Southampton Statistical Sciences Research Institute
University of Southampton

Small Area Estimation: A Review of Methods Based on the Application of Mixed Models

Ayoub Saei, Ray Chambers

Abstract

This is the review component of the report on small area estimation theory that was prepared as part of Southampton's involvement in the EURAREA "Enhancing small area estimation techniques to meet European Needs" project IST 2000-26290 in the European Union's Fifth Research And Technological Development Framework Programme.

S^{3}RI Methodology Working Paper M03/16

Small Area Estimation: A Review of Methods Based on the Application of Mixed Models

Ayoub Saei and Ray Chambers
Southampton Statistical Sciences Research Institute
University of Southampton
Highfield, Southampton
SO17 1BJ
United Kingdom

September 2003

Abstract

This is the review component of the report on small area estimation theory that was prepared as part of Southampton's involvement in the EURAREA "Enhancing small area estimation techniques to meet European Needs" project IST 2000-26290 in the European Union's Fifth Research And Technological Development Framework Programme

1 Introduction

Efficient estimation of population characteristics for subnational domains is an important aim for many statistical agencies. Such domains are defined as any subdivision of the population for the variable of the interest. However, geographically-based domains, like regions, states, counties, wards and metropolitan areas are typically of most interest. A traditional approach to estimation for such domains is based on application of classical design-based survey sampling methods. Estimates based on this approach are often called direct estimates in the literature. However, sample sizes are typically small or even zero within the domains/areas of interest. This results in the direct estimators having large variances. When there are no sample observations in some of the small areas of interest, direct estimators cannot even be calculated. Small area estimation theory is concerned with resolving these problems.

Often there is auxiliary information that can be used to define estimators for small areas. In some cases these are values of the variable of interest in other, similar, areas, or past values of this variable in the same area or values of other variables that are related to the variable of interest. Estimation and inference approaches based on using this auxiliary information are called indirect or model-based. Methods based on the use of auxiliary information have been characterized in the statistical literature as ″borrowing strength” from the relationship between the values of the response variables and the auxiliary information. Model-based methods have long history, but have only received attention in the last few decades as defining an approach to estimating small area characteristics. In this context, two main ideas have been used in developing models for use in small area estimation. These either assume that the interdomain variability in the response variable can be explained entirely in terms of corresponding variability in the auxiliary information, leading to so-called fixed effect models, or require the assumption that "unexplained" domain specific variability remains even after accounting for the auxiliary information, leading to mixed models incorporating domain specific random effects.

Fixed effect models explain inter-domain variation in the response variable of interest entirely in terms of variation in known factors. Such models have been the mainstay of statistical analysis since the pioneering work of Legendre (1806). Estimates of small area characterics based on fixed effect models are referred to as synthetic estimators (Levy and

French, 1977), composite estimators (Schaible et al, 1977), and prediction estimators (Holt et al, 1979; Sarndal, 1984; and Marker, 1999).

Mixed models also have a long history, but have received special interest only in the last few decades. This is partly due to the heavy computational burden of estimation methods used with such models. Recent developments in computing hardware, software and estimation methods have however led to increased attention being paid to the use of mixed models for data analysis.

Linear mixed models have a wide range of applications. In particular, the ability to predict linear combination of fixed and random effects is one the more attractive properties of such models. In a series of papers, Henderson (1948, 1949, 1959, 1963, 1973, 1975) developed the best linear unbiased prediction (BLUP) method for mixed models. In this case "best" stands for minimum mean square error among all linear unbiased predictors, "linear" means that the predictor is a linear combination of the response variable values and "unbiased" means that expected value of the prediction error (predicted value of variable - actual value of variable) is zero.

The BLUP method has become a powerful and widely used procedure for fitting models for genetic trends in animal populations based on traits measured on both the continuous and the categorical scale. However, the BLUP methods described in Henderson (1949-75) assumed that the variances associated with random effects in the mixed model (the variance components) are known. In practice of course such variance components are unknown and have to be estimated from the data. There are several methods for estimating variance components. Harville (1977) reviews these methods, including maximum likelihood and residual maximum likelihood and three other methods suggested by Henderson. The predictor obtained from the BLUP when unknown variance components are replaced by associated estimators is called the empirical best linear unbiased predictor (EBLUP) and is described in Harville (1990), Robinson (1990) and Harville (1991).

Over the last decade several important approaches have been developed for estimating/predicting the value of a linear combination of fixed and random effects in discrete response data. In virtually all of these random effects are assumed to be normally distributed. Schall (1991), Breslow and Clayton (1993), McGilchrist and Aisbett (1991) and McGilchrist

(1994) have extended the empirical best linear predictor (EBLUP) to generalized linear models. Wolfinger (1993) and Wolfinger and O'Connell (1993) developed essentially the same computational algorithm but via a different approach. Three likelihood expansion techniques including the Solomon and Cox (1992) approach were compared in Breslow and Lin (1995) and Lin and Breslow (1996). Zeger and Karim (1991) introduced a Gibbs sampling approach for generalized linear mixed models. Monte Carlo EM (MCEM) and Monte Carlo Newton-Raphson (MCNR) computation approaches are described in McCulloch (1994) and McCulloch (1997) respectively.

Mixed models have been used to improve estimation of small area characteristics of small area based on survey sampling or census data by Fay and Herriot (1979), Ghosh and Rao (1994), Rao (1999) and Pfeffermann (1999). In these applications, the mixed model derives from the concept that the vector of finite population values is a realisation of a superpopulation. In this context, estimation of a small area mean is equivalent to prediction of the realization of the unobservable random area effect in a linear mixed model for the superpopulation distribution of the variable defining this mean. See Valliant et al (2000) for a discussion of the distinction between the traditional interpretation of model-based prediction and its application in survey sampling.

In addition to EBLUP, empirical Bayes (EB) and hierarchical Bayes (HB) estimation and inference methods have been also applied to small area estimation. Under the EB approach, Bayes estimation and inferential approaches are used in which posterior distributions are estimated from data. Under the HB approach, unknown model parameters (including variance components) are treated as random, with values drawn from specified prior distributions. Posterior distributions for the small area characteristics of interest are then obtained by integrating over these priors, with inferences based on these posterior distributions. Ghosh and Rao (1994) review the application of these estimation methods in small area estimation. Maiti (1998) has used non-informative priors for hyperparameters when applying HB methods. You and Rao (2000) have used HB methods to estimate small area means under random effect models.

Many surveys or censuses are repeated over time, and this means that auxiliary information from past values of a variable of interest can be used to improve current estimates

of this variable. This "borrowing of strength over time" has been used to improve estimators for small areas. Starting with the work of Scott and Smith (1974), Pfeffermann and Burck (1990), Tiller (1991) Rao and You (1992), Singh et al (1994), and Ghosh et al (1996) have used time series models and associated estimation methods to improve estimators for small areas. In particular, Datta et al (1999) have used times series models in estimating State-level unemployment rates for the US.

Often the survey or census data from which small area estimates are needed are discrete or categorical. A general approach for small area estimation based on generalized linear models is described in Ghosh et al (1998). Malec et al (1999) have extended the models described in Malec et al (1997) by including an oversampling component in the likelihood. Farrell et al (1997) extend the mixed logistic model of MacGibbon and Tomberlin (1989). Moura and Migon (2001) further extend this model, introducing a component to account for spatially correlated structure in the binary response data.

The naïve mean square error estimator suggested by Henderson (1975) underestimates the true mean square error of small area estimators based on mixed effect models. Kacker and Harville (1984) introduced an estimator for the mean square error of an estimator of a small area mean based on an approximation to its true mean squared error under such models. Prasad and Rao (1990) developed an approximation to the Kacker and Harville estimator. Harville and Jeske (1992) also developed approximations to the mean square error of predictors. Rivest and Belmonte (2000) have developed the conditional mean square error of a small area estimator.

The aim of this report is to review small area estimation and inference methods based on the application of mixed models within a frequentist environment. Section 2 starts by reviewing the mixed models that have been used in small area estimation. Section 3 then describes the different estimation and inference methods that have been used with these models. Details of the BLUP approach are set out in subsection 3.1, while subsection 3.2 reviews the application of empirical best linear unbiased prediction (EBLUP) methods to small area estimation. Extensions to Generalized Linear Mixed Models (GLMMs) are reviewed in subsection 3.3. Section 4 concludes the report with a review of modern methods for estimating the mean square error of small area estimators based on mixed models.

2 Models for Small Area Means

Let y_{dti} represent the i^{th} population value for a characteristic of interest within a "cell" t within an area d (i = 1,2,…, N_{dt} ; t = 1,2,…,T; d = 1,2,…,D). In general the "cells" can be defined quite freely, and in many cases will include a time component. Similarly, we define vectors $\mathbf{X}_{dti}$ which contain corresponding auxiliary population information (covariates). The population and sample sizes for the area i are assumed known and are given by $N_d = \sum_{t=1}^{T} N_{dt}$ and $n_d = \sum_{t=1}^{T} n_{dt}$ respectively. The objective is to predict the value of the small area mean $\bar{y}_d = N_d^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N_{dt}} y_{dti}$.

Many different types of models have been put forward as a basis for this prediction. In this review we focus on two basic classes of models, corresponding to fixed effect and random effect specifications. The fixed effect specification assumes that all systematic variability between individuals from different areas can be explained in terms of variability in covariates. In contrast the random effects specification assumes that significant systematic variation between small areas remains after covariates are accounted for in the model, and models this via the addition of area specific random coefficients. Models that do not include such random coefficients are referred to as fixed effect below, while those that contain random coefficients (in addition to fixed effects) are referred to as mixed below.

2.1 Linear Fixed Effect Models

Let $\mathbf{y} = \{ y_{dti} \}$ be the population vector of response variable values, with $\mathbf{y}_s$ denoting the values observed in the sample, and let $\mathbf{X} = \{ \mathbf{X}_{dti} \}$ be a corresponding matrix of covariate values known for all units in the population, with $\mathbf{X}_s$ denoting the corresponding sample component of this quantity. A large class of fixed effect models for a population can then be represented by the linear form

(2.1.1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{e}$ denotes a population vector of independent random "noise" components. A fundamental assumption is then that the model (2.1.1) also applies to the sample data. That is, one can write

$$\mathbf{y}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{e}_s$$

where the components of $\mathbf{e}_s$ are still independent random "noise" values. It is clear that the sample data can therefore be used to calculate an efficient estimator of $\boldsymbol{\beta}$. Denote this estimator by $\hat{\boldsymbol{\beta}}$. An obvious predictor of y_{dti} for a non-sample unit in area d is $\mathbf{X}'_{dti}\hat{\boldsymbol{\beta}}$ and the corresponding predictor of the overall population mean of y in area d is

$$\hat{\bar{y}}_d = N_d^{-1}\left(\sum_{t=1}^{T}\sum_{i=1}^{n_{dt}} y_{dti} + \sum_{t=1}^{T}\sum_{i=n_{dt}+1}^{N_{dt}} \mathbf{X}'_{dti}\hat{\boldsymbol{\beta}}\right). \tag{2.1.2}$$

Note that (2.1.2) does not require that there be any sample units in area d. Its basic assumption is that the relationship between $\mathbf{y}$ and $\mathbf{X}$ specified in (2.1.1) above remains the same irrespective of area.

