Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Small area estimation using multilevel models

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Canadä

by Fernando A.S. Moura and David Holt

June 1999

Statistics Canada Statistique Canada

Canadä

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 73-80
Statistics Canada, Catalogue No. 12-001

Small area estimation using multilevel models

Fernando A.S. Moura and David Holt [1]

Abstract

In this paper a general multilevel model framework is used to provide estimates for small areas using survey data. This class of models allows for variation between areas because of: (i) differences in the distributions of unit level variables between areas, (ii) differences in the distribution of area level variables between areas and (iii) area specific components of variance which make provision for additional local variation which cannot be explained by unit-level or area-level covariates. Small area estimators are derived for this multilevel model formulation and an approximation to the mean square error (MSE) of each small area estimate for this general class of mixed models is provided together with an estimator of this MSE. Both the approximations to the MSE and the estimator of MSE take into account three sources of variation: (i) the prediction MSE assuming that both the fixed and components of variance terms in the multilevel model are known, (ii) the additional component due to the fact that the fixed coefficients must be estimated, and (iii) the further component due to the fact that the components of variance in the model must be estimated. The proposed methods are estimated using a large data set as a basis for numerical investigation. The results confirm that the extra components of variance contained in multilevel models as well as small area covariates can improve small area estimates and that the MSE approximation and estimator are satisfactory.

Key Words: Small area estimation; Mixed models; Multilevel Models; EBLUE.

1. Introduction

The need for small area (and small domain) estimates from survey data has long been recognized. The difficulty with the production of such estimates is that for most, if not all, small areas, the sample size achieved by a survey designed for national purposes is too small for direct estimates to be made with acceptable precision. Early attempts to tackle this problem using methods such as synthetic estimation (Gonzalez 1973) involved the use of auxiliary information and the pooling of information across small areas. An excellent review and bibliography are given by Ghosh and Rao (1994).

Empirical studies show that such methods made too little provision for local variation and consequently the resulting small area estimates were shrunk too far towards a predicted mean. More recent approaches (*e.g.*, Battese and Fuller 1981 and Battese, Harter and Fuller 1988) use some components of variance model, or equivalent, to provide for local variation. Empirical studies show the superiority of this approach (*e.g.*, Prasad and Rao 1990).

This paper proposes a general multilevel model framework for small area estimation. This involves the potential to use auxiliary information at both the unit and small area level. In addition any of the regression parameters, rather than just the intercept as proposed by Battese and Fuller (1981), may be treated as varying randomly between small areas. The local variation is provided for by using differences between the means of unit level auxiliary variables, the small area level variables, and the various components of variance which allow variation between areas.

For this general model, the small area predictor is obtained. In addition, an approximation to the mean square error (MSE) of each separate small area prediction and an estimator of this MSE are developed.

The numerical study, based on a large data set from Brazil shows that such models may be useful for predicting small area estimates. The robustness of the approach to misspecification of the variance-covariance matrix of the small area random effects and misspecification of small area covariates are also investigated. Further numerical results demonstrate the success of the MSE approximation and its estimator.

2. The multilevel model framework

2.1 Introduction

We consider the following multilevel model for predicting the small area means:

$$Y_i = X_i \beta_i + \varepsilon_i$$

$$\beta_i = Z_i \gamma + \nu_i \quad i = 1, \ldots, m \tag{2.1}$$

where Y_i is the vector of length n_i for the characteristic of interest for the sample units in the i^{th} small area, $i = 1, \ldots, m$; X_i is the matrix of explanatory variables at sample unit level; Z_i is the design matrix of small area variables; γ is the vector of length q of fixed coefficients and $v_i = (v_{i0}, \ldots, v_{ip})^T$ is the vector of length $(p+1)$ of random effects for the i^{th} small area. We assume the

[1] Fernando A.S. Moura, Instituto de Matemática, UFRJ, Rio de Janeiro, Brazil, CP: 68530, CEP: 21941-590. E-mail: fmoura@dme.ufrj.br; David Holt, Office for National Statistics, 1 Drummond Gate, London, SW1P 2QQ. E-mail: tholt@ons.go.uk.

following about the distribution of the random vectors: (a) the v_i are independent between small areas and have a joint distribution within each small area with $E(v_i) = 0$ and (b) $V(v_i) = \Omega$ (b) The ε_i's and v_i's are independent and $V(\varepsilon_i) = \sigma^2 I.$

For the whole population (2.1) applies with n_i replaced by $N_i,$ the small area population sizes.

The set of *m* equations in (2.1) can be concisely written by stacking them as

$$Y = XZ\gamma + X\nu + \varepsilon. \tag{2.2}$$

It is worth noting that the random intercept model (see section 2.3) can be regarded as a special case of the model (2.1) where Z_i is equal to the identity matrix for each small area and Ω has all terms constrained to be zero except the one corresponding to the variance of the intercept term. Other intermediate models exist, for instance, when Ω is diagonal so that the small area regression coefficients are random but uncorrelated between covariates.

