Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Local polynomial estimation for a small area mean under informative sampling

by Marius Stefan and Michael A. Hidiroglou

Release date: June 30, 2020

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2020
Vol. 46, No. 1, pp. 29-49
Statistics Canada, Catalogue No. 12-001-X

Local polynomial estimation for a small area mean under informative sampling

Marius Stefan and Michael A. Hidiroglou[1]

Abstract

Model-based methods are required to estimate small area parameters of interest, such as totals and means, when traditional direct estimation methods cannot provide adequate precision. Unit level and area level models are the most commonly used ones in practice. In the case of the unit level model, efficient model-based estimators can be obtained if the sample design is such that the sample and population models coincide: that is, the sampling design is non-informative for the model. If on the other hand, the sampling design is informative for the model, the selection probabilities will be related to the variable of interest, even after conditioning on the available auxiliary data. This will imply that the population model no longer holds for the sample. Pfeffermann and Sverchkov (2007) used the relationships between the population and sample distribution of the study variable to obtain approximately unbiased semi-parametric predictors of the area means under informative sampling schemes. Their procedure is valid for both sampled and non-sampled areas. Verret, Rao and Hidiroglou (2015) studied alternative procedures that incorporate a suitable function of the unit selection probabilities as an additional auxiliary variable. Their procedure resulted in approximately unbiased empirical best linear unbiased prediction (EBLUP) estimators for the small area means. In this paper, we extend the Verret et al. (2015) procedure by not assuming anything about the inclusion probabilities. Rather, we incorporate them into the unit level model via a smooth function of the inclusion probabilities. This function is estimated via a local approximation resulting in a local polynomial estimator. A conditional bootstrap method is proposed for the estimation of mean squared error (MSE) of the local polynomial and EBLUP estimators. The bias and efficiency properties of the local polynomial estimator are investigated via a simulation. Results for the bootstrap estimator of MSE are also presented.

Key Words: Local polynomial estimation; EBLUP estimation; Augmented model; Nested error model; Informative sampling; Conditional bootstrap.

1 Introduction

Population totals and means are often required for small subpopulations (or areas). When the inference is based on the area specific sample data, the resulting small area parameter estimators (direct estimators) are not of adequate precision due to the small area specific sample sizes. As a result, it becomes necessary to borrow strength across areas. Indirect estimators (predictors) that borrow strength are obtained when a model is used for the population of small areas. The model provides a link to related small areas. As a consequence, a model-based small area indirect estimator uses all the observations in the national sample, as well as the observations from the small area.

Suppose that the population of interest, U of size N, consists of M non-overlapping areas with N_i units in the i^{th} small area U_i $(i = 1, \ldots, M)$. A sample, s, of m areas is first selected using a specified sampling scheme with inclusion probabilities $\pi_i = mp_i\,(i = 1, \ldots, M)$, where p_i denotes the selection probability of small area i. Subsamples s_i of specified sizes n_i are independently selected from each small area U_i according to a specified sampling design with selection probabilities $p_{j|i}$ $\left(\sum_{j=1}^{N_i} p_{j|i} = 1\right)$. The inclusion probabilities are $\pi_{j|i} = n_i p_{j|i}$ with sampling weights $w_{j|i} = \pi_{j|i}^{-1}$. We consider the selection probabilities $p_{j|i}$ proportional to a size measure, c_{ij}, related to the response

1. Marius Stefan, Faculty of Applied Sciences, Polytechnic University of Bucharest, Splaiul Independentei, nr. 313. E-mail: mastefan@gmail.com; Michael A. Hidiroglou, Statistics Canada Alumni. E-mail: hidirog@yahoo.ca.

variable y_{ij}: that is $p_{j|i} = c_{ij} \Big/ \sum_{k=1}^{N_i} c_{ik}$. We assume that all small areas are sampled, that is $m = M$. The resulting overall sample size is $n = \sum_{i=1}^{M} n_i$.

The basic population nested error regression model introduced by Battese, Harter and Fuller (1988) is given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + e_{ij}, \; j = 1, \ldots, N_i; \; i = 1, \ldots, M, \tag{1.1}$$

where y_{ij} is the value of the response variable for unit j in small area i, $\mathbf{x}_{ij} = \left(1, x_{ij1}, \ldots, x_{ijp}\right)^T$ is the vector of covariates, $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)^T$ is the vector of fixed effects, and $v_i \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$ are the random small area effects independent of the unit level errors $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right)$. The estimation of small area means, $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, is of primary interest.

If the sampling design is non-informative for the model, that is if the model (1.1) holds for the sample, then efficient model-based estimators of the small area means $\bar{Y}_i$ can be obtained using empirical best linear unbiased prediction (EBLUP) (see Rao and Molina, 2015, Chapter 6 for an excellent account of the procedure). In this case, both the sample and population models coincide, allowing the use of (1.1) on the sample data to estimate $\bar{Y}_i$.

If the selection probability $p_{j|i}$ is related to y_{ij} even after conditioning on $\mathbf{x}_{ij}$, the sampling design is informative and the model (1.1) no longer holds for the sample. Consequently, the EBLUP estimator, that is based on (1.1) for the sample, may be heavily biased. It is, therefore, necessary to develop estimators that can account for sample selection, thereby reducing estimation bias. To this end, Verret et al. (2015) augmented model (1.1) by including the variable $g\left(p_{j|i}\right)$, where $g\left(p_{j|i}\right)$ is a specified function of the probability $p_{j|i}$. Their model for the sample is given by

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta}_0 + g\left(p_{j|i}\right) \delta_0 + v_{0i} + e_{0ij}, \; j = 1, \ldots, n_i; \; i = 1, \ldots, M, \tag{1.2}$$

where $v_{0i} \overset{\text{iid}}{\sim} N\left(0, \sigma_{0v}^2\right)$ and independent of $e_{0ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_{0e}^2\right)$, and $\boldsymbol{\beta}_0 = \left(\beta_{00}, \beta_{01}, \ldots, \beta_{0p}\right)^T$. Verret et al. (2015) checked the adequacy of (1.2) after fitting the model to sample data $\left(y_{ij}, \mathbf{x}_{ij}, p_{j|i}\right)$, $j = 1, \ldots, n_i; \; i = 1, \ldots, M$, for different choices of $g(\cdot)$ that provide the best fit to the data. They suggested the following four possibilities for the choice of $g\left(p_{j|i}\right)$: $p_{j|i}, \log\left(p_{j|i}\right), w_{j|i} = \left(n_i p_{j|i}\right)^{-1}$ and $n_i w_{j|i} = p_{j|i}^{-1}$. Since their sample model is parametric, the EBLUP theory can be used to estimate the relevant parameters using model (1.2).

Verret et al. (2015) illustrated via a simulation that the resulting EBLUP estimator, denoted as $\hat{\bar{Y}}_i^{\text{VRH}}$, obtained under (1.2), performs well under informative sampling design by reducing both bias and mean squared error as compared to the EBLUP estimator, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, obtained from the sample data under the non-augmented model (1.1). Their simulation study compared their approach to the one used in Pfeffermann and Sverchkov (2007). Their simulation results showed that the bias-adjusted estimator of Pfeffermann and Sverchkov (2007) performed well under informative sampling in terms of bias, but that its MSE is significantly larger than the corresponding MSE of the EBLUP estimator based on the augmented model.

In this paper, we make no assumptions concerning the form of the function $g\left(p_{j|i}\right)$. Instead, we incorporate the $p_{j|i}$'s into the model (1.1) via an unknown smooth function $m_0\left(p_{j|i}\right)$. Our smooth function $m_0(\cdot)$ does not have a parametric form such as the one in Verret et al. (2015). We suppose that $m_0(\cdot)$ can be locally approximated by a polynomial of order q. For each point l in small area U_k, the corresponding polynomial is obtained by the Taylor expansion of $m_0\left(p_{j|i}\right)$ in a neighbourhood of $p_{l|k}$. For each point (l, k) in the population, we replace $m_0\left(p_{j|i}\right)$ by the corresponding parametric approximation and fit the resulting model just as in parametric fitting. We refer to this method as parametric polynomial localization.

This local approximation results in an augmented model that is semi-parametric. Such models have been applied to small area estimation by Opsomer, Claeskens, Ranalli, Kauermann and Breidt (2008). These authors chose a technique based on penalized splines to estimate the non-parametric part of their models. Breidt and Opsomer (2000) and Breidt, Opsomer, Johnson and Ranalli (2007) used the local polynomial technique in survey sampling theory to construct model-assisted estimators. Their estimators were based on non-parametric models without random effects. To the best of our knowledge, the estimation of a small area mean $\bar{Y}_i$, based on a local polynomial technique under semiparametric models has hardly been investigated.

The paper is structured as follows. Section 2 provides a review of two methods that result in estimators that account for sample selection: these methods were developed by Pfeffermann and Sverchkov (2007) and by Verret et al. (2015). In Section 3, we present a three-step procedure to estimate the proposed semi-parametric augmented model and the small area mean $\bar{Y}_i$ using a local polynomial approximation. We label the resulting estimator of the small area mean as $\hat{\bar{Y}}_i^{\mathrm{LP}}$. The mean squared error (or MSE) of $\hat{\bar{Y}}_i^{\mathrm{LP}}$ is estimated in Section 4 by a parametric conditional bootstrap method. The conditional bootstrap method is also used to estimate the MSE of EBLUP estimators obtained under augmented model (1.2). In Section 5, we conduct a simulation study under the design-model (or pm) framework to compare the bias and MSE of the new estimator $\hat{\bar{Y}}_i^{\mathrm{LP}}$ to the EBLUP estimator, as well as to the two estimators discussed in Verret et al. (2015). We also study the performance of the conditional bootstrap procedure in estimating the MSE of the proposed local polynomial and EBLUP estimators studied in Verret et al. (2015). The performance is evaluated in terms of mean relative bias and mean confidence interval level. Concluding remarks are given in Section 6.

