Estimation of Randomization Mean Square Error in Small Area Estimation

Danny Pfeffermann[(1),(2),(3)] and Dano Ben-hur[(1)]

Abstract

In this article we propose a new method for estimating the randomization (design-based) mean squared error (DMSE) of model-dependent small area predictors. Analogously to classical survey sampling theory, the DMSE considers the finite population values as fixed numbers and accounts for the MSE of small area predictors over all possible sample selections. The proposed method models the true DMSE as computed for synthetic populations and samples drawn from them, as a function of known statistics, and then applies the model to the original sample. Several simulation studies for the linear area level model of Fay and Herriot (1979) and the unit-level mixed logistic model of MacGibbon and Tomberlin (1989) illustrate the performance of the proposed method and compare it to the performance of other DMSE estimators proposed in the literature.

Key words: Area-level model, Design MSE, Mixed logistic model, Model-based MSE

(1)- Central Bureau of Statistics, Israel
(2)- Hebrew University of Jerusalem, Israel
(3)- University of Southampton, United Kingdom

1. Introduction

The term small area estimation (SAE) encompasses a set of statistical procedures for estimating area parameters such as totals, proportions or even distribution functions, for areas or domains for which only small samples, and in some cases no samples are available. A direct estimator for a target parameter, based only on observations from that area can be very inaccurate as a result of the small sample size. Consequently, over the last 4 decades indirect model-based predictors, which borrow information across areas or over time, have been developed. The models typically contain random effects, which are aimed to account for the unexplained variability of the true area parameters or individual observations not accounted for by known covariates.

As with any other problem in statistics, the production of an estimator is only part of the inference, and one needs to provide also a measure of its reliability. A common measure is an estimate of the mean squared error (MSE), which in the case of SAE is required for every area separately. Model-based prediction MSEs (PMSE) of small area predictors account for all sources of variation, including the distribution of the hypothetical random effects included in the model. This implies that the target area parameters are considered as random, which is different from classical survey sampling applications under which the finite population values and hence the target parameters are viewed as fixed values. Users of sample survey (official statistics) estimates are familiar with measures of error, which only account for the variability originating from the randomness of the sample selection (known as the randomization distribution). In other words, these users are accustomed to estimates of the design-based (randomization) MSE (denoted hereafter as DMSE), over all possible sample selections, with the population values of the survey variables (and hence the values of the target parameters), held fixed. Estimation and publication of the DMSE (or its square root) is a common routine in national statistical offices all over the world.

In this article we propose a new procedure for estimation of the DMSE in SAE, with special attention to the area-level model (Fay and Herriot, 1979), and the unit-level mixed logistic model (MacGibbon and Tomberlin, 1989). The procedure consists of modelling the true DMSE computed for synthetic populations and samples generated from the underlying model as a function of known statistics, and then applying the model to the original sample. It relies in part on the procedure for bias correction proposed in Pfeffermann and Correa (2012), with appropriate modifications. We illustrate the performance of the proposed method and compare it to other DMSE estimators proposed in the literature using simulated data.

2. Literature review

2.1 Model-based estimation of prediction MSE

Denote the true target parameter in area i by θ_i (hypothetical parameter like expectation or percentile, or its finite population counterpart like the realized area average). Denote by $\hat{\theta}_i$ its (model-based) predictor. The model-based PMSE is defined as,

$$PMSE(\hat{\theta}_i) = E_M[(\hat{\theta}_i - \theta_i)^2], \tag{2.1}$$

where the subscript M signifies that the expectation accounts for all sources of variation, including the distribution of the random effects included in the model.

Estimators of the PMSE for the area-level estimators of Fay and Herriot (1979) and the unit level estimators of Battese et al. (1988), were developed by Prasad and Rao (PR, 1990), for the case where the variance of the random effects is estimated by the Method of Moments. Datta and Lahiri (2000) extend the procedure of PR to more general mixed linear models and for the case where the unknown variance components are estimated by maximum likelihood estimators (MLE). Other extensions have been proposed by Datta et al. (2005) and Das et al. (2004). All the above MSE estimators have bias of desired order $o(1/m)$, where m is the number of sampled areas.

Resampling methods for estimation of model-based PMSE with the same order of bias have been proposed by Jiang et al. (2002) and Lohr and Rao (2009), based on the jackknife method, and by Hall and Maiti (2006) based on parametric bootstrap. (Hall and Maiti (2006) only prove second-order unbiasedness for the basic bias corrected estimator but not for the tuned estimators that are strictly positive. See Section 3 for more details.)

Pfeffermann and Correa (2012) developed a method for bias correction, which models the error of a given predictor as a function of the corresponding estimators obtained from bootstrap samples and the original estimator and bootstrap estimators of the parameters governing the model fitted to the sample data. The method is applied for estimating the PMSE of the empirical best predictors under the mixed logistic model. Our proposed procedure for DMSE estimation is based in part on this method. The advantage of resampling methods over parametric estimators is that they are applicable for more general mixed models, although they are not fully nonparametric because they require computing the model-dependent estimators for each new sample.

A third approach for assessing the prediction error under the model is to follow the Bayesian paradigm, in which case the predictor is commonly defined by the posterior mean and its PMSE is estimated by the posterior variance. For more details of these and other methods of estimation of model-based PMSE, see Pfeffermann (2013) and the book of Rao and Molina (2015).

2.2 Estimation of Design-based MSE of model-based predictors

The methods mentioned in Section 2.1 are model dependent, attempting to estimate the PMSE by accounting for all sources of error. As mentioned in the introduction, an alternative approach, which is more appealing to users of small area estimators, is to estimate the MSE with respect to randomization distribution over all possible sample selections, but with the true population values held fixed. The problem with this approach in the context of SAE is that estimation of the design-based MSE (DMSE) with acceptable level of accuracy is very complicated because of the small sample sizes in some or all the areas. Consequently, it was recommended in the past to average the MSE estimators over many small areas to get a stable estimator. See Rao and Molina (2015, section 3.2.5) for several averaging procedures applicable when estimating the DMSE of synthetic estimators. However, as already stated, a small area predictor and an estimator of its DMSE is required for every area separately, and not just as an average over many areas. Only few attempts to tackle this problem have been reported in the literature.

Following the previous notation, the DMSE is defined as,

$$\mathrm{DMSE}(\hat{\theta}_i) = E_D[(\hat{\theta}_i - \theta_i)^2 \mid F], \tag{2.2}$$

where $F = F_1 \cup ... \cup F_m$ is the collection of all the vector values of the survey variables under consideration in the finite population, which in the present context is classified into the m areas of interest, $\theta_i = g(F_i)$ is a function of F_i corresponding to area i like the true area mean or proportion of the target variable of interest, and $\hat{\theta}_i$ is an estimator of θ_i. When $\hat{\theta}_i$ is computed based only on the sample of values from the area, it is referred to in the SAE literature as a direct estimator. Model dependent estimators use also data from other areas and are referred to as indirect estimators. The expectation operator E_D is with respect to the randomization distribution over all possible sample selections from F. In what follows we review the approaches proposed in the literature for DMSE estimation.

2.2.1. *Area level models*

Rivest and Belmonte (2000) consider area level models under which the available data consist of a set of direct estimators, $y = (y_1,...,y_m)'$. The authors assume that $y \mid \theta \sim N_m(\theta, \Sigma)$, the m-variate normal distribution with mean vector θ and known covariance matrix Σ. In order to improve the accuracy of the direct estimators, the authors consider shrinkage-based estimators of the form,

$$\hat{\theta}_i = y_i + g_i(y_1,...,y_m),\ i = 1,...,m. \tag{2.3}$$

Let $g(y)^t = [g_1(y),....,g_m(y)]$. For known hyper-parameter values featuring in $g(y)^t$, the authors propose estimating the DMSE unbiasedly as,

$$\hat{DMSE}(\hat{\theta}) = \Sigma + \Sigma \nabla g(y) + \nabla g(y)^t \Sigma + g(y)g(y)^t, \tag{2.4}$$

where $\nabla g(y)$ is the $m \times m$ matrix with the $(i,j)^{th}$ element given by $g_{ij}(y) = \partial g_i(y) / \partial y_j$. The derivation of (2.4) relies on the normality of the direct estimators given the true area means. Consider, for example, the Fay Herriot (FH, 1979) model,

$$y_i = \theta_i + e_i;\ \theta_i = x_i'\beta + u_i;\ u_i \sim N(0,\sigma_u^2),\ e_i \sim N(0,\sigma_{D_i}^2), \tag{2.5}$$

where y_i denotes the direct estimator for area i, e_i is the sampling error and u_i the corresponding random effect. For known parameter values, the F-H estimator is,

$$\hat{\theta}_i^{FH} = \gamma_i y_i + (1-\gamma_i) x_i'\beta;\quad \gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_{D_i}^2}. \tag{2.6}$$

Clearly, the F-H estimator is a special case of the estimators considered by Rivest and Belmonte (2000) and the DMSE estimator takes in this case the form,

$$\hat{DMSE}(\hat{\theta}_i^{FH}) = \gamma_i \sigma_{D_i}^2 + (1-\gamma_i)^2 [(y_i - x_i'\beta)^2 - (\sigma_{D_i}^2 + \sigma_u^2)]. \tag{2.7}$$

Notice that unlike (2.4), the derivation of (2.7) does not require normality of $y_i \mid \theta_i$. The estimator (2.7) is an unbiased estimator of the DMSE with respect to the randomization distribution, but it is very variable and can take negative values. This would particularly be the case if the sampling variance, $\sigma_{D_i}^2$, is large, the basic problem of SAE.