Holt (in Ghosh 1994, page 82) observes that the advantage of the model (2.1) over other competitors is that it effectively integrates the use of unit level and area level covariates into a single model. Besides the use of extra random effects for the regression coefficients gives greater flexibility in situations where it is not appropriate to assume the same slope coefficients apply for all small areas.

2.2 Fixed and component of variance parameter estimates

The fixed and components of variance parameters in the model (2.1) are γ and $\theta = ([\text{Vech}(\Omega)]^T, \sigma^2)^T$ respectively. Various methods for estimating these model parameters in the case of a general mixed linear model are available. Most of them, based on iterative algorithms, lead to the maximum likelihood estimator (MLE) or the restricted maximum likelihood estimator (RMLE) under certain regularity conditions.

Goldstein (1986) shows how consistent estimators can be obtained by applying iterative generalised least squares procedures (IGLS). He also proved its equivalence to the maximum likelihood estimator under normality. Later Goldstein (1989) proposed a slight modification of his algorithm (namely, restricted iterative generalised least squares (RIGLS)) which is equivalent to RMLE under normality. Unlike the IGLS estimates, the RIGLS estimation procedures provide unbiased estimates of the component of variance parameters by taking into account the loss in degrees of freedom resulting from estimating the fixed parameters.

This work is confined to the RIGLS approach as in Goldstein (1989). The RIGLS procedure is described in details in Appendix A.

2.3 The estimator of the small area mean

Assuming the model (2.1) and considering that the population size N_i in the i^{th} small area is large, we can write the mean for the i^{th} small area as

$$\mu_i = \bar{X}_i^T Z_i \gamma + \bar{X}_i^T \nu_i \tag{2.3}$$

where $\bar{X}_i$ is the $(p+1)$ population mean vector for the i^{th} small area.

An estimator of μ_i may be obtained by plugging the RIGLS estimators of γ and θ in the respective terms of equation (2.3), where the predictor of the i^{th} small area random effect v_i is given by $\hat{v}_i = \hat{\Omega} X_i^T \hat{V}_i^{-1}(Y_i - X_i Z_i \hat{\gamma})$ where $\hat{V}_i^{-1} = \hat{\sigma}^{-2} I - \hat{\sigma}^{-4} X_i \hat{\Omega} \hat{G}_i^{-1} X_i^T$ and $\hat{G}_i^{-1} = (I + \hat{\sigma}^{-2} X_i^T X_i \hat{\Omega})^{-1}.$

This estimator of μ_i is known as Empirical Best Linear Unbiased Estimator (EBLUE)

$$\hat{\mu}_i = \bar{X}_i^T Z_i \hat{\gamma} + \bar{X}_i^T \hat{v}_i. \tag{2.4}$$

Battese *et al.*, (1981, 1988) propose and apply a random intercept model to provide small area estimates. In this case, the Empirical Best Linear Unbiased Estimator is

$$\hat{\mu}_{i(\text{RI})} = \bar{X}_i^T \hat{\beta} + \hat{v}_{i0}.$$

We use the label (RI) to imply a random intercept model since only the intercept of each small area is random while the other components of β remain fixed.

2.4 Approximation to the Mean Square Error (MSE)

Kackar and Harville (1984) show that, if $\hat{\theta}$ is a translation invariant estimator of θ and the random terms are normally distributed, the mean square error of a predictor of a linear combination of a fixed and random effect can be decomposed into two terms. The first one is due to the variability in estimating the fixed parameters when the components of variance are known, the second term comes from estimating the components of variance.

Since under normality the RIGLS estimator is equivalent to the RMLE estimator and the RMLE is translation-invariant, Kackar and Harville's (1984) results can be applied to the small area means estimators $\hat{\mu}_i,\ i = 1, \ldots, m$:

$$\text{MSE}(\hat{\mu}_i) = E[\hat{\mu}_i - \mu_i]^2 = E[\tilde{\mu}_i - \mu_i]^2 + E[\hat{\mu}_i - \tilde{\mu}_i]^2 \tag{2.5}$$

where $\tilde{\mu}_i$ is the BLUE of μ_i.