This prediction-based approach to small area estimation is discussed further in Holt et al (1979). Ghosh and Rao (1994) review this approach and Ericksen (1974 and 1975) apply the prediction approach to obtain estimates of population change from 1960 to 1970 for 2586 counties in the US. Levy (1971) and Gonzalez and Hoza (1978) use a regression model-based prediction approach to estimate unemployment in small areas.

Alternatively, classical design-based sampling techniques can be applied to obtain an estimate for the small area mean. Such estimation methods are called direct in the literature. This approach assumes that there are sample units in each small area of interest. The estimator for the mean of y in small area i is then

$$\hat{\bar{y}}_d = \left(\sum_{t=1}^{T}\sum_{i=1}^{n_d} w_{dti}\right)^{-1}\left(\sum_{t=1}^{T}\sum_{i=1}^{n_d} w_{dti}\, y_{dti}\right) = \bar{y}_{wd} \tag{2.1.3}$$

where w_{dti} is the "sample weight" associated with the sample unit indexed by dti. From a modelling perspective, using (2.1.3) corresponds to replacing (2.1.1) by the more restrictive ANOVA-type model

$$y_{dti} = \beta_d + e_{dti} \tag{2.1.4}$$

for each d =1,2,…,D and Var(e_{dti}) = $\sigma^2 w_{dti}^{-1}$. This is a simple form of the fixed effect model (2.1.1) with the matrix **X** defined by vectors that "pick out" the different small areas.

There are two basic problems with this approach. First, from a model-based perspective it assumes that all units in a small area have the same expected value of y. This ignores the variability in y "explained" by the different values in **X**. Second, it requires that the sample size in any small area be large enough so that the sample weighted mean (2.1.3) above can (i) be calculated, and (ii) be a reasonably efficient estimator of the small area population mean. Neither of these conditions hold in practice. Consequently, direct estimators are of little interest in small area estimation.

An extension of the direct estimation approach that "bridges the gap" between (2.1.2) and (2.1.3) is based on the model-assisted approach to design-based sampling theory. This estimates the regression parameter $\boldsymbol{\beta}$ in (2.1.1) using the sample weights w_{dti} to obtain a weighted estimate $\hat{\boldsymbol{\beta}}_w$ (which is not always efficient) and then estimates the small area mean of interest by

$$
\begin{aligned}
\hat{\bar{y}}_d &= N_d^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N_{dt}} \mathbf{X}'_{dti} \hat{\boldsymbol{\beta}}_w + \left(\sum_{t=1}^{T} \sum_{i=1}^{n_{dt}} w_{dti} \right)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n_{dt}} w_{dti} (y_{dti} - \mathbf{X}'_{dti} \hat{\boldsymbol{\beta}}_w) \\
&= \left(\sum_{t=1}^{T} \sum_{i=1}^{n_{dt}} w_{dti} \right)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n_{dt}} w_{dti} y_{dti} \\
&\quad + N_d^{-1} \sum_{t=1}^{T} \sum_{i=1}^{N_{dt}} \mathbf{X}'_{dti} \hat{\boldsymbol{\beta}}_w - \left(\sum_{t=1}^{T} \sum_{i=1}^{n_{dt}} w_{dti} \right)^{-1} \sum_{t=1}^{T} \sum_{i=1}^{n_{dt}} w_{dti} \mathbf{X}'_{dti} \hat{\boldsymbol{\beta}}_w \\
&= \bar{y}_{wd} + (\bar{\mathbf{X}}_d - \bar{\mathbf{X}}_{wd})' \hat{\boldsymbol{\beta}}_w
\end{aligned}
\qquad (2.1.5)
$$

The estimator defined by (2.1.5) is often referred to as a generalised regression estimator or GREG estimator. Technically it still requires that there is sample in every small area of interest, but this requirement is often relaxed and a slightly modified version of (2.1.5) is calculated, defined by $\hat{\bar{y}}_d = \bar{\mathbf{X}}'_d \hat{\boldsymbol{\beta}}_w$ in small areas where n_d is zero (or very small). Clearly this corresponds to using an inefficient version of the model-based estimator (2.1.2) in such areas and the GREG estimator (2.1.5) where the sample size is "reasonable". See Sarndal (1984) and Hidiroglou and Sarndal (1985).

Early developments in small area estimation were based on simple apportionment methods. That is, reasonably accurate means for large subgroups in the population of interest

were calculated, and these were then apportioned within a small area of interest according to the proportional representation of the different subgroups in the small area. (Such methods were called synthetic when they were first introduced in NCHS (1968), but now this name is used to describe any model-based small area estimate based on a fixed effects model.) In our notation an apportionment-based estimator for the mean of y in small area d is

$$\hat{\bar{y}}_d = \sum_{t=1}^{T} \left(\frac{N_{dt} \bar{x}_{dt}}{N_d \bar{x}_d} \right) \hat{\bar{y}}_t \tag{2.1.6}$$

where $\hat{\bar{y}}_t$ denotes the estimate of the population average of y in subgroup t and it is assumed that the covariate x is scalar. When x is identically equal to one it is easy to see that use of (2.1.6) essentially corresponds to the assumption of an ANOVA model for y within subgroups (not areas) in the population. A popular choice however is where x is the value of y from a previous census. In this case this estimator does not have a straightforward linear model interpretation. For example, suppose it is reasonable to assume that $E(y_{dti}) = \beta x_{dti}$ and $\hat{\bar{y}}_t$ is unbiased for $\bar{y}_t$ under this model. Then (2.1.6) is biased, since

$$E(\hat{\bar{y}}_d - \bar{y}_d) = \beta N_d^{-1} \sum_{t=1}^{T} N_{dt} \bar{x}_{dt} \left(\frac{\bar{x}_t}{\bar{x}_d} - 1 \right)$$

is not zero in general. Design-based properties of these estimators are developed in papers by Gonzalez and Waksberg (1978) and Levy and French (1977). Assuming that a design-unbiased direct estimator of $\bar{y}_d$ is available, Gonzalez and Waksberg (1973) developed a design-unbiased estimator of the design mean square error (p-MSE) of the apportionment estimator. The performance of this estimators is investigated further in Levy (1971), who uses this approach to estimate death rates, in Namekata et al (1975), and in Schaible et al (1977) who use it to estimate regional unemployment rates. Laake (1979) compares the prediction approach with the apportionment approach. Marker (1999) derives the apportionment estimator as form of GREG estimator.

A standard approach in small area estimation is to construct a linear combination of estimators suggested by different approaches. Such estimators are referred to as composite estimators. Schaible et al (1977) proposed a composite estimator based on a linear combination of the unbiased direct estimator and the model-based prediction estimator. This is an estimator of the form

$$\hat{\bar{y}}_d = \gamma_d \hat{\bar{y}}_{1d} + (1-\gamma_d)\hat{\bar{y}}_{2d}$$

where $\hat{\bar{y}}_{1d}$ and $\hat{\bar{y}}_{2d}$ are the direct and model-based estimators of the small area mean $\bar{y}_d$. An obvious issue is choice of the coefficient γ_d. Purcell and Kish (1980) also consider this approach, but use a common value of γ when defining the composite estimator.

2.2 Linear Mixed Models

The linear model (2.1.1) assumes that all the systematic variability in the population values making up **y** is explained by the variation in the values in **X**. This validates the assumption that the error vector **e** is "noise". However, in practice the observed sample residuals $\mathbf{r}_s = \mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}$ do not look like noise, and often contain significant between area variation. This implies that the model (2.1.1) is misspecified, and that there are "missing" covariates in the model, whose values vary from one area to another. If there are relatively few small areas making up the population, so that the sample size within any one is reasonable, then this misspecification is easily handled. If **X** contains an intercept term, then this can be replaced by an area factor. A little consideration shows that this leads to the estimator

$$\hat{\bar{y}}_d = \bar{y}_{sd} + (\bar{\mathbf{X}}_d - \bar{\mathbf{X}}_{sd})'\hat{\boldsymbol{\beta}}$$

where now **X** and $\boldsymbol{\beta}$ exclude the intercept term. If **X** does not contain an intercept term (so the regression is through the origin) then (2.1.1) can be replaced by

(2.2.1) $$\mathbf{y}_d = \mathbf{X}_d'(\boldsymbol{\beta} + \boldsymbol{\gamma}_d) + \mathbf{e}_d$$

where the parameter $\boldsymbol{\gamma}_d$ represents the deviation from the overall model (2.1.1) in small area *d*, and the $\boldsymbol{\gamma}_d$ sum to zero across all such areas. This type of constrained model forms the basis of the small area modelling approach adopted for underenumeration estimation in the 2001 UK Census (Abbott *et al*, 2000). The model itself can be fitted to the sample data in the small areas using standard regression software (e.g. SAS).

In many cases, however, there are many small areas, with any particular area containing only a small number of sample units. In such cases introducing a fixed effect for each small area in the model can lead to considerable loss of efficiency (the inefficient direct estimator is, of course, the archetype such area effect-based estimator). In these cases therefore the basic approach replaces fixed effects for the small areas by random effects,

working on the assumption that the small areas can be considered an exchangeable set of population subgroups once unit and area specific covariate information is taken into account. In practice this means (2.1.1) is replaced by the linear mixed model

(2.2.2) $\mathbf{y}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u} + \mathbf{e}_s$

where $\mathbf{Z}_s$ is a matrix of known covariates (often area specific rather than unit specific) and $\mathbf{u}$ is a random vector with zero mean and unknown covariance matrix $\boldsymbol{\Omega}$. This vector of "random effects" is assumed to be uncorrelated with the error vector $\mathbf{e}_s$, which is still assumed to be "white noise" with a zero mean vector and a covariance matrix proportional to a known diagonal matrix $\mathbf{W}_s$, with unknown constant of proportionality σ^2. It immediately follows from these assumptions that

(2.2.3) $\mathrm{E}(\mathbf{y}_s \mid \mathbf{X}_s) = \mathbf{X}_s\boldsymbol{\beta}$

(2.2.4) $Var(\mathbf{y}_s \mid \mathbf{X}_s) = \mathbf{Z}_s\boldsymbol{\Omega}\mathbf{Z}_s' + \sigma^2\mathbf{W}_s$.

The components of the matrix $\boldsymbol{\Omega}$ and σ^2 are called variance components and the model defined by (2.2.3) and (2.2.4) is often referred to as a variance components model. Note that the same variance components structure applies to the population as a whole (and, by subtraction, to the non-sample units). All one needs to do is to delete the "*s*" index in (2.2.3) and (2.2.4). Typically, specialist software is needed to fit models like (2.2.2)-(2.2.3).