2 Existing methods

Suppose that the population model (1.1) holds for the sample. Let $\bar{\mathbf{X}}_i$ be the area mean of the population values $\mathbf{x}_{ij}$. Then the EBLUP estimator of $\mu_i = \bar{\mathbf{X}}_i^T\boldsymbol{\beta} + v_i$ is given by

$$\hat{\mu}_i^{\mathrm{EBLUP}} = \bar{\mathbf{X}}_i^T\hat{\boldsymbol{\beta}} + \hat{v}_i = \hat{\gamma}_i\bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{\gamma}_i\bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}, \tag{2.1}$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2 \Big/ \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_i\right)$, $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij} \Big/ n_i$, $\bar{\mathbf{x}}_i = \sum_{j=1}^{n_i} \mathbf{x}_{ij} \Big/ n_i$ are the unweighted sample means of the response variable y and the covariates $\mathbf{x}$, and $\hat{v}_i = \hat{\gamma}_i\left(\bar{y}_i - \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}\right)$. The estimator of the regression vector $\boldsymbol{\beta}$ in (1.1) is

$$\hat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{x}_{ij}\left(\mathbf{x}_{ij} - \hat{\gamma}_i\bar{\mathbf{x}}_i\right)^T\right\}^{-1}\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \hat{\gamma}_i\bar{\mathbf{x}}_i\right)y_{ij}\right\}. \tag{2.2}$$

The estimated variance components $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ are obtained by the Henderson method of fitting of constants (HFC) or restricted maximum likelihood (REML) (see Battese et al., 1988 and Chapter 7 in Rao and Molina, 2015). The EBLUP estimator of the area mean $\bar{Y}_i$ may be written in terms of $\hat{\mu}_i^{\text{EBLUP}}$ as

$$\hat{\bar{Y}}_i^{\text{EBLUP}} = \frac{1}{N_i}\left[\left(N_i - n_i\right)\hat{\mu}_i^{\text{EBLUP}} + n_i\left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}\right\}\right]. \tag{2.3}$$

Note that $\hat{\bar{Y}}_i^{\text{EBLUP}} \approx \hat{\mu}_i^{\text{EBLUP}}$ if the sampling fraction n_i / N_i is sufficiently small. The EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is design consistent under simple random sampling (SRS) or stratified SRS with proportional allocation within small area U_i, leading to equal $p_{j|i}$'s.

Pfeffermann and Sverchkov (2007) studied the estimation of small area means under informative sampling, assuming the following model for the sample data

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\alpha} + u_i + h_{ij};\ \ j = 1, \ldots, n_i;\ \ i = 1, \ldots, M, \tag{2.4}$$

where $u_i \overset{\text{iid}}{\sim} N\left(0, \sigma_u^2\right)$, and $h_{ij}\left| j \in s_i \right. \overset{\text{iid}}{\sim} N\left(0, \sigma_h^2\right)$. They assumed that the unit design weight $w_{j|i} = \pi_{j|i}^{-1}$ is random with conditional expectation

$$\begin{aligned} E_{si}\left(w_{j|i}\left|\mathbf{x}_{ij}, y_{ij}, v_i\right.\right) &= E_{si}\left(w_{j|i}\left|\mathbf{x}_{ij}, y_{ij}\right.\right) \\ &= k_i\exp\left(\mathbf{x}_{ij}^T\mathbf{a} + dy_{ij}\right), \end{aligned} \tag{2.5}$$

where $\mathbf{a}$ and d are fixed unknown constants and

$$k_i = \frac{N_i}{n_i}\left\{\sum_{j=1}^{N_i}\exp\left(-\mathbf{x}_{ij}^T\mathbf{a} - dy_{ij}\right) / N_i\right\}.$$

The Pfeffermann and Sverchkov (2007) estimator of $\bar{Y}_i$ provides protection against informative sampling supposing that this assumption holds. The estimator is given by

$$\hat{\bar{Y}}_i^{\text{PS}} = \frac{1}{N_i}\left[\left(N_i - n_i\right)\hat{\mu}_{iu}^{\text{EBLUP}} + n_i\left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\alpha}}\right\} + \left(N_i - n_i\right)\hat{d}\hat{\sigma}_h^2\right], \tag{2.6}$$

where $\hat{\mu}_{iu}^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T\hat{\boldsymbol{\alpha}} + \hat{u}_i$ is the EBLUP estimator of $\mu_{iu} = \bar{\mathbf{X}}_i^T\boldsymbol{\alpha} + u_i$ under the sample model (2.4) and $\hat{d}$ is an estimator of d in the model (2.5) for the weights $w_{j|i}$. The last term in (2.6) corrects for any bias due to informative sampling under (2.5). Pfeffermann and Sverchkov (2007) obtained the estimator $\hat{d}$ of d in (2.5) by regressing the sampling weights $w_{j|i}$ on $k_i\exp\left(\mathbf{x}_{ij}^T\mathbf{a} + dy_{ij}\right)$. The coefficients k_i, $\mathbf{a}$ and d may be estimated by fitting the model (2.5) using the NLIN procedure in SAS or function `nls` in Splus. This involves iterative calculations and the initial values for $\mathbf{a}$ and d are obtained by regressing $\log\left(w_{j|i}\right)$ on $\mathbf{x}_{ij}$ and y_{ij}. Initial values for $\hat{k}_i$, $i = 1, \ldots, M$ are taken as $k_i = N_i / n_i$.

The Verret et al. (2015) estimator is obtained when the EBLUP theory is applied to model (1.2). Let $\mathbf{x}_{ij}^{\text{aug}} = \left(\mathbf{x}_{ij}^T, g\left(p_{j|i}\right)\right)^T$ be the vector $\mathbf{x}_{ij}$ augmented by the variable $g\left(p_{j|i}\right)$, $\bar{G}_i$ the area mean of the population values $g\left(p_{j|i}\right)$, and $\mu_{0i} = \bar{\mathbf{X}}_i^T\boldsymbol{\beta}_0 + \bar{G}_i\delta_0 + v_{0i}$. The EBLUP estimator of μ_{0i} is given by

$$\hat{\mu}_{0i}^{\text{EBLUP}} = \bar{\mathbf{X}}_i^T\hat{\boldsymbol{\beta}}_0 + \bar{G}_i\hat{\delta}_0 + \hat{v}_{0i} = \hat{\gamma}_{0i}\bar{y}_i + \left(\bar{\mathbf{X}}_i - \hat{\gamma}_{0i}\bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}_0 + \left(\bar{G}_i - \hat{\gamma}_{0i}\bar{g}_i\right)\hat{\delta}_0, \tag{2.7}$$

where $\hat{\gamma}_{0i} = \hat{\sigma}_{0v}^2 \big/ \left(\hat{\sigma}_{0v}^2 + \hat{\sigma}_{0e}^2 / n_i\right)$, $\bar{g}_i = \sum_{j=1}^{n_i} g\left(p_{j|i}\right) \big/ n_i$ and $\hat{v}_{0i} = \hat{\gamma}_{0i}\left(\bar{y}_i - \bar{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}_0 - \bar{g}_i\hat{\delta}_0\right)$. The parameters, $(\boldsymbol{\beta}_0, \delta_0)$ are estimated by

$$\left(\hat{\boldsymbol{\beta}}_0^T, \hat{\delta}_0\right)^T = \left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\mathbf{x}_{ij}^{\text{aug}}\left(\mathbf{x}_{ij}^{\text{aug}} - \hat{\gamma}_{0i}\bar{\mathbf{x}}_i^{\text{aug}}\right)^T\right\}^{-1}\left\{\sum_{i=1}^{M}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij}^{\text{aug}} - \hat{\gamma}_{0i}\bar{\mathbf{x}}_i^{\text{aug}}\right)y_{ij}\right\}, \tag{2.8}$$

with $\bar{\mathbf{x}}_i^{\text{aug}} = \sum_{j=1}^{n_i}\mathbf{x}_{ij}^{\text{aug}} \Big/ n_i = \left(\bar{\mathbf{x}}_i^T, \bar{g}_i\right)^T$. The model parameters $\left(\hat{\sigma}_{0v}^2, \hat{\sigma}_{0e}^2\right)$ are estimated by HFC or REML method. The estimator of the area mean $\bar{Y}_i$, denoted $\hat{\bar{Y}}_i^{\text{VRH}}$, may be written in terms of $\hat{\mu}_{0i}^{\text{EBLUP}}$ as

$$\hat{\bar{Y}}_i^{\text{VRH}} = \frac{1}{N_i}\left[\left(N_i - n_i\right)\hat{\mu}_{0i}^{\text{EBLUP}} + n_i\left\{\bar{y}_i + \left(\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i\right)^T\hat{\boldsymbol{\beta}}_0 + \left(\bar{G}_i - \bar{g}_i\right)^T\hat{\delta}_0\right\}\right]. \tag{2.9}$$

3 The local polynomial estimator

3.1 The estimation of a small area mean

The objective is to estimate the mean $\bar{Y}_i$ for small area U_i for $i = 1, \ldots, M$. Splitting the population U_i into observed units in the sample, s_i of size n_i, and non-observed units in the non-sampled portion, $\bar{s}_i = U_i / s_i$ of size $N_i - n_i$, we can express $\bar{Y}_i$ as

$$\bar{Y}_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in\bar{s}_i} y_{ij}\right). \tag{3.1}$$

Given that we do not know the y values for the non-observed units in sets $\bar{s}_i$ for $i = 1, \ldots, M$, we need to estimate them. Denoting as $\hat{y}_{ij}$ the estimator of y_{ij} for such units, the resulting estimator of the mean $\bar{Y}_i$ is

$$\hat{\bar{Y}}_i = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij} + \sum_{j\in\bar{s}_i}\hat{y}_{ij}\right). \tag{3.2}$$

We obtain estimators $\hat{y}_{ij}$ of y_{ij}, for $j \in \bar{s}_i$, based on an augmented model that includes an unknown smooth function of the selection probabilities $p_{j|i}$, denoted $m_0\left(p_{j|i}\right)$. The proposed augmented semi-parametric sample model is given by

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T\boldsymbol{\beta}_1 + m_0\left(p_{j|i}\right) + v_{1i} + e_{1ij},\ j = 1, \ldots, n_i;\ i = 1, \ldots, M, \tag{3.3}$$

where $v_{1i} \overset{\text{iid}}{\sim} N(0, \sigma_{1v}^2)$ and independent of $e_{1ij} \overset{\text{iid}}{\sim} N(0, \sigma_{1e}^2)$. The vector $\tilde{\mathbf{x}}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$ in model (3.3) represents the covariates $\mathbf{x}_{ij}$ without a constant (i.e., the intercept) and $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^T$ a vector of fixed effects. Model (3.3) is semi-parametric as the response variable y_{ij} depends linearly on the vector of auxiliary variables, $\tilde{\mathbf{x}}_{ij}$, and the probability of selection $p_{j|i}$ enters non-parametrically through the smooth function $m_0(\cdot)$.