The first term of (2.5), that is $\text{MSE}[\tilde{\mu}_i],$ can be obtained by direct calculation as

$$\begin{aligned}\text{MSE}(\tilde{\mu}_i) &= \bar{X}_i^T (G_i^{-1})^T \Omega \bar{X}_i \\ &\quad + \sigma^2 \bar{X}_i (G_i^{-1})^T Z_i \left(\sum_{i=1}^{m} Z_i^T G_i^{-1} X_i^T X_i Z_i \right)^{-1} Z_i^T G_i^{-1} \bar{X}_i \end{aligned} \tag{2.6}$$

where $G_i = I + \sigma^2 X_i^T X_i \Omega.$ Kackar and Harville (1984) point out that the second term of (2.6) is not tractable, except for special cases, and propose an approximation to it. Prasad and Rao (1990) propose an approximation to this second term and work out the details of their approximation for three particular cases: the random intercept model, random regression coefficient model and the Fay-Herriot model. They also give some regularity conditions for their approximation to be of the second order, and prove that their MSE approximation for the Fay-Herriot model is of the second order. Nevertheless, it seems to be more difficult to give general conditions for more complex models such as model (2.1).

Applying Prasad and Rao's approach, an approximation to the second term of (2.5) is developed in Appendix B.

It is worth noting that the MSE approximation of $(\hat{\mu}_i)$ can be decomposed into three terms:

$$\text{MSE}(\hat{\mu}_i) \approx T_1 + T_2 + T_3 \qquad (2.7)$$

where T_1 and T_2 are respectively the first and the second term of equation (2.6) and T_3 is described in Appendix B.

The term T_1 is the variability of $\hat{\mu}_i$ when all parameters are known, the second term T_2 is due to estimating the fixed effects and the third term T_3 comes from estimating the components of variance.

When sampling fractions are not negligible, estimators of the small area means can be built in the spirit of the finite population approach by predicting specifically for the non-sampled units:

$$\hat{\mu}_i^F = f_i\,\bar{y}_i + (\bar{X}_i - f_i\,\bar{x}_i)^T (Z_i\hat{\gamma} + \hat{v}_i) \qquad (2.8)$$

where the superscript F indicates that a correction for the finite population sampling fraction f_i was used; $\bar{x}_i$ is the $(p+1)$ vector of sample means.

The $\text{MSE}(\hat{\mu}_i^F)$ can be obtained by noting that

$$\hat{\mu}_i^F - \bar{Y}_i = (1-f_i)[(\bar{X}_i^C)^T (Z_i(\hat{\gamma}-\gamma) + \hat{v}_i - v_i - \bar{\varepsilon}_i^C)]$$

where $\bar{X}_i^C = (1-f_i)^{-1}(\bar{X}_i - f_i\bar{x}_i)$ and $\bar{\varepsilon}_i^C$ is the mean of ε_{ij} for the non-sampled units in the i^{th} small area. Therefore

$$\text{MSE}(\hat{\mu}_i^F) = (1-f_i)^2[\text{MSE}^*(\hat{\mu}_i) + N_i^{-1}(1-f_i)^{-1}\sigma^2] \quad (2.9)$$

where $\text{MSE}^*(\hat{\mu}_i)$ is the equation (2.7) with $\bar{X}_i$ replaced by $\bar{X}_i^C$.

2.5 Estimation of mean square error

It is common practice to estimate the MSE of a linear combination of the fixed and random effects in a mixed model as in (2.1) by replacing estimates of the components of variance respectively in the expression of MSE. This estimator ignores the contribution to MSE due to estimating the components of variance parameters. Several studies (see for example Singh, Stukel and Pfeffermann 1998 or Harville and Jeske 1992) argue that this procedure tends to underestimate the MSE. Prasad and Rao (1990) reported a simulation study which showed that the use of this "naive" estimator leads to severe downwards bias. They also showed for the Fay-Herriot model (a special case of the model (2.1)), using "truncated Henderson" estimates for the variance components, that

$$E(\hat{T}_1) = T_1 - T_3 + o(m^{-1});$$

$$E(\hat{T}_2) = T_2 + o(m^{-1});$$

$$E(\hat{T}_3) = T_3 + o(m^{-1}).$$

Harville and Jeske (1992) establish some conditions for the unbiasedness of Prasad and Rao's mean square error estimator. However, considering the more general model (2.1), again it seems more difficult to give general conditions for which the order of bias of Prasad and Rao's estimator is $o(m^{-1})$, especially if iterative procedures as RIGLS are used to obtain the parameter estimates.

Nevertheless, motivated by the simulation study summarised in Section 3.4 and an extensive simulation study described in Moura (1994), we propose to use an estimator similar to Prasad and Rao's for $\text{MSE}(\hat{\mu}_i)$:

$$\text{M}\hat{\text{S}}\text{E} = \hat{T}_1 + \hat{T}_2 + 2\hat{T}_3 \qquad (2.10)$$

where $\hat{T}_i$ are obtained from (2.5) by replacing σ^2 and Ω by their respective RIGLS estimators.