Variance components models have a long history, dating back to Airy (1861). However, their application to small area estimation is relatively recent, with interest focussing on prediction of finite population quantities (e.g. small area means) under these models. In this context Royall (1976) gives a very general result that allows one to write down the Best Linear Unbiased Predictor (BLUP) of any linear function $\theta = \mathbf{a}'\mathbf{y}$ of the population y-values under the model (2.1.1) with <u>arbitrary</u> correlation of sample and non-sample units. Here $\mathbf{a}$ is a <u>known</u> vector of constants of the same dimension as the vector $\mathbf{y}$ of population y-values. Using Royall's result, the BLUP of θ is

(2.2.5) $\hat{\theta} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{V}_{rs}\mathbf{V}_{ss}^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})]$

where a subscript of *s* denotes restriction to the sample units, a subscript of *r* denotes restriction to the non-sample units, and $\mathbf{V}_{ss} = Var(\mathbf{y}_s \mid \mathbf{X}_s)$, $\mathbf{V}_{rs} = Cov(\mathbf{y}_r, \mathbf{y}_s \mid \mathbf{X}_r, \mathbf{X}_s)$. The vector $\hat{\boldsymbol{\beta}}$ in this formula is the Best Linear Unbiased Estimator (BLUE) of the parameter $\boldsymbol{\beta}$ in (2.1.1), defined by the GLS estimator

(2.2.6) $$\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}_s'\mathbf{V}_{ss}^{-1}\mathbf{y}_s).$$

In the context of prediction of the mean of y in small area i the quantity θ is defined by a-values that are zero for population units not in small area i and equal to $1/N_d$ for units in this small area. Note that calculation of (2.2.5) requires specification of the correlation structure of the residuals associated with (2.1.1), so this model needs to be extended appropriately. Also, since in any practical situation the coefficients of $\mathbf{V}_{rs}$ and $\mathbf{V}_{ss}$ will be unknown (or at least depend on unknown model parameters), the BLUP (2.2.5) cannot be used in practice. When estimates of these quantities are substituted in (2.2.5) the resulting estimator is often referred to as an Empirical BLUP, or EBLUP.

In order to motivate use of (2.2.5) in the context of a variance components model, we focus on a very simple form of (2.2.2) that has been applied quite widely in small area estimation. This is where there are no subgroup differences (so we can drop the "t" subscript) $\mathbf{X} = \mathbf{1}_n$, $\mathbf{u}$ is a random vector of dimension equal to the number (m) of small areas with $\boldsymbol{\Omega} = \omega^2\mathbf{I}_m$, $\mathbf{Z}_s = diag(\mathbf{1}_{sd})$ is a matrix that "picks out" the component of $\mathbf{u}$ corresponding to any particular small area and $\mathbf{W}_s = \sigma^2\mathbf{I}_n$. Here $\mathbf{I}_a$ denotes the identity matrix of order a and $\mathbf{1}_a$ denotes a vector of ones of length a. In this case (2.2.2) is equivalent to

(2.2.7) $$y_{di} = \beta + \mathrm{u}_d + e_{di}.$$

This model corresponds to the assumption of a common overall mean β, but where area specific sample means vary independently around this common mean, with variance equal to $\omega^2 + n_d^{-1}\sigma^2$. Here ω^2 is the variance of the random area effect u_d. It is straightforward to see that the BLUE of β is then a weighted average of these area-specific means, given by

$$\hat{\beta} = \left(\sum\nolimits_d (\omega^2 + n_d^{-1}\sigma^2)^{-1}\right)^{-1} \sum\nolimits_d (\omega^2 + n_d^{-1}\sigma^2)^{-1}\bar{y}_{sd}.$$

In order to construct the BLUP for the mean of y in area d we apply Royall's formula (2.2.5), noting that, under (2.2.7)

$$\mathbf{V}_{rs} = diag(\omega^2\mathbf{1}_{rd}\mathbf{1}_{sd}')$$

$$\mathbf{V}_{ss} = diag(\omega^2\mathbf{1}_{sd}\mathbf{1}_{sd}' + \sigma^2\mathbf{I}_{sd}).$$

It follows

$$\mathbf{V}_{rs}\mathbf{V}_{ss}^{-1} = diag(\omega^2\mathbf{1}_{rd}\mathbf{1}_{sd}'[\omega^2\mathbf{1}_{sd}\mathbf{1}_{sd}' + \sigma^2\mathbf{I}_{sd}]^{-1})$$

where straightforward manipulations can be used to show

$$[\omega^2 \mathbf{1}_{sd}\mathbf{1}'_{sd} + \sigma^2 \mathbf{I}_{sd}]^{-1} = \sigma^{-2}[\mathbf{I}_{sd} - \frac{\omega^2}{\sigma^2 + n_d\omega^2}\mathbf{1}_{sd}\mathbf{1}'_{sd}].$$

Substituting these results in (2.2.5) and taking the components of the vector **a** there as being zero for populations units outside area *i* and N_d^{-1} for units in the area, we see that the BLUP of the area *i* mean of *y* is

(2.2.8) $$\hat{\bar{y}}_d = N_d^{-1}\left(n_d \bar{y}_{sd} + (N_d - n_d)\hat{\beta} + (N_d - n_d)\left(\frac{n_d\omega^2}{\sigma^2 + n_d\omega^2}\right)(\bar{y}_{sd} - \hat{\beta})\right).$$

Note that when ω^2 is zero (there is no area effect) $\hat{\beta}$ reduces to the overall sample mean of *y* and (2.2.8) reduces to the "standard" prediction estimator of $\bar{y}_d$ under a common mean and variance model for the distribution of the population values of *y*. Note also that implicit in (2.2.8) is the BLUP of the random area effect u_i.

$$\hat{u}_d = \left(\frac{n_d\omega^2}{\sigma^2 + n_d\omega^2}\right)(\bar{y}_{sd} - \hat{\beta}).$$

This is sometimes referred to as the "shrunken" area *i* residual.

Extending the basic ideas set out in Ericksen (1973, 1974) and Madow and Hansen (1975), Fay and Herriot (1979) used (2.2.2), with **Z** just containing an intercept term (so the model only includes a random area effect), to estimate average incomes for small areas (population less than 1000) using 1970 US Census data. In particular, their model assumed that both the error term and the random effects followed independent normal distributions. This model was also used by Ericksen and Kadane (1985) to produce small area population produce estimates. Battese and Fuller (1981) and Battese et al. (1988) applied the same model to estimate mean crop acreage for 12 counties (small areas) in Iowa. A similar model was applied to estimation of population changes for local areas by Ericksen (1973), while Kott (1989) used it to investigate robust small domain estimation. Prasad and Rao (1990) used (2.2.2) with $\mathbf{X}_d = \mathbf{Z}_d$ to develop estimators of the mean square error of the BLUP. Datta and Ghosh (1991) extended the models considered in Prasad and Rao (1990) and in Battese et al. (1988) to cross-classified data and to two stage sampling. In the latter case this was accomplished by the introduction of separate random effects for the different stages of sampling. Hulting and Harville (1991) discuss all these models in their comparison of Bayesian and frequentist approaches to the small area estimation.

In the context of estimating farmland values in the United States, Pfeffermann and Barnard (1991) use the model (2.2.2). Their response variable is the value reported the Agricultural and Land Values Survey (ALVS) and they assume that land values reported by farmers residing in the same county are distributed randomly around the county mean. The county means themselves were modelled as functions of auxiliary variables, representing known county characteristics and random state and county effects. Ghosh and Rao (1994) give an excellent review of all these developments.

The general model (2.2.2) can be interpreted as a multilevel model (Goldstein, 1995) and Moura (1994) and Moura and Holt (1999) use this framework to develop estimates for small area means. Lahiri and Rao (1995) also consider the Fay-Herriot version of (2.2.2) and show that an estimator of the mean square error of the BLUP, derived by Prasad and Rao (1990) under an assumption of normality of small area means, is robust to this assumption. Datta and Lahiri (1995) also study the Fay-Herriot version of (2.2.2) and propose a robust hierarchical Bayes technique for estimating small area means when one or more outliers are present. The performance of Bayes and frequentist methods for estimating the mean squared error of small area estimators based on the model (2.2.2) is discussed in Singh et al (1998). Prasad and Rao (1999) develop a model-assisted estimator of a small area mean under the simple model (2.2.7) and Stukel and Rao (1999) have obtained estimators for small area means under nested error regression models. Hierarchical Bayes estimation of small area means under multilevel models is described in You and Rao (2000), while Rao (2001) discusses small area estimation with applications to agriculture based on the use of linear mixed models. The impact of outliers on small area estimation under linear mixed models has been investigated by Li (2000), who derives robust methods for estimating small area means under the version of (2.2.2) investigated by Battese et al. (1988).

Small area estimation methods that "borrow strength over time" have also been investigated. Beginning with the seminal work of Scott and Smith (1974), times series models have been fitted to aggregate survey data in the context of estimating corresponding population means. Scott et al (1977), Smith (1978), Jones (1980), Binder and Dick (1986, 1989), Binder and Hidiroglou (1986), Tiller (1989), Bell and Hillmer (1990) and Pfeffermann (1991) are relevant references.

Tiller (1991) uses a time series modelling approach to estimate state level unemployment rates from the Current Population Survey. This approach is based on a time series model for the true unemployment rate that includes auxiliary variables and an ARMA model for the sampling errors for the survey-based estimates of these rates. Pfeffermann and Burck (1991) discuss a general class of models that combine time series and cross-sectional data in order to estimate small area means. Their approach assumes a random walk model for the model regression coefficients and uses the Kalman filter to obtain estimates of these coefficients as well as the associated variance components. Singh at al (1994) extend the model considered by Fay-Herriot (1979) to repeated survey data and use EBLUP methods to obtain estimates of small area means. Their simulation results supported a combined cross-sectional and time series approach to modelling for small area estimation. Rao and Yu (1994) also extend the Fay-Herriot model to cross-sectional and time series data. Their proposed model is of the form

$$\mathbf{y}_{dt} = \mathbf{X}_{dt}\boldsymbol{\beta} + \mathbf{Z}_{dt}\mathbf{u}_{1d} + u_{2dt}\mathbf{1}_{dt} + \mathbf{e}_{dt}$$

where $\mathbf{u}_{1d}$ is a cross-sectional random effect that does not vary with time, while u_{2dt} is the outcome of a AR(1) process. Pfeffermann and Burck (1990) and Singh et al (1991) consider similar models, but assume specific models for the sampling error. Ghosh et al (1996) also consider a time specific random effects model with the time effects following a random walk. Datta et al (1999) use hierarchical Bayes methods to estimate State level unemployment rates in the US using the Rao and Yu (1994) model and generalise the Fay-Herriot (1979) model to time series data. Pfeffermann (1999), Rao (1999) and Marker (1999) are reviews of these developments.