We assume that model (3.3) has a similar covariance structure with the one associated with model (1.2): the small area effects v_{1i} and random errors e_{1ij} are iid, normally distributed and independently of one another. However, the semi-parametric model (3.3) is more flexible than the parametric model (1.2), as it does not force the function $m_0(p_{j|i})$ to be of a specific form. There is a disadvantage to this set-up. Since model (3.3) is not a linear mixed model, the general EBLUP theory given in Section 2 cannot be applied directly to obtain estimators of $m_0(p_{j|i})$, $\boldsymbol{\beta}_1$ and v_{1i}. Consequently, we propose to estimate (3.3) by combining the EBLUP theory for linear mixed models and the local polynomial technique (Fan and Gijbels, 1996).

We estimate (3.3) in three steps. In the first step, we obtain estimates of $m_0(p_{j|i})$, $\hat{m}_0(p_{j|i})$, $j = 1, \ldots, N_i$, $i = 1, \ldots, M$, for all units in the population. These estimates are local in character as they are based on the local polynomial technique. Estimates $\hat{m}_0(p_{j|i})$, $j \in s_i$ for the observed units are then used in the second step to obtain global estimators of $\boldsymbol{\beta}_1$ and v_{1i}, $i = 1, \ldots, M$. We denote these estimators as $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$, $i = 1, \ldots, M$. Finally, in the third step, we use the local estimators $\hat{m}_0(p_{j|i})$ for the unobserved units, obtained in the first step, and the global estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$ obtained in the second step, to estimate y_{ij} for $j \in \overline{s}_i$ and $i = 1, \ldots, M$. The resulting estimators of y_{ij}, denoted as $\hat{y}_{ij}$, are

$$\hat{y}_{ij} = \tilde{\mathbf{x}}_{ij}^T \hat{\boldsymbol{\beta}}_{\text{glo},1} + \hat{m}_0(p_{j|i}) + \hat{v}_{\text{glo},1i}, \quad j \in \overline{s}_i. \tag{3.4}$$

The $\hat{y}_{ij}$'s are incorporated into equation (3.2) to obtain the estimator of the small area mean $\hat{\bar{Y}}_i$.

We now proceed to describe the first step in more detail. Following Ruppert and Matteson (2015), we estimate the values of the unknown function $m_0(p_{l|k})$ for all units $l \in U_k$ and small areas k, with $k = 1, \ldots, M$, by using local polynomial regression. Local polynomial regression is based on the principle that a smooth function can be approximated locally by a low-degree polynomial. We approximate $m_0(p_{j|i})$ in model (3.3) by a q^{th}-degree polynomial, say $m_1(p_{j|i})$, using a Taylor expansion around $p_{l|k}$. The approximation is given by

$$m_1(p_{j|i}) = m_0(p_{l|k}) + \sum_{a=1}^{q} \frac{1}{a!} m_0(p_{l|k})^{(a)} (p_{j|i} - p_{l|k})^a, \quad j \in s_i;\ i = 1, \ldots, M, \tag{3.5}$$

where $m_0(p_{l|k})^{(a)}$ is the a^{th} derivative of $m_0(p_{j|i})$ evaluated at $p_{l|k}$. The function $m_1(p_{j|i})$ depends on $l \in U_k$, but we suppress this dependence to simplify the notation.

For each point $p_{l|k}$, $l \in U_k$; $k = 1, \ldots, M$, in model (3.3) we replace $m_0(p_{j|i})$ by its approximation $m_1(p_{j|i})$ given by (3.5). The resulting model is given by

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}_1 + m_0\left(p_{l|k}\right) + \sum_{a=1}^{q} \frac{1}{a!} m_0\left(p_{l|k}\right)^{(a)} \left(p_{j|i} - p_{l|k}\right)^a + v_{1i} + e_{1ij}, \quad j \in s_i;\ i = 1, \ldots, M. \quad (3.6)$$

Model (3.6) is an approximate local model for (3.3) depending on the point $l \in U_k$ of the population. Estimates of $\boldsymbol{\beta}_1$ and v_{1i} based on (3.6) will be denoted by $\hat{\boldsymbol{\beta}}_{\text{loc},1}$ and $\hat{v}_{\text{loc},1i}$. Notice that (3.6) allows the estimation of $m_0\left(p_{l|k}\right)$, the value of the smooth function $m_0(\cdot)$ at a point $p_{l|k}$. We express (3.6) as

$$y_{ij} = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}_1 + u_0 + \sum_{a=1}^{q} u_a \left(p_{j|i} - p_{l|k}\right)^a + v_{1i} + e_{1ij}:\ j \in s_i;\ i = 1, \ldots, M, \quad (3.7)$$

where $u_a = m_0\left(p_{l|k}\right)^{(a)} \big/ a!$ for $a = 0, \ldots, q$. Model (3.7) is a linear mixed model with fixed parameters $\left(\boldsymbol{\beta}_1,\ u_0, \ldots, u_q\right)$ and random small area effects $v_{1i},\ i = 1, \ldots, M$.

Let $\hat{u}_0$ be an estimator of u_0 obtained by fitting model (3.7). An approximate estimator of $m_0\left(p_{l|k}\right) = u_0$ is given by $\hat{m}_0\left(p_{l|k}\right) = \hat{u}_0$. Since we require estimators of $m_0\left(p_{l|k}\right)$ for $l \in U_k$ and $k = 1, \ldots, M$, we use $N = \sum_{i=1}^{M} N_i$ models (3.7). As pointed out by an Associate Editor, if N is large, estimating the values of $m_0(\cdot)$ for all points in the population can be computationally intensive.

It is more convenient to work with matrix notation. To this end, we define $\mathbf{y}_i = \left(y_{i1}, \ldots, y_{in_i}\right)^T$, $\tilde{\mathbf{X}}_i = \left(\tilde{\mathbf{x}}_{i1}^T, \ldots, \tilde{\mathbf{x}}_{in_i}^T\right)^T$, $\mathbf{m}_{0,i} = \left(m_0\left(p_{1|i}\right), \ldots, m_0\left(p_{n_i|i}\right)\right)^T$, $\mathbf{v}_1 = \left(v_{11}, \ldots, v_{1M}\right)^T$ and $\mathbf{e}_{1i} = \left(e_{1i1}, \ldots, e_{1in_i}\right)^T$. Model (3.3) can be expressed in a matrix form by stacking the observations, and the resulting equation is

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{m}_0 + \mathbf{Z}\mathbf{v}_1 + \mathbf{e}_1, \quad (3.8)$$

where $\mathbf{y} = \text{col}_{1 \le i \le M}\left(\mathbf{y}_i\right)$, $\tilde{\mathbf{X}} = \text{col}_{1 \le i \le M}\left(\tilde{\mathbf{X}}_i\right)$, $\mathbf{m}_0 = \text{col}_{1 \le i \le M}\left(\mathbf{m}_{0,i}\right)$, $\mathbf{Z} = \text{diag}_{1 \le i \le M}\left\{\mathbf{1}_{n_i}\right\}$ and $\mathbf{e}_1 = \text{col}_{1 \le i \le M}\left(\mathbf{e}_{1i}\right)$.

For unit l in small area U_k, we define the $n \times (q + 1)$ matrix:

$$\mathbf{Q} = \begin{pmatrix} 1 & \left(p_{1|1} - p_{l|k}\right) & \cdots & \left(p_{1|1} - p_{l|k}\right)^q \\ \vdots & \vdots & \cdots & \vdots \\ 1 & \left(p_{n_M|M} - p_{l|k}\right) & \cdots & \left(p_{n_M|M} - p_{l|k}\right)^q \end{pmatrix},$$

where $n = \sum_{i=1}^{M} n_i$ is the total sample size. Let $\mathbf{u} = \left(m_0\left(p_{l|k}\right), m_0^{(1)}\left(p_{l|k}\right)/1!, \ldots, m_0^{(q)}\left(p_{l|k}\right)/q!\right)^T$ represent the vector of derivatives of the function $m_0(\cdot)$ evaluated at $p_{l|k}$. The terms $\mathbf{Q}$ and $\mathbf{u}$ depend on the unit $l \in U_k$ where the localization is realized. We omitted their dependence on the unit l from small area U_k in order not to burden the notation. We define vector $\mathbf{m}_1$ obtained by stacking the n values of the function $m_1(\cdot)$ defined by (3.5). That is, $\mathbf{m}_1 = \text{col}_{1 \le i \le M}(\mathbf{m}_{1,i})$ with $\mathbf{m}_{1,i} = (m_1(p_{1|i}), \ldots, m_1(p_{n_i|i}))^T$. This allows to approximate $\mathbf{m}_0$ by $\mathbf{m}_0 \approx \mathbf{m}_1$. The vector $\mathbf{m}_1$ is given by $\mathbf{m}_1 = \mathbf{Q}\mathbf{u}$. It then follows that an approximation to (3.8) in a neighbourhood of $l \in U_k$ is

$$\mathbf{y} = \tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{Q}\mathbf{u} + \mathbf{Z}\mathbf{v}_1 + \mathbf{e}_1. \quad (3.9)$$

Equations (3.8) and (3.9) are the matrix form equivalents of equations (3.3) and (3.7), respectively. The matrix $\tilde{\mathbf{X}}$ in (3.9) does not include the constant term that represents the intercept, because this term is

already included in $\mathbf{Q}$. Equation (3.9) is a standard linear mixed effects model with fixed parameters $\boldsymbol{\beta}_{\text{fixed}} = (\boldsymbol{\beta}_1^T, \mathbf{u}^T)^T$ and random small area effects $\mathbf{v}_1$. We denote by $V(\mathbf{v}_1) = \mathbf{G} = \sigma_{1v}^2 \mathbf{I}_M$, $V(\mathbf{e}_{1i}) = \mathbf{R}_i = \sigma_{1e}^2 \mathbf{I}_{n_i}$ and $V(\mathbf{e}_1) = \mathbf{R} = \text{diag}_{1 \le i \le M}\{\mathbf{R}_i\}$ as the respective covariance matrices of $\mathbf{v}_1$, $\mathbf{e}_{1i}$ and $\mathbf{e}_1$. The covariance matrix of $\mathbf{y}_i$ is given by $V(\mathbf{y}_i) = \mathbf{V}_i = \sigma_{1v}^2 \mathbf{J}_{n_i} + \sigma_{1e}^2 \mathbf{I}_{n_i}$. The matrices $\mathbf{I}_M$ and $\mathbf{I}_{n_i}$ are the identity matrices of order M and n_i respectively, whereas $\mathbf{J}_{n_i}$ is the square matrix of order n_i with all its elements equal to 1. It follows that $V(\mathbf{y}) = \mathbf{V} = \text{diag}_{1 \le i \le M}\{\mathbf{V}_i\}$.

