Electronic Journal of Statistics
Vol. 7 (2013) 1835–1855
ISSN: 1935-7524
DOI: 10.1214/13-EJS827

Small area estimation by splitting the sampling weights

Toky Randrianasolo

Dynamiques Économiques et Sociales des Transports (DEST), Institut Français des sciences et technologies des transports, de l'aménagement et des réseaux (Ifsttar)
14 - 20 Boulevard Newton, F-77447 Marne-la-Vallée Cedex 2, France;
University of Neuchâtel
e-mail: toky.randrianasolo@ifsttar.fr ; toky.randrianasolo@unine.ch

and

Yves Tillé

Institute of Statistics, University of Neuchâtel
Pierre à Mazel 7, CH-2000 Neuchâtel, Switzerland
e-mail: yves.tille@unine.ch

Abstract: A new method is proposed for small area estimation. The principle is based upon the splitting of the sampling weights between the areas. A matrix of weights is defined. Each column of this matrix enables us to estimate the total of the variables of interest at the level of an area. This method automatically satisfies the coherence property between the local estimates and the overall estimate. Moreover, the local estimators are calibrated on auxiliary information available at the level of the small areas. This methodology also enables the use of composite estimators that are weighted means between a direct estimator and a synthetic estimator. Once the weights are computed, the estimates can be easily computed for any variable of interest. A set of simulations shows the interest of the proposed method.

AMS 2000 subject classifications: primary 62D05.
Keywords and phrases: Indirect estimator, matrix calibration, weights.

Received March 2013.

1. Introduction

Three main families of estimators are usually used by statisticians to increase the quality of estimates at the level of small areas: direct estimators and indirect estimators based on implicit or explicit models (see, for instance, Rao, 2003).

The direct estimators are based on the survey data provided only by the considered area. When available, the auxiliary information only depends on the units of the area. The family of direct estimators gathers the estimator proposed by Horvitz and Thompson (1952) also called π-estimator, the generalized regression estimators (Särndal, Swensson and Wretman, 1992) and the calibration estimators (Deville and Särndal, 1992). The main problem with this family of estimators is the increasing variance when the area size decreases. A wise choice of auxiliary information can reduce the variance.

The indirect estimators can depend on all the sampled units for the estimation of a particular area. These estimators are based on the notion of deriving strength from space because a unit of a given area can help for the estimation of any area. Two sub-families of estimators are distinguished: synthetic and composite estimators.

According to Gonzalez (1973), "an unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates". Practically, these estimators are based on the hypothesis of the equality of a parameter within the areas and in the overall population. The synthetic estimator can thus be a prediction by means of a linear model of a total based upon the assumption that the regression coefficient does not vary from an area to another. These estimators generally have a low variance since they depend on all the observations, i.e. the size of the sample is thus large. Unfortunately, these estimators can miss a specificity of a given area. The composite estimators are weighted means of a direct estimator and a synthetic one. For each area, the weight of the mean can be chosen by minimizing the mean squared error.

The indirect estimators built from explicit modeling are based on linear mixed models, generalized linear mixed models and Bayesian procedures which derive Best Linear Unbiased predictors (BLUP), Empirical Best Linear Unbiased Predictors (EBLUP) and empirical Bayes estimators (see, among others, Fuller and Battese, 1973; Prasad and Rao, 1990, 1999; Rao, 2003). The most famous model using linear mixed models is the one developed by Fay and Herriot (1979). The authors begin by modeling a function of the mean in a given area, explained on the one hand, by the auxiliary information, and on the other hand, by a random part explaining the variability across areas that are not considered in the auxiliary information. Then, they show that the BLUP is a composite estimator.

In small area estimation, an important factor to bear in mind is that local estimates are not always consistent with the overall population estimate. Indeed, in general, the sum of estimates at the level of small areas does not coincide with the estimate at the level of the overall population. In order to satisfy a benchmarking property, Prasad and Rao (1999) propose a two-step procedure to obtain a pseudo-EBLUP of a small area mean. They combine area models using survey weights with unit level models. Using a nested error regression model, You and Rao (2002) developed a method that provides coherent estimates thanks to a skillful variable change in the regression and the use of sampling weights so as to build a pseudo-EBLUP. Also, under a nested error regression model, You and Rao (2003) use a pseudo-hierarchical Bayes approach to obtain posterior estimators of small area means. Ugarte, Militino and Goicoa (2009) propose an EBLUP based upon a linear mixed model with restrictions. They force the sum of small area estimates to equal the calculated estimate of the overall population using a synthetic estimator.