From equation (2.9) we can also obtain an estimator for $\text{MSE}(\hat{\mu}_i^F)$ as follows:

$$\text{M}\hat{\text{S}}\text{E}(\hat{\mu}_i^F) = (1-f_i)^2[\text{M}\hat{\text{S}}\text{E}^*(\hat{\mu}_i) + N_i^{-1}(1-f_i)^{-1}\hat{\sigma}^2] \quad (2.11)$$

where $\text{M}\hat{\text{S}}\text{E}^*(\hat{\mu}_i)$ is the equation (2.10) with $\bar{X}_i$ replaced by $\bar{X}_i^C$.

3. A model-based numerical investigation

3.1 Comparison of the estimators

In order to investigate the properties of alternative estimators, data was used from 38,740 households in the enumeration districts in one county in Brazil. The Head of Household's income was treated as the dependent variable. Two unit level independent variables were identified as the educational attainment of the Head of Household (ordinal scale of $0-5$) and the number of rooms in the household $(1-11+)$.

The assumed model is

$$Y_{ij} = \beta_{i0} + \beta_{i1}x_{1ij} + \beta_{i2}x_{2ij} + \varepsilon_{ij} \;\; i=1, ..., m; \; j=1, ..., N_i$$

$$\beta_{i0} = \gamma_{00} + v_{i0}; \;\; \beta_{i1} = \gamma_{10} + v_{i1}; \;\; \beta_{i2} = \gamma_{20} + v_{i2} \qquad (3.1)$$

where x_1 and x_2 respectively represent the number of rooms and the educational attainment of the head of the household (centred about their respective population means).

The parameter values for the fit model and their respective standard errors are

$\gamma_{00} = 8.456(0.108)$ $\gamma_{10} = 1.223(0.046)$ $\gamma_{20} = 2.596(0.086)$

$\sigma_{00} = 1.385(0.194)$ $\sigma_{01} = 0.354(0.66)$ $\sigma_{02} = 0.492(0.117)$

$\sigma_{12} = 0.333(0.054)$ $\sigma_{11} = 0.234(0.35)$ $\sigma_{22} = 0.926(0.124)$

$\sigma^2 = 47.74(0.345)$.

To carry out numerical investigations within the model-based framework a simulation was carried out keeping the enumeration district identifiers and the values of the two explanatory variables (X) fixed. Initially the area population means $\bar{X}_{1i}$ and $\bar{X}_{2i}$ were calculated for the whole data set and a randomly selected subsample of 10% of records from each small area was identified. This same subset was retained throughout the simulations (the Simulation subset).

The data generation for the simulations was carried out in two stages using a data generation model which was the General Model (G), the Diagonal Model (D), the Random Intercept Model (RI) as appropriate. In the first case the parameter values were taken from the estimates mentioned earlier. In the second case the off-diagonal terms were set to zero, in the third case only $\sigma_{00} = 1.385$ was non-zero.

The first stage of the data simulation process was to generate the level 2 random terms (that is, the non-zero elements of ν_{i0} and ν_{i1} and ν_{i2}) depending on the choice of the data generation model. These random terms were Normally distributed (jointly Normal in the case of the General Data Generation Model and the Diagonal Data Generation Model). At this stage the expected value of the mean for the i^{th} area conditional on the area level random effects generated by the model $m_1 =$ G, D, RI in the r^{th} simulation could be obtained:

$$\mu_{im_1}^{(r)} = \beta_{0i}^{(r)} + \beta_{1i}^{(r)} \bar{X}_{1i} + \beta_{2i}^{(r)} \bar{X}_{2i}.$$

At the second stage of the data simulation process, unit values (Y_{ij}) were created for each of the data generation models. Having generated the data for the simulation subset under one of the data generation models, all three of the estimation models (G, D and RI) could be fitted to the simulated data to obtain parameter estimates and predictors for the small area means.

For each data generation model $m_1 =$ G, D, RI the whole simulation process was repeated $R = 5{,}000$ times to yield a set of small area means $\mu_{im_1}^{(r)}$ and predicted means $\hat{\mu}_{im_1, m_2}^{(r)}$, $r = 1, \ldots, R$ for each small area, i, $i = 1, \ldots, m$ and for the three estimation models: $m_2 =$ G, D, RI. For each small area and for data generated under model $m_1 =$ G, D, RI, the Mean Square Error (MSE) of the prediction process for each estimation model m_2 may be defined as

$$\text{MSE}[\hat{\mu}_{im_1, m_2}] = R^{-1} \sum_{r=1}^{R} (\hat{\mu}_{im_1, m_2}^{(r)} - \mu_{im_1}^{(r)})^2$$

and the absolute relative error (ARE) by

$$\text{ARE}[\hat{\mu}_{im_1, m_2}] = R^{-1} \sum_{r=1}^{R} |\, \mu_{im_1, m_2}^{(r)} - \mu_{im_1}^{(r)}| \,/ \mu_{im_1}^{(r)}.$$