Assume that $\mathbf{V}$ is known and that $\mathbf{v}_1$ and $\mathbf{e}_1$ are normally distributed. Using classical EBLUP theory, estimators of $\boldsymbol{\beta}_{\text{fixed}}$ and $\mathbf{v}_1$ can be obtained by minimizing

$$\Phi = (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Q}\mathbf{u} - \mathbf{Z}\mathbf{v}_1)^T \mathbf{R}^{-1} (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Q}\mathbf{u} - \mathbf{Z}\mathbf{v}_1) + \mathbf{v}_1^T \mathbf{G}^{-1} \mathbf{v}_1.$$

Note that all the observations that are included in Φ are equally weighted. However, we need to modify Φ to be in line with how local polynomial estimation is carried out. To this end, referring back to equation (3.7), we estimate its parameters by associating kernel weights $K\left((p_{j|i} - p_{l|k})/h\right)/h$ to each sampled unit $j \in s_i$; $i = 1, \ldots, M$. These kernel weights are chosen so as to give a larger weight to the sample points that are close to $l \in U_k$, and a smaller weight to those that are further away. The weight $K(\cdot)$ is a probability density function and h is a bandwidth controlling the size of the local neighbourhood. We explain in Section 3.2 how an optimal bandwidth can be obtained. Let $\mathbf{W}$ be the $n \times n$ diagonal matrix of kernel weights given by

$$\mathbf{W} = \underset{\substack{1 \le j \le n_i \\ 1 \le i \le M}}{\text{diag}} \left\{ \frac{1}{h} K\left(\frac{p_{j|i} - p_{l|k}}{h} \right) \right\}.$$

The matrix $\mathbf{W}$ depends on unit l from small area U_k and the bandwidth h. We do not include the subscripts $l \in U_k$ and h in the definition of the matrix $\mathbf{W}$, in order not to burden notation. Following Wu and Zhang (2002), the incorporation of the kernel weights in Φ lead us to minimize Φ_W where

$$\Phi_W = (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Q}\mathbf{u} - \mathbf{Z}\mathbf{v}_1)^T \mathbf{W}^{1/2} \mathbf{R}^{-1} \mathbf{W}^{1/2} (\mathbf{y} - \tilde{\mathbf{X}}\boldsymbol{\beta}_1 - \mathbf{Q}\mathbf{u} - \mathbf{Z}\mathbf{v}_1) + \mathbf{v}_1^T \mathbf{G}^{-1} \mathbf{v}_1,$$

and $\mathbf{W}^{1/2}$ represents the square root of the matrix $\mathbf{W}$.

Estimating the parameters of (3.9) by minimizing Φ_W is equivalent to estimating those given by

$$\mathbf{W}^{1/2}\mathbf{y} = \mathbf{W}^{1/2}\tilde{\mathbf{X}}\boldsymbol{\beta}_1 + \mathbf{W}^{1/2}\mathbf{Q}\mathbf{u} + \mathbf{W}^{1/2}\mathbf{Z}\mathbf{v}_1 + \mathbf{e}_1. \tag{3.10}$$

The weighted EBLUP based on (3.9) with the matrix of weights given by $\mathbf{W}$ corresponds to a classical EBLUP obtained from model (3.10). Define $\mathbf{y}_w = \mathbf{W}^{1/2}\mathbf{y}$, $\mathbf{X}_w = [\mathbf{W}^{1/2}\tilde{\mathbf{X}}, \mathbf{W}^{1/2}\mathbf{Q}]$ and $\mathbf{Z}_w = \mathbf{W}^{1/2}\mathbf{Z}$. Equation (3.10) can be rewritten as

$$\mathbf{y}_w = \mathbf{X}_w\boldsymbol{\beta}_{\text{fixed}} + \mathbf{Z}_w\mathbf{v}_1 + \mathbf{e}_1. \tag{3.11}$$

Let $\hat{\boldsymbol{\beta}}_{\text{loc, fixed}} = (\hat{\boldsymbol{\beta}}_{\text{loc},1}^T, \hat{\mathbf{u}}^T)^T$ and $\hat{\mathbf{v}}_{\text{loc},1} = (\hat{v}_{\text{loc},11}, \ldots, \hat{v}_{\text{loc},1M})^T$ be the EBLUP estimators of the fixed and random effects of (3.11). The estimators $\hat{\boldsymbol{\beta}}_{\text{loc, fixed}}$ and $\hat{\mathbf{v}}_{\text{loc},1}$ are based on local estimators of the variance components $(\sigma_{1v}^2, \sigma_{1e}^2)$. The estimators of these components, denoted as $(\hat{\sigma}_{\text{loc},1v}^2, \hat{\sigma}_{\text{loc},1e}^2)$, are

obtained using HFC or REML methods under model (3.11). Given that $\mathbf{u} = (m_0(p_{l|k}), m_0^{(1)}(p_{l|k})/1!, \ldots, m_0^{(q)}(p_{l|k})/q!)^T$, an estimator $\hat{m}_0(p_{l|k})$ of $m_0(p_{l|k})$ is the first component $\hat{u}_0$ of $\hat{\mathbf{u}}$.

Notice that $\hat{\boldsymbol{\beta}}_{\text{loc},1}$, $\hat{m}_0(p_{l|k})$ and $\hat{v}_{\text{loc},1k}$ could be used to obtain local estimates $\hat{y}_{\text{loc},kl}$ for the unknown value y_{kl}, where $\hat{y}_{\text{loc},kl} = \tilde{\mathbf{x}}_{kl}^T \hat{\boldsymbol{\beta}}_{\text{loc},1} + \hat{m}_0(p_{l|k}) + \hat{v}_{\text{loc},1k}$ for $l \in \overline{s}_k$. However, a referee pointed out that, in practice, this methodology would not likely to be well behaved because it requires a strong balance of the small areas across the range of the probabilities $p_{l|k}$. If this balance is not respected, the resulting estimation would suffer severely from this localization. As a consequence, we opted for a global estimation of $\boldsymbol{\beta}_1$ and $\mathbf{v}_1$.

We now explain the second step of our procedure. Parameters $\boldsymbol{\beta}_1$ and $\mathbf{v}_1$ can be estimated globally based on the estimations $\hat{m}_0(p_{j|i})$ and the auxiliary data $\tilde{\mathbf{x}}_{ij}$ associated with the sample units. For $j \in s_i$ and $i = 1, \ldots, M$, define a new variable, say ξ, as

$$\xi_{ij} = y_{ij} - \hat{m}_0(p_{j|i}),\ j \in s_i;\ i = 1, \ldots, M.$$

The n values ξ_{ij} represent the differences between the observed y_{ij}'s and their local estimators $\hat{m}_0(p_{j|i})$. Using model (3.3), ξ satisfies the following model

$$\xi_{ij} = \tilde{\mathbf{x}}_{ij}^T \boldsymbol{\beta}_{\text{glo},1} + v_{\text{glo},1i} + e_{\text{glo},1ij},\ j \in s_i;\ i = 1, \ldots, M, \tag{3.12}$$

where $v_{\text{glo},1i} \sim N(0, \sigma^2_{\text{glo},1v})$ and $e_{\text{glo},1ij} \sim N(0, \sigma^2_{\text{glo},1e})$. The subscript glo indicates that (3.12) is a global model.

Given that (3.12) represents a parametric linear mixed effects model, we can use the classical (unweighted) EBLUP to estimate its parameters. Let $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$ be the respective empirical best linear unbiased estimators of $\boldsymbol{\beta}_{\text{glo},1}$ and $v_{\text{glo},1i}$. Let $(\hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ be the estimators of the variance components $(\sigma^2_{\text{glo},1v}, \sigma^2_{\text{glo},1e})$ where HFC or REML can be used to estimate these parameters. We estimate $(\boldsymbol{\beta}_1, v_{1i}, \sigma^2_{1v}, \sigma^2_{1e})$ of model (3.3) by $(\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{v}_{\text{glo},1i}, \hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ using model (3.12). The global estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$, $\hat{v}_{\text{glo},1i}$ and $(\hat{\sigma}^2_{\text{glo},1v}, \hat{\sigma}^2_{\text{glo},1e})$ are free of bias caused by informative sampling design because ξ_{ij} is no longer related to the $p_{j|i}$'s after conditioning on $\mathbf{x}_{ij}$.

The third step estimates the non observed y_{ij} values, for $j \in \overline{s}_i$ and $i = 1, \ldots, M$, by plugging into equation (3.4): i. the local estimators $\hat{m}_0(p_{j|i})$ for $j \in \overline{s}_i$, obtained in the first step, and ii. the global estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$ and $\hat{v}_{\text{glo},1i}$ obtained in the second step. The resulting $\hat{y}_{ij}$'s, for $j \in \overline{s}_i$, are inserted into (3.2) to compute the estimator $\hat{\bar{Y}}_i$. Note that $\hat{\bar{Y}}_i$ requires $\tilde{\mathbf{x}}_{ij}$ and $p_{j|i}$ are known for all the units of the population. A referee pointed out that, in practice, this assumption may limit the applicability of the proposed procedure. This could be remedied if National Statistical Offices provided access to the selection probabilities of all units, as they may be needed in applications such as this one.

3.2 Bandwidth selection

Local polynomials require the specification of the kernel $K(\cdot)$, the order of the polynomial fit q, as well as the bandwidth h. Fan and Gijbels (1996) state that values of q larger than 1 do not bring a

significant improvement as compared to the linear fit $(q = 1)$. Fan and Gijbels (1996) also state that the choice of h is far more important than the degree of the polynomial. In what follows, we use a normal density kernel, and chose q equal to one, as this leads to satisfactory results for most applications.

