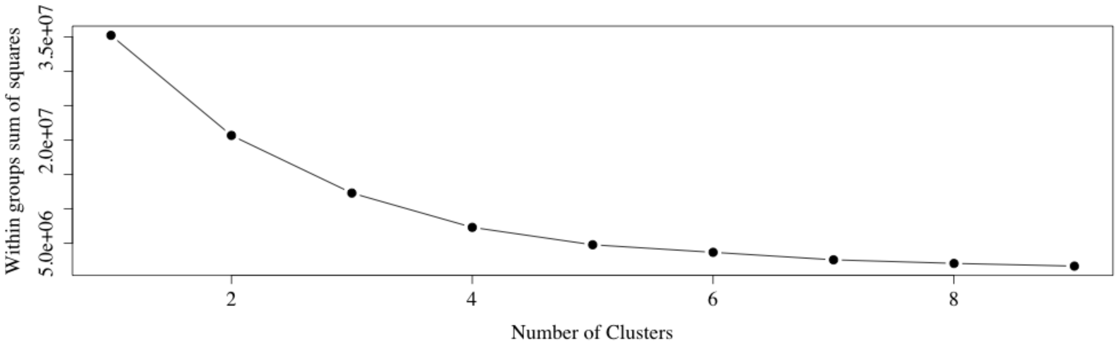


Figure 1: The effect of number of clusters on within groups sums of squares in the k-means clustering

There are a few approaches to cluster small areas. For example, the k-means clustering is an approach that minimizes the Euclidian distance of the covariate from the mean of covariates belonging to a cluster (Hartigan and Wong, 1979). Another approach is to use the hierarchical clustering (Ward Jr, 1963) based on the Lance-Williams algorithm through the squared Euclidian distance between covariates. The Silhouette method (Rousseeuw, 1987) also gives the 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 are far from the optimum value, using Tukey's method, we can always combine clusters which are not significantly different. So, as Figure 1 shows, we cluster small areas in 7 groups using the hierarchical clustering and model the 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 $n_c = (165, 521, 333, 64, 252, 9, 18)$. Figure 2 shows the 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 $(11.40, 0.36, 2.87, 4.05, 1.76, 420.77, 36.17)$. The assumption of the equality of the variance components is rejected with $\chi^2_4 = 90.80$ with the corresponding p-value of zero. The residual is defined as

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

where $\hat{\theta}_{j_l}$ is obtained using (4.1) and substituting the unknown parameters from either simple or complete clustering methods for $j_l = 1, \ldots, n_{c_l}$, $l = 1, \ldots, 7$. Figures 3 and 4 show boxplots of the residual variable belonging to different clusters using the simple and complete methods, respectively. After implementing the complete clustering, the behaviour of residuals in different clusters is more homogenous in Figure 4. By using Tukey's method, we have 5 combined clusters where the modified number of small areas in clusters are $n_c = (165, 521, 649, 9, 18)$ with the variance components $(11.40, 0.36, 2.89, 420.77, 36.17)$.

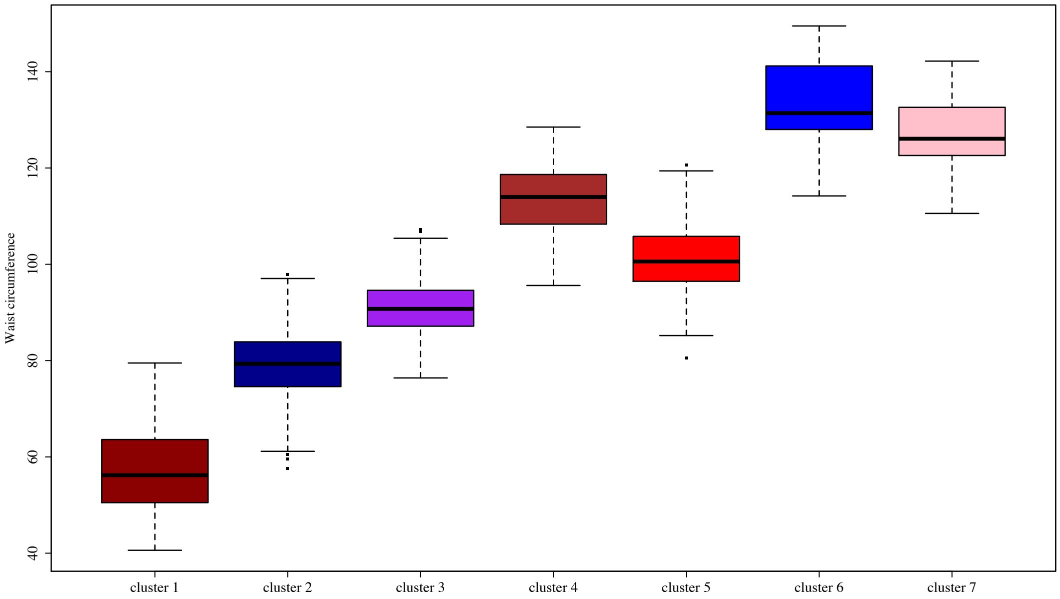


Figure 2: The boxplot of the response variable (waist circumference) in different clusters after implementing the hierarchical method.