In this paper, a new approach is proposed. The method consists of splitting the sampling weights of the overall population estimator to construct local estimators. Each weighting system corresponds to a small area. The idea is to

define weights for the areas that depend on the sampled units as well as the other small areas. Each statistical unit can contribute to all the small areas.

In order to satisfy a benchmarking principle, the sum of the weights of a particular unit relative to each area must be equal to its global weight, which automatically implies that the global estimator is the sum of the local estimates. Furthermore, the weights are calibrated in such a way that, in each area, the estimates of the totals are equal to the population total for the auxiliary variables that are known at the level of the small areas.

The main tool of this method is a matrix **Q** for which the number of rows is equal to the number of units and the number of columns is equal to the number of small areas. This matrix embodies the way the global weights are split into the areas. The benchmarking principle is obtained by the simple fact that the sum of the elements of each row is equal to 1. For each unit, two kinds of contributions are distinguished: its contribution to its own area ("auto-contribution") and its contribution to the other areas ("extra-contribution"). A consequence of the benchmarking principle is that the more a unit contributes to its own area, the less it contributes to the other ones. A composite estimator is built with an "area" part by a direct estimation and with an "extra-contribution" part, given the **Q** probability matrix. At the level of each small area, these two parts are balanced with a parameter (tuning constant) obtained by minimizing the statistical dispersion of the variable of interest.

This paper is structured as follows. In Section 2, the notation is defined. The direct estimators are presented in Section 3. In Section 4, the weight splitting estimation is developed. Next, in Section 5, the composite estimator is presented. The choice of the tuning constant is discussed in Section 6. A simulation study is presented in Section 7, and the paper ends with some brief concluding remarks in Section 8.

2. Notation

Consider a finite population U of N statistical units belonging to D disjoint areas $\{A_1, \ldots, A_d, \ldots, A_D\}$ of sizes $\{N_1, \ldots, N_d, \ldots, N_D\}$. The units can be identified by a label $k \in \{1, \ldots, k, \ldots, N\}$. Consider also J auxiliary variables $x_1, \ldots, x_j, \ldots, x_J$. We are interested in estimating the overall total

$$t_y = \sum_{k \in U} y_k.$$

of the interest variable y. Moreover, we want to estimate the total of y in each area, i.e. for area A_d

$$t_y^d = \sum_{k \in U \cap A_d} y_k.$$

The estimation is based upon the availability of J auxiliary variables. The values of the jth variable for all the units are denoted $x_{1j}, \ldots, x_{kj}, \ldots, x_{Nj}$. The values of the J variables of unit k are denoted by the column vector of $\mathbb{R}^J$

$$\mathbf{x}_k = \begin{pmatrix} x_{k1} & \cdots & x_{kj} & \cdots & x_{kJ} \end{pmatrix}^\top.$$

The J variables of the population are represented by $N \times J$ matrix

$$\mathbf{X}_U = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_k & \cdots & \mathbf{x}_N \end{pmatrix}^\top.$$

The totals of the J variables for the population are represented by the row vector of size J

$$\mathbf{t}_x = \begin{pmatrix} \sum_{k\in U} x_{k1} & \cdots & \sum_{k\in U} x_{kj} & \cdots & \sum_{k\in U} x_{kJ} \end{pmatrix}.$$

The totals of the J variables for the population belonging to area A_d are represented by the row vector of size J

$$\mathbf{t}_x^d = \begin{pmatrix} \sum_{k\in U\cap A_d} x_{k1} & \cdots & \sum_{k\in U\cap A_d} x_{kj} & \cdots & \sum_{k\in U\cap A_d} x_{kJ} \end{pmatrix},$$

and the totals of the J variables for the population belonging to each of the areas are represented by the matrix of size $D \times J$

$$\mathbf{t}_x^A = \begin{pmatrix} \mathbf{t}_x^1 \\ \vdots \\ \mathbf{t}_x^d \\ \vdots \\ \mathbf{t}_x^D \end{pmatrix}. \tag{2.1}$$

Below, matrix $\mathbf{t}_x^A$ is supposed to be known. This matrix can thus be used to improve estimation of the totals in domains t_y^d.