The optimal h is determined using the cross-validation method (CV). For a given h, compute the estimator of y_{ij} given by (3.4) using the sample that remains after the j^{th} unit has been removed from s_i. Denoting the resulting estimator of y_{ij} as $\tilde{y}_{ij}$, we follow Wu and Zhang (2002) and define the CV criterion as

$$\text{CV}(h) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{n_i}\sum_{j\in s_i}\left(y_{ij} - \tilde{y}_{ij}\right)^2.$$

The term $1/n_i$ takes into account the number of observations within small area U_i. The optimal bandwidth h_{opt} is obtained by minimizing the $\text{CV}(h)$. Given h_{opt}, the local polynomial estimator of the small area mean $\bar{Y}_i$ given by (3.2) is denoted as $\hat{\bar{Y}}_i^{\text{LP}}$.

4 MSE estimation based on the bootstrap

The MSE estimation of small area estimators is a challenging problem even in the case of classical EBLUP estimators. The general EBLUP theory provides a closed form approximation to $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ based on a linearization method. Using this approximation, an estimator for $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ can be obtained (see Prasad and Rao, 1990 for details). Verret et al. (2015) used the closed form approximation to estimate the mean squared error estimator for $\hat{\bar{Y}}_i^{\text{VRH}}$ given in (2.9). This was possible because estimator $\hat{\bar{Y}}_i^{\text{VRH}}$ is a standard EBLUP obtained under a linear mixed model that includes the additional known variable $g\left(p_{j|i}\right)$. No new theory is needed to estimate the MSE of $\hat{\bar{Y}}_i^{\text{VRH}}$. In our case, given the repeated local estimation of model (3.6), it is not possible to obtain a closed-form approximation to the mean squared error of $\hat{\bar{Y}}_i^{\text{LP}}$, $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$, nor for its estimator $\text{mse}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$. We used two variants of the bootstrap procedure to estimate the MSE of the small area estimators that we have discussed so far. For estimating the MSE of $\hat{\bar{Y}}_i^{\text{EBLUP}}$, we used an *unconditional* bootstrap, whereas for $\hat{\bar{Y}}_i^{\text{LP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$, we used a *conditional* bootstrap. We proceed to describe how each bootstrap type is computed.

We first describe the unconditional bootstrap. This is a variant of the parametric bootstrap of Hall and Maiti (2006), proposed by González-Manteiga, Lombardia, Molina, Morales and Santamaria (2008). This procedure can be used for estimating the MSE of $\hat{\bar{Y}}_i^{\text{EBLUP}}$ that is based on model (1.1) because the estimates of the various parameters in model (1.1) do not depend on the selection probabilities $p_{j|i}$: $j \in s_i$; $i = 1, \ldots, M$. The y values are predicted by generating $v_i^* \sim N\left(0, \hat{\sigma}_v^2\right)$ and $e_{ij}^* \sim N\left(0, \hat{\sigma}_e^2\right)$, where $\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ are the HFC or REML estimators of $\left(\sigma_v^2, \sigma_e^2\right)$. Using the EBLUP estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, bootstrap values of y_{ij} are obtained as

$$y_{ij}^* = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + v_i^* + e_{ij}^*,\ j \in U_i;\ i = 1, \ldots, M. \tag{4.1}$$

The bootstrap version of the target parameter $\bar{Y}_i$ is computed as $\bar{Y}_i^* = N_i^{-1}\sum_{j=1}^{N_i} y_{ij}^*$. The bootstrap version of the EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is given by

$$\hat{\bar{Y}}_i^{\text{EBLUP*}} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{ij}^* + \sum_{j\in \bar{s}_i} \hat{y}_{ij}^*\right),$$

where $\hat{y}_{ij}^* = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}^* + \hat{v}_i^*$ and $\left(\hat{\boldsymbol{\beta}}^*, \hat{v}_i^*\right)$ are the EBLUP estimators of $(\boldsymbol{\beta}, v_i)$ that are based on $\left(y_{ij}^*, \mathbf{x}_{ij}\right)$, $j \in s_i$, for $i = 1, \ldots, M$. Repeating the above procedure B times, the bootstrap estimator of $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ is

$$\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_i^{\text{EBLUP*}}(b) - \bar{Y}_i^*(b)\right)^2, \tag{4.2}$$

where $\hat{\bar{Y}}_i^{\text{EBLUP*}}(b)$ and $\bar{Y}_i^*(b)$ are the values of $\hat{\bar{Y}}_i^{\text{EBLUP*}}$ and $\bar{Y}_i^*$ for the b^{th} bootstrap replicate. Since the estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_v^2, \hat{\sigma}_e^2\right)$ are severely biased due to the informative sampling design, we expect that $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$ will be a biased estimator of $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{EBLUP}}\right)$. This is because it is based on the population model (1.1), and that this model does not hold for the sample.

We now turn to the estimation of $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$ via the conditional bootstrap. Recall that $\hat{\bar{Y}}_i^{\text{LP}}$ is based on the augmented model (3.3). It is therefore natural to use this model when we estimate the precision of the local polynomial estimator. It is not possible to use the parametric unconditional bootstrap as it would require the generation of bootstrap values $\left(y_{ij}^*, p_{j|i}^*\right)$ for both y_{ij} and $p_{j|i}$, and this would imply that we would need to know how the y_{ij}'s are related to the selection probabilities $p_{j|i}$. As the Associate Editor pointed out, the exact relationship between y_{ij} and $p_{j|i}$ is not known in practice. We therefore opted to keep the selection probabilities $p_{j|i}$ associated with the initial sample, and generate bootstrap values only for the response variable y_{ij}. The resulting bootstrap is conditional on $p_{j|i}$, $j \in U_i$; $i = 1, \ldots, M$, and it is for this reason that we label it as *conditional parametric bootstrap*. It has been used by Rao, Sinha and Dumitrescu (2014), and more recently by Chatrchi (2018) to estimate the MSE under a penalized spline mixed model.

In our context, for estimating $\text{MSE}\left(\hat{\bar{Y}}_i^{\text{LP}}\right)$, we proceed as follows. We generate $v_{1i}^* \sim N\left(0, \hat{\sigma}_{\text{glo},1v}^2\right)$ and $e_{1ij}^* \sim N\left(0, \hat{\sigma}_{\text{glo},1e}^2\right)$, and obtain the bootstrap responses

$$y_{1ij}^* = \tilde{\mathbf{x}}_{ij}^T\hat{\boldsymbol{\beta}}_{\text{glo},1} + \hat{m}_0\left(p_{j|i}\right) + v_{1i}^* + e_{1ij}^*, \; j \in U_i; \; i = 1, \ldots, M. \tag{4.3}$$

The $\hat{m}_0\left(p_{j|i}\right)$'s were estimated using the local model (3.6). The triplet $\left(\hat{\boldsymbol{\beta}}_{\text{glo},1}, \hat{\sigma}_{\text{glo},1v}^2, \hat{\sigma}_{\text{glo},1e}^2\right)$ was estimated using the global model (3.12) and the sample data $\left(y_{ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$. The population bootstrap mean is $\bar{Y}_{1i}^* = N_i^{-1}\sum_{j=1}^{N_i} y_{1ij}^*$. Let $\hat{\boldsymbol{\beta}}_{\text{glo},1}^*$, $\hat{m}_0^*\left(p_{j|i}\right)$ and $\hat{v}_{\text{glo},1i}^*$ be bootstrap versions of estimators $\hat{\boldsymbol{\beta}}_{\text{glo},1}$, $\hat{m}_0\left(p_{j|i}\right)$ and $\hat{v}_{\text{glo},1i}$, that are based on bootstrap data $\left(y_{1ij}^*, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$ and the h_{opt} obtained with the original data set $\left(y_{ij}, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$. We did not re-compute the optimal h_{opt}^* associated with $\left(y_{1ij}^*, \tilde{\mathbf{x}}_{ij}, p_{j|i}\right)$, $j \in s_i$; $i = 1, \ldots, M$, as it would result in far too many computations in the Monte Carlo study. The bootstrap procedure is therefore conditional on $p_{j|i}$, $j \in U_i$; $i = 1, \ldots, M$ and h_{opt} obtained with the initial sample. Given that $\bar{s}_i$ is the set of non-sampled units in area i, the predicted bootstrap values $\hat{y}_{1ij}^*$ for $j \in \bar{s}_i$, are obtained as

$$\hat{y}_{1ij}^* = \tilde{\mathbf{x}}_{ij}^T\hat{\boldsymbol{\beta}}_{\text{glo},1}^* + \hat{m}_0^*\left(p_{j|i}\right) + \hat{v}_{\text{glo},1i}^*. \tag{4.4}$$

The resulting estimator of $\bar{Y}_{1i}^{*}$ is

$$\hat{\bar{Y}}_{1i}^{*} = \frac{1}{N_i}\left(\sum_{j\in s_i} y_{1ij}^{*} + \sum_{j\in \bar{s}_i} \hat{y}_{1ij}^{*}\right).$$

Repeating the above procedure B times, the conditional bootstrap estimator of MSE of the local polynomial estimator of $\bar{Y}_i$ is given by

$$\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i^{\text{LP}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\bar{Y}}_{1i}^{*}(b) - \bar{Y}_{1i}^{*}(b)\right)^2, \tag{4.5}$$

where $\hat{\bar{Y}}_{1i}^{*}(b)$ and $\bar{Y}_{1i}^{*}(b)$ are the values of $\hat{\bar{Y}}_{1i}^{*}$ and $\bar{Y}_{1i}^{*}$ for the b^{th} bootstrap replicate.

The conditional bootstrap can also be used for estimating the mean squared error of an EBLUP estimator, $\hat{\bar{Y}}_i^{\text{VRH}}$, based on the augmented model (1.2) proposed by Verret et al. (2015). We included this procedure in the simulation given in Section 5, to get an idea of how the resulting MSE estimators compare to those obtained for $\hat{\bar{Y}}_i^{\text{LP}}$. The steps for obtaining the $\text{mse}\left(\hat{\bar{Y}}_i^{\text{VRH}}\right)$ are similar to those used for obtaining the mse of the local polynomial estimator $\hat{\bar{Y}}_i^{\text{LP}}$. In this case, bootstrap values for the responses y_{ij} are based on the augmented model (1.2) and the estimators $\left(\hat{\boldsymbol{\beta}}_0, \hat{\delta}_0\right)$ and $\left(\hat{\sigma}_{0v}^2, \hat{\sigma}_{0e}^2\right)$ obtained when the classical EBLUP theory is used with the sample data $\left(y_{ij}, \mathbf{x}_{ij}, g\left(p_{j|i}\right)\right)$, $j \in s_i$; $i = 1, \ldots, M$.