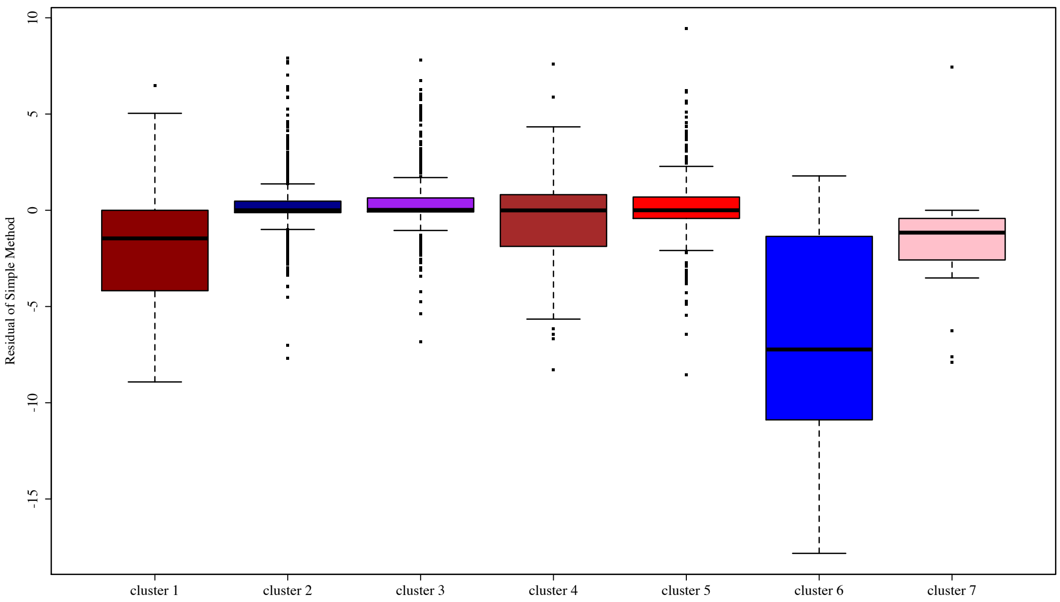


Figure 3: The boxplot of the residual variable belonging to different clusters after implementing the simple method (Fay-Herriot model).

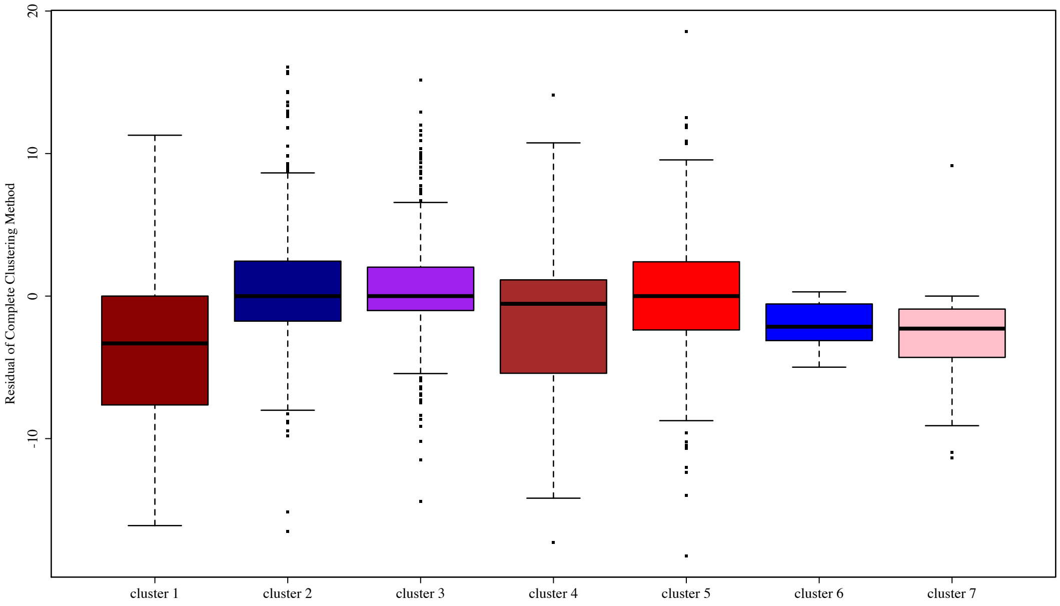


Figure 4: The boxplot of the residual variable belonging to different clusters after implementing the complete method.