So far, the emphasis has been on estimating a single characteristic (e.g. a small area mean). However, estimation of multiple characteristics is also of interest in small area estimation. Datta et al (1999) consider corn and soybeans output as two related characteristics and use nested regression models to develop estimates of multivariate small area means based on EBLUP and EB methods. They also obtain a second order approximation (based on Prasad and Rao, 1990) to the mean squared error of the EBLUP estimators. Their approach is based on the model considered by Battese et al (1988) and their results demonstrate the superiority of multivariate models over univariate models. Datta et al (1996) apply multivariate area level

models to obtain estimates of median income for four-person families in the U.S. Univariate shrinkage (composite) methods for estimating small area means and rates are extended to multivariate shrinkage methods by Longford (1999), who also illustrates the advantages of multivariate shrinkage over univariate shrinkage methods. Finally, Fuller and Harter (1987) discuss methods for estimating multivariate small area means.

2.3 Non-Linear Mixed Models

Extension of small area estimation methods to nonlinear models has largely been carried out via Bayesian methods. MacGibbon and Tomberlin (1989) use EB methods to estimate small area proportions. Their analysis uses simulated data based on the binomial-logistic mixed model corresponding to (2.2.2). This is defined by

(2.3.1) $$\text{logit}(\pi_d) = \mathbf{X}_d\boldsymbol{\beta} + \mathbf{Z}_d\mathbf{u}$$

where $\boldsymbol{\pi}_d$ is a vector of probabilities associated with the units in small area i and the individual values in small area i are assumed to be independent Bernoulli outcomes based on these probabilities.

In the absence of population microdata, Farrell et al (1997) use a second order approximation to the EB method to develop estimates of small area proportions based on MacGibbon and Tomberlin's model. They use bootstrap methods to account for uncertainty caused by estimation of prior parameters. Malec et al (1997) apply HB and EB methods to the model (2.3.1) to estimate the proportion of individuals who visit a doctor at least once within 12 months. Farrell (1997) extends the EB methods in Farrell et al (1997) to multinomial and ordinal models.

A general HB approach to small area estimation based on the generalized linear model is described in Ghosh et al (1998). Their approach also applies to models with spatial correlation. Maiti (1998) uses non-informative priors for hyperparameters when applying HB estimation methods to estimation of mortality rates for disease mapping. Malec et al (1999) extend the hierarchical model of Malec et al (1997) by including an oversampling component in the likelihood. They then use EB estimation methods to estimate overweight prevalence in U.S. You and Rao (2000) have develop HB methods for estimating small area means under multilevel models. Moura and Migon (2001) extend the mixed effects logistic model of

MacGibbon and Tomberlin (1989) by introducing a further component to account for spatial correlation in the binary response data. They then use the model selection methodology proposed by Gelfand and Ghosh (1998) to select a best model.

3 Estimation Methods for Models with Random Effects

There are three standard approaches to small area estimation based on models with random effects. These are Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB) and Hierarchical Bayes (HB). The EBLUP method usually requires maximum likelihood (ML) or residual (restricted) maximum likelihood (REML) estimation of parameters associated with random area effects, and is identical to the EB and HB approaches in some situations. In this report we focus on the frequentist approach to small area estimation and so we describe the ML and REML methods used in the EBLUP approach in more detail in the following sections.

Althogh we do not treat Bayesian methods in any detail in this review, Empirical Bayes (EB) and Hierarchical Bayes (HB) methods form an important and widely used class of inferential techniques used in small area estimation. Consequently it is appropriate that we briefly describe how they are typically applied. In the EB method, the joint prior and posterior distributions of the small area quantities of interest are first estimated from the sample data. Typically, these depend on unknown model parameters, which are then estimated via the usual techniques eg. ML or REML. These estimates are "plugged into" the posterior distribution of the small area quantities of interest, with inference about any particular small area quantity based on the marginal posterior distribution of this quantity.

A basic problem with the EB method is the necessity to account for the uncertainty in estimating the prior/posterior. Several approaches have been proposed to account for this uncertainty, with the delta and bootstrap methods being the most popular. Deely and Lindley (1981), Morris (1983a, b) and Kass and Steffey (1989) discuss the delta methods, while Laird and Louis (1987) propose a bootstrap method. For more detail on the EB method see Morris (1983).

In the HB method, the fixed effect parameter β and the variance components are treated as random. The method assumes a joint prior distribution for these parameters, and

Bayes Theorem is used to define the joint posterior distribution of the small area quantities of interest. This is usually done in several stages. For example Fay-Herriot (1979) first derive the posterior distribution of $\boldsymbol{\beta}$ given the variance component σ_u^2 and then the posterior distribution of σ_u^2. For specified variance components the HB estimate of a small area mean typically coincides with the BLUP and EB estimate of this quantity

The HB method has the advantage that because modelling is carried out in stages, each stage can be relatively simple and easy to understand, even though the entire model fitting process can be rather complicated. Furthermore, under some assumptions concerning the prior distribution of the model parameters, the HB estimators have smaller MSE than corresponding BLUP-based estimators (Ghosh and Rao, 1994).

However, the HB method also has two major disadvantages. The first is its lack of robustness to specification of the prior distribution. The other is its computational complexity. Typically, application of the HB method requires simulation-based methods such as Markov Chain Monte Carlo (MCMC) to be used in order to approximate the posterior distribution and hence the posterior means and variances of the small area quantities of interest.

3.1 Best Linear Unbiased Prediction (BLUP)

In section 2.2 we introduced the linear mixed model (2.2.2) as a way of characterising between area variation in the values of the characteristic *Y*. In particular, the aim is to predict/estimate a linear function $\mathbf{a}'\mathbf{y}$ of the population *y*-values given this model. We assume that the vector $\mathbf{y}$ can be partitioned as $\mathbf{y} = [\mathbf{y}_s', \mathbf{y}_r']'$ where subscripts of *s* and *r* corresponding to sample and non-sample population units. These subscripts will be used below to denote conformable partitions of other vectors and matrices. Thus, the vector $\mathbf{a}$ is partitioned conformably as $\mathbf{a} = [\mathbf{a}_s', \mathbf{a}_r']'$. The linear function $\mathbf{a}'\mathbf{y}$ can then be written

(3.1.1) $$\mathbf{a}'\mathbf{y} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'\mathbf{y}_r.$$

The first term in (3.1.1) depends only on the sample values and is known after the sample is observed. The second term, which depends on the non-sample values, is unknown. The problem of using the sample values $\mathbf{y}_s$ to predict the linear function $\mathbf{a}'\mathbf{y}$ therefore becomes the problem of predicting the linear function $\mathbf{a}_r'\mathbf{y}_r$ under (2.2.2). In turn, this is a special case of

the more general problem of predicting the linear function $\tau = \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u})$ given the linear mixed model

(3.1.2) $\mathbf{y}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u} + \mathbf{e}_s$

where the vector of random effects $\mathbf{u}$ are partitioned into m subvectors $\mathbf{u} = [\mathbf{u}_1', \mathbf{u}_2', \ldots, \mathbf{u}_m']'$, with $\mathbf{Z}_s$ partitioned conformably as $\mathbf{Z}_s = [\mathbf{Z}_{s1}, \ldots, \mathbf{Z}_{sm}]$, and $\mathrm{Var}(\mathbf{u}) = \sigma^2\boldsymbol{\Omega} = \sigma^2 blk\text{-}diag(\varphi_j\boldsymbol{\Omega}_j)$ with $\varphi_j = \sigma_j^2/\sigma^2$. The random errors $\mathbf{e}_s$ are assumed to be independent of $\mathbf{u}$ with variance $\sigma^2\mathbf{W}_s$. The variance-covariance matrix of $\mathbf{y}_s$ is therefore $\mathrm{Var}(\mathbf{y}_s) = \sigma^2(\mathbf{W}_s + \mathbf{Z}_s\boldsymbol{\Omega}\mathbf{Z}_s') = \sigma^2\boldsymbol{\Sigma}_s$.

Following Henderson (1963), we consider predictors of τ that are linear functions of $\mathbf{y}_s$, i.e.,

(3.1.3) $\tilde{\tau} = \boldsymbol{\Delta}_s'\mathbf{y}_s + \nabla$

where $\boldsymbol{\Delta}$ and ∇ are vector and scalar respectively. Unbiasedness of $\tilde{\tau}$ under (3.1.2) implies that

(3.1.4) $E(\tilde{\tau}) = \boldsymbol{\Delta}_s'\mathbf{X}_s\boldsymbol{\beta} + \nabla = E(\tau) = \mathbf{a}_r'\mathbf{X}_r\boldsymbol{\beta}$.

For this to hold in general we must have $\nabla = 0$ and

(3.1.5) $\boldsymbol{\Delta}_s'\mathbf{X}_s = \mathbf{a}_r'\mathbf{X}_r$.

Put $\mathbf{C}_s = \mathrm{Cov}(\mathbf{y}_s, \mathbf{u})$. The vector $\boldsymbol{\Delta}_s$ is then defined by solving the constrained optimisation problem

(3.1.6) Minimise $Var(\boldsymbol{\Delta}_s'\mathbf{y}_s - \tau) = \sigma^2\boldsymbol{\Delta}_s'\boldsymbol{\Sigma}_s\boldsymbol{\Delta}_s + Var(\tau) - 2\boldsymbol{\Delta}_s'\mathbf{C}_s\mathbf{Z}_r'\mathbf{a}_r$

subject to: $\boldsymbol{\Delta}_s'\mathbf{X}_s = \mathbf{a}_r'\mathbf{X}_r$.

The Lagrangian for this problem is

(3.1.7) $L(\boldsymbol{\Delta}_s, \boldsymbol{\lambda}) = \sigma^2\boldsymbol{\Delta}_s'\boldsymbol{\Sigma}_s\boldsymbol{\Delta}_s - 2\boldsymbol{\Delta}_s'\mathbf{C}_s\mathbf{Z}_r'\mathbf{a}_r - 2(\boldsymbol{\Delta}_s'\mathbf{X}_s - \mathbf{a}_r'\mathbf{X}_r)\boldsymbol{\lambda}$.