For comparative purposes we contrast the properties of each estimator with those of the estimator which is the same as the data generation model. Hence we define the Ratio of Mean Square Error (RMSE):

$$\text{RMSE}_{m_2, m_1} = \left\{ \sum_{i=1}^{m} \text{MSE}[\hat{\mu}_{im_1, m_2}] \right\} \Big/ \left\{ \sum_{i=1}^{m} \text{MSE}[\hat{\mu}_{im_1, m_1}] \right\} \times 100$$

and the Ratio of Absolute Relative Error (RARE):

$$\text{RARE}_{m_2, m_1} = \left\{ \sum_{i=1}^{m} \text{ARE}[\hat{\mu}_{im_1, m_2}] \right\} \Big/ \left\{ \sum_{i=1}^{m} \text{ARE}[\hat{\mu}_{im_1, m_1}] \right\} \times 100.$$

It will be seen that when the data are generated from a simpler model (*e.g.*, RI) the more complex estimation procedures do not suffer any appreciable worsening of efficiency or bias. On the other hand when the data are generated from a more complex model the simpler estimators have inferior properties. However the difference between the Diagonal and General estimators is much less than between these and the Random Intercept Estimator. From Table 1 one would conclude that it is worth introducing additional random coefficients of some kind, beyond the simple Random Intercept model assumptions, but not necessarily the full General Model.

Table 1
Ratios of mean square errors and ratios of absolute relative errors (in parentheses) for the three estimators and three data generation models

Estimator	Data Generation Model		
	G	D	RI
General (G)	100.0 (100.0)	101.8 (100.9)	101.2 (100.6)
Diagonal (D)	108.8 (82.6)	100.0 (100.0)	100.2 (100.1)
R. Intercept (RI)	131.9 (176.9)	109.1 (105.6)	100.0 (100.0)

The summary measures in Table 1 are average properties over all small areas. A careful analysis of the MSE performance of the estimators for each small area shows that there is a modest increase in the MSE for the Diagonal Estimator compared to the General Estimator for all areas, whereas for the Random Intercept estimator a relatively small number of areas exhibit a substantial increase in MSE. A similar pattern occurs between the Diagonal and Random Intercept estimator when the Diagonal Data Generation Model is used.

3.2 Introducing a small-area level covariate

In this section an attempt is made to investigate the impact on small area estimates of introducing an area covariate *Z*. Unfortunately for the data set used, it was not possible to identify a single contextual area level covariate which had a substantial effect on the multilevel models. Nevertheless, the number of cars per household in each small area was a useful covariate for the random coefficients for the individual level random slopes coefficients for “Room” and “Edu”, but not for the random intercept term. This was observed after some preliminary model fit analysis on the real data. Although the “numbers of cars” was the best small area level covariate found to explain between area variation, it was not as powerful at the individual level as “Room” and “Edu”, the individual level covariates chosen.

The model above with the small area covariate *Z* can be written as

$$Y_{ij} = \beta_{i0} + \beta_{i1} x_{1ij} + \beta_{i2} x_{2ij} + \varepsilon_{ij} \quad i = 1, \ldots, m; \; j = 1, \ldots, N_i$$

$$\beta_{i0} = \gamma_{00} + \nu_{i0}; \; \beta_{i1} = \gamma_{10} + \gamma_{11} z_i + \nu_{i1}; \; \beta_{i2} = \gamma_{20} + \gamma_{21} z_i + \nu_{i2} \quad (3.2)$$

The small area random effects were assumed uncorrelated in order to avoid convergence failure in the simulation study.

Table 2 reports the parameter estimates and their respective standard errors obtained by fitting the Diagonal Model with the *Z* covariate (3.2) and without the *Z* covariate (2.1). It is worth noting the significant reduction of all the components of variance estimates, except $\hat{\sigma}_{00}$ and $\hat{\sigma}^2$, after introducing the explanatory area covariate *Z*.

In order to investigate the effect of misspecification of the *Z* variable, the model based simulation procedure described in section 3.1 was applied to the two models above, where the data generation was done according to the parameters presented in Table 2. Table 3 summarises the simulation results.

It is worth noting that in both cases there is a significant loss of efficiency by using an unsuitable estimator. It can also be seen from an individual analysis of MSE for each small area that a considerable gain in efficiency is achieved with the introduction of a small area covariate *Z* over the diagonal model. For many small areas the MSE of the Diagonal with *Z* is significantly less than the MSE of the corresponding estimator without *Z*. Even for those few areas in which the MSE of the Diagonal with *Z* is unchanged or even slightly increased by the introduction of *Z*, the difference is not appreciable.