Rao et al. (2018) propose estimating the DMSE of the empirical FH estimator as a weighted average of a design-unbiased DMSE estimator and an estimator of the (unconditional) model-dependent PMSE of correct order (See Section 1). For the F-H model, the weighting coefficients are $\hat{\gamma}_i$ and $(1-\hat{\gamma}_i)$, where $\hat{\gamma}_i = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \sigma_{D_i}^2)^{-1}$. The resulting, (composite) DMSE estimator is thus,

$$\hat{DMSE}_1(\hat{\theta}_i^{FH}) = \hat{\gamma}_i \hat{DMSE}(\hat{\theta}_i^{FH}) + (1-\hat{\gamma}_i)\hat{PMSE}(\hat{\theta}_i^{FH}), \quad (2.8)$$

where $\hat{DMSE}(\hat{\theta}_i^{FH})$ is the design unbiased estimator for the case where the model parameters are estimated by restricted maximum likelihood (REML); see Datta et el. (2011) for derivation of the unbiased estimator. The rationale of using an estimator of the model PMSE to estimate the DMSE is that under the model and for known hyper-parameter values, $E_{\theta_i}[DMSE(\hat{\theta}_i^{FH} \mid \theta_i)] = \sigma_{D_i}^2 \gamma_i$ $= PMSE(\hat{\theta}_i^{FH})$, so one estimates the DMSE by an estimator of its model expectation. For small estimates $\hat{\gamma}_i$ (small area sample size), Rao et el. (2018) propose replacing in (2.8) $\hat{\gamma}_i$ by $\sqrt{\hat{\gamma}_i}$ and $(1-\hat{\gamma}_i)$ by $(1-\sqrt{\hat{\gamma}_i})$. The resulting DMSE estimator is then,

$$\hat{DMSE}_2(\hat{\theta}_i^{FH}) = \sqrt{\hat{\gamma}_i}\, \hat{DMSE}(\hat{\theta}_i^{FH}) + (1-\sqrt{\hat{\gamma}_i})\hat{PMSE}(\hat{\theta}_i^{FH}). \quad (2.9)$$

The estimator (2.9) gives more weight to $\hat{DMSE}(\hat{\theta}_i^{FH})$ than (2.8), thus reducing the overall design bias, but at the expense of possibly increasing the design MSE.

The estimators in (2.8) and (2.9) can take negative values because $\hat{DMSE}(\hat{\theta}_i^{FH})$ can be negative. Thus, a third modification considered by the authors is to replace the estimators by the model-dependent PMSE estimator when they take negative values.

Brakel et al. (2016) propose a design-based variance (DVAR) estimator of the empirical FH predictor for the case where the auxiliary variables x_i are also estimated from another survey. The empirical predictor they consider is,

$$\hat{\theta}_{i,\hat{x}}^{FH} = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i)\hat{x}_i'\hat{\beta}_{G\hat{L}S};\ \hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \sigma_{D_i}^2}, \quad (2.10)$$

where the estimates $\hat{\sigma}_u^2$ and hence $\hat{\beta}_{G\hat{L}S}$ are obtained under the model. (The sampling error variances, $\sigma_{D_i}^2$, are estimated externally and assumed known, a common practice when using the FH model). The authors develop a first order Taylor approximation for the design variance of $\hat{x}_i'\hat{\beta}_{G\hat{L}S}$,

where $\hat{\beta}_{G\hat{L}S} = (\sum_{i=1}^{m} \hat{\gamma}_i \hat{x}_i \hat{x}_i')^{-1} \sum_{i=1}^{m} \hat{\gamma}_i \hat{x}_i y_i = \hat{T}^{-1} t$ is the empirical GLS estimator, and consequently, the following estimator for $DVAR(\hat{\theta}_{i,\hat{x}}^{FH})$,

$$\begin{aligned} D\hat{V}AR(\hat{\theta}_{i,\hat{x}}^{FH}) = \hat{\gamma}_i^2 \, \hat{\text{var}}(y_i) + (1-\hat{\gamma}_i)^2 \left\{ \sum_{j=1}^{m} \hat{B}_{i,j} \, \hat{\text{cov}}(\hat{x}_j) \hat{B}_{i,j}^T + \sum_{j=1}^{m} \hat{C}_{i,j}^2 \, \hat{\text{var}}(y_i) \right\} \\ + 2\hat{\gamma}_i (1-\hat{\gamma}_i) \hat{C}_{i,i} \, \hat{\text{var}}(y_i), \end{aligned} \tag{2.11}$$

where $\hat{B}_{i,j} = (\delta_{i,j} - \hat{\gamma}_j \hat{x}_i^T \hat{T}^{-1} \hat{x}_j) \hat{\beta}_{G\hat{L}S} + \hat{\gamma}_j (y_j - \hat{x}_j^T \hat{\beta}_{G\hat{L}S}) \hat{x}_i' \hat{T}^{-1}$, $\delta_{ij} = 1(0)$ if $i = j (i \neq j)$, and $\hat{C}_{i,j} = \hat{x}_i' \hat{T}^{-1} \hat{x}_j \hat{\gamma}_j$.

The estimator (2.11) conditions on the estimator $\hat{\sigma}_u^2$ of the random effects as obtained from the original sample and hence conditions on the estimates $\hat{\gamma}_i$, but it accounts for the sampling errors of the direct estimator and of the estimates $\hat{x}_j$. Notice, however, that the estimator does not account for the design-bias of the small area predictor and hence it is not considered in the simulation study in Section 4.

2.2.2. *Unit level models*

Unit level models are applicable for the case where individual observations, $y_{ij}, x_{ij}, j = 1,...,n_i$, are available for every sampled area $i = 1,...,m$. Molina and Strzalkowska-Kominiak (2017) consider the case of a binary response, $y_{ij} \in \{0,1\}$ and assume the generalized linear mixed model (GLMM),

$$y_{ij} \mid p_{ij} \sim \text{Bernoulli}(p_{ij}); \; \text{logit}(p_{ij}) = x_{ij}'\beta + u_i, \; u_i \sim N(0, \sigma_u^2). \tag{2.12}$$

The target parameters are the true area proportions,

$P_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij} = \frac{1}{N_i} (\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} y_{ij}), \; i = 1,...,m$, where s_i and $\bar{s}_i$ define respectively the sample and out of sample units in area i and N_i is the area size. Several model-based predictors of the form,

$$\hat{P}_i = \frac{1}{N_i} (\sum_{j \in S_i} y_{ij} + \sum_{j \in \bar{S}_i} \hat{y}_{ij}), \; i = 1,...,m \tag{2.13}$$

are considered, with the predictors $\hat{y}_{ij}$ obtained under alternative model approximations. However, for DMSE estimation, the paper restricts to the case where the model fitted is the basic unit level model of Battese et al. (1988),

$$y_{ij} = x'_{ij}\beta + u_i + \varepsilon_{ij};\ u_i \sim N(0, \sigma_u^2),\ \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \tag{2.14}$$

such that in (2.13), $\hat{y}_{ij} = x'_i\hat{\beta} + \hat{u}_i;\ j \in \bar{s}_i$, as computed under the model. The authors show by use of Taylor approximation that the relative error due to the use of the LMM (2.14) instead of the GLMM (2.12) is below 10%, if the true probabilities are in the interval [0.3, 0.7].

They consider three estimators for the DMSE of the predictor $\hat{P}_i$ obtained under the model (2.14). The first estimator is a nonparametric bootstrap (NPB) estimator, obtained by first replicating each sample observation (x_{ij}, y_{ij}) w_{ij} times, where $w_{ij} \cong (1/\pi_{j|i})$ is the rounded calibrated sampling weight, yielding the pseudo population $\{(y_{ij}^*, x_{ij}^*),\ i = 1,...,m,\ j = 1,...,\langle \hat{N}_i \rangle\}$, where $\langle \hat{N}_i \rangle = (\sum_{j \in s_i} w_{ij})$ is the closest integer of $\hat{N}_i = \sum_{j \in s_i} w_{ij}$, and the bootstrap proportions, $P_i^* = \frac{1}{\hat{N}_i}\sum_{j=1}^{\hat{N}_i} y_{ij}^*,\ \ i = 1,...,m$. In the next step, B samples are drawn from each area of replicated measurements and the LMM (2.14) is fitted, yielding the estimates $\hat{P}_{i,b}^{*,NPB}$. The nonparametric bootstrap DMSE estimator is calculated under this method as,

$$\hat{DMSE}_{NPB}(\hat{P}_i) = \left(1 - \frac{n_i}{N_i}\right)\frac{1}{B}\sum_{b=1}^{B}\left(\hat{P}_{i,b}^{*,NPB} - P_i^*\right)^2. \tag{2.15}$$

The second estimator is a composite estimator of the NPB estimator, and the PMSE estimator (under the model 2.13),

$$\hat{DMSE}_{COM}(\hat{P}_i) = \hat{\gamma}_i \hat{DMSE}_{NPB}(\hat{P}_i) + (1 - \hat{\gamma}_i)\hat{PMSE}(\hat{P}_i), \tag{2.16}$$

where $\hat{\gamma}_i = \hat{\sigma}_u^2(\hat{\sigma}_u^2 + \sigma_{D_i}^2)^{-1}$ is calculated from the original sample and $\hat{PMSE}(\hat{P}_i)$ is the PMSE estimator under the model. See the explanation after Eq. (2.8) for the rationale of the estimator (2.16).