Differentiating (3.1.7) with respect to $\boldsymbol{\Delta}_s$ and $\boldsymbol{\lambda}$ and equating to zero yields the estimating equations

(3.1.8) $$\begin{bmatrix} \sigma^2\boldsymbol{\Sigma}_s & \mathbf{X}_s \\ \mathbf{X}_s' & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\Delta}_s \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{C}_s\mathbf{Z}_r'\mathbf{a}_r \\ \mathbf{X}_r'\mathbf{a}_r \end{bmatrix}.$$

Using Searle (1982, pg 261-262), the inverse of the matrix of coefficients on the left hand side of (3.1.8) is given by

(3.1.9) $$\begin{bmatrix} \sigma^{-2}\boldsymbol{\Sigma}_s^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\sigma^{-2}\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s \\ \mathbf{I} \end{bmatrix}(-\sigma^{-2}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\begin{bmatrix} -\sigma^{-2}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1} & \mathbf{I} \end{bmatrix}.$$

Using (3.1.9), the optimum value of $\boldsymbol{\Delta}_s$ is therefore

$$
\begin{aligned}
\Delta_s^{opt} &= \sigma^{-2}\boldsymbol{\Sigma}_s^{-1}\mathbf{C}_s\mathbf{Z}_r'\mathbf{a}_r \\
&\quad + \boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}(\mathbf{X}_r'\mathbf{a}_r - \sigma^{-2}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{C}_s\mathbf{Z}_r'\mathbf{a}_r) \\
&= \boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_r'\mathbf{a}_r \\
&\quad + \sigma^{-2}\boldsymbol{\Sigma}_s^{-1}[I - \mathbf{X}_s(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}]\mathbf{C}_s\mathbf{Z}_r'\mathbf{a}_r
\end{aligned}
\tag{3.1.10}
$$

and so the BLUP of τ is

$$
\tilde{\tau} = \mathbf{a}_r'[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \sigma^{-2}\mathbf{Z}_r\mathbf{C}_s'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})] \tag{3.1.11}
$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{y}_s$ is Aitken's generalized least squares (GLS) estimator of $\boldsymbol{\beta}$ under (3.1.2). It immediately follows that the BLUP of $\theta = \mathbf{a}'\mathbf{y}$ is

$$
\tilde{\theta} = \mathbf{a}_s'\mathbf{y}_s + \mathbf{a}_r'[\mathbf{X}_r\hat{\boldsymbol{\beta}} + \sigma^{-2}\mathbf{Z}_r\mathbf{C}_s'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})]. \tag{3.1.12}
$$

Substituting $\mathbf{C}_{rs} = Cov(\mathbf{y}_r, \mathbf{y}_s)$ for $\mathbf{Z}_r\mathbf{C}_s'$ in (3.1.12) leads to the BLUP for θ derived by Royall (1976), while setting $\mathbf{X}_r$ and $\mathbf{a}_s$ to zero, $\mathbf{a}_r$ to one and $\mathbf{Z}_r$ to the identity matrix in (3.1.11) leads to the BLUP $\hat{\mathbf{u}} = \sigma^{-2}\mathbf{C}_s'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}})$ obtained by Henderson (1963). The BLUP of linear combination of fixed and mixed effects set out in Hoderson (1975) is also a special case of (3.1.11).

3.2 Empirical Best Linear Unbiased Prediction (EBLUP)

The BLUP development set out in the preceding section assumes variance components are known. In practice of course, this is hardly ever the case. We therefore need to estimate these variance components from the sample data. The empirical best linear unbiased prediction (EBLUP) method replaces the unknown variance components in BLUP by these estimates.

Henderson (1963) introduced three methods (I, II and III) of estimating variance components. These were in common use up to the 1970s, when advances in computing power led to the introduction of maximum likelihood (ML) and residual maximum likelihood (REML) estimators. In this section we describe these estimation methods.

Henderson (1975) showed that substituting estimated values of variance components in the BLUP led to biased predictions. However, Kackar and Harville (1981) showed that a two-stage approach (first estimate variance components, then use these to estimate and predict fixed parameters and random components) leads to unbiased predictors provided the distribution of the data vector is symmetric about its expected value and provided the variance

component estimators are translation invariant and are even functions of the data vector. They showed that the ML and REML variance component estimators have these properties.

3.2.1 Maximum Likelihood Estimation of Variance Components (ML)

The ML approach requires parametric specification of the distribution of the random component in a mixed model. A standard assumption is that this component is normally distributed. With this extra assumption, we can write down the log-likelihood function generated by the observation vector $\mathbf{y}_s$ under the linear mixed model (3.1.2) as

$$l = -(1/2)[n\ln(2\pi\sigma^2) + \ln|\boldsymbol{\Sigma}_s| + \sigma^{-2}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})]$$

Differentiation of this log-likelihood function with respect to the parameters $\boldsymbol{\beta}, \sigma^2$ and φ_j leads to the ML score functions

$$\begin{aligned} \partial l/\partial\boldsymbol{\beta} &= \sigma^{-2}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \\ \partial l/\partial\sigma^2 &= -(1/2)[n\sigma^{-2} - \sigma^{-4}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) \\ \partial l/\partial\varphi_j &= -(1/2)[tr(\boldsymbol{\Sigma}_s^{-1}\mathbf{Z}_{sj}\boldsymbol{\Omega}_j\mathbf{Z}_{sj}') \\ &\quad - \sigma^{-2}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\boldsymbol{\Sigma}_s^{-1}\mathbf{Z}_{sj}\boldsymbol{\Omega}_j\mathbf{Z}_{sj}'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})] \end{aligned} \tag{3.2.1.1}$$

Equating these score functions to zero yields the ML estimating equations for $\boldsymbol{\beta}, \sigma^2$ and φ_j.

Given the MLEs for σ^2 and φ_j, and hence the MLE $\hat{\boldsymbol{\Sigma}}_s$ for $\boldsymbol{\Sigma}_s$, it is clear that the MLE for $\boldsymbol{\beta}$ is just the GLS estimator of this parameter defined by $\hat{\boldsymbol{\Sigma}}_s$. That is,

$$\hat{\boldsymbol{\beta}}_{ML} = (\mathbf{X}_s'\hat{\boldsymbol{\Sigma}}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\hat{\boldsymbol{\Sigma}}_s^{-1}\mathbf{y}_s.$$

However, the estimating equations for the variance components have no analytic solution and so have to be solved numerically.

3.2.2 Residual Maximum Likelihood of Variance Components (REML)

For the linear mixed model (3.1.2), the expectations of the component score functions for σ^2 and φ_j defined by (3.2.1.1) are zero. However, if $\hat{\boldsymbol{\beta}}_{ML}$ is substituted for $\boldsymbol{\beta}$ in these functions, then these expectations are no longer zero. For example, the expected value of the score function for σ^2 when $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}_{ML}$ is $-(p/2)\sigma^2$ and so the ML estimator for this parameter is biased. On the other hand, an unbiased estimator for σ^2 is defined by

$$(n-p)\sigma^{-2} - \sigma^{-4}(\mathbf{y}_s - \mathbf{X}_s\hat{\boldsymbol{\beta}}_{ML})'\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}_{ML}) = 0 \tag{3.2.2.1}$$

where n is sample size and p is the rank of the matrix $\mathbf{X}_s$. Similarly, we can show that when $\boldsymbol{\beta}$ is replaced by $\hat{\boldsymbol{\beta}}_{ML}$ the expectation of the score function for φ_j in (3.2.1.1) is not zero but $-(1/2)\partial(\ln|\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s|/\partial\varphi_j$. Consequently, an unbiased estimator for φ_j is defined by

$$\frac{\partial \ln|\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s|}{\partial\varphi_j}+\frac{\partial \ln|\boldsymbol{\Sigma}|}{\partial\varphi_j}+\frac{1}{\sigma^2}(\mathbf{y}_s-\mathbf{X}_s\hat{\boldsymbol{\beta}}_{ML})'\frac{\partial\boldsymbol{\Sigma}_s^{-1}}{\partial\varphi_j}(\mathbf{y}_s-\mathbf{X}_s\hat{\boldsymbol{\beta}}_{ML})=0. \tag{3.2.2.2}$$

The Residual Maximum Likelihood (REML) approach uses the above idea to reduce bias when estimating variance components. In particular, this approach starts by first transforming $\mathbf{y}_s$ into two independent vectors $\mathbf{y}_{s1}=\mathbf{K}_1\mathbf{y}_s$ and $\mathbf{y}_{s2}=\mathbf{K}_2\mathbf{y}_s$. The $\mathbf{y}_{s1}$ vector has a distribution that does not depend on the fixed effect $\boldsymbol{\beta}$ and hence satisfies $E(\mathbf{K}_1\mathbf{y}_s)=0$, i.e. $\mathbf{K}_1\mathbf{X}_s=\mathbf{0}$, while the $\mathbf{y}_{s2}$ vector is independent of $\mathbf{y}_{s1}$ and satisfies $\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_2'=\mathbf{0}$. The matrix $\mathbf{K}_1$ is chosen to have maximum rank, i.e. $n-p$, and so the rank of $\mathbf{K}_2$ is p. The likelihood function of $\mathbf{y}_s$ is then the product of the likelihoods of $\mathbf{y}_{s1}$ and $\mathbf{y}_{s2}$. REML estimators of variance components are maximum likelihood estimators based on $\mathbf{y}_{s1}$. That is, the REML method estimates variance components by maximising the log-likelihood function defined by $\mathbf{y}_{s1}=\mathbf{K}_1\mathbf{y}_s$,

$$l_{REML}=-(1/2)[(n-p)\ln 2\pi\sigma^2+|\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1|+\sigma^{-2}y_s'\mathbf{K}_1(\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1)^{-1}\mathbf{K}_1\mathbf{y}_s]$$

where the matrix $\mathbf{K}_1=\mathbf{W}_s^{-1}-\mathbf{W}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\mathbf{W}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{W}_s^{-1}$. Note that if $\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1$ is not of full rank, then $|\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1|$ must be interpreted as the determinant of its linearly independent rows and columns. Given this definition of $\mathbf{K}_1$, the matrix $\mathbf{K}_2$ is defined as $\mathbf{K}_2=\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}$. The log-likelihood function defined by $\mathbf{y}_{s2}=\mathbf{K}_2\mathbf{y}_s$ is

$$\begin{aligned}l_L&=-(1/2)[p\ln 2\pi\sigma^2+\ln|\mathbf{K}_2\boldsymbol{\Sigma}_s\mathbf{K}_2'|\\&\quad+\sigma^{-2}(\mathbf{y}_{s2}-E(\mathbf{y}_{s2}))'(\mathbf{K}_2\boldsymbol{\Sigma}_s\mathbf{K}_2')^{-1}(\mathbf{y}_{s2}-E(\mathbf{y}_{s2}))]\\&=-(1/2)[p\ln 2\pi\sigma^2+\ln|\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s|\\&\quad+\sigma^{-2}(\mathbf{y}_s-\mathbf{X}_s\boldsymbol{\beta})'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s\boldsymbol{\Sigma}_s^{-1}(\mathbf{y}_s-\mathbf{X}_s\boldsymbol{\beta})]\end{aligned}$$

For given values of the variance components, $\boldsymbol{\beta}$ is estimated by maximizing l_L, leading to $\hat{\boldsymbol{\beta}}=(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{y}_s$.