Using the complete clustering approach, explained in (2.2) where we derived estimates of variance components using (3.1), we observe more than 49% improvements over the simple approach (Fay-Herriot Model) in the mspe in 70% of small areas. In 10.00% of small areas, the mspe of the simple clustering approach is up to 62.24 times more than the mspe of the complete clustering approach. On average we get the 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 combined clustering approaches have the same performance in over 80% of small areas in terms of the mspe. In order to compare the performance of the proposed methods with the simple clustering approach, we first calculate the 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, the usage of either the complete or combined clustering schemes are justifiable. However, the MMM estimates of the variance components for three clusters, 2.87, 4.05, 1.76, are very close. Because of this similarity, the combined clustering scheme merges the corresponding clusters.

Table 1: The comparison between the complete, combined, and simple clustering in terms of the quantiles of their corresponding mspe

	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

Figures 5 and 6 present the mspe and predicted values of small area means, respectively.

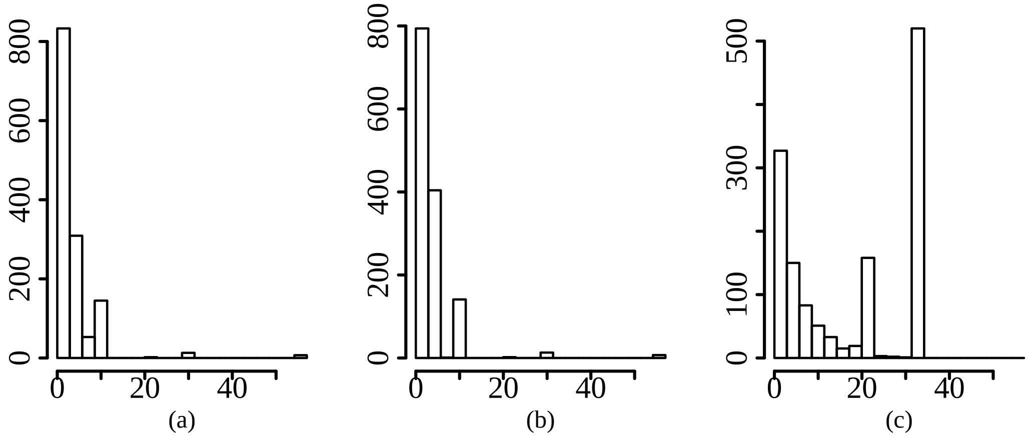


Figure 5: Histogram of the mspe's: (a) the complete clustering method, (b) the combined clustering method, (c) the simple method

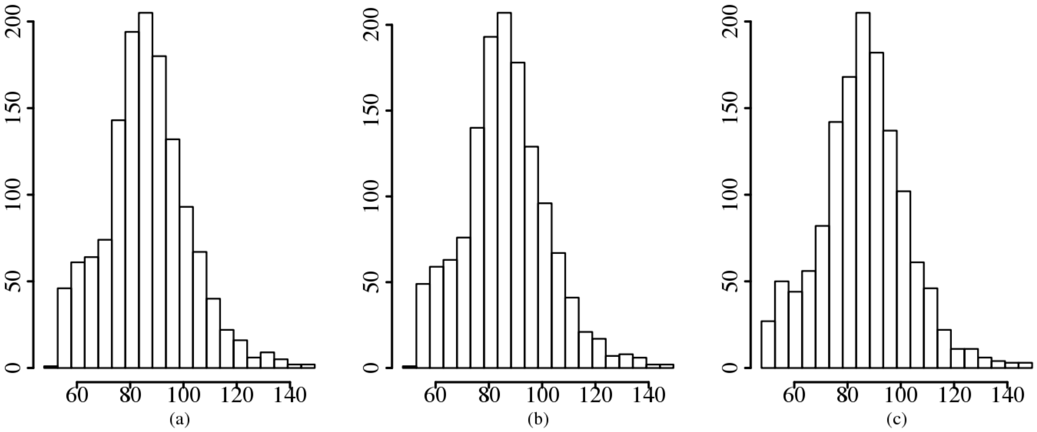


Figure 6: Histogram of predicted small area means: (a) the complete clustering method, (b) the combined clustering method, (c) the simple method