5 Simulation study

The set-up of the simulation study follows the one used in Verret et al. (2015). We considered a population with $M = 15$ small areas and $N_i = 15$ units within each small area. The relatively small number of small areas and units within areas were chosen so as to alleviate the computational burden. We used a single auxiliary variable x. The population x-values were generated from a gamma distribution with mean 10 and variance 50. The population y_{ij}-values were generated by the following model

$$y_{ij} = 4 + x_{ij} + v_i + e_{ij};\ i = 1, \ldots, 15;\ j = 1, \ldots, 15, \tag{5.1}$$

where $v_i \overset{\text{iid}}{\sim} N\left(0, \sigma_v^2\right)$ and $e_{ij} \overset{\text{iid}}{\sim} N\left(0, \sigma_e^2\right)$ with $\sigma_v^2 = 0.5$ and $\sigma_e^2 = 2$.

We considered a single sample size, $n_i = 3$, within a small area. We used Conditional Poisson Sampling (CPS) to select unequal probability samples within the small areas, with probabilities proportional to specified sizes c_{ij} (see Tillé, 2006, Chapter 5). We considered two different choices of the sizes c_{ij} in the simulation study. The first choice uses

$$c_{ij} = \exp\left[\frac{1}{3}\left\{-\frac{(v_i + e_{ij})}{\sigma_e} + \frac{\delta_{ij}}{5}\right\}\right], \tag{5.2}$$

where $\delta_{ij} \overset{\text{iid}}{\sim} N(0, 1)$. The size measures (5.2) are equivalent to those used by Pfeffermann and Sverchkov (2007) in their simulation study and satisfy the relationship (2.5) on the weights $w_{j|i} = \pi_{j|i}^{-1}$.

The second choice of size measures, following Asparouhov (2006), involves two different types of size measures: invariant (I) and non-invariant (NI). For the invariant case, c_{ij} is independent of v_i given $\mathbf{x}_{ij}$; otherwise, it is called non-invariant. Invariant size measures are given by

$$c_{ij} = \left(1 + \exp\left\{-\tau\left(\frac{1}{\alpha}e_{ij} + \sqrt{1 - \frac{1}{\alpha^2}}\, e_{ij}^*\right)\right\}\right)^{-1}. \quad (5.3)$$

Non-invariant size measures are taken as

$$c_{ij} = \left(1 + \exp\left\{-\tau\left[\frac{1}{\alpha}(v_i + e_{ij}) + \sqrt{1 - \frac{1}{\alpha^2}}\,(v_i^* + e_{ij}^*)\right]\right\}\right)^{-1}, \quad (5.4)$$

where the random pair (v_i^*, e_{ij}^*) is generated independently of (v_i, e_{ij}) using the same distributions as v_i and e_{ij}. These size measures were used by Asparouhov (2006). The coefficient τ controls for the variation of the weights and the value α controls the level of informativeness of the sampling design. We chose $\tau = 0.5$ and $\alpha = 1, 2, 3$ and ∞ corresponding to several levels of informativeness generated by c_{ij} in (5.3) and (5.4). Increasing α decreases informativeness, with $\alpha = \infty$ corresponding to non-informative sampling. If some of the $\pi_{j|i}$'s exceeded one, they were set to one, and the probabilities were recomputed for the remaining units.

5.1 Performance of the local polynomial estimator of $\bar{Y}_i$

We compared the bias and mean squared error of the estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$. The EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ based on (1.1) assumes that the sample model coincides with the population model, thereby ignoring the informativeness of the sampling design. We studied two versions of $\hat{\bar{Y}}_i^{\text{VRH}}$ investigated by Verret et al. (2015) for various choices of $g(\cdot)$ that account for informativeness. They are EBLUP estimators based on the augmented sample model (1.2). They are denoted as $\hat{\bar{Y}}_i^{\text{VRH1}}$ when $g(p_{j|i}) = p_{j|i}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$ when $g(p_{j|i}) = \log(p_{j|i})$. We report results only for these g functions, as they outperform others given in Verret et al. (2015). Finally, $\hat{\bar{Y}}_i^{\text{LP}}$ represents our new local polynomial estimator.

The bias and the mean squared error of the estimators were computed using $R = 1{,}000$ simulated samples selected under a design-model approach. For each run, $r = 1, \ldots, R$, we first generated the population y_{ij}-values under the population model (5.1) and computed $\bar{Y}_i^{(r)}$, the mean of the small area i in the r^{th} generated population. Samples of sizes $n_i = 3$ were then selected within the small areas using CPS with probabilities proportional to specified sizes $c_{ij}^{(r)}$ given by (5.2) for the Pfeffermann and Sverchkov (2007) (PS) size measures, and (5.3) and (5.4) corresponding to the invariant and non-invariant cases in the case of the Asparouhov (2006) (AP) size measures. From each simulated sample r $(r = 1, \ldots, R)$, the estimates $\hat{\bar{Y}}_i^{\text{EBLUP}(r)}$, $\hat{\bar{Y}}_i^{\text{VRH1}(r)}$, $\hat{\bar{Y}}_i^{\text{VRH2}(r)}$ and $\hat{\bar{Y}}_i^{\text{LP}(r)}$ were computed for each small area U_i. An optimal bandwidth $h_{\text{opt}}^{(r)}$ was found for $\hat{\bar{Y}}_i^{\text{LP}(r)}$ using the cross-validation criterion. A grid of the form (0.01, 0.02, 0.03, …, 0.15) covered the possible values for $h_{\text{opt}}^{(r)}$ in populations generated by (5.1).

For a given estimator of the small area mean $\bar{Y}_i$, we considered the following performance measures:

Average Absolute Bias

$$\overline{\text{AB}} = \frac{1}{M}\sum_{i=1}^{M} \text{AB}_i,$$

where

$$\mathrm{AB}_i = \left| \frac{1}{R} \sum_{r=1}^{R} \left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)} \right) \right|.$$

Average Root Mean Squared Error

$$\overline{\mathrm{RMSE}} = \frac{1}{M} \sum_{i=1}^{M} \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)} \right)^2}.$$

Table 5.1 reports on the average absolute bias $\left(\overline{\mathrm{AB}}\right)$ of estimators $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, $\hat{\bar{Y}}_i^{\mathrm{VRH1}}$, $\hat{\bar{Y}}_i^{\mathrm{VRH2}}$ and $\hat{\bar{Y}}_i^{\mathrm{LP}}$ under the PS size measures (5.2) and AP size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞.

Table 5.1
Average absolute bias $\left(\overline{\mathrm{AB}}\right)$ for the PS and AP size measures

Generation of $p_{j\|i}$ \ Estimator			$\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ without $g(p_{j\|i})$	$\hat{\bar{Y}}_i^{\mathrm{VRH1}}$ $g(p_{j\|i}) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\mathrm{VRH2}}$ $g(p_{j\|i}) = \log(p_{j\|i})$	$\hat{\bar{Y}}_i^{\mathrm{LP}}$ $m_0(p_{j\|i})$
PS			0.309	0.020	0.004	0.011
AP	$\alpha = 1$	I	0.431	0.002	0.036	0.004
		NI	0.425	0.010	0.035	0.005
	$\alpha = 2$	I	0.206	0.017	0.022	0.024
		NI	0.219	0.019	0.016	0.016
	$\alpha = 3$	I	0.139	0.005	0.012	0.033
		NI	0.137	0.008	0.013	0.019
	$\alpha = \infty$	I	0.008	0.008	0.008	0.026
		NI	0.006	0.006	0.006	0.021

As observed in Verret et al. (2015), the $\overline{\mathrm{AB}}$ of the EBLUP estimator $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ with just the auxiliary variable x, is quite a bit larger than those based on the augmented models ($p_{j|i}$ and $\log(p_{j|i})$), and the local polynomial method. This holds regardless of how the size measures have been generated (PS or AP). The $\overline{\mathrm{AB}}$ of $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ attains its highest value (0.431) when the design is very informative ($\alpha = 1$), and decreases as α increases. This observation also holds for the estimators based on the augmented models. The inclusion of $p_{j|i}$ or $\log(p_{j|i})$, as an augmenting variable, in the model results in small $\overline{\mathrm{AB}}$'s, with the highest being 0.036. Comparing the $\overline{\mathrm{AB}}$'s of the local polynomial estimator $\hat{\bar{Y}}_i^{\mathrm{LP}}$ to those associated with the VRH augmented models, we observe that they are comparable for $\alpha = 1$ and $\alpha = 2$, and slightly larger for $\alpha \geq 3$.

Table 5.2 reports the simulation results on the average root mean squared error $\left(\overline{\mathrm{RMSE}}\right)$ of the estimators for both the PS size measures (5.2) and the AP size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞. The EBLUP, $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$, based on model (1.1) without the augmenting variable $g(p_{j|i})$, has the largest $\overline{\mathrm{RMSE}}$'s (0.740 for I and 0.752 for NI) for the AP size measures corresponding to $\alpha = 1$, and 0.685 for the PS size measure. The $\overline{\mathrm{RMSE}}$ decreases as α increases: 0.608 for I and 0.610 for NI in the case of non-informative sampling ($\alpha = \infty$). The $\overline{\mathrm{RMSE}}$'s for $\hat{\bar{Y}}_i^{\mathrm{VRH1}}$, $\hat{\bar{Y}}_i^{\mathrm{VRH2}}$ and $\hat{\bar{Y}}_i^{\mathrm{LP}}$ are significantly smaller than those associated with $\hat{\bar{Y}}_i^{\mathrm{EBLUP}}$ when sampling is very informative ($\alpha = 1$) and for the PS size

measure. There are small differences in terms of $\overline{\text{RMSE}}$ between our non-parametric approach and the parametric approach in Verret et al. (2015).