To gain insight into the REML method consider the simple standard linear regression/ANOVA model, i.e. $\mathbf{y}_s=\mathbf{X}_s\boldsymbol{\beta}+\mathbf{e}$. Assuming that the random error vector $\mathbf{e}$ is a multivariate normal with zero mean and variance of $\sigma^2\mathbf{I}$, the log-likelihood functions l_L and l_{REML} simplify to

$$l_L = -(1/2)[p\ln 2\pi\sigma^2 + \ln|\mathbf{X}_s'\mathbf{X}_s| + \sigma^{-2}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})]$$

$$l_{REML} = -(1/2)[(n-p)\ln 2\pi\sigma^2 + \ln|\mathbf{K}_1| + \sigma^{-2}\mathbf{y}_s'\mathbf{K}_1\mathbf{y}_s]$$

where $\mathbf{K}_1 = \mathbf{I} - \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'$. In this special case, the REML estimator of σ^2 is the unbiased $\hat{\sigma}^2_{REML} = \mathbf{y}_s'\mathbf{K}_1\mathbf{y}_s/(n-p) = (\mathbf{y}_s'[\mathbf{I} - \mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s]\mathbf{y}_s)/(n-p)$.

Following Patterson and Thompson (1971), the residual maximum likelihood estimates of the variance components σ^2 and φ_j are the values of these parameters that maximise l_{REML}. Differentiation of l_{REML} with respect to these variance components yields the REML estimating equations

$$\partial l_{REML}/\partial\sigma^2 = -(1/2)[(n-p)\sigma^{-2} - \sigma^{-4}y_s'\mathbf{K}_1(\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1)^{-1}\mathbf{K}_1\mathbf{y}_s]$$

$$\partial l_{REML}/\partial\varphi_j = -(1/2)[\{\mathrm{tr}(\mathbf{K}_{\varphi_j}) - \sigma^{-2}\mathbf{y}_s'\mathbf{K}_{\varphi_j}\mathbf{K}_1(\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1)^{-1}\mathbf{K}_1\mathbf{y}_s]$$

where $\mathbf{K}_{\varphi_j} = \mathbf{K}_1(\mathbf{K}_1\boldsymbol{\Sigma}_s\mathbf{K}_1)^{-1}\mathbf{K}_1\partial\boldsymbol{\Sigma}_s/\partial\varphi_j$. The estimating equations for the variance components have no analytic solution and so have to be solved numerically.

In order to motivate use of the REML method in the context of small area estimation, we use the simple model (2.2.7) to illustrate the steps in the iterative calculation of the variance components. In this case $\varphi_j = \varphi = \omega^2/\sigma^2$ and the steps in the iterations for this special case are as follows:

1. Set $k = 0$ and assign an initial value φ_0 to the variance component φ.
2. Set $k = k + 1$
3. Put $\gamma_d^{(k)} = \mathrm{n}_d\hat{\varphi}_{k-1}(1 + \mathrm{n}_d\hat{\varphi}_{k-1})^{-1}$.
4. Calculate $\hat{\beta}_k = \left[\sum_d \gamma_d^{(k)}\right]^{-1}\sum_d \gamma_d^{(k)}\overline{\mathrm{y}}_{\mathrm{sd}}$ and $\hat{\mathbf{u}}_k = (\mathrm{diag}(\gamma_d^{(k)}))(\overline{\mathbf{y}}_s - \mathbf{1}_D\hat{\beta}_k)$. Here $\overline{\mathbf{y}}_s$ is the vector of sample means for the m small areas.
5. Calculate $\hat{\sigma}_k^2 = (n-1)^{-1}\mathbf{y}_s'(\mathbf{y}_s - \mathbf{1}_n\hat{\beta}_k - \mathbf{Z}_s\hat{\mathbf{u}}_k)$, where $\mathbf{Z}_s$ is a $n \times D$ matrix that "picks out" the small areas.
6. Calculate $\hat{\varphi}_k = (D - r_k)^{-1}\hat{\sigma}_k^{-2}\hat{\mathbf{u}}_k'\hat{\mathbf{u}}_k$

   where $r_k = \varphi_{k-1}^{-1}tr[\mathbf{G}_k + \mathbf{G}_k\mathbf{Z}_s'\mathbf{1}_n(\mathbf{1}_n'\boldsymbol{\Sigma}_k^{-1}\mathbf{1}_n)^{-1}\mathbf{1}_n'\mathbf{Z}_s\mathbf{G}_k]$

   $G_k = diag(\gamma_d^{(k)}/n_d)$ and $\boldsymbol{\Sigma}_k = blk - diag(\mathbf{I}_{n_d} - \gamma_d^{(k)}/n_d\mathbf{1}_{n_d}\mathbf{1}_{n_d}')$.
7. Return to step 2 and repeat the procedure until the estimates converge.
8. The Empirical BLUP (EBLUP) of the area i mean of Y is then

$$\hat{\bar{y}}_d = N_i^{-1}[n_d \bar{y}_{sd} + (N_d - n_d)\hat{\beta} + (N_d - n_d)\hat{\gamma}_d(\bar{y}_{sd} - \hat{\beta})]$$

where $\hat{\gamma}_d = n_d\hat{\varphi}(1 + n_d\hat{\varphi})^{-1}$.

3.3 Generalized Linear Mixed Models (GLMM)

The data underpinning small area estimates are often categorical and generalized linear models (GLMs), see Nelder and Wedderburn (1972), are a standard technique for analysing such data. In the case of small area estimation we use the Generalized Linear Mixed Model (GLMM) extension of the GLM to allow for correlation within the small areas of interest. In the GLMM $E(\mathbf{y}_s \mid \mathbf{u}) = h(\boldsymbol{\eta}_s)$ for a specified function h. Here $\mathbf{u}$ denotes the vector of random area effects and $\boldsymbol{\eta}_s = \mathbf{X}_s\boldsymbol{\beta} + \mathbf{Z}_s\mathbf{u}$. As usual, we assume $\mathbf{u}$ is normally distributed with zero mean vector and variance-covariance matrix Var($\mathbf{u}$) = $\boldsymbol{\Omega}$ = *blk-diag*(φ_j $\boldsymbol{\Omega}_j$). Let f($\mathbf{y}_s$;$\boldsymbol{\beta}$|$\mathbf{u}$) be the probability density function of $\mathbf{y}_s$ conditional on $\mathbf{u}$. The log-likelihood of $\mathbf{y}_s$ conditional on $\mathbf{u}$ is then $l_1 = \ln f(\mathbf{y}_s;\boldsymbol{\beta} \mid \mathbf{u})$, while the logarithm of the probability density function of $\mathbf{u}$ is

$$l_2 = -(1/2)\sum_{j=1}^{J}[\nu_j \ln \pi\varphi_j + \ln\left|\boldsymbol{\Omega}_j\right| + \varphi_j^{-1}\mathbf{u}_j\boldsymbol{\Omega}_j^{-1}\mathbf{u}_j].$$

where v_j is the rank of matrix $\mathbf{Z}_{sj}$. The function l_2 can be thought of as a penalty function, and the parameter values that maximise l_1+l_2 are sometimes referred to as penalised likelihood (PL) estimates.

Key references for development GLMM theory are Schall (1991), Solomon and Cox (1992), Wolfinger (1993), Wolfinger and O'Connell (1993), Breslow and Clayton (1993), McGilchrist (1994) and Engel and Keen (1994). Breslow and Lin (1995) and Lin and Breslow (1996) review the different estimation approaches introduced by these authors.

A typical example of discrete data is a binomial response. In this case the vector of small area counts $\mathbf{y}_s = [\mathbf{y}_{s1}, \mathbf{y}_{s2},...,\mathbf{y}_{sD}]'$ is assumed to follow a binomial distribution with log-likelihood conditional on the random components $\mathbf{u}$ given by

$$l_1 = \text{constant} + \sum_d [y_{sd} \ln\theta_d + (n_d - y_{sd}) \ln(1 - \theta_d)]$$

where y_{sd} is distributed as binomial with parameters n_d and θ_d. The probabilities θ_d are assumed to satisfy $\theta_d = \exp(\eta_d)/[1 + \exp(\eta_d)]$ where $\eta_d = \mathbf{X}_d'\boldsymbol{\beta} + \mathbf{Z}_d'\mathbf{u}$. This is usually called a mixed logistic regression model. The $\mathbf{X}_d$ and $\mathbf{Z}_d$ are vectors of area level auxiliary covariates.

For response data satisfying the mixed logistic model $\boldsymbol{\eta}_{sd} = (\beta + u_d)\mathbf{1}_{sd}$, an estimate of the proportion of successes in area d is then given by

$$\hat{\theta}_d = N_d^{-1}[n_d p_{sd} + (N_d - n_d)(\hat{\beta} + \hat{u}_d)]$$

where p_{sd} is the proportion of successes in the sample in small area d. Here $\hat{\beta}$ and $\hat{u}_d$ correspond to the estimate of β and the predicted value of u_d obtained from fitting the mixed logistic model to the sample data in the small areas of interest.

4 Estimation of Mean Square Error (MSE)

An important measure of the "quality" of a statistical estimator is its mean square error (MSE), and so it is important that any small area estimate be accompanied by an estimate of its MSE. In this section therefore we review methodology for estimating the MSE of predictors of $\tau = \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\mathbf{u})$. All of these predictors can be expressed in the form $\hat{\tau} = \mathbf{a}_r'(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}})$ for suitably chosen $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$.

To start, assume $\boldsymbol{\beta}$ and the variance components σ^2, $\boldsymbol{\varphi} = (\varphi_j)$ are specified, and so $\hat{\tau} = \mathbf{a}_r'(\mathbf{X}_r\boldsymbol{\beta} + \mathbf{Z}_r\hat{\mathbf{u}})$ where $\hat{\mathbf{u}}$ is the BLUP of $\mathbf{u}$ based on these specified values of the variance components. Then

$$MSE(\hat{\tau}) = \sigma^2\mathbf{a}_r'\mathbf{Z}_r\mathbf{T}^*\mathbf{Z}_r'\mathbf{a}_r = g_1(\boldsymbol{\varphi}) \tag{4.1}$$

where $\mathbf{T}^* = (\boldsymbol{\Omega}^{-1} + \mathbf{Z}_s\mathbf{W}_s^{-1}\mathbf{Z}_s')^{-1}$. For the simple model (2.2.7), this MSE becomes $MSE(\hat{\tau}) = (1 - f_d)^2\sigma^2 n_d^{-1}\gamma_d$ where $\gamma_d = \sigma_u^2(\sigma^2 n_d^{-1} + \sigma_u^2)^{-1}$, $f_d = n_d N_d^{-1}$ and $(1 - f_d)$ is the finite population correction factor.