6. Simulation Studies

In this section, we consider different scenarios of the sampling variances, D_{j_l}'s, and the variance components, $\sigma^2_{u_l}$'s, for $j_l = 1, \ldots, n_{c_l}$ 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. To this end, the empirical MSPE (EMSPE) of small area means using different clustering schemes is calculated. Further, we evaluate the performance of the complete and combined clustering approaches on the estimation of MSPE, mspe, using RB. In Section 5, an unbiased estimator of the sampling variance inside the small area was proposed. We assess the performance of this estimator using simulation studies.

6.1. Evaluating the test statistic and the effect of the complete and combined clustering on EMSPE and mspe

For the purpose of simulation studies, we use an area-level dataset regarding the prescription costs from Union R$\acute{e}$gionale des Caisses d'Assurance Maladie (URCAM) of the Midi-Pyr$\acute{e}$n$\acute{e}$es Region in the south-west of France, during the period January-December, 1999. The dataset 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 to 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 dataset 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 dataset in a spatial analysis to predict the average prescription amount in each canton.

In this work, we consider small area estimation to conduct the simulation studies using this dataset. The area-level small area model provides an appropriate link between different small areas in order 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 while 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 based on the percentage of patients over 70 years. We expect patients in similar aging groups to have reasonably similar prescription costs as they need similar number of follow-up visits, etc. Another important factor is the type of medication that is taken by the patient. Apparently, people in similar aging groups need similar supplementary treatments (Speros, 2009).

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

Consider $n_c = (33, 37, 32, 14, 34, 65, 4, 13, 18, 18)$. Let x_{j_l} denote the j'th covariate in the l'th 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. 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 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 (REML) estimates to predict small area means and also estimate the MSPE. We performed simulation studies in order to compare our proposed estimates of the variance components with the REML estimates of the variance components (Cressie, 1992; Rao and Molina, 2015) in terms of mean squared error (MSE). The aim is to show although MMM estimator of the variance components are based on the method of moments, their performance is comparable to REML estimates of the variance components. Tables 2 and 3 show that the two methods have almost the same performance using the MSE criterion. As Table 3 shows the MSE's 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 get a smaller MSE of variance components for larger cluster sizes.

Table 2: MSE of the REML and MMM estimates for different variance components

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

Table 3: MSE of the REML and MMM estimates for similar variance components in some clusters

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

In order to calculate the power of the proposed test, 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 the equality of variance components if 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 different set-ups.

As it is shown in Table 4, when the difference between D_{jl}'s and $\sigma^2_{u_l}$'s gets larger, the test becomes more powerful. Moreover, when the similarities between the $\sigma^2_{u_l}$'s, for $l = 1, \dots, k$, increase, the power of the test decreases. Generally, removing clusters from the analysis may 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 one, we expect to have a higher power because of 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 significance level $\alpha = 0.05$ (Table 6). As it is explained in Section 2, this is due to the fact that the precision of the estimate of the $\sigma^2_{u_l}$'s, for $l = 1, \dots, k$ depends highly on the number of small areas in each cluster. In our set-up, the 7'th cluster contains four small areas while the 6'th cluster contains 65 small areas. So, even though the variance components are the same in both, the null hypothesis of the equality might be rejected because of the difference between n_{c_6} and n_{c_7}.

Table 4: The power of the test statistic for different values of $\sigma^2_{u_l}$'s and D_j's

	D_{jl}			
σ^2_u	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

We also perform simulation studies in order to evaluate the performance of our proposed method

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

	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

Table 6: The percentage of times the null hypothesis is rejected by mistake

	D_{j_l}			
σ_u^2	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

in terms of the EMSPE. To this end, we calculate the EBLUP of small area means, $\hat{\theta}_{j_l}^{(r)}$'s, by finding the MMM estimates of the variance components and substituting them in (4.1). The EMSPE is given as follows