Table 5.2
Average root mean squared error $\left(\overline{\text{RMSE}}\right)$ for the PS and AP size measures

Generation of $p_{j\|i}$		Estimator	$\hat{\bar{Y}}_i^{\text{EBLUP}}$ without $g(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g(p_{j\|i}) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g(p_{j\|i}) = \log(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{LP}}$ $m_0(p_{j\|i})$
PS			0.685	0.229	0.200	0.200
AP	$\alpha = 1$	I	0.740	0.089	0.170	0.087
		NI	0.752	0.158	0.200	0.149
	$\alpha = 2$	I	0.644	0.562	0.568	0.557
		NI	0.650	0.557	0.555	0.555
	$\alpha = 3$	I	0.617	0.588	0.591	0.612
		NI	0.619	0.587	0.589	0.607
	$\alpha = \infty$	I	0.608	0.619	0.621	0.626
		NI	0.610	0.622	0.625	0.629

When the sampling is less informative $(\alpha = 3)$, the local linear estimator $\hat{\bar{Y}}_i^{\text{LP}}$ is better than $\hat{\bar{Y}}_i^{\text{EBLUP}}$, but its $\overline{\text{RMSE}}$ is slightly larger than those associated with the parametric estimators $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$. In this case, we observe that the estimated function $m_0(p_{j|i})$ is close to a flat line, and this implies that the local linear approximation is not as appropriate. This explains why $\hat{\bar{Y}}_i^{\text{LP}}$ is slightly worse than $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$ when the level of informativeness of the sampling is low. A local polynomial estimator performs well when the function $m_0(\cdot)$ is meaningfully non-constant.

When the sample is non-informative $(\alpha = \infty)$, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is better than $\hat{\bar{Y}}_i^{\text{VRH1}}$, $\hat{\bar{Y}}_i^{\text{VRH2}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$ in both invariant and non-invariant case. This conclusion is somewhat different from that of Verret et al. (2015) where for $\alpha = \infty$ their estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH1}}$ and $\hat{\bar{Y}}_i^{\text{VRH2}}$ have equal $\overline{\text{AB}}$ and $\overline{\text{RMSE}}$ values. Verret et al. (2015) used both larger populations and samples, and this may explain why their augmented models produced estimators as good as the population model under non-informative sampling designs. Under our simulation set-up, we found that the $\overline{\text{AB}}$ and $\overline{\text{RMSE}}$ of the EBLUP are small for α values larger than 6: this corresponds to a sample design that is almost non-informative. In this case, we recommend using EBLUP.

5.2 Performance of the MSE estimators

We now turn to the performance of the bootstrap procedures for estimating the MSEs of the EBLUP, VRH and local polynomial estimators. Let $\hat{\bar{Y}}_i$ be an estimator of $\bar{Y}_i$ and $\text{mse}_{\text{boot}}\left(\hat{\bar{Y}}_i\right)$ be the bootstrap estimator of $\text{MSE}\left(\hat{\bar{Y}}_i\right)$. From $R = 1{,}000$ simulated populations and samples, we first computed measures of MSE values as

$$\text{MSE}\left(\hat{\bar{Y}}_i\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\bar{Y}}_i^{(r)} - \bar{Y}_i^{(r)}\right)^2,$$

where $\overline{Y}_i^{(r)}$ is the true mean, and $\hat{\overline{Y}}_i^{(r)}$ is the value of the estimator for the r^{th} population. Let $\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i\right)$ be the bootstrap estimator of $\text{MSE}\left(\hat{\overline{Y}}_i\right)$. It is denoted as $\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i^{\text{EBLUP}}\right)$ for the EBLUP estimator $\hat{\overline{Y}}_i^{\text{EBLUP}}$, and corresponds to the parametric (unconditional) bootstrap method given by equation (4.2). For our local polynomial estimator $\hat{\overline{Y}}_i^{\text{LP}}$ and the Verret et al. (2015) estimators, $\hat{\overline{Y}}_i^{\text{VRH1}}$ and $\hat{\overline{Y}}_i^{\text{VRH2}}$, the mse values, denoted as $\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i^{\text{LP}}\right)$ and $\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i^{\text{VRH}j}\right)$, for $j = 1$ and $j = 2$ respectively, are computed using the conditional parametric bootstrap method of Section 4. For each selected sample in the r^{th} simulated population $(r = 1, \ldots, R)$, we used $B = 400$ bootstraps to compute the r^{th} value of $\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i\right)$, that we denote as $\text{mse}_{\text{boot}}^{(r)}\left(\hat{\overline{Y}}_i\right)$. We considered two measures to evaluate the performance of $\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i\right)$: average absolute relative bias and average confidence interval. These measures are defined as follows:

Average Absolute Relative Bias:

$$\overline{\text{ARB}} = \frac{1}{M}\sum_{i=1}^{M}\left|\frac{E\left(\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i\right)\right)}{\text{MSE}\left(\hat{\overline{Y}}_i\right)} - 1\right|,$$

where

$$E\left(\text{mse}_{\text{boot}}\left(\hat{\overline{Y}}_i\right)\right) = \frac{1}{R}\sum_{r=1}^{R}\text{mse}_{\text{boot}}^{(r)}\left(\hat{\overline{Y}}\right).$$

Average Confidence Level:

$$\overline{\text{CL}} = \frac{1}{M}\sum_{i=1}^{M}\text{CL}_i,$$

where $\text{CL}_i = R^{-1}\sum_{r=1}^{R}\text{I}\left(\overline{Y}_i^{(r)} \in \text{IC}^{(r)}\right)$ and $\text{IC}^{(r)} = \left[\hat{\overline{Y}}_i^{(r)} \pm 1.96\sqrt{\text{mse}_{\text{boot}}^{(r)}\left(\hat{\overline{Y}}_i\right)}\right]$.

Table 5.3 reports simulation results on the average relative bias $\left(\overline{\text{ARB}}\right)$ of the MSE estimators for both the PS size measures (5.2) and Asparouhov size measures (5.3 and 5.4) for $\alpha = 1, 2, 3$ and ∞.

Table 5.3
Average relative bias (%) of mse $\left(\overline{\text{ARB}}\right)$ for the PS and AP size measures

Generation of $p_{j\|i}$ \ Estimator			$\hat{\overline{Y}}_i^{\text{EBLUP}}$ without $g\left(p_{j\|i}\right)$	$\hat{\overline{Y}}_i^{\text{VRH1}}$ $g\left(p_{j\|i}\right) = p_{j\|i}$	$\hat{\overline{Y}}_i^{\text{VRH2}}$ $g\left(p_{j\|i}\right) = \log\left(p_{j\|i}\right)$	$\hat{\overline{Y}}_i^{\text{LP}}$ $m_0\left(p_{j\|i}\right)$
PS			25.4	3.9	3.4	7.7
AP	$\alpha = 1$	I	39.9	9.7	14.4	7.5
		NI	46.6	4.1	8.7	10.0
	$\alpha = 2$	I	16.0	2.9	3.8	5.9
		NI	21.4	3.8	3.5	5.8
	$\alpha = 3$	I	13.4	6.1	6.4	5.8
		NI	15.4	7.3	7.4	8.8
	$\alpha = \infty$	I	4.6	4.2	4.5	6.2
		NI	6.1	6.4	6.3	6.9

The $\overline{\text{ARB}}$ of $\hat{\bar{Y}}_i^{\text{EBLUP}}$, based on the model without the augmenting variable $g(p_{j|i})$, is very large when the sampling is very informative ($\alpha = 1$): 39.9% for I and 46.6% for NI. The $\overline{\text{ARB}}$ gradually decreases to around 5% under non-informative sampling ($\alpha = \infty$). The $\overline{\text{ARB}}$'s of both the parametric and non-parametric estimators are smaller in general than 10%, with the exception of 14.4% for the $\hat{\bar{Y}}_i^{\text{VRH2}}$ estimator that uses $\log(p_{j|i})$ as an augmenting variable.

Table 5.4 reports simulation results on the average confidence level $(\overline{\text{CL}})$ associated with the MSE estimators for both the PS size measures (5.2) and the AP size measures (5.3 and 5.4) for $\alpha = 1$, 2, 3 and ∞ and nominal level of 0.95.

Table 5.4
Average confidence level of mse $(\overline{\text{CL}})$ for the PS and AP size measures

Generation of $p_{j\|i}$ / Estimator			$\hat{\bar{Y}}_i^{\text{EBLUP}}$ without $g(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{VRH1}}$ $g(p_{j\|i}) = p_{j\|i}$	$\hat{\bar{Y}}_i^{\text{VRH2}}$ $g(p_{j\|i}) = \log(p_{j\|i})$	$\hat{\bar{Y}}_i^{\text{LP}}$ $m_0(p_{j\|i})$
PS			0.898	0.937	0.941	0.936
AP	$\alpha = 1$	I	0.856	0.918	0.908	0.928
		NI	0.834	0.930	0.920	0.934
	$\alpha = 2$	I	0.916	0.937	0.936	0.932
		NI	0.907	0.936	0.933	0.936
	$\alpha = 3$	I	0.922	0.927	0.926	0.934
		NI	0.918	0.930	0.933	0.926
	$\alpha = \infty$	I	0.937	0.935	0.935	0.938
		NI	0.934	0.934	0.933	0.931

The EBLUP estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ has the worst coverage when the sample design is very informative. The coverage improves as the design becomes less informative. The coverage of the other estimators is between 93% and 95%, with the exception of $\hat{\bar{Y}}_i^{\text{VRH2}}$ (the one that includes $\log(p_{j|i})$) with coverage slightly lower.

5.3 Inclusion of an augmenting variable

The local polynomial approach results in an automatic way of obtaining a reasonable augmented model that is a function of the selection probabilities $p_{j|i}$. However, given that one does not know whether the design is informative or not, should we always include an augmenting variable in the model? If the sample design is not informative it is reasonable to use model (1.1). Note that in this case, including the augmenting variables, $p_{j|i}$ or $\log(p_{j|i})$, has a very small impact either on the absolute relative bias of the estimator and absolute relative bias of the estimated MSE. A similar conclusion was obtained in Verret et al. (2015) who used a larger population and sample size.

The same question arises with respect to the use of the local polynomial procedure. In this case, the conclusions are not quite as clear. If the design is very informative, the local polynomial approach gains in terms of absolute bias and mean squared error when $\alpha = 1$ or $\alpha = 2$. When the sampling design is less

informative $(\alpha = 3)$ the parametric approach in Verret et al. (2015) is the better choice, but by a very small margin.