Table 2
Parameter estimates and standard errors for general model with area level covariate: Demographic data

Parameter	Diagonal Model with *Z*	Diagonal Model
γ_{00}	8.442(0.112)	8.688(0.136)
γ_{10}	0.451(0.179)	1.321(0.085)
γ_{20}	0.744(0.272)	2.636(0.134)
γ_{11}	3.779(0.507)	-
γ_{22}	1.659(0.323)	-
σ_{00}	0.745(0.308)	0.637(0.303)
σ_{11}	0.237(0.083)	0.471(0.116)
σ_{22}	0.700(0.197)	1.472(0.295)
σ^2	44.00(1.05)	44.01(1.05)

Table 3
Ratios of mean square errors and ratios of absolute relative errors (in parentheses) for the diagonal and the diagonal with *Z* estimators under the two respective data generation models

Estimator	Data Generation Model	
	Diagonal	Diagonal with *Z*
Diagonal	100.0	110.3
	(100.0)	(125.4)
Diagonal with *Z*	126.2	100.0
	(107.5)	(100.0)

3.3 Comparisons with regression estimator

One essential advantage of the multilevel models over regression models is to recognize that groups (here the small areas) share common features; they are not completely independent as could be assumed, for example by using separate linear regression model for each small area. Nevertheless, the relatively small intraclass correlation observed for the data set used plus the fact that each small area has on average 28 units, could make one think that in this case the use of the multilevel model would not result in great improvement in the small area estimators. However, it is gratifying to know that even in these circumstances the multilevel model small area estimator performs on average better than the synthetic separate regression estimator, under either the multilevel model or even under the regression model. Table 4 illustrates this finding.

The multilevel data generation model used was the General one with the parameters given in section 3.1. The parameters used in the data generation regression model were obtained by fitting a separate regression for each small area.

It can be seen from Table 4 that the Separate Regression estimator which does not explore the difference of small areas through small area random effects shows substantial loss of efficiency when compared with the General estimator.

Table 4
Ratios of mean square errors and ratios of absolute relative errors (in parentheses) for the general and the separate regression estimators under the two respective data generation models

Estimator	Data Generation Model	
	General	Separate Regression
General	100.0	88.1
	(100.0)	(83.1)
Separated Regression	247.6	100.0
	(154.7)	(100.0)

Figure 1 illustrates this fact by showing a plot of the ratio of mean square error between the General estimator and the Separate Regression estimator for each small area. To demonstrate the effect of the small area sample size on the efficiency, the ratio of the MSEs is plotted against the sample size for each small area. It is clear from Figure 1 that the gain in efficiency tends to decrease as the sample size increases.

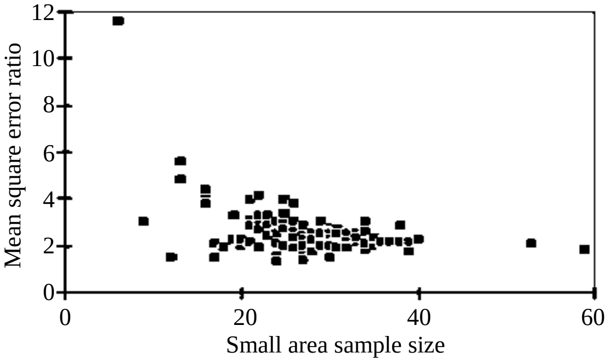

Figure 1 Model-based efficiencies of the general estimator compared with the separate regression estimator for each small area

3.4 An evaluation of the MSE approximation and the MSE estimator

From the simulation results we may investigate the properties of the MSE approximation (2.7). If we consider the General estimator when the General Data Generation model is used the MSE approximation appears to be very good. The average underestimation of the MSE approximation was 0.31% of the MSE value with a range from the largest underestimate of 5.4% of the MSE value through to a largest overestimate of 4.8% of the MSE value. For the situation considered here T_1 contributed on average 94.6% of the total variation and T_3 a further 4.3%. Given the large component of variance due to σ^2, these results are not unexpected. For individual areas the component T_1 varied between 87.4% to 99.1% of the total and T_3 varied between 0.7% and 10.5% of the total. The component T_2 never contributed more than 2.2% of the total MSE for any area.

We also investigated the performance of the MSE estimator represented by equation (2.10) against the "naive" estimator of the MSE, which does not consider the last term of (2.10). The average Root Mean Squared Error of the proposed MSE estimator is 17.5% ranging from 4.7% to 32.3%, while for the naive estimator the average is 20.9% ranging from 5.2% to 47.5%. The MSE estimator is on average unbiased while the naive MSE estimator underestimates the MSE on average by 9.1%, its relative bias ranging from -23.5% to -0.9%. Our results agree with others, see Singh, Stukel and Pfeffermann (1998) and Prasad and Rao (1990), which show that the naive estimator can exhibit severe bias.

4. Discussion

Prasad and Rao (1990) and Battese *et al.*, (1981, 1988) have demonstrated that models which include small area specific components of variance can provide greatly improved small area estimators. Some of the numerical results in this paper show that within the model-based simulation framework even better estimators can be obtained by allowing the small area slopes as well as the intercept to be random.