The third estimator considered by Molina and Strzalkowska-Kominiak (2017) is a "parametric design bootstrap estimator", obtained by generating a (single) population $\{(y_{ij}^{pb}, x_{ij}),\ i = 1,...,m,\ j = 1,...,N_i\}$ under the model (2.14) with estimated parameters $\hat{\beta}, \hat{\sigma}_u^2$ obtained from the original sample, drawing bootstrap samples as under the first method and computing the model-based predictors $\hat{P}_{i,b}^{PB}$ for each bootstrap sample and the regression estimators $\hat{P}_i^{reg} = \bar{y}_i + (\bar{X}_i - \bar{x}_i)'\hat{\beta}$ for each area, where

$(\bar{y}_i, \bar{x}_i) = (\frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij})$ are the sample means and $\bar{X}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$ is the true area mean in area i. The DMSE estimator is defined as,

$$\hat{DMSE}_{PD}(\hat{P}_i) = \hat{\gamma}_i \frac{1}{B}\sum_{b=1}^{B}(\hat{P}_{i,b}^{PB} - \hat{P}_i^{reg})^2 + (1-\hat{\gamma}_i)\frac{1}{B}\sum_{b=1}^{B}(\hat{P}_{i,b}^{PB} - \hat{P}_i)^2, \quad (2.17)$$

where $\hat{P}_i$ is the model based predictor calculated from the original sample. The rationale of this estimator is that for areas with large sample sizes (large $\hat{\gamma}_i$), the regression estimator $\hat{P}_i^{reg}$ is more reliable as an estimator of the true unknown proportion, particularly when the working model (2.14) is not correct, but the model-based predictor $\hat{P}_i$ is more stable for areas with small sample sizes (small $\hat{\gamma}_i$).

Molina and Strzalkowska-Kominiak (2017) conclude, based on a simulation experiment, that the composite estimator (2.16) and the parametric design bootstrap estimator (2.17) have an acceptable quality for estimation of the DMSE of the EBLUP estimator $\hat{P}_i$ under the model (2.14).

Rao et al. (2018) likewise consider the basic unit-level model (2.14), with the target area parameter being $\bar{Y}_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$. For known parameters $(\beta, \sigma_u^2, \sigma_\varepsilon^2)$, the best model-based predictor of the area mean $\bar{Y}_i$ is in this case,

$$\hat{\bar{Y}}_i^B = E[\bar{Y}_i \mid \{(y_{ij}, x_{ij}),\ j = 1,...,n_i\}] = \bar{X}_i'\beta + a_i(\bar{y}_i - \bar{x}_i'\beta), \quad (2.18)$$

where $a_i = (1-f_i)\gamma_i + f_i;\ f_i = (n_i / N_i)$ and $\gamma_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2 / n_i}$. By (2.18),

$\hat{\bar{Y}}_i^B - \bar{Y}_i = a_i(\bar{u}_i - \bar{U}_i) - (1-a_i)\bar{U}_i$, where $\bar{u}_i, \bar{U}_i$ are the area sample and population means of the values $u_{ij} = (y_{ij} - x_{ij}'\beta)$ and hence under simple random sampling without replacement (SRSWR) within each area,

$$\mathrm{DMSE}(\hat{\bar{Y}}_i^B) = a_i^2 V_d(\bar{u}_i) + (1-a_i)^2 \bar{U}_i^2, \quad (2.19)$$

where $V_d(\bar{u}_i) = \frac{1}{n_i}(1-f_i)S_{ui}^2;\ S_{ui}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(u_{ij} - \bar{U}_i)^2$.

A design-unbiased estimator of the DMSE is obtained by estimating unbiasedly S_{ui}^2 by $s_{ui}^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (u_{ij} - \bar{u}_i)^2$ and $\bar{U}_i^2$ by $\hat{\bar{U}}_i^2 = n_i^{-1} \sum_{j=1}^{n_i} u_{ij}^2 - N_i^{-1}(N_i - 1) s_{ui}^2$, yielding,

$$\hat{DMSE}(\hat{\bar{Y}}_i^B) = a_i^2 n_i^{-1}(1 - f_i) s_{ui}^2 + (1 - a_i)^2 \hat{\bar{U}}_i^2 . \tag{2.20}$$

The estimator (2.20) refers to the best predictor $\hat{\bar{Y}}_i^B$, which assumes known parameter values. In practice, the unknown parameters are replaced by sample estimates, yielding the empirical best (EB) estimator, $\hat{\bar{Y}}_i^{EB}$. A naïve estimator of $DMSE(\hat{\bar{Y}}_i^{EB})$ is obtained by replacing the unknown parameters in the right hand side of (2.20) by their sample estimates, but this estimator ignores the error resulting from the estimation of the model parameters. Hence, the authors propose the use of the composite estimator,

$$\hat{DMSE}(\hat{\bar{Y}}_i^{EB}) = \hat{\gamma}_i \hat{DMSE}(\hat{\bar{Y}}_i^{EB} \mid \hat{\beta}, \hat{\sigma}_u^2 \hat{\sigma}_\varepsilon^2) + (1 - \hat{\gamma}_i) \hat{PMSE}(\hat{\bar{Y}}_i^{EB}) , \tag{2.21}$$

where $\hat{DMSE}(\hat{\bar{Y}}_i^{EB} \mid \hat{\beta}, \hat{\sigma}_u^2 \hat{\sigma}_\varepsilon^2)$ is the naïve DMSE estimator and $\hat{PMSE}(\hat{\bar{Y}}_i^{EB})$ is the model-based PMSE estimator, similarly to (2.8), (2.16) and (2.17).

3. A New procedure for estimation of the randomization MSE

3.1 Some elementary estimators

To simplify the details of our proposed procedure, we consider in this section the FH (area-level) model (2.5). In the simulation study of Section 4, we apply the procedure to this model. In another simulation study in Section 5, we apply the procedure to the unit-level mixed logistic model of MacGibbon and Tomberlin (1989). To the best of our knowledge, DMSE estimation under the latter model has not been investigated in the literature so far. The proposed procedure follows the method of Pfeffermann and Correa (2012) and consists of modeling the DMSE and then applying the model to the original sample data.

For known model parameters $(\sigma_u^2, \sigma_{Di}^2, \beta)$, the FH estimator is presented in (2.6). Simple calculations show that the DMSE is in this case,

$$\lambda_i(\gamma_i, \beta, \sigma_{D_i}^2) = DMSE(\hat{\theta}_i) = \gamma_i{}^2 \sigma_{D_i}^2 + (1 - \gamma_i)^2 (\theta_i - x_i'\beta)^2 . \tag{3.1}$$

A design-unbiased estimator of $\lambda_i(\gamma_i, \beta, \sigma^2_{D_i})$ is thus, $\hat{\lambda}_i^{UB}(\gamma_i, \beta, \sigma^2_{D_i}) = (2\gamma_i - 1)\sigma^2_{D_i} + (1-\gamma_i)^2(y_i - x_i'\beta)^2$. Hence, for large number of areas, an approximately design unbiased estimator of $\lambda_i(\gamma_i, \beta, \sigma^2_{D_i})$ is,

$$\hat{\lambda}_i^{UB} = (2\hat{\gamma}_i - 1)\sigma^2_{D_i} + (1-\hat{\gamma}_i)^2(y_i - x_i'\hat{\beta}_{G\hat{L}S})^2, \tag{3.2}$$

where $\hat{\beta}_{G\hat{L}S}$ is the GLS estimator but with σ_u^2 replaced by $\hat{\sigma}_u^2$. (The estimators $(\hat{\gamma}_i, \hat{\beta}_{G\hat{L}S})$ are consistent for (γ_i, β) as the number of areas increases. As noted before, the sampling variance σ^2_{Di} is taken as known.) However, the estimator (3.2) is very unstable if the sampling variance $\sigma^2_{D_i}$ is large, as is commonly the case with small sample size.

A naïve DMSE estimator of $DMSE(\hat{\theta}_i)$ is obtained by replacing the unknown model parameters (σ_u^2, β) in (3.1) by sample estimates, and θ_i by its empirical FH estimate yielding,

$$\begin{aligned} \tilde{\lambda}_i^{Na} &= DMSE(\hat{\sigma}_u^2, \hat{\beta}, \hat{\theta}_i) = \hat{\gamma}_i{}^2\sigma^2_{Di} + (1-\hat{\gamma}_i)^2(\hat{\theta}_i^{FH} - x_i'\hat{\beta})^2 \\ &= \hat{\gamma}_i{}^2\sigma_{D_i}{}^2 + (1-\hat{\gamma}_i)^2\hat{\gamma}_i^2(y_i - x_i'\hat{\beta})^2. \end{aligned} \tag{3.3}$$

Remark 1. The estimator (3.3) is biased even for large samples since $(\hat{\theta}_i^{FH} - x_i'\hat{\beta})^2 = [\hat{\gamma}_i(y_i - x_i')]^2 = \hat{u}_i^2$ does not converge to $(\theta_i - x_i'\beta)^2 = u_i^2$. As with the estimator $\hat{\lambda}_i^{UB}$ in (3.2), this estimator does not account for the estimation of σ_u^2 and β.