A sample s is a subset of U, and a sampling design $p(s)$ is a probability distribution on all possible samples that can be drawn from U, such that

$$p(s) \geq 0, \text{ and } \sum_{s\subset U} p(s) = 1.$$

For a given sampling design $p(s)$, a sample s is the realization of a random sample S, i.e. $\Pr(S = s) = p(s)$ for all $s \subset U$. We note $\{n_1, \ldots, n_d, \ldots, n_D\}$ the sizes of the areas $\{\#(S \cap A_1), \ldots, \#(S \cap A_d), \ldots, \#(S \cap A_D)\}$. The first order inclusion probability of unit k is denoted by $\pi_k = \Pr(k \in S)$.

3. Direct estimation

At the level of an area, direct estimation consists of building an estimator of t_y^d without using any information outside of the given area. Then, in a direct estimation, a unit can only contribute to its own area. For instance, the Horvitz and Thompson (1952) estimator (or π-estimator) which directly uses the sample weights $1/\pi_k$ is a direct estimator. A calibrated estimator is also a direct estimator.

Let $t_y = \sum_{k\in U} y_k$ be the total of the quantitative variable y. The "Horvitz and Thompson (1952) estimator" or "π-estimator" of t_y is defined by

$$\hat{t}_{y,\pi} = \sum_{k\in S} \frac{y_k}{\pi_k}.$$

At the level of area A_d, the Horvitz and Thompson (1952) estimator of the total of the variable y denoted by $t_y^d = \sum_{k \in U \cap A_d} y_k$ is given by the quantity:

$$\hat{t}_{y,\pi}^d = \sum_{k \in S \cap A_d} \frac{y_k}{\pi_k}.$$

This method of calibration was formalized by Deville and Särndal (1992) who gave a common framework for the calibration estimation and the properties of these estimators. Suppose that a vector of totals $\mathbf{t}_x$ of J auxiliary variables is known at the level of the population. The calibration estimator of the total of the quantitative variable y depends on a weighting system w_k. The estimator is defined by

$$\hat{t}_{y,w} = \sum_{k \in S} w_k y_k,$$

where the w_k weighting system depends on the sample S and satisfies the calibration equation:

$$\sum_{k \in S} w_k \mathbf{x}_k^\top = \mathbf{t}_x. \tag{3.1}$$

The calibration equation (3.1) means that the weighting system of the calibrated estimator must reproduce exactly the values of the totals of the auxiliary variables that are known at the population level.

The weights w_k are computed in such a way to be as close as possible to the Horvitz-Thompson weights $d_k = 1/\pi_k$. Moreover, the weights must satisfy the calibration equation (3.1). In order to find such weights, Deville and Särndal (1992) propose several pseudo-distances denoted by $G_k(w_k, d_k)$ that is assumed to be positive, derivable, strictly convex with regard to w_k and such that $G_k(d_k, d_k) = 0$ for all $k \in U$. The w_k weights are obtained by minimizing the quantity

$$\sum_{k \in S} \frac{G_k(w_k, d_k)}{q_k}$$

subject to the constraints of the calibration equation given in (3.1). The ${q_k}^{-1}$ are coefficients which determine the importance of each unit in the calculation of the distance.

Many different distances can be used and are discussed in Deville and Särndal (1992). In general, if $g_k(w_k, d_k)$ denotes the derivative of $G_k(w_k, d_k)$ with respect to w_k, then the weights are defined by

$$w_k = d_k F_k(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k), \tag{3.2}$$

where $\boldsymbol{\lambda}$ is the vector of Lagrangian multiplier and $d_k F_k(.)$ is the reciprocal function of $g_k(., d_k)$. The value of the Lagrangian multipliers $\boldsymbol{\lambda}$ can be identified by inserting (3.2) in (3.1) and by solving the calibration equation by the Newton-Raphson method.