The overall conclusions from this investigation for this set of parameter values are that: a component of variance model more complex than the Random Intercept estimator is beneficial; overspecification of the model (*e.g.*, using the General estimator with data generated under the Random Intercept Model) does not lead to serious loss of efficiency; the use of small area covariates can also improve the small area estimates; and the use of multilevel models should be preferred rather than the Separate Regression Model. The simulation study confirms that the MSE approximation appears to be precise and the MSE estimation is approximately unbiased, reflecting the variation in MSE between areas, but further theoretical investigation about the exact order of the approximation should be done.

Clearly model fitting and diagnostics are crucial. If we apply a general mixed model in circumstances where it is only a poor fit to the data, then the results may be disappointing. Considerably more investigation is needed to understand what characteristics of specific small areas are likely to provide efficiency gains if general mixed models are used rather than simpler models.

Acknowledgements

We would like to thank the referees and the Editor for their helpful comments on the earlier version of this paper.

Appendix A

Restricted iterative generalized procedure

The generalized least squares estimator of γ in the model (2.1) is given by

$$\tilde{\gamma} = (Z^T X^T V^{-1} XZ)^{-1} (Z^T X^T V^{-1} Y)$$

$$= \left(\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} X_i Z_i \right)^{-1} \left(\sum_{i=1}^{m} Z_i^T X_i^T V_i^{-1} Y_i \right) \quad \text{(A.1)}$$

where $V = \text{Diag}(V_1, ..., V_m)$ and $V_i = \sigma^2 I + X^T \Omega X$ is the covariance matrix of Y_i, $i = 1, ..., m$.

However, V is assumed to be a function of unknown parameters, thus γ cannot be estimated using (A.1). On the other hand, if γ is known then

$$Y^* = \text{vech}[(Y - XZ\gamma)(Y - XZ\gamma)^T] \quad \text{(A.2)}$$

is an unbiased estimator of $\text{vech}(V)$. Furthermore $\text{vech}(V)$ is a linear function of θ. Then we can consider the following linear model:

$$Y^* = F\theta + \xi. \quad \text{(A.3)}$$

Where $F = \partial\text{vech}(V)/\partial\theta$ and ξ is a random variable with mean $O = (0, ..., 0)$ and the covariance of ξ is given by $V_\xi = 2\varphi_n (V \otimes V)\varphi_n^T$. The matrix φ_n is any linear transformation of $\text{vec}(A)$ into $\text{vech}(A)$, and A is any $n \times n$ matrix such that $\text{vech}(A) = \varphi_n \text{vec}(A)$, see Fuller (1987) for further details. Then, assuming that F has full rank and V_ξ is known and non-singular, it may be shown that the Generalized Least Square Estimator of θ is given by

$$\ddot{\theta}_a = \text{cov}(\ddot{\theta}_a) \left(\frac{\partial \text{vec}(V)}{\partial\theta} \right)^T \left(\frac{1}{2} V^{-1} \otimes V^{-1} \right) \text{vec}(\tilde{Y}\tilde{Y}^T) \quad \text{(A.4)}$$

where

$$\text{cov}(\ddot{\theta}_a) = \left[\left(\frac{\partial \text{vec}(V)}{\partial\theta} \right)^T \left(\frac{1}{2} V^{-1} \otimes V^{-1} \right) \left(\frac{\partial \text{vec}(V)}{\partial\theta} \right) \right]^{-1}$$

and

$$\tilde{Y} = Y - XZ\gamma.$$

Note that $\ddot{\theta}_a$ depends on θ and γ, so both may be iteratively estimated. The IGLS procedure starts with an initial estimate of V (that is, setting initial values of θ) which produces an estimate of γ. Hence replacing the initial estimate of V together with the estimate of γ in (A.1) provides an improved estimate of θ. In most cases convergence is achieved after a few iterations between equations (A.1) and (A.4), although it is not always guaranteed.

The RIGLS approach is based on the fact that if γ is estimated by using generalised least squares with V known then

$$E[(Y - XZ\hat{\gamma})(Y - XZ\hat{\gamma})^T] = V - XZ(Z^T X^T V^{-1} XZ)^{-1} Z^T X^T.$$

The equation above suggests that we use

$$(Y - XZ\gamma)(Y - XZ\gamma)^T + XZ(Z^T X^T V^{-1} XZ) Z^T X^T \quad \text{(A.5)}$$

instead of $(Y - XZ\gamma)(Y - XZ\gamma)^T$ at each iteration cycle described above in order to obtain an approximately unbiased estimator of V and consequently of θ.

As pointed out by Goldstein (1986, 1989), if we start with a consistent estimate of γ, say the ordinary least squares estimator, then the final estimates will be consistent providing finite fourth moments exist.