Now suppose $\boldsymbol{\beta}$ is replaced by its WLS estimator $\hat{\boldsymbol{\beta}}$, but the variance components σ^2, $\boldsymbol{\varphi}$ = (φ_j) are still assumed known. In this case $\hat{\tau} = \mathbf{a}_r'(\mathbf{X}_r\hat{\boldsymbol{\beta}} + \mathbf{Z}_r\hat{\mathbf{u}})$ is the BLUP of τ, with $\hat{\mathbf{u}}$ defined as before, and its MSE is

$$\begin{aligned} MSE(\hat{\tau}) &= \begin{pmatrix}\mathbf{a}_r'\mathbf{X}_r & \mathbf{a}_r'\mathbf{Z}_r\end{pmatrix} E\left(\begin{bmatrix}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ (\hat{\mathbf{u}} - \mathbf{u})\end{bmatrix}\begin{bmatrix}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ (\hat{\mathbf{u}} - \mathbf{u})\end{bmatrix}'\right)\begin{pmatrix}\mathbf{X}_r'\mathbf{a}_r \\ \mathbf{Z}_r'\mathbf{a}_r\end{pmatrix} \\ &= \sigma^2[\mathbf{a}_r'\mathbf{X}_r - \mathbf{a}_r'\mathbf{Z}_r\mathbf{T}^*\mathbf{Z}_s'\mathbf{W}_s^{-1}\mathbf{X}_s](\mathbf{X}_s\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}[\mathbf{X}_r'\mathbf{a}_r - \mathbf{X}_s'\mathbf{W}_s^{-1}\mathbf{Z}_s\mathbf{T}^*\mathbf{Z}_r'\mathbf{a}_r] \\ &\quad + \sigma^2\mathbf{a}_r'\mathbf{Z}_r\mathbf{T}^*\mathbf{Z}_r'\mathbf{a}_r \\ &= g_2(\boldsymbol{\varphi}) + g_1(\boldsymbol{\varphi}) \end{aligned} \tag{4.2}$$

where $g_1(\boldsymbol{\varphi})$ was defined in (4.1). Note that under the simple model (2.2.7)

$$g_2(\boldsymbol{\varphi}) = \sigma^2 n^{-1}(1-f_d)^2(1-\gamma_d)^2\left[1-n^{-1}\sum_{d=1}^{D} n_d\gamma_d\right]^{-1}.$$

Finally, we consider the EBLUP situation, i.e. where both $\boldsymbol{\beta}$ and the variance components are replaced by estimators. For estimated variance components the MSE of the EBLUP is

$$\begin{aligned} MSE(\hat{\tau}_{EBLUP}) &= E(\hat{\tau}_{EBLUP} - \hat{\tau}_{BLUP} + \hat{\tau}_{BLUP} - \tau)^2 \\ &= MSE(\hat{\tau}_{BLUP}) + E(\hat{\tau}_{EBLUP} - \hat{\tau}_{BLUP})^2 \\ &\quad + 2E(\hat{\tau}_{BLUP} - \tau)(\hat{\tau}_{EBLUP} - \hat{\tau}_{BLUP}). \end{aligned} \tag{4.3}$$

The first term on the right hand side of (4.3) is given by equation (4.2). A naive estimator of the MSE of the EBLUP is therefore defined by disregarding last two terms on the right hand side of (4.3) and replacing the unknown variance components in (4.2) by suitable estimators. However, Kackar and Harville (1984) show that although the third component on the right hand side of (4.3) is neglible, the second is not. They develop an approximation to the second term on the right hand side of (4.3), allowing them to write

$$MSE(\hat{\tau}_{EBLUP}) \cong MSE(\hat{\tau}_{BLUP}) + tr\left[Var\left(\partial\hat{\tau}_{BLUP}/\partial\boldsymbol{\varphi}\right)\mathbf{B}(\boldsymbol{\varphi})\right] \tag{4.4}$$

where $\mathbf{B}(\boldsymbol{\varphi})$ is the MSE matrix $E[(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi})']$. Since

$$\partial\hat{\tau}_{BLUP}/\partial\boldsymbol{\varphi} = \mathbf{a}_r'\boldsymbol{\Delta}(\partial\mathbf{A}/\partial\boldsymbol{\varphi})p(\boldsymbol{\varphi})\mathbf{y}_s$$

where

$$p(\boldsymbol{\varphi}) = \boldsymbol{\Sigma}_s^{-1} - \boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}$$

and

$$\boldsymbol{\Delta} = (\mathbf{Z}_r\mathbf{T}^*\mathbf{Z}_s' - \mathbf{X}_r)(\mathbf{X}_s'\boldsymbol{\Sigma}_s^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s'\mathbf{Z}_s\mathbf{T}^*\boldsymbol{\Omega}^{-1} + \mathbf{Z}_r\mathbf{T}^*\boldsymbol{\Omega}^{-1}$$

it follows

$$Var\left(\partial\hat{\tau}_{BLUP}/\partial\boldsymbol{\varphi}\right) = \mathbf{a}_r'\boldsymbol{\Delta}(\partial\boldsymbol{\Omega}/\partial\boldsymbol{\varphi})\mathbf{Z}_s'p(\boldsymbol{\varphi})\mathbf{Z}_s(\partial\boldsymbol{\Omega}/\partial\boldsymbol{\varphi})'\boldsymbol{\Delta}'\mathbf{a}_r.$$

The matrix $\mathbf{B}(\boldsymbol{\varphi})$ on the other hand can be approximated by the asymptotic variance-covariance matrix of the ML/REML estimators of the variance components.

Prasad and Rao (1990) investigated the application of (4.4) to small area estimation. Their study showed that the MSE estimator based on a "plug-in" approximation of (4.4) underestimates the true MSE. They therefore introduced a second order approximation to the left hand side of (4.3) of the form

$$\begin{aligned} MSE(\hat{\tau}_{EBLUP}) &\cong MSE(\hat{\tau}_{BLUP}) + \sigma^2 tr\left[(\partial\mathbf{b}/\partial\boldsymbol{\varphi})\boldsymbol{\Sigma}_s(\partial\mathbf{b}/\partial\boldsymbol{\varphi})'\mathbf{B}(\boldsymbol{\varphi})\right] \\ &= g_1(\boldsymbol{\varphi}) + g_2(\boldsymbol{\varphi}) + g_3(\boldsymbol{\varphi}) \end{aligned} \tag{4.5}$$

where $\mathbf{b} = \mathbf{a}_r'\mathbf{Z}_r\boldsymbol{\Omega}\mathbf{Z}_s'\boldsymbol{\Sigma}_s^{-1}$. Their paper also contains simplifications of this estimator for the special cases of nested error regression, random coefficient regression and Fay-Herriot type small area models. Alternative approximations to (4.3) are given in Harville and Jeske (1992), while Hulting and Harville (1991) use (4.5) to obtain interval predictors.

For the simple model (2.2.7), the first two components of the MSE of the EBLUP have been derived above. If the variance components are estimated via ML then

$$g_3(\varphi) = 2n\sigma^2 n_d^{-2}(1-f_d)^2(n_d^{-1}+\varphi)^{-3}[n\sum\nolimits_d (n_d^{-1}+\varphi)^2 - (\sum\nolimits_d n_d^{-1}+\varphi)^2]^{-1}.$$

On the other hand, if the variance components are estimated via REML, then

$$g_3(\varphi) = 2n\sigma^2 n_d^{-2}(n_d^{-1}+\varphi)^{-3}[na_{22} - a_{12}]^{-1}$$

where

$$a_{12} = \sum\nolimits_d (n_d^{-1}+\varphi)^{-1} + \sum\nolimits_d \Delta_d$$

$$a_{22} = \sum\nolimits_d (n_d^{-1}+\varphi)^{-1} + \left(\varphi^{-1} - \sum\nolimits_d \Delta_d\right)^2 + \varphi^{-1} - 2\sum\nolimits_d \Delta_d (1+n_d\varphi)^{-1}$$

and

$$\Delta_d = (n_d^{-1}+\varphi)^{-2} / \sum\nolimits_j (n_j^{-1}+\varphi)^{-1}.$$

Pfeffermann and Barnard (1991) use (4.5) to compare the performance of different small area models. Rao and Yu (1994) develop (4.5) for the situation where the small area estimators are based on a model that includes both time series and cross-sectional effects. Lahiri and Rao (1995) show that (4.5) is robust to nonormality of small area effects for the Fay-Herriot (1979) model. The performance of (4.5) under the nested regression model is studied in Singh et al (1998). Moura and Holt (1999) develop (4.5) for small area estimators based on multilevel models, while Rao (1999) and Pfeffermann (1999) review the application of MSE estimators based on (4.5). Other papers that are based on (4.5) are Prasad and Rao (1999) and Stukel and Rao (1999).

Finally, we note that Booth and Hobert (1998) propose conditional mean square error (CMSEP) as a general measure of prediction variance for linear combinations of the fixed and random effects in generalized linear mixed models, where the conditioning is with respect to the random effects in the model. Rivest and Belmonte (2000) adapt this approach to estimation of mean square error for small area estimators.

Bibliography

Abbott, O., Brown, J., Chambers, R. and Cruddas, M. (2000). Small area population estimation in the 2001 UK One Number Census. *One Number Census Steering Committee Paper 00/03B*, Office for National Statistics, UK.

Airy, G.B. (1861). On the Algebrical and numerical Theory of Errors of Observations and the Combinations of Observations. *London: MacMillan*.

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

Bell, W.R. and Hillmer, S.C. (1990). The time series approach to estimation for repeated surveys. *Survey Methodology* **16**, 29-45.

Binder, D.A. and Dick, J.P (1989). Modelling and estimation for repeated surveys. *Survey Methodology* **15**, 29-45.

Binder, D.A. and Hidiroglou, M.A. (1986). Sampling in time. In Handbook of Statistics, Vol 6, (Eds, Krishnaiha P.R. and Rao, C.R.), Amesterdam: Elsevier Science, 187-211.

Booth, J.G and Hobert, J.P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 262-272

Breslow, N.E. and Clayton, D.G. (1993). Approximation inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**, 81-92.

Datta, G.S., Day, B. and Basawa, I. (1999). Empirical best linear unbiased and empirical Bayes prediction in multivariate small area estimation. *Journal of Statistical Planning and Inference* **75**, 269-279.

Datta, G.S. and Ghosh, M. (1991). Bayesain prediction in linear models: application to small area estimation. *The Annals of Statistics* **19**, 1748-1770.

Datta, G.S. and Ghosh, M., Nangia, N. and Natarajan, K. (1996). Estimation of median income of four-person families: a Bayesian approach. In: Berry, D.A., Chaloner, K.M., Geweke, J.K. (Eds), Bayesian Analysis in Statistics and Econometrics. Wiley, New York, pp. 129-140

Datta, G.S. and Lahiri, P. (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis* **54**, 310-328.

Datta, G.S., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment artes for the States of the U.S. *Journal of the American Statistical Association* **94**, 1074 1082.

Deely, J.J and Lindley, D.V. (1981). Bayes empirical Bayes. *Journal of the American Statistical Association* **76**, 833-841.

Ericksen, E.P. (1974). A regression method for estimating population changes of local areas. *Journal of the American Statistical Association* **69**, 867-875.

Engel, B. and Kenn, A. (1994). A simple approach for analysis of generalized linear mixed models. *Statistica Neerlandica*, **49**, 1-22.

Ericksen, E.P. (1975). A method for combining sample survey data and symptomatic indictors to obtain population estimate for local areas. *Demography* **10**, 137-160.