Below, we mainly use the raking-ratio method which is defined by means of the Kullback-Leibler measure. In this case, $q_k = 1, k \in S$

$$G_k(w_k, d_k) = w_k \log \frac{w_k}{d_k} + d_k - w_k,$$

$$g_k(w_k, d_k) = \log \frac{w_k}{d_k},$$

and

$$w_k = d_k F_k(q_k \boldsymbol{\lambda}^\top \mathbf{x}_k) = d_k \exp(\boldsymbol{\lambda}^\top \mathbf{x}_k).$$

At the level of small areas, a direct estimator cannot be used when the sample size within the area is small because its variance becomes very large. For a given area A_d of size n_d, the variance of a direct estimator is $\mathcal{O}(1/n_d)$. The smaller the size n_d, the larger the variance. Hence, the quality of small area direct estimates is debatable. When the size of a given area is not large enough to have a satisfactory direct estimation, we attempt to improve the quality of the estimates by borrowing information at the level of the other areas. And so we use the method of splitting the sampling weights.

4. Weight splitting or extra-contribution estimation

4.1. Constraints on the split weights

The proposed method consists of splitting the weights into the areas. In a direct estimator, only the weights of the units that belong to an area A_d can contribute to a local estimation at the level of A_d, which is called "auto-contribution part". In the weight splitting method, any unit can contribute to the estimation of any area through a certain weight, which is called "extra-contribution part".

Suppose that a weighting system has already been computed for the overall estimation. These weights can be the inverse of the inclusion probabilities or can be obtained by means of a calibration procedure on the total $\mathbf{t}_x$. The main idea of the proposed weight-splitting approach is to build a weight w_{kd} which depends both on unit k and area A_d. This weight is defined as the product of the basic w_k weight with a splitting coefficient q_{kd} that distributes the weights in the areas, i.e.

$$w_{kd} = w_k q_{kd}, \text{ for all } k \in S \text{ and for } d = 1, \ldots, D.$$

The $D \times J$ matrix $\mathbf{t}_x^A$ of the D totals of the auxiliary variables, given in (2.1) is supposed to be known. The knowledge of this auxiliary information at the level of the areas is used to calibrate the weighting system.

More precisely we would like to have weights for the areas that are calibrated on the totals of the areas, i.e.

$$\sum_{k \in S} w_{kd} \mathbf{x}_k^\top = \sum_{k \in S} w_k q_{kd} \mathbf{x}_k^\top = \mathbf{t}_x^d, \text{ for all } d, = 1, \ldots, D. \tag{4.1}$$

Moreover, we want to impose a coherence between the sum of small areas estimates and the overall estimates. Since $w_{kd} = w_k q_{kd}$, the coherence

$$\sum_{d=1}^{D} \sum_{k \in S} w_{kd} \mathbf{x}_k^\top = \sum_{k \in S} w_k \mathbf{x}_k^\top$$

can be obtained if

$$\sum_{d=1}^{D} w_{kd} = \sum_{d=1}^{D} w_k q_{kd} = w_k, \text{ for all } k \in S.$$

The splitting coefficients must thus satisfy

$$\sum_{d=1}^{D} q_{kd} = 1. \tag{4.2}$$

To sum up, the $n \times D$ weights q_{kd} must satisfy $D \times J + n$ constraints:

- the $D \times J$ constraints of calibration on the totals of the areas given in (4.1),
- the n constraints of consistency given in (4.2).

Note also that the constraints of consistency given in (4.2) are based on an important practical interpretation. If, in small area estimation, the estimate is written in the form of a weighting system, the constraints of consistency imply that units that strongly contribute to other areas contribute less to their own area. Extra-contribution works like an exchange of information, if there are few units in an area, this area needs to borrow strength from other areas, but in exchange, the units of this small area must contribute more to the other areas.

The weights q_{kd} can be gathered in a $\mathbf{Q}$ matrix with n rows and D columns. The sum of the elements in each row is thus equal to 1:

$$\mathbf{Q} = \begin{pmatrix} q_{11} & \cdots & q_{1d} & \cdots & q_{1D} \\ \vdots & & \vdots & & \vdots \\ q_{k1} & \cdots & q_{kd} & \cdots & q_{kD} \\ \vdots & & \vdots & & \vdots \\ q_{n1} & \cdots & q_{nd} & \cdots & q_{nD} \end{pmatrix},$$

which can be written

$$\mathbf{Q}\mathbf{1}_D = \mathbf{1}_n,$$

where $\mathbf{1}_D$ (resp. $\mathbf{1}_n$) is a column vector of D ones (resp. of n ones).