In a practical situation, the value of α is not known and the decision to use the augmenting variable in a parametric or nonparametric model should be taken. To this end, we follow the suggested procedure in Verret et al. (2015) to provide some guidelines on how to decide on this choice for an arbitrary data set. Define $u_{ij} = v_i + e_{ij}$, and fit the following model $y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij}$ to the sample data by ordinary least squares (OLS). The residuals are $\tilde{u}_{ij} = y_{ij} - \tilde{\beta}_0 - \tilde{\beta}_1 x_{ij}$, where $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are the OLS estimators of β_0 and β_1 respectively. Figure 5.1 displays residual plots of $\left(\hat{m}_0\left(p_{j|i}\right), \tilde{u}_{ij}\right)$, $j = 1, \ldots, N_i$; $i = 1, \ldots, M$ for the AP measures $\alpha = 1, 2, 3$ and ∞ in the invariant case. For $\alpha = 1$, the relationship between $\tilde{u}_{ij}$ and $\hat{m}_0\left(p_{j|i}\right)$ is clearly linear, suggesting that the design is informative. As α increases, the design is less informative. Note that $m_0\left(p_{j|i}\right)$ is constant when $\alpha = \infty$. Similar observations hold for the non-invariant case. For the PS size measures the graph resembled the one given in Figure 5.1 when $\alpha = 1$.

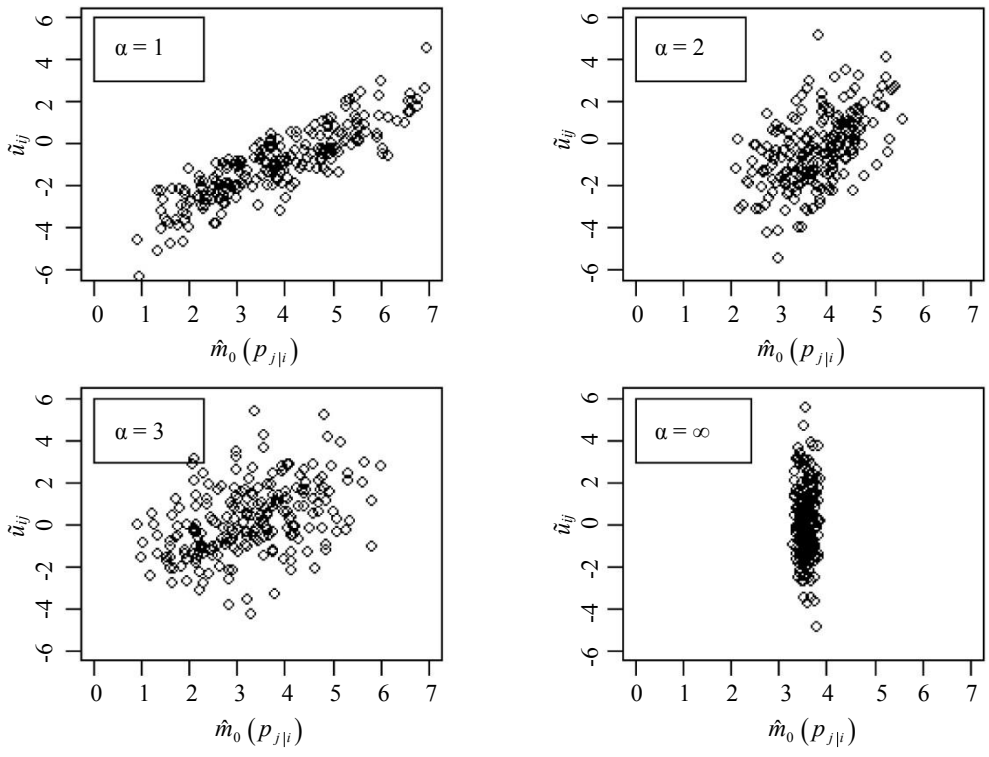

Figure 5.1 Residual plots for the population: AP invariant size measures.

Table 5.5 provides the estimated correlation coefficients, $\hat{\rho} = \text{cor}\left(\tilde{u}_{ij}, \hat{m}_0\left(p_{j|i}\right)\right)$, for PS and AP size measures for $\alpha = 1, 2, 3$ and ∞.

Table 5.5
Estimated correlation coefficient $\hat{\rho} = \text{cor}\left(\tilde{u}_{ij}, \hat{m}_0\left(p_{j|i}\right)\right)$ for the PS and AP size measures

Estimated correlation coefficient	AP								PS
	$\alpha = 1$		$\alpha = 2$		$\alpha = 3$		$\alpha = \infty$		
	I	NI	I	NI	I	NI	I	NI	
$\hat{\rho}$	0.870	0.850	0.450	0.510	0.240	0.210	0.007	0.001	0.800

In terms of $\overline{\text{RMSE}}$, we noticed in Section 5.1 that $\hat{\bar{Y}}_i^{\text{EBLUP}}$ is better than the estimators based on augmented models for $\alpha \geq 6$. Results not presented in Table 5.5 show that for $\alpha \geq 6$, the absolute value of the correlation coefficient is less than 0.1. On the basis of this limited simulation, a user could decide on the choice of the estimator to use for a real data set as follows: i. If $|\hat{\rho}|$ is larger than 0.5, use $\hat{\bar{Y}}_i^{\text{LP}}$; ii. If $|\hat{\rho}|$ is less than 0.1, use $\hat{\bar{Y}}_i^{\text{EBLUP}}$; iii. otherwise use $\hat{\bar{Y}}_i^{\text{VRH1}}$ or $\hat{\bar{Y}}_i^{\text{VRH2}}$.

6 Concluding remarks

In this paper, we studied the estimation of a small area mean under informative sampling by using an augmented model approach where the augmenting variable is a smooth function $m_0\left(p_{j|i}\right)$ of the selection probability $p_{j|i}$. Our augmented model is semi-parametric. It differs from Verret et al. (2015), in that nothing was assumed about the augmenting function $m_0(\cdot)$.

We proposed a three-step procedure to estimate the augmented semi-parametric model. Firstly, local polynomial fits were estimated for each unit of the population (sampled and non-sampled). Secondly, given these local fits a new dependent variable was defined to obtain global estimators of the regression parameters and the small area effects. The resulting estimators were used to compute the predicted values of the dependent variable, y, for all non-sampled units. Finally, using the observed sample values of y, and the predicted values of y, we computed the local polynomial estimator $\hat{\bar{Y}}_i^{\text{LP}}$ for the small area mean $\bar{Y}_i$.

We adopted the conditional parametric bootstrap method to estimate the mean squared error of the newly proposed estimator. The conditional bootstrap is a modified version of the parametric bootstrap estimator method of Hall and Maiti (2006).

We carried out a simulation study to compare the bias and mean squared error performance of the usual EBLUP, $\hat{\bar{Y}}_i^{\text{EBLUP}}$, the augmented EBLUP of Verret et al. (2015), $\hat{\bar{Y}}_i^{\text{VRH}}$, and the proposed local polynomial estimator, $\hat{\bar{Y}}_i^{\text{LP}}$. As expected, $\hat{\bar{Y}}_i^{\text{EBLUP}}$ exhibited large bias under informative sampling. The new estimator $\hat{\bar{Y}}_i^{\text{LP}}$ had equal or smaller MSE than $\hat{\bar{Y}}_i^{\text{VRH}}$ when the sample design was highly informative. If the sample design is less informative, it is better to use one of the two estimators in Verret et al. (2015): that is, augment the basic model with either $p_{j|i}$ or $\log\left(p_{j|i}\right)$. Note that in doing so, the gains are very small. If the sampling design is very slightly or not at all informative, then estimator $\hat{\bar{Y}}_i^{\text{EBLUP}}$ based on the population model should be used.

We also evaluated the performance of the mean squared error bootstrap estimation for the estimators $\hat{\bar{Y}}_i^{\text{EBLUP}}$, $\hat{\bar{Y}}_i^{\text{VRH}}$ and $\hat{\bar{Y}}_i^{\text{LP}}$, in terms of average absolute relative bias $\left(\overline{\text{ARB}}\right)$ and average confidence level $\left(\overline{\text{CL}}\right)$. The conditional bootstrap provides a good way to estimate the mean squared errors.

The advantage of the local polynomial approach is that it provides an automatic way of augmenting the model when the design is informative. Its biggest disadvantage is its computational burden both in terms of parameter estimation and associated reliability. The procedure outlined in Section 5.3 suggests a way to determine whether it is worth using it or not. An alternative approach is to augment the unit level model with a P-spline term of selection probabilities to account for the informativeness of the sampling design. This approach has been recently studied by Chatrchi (2018).

Acknowledgements

We would like to thank J.N.K. Rao for suggesting the conditional bootstrap for estimating the mean squared error of the polynomial estimator, and commenting on the article. We would also like to thank the Associate Editor and a referee for their constructive comments that have improved the paper.

References

Asparouhov, T. (2006). General multi-level modelling with sampling weights. *Communication in Statistics, Theory and Methods*, 439-460.

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

Breidt, F.J., and Opsomer, J.D. (2000). Local polynomial regression estimators in survey sampling. *Annals of Statistics*, 28, 4, 1026-1053.

Breidt, F.J., Opsomer, J.D., Johnson, A.A. and Ranalli, M.G. (2007). Semiparametric model-assisted estimation for natural resource surveys. *Survey Methodology*, 33, 1, 35-44. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2007001/article/9850-eng.pdf.

Chatrchi, G. (2018). *Small Area Estimation: Informative Sampling and Two-Fold Models*. Unpublished Ph.D. Thesis, Carleton University, Ottawa, Canada.

Fan, J., and Gijbels, I. (1996). *Local Polynomial Modelling and its Applications*.London: Chapman and Hall.

González-Manteiga, W., Lombardia, M.J., Molina, I., Morales, D. and Santamaria, L. (2008). Bootstrap mean squared error of a small area EBLUP. *Journal of Statistical Computation and Simulation*, 78, 5, 443-462.

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

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 265-286.

Pfeffermann, D., and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 480, 1427-1439.

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, 409, 163-171.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Rao, J.N.K., Sinha, S.K. and Dumitrescu, L. (2014). Robust small area estimation under semi-parametric mixed models. *The Canadian Journal of Statistics*, 42, 126-141.

Ruppert, D., and Matteson, D.E. (2015). *Statistics and Data Analysis for Financial Engineering with R Examples: 2nd Ed.*, New York: Springer.

Tillé, Y. (2006). *Sampling Algorithms:* New York: Springer.

Verret, F., Rao, J.N.K. and Hidiroglou, M.A. (2015). Model-based small area estimation under informative sampling. *Survey Methodology*, 41, 2, 333-347. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2015002/article/14248-eng.pdf.

Wu, H., and Zhang, J.T. (2002). Local polynomial mixed-effects models for longitudinal data. *Journal of the American Statistical Association*, 97, 459, 883-897.