Remark 2. A naïve estimator of the form (3.3) is not applicable for models or estimators for which the DMSE with known model parameters does not have an analytical expression. This is the case with the unit level logistic model considered later.

3.2 Proposed procedure

In this section we describe our proposed procedure for DMSE estimation for the case of the area-level model which, as mentioned before, consists of modelling the DMSE and then fitting the model to the original (actual) sample. In Section 5 we describe and illustrate the application of the procedure for the case of the mixed logistic model. The procedure accounts for the variability resulting from the estimation of the model parameters. It consists of the following 7 or 10 simple steps:

Step 1. Estimate $(\hat{\sigma}_u^2, \hat{\beta})$ from the original sample. Generate a large number of *R* values $\hat{\sigma}_{u,r}^2, \hat{\beta}_r$ from a neighborhood around $\hat{\sigma}_u^2, \hat{\beta}$ that is expected to include the true values underlying the hypothetical model generating the population values.

Step 2. Generate pseudo area means $\theta_{ri} = x_i'\hat{\beta}_r + u_{ri}$; $u_{ri} \sim N(0, \hat{\sigma}_{ur}^2)$, $r = 1,...,R; i = 1,...,m$, using the same covariates as in the actual population.

Step 3. For each pseudo population of area means, generate *J* parametric bootstrap samples, $y_{rij} = \theta_{ri} + e_{rij} = x_i'\hat{\beta}_r + u_{ri} + e_{ri}$; $e_{rij} \sim N(0, \sigma_{D_i}^2)$; $j = 1,...,J$, $r = 1,...,R$, $i = 1,...,m$ (*J* large).

Step 4. For each bootstrap sample, re-estimate $\hat{\beta}_{rj}, \hat{\sigma}_{urj}^2$ and compute the FH predictor, $\hat{\theta}_{rij} = \hat{\gamma}_{rij} y_{rij} + (1 - \hat{\gamma}_{rij}) x_i'\hat{\beta}_{rj}$; $\hat{\gamma}_{rij} = \hat{\sigma}_{urj}^2 (\hat{\sigma}_{urj}^2 + \sigma_{D_i}^2)^{-1}$.

Step 5. Approximate the true DMSE of the FH predictor, $\hat{\theta}_{ri} = \hat{\gamma}_{ri} y_{ri} + (1 - \hat{\gamma}_{ri}) x_i'\hat{\beta}_r$; $\hat{\gamma}_{ri} = \hat{\sigma}_{ur}^2 (\hat{\sigma}_{ur}^2 + \sigma_{D_i}^2)^{-1}$ as,

$$DMSE_{ri}(\hat{\theta}_{ri}) = DMSE_{ri}^{BS1}(\hat{\theta}_{ri}) = \frac{1}{J}\sum_{j=1}^{J} (\hat{\theta}_{rij} - \theta_{ri})^2 . \quad (3.4)$$

Remark 3. For *R*=1, Steps 1-5 correspond to the 1st stage of the double bootstrap estimator of Hall and Maiti (2006), but restricted to the randomization distribution. Consequently, what we refer to as the true DMSE may be viewed as a plausible estimate of $DMSE_i(\hat{\theta}_i)$, the target of estimation. See Section 4.2.

Remark 4. The FH model (2.5) does not specify how the individual measurements defining the finite population values in F (Eq. 2.2) are generated. Instead, it defines how the finite population means and the direct estimators are obtained. By definition, the finite population means are held fixed when computing the DMSE. The derivation of (3.4) follows exactly the same steps. For each replication *r* we obtain a single set of area means $\{\theta_{ri}\}$ from the same model, which are held fixed when approximating the DMSE by generating many direct sample estimates from their hypothesized distribution, and using them for computing the corresponding FH estimates. Conditioning on *F* is equivalent in this case to conditioning on the finite population means computed from *F*. See the paragraph following Eq. (2.2).

The following 3 steps are optional and found unnecessary in our simulation experiments:

Step 6. For each pseudo sample (Steps 3 and 4), generate a large number *B* of bootstrap samples, $y_{rij}^b = \theta_{rij} + e_{rij}^b$; $e_{rij}^b \sim N(0, \sigma_{Di}^2)$, where $\theta_{rij} = x_i'\hat{\beta}_{rj} + u_{rij}$; $u_{rij} \sim N(0, \hat{\sigma}_{urj}^2)$.

Step 7. Estimate $\hat{\beta}_{rj}^{(b)}, \hat{\sigma}_{urj}^{2(b)}$ for each bootstrap sample *b* and compute the FH predictor, $\hat{\theta}_{rij}^{(b)} = \hat{\gamma}_{rij}^{(b)} y_{rij}^b + (1-\hat{\gamma}_{rij}^{(b)}) x_i'\hat{\beta}_{rj}^{(b)}$; $\hat{\gamma}_{rij}^{(b)} = \hat{\sigma}_{urj}^{2(b)}(\hat{\sigma}_{urj}^{2(b)} + \sigma_{D_i}^2)^{-1}$.

Step 8. Compute the (double bootstrap) DMSE estimator,

$$DMSE_{ri}^{BS2}(\hat{\theta}_{ri}) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_{rij}^{(b)} - \theta_{rij})^2 . \tag{3.5}$$

Remark 5. The estimator (3.5) corresponds to the 2st stage of the double bootstrap estimator of Hall and Maiti (2006), but restricted to the randomization distribution.

Let *R*=1. Denote, $D_i^{BS1} = DMSE_i^{BS1}(\hat{\theta}_i)$ and $D_i^{BS2} = DMSE_i^{BS2}(\hat{\theta}_i)$, where $DMSE_i^{BS1}(\hat{\theta}_i)$ and $DMSE_i^{BS2}(\hat{\theta}_i)$ are defined by (3.4) and (3.5) for the case *R*=1, in which case $(\hat{\sigma}_{ur}^2, \hat{\beta}_r) = (\hat{\sigma}_u^2, \hat{\beta})$, the estimates obtained from the original sample. A plausible estimator of $DMSE(\hat{\theta}_i)$, resulting from the double-bootstrap bias correction of Hall and Maiti (2006) but restricted to the randomization distribution is,

$$\hat{DMSE}^{\mathrm{BS}}(\hat{\theta}_i) = \begin{cases} D_i^{BS1} + (D_i^{BS1} - D_i^{BS2}), & \text{if } D_i^{BS1} \geq D_i^{BS2} \\ D_i^{BS1}\exp[(D_i^{BS1} - D_i^{BS2})/D_i^{BS2}], & \text{if } D_i^{BS1} < D_i^{BS2}. \end{cases} \tag{3.6}$$

We consider the estimator (3.6) in the simulation study.

Step 9. Search for a function $q_l(\cdot) = \hat{DMSE}_{q_l,ri}(\hat{\theta}_{ri})$ of known predictors, which best estimates $DMSE_{ri}(\hat{\theta}_{ri})$(Eq. 3.4) among plausible functions $q_l(\cdot)$. In the simulation study of Section 4, which considers the case of the area level model we consider as possible predictors, $DMSE_{ri}^{BS2}(\hat{\theta}_{ri} \mid \theta_{ri}), \hat{PMSE}_{ri}(\hat{\theta}_{ri}), x_i'\hat{\beta}_r, \hat{\sigma}_{ur}^2, \sigma_{D_i}^2, \hat{\gamma}_{ri}, \hat{\gamma}_{ri}^2, (1\text{-}\hat{\gamma}_{ri})^2, \hat{\theta}_{ri}, (\hat{\theta}_{ri} - x_i'\hat{\beta}_r)^2, (y_{ri} - x_i'\hat{\beta}_r)^2$, and interactions between them. Obviously, other predictors can be considered, depending on the underlying model. See Section 5 for the predictors considered for the case of the mixed logistic model (2.12).

Our proposed search procedure is based on cross-validation techniques (see below). First select the "best predictors" for each candidate function by use of stepwise regression, using the observations allocated to the training group, and then choose the best function based on the observations in the validation group. Notice that the total number of observations is $T = m \times R$.

Step 10. Apply the chosen function to the original sample to obtain an estimator of the DMSE of the empirical FH estimator, $\hat{\theta}_i^{FH} = \hat{\gamma}_i y_i + (1-\hat{\gamma}_i) x_i' \hat{\beta}$, which is based on the original sample.

We mentioned in Step 9 that we propose using a cross validation technique for searching the "best" function. In our simulation study we used the following procedure.

Split randomly the *T* vectors of estimated and predictors' values,

$a_{ri} = (DMSE_{ri}(\hat{\theta}_{ri}); DMSE_{ri}^{BS2}(\hat{\theta}_{ri}), PM\hat{S}E_{ri}(\hat{\theta}_{ri}), x_i'\hat{\beta}_r, \hat{\sigma}_{ur}^2, \sigma_{D_i}^2, \hat{\gamma}_{ri}, \hat{\gamma}_{ri}^2, (1-\hat{\gamma}_{ri})^2, \hat{\theta}_{ri}, (\hat{\theta}_{ri} - x_i'\hat{\beta}_r)^2, (y_{ri} - x_i'\hat{\beta}_r)^2$ into a *training group*, G_{Tr}, of size T_r and a *validation group*, G_{Va}, of size $V_r = T - T_r$. Eligible functions $q_l(\cdot)$, $l = 1,...,L$ are estimated in G_{Tr}, and then compared in G_{Va}. Choose the function $D\hat{M}SE_{qm}$ minimizing the squared differences, $D\hat{M}SE_{q_l,ri} = \frac{1}{V_r}\sum_{a_{ri} \in G_{Va}} [D\hat{M}SE_{q_l,ri}(\hat{\theta}_{ri}) - DMSE_{ri}(\hat{\theta}_{ri})]^2$;

$$D\hat{M}SE_{qm} = \min_{q_l}(D\hat{M}SE_{q_l,ri}) = \min_{q_l}\{\frac{1}{V_r}\sum_{a_{ri} \in G_{Va}} [D\hat{M}SE_{q_l,ri}(\hat{\theta}_{ri}) - DMSE_{ri}(\hat{\theta}_{ri})]^2\}. \quad (3.7)$$

Obviously, other loss functions can be considered.