It is worth noting that it is possible for the above procedure to yield negative estimates of variances. This problem can be avoided by imposing constraints at each iteration. For further details on this issue see Goldstein (1986).

Appendix B

An approximation to $E[\hat{\mu}_i - \hat{\mu}_i]^2$

Prasad and Rao (1990), based on Kachar and Harville (1984), developed a second order approximation to the second term of (2.5) under some regularity conditions:

$$E[\hat{\mu}_i - \tilde{\mu}_i]^2 \approx T_3 =$$

$$tr\left[\left(\frac{\partial d_i}{\partial\theta} \right) V \left(\frac{\partial d_i}{\partial\theta} \right)^T E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] \right] \quad \text{(B.1)}$$

where, for the model (2.1), $d_i = \bar{X}_i^T K_i (I \otimes \Omega) X^T V^{-1}$, $K_i = [0, ..., I, ...0]$, is the $(p+1) \times (p+1)m$ matrix with the identity matrix I of order $p+1$ in the i^{th} position and 0 as the null matrix of order $p+1$, and $\hat{\theta}$ is any translation-invariant estimator of $\theta = (\theta_1, ..., \theta_s)$ where $\theta_s = \sigma^2$ and θ_k; $k = 1, ..., s-1$ are the distinct elements of Ω. Goldstein (1989) proves that under normality of the random terms of model (2.1), the RIGLS estimator of θ is equivalent to the Restricted Maximum Likelihood Estimator (RMLE), which is translation invariant.

Let us approximate $E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T]$ to the asymptotic covariance matrix of the RMLE estimator (B). The jk^{th} element of B^{-1} is given by (see Harville 1977)

$$b_{jk}^* = Tr\left(\sum_{i=1}^{m} P_i \frac{\partial V}{\partial\theta_j} P_i \frac{\partial V}{\partial\theta_K} \right)$$

for j and $k = 1, ..., s$ where $P_i = V_i^{-1} - V_i^{-1} X_i Z_i (\sum_{i=1}^m Z_i^T X_i^T V_i^{-1} X_i Z_i) Z_i^T X_i^T V_i^{-1}$. Let $b_{j,k}$ be jk^{th} element of B. After some matrix algebra, it can be shown that

$$T_3 = \overline{X}_i^T (G_i^{-1})^T \left(\sum_{j=1}^{s-1} \sum_{k=1}^{s-1} b_{jk} \Delta_j C_i \Delta_k^T \right) G_i^{-1} \overline{X}_i$$

$$-2\overline{X}_i^T (G_i^{-1})^T \left(\sum_{j=1}^{s-1} b_{j,s} \Delta_j \right) R_i \, \Omega \, \overline{X}_i + b_{ss} \overline{X}_i^T \Omega S_i \, \Omega \overline{X}_i \quad \text{(B.2)}$$

where $C_i = \sigma^{-2} G_i^{-1} X_i^T X_i$; $R_i = \sigma^{-4} G_i^{-2} X_i^T X_i$; $S_i = \sigma^{-6} G_i^{-3} X_i^T X_i$; and

$$\Delta_k = \frac{\partial \Omega}{\partial \theta_k} \quad k = 1, \ldots, s-1$$

is the $s-1$ square derivative matrix with respect to θ_k; $k = 1, \ldots, s-1$.

References

Battese, G.E., and Fuller, W.A. (1981). Prediction of county crop areas using survey and satellite data. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 500-505.

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

Fuller, W.A. (1987). *Measurement Error Models*. Chichester: John Wiley & Sons, Inc.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalised least squares estimation. *Biometrika*, 73, 43-56.

Goldstein, H. (1989). Restricted unbiased iterative generalised least squares estimation. *Biometrika*, 76, 622-623.

Gonzales, M.E. (1973). Use and evaluation of synthetic estimates. *Proceedings of the Social Statistics Section*, American Statistical Association, 33-36.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Harville, D.A. (1977). Maximum likelihood approach to variance component estimation and related problems. *Journal of the American Statistical Association*, 72, 320-340.

Harville, D.A., and Jeske, D.R. (1992). Mean squared error of estimation on prediction under a general linear model. *Journal of the American Statistical Association*, 87, 724-731.

Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423-447.

Holt, D., and Moura, F. (1993a). Mixed models for making small area estimates. In: *Small Area Statistics and Survey Design*, (Eds., G. Kalton, J. Kordos and R. Platek) 1, 221-231. Warsaw: Central Statistical Office.

Holt, D., and Moura, F. (1993b). Small area estimation using multilevel models. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 21-31.

Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79, 853-862.

Longford, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika*, 79, 817-827.

Moura, F.A.S. (1994). Small Area Estimation Using Multilevel Models. University of Southampton. Unpublished Ph.D. Thesis.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of the mean squared error of small - Area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Singh, A.C., Stukel, D. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, Series B, 377-396.