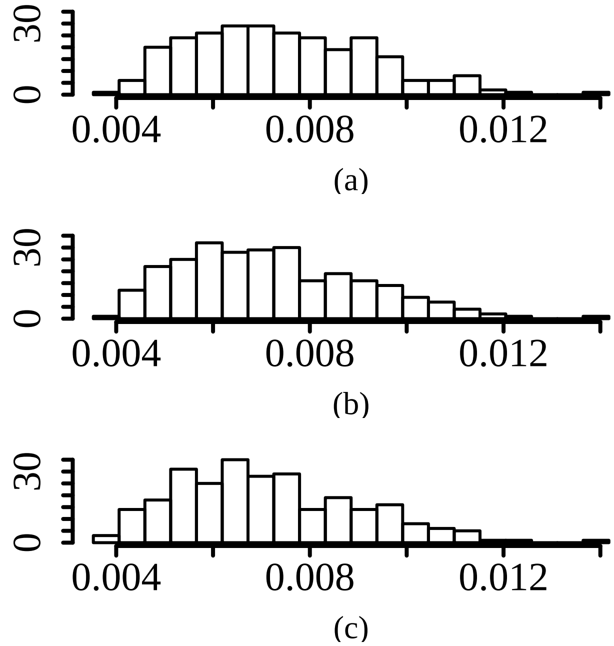
$$\text{EMSPE}(\theta_{j_l}) = \tfrac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{j_l}^{(r)} - \theta_{j_l}^{(r)})^2, \text{ for } j_l = 1, \ldots, n_{c_l} \text{ and } l = 1, \ldots, k.$$

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

We compare the EMSPE of the and complete an combined clustering, EMSPE_c and EMSPE_{cb}, with the simple method, EMSPE_s, by finding for e.g. the following ratio for each small area

$$\frac{\text{EMSPE}_s}{\text{EMSPE}_c}.$$

Values larger than one indicate that the complete clustering reduces the true MSPE for small

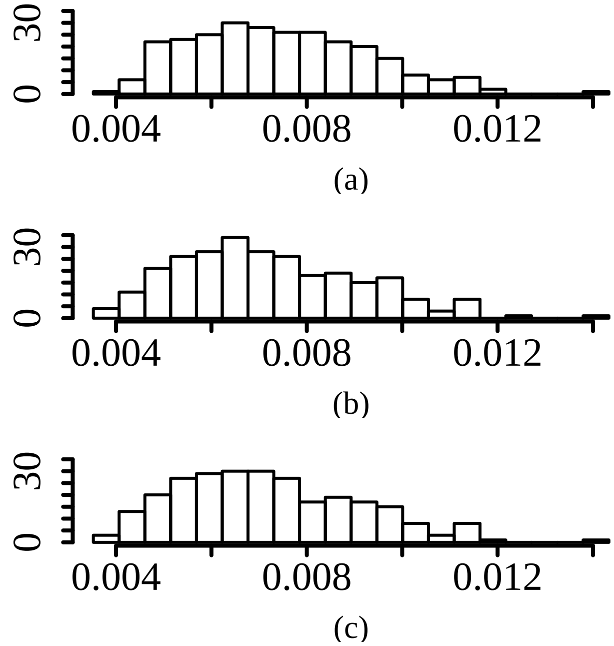


(a)

(b)

(c)

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



(a)

(b)

(c)

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

areas. We find the ratios of $\frac{\text{EMSPE}_s}{\text{EMSPE}_{cb}}$ and $\frac{\text{EMSPE}_{cb}}{\text{EMSPE}_c}$ to compare EMSPE of the combined version of the clustering, EMSPE_{cb}, after implementing Tukey's method, with the simple method and the complete clustering.

We observe that when σ_l^2's 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 gets up to 67% with the minimum of −8% compared to the simple approach. We get larger improvements for areas belonging to clusters with variance components far from their overall average (25.1) while the negative improvement shows a scattered pattern. For combined clusters, we gain up to 49% improvement with the minimum of −10% compared to the simple approach. Also, the complete clustering performs better than the combined one for up to 18% and the minimum of −2%. Similar to the complete clustering scheme, in the case of combined clustering, the larger improvement happens for areas belonging to clusters with variance components far from the overall average. Table 7 summarizes the results for $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$.

Table 7: Comparison of EMSPE of predictors of small area means using different approaches based on their deciles: (a) $\frac{\text{EMSPE}_s}{\text{EMSPE}_c}$, (b) $\frac{\text{EMSPE}_s}{\text{EMSPE}_{cb}}$, (c) $\frac{\text{EMSPE}_{cb}}{\text{EMSPE}_c}$ when $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$ and $D_{j_l} \sim \text{Uniform}(0.25, 0.5)$.