4. Simulation study for the area-level model

In this section we report the results of two simulation experiments, aimed to illustrate the performance of the proposed procedure in the case of the area level model (2.5), and compare it to other procedures proposed in the literature, reviewed in Section 2. In Section 5 we report the results of another simulation experiment for the case of the generalized linear mixed logistic model (2.12).

4.1 Simulation set-up

First experiment: Following Datta et al. (2005), we generated a single set of covariates $x_i \sim U[0,10]$ for 250 areas, and used them to divide the areas into 5 groups of 50 areas in each group, based on the ascending values of x. Next we generated area means, θ_i, as $\theta_i = 20+5x_i + u_i$; $u_i \sim N(0,100)$, and sample values $y_i = \theta_i + e_i$; $e_i \sim N(0, \sigma_{Di}^2)$, $i = 1,...,250$. The sampling variances, σ_{Di}^2, are the same for each group but differ between the groups; $\sigma_{D1}^2 = 30, \sigma_{D2}^2 = 40, \sigma_{D3}^2 = 50, \sigma_{D4}^2 = 60, \sigma_{D5}^2 = 70$.

Second experiment: Following Rao et al. (2018), we generated another single set of covariates $x_i \sim N(-1, 1)$ for 30 areas, and then classified the areas at random into 5 groups of 6 areas in each

group, similarly to the first experiment. For this experiment we generated area means as $\theta_i = 1+x_i + u_i$; $u_i \sim N(0,1)$, and sample values $y_i = \theta_i + e_i$; $e_i \sim N(0, \sigma^2_{Di})$, $i = 1,.....,30$. The sampling variances are again the same for each group but differ between the groups; $\sigma^2_{D1} = 0.2,\ \sigma^2_{D2} = 0.4,\ \sigma^2_{D3} = 0.5,\ \sigma^2_{D4} = 0.6,\ \sigma^2_{D5} = 2$.

4.2 DMSE estimators considered in simulation experiments

1- The "unbiased" estimator $\hat{\text{DMSE}}\_\text{UB}$ (the approximately unbiased estimator $\hat{\lambda}_i^{UB}$, Eq. 3.2).

2- The naive estimator $\hat{\text{DMSE}}\_\text{Na}$ (the naive estimator $\tilde{\lambda}_i^{Na}$, Eq. 3.3).

3- The estimators $\hat{\text{DMSE}}\_1$ (the estimator $DMSE_1(\hat{\theta}_i^{FH})$, Eq. 2.8) and $\hat{\text{DMSE}}\_2$ (the estimator $DMSE_2(\hat{\theta}_i^{FH})$, Eq. 2.9), as proposed by Rao et al. (2018), with the model parameters estimated by REML. However, in order to facilitate the computations, we replaced the exact design unbiased estimator $\hat{DMSE}(\hat{\theta}_i^{FH})$ in (2.9) by the approximately unbiased estimator $\hat{\lambda}_i^{UB}$ (Eq. 3.2). The estimator $\hat{PMSE}(\hat{\theta}_i^{FH})$ has been computed as in Datta and Lahiri (2000).

4- The estimators $DMSE_{ri}^{BS1}(\hat{\theta}_{ri})$, $DMSE_i^{BS2}$ and $\hat{DMSE}^{\text{BS}}$, (Eqs. 3.4-3.6) with *R*=1).

5- The proposed estimator $\hat{DMSE}_{qm}$ (Eq. 3.7). (See below for the functions considered).

4.3 Functions considered for application of the proposed DMSE estimator

For application of our proposed procedure, we generated *R*=100 pseudo populations (Step 1), *J*=200 bootstrap samples for each pseudo population (Step 3), and *B*=200 second-stage bootstrap samples (Step 6).

Denote for convenience, $DMSE_{ri}(\hat{\theta}_{ri}) = D_i$, $\hat{\beta}_r = \hat{\beta}$, $\hat{\gamma}_{ri} = \hat{\sigma}^2_{ur}(\hat{\sigma}^2_{ur} + \sigma^2_{D_i})^{-1} = \hat{\gamma}_i$, $\hat{\theta}_{ri} = \hat{\theta}_i$ (See Step 5).

We considered 5 possible definitions (transformations) of the dependent variable D_i:

$$D_i,\ log(D_i),\ arcsin\left[\frac{D_i}{100}\right],\ \frac{1}{D_i},\ \sqrt{D_i},\ \frac{1}{\sqrt{D_i}}.$$

Application of the proposed cross-validation procedure (Step 9) for the first experiment resulted in the following "best function":

$$\log(D_i) = \alpha_0 + \alpha_1\hat{\gamma}_i^2\sigma^2_{Di} + \alpha_2\hat{\gamma}_i\sigma^2_{Di} + \alpha_3\sigma^2_{Di} + \alpha_4(1-\hat{\gamma}_i)^2(\hat{\theta}_i - x_i'\hat{\beta})^2. \qquad (4.1)$$

The "best function" for the second experiment turned out to be:

$$\log(D_i) = \alpha_0 + \alpha_1 \hat{\gamma}_i \sigma_{Di}^2 + \alpha_2 (1-\hat{\gamma}_i)^2 (y_i - x_i'\hat{\beta})^2 + \alpha_3 (\hat{\theta}_i - x_i'\hat{\beta})^2 . \qquad (4.2)$$

Remark 6. For each function we computed the DMSE estimators as $\hat{DMSE} = E(D_i \mid \text{the predictors})$.

4.4. Performance assessment

In order to assess the performance of the DMSE estimators listed in Section 4.2 and compare them, we computed for each area i and each estimator the absolute relative error (ARE) defined as:

$$ARE_i = \frac{\left| \hat{DMSE}_i - DMSE_i \right|}{DMSE_i} \quad . \qquad (4.3)$$

4.5. Results

Figure 1 exhibits the "true" (empirical) DMSE and PMSE of the FH estimator in the first experiment. The two MSE measures have been computed as follows:

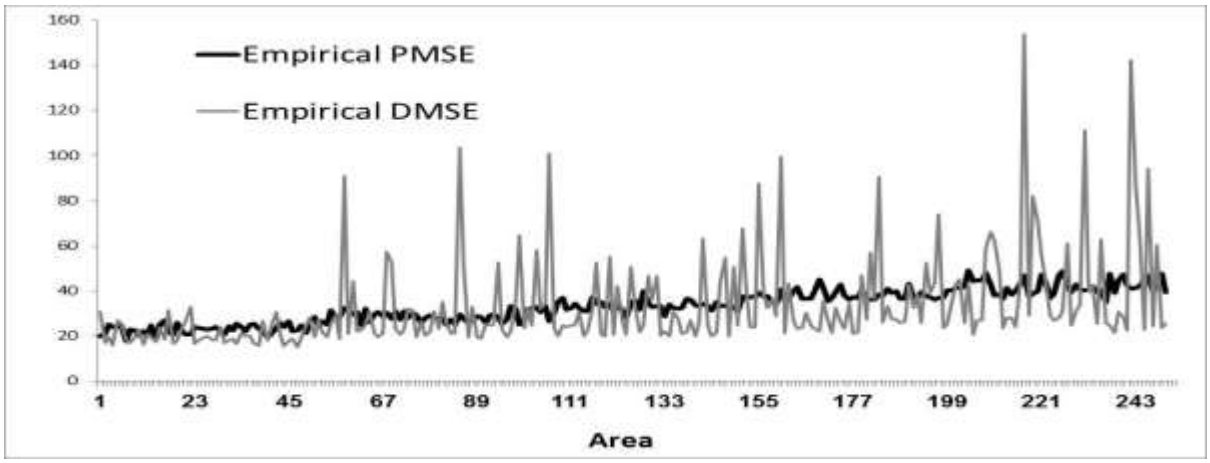
Empirical DMSE: *K*=10,000 simulated samples $\{y_i^k, i=1,...,250; k=1,...,K\}$ were generated as $y_i^k = \theta_i + e_i^k$, with $\theta_i = 1 + x_i + u_i$; $u_i \sim N(0,1)$, $i=1,...,250$ generated only once and held fixed, and the sampling errors e_i^k generated from $N(0, \sigma_{Di}^2)$. For each simulated sample we computed the FH estimator $\hat{\theta}_i^{FH(k)}$, and then the empirical DMSE,

$$DMSE_i = \frac{1}{K} \sum_{k=1}^{K} (\hat{\theta}_i^{FH(k)} - \theta_i)^2 \qquad (4.4)$$