Ericksen, E.P. and Kadane, J. (1985). Estimating the population in a census year. *Journal of the American Statistical Association* **80**, 98-109.

Farrell, P.J. (1997). Empirical Bayes estimation of small area proportions based on ordinal outcome variables. *Survey Methodology* **23**, 119-126.

Farrell, P.J., MacGibbon, B. and Tomberlin, T. J. (1997). Empirical Bayes small-area estimation using logistic regression models and summery statistics. *Journal of Business & Economic Statistics* **15**, 101-108.

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.

Fuller, W.A. and Battese, G.E. (1973). Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association* **68**, 626-632.

Fuller, W.A and Harter, R.M. (1987). The multivariate components of variance model for small area estimation. In: Platek, R. Rao, J.N.K., Sarndal, C.E., Sigh, M.P. (ed.), Small Area Statistics. Wiley, New York 103-123.

Ghosh, M. and Rao, J.N.k. (1994). Small area estimation: an appraisal. *Statistical Sciences* **9**, 55-93.

Ghosh, M., Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: a Bayesian time series approach. *Journal of the American Statistical Association* **91**, 1423-1431.

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association* **93**, 273-282.

Gonzalez, M.E. and Hoza, C. (1978). Small-area estimation with application to unemployment and housing estimates. *Journal of the American Statistical Association* **73**, 7-15.

Gonzalez, M.E. and Waksberg, J. (1973). Estimation of the error of synthetic estimates. Presented at first meeting of the International Association of Survey Statisticians in Vienna Austria on August 18-25.

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

Harville, D.A. (1990). Discussion on Robinson paper, That BLUP is a good thing: the estimation of random effects. *Statistical Science*. **6**, 15-51.

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

Henderson, C.R. (1948). Estimation of general, specific and maternal combining abilities in crosses among inbred lines of swine. Ph.D. Thesis, Iowa State University, Ames, Iowa.

Henderson, C.R. (1949). Estimation of changes in herd environment (Abstract). *J. Dairy Sci*. 32:706.

Henderson, C.R. (1950). Estimation genetics parameters (Abstract). *The Annals of Mathematical Statistics* **21**, 309-310

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-252.

Henderson, C.R. (1963). Selection index and expected genetic advance. In Statistical Genetics and Plant Breeding (W.D. Hanson and H.F. Robinson, eds.), 141-163. National Academy of Sciences and National Research Council Publication No. 982, Washington, D.C.

Henderson, C.R. (1973). Sire evaluation and genetic trends. *Proceedings of the Animal Breeding and Genetics Symposium in Honour of Dr. Jay L. Lush* 10-41. Amer. Soc. Animal Sci.-Amer. Dairy Sci. Assn.-Poultry Sci. Assn. Champaigne, Illinois.

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

Henderson, C.R., Kempthorne, O., Searle, S.R., and von Krosigk, C.M. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192-218.

Hidiroglou, M.A. and Sorndal, C.E. (1985). An empirical study of some regression estimators for small domains. *Survey Methodology*, **11**, 65-77.

Holt, D., Smith, T.M.F. and Tomberlin (1979). A model-based approach to estimation for small subgroups of a population. *Journal of the American Statistical Association* **74**, 405-410.

Hulting, F.L. and Harville, D.A. (1991). Some Bayesian and non-Bayesian procedures for the analysis of comparative experiments and for small-area estimation: computational aspects, frequentist properties, and relationships. *Journal of the American Statistical Association* **86**, 557-567

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of Royal Statistical Society Series B*, **42**, 221-226

Kacker, R.N. and Harville, D.A. (1981). Unbiased of two-stage estimation and prediction procedure for mixed linear models. *Communications in Statistics- Theory and Methods* **A10**, 1249-1261.

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

Kass, R.E. and Steffey, D. (1989). Approximate Bayesian inference in conditionally independent hierarchical models (parametric empirical Bayes models). *Journal of the American Statistical Association* **84**, 717-726.

Kott, P. (1989). Robust small domain estimating using random effects modelling. *Survey Methodology* **15**, 3-12.

Laake, P. (1979). A predictive approach to subdomain estimation in finite populations. *Journal of the American Statistical Association* **74**, 355-358.

Legendre, E.L. (1806) *Nouvalles Méthodes pou la Determination des Obites Cométes: avec un Supplément Contenant Divers Perfectionnements de ces Méthodes leur Application aux deux Cométes de* 1805. Coucier, Paris.

Levy, P.S. (1971). The use of mortality data in evaluating synthetic estimates. *Proceedings of the American Statistical Association* 328-331.

Levy, P.S., and French, D.K. (1977) Synthetic estimation of state health characteristics based on the health interview survey. *Vital and Health Statistics: series 2, No. 75*, DHEW publication (PHS) 78-1349. Washington: U.S Government Printing Office.

Li, B. (2001). Robust prediction from linear mixed-effects models with applications in small area estimation. PhD thesis, University of California at Davis.

Longford, N.T. (1999). Multivariate shrinkage estimation of small area means and proportions. *Journal of Royal Statistical Society Series A* **162**, 227-245.

Lahiri, P. and Rao, J.N.K. (1995). Robust estimation of means squared error of small area estimators. *Journal of the American Statistical Association* **90**, 758-766

Laird, N.M. and Louis, T.A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association* **82**, 739-750

Lin, X. and Breslow, N.E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007-1016.

MacGibbon, B. and Tomberlin, T.J. (1989). Small area estimates of proportions via empirical Bayes techniques. *Survey Methodology* **15**, 237-252.

Madow, W.J. and Hansen, M.H. (1975). On statistical models and estimation in sample surveys. Contributed paper to 40th session of the International Statistical Institute, Warsaw, Poland, 554-557.

Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *J. Statistical Planning and Inference* **69**, 339-348.

McCulloch, C.E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89**, 330-335.

McCulloch, C.E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assn.* 92, 162-170.

McGilchrist, C.A. (1994). Estimation in generalized mixed models. *Journal of the Royal Statistical Society* **B**, **56**, 61-69.

McGilchrist, C.A. and Aisbett, C.W. (1991). Restricted BLUP for mixed linear models. *BiometricalJournal*. **33**, 131-141.

Malec, D., Sedransk, J., Moriarity, C.L. and Leclere, F.B (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association* **92**, 815-826.

Malec, D., Davis, W.W and Cao, X. (1999). Model-based small area estimates of overweight prevalence using sample selection adjustment. *Statistics in Medicine* **18**, 3189-3200.

Marker, D. (1999). Organization of small area estimators using a generalized linear regression framework. *Journal of Official Statistics*. **15**, 1-24.

Morris, C.N. (1983). Parametric empirical Bayes inference theory and applications. *Journal of the American Statistical Association* **78**, 47-65

Moura, F.A.S. (1994). Small area estimation using multilevel models. PhD. Thesis, Southampton University, UK.

Moura, F.A.S. and Holt, D. (1999). Small area estimation using multilevel models. *Survey Methodology* **25**, 73-80.

Moura, F.A.S. and Mignon, H.S. (2001). Bayesian models for small area proportions. *Working Paper 140*, Universidade Federal do Rio De Janeiro, Instituto De Matematica, Departmento de Metodos Estatisticos.

Namekata, T., Levy, P.S. and O'Rourke, T.W. (1975). Synthetic estimates of work loss disability for each State and District of Columbia. *Public Health Reports* **90**, 532-538.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A*, **135**, 370-384.

Patterson, H.D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business & Economics Statistics* **9**, 163-175.

Pfeffermann, D. (1999). Small area estimation - Big developments. *Keynote Paper, Conference on Small Area Statistics*, Riga, Latvia, August 1999.

Pfeffermann, D and Bernard, C. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economics Statistics* **9**, 73-84.

Pfeffermann, D. and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data. *Survey Methodology* **16**, 217-237.

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.

Prasad, N.G.N and Rao, J.N.K. (1999). On robust small area estimating using a simple random effects model. *Survey Methodology* **25**, 67-72.

Purcell, N.J and Kish, L. (1980). Postcensal estimates for local areas (domains). *International Statistical Review* **48**, 3-18.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, **25** 175-186.

Rao, J.N.K. (2001). Small area estimation with application to agriculture. Presented at *Conference in Celebration of Wayne A. Fuller's 70th Birthday*, Iowa State University, Ames, Iowa, June 2001.

Rao, J.N.K. and Yu, M. (1994) Small-area estimation by combining time-series and cross-sectional data. *The Canadian Journal of Statistics* **22**, 511-528.

Rivest, L.P. and Belmonte, E. (2000) A conditional mean squared error of small area estimators. *Survey Methodology*, **26**, 67-78.

Rao, J.N.K. and Yu, M. (1992). Small area estimation combining time series and cross-sectional data. *Proceeding Survey Research Methods Section. American Statistical association*, 1-9.

Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science*. **6**, 15-51.

Royall, R.M. (1976). The linear least-squares prediction approach to two-stage sampling. *Journal of the American Statistical Association* **71**, 657-664.

Sarndal, C. E. (1984). Design-consistent versus model-dependent estimation for small domains. *Journal of the American Statistical Association* **79**, 624-631.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. Wiley and Sons, Inc.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

Schaible, W.L., Brock, D.B. and Schank, G.A. (1977). An empirical comparison of simple inflation, synthetic and composite estimators for small area statistics. *Proceedings of the American Statistical Association* 1017-1021.

Scott, A. J. and Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods. *Journal of the American Statistical Association* **69**, 674-678.

Scott, A. J., Smith, T.M.F. and Jones, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review* **45**, 13-28.

Singh, A.C. and Mantel (1991). State space composite estimation for small areas. *Proceedings: Symposium 91: Spatial Issues in Statistics*, Statistics Canada, Ottawa.

Singh, A.C., Mantel, H.J. and Thomas, B.W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, **20**, 33-43.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society Series B* **60**, 377-396.

Smith, T.M.F (1987). Principles and problems in the analysis of repeated surveys. In *Survey Sampling and Measurement*, Ed. N.K. Namboodiri, New York: Academic Press, 201-216.

Solomon, P.J. and Cox, D.R. (1992). Nonlinear components of variance models. *Biometrika* **79**, 1-11.

Stukel, D.M. and Rao, J.N.K. (1999). On small-area estimation under two-fold nested error regression models. *Journal of Statistical Planning and Inference* **78**, 131-147.

Tiller, R.B. (1991). Times series modelling of sample survey data from the U.S. Current Population Survey. *Journal of Official Statistics* **8**, 149-166.

Valliant, R., Dortman, A.H. and Royall R.M. (2000). *Finite Population Sampling and Inference.* New York: John Wiley.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* **80**, 791-795.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233-243.

You, Y. and Rao, J.N.K. (2000). Hierarchical Bayes estimation of small area means using multi-level models. *Survey Methodology*, **26** 173-181.

Zeger, S.L. and Karim, M.R. (1991). Generalized linear models with random effect; A Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79-86.