	(a)	(b)	(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

In order to have a numerical evaluation of the performance of (4.3), we use the RB defined by

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

where $\text{E}(\text{mspe}_{j_l})$, for $j_l = 1, \dots, n_{c_l}$ and $l = 1, \dots, k$, is the average of obtained values from (4.3) over

$R = 5000$ iterations. Tables 8 and 9 give the summary statistics of the RB and variance of the EMSPE for different approaches when $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$ and $D_{jl} \sim \text{Uniform}(0.25, 0.5)$. Our findings indicate that we get small RB for three approaches. In particular, the value of $|RB|$ is less than 0.55, 0.57, 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 the simple, combined, and complete clustering methods have the same performance in terms of the coefficient of variation of mspe.

Table 8: The summary statistics for the RB of the estimator of MSPE of small area means using different approaches for $\sigma_u^2 = (1, 8, 25, 5, 64,\ 10, 49, 36, 40, 13)$: (a) the complete clustering approach, (b) the combined clustering approach, and (c) the simple approach

	(a)	(b)	(c)
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

We now consider the simulation studies for $D_{jl} \sim \text{Uniform}(0.25, 0.5)$ and σ_u^2 =(1, 8, 25, 25, 64, 25, 49, 40, 40, 13) to evaluate to what extend 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). The maximum improvement of 66% for the complete clustering and the minimum of -4% are obtained compared to the simple approach. For the combined version, the maximum improvement of 62% and the minimum of -4% are obtained compared to the simple approach. Also, the complete clustering performs better than the combined one for up to 10% and the minimum of -4%. Similar to Table 7, we are not able to determine a specific trend for

Table 9: Comparison of mspe of predictors of small area means using different approaches based on deciles of their variance: (a) mspe$_c$, (b) mspe$_{cb}$, (c) mspe$_s$ when $\sigma_u^2 = (1, 8, 25, 5, 64, 10, 49, 36, 40, 13)$ and $D_{j_l} \sim$ Uniform$(0.25, 0.5)$.

	(a)	(b)	(c)
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

the negative improvement. The highest 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 using different approaches when $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$. The complete and combined clustering have similar performance with the $|RB| \leq 0.56$ while the simple approach results in $|RB| \leq 0.57$. All three methods have almost the same performance in terms of the variance of mspe. Our analysis shows the simple, combined, and complete clustering methods have the same performance in terms of the coefficient of variation of mspe.

6.2. Assessing the performance of the proposed estimator of D_i

In Formula (5.2), we proposed an estimator of D_i's for the unit level data. In order to evaluate the performance of this estimator, we implement 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

Table 10: Comparison of the EMSPE of predictors of small area means using different approaches based on their deciles: (a) $\frac{\text{EMSPE}_s}{\text{EMSPE}_c}$, (b) $\frac{\text{EMSPE}_s}{\text{EMSPE}_{cb}}$, (c) $\frac{\text{EMSPE}_{cb}}{\text{EMSPE}_c}$ when $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$ and $D_{jl} \sim \text{Uniform}(0.25, 0.5)$.

	(a)	(b)	(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

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_i's ranging from 4.58×10^{-16} to 61.63 in order to generate the data. In each iteration, the sampling variance of the mean of the response variable in each small area is calculated. Figure 9 shows the true and estimated values of D_i's using Formula (5.2). Figure 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 obtained from all sampled units. Mathematically, (5.1) is an unbiased estimator of σ_e^2. Expanding (5.1) results in

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

We perform simulation studies to check the magnitude of the first term in (6.2), $\frac{1}{n_i-1}[2\beta_1 \sum_{j=1}^{n_i}(x_{ij} - x_i)(e_{ij} - e_i)]$. Figure 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 (6.2) is negligible.

Table 11: The summary statistics for the RB of the estimator of MSPE of small area means using different approaches for $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$: (a) the complete clustering approach, (b) the combined clustering approach, and (c) the simple approach

	(a)	(b)	(c)
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

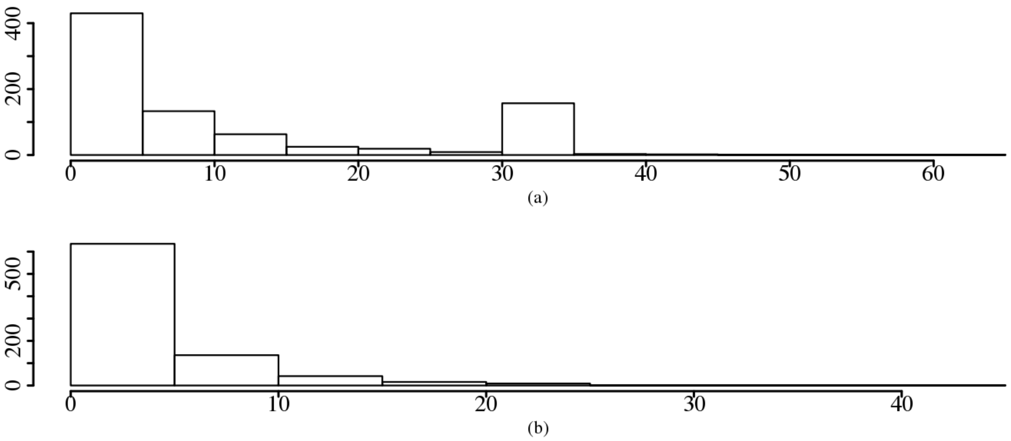


Figure 9: (a) histogram of the true D_i's vs. (b) histogram of the estimates of D_i's