Empirical PMSE: We generated *T*=10,000 samples $\{y_i^t, i=1,...,250, t=1,...,T\}$ as $y_i^t = 1 + x_i + u_i + e_i^t$ $= \theta_i + e_i^t$, where $u_i \sim N(0,1)$ and $e_i \sim N(0, \sigma_{Di}^2)$. (New random effects and sampling errors for each simulated sample.) For each simulated sample $\{(y_i^{(t)}, \theta_i^{(t)}, x_i), i=1,...,250\}$ we computed $\hat{\theta}_i^{FH(t)}$ and then the empirical PMSE:

$$PMSE_i = \frac{1}{T} \sum_{t=1}^{T} (\hat{\theta}_i^{FH(t)} - \theta_i)^2 . \qquad (4.5)$$

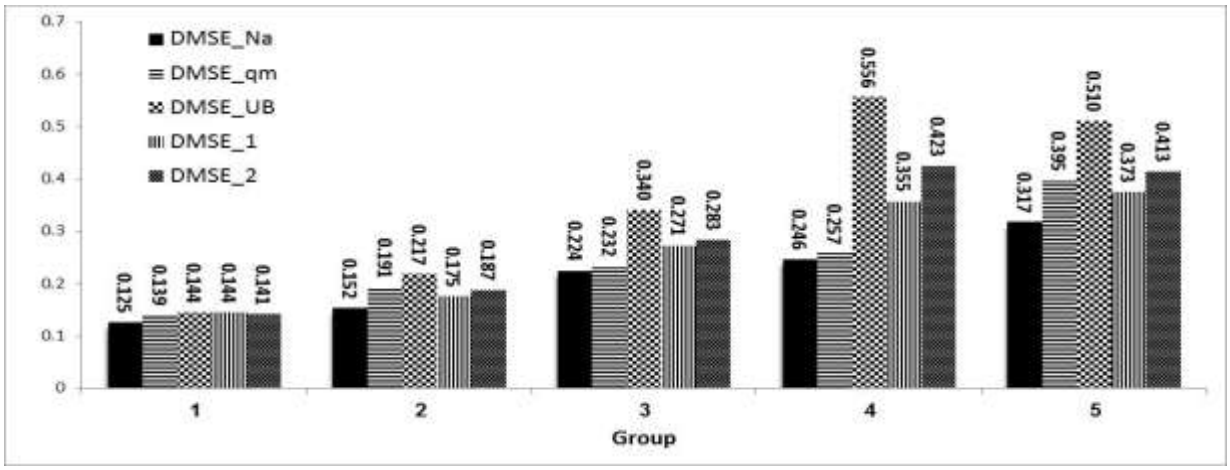
Figure 1. Empirical DMSE and PMSE (first experiment).
Areas ordered by the ascending values of x



As noted before and clearly seen in the graph, the DMSEs fluctuate around the PMSEs, with a few isolated large values, resulting from corresponding large u_i^2, which are integrated out when computing the PMSEs. (As implied by Eq. 3.1, the DMSE increases as the square of the random effect and/or the sampling variance increase.)

Figure 2 shows averages of the AREs (Eq. 4.3) within the 5 groups of equal sampling variances, for five of the estimators listed in Section 4.2: $DMSE\_Na$, $DMSE\_UB$, $DMSE\_1$, $DMSE\_2$ and the newly proposed estimator $DMSE\_qm$. (We omit the "hats" from the notation for convenience.) The AREs of the three bootstrap estimators listed in Section 4.2 are not shown, because they are much higher than the AREs of the estimators shown, even higher than the ARE of the unbiased estimator.

Figure 2. Average of ARE statistics within the 5 groups (first experiment).
Groups ordered by the ascending values of the sampling error variances.

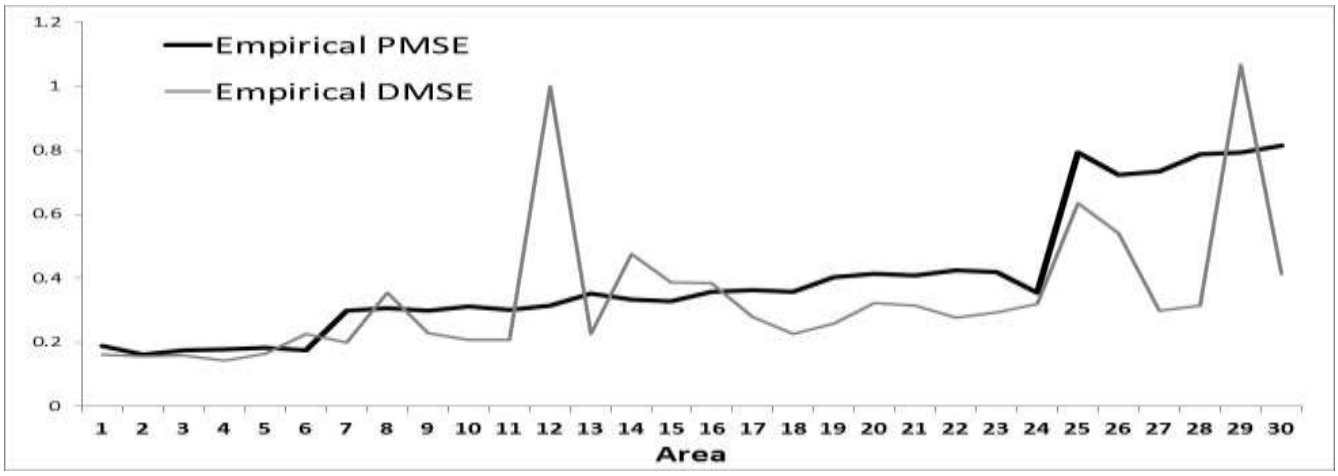


The averages $Av(\mathrm{u}_i^2)$ within the five groups are 68.4, 120.7, 89.4, 92.6, and 118.8, respectively. $(\sigma_U^2 = 100)$.

The first notable, although expected outcome revealed from Figure 2 is that the AREs of all the DMSE estimators generally increase, as the sampling variance increases. Among the five estimators, the Naïve estimator performs best in all the groups and the unbiased estimator performs the worst. The two estimators of Rao et al. (2018) perform similarly in the first three groups with the small sampling variances, but the first estimator with the weighting coefficients $\hat{\gamma}_i$ and $(1-\hat{\gamma}_i)$ performs better in Groups 4 and 5. As explained in Section 2.2.1, this outcome is explained by the fact that for large sampling variances, the first estimator assigns more weight to the model estimator $P\hat{M}SE(\hat{\theta}_i^{FH})$, which is much more stable in this case than the approximately design-unbiased estimator of the DMSE. The proposed estimator performs quite similarly to the two estimators of Rao et al. (2018) in groups 1,2 and 5, but outperforms them in Groups 3 and 4. It performs similarly to the naive estimator except in the last group with the large sampling error variances. The fact that the naïve estimator performs best in this experiment is not surprising. With 250 areas, the unknown model parameters (σ_u^2, β) are estimated almost perfectly, and so the empirical FH estimator is very close to the estimator that uses the true model parameters, implying that the naive estimator is close in this case to the true DMSE of the empirical estimator.

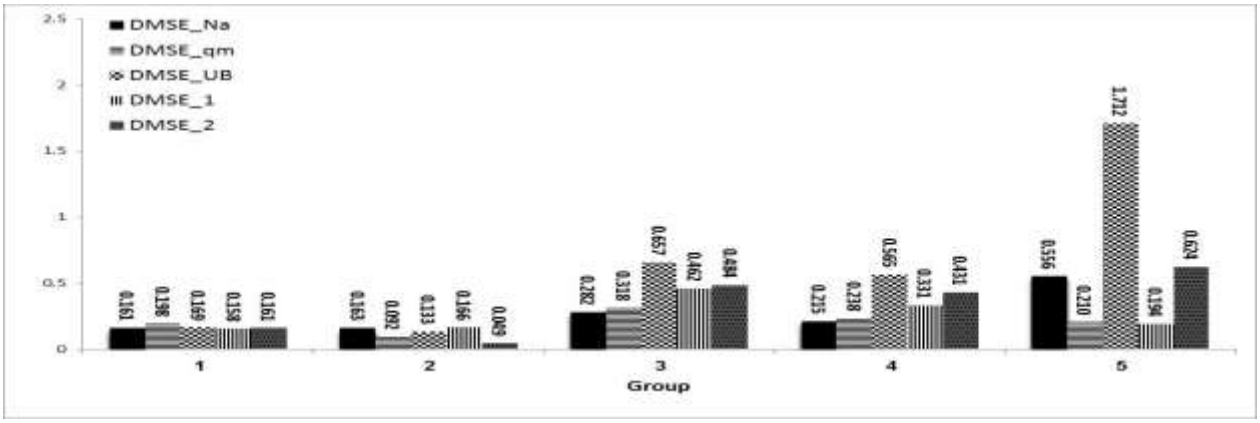
The next two figures summarize the results obtained for the second experiment described in Section 4.1. The empirical ("true") DMSE and PMSE have been calculated similarly to the first experiment.

Figure 3. Empirical DMSE and PMSE (second experiment). Areas ordered by the ascending values of x



The empirical DMSEs again fluctuate around the empirical PMSE with a few "outlying" values, which are explained by a large value $u_{12}^2 = 5.04$ in Area 12 and very small values $u_{27}^2 = 0.086$, $u_{28}^2 = 0.029$, $u_{30}^2 = 0.022$ in Areas 27, 28 and 30. respectively. $(\sigma_U^2 = 1)$. The averages $Av(\mathrm{u}_i^2)$ within the five groups are in this case 0.802, 1.417, 0.670, 0.371 and 0.573 respectively.