The constraints given in (4.1) can be rewritten with a matrix notation:

$$\begin{pmatrix} q_{11} & \cdots & q_{k1} & \cdots & q_{n1} \\ \vdots & & \vdots & & \vdots \\ q_{1d} & \cdots & q_{kd} & \cdots & q_{nd} \\ \vdots & & \vdots & & \vdots \\ q_{1D} & \cdots & q_{kD} & \cdots & q_{nD} \end{pmatrix} \begin{pmatrix} w_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & w_k & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & w_n \end{pmatrix} \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_k^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} \mathbf{t}_x^1 \\ \vdots \\ \mathbf{t}_x^d \\ \vdots \\ \mathbf{t}_x^D \end{pmatrix}.$$

In short, the two following properties must be satisfied:

1. the sum of the rows of $\mathbf{Q}$ must equal 1,
2. $\mathbf{Q}^\top \operatorname{diag}(w_1, \ldots, w_k, \ldots, w_n)\mathbf{X}_S = \mathbf{t}_x^A$, where $\mathbf{X}_S = (x_{kj})_{k \in S, j=1,\ldots,J}$.

Each coefficient q_{kd} embodies the contribution of unit k to the estimator for area A_d.

4.2. Computation of matrix Q

In order to compute a matrix $\mathbf{Q}$ which satisfies both the coherence property and the totals at the level of each area, we propose a simple algorithm which repeats two successive calibrations. The columns on the totals of the areas are calibrated on the values of the x-variables. Next, the rows of the matrix are calibrated again to ensure that the sum is equal to 1. These steps are repeated until convergence. This method is a generalization of the raking-ratio method that enables us to calibrate a contingency table on marginal totals (see for instance Ireland and Kullback, 1968; Arora and Brackstone, 1977).

More specifically, the algorithm begins with the initialization of the matrix by

$$\mathbf{Q}^{\{0\}} = \begin{pmatrix} \frac{N_1}{N} & \cdots & \frac{N_d}{N} & \cdots & \frac{N_D}{N} \\ \vdots & & \vdots & & \vdots \\ \frac{N_1}{N} & \cdots & \frac{N_d}{N} & \cdots & \frac{N_D}{N} \end{pmatrix}.$$

This first matrix of splitting coefficients simply shares the weights proportionally to the size of the areas in the population.

Next, at step $2t$, for $t = 1, 2, 3, \ldots$ the following two operations are repeated:

- Each column of $\mathbf{Q}^{\{2t-2\}}$ is calibrated on the vector of known totals of each area by solving in $\boldsymbol{\lambda}$ for each area A_d, $d = 1 \ldots D$ the equations system:

$$\mathbf{t}_x^d = \sum_{k \in S} w_k q_{kd}^{\{2t-2\}} \mathbf{x}_k^\top \exp\left(\mathbf{x}_k \boldsymbol{\lambda}_d\right)$$

  The coefficients of the new matrix $\mathbf{Q}^{\{2t-1\}}$ can be obtained by:

$$q_{kd}^{\{2t-1\}} = q_{kd}^{\{2t-2\}} \exp(\mathbf{x}_k \boldsymbol{\lambda}_d).$$

- Then the sum of the rows is calibrated so as to equal 1:

$$q_{kd}^{\{2t\}} = \frac{q_{kd}^{\{2t-1\}}}{\sum_{d=1}^{D} q_{kd}^{\{2t-1\}}}.$$

The iteration stops when the sum of the rows is almost equal to 1 after a column calibration, or more specifically when

$$\sum_{k=1}^{n} \left| \sum_{d=1}^{D} q_{kd}^{\{2t-1\}} - 1 \right| < \varepsilon,$$

where ε is a sufficiently small positive real.