7. Concluding Remarks

In small area estimation, the ultimate goal is to find reliable estimates of parameters of small areas while 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

Table 12: Comparison of mspe of predictors of small area means using different approaches based on deciles of their variance: (a) mspe$_c$, (b) mspe$_{cb}$, (c) mspe$_s$ when $\sigma_u^2 = (1, 8, 25, 25, 64, 25, 49, 40, 40, 13)$ and $D_{j_l} \sim$ Uniform$(0.25, 0.5)$.

	(a)	(b)	(c)
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

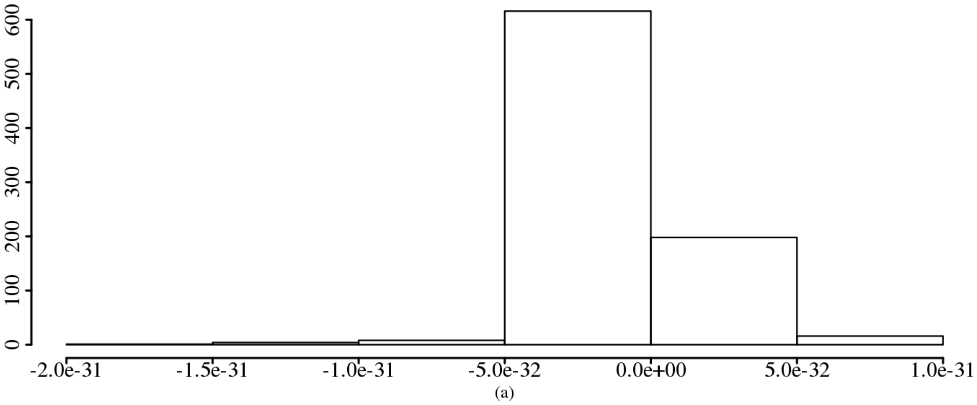


Figure 10: Histogram of the neglected terms in estimation of $\sigma_{e_i}^2$'s given in (6.2).

areas for the purpose of prediction. In this paper, the main interest lies in predicting small area means while the precision of the predictor is quantified using the mspe.

Clustering small areas using the Euclidean distance between covariates is proposed. The goal is to get more accurate predictions of small area means. To this end, 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 the equality of variance components in all clusters. In order to compare the performance of the new predictors with the simple predictor of small area means, the EMSPE of three methods are calculated using simulations. The results show improvement in terms of the EMSPE specially when the difference between variance components is significant. The simulation studies (not shown here) indicate of 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 analyzed corresponding to the unit level model. In order to make it the area level model and implement the methodologies developed here, we obtain the mean of the response variable and the covariate. Using the complete clustering approach, mspe, the estimated MSPE, is on average 19.40 times smaller than mspe obtained using the simple method. The estimated values of variance components for this dataset are significantly different. Tukey's method is implemented and similar clusters in terms of the variance components are merged. We propose to consider the combined clustering for this data set due to 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 in order to better account the inherent differences between areas and most likely increase the precision of the small area mean prediction. We developed our methodology based on 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 using our proposed estimator. Addressing this limitation is of importance and will make the proposed methodology more general. Another future work of interest is extending the results of this paper to linear mixed models with discrete covariates while clustering is done by using other similarity measures such as the Gower distance.

Acknowledgment:

We would like to thank two anonymous referees and an associated editor for their constructive comments on an earlier version of the paper. Mohammad Jafari Jozani and Mahmoud Torabi gratefully acknowledge partial support from NSERC.