Figure 4. Average of ARE statistics within the 5 groups (second experiment). Groups ordered by the ascending values of the sampling error variances.



For this experiment the naive DMSE estimator no longer outperforms the other estimators because with only 30 areas, the empirical FH estimators differ in a noticeable way from the estimators based on the true model parameters and hence the naive estimator no longer accounts properly for the use of estimated parameters. See the discussion following Figure 2. Overall, the proposed estimator performs best in this experiment across the groups. As with the first experiment, the first estimator of Rao et al. (2018) outperforms the second estimator in the last two groups with the large sampling variance. Here again, none of the bootstrap estimators performs satisfactorily.

5. Simulation study for the generalized linear mixed logistic model

In this section we report the results of our third simulation experiment, aimed to illustrate the performance of the proposed procedure for the case of the unit-level generalized linear mixed model defined by (2.12).

5.1 Simulation set-up and estimation of target proportions

Following Pfeffermann et al. (2012), we generated a population of *m*=30 areas with $N_i = 1{,}000$ units in each area. For each unit j in area i, we generated covariate values

$x_{1ij} \sim Ber(0.5),\ x_{2ij} \sim U(20,40)$, and binary responses $y_{ij} \sim Ber(p_{ij})$ with $p_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i)}{1+\exp(\beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_i)}$; $\beta_0 = -1,\ \beta_1 = 0.5,\ \beta_2 = 0.1$ and $u_i \sim N(0, \sigma_u^2 = 1)$. We divided the areas into 5 groups with 6 areas in each group. Next we drew simple random samples without replacement from each area, such that the sample sizes are the same for the areas in the same group but differ between the groups; the sample sizes in the five groups are 10, 50, 100, 200, 400.

The target area parameters are in this case the true proportions $P_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$. The estimators are computed as $\hat{P}_i = \frac{1}{N_i}(\sum_{j\in s_i} y_{ij} + \sum_{k\in \bar{s}_i} \hat{p}_{ik})$, where s_i and $\bar{s}_i$ define correspondingly the sample and non-sample units in the area and $\hat{p}_{ik}$ defines the empirical best predictor of p_{ij} given the observed data, with the unknown β-coefficients replaced by their sample estimates. See Pfeffermann and Correa (2012) for more details. For the present application we used the SAS procedure *nlmixed* for producing the estimates $\hat{\beta}$ and $\hat{u}_i$ (for each area), and then estimated,

$$\hat{p}_{ik} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + \hat{\beta}_2 x_{2ij} + \hat{u}_i)}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1ij} + \hat{\beta}_2 x_{2ij} + \hat{u}_i)} \text{ for } k \in \bar{s}_i . \quad (5.1)$$

As mentioned earlier, we are not aware of any attempt to estimate the DMSE for this, much more complicated model. (Molina and Strzalkowska-Kominiak (2018) approximate the mixed logistic model by the mixed linear model.) We considered therefore the following DMSE estimators.

1- The bootstrap estimators $DMSE_i^{BS1}(\hat{P}_i)$, $DMSE_i^{BS2}(\hat{P}_i)$, $DMSE_i^{BS}(\hat{P}_i)$, (Eqs. 3.4-3.6 with R=1, see also below).

2- The proposed estimator $\hat{DMSE}_{qm}$. (See below for the functions considered).

For application of the proposed method to unit-level models, the computation of new estimators based on new samples requires at each step to first generate a corresponding new synthetic population and then sample from that population many times. Following, we outline the main steps of our proposed procedure for the present model:

Step 1. Same as before.

Step 2. Compute $p_{rij} = \frac{\exp(x_{ij}'\hat{\beta}_r + u_{ri})}{1+\exp(x_{ij}'\hat{\beta}_r + u_{ri})}$; $u_{ri} \sim N(0, \hat{\sigma}_{ur}^2)$, $r = 1,...,R,\ i = 1,...,m,\ j = 1,...,N_i$, using the same covariates as in the actual population.

Step 3. Generate *R* populations from $y_{rij} \sim Ber(P_{rij})$ and calculate, $P_{ri} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{rij}$.

Step 4. For each synthetic population r, sample *T* simple random samples without replacement (SRSWOR) and re-estimate $(\hat{\beta}_{rt}, \hat{\sigma}_{urt}^2, \hat{u}_{rit})$ and $\hat{P}_{rit} = \frac{1}{N_i}(\sum_{j\in s_{rit}} y_{rij} + \sum_{k\in \bar{s}_{rit}} \hat{p}_{rikt})$, where $\hat{p}_{rikt} = \frac{\exp(x_{ik}'\hat{\beta}_{rt} + \hat{u}_{rit})}{1+\exp(x_{ik}'\hat{\beta}_{rt} + \hat{u}_{rit})}$. The sample sizes are the same as for the original (actual) sample.

Step 5. Approximate the DMSE of $\hat{P}_{ri}$ as $DMSE_{ri}(\hat{P}_{ri}) = DMSE_{ri}^{BS1}(\hat{P}_{ri}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{P}_{rit} - P_{ri})^2$.

Step 6. For each pseudo sample (Step 4), compute $p_{rijt} = \frac{\exp(x_{ij}'\hat{\beta}_{rt} + u_{rit})}{1+\exp(x_{ij}'\hat{\beta}_{rt} + u_{rit})}; u_{rit} \sim N(0, \hat{\sigma}_{urt}^2)$, using the same covariates as in the actual population. Generate *T* populations $y_{rijt} \sim Ber(p_{rijt})$ and calculate $P_{rit} = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{rijt}$.

Step 7. Draw *B* samples from each synthetic population by use of SRSWOR and re-estimate $(\hat{\beta}_{rt}^{(b)}, \hat{\sigma}_{urt}^{2(b)}, \hat{u}_{rit}^{(b)})$ and $\hat{p}_{rikt}^{(b)} = \frac{\exp(x_{ik}'\hat{\beta}_{rt}^{(b)} + \hat{u}_{rit}^{(b)})}{1+\exp(x_{ik}'\hat{\beta}_{rt}^{(b)} + \hat{u}_{rit}^{(b)})}$. Compute $\hat{P}_{rit}^{(b)} = \frac{1}{N_i}(\sum_{j\in s_{rit}} y_{rijt}^{(b)} + \sum_{k\in \bar{s}_{rit}} \hat{p}_{rikt}^{(b)})$.

Step 8. Compute the (double bootstrap) DMSE estimator,

$$DMSE_{ri}^{BS2}(\hat{P}_{ri}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{B}\sum_{b=1}^{B}(\hat{P}_{rit}^{(b)} - P_{ri})^2 . \quad (5.2)$$

Steps 9 and 10. Same as before.

We generated *R*=100 pseudo populations (Step 3), *T*=200 bootstrap samples for each pseudo population (Step 4), and *B*=200 second-stage bootstrap samples (Step 6).

5.2. Functions considered for application of proposed DMSE estimation

Denote, as before, $DMSE_{ri}(\hat{P}_{ri}) = D_i$, $\hat{\beta}_r = \hat{\beta}$, $\hat{P}_{ri} = \hat{P}_i$ (Step 5). We considered the same transformations of the dependent variable D_i as for the area level model, i.e.,

$D_i, log(D_i), arcsin\left[\frac{D_i}{100}\right], \frac{1}{D_i}, \sqrt{D_i}, \frac{1}{\sqrt{D_i}}$. As possible predictors we considered,

$DMSE_{ri}^{BS2}(\hat{P}_{ri}), PMSE_{ri}^{BS2}(\hat{P}_{ri}), \bar{x}_i'\hat{\beta}_r, \bar{y}_{ri}, (\bar{y}_{ri} - \bar{x}_i'\hat{\beta}_r)^2, \frac{\hat{P}_{ri}(1-\hat{P}_{ri})}{n_{ri}}, \hat{P}_{ri}, \hat{\sigma}_{ur}^2$ and interactions between them,

where $\bar{y}_{ri}$ is the simple sample proportion $\tilde{P}_{ri} = \frac{1}{n_{ri}} \sum_{j \in s_{ri}} y_{rij}$ and $n_{ri} = n_i$ denotes the i^{th} area sample size. We estimated $P\hat{M}SE_{ri}^{BS2}(\hat{P}_{ri})$ similarly to the computation of $D\hat{M}SE_{ri}^{BS2}(\hat{P}_{ri})$, except that each bootstrap sample is now sampled from a different synthetic population with random effects, which vary from one population to the other.

Application of the proposed cross-validation procedure (Step 9) in this experiment resulted in the following "best function", which was applied to the original (actual) sample in Step 10:

$$D_i = \alpha_0 + \alpha_1 \frac{\hat{P}_i(1-\hat{P}_i)}{n_i} + \alpha_2 \bar{x}_i'\hat{\beta} + \alpha_3 \bar{y}_i + \alpha_4 (\bar{y}_i - \bar{x}_i\hat{\beta})^2 . \quad (5.3)$$

Note that the predictor $\hat{P}_i(1-\hat{P}_i)/n_i$ is an estimator of the variance of the sample proportion $\tilde{P}_i$ in Area i, so the fact that it is included in the best function is not surprising.

For this model, the relative error of the DMSE estimators can be very small and hence we show for each area the absolute error $\mathrm{AE}_i = | D\hat{M}SE_i - DMSE_i |$ rather than the relative error.

Figure 5 shows the empirical DMSE and PMSE of the empirical best predictor under this model. Also shown for comparison are the empirical DMSE and PMSE of the sample proportions $\tilde{P}_i = \frac{1}{n_i} \sum_{j \in s_i} y_{ij}$ (the direct estimator). See Section 4.5 for computation of the empirical DMSE and PMSE, with straightforward modifications for the present model. Notice that $\tilde{P}_i$ is design-unbiased for P_i, and hence also design-model-unbiased (over all possible population values and sample selections).

Figure 5. Empirical DMSE and PMSE of empirical best predictor and of sample proportion. Mixed logistic model. Areas ordered by the ascending values of the sample sizes.

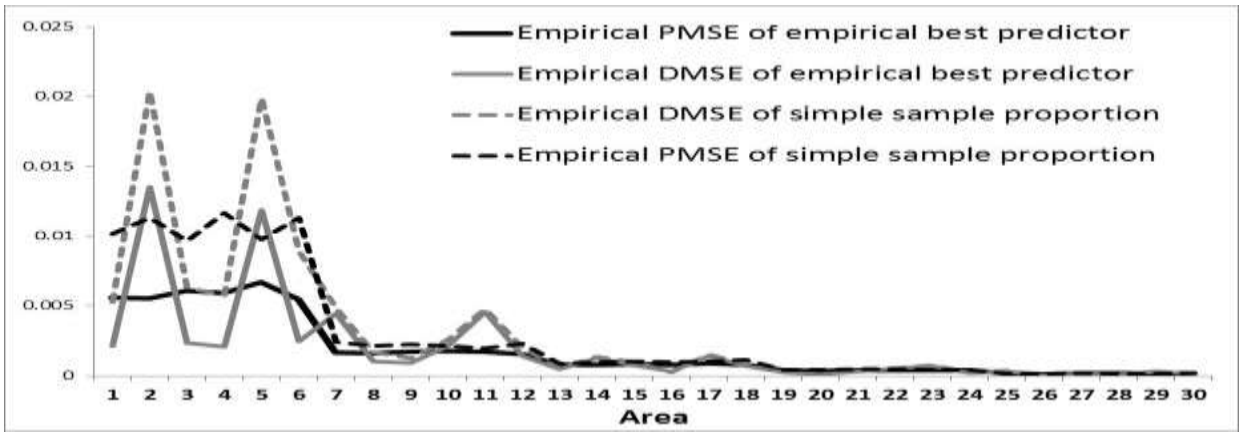
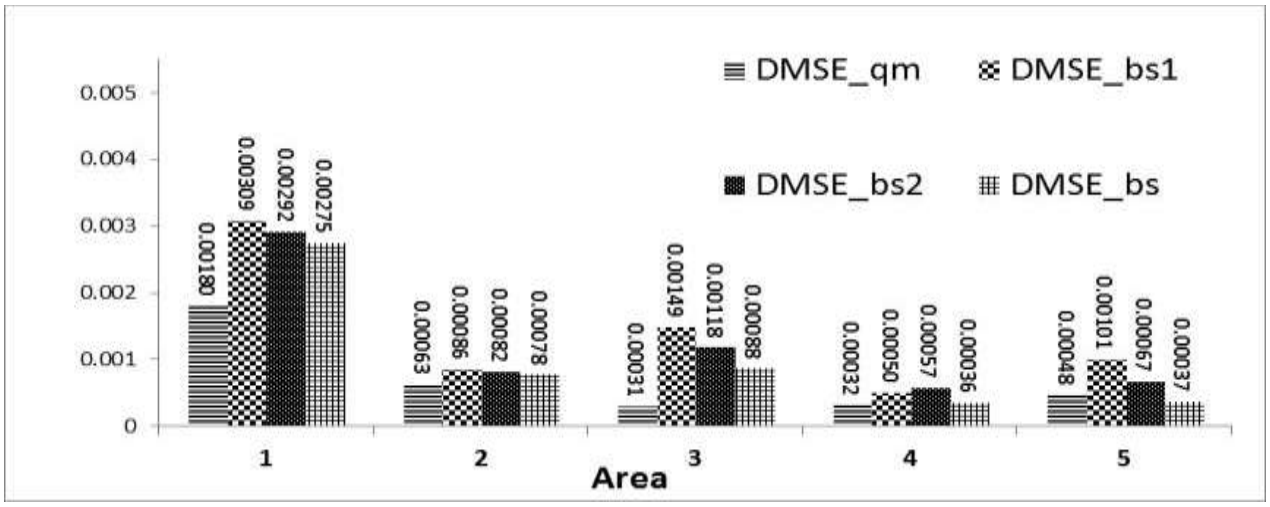


Figure 5 reveals a similar picture to Figures 1 and 3. As can be seen, the DMSE and PMSE of the empirical best predictor and of the sample proportion reduce, as the area sample sizes increase. Notice also that the DMSE and PMSE of the empirical best predictor and of the sample proportion behave similarly, but the empirical best predictor outperforms the sample proportion, both in terms of the DMSE and the PMSE, particularly in the areas with the small sample sizes.

Figure 6. Average of AE statistics within the 5 groups.



Averages of $Av(\mathrm{u}_i^2)$ within the groups: 0.773, 1.238, 0.572, 1.271 and 1.787 respectively $(\sigma_u^2 = 1)$.

Two notable outcomes emerging from Figure 6 are:

1- The proposed DMSE estimator performs better than the bootstrap estimators, particularly in the areas with the smaller absolute values of the random effects,

2- The estimator $DMSE_i^{BS2}(\hat{P}_i)$ outperforms $DMSE_i^{BS1}(\hat{P}_i)$, with both estimators being dominated by $\hat{DMSE}_i^{BS}(\hat{P}_i)$, which agrees with the conclusions of Hall and Martin (1988).

6. Concluding Remarks

In this article we propose a new method for estimating the DMSE of model-dependent predictors in small area estimation. The notable feature of this method is that it does not require the use of any approximately unbiased DMSE estimator, and it is applicable in principle to any model or estimator. It is more computational intensive than some of the other procedures considered in this article, but this should not introduce any difficulties in a real application. The method is shown to perform well, in contrast to common belief that it is practically impossible to estimate the DMSE of model dependent estimators, unless in areas with sufficiently large sample sizes.

There are two open questions underlying the use of the proposed procedure. The first regarding its theoretical properties and in particular, the order of bias as the number of areas with observations increases, and the second regarding its robustness to possible deviations from the working model used for its application. It would seem that second order bias can be established following similar lines to the proof in Pfeffermann and Correa (2012), as the proposed method is similar in nature to the method developed in the later article for estimating PMSEs. As for the second question, we have not yet studied the robustness of the procedure, but as we already commented in Section 2.1, basically all the model-dependent, as well as the resampling methods for MSE estimation proposed in the literature assume the "correctness" of the model assumed to generate the population and sample measurements. The same is also true for the other DMSE estimators reviewed in the present article, except for the first bootstrap estimator of Molina and Strzalkowska-Kominiak (2017), so that our proposed procedure is no exception in this regard. Still, the robustness of the procedure should be studied, at least via simulation experiments.

Finally, it is quite obvious that the proposed procedure can be improved by enlarging the group of plausible functions and by applying more advanced techniques for selecting the best function in the group. In the present article we focused more on the basic principles of the method, rather than on refinements of its application.

7. References

BATTESE, G. E., HARTER, R. M. and FULLER, W. A. (1988). An error components model for prediction of county crop area using survey and satellite data. *Journal of the American Statistical Association* **83** 28–36.

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.

DATTA, G. S. and LAHIRI, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica* **10** 613–627.

DATTA, G. S., RAO, J. N. K. and SMITH, D. D. (2005). On measuring the variability of small area estimators under a basic area level model. *Biometrika* **92** 183–196.

DATTA, G. S., KUBOKAWA, T., MOLINA, I. and RAO, J.N.K. (2011). Estimation of mean squared errors of model-based small area estimators. *Test*, 20, 367-388.

DAS, K., JIANG, J. and RAO, J.N.K. (2004). Mean squared error of empirical predictor. *Annals of Statistics*, **32**, 818-840.

HALL, P. and MAITI, T. (2006). On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society Series B* **68** 221- 238.

HALL, P. & MARTIN, M. A. (1988). On bootstrap resampling and iteration. *Biometrika* **75**, 661–71.

JIANG, J., LAHIRI, P. S. and WAN, S. M. (2002). A unified jackknife theory for empirical best prediction with M-estimation. *The Annals of Statistics* **30** 1782–1810.

LOHR, S. L. and RAO, J. N. K. (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. *Biometrika* **96** 457-468.

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

MOLINA, I. and STRZALKOWSKA-KOMINIAK, E. (2017). Estimation of proportions in small areas: application to the labor force using the Swiss Census Structural Survey. Unpublished report.

PFEFFERMANN, D. and CORREA, S. (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. *Biometrika* **99** 1-16.

PFEFFERMANN, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science* **28** 40-68.

PRASAD, N. G. N. and RAO, J. N. K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association* **85** 163–171.

RAO, J.N.K., and MOLINA, I. (2015). *Small Area Estimation*, 2nd Edition, Wiley.

RAO, J.N.K., RUBIN-BLEUER, S. and ESTEVAO, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. Unpublished technical report.

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

VAN den BRAKEL, J., BUELENS B. and BOONSTRA. J. B. (2016). Small area estimation to quantify discontinuities in repeated sample surveys. *J. R. Statist. Soc.* **179**, 229-250.