A. Appendix

Proof of Lemma 1. Note that

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

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

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

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

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\delta &= (\tfrac{1}{m}\mathbf{X}'\mathbf{X})^{-1}(\tfrac{1}{m}\mathbf{X}'\delta)\\ &= (\tfrac{1}{m}\mathbf{X}'\mathbf{X})^{-1}\begin{bmatrix}\frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}}\delta_{j_l}\\ \frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}}X_{j_l 1}\delta_{j_l}\\ \vdots\\ \frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}}X_{j_l p}\delta_{j_l}\end{bmatrix}.\end{aligned}$$

Let $S_{j_l i} = X_{j_l}\delta_{j_l}$ 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 $\mathrm{E}(X_i) = \mu_i$ and $var(X_i) = \sigma_i^2$. Since $\mathbf{X}$ and δ are independent, we have

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

and,

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

On the other hand, $\exists M > 0$ such that $\sigma^2_{u_l} + D_j \leq M$ for $j_l = 1, \ldots, n_{c_l}$ and $l = 1, \ldots, k$. One can easily show that

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

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

$$\frac{1}{m}\sum_{l=1}^{k}\sum_{j_l=1}^{n_{c_l}} S_{j_l i} \xrightarrow{a.s.} 0.$$

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

Proof of Theorem 1. Note that

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

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

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

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

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

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

$$\mathrm{E}(Z_{j_l}) = \sigma^2_{u_l} \quad \text{and} \quad var(Z_{j_l}) = 2(\sigma^2_{u_l} + D_{j_l})^2.$$

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

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

where $\epsilon > 0$ and $\mathbf{1}$ is the indicator function. To this end, we first expand the expectation term. Let

$t = (y_{j_l} - \mathbf{X}_{j_l}\beta)$. Based on (A.1), we have

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

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

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

Noting that $s_{n_{c_l}} \longrightarrow \infty$ as $n_{c_l} \longrightarrow \infty$, (A.4) goes to zero. Thus, (A.2) holds. Since $\hat{\sigma}^2_{u_l} = \frac{1}{n_{c_l}} \sum_{j_l=1}^{n_{c_l}} Z_{j_l}$, the asymptotic distribution of $\hat{\sigma}^2_{u_l}$ easily obtained as follows

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

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

References

Billingsley, P. 2008. Probability and measure. John Wiley & Sons.

Cressie, N., Perrin, O., and Thomas-Agnan, C. 2005. Likelihood-based estimation for gaussian MRFs. *Statistical Methodology* 2:1–16.

Cressie, N. A. 1992. REML estimation in empirical bayes smoothing of census undercount. *Survey Methodology* 18:75–94.

Cressie, N. A., Perrin, O., and Thomas-Agnan, C. 2006. Doctors prescribing patterns in the midi-pyrénées rregion of france: Point-process aggregation, pp. 183–195. *In* Case Studies in Spatial Point Process Modeling. 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. *Journal of the American Statistical Association* 106:362–374.

Datta, G. S. and Mandal, A. 2015. Small area estimation with uncertain random effects. *Journal of the American Statistical Association* 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. *Journal of the American Statistical Association* 74:269–277.

Hartigan, J. A. and Wong, M. A. 1979. Algorithm as 136: A k-means clustering algorithm. *Applied statistics* pp. 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. *Canadian Journal of Statistics* 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. *Computational Statistics & Data Analysis* 53:3016–3032.

Macieira-Coelho, A. 1986. Cancer and aging. *Experimental gerontology* 21:483–495.

Maiti, T., Ren, H., Dass, S. C., Lim, C., and Maier, K. S. 2011. Clustering-based small area estimation: An application to meap data. *Calcutta Statistical Association Bulletin* 66:73–93.

Pfeffermann, D. 2002. Small area estimation-new developments and directions. *International Statistical Review* 70:125–143.

Pfeffermann, D. and Sverchkov, M. 2005. Small area estimation under informative sampling. *Statistics in Transition* 7:675–684.

Pfeffermann, D. and Sverchkov, M. 2007. Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association* 102:1427–1439.

Prasad, N. and Rao, J. 1990. The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association* 85:163–171.

Rao, J. N. and Molina, I. (2015). Small area estimation, Second Edition. Wiley.

Rigby, R. A. and Stasinopoulos, D. M. 2005. Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54:507–554.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* 20:53–65.

Speros, C. 2009. More than words: Promoting health literacy in older adults. *OJIN: The Online Journal of Issues in Nursing* Vol. 14, No. 3, Manuscript 5.

Torkashvand, E., Jafari Jozani, M., and Torabi, M. 2015. Pseudo-empirical bayes estimation of small area means based on the james-stein estimation in linear regression models with functional measurement errors. *Canadian Journal of Statistics* 43:265–287.

Tukey, J. W. 1949. Comparing individual means in the analysis of variance. *Biometrics* pp. 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. *Journal of the American Statistical Association* 98:716–723.

Ward Jr, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association* 58:236–244.