The use of an exponential calibration function guarantees that the weights q_{kd} remain nonnegative at each step of the method. Once the $\mathbf{Q}$ matrix is computed, the totals of areas for any variable of interest can be estimated. At the level of a given area A_d, the extra-contribution estimator of the total of a quantitative variable y is defined by

$$\hat{t}_{y,q}^d = \sum_{k \in S} w_k q_{kd} y_k. \tag{4.3}$$

The sum of the area estimators is always equal to the estimator in the population.

5. Composite estimator

The weight splitting method is a synthetic estimator because all the statistical units can contribute to each area. In order to avoid to miss some area specificities, a composite estimator can be obtained by mixing the direct estimator with the extra-contribution estimator so as to propose a method where each area is estimated by a part of auto-contribution built from a direct estimation and by a part of extra-contribution built thanks to the **Q** matrix.

A matrix **C** of new composite weights is constructed by means of weights α_d for $d = 1, \ldots, D$. The procedure starts with the construction of a $n \times D$ matrix $\mathbf{G} = (g_{kd})$, where

$$g_{kd} = \alpha_d q_{kd} + (1 - \alpha_d)\mathbb{1}_{\{k \in A_d\}}, k \in S, d = 1, \ldots, D,$$

where $\mathbb{1}_{\{C\}}$ equals 1, if condition C is true and 0 otherwise. The coefficients α_d depend on the areas. The smaller the area, the larger the α_d. It is desirable that, in large areas, the estimator depends more on the units of these areas. Whereas in small areas, the estimator depends more on the units outside of these areas.

Next, matrix **G** is calibrated again on the two sets of constraints in such a way that the totals of the areas are reproduced for the auxiliary variables and that the sum of the rows equals 1. The algorithm described in Section 4.2 is thus applied again. We obtain matrix $\mathbf{C} = (c_{kd})$ the elements of which can be written as $c_{kd} = g_{kd}h_{kd}$ where the h_{kd} are the matrix calibration adjustments.

Considering a quantitative variable of interest y, the composite estimator of the total of y at the level of A_d is a weighted average given by:

$$\begin{aligned} \hat{t}^d_{y,c} &= \sum_{k \in S} c_{kd} w_k y_k \\ &= \alpha_d \sum_{k \in S} h_{kd} q_{kd} w_k y_k + (1 - \alpha_d) \sum_{k \in S \cap A_d} h_{kd} w_k y_k. \end{aligned} \tag{5.1}$$

The estimator is a weighted average of two terms. The first one is a synthetic estimator that depends on all the statistical units of the sample. The second one is a direct estimator that only depends on the selected units in the small area.

6. Determination of a tuning constant α_d

6.1. Approximation of the variance of the composite estimator

In order to obtain a reasonable value for the tuning constants, one can use heuristic reasoning. Since the first term of the composite estimator given in (5.1) depends on all the units, we assume that its variance can be written $\sigma^2_{d,1}/n$. Since the second term only depends on the units that belong to A_d, we assume that its variance is equal to $\sigma^2_{d,2}/n_d$.

Moreover, if we assume that the correlation coefficient between the first and the second term is equal to ρ, the variance of the composite estimator is equal to the following quantity:

$$\operatorname{Var}(\hat{t}_{y,c}^{d}) = \alpha_d^2 \frac{\sigma_{d,1}^2}{n} + (1-\alpha_d)^2 \frac{\sigma_{d,2}^2}{n_d} + 2\alpha_d(1-\alpha_d)\rho \frac{\sigma_{d,1}\sigma_{d,2}}{\sqrt{n_d n}}.$$

If we assume that the covariance term is negligible,

$$\operatorname{Var}(\hat{t}_{y,c}^{d}) \approx \alpha_d^2 \frac{\sigma_{d,1}^2}{n} + (1-\alpha_d)^2 \frac{\sigma_{d,2}^2}{n_d}. \tag{6.1}$$

This kind of approximation is done for composite estimators for instance in Rao (2003, p. 57).

By setting the derivative of (6.1) with respect to α_d to zero, we obtain:

$$\alpha_d \frac{\sigma_{d,1}^2}{n} - (1-\alpha_d)\frac{\sigma_{d,2}^2}{n_d} = 0.$$

The value for α_d that minimizes (6.1) is then given by:

$$\alpha_d(n_d) \approx \frac{1}{1 + \frac{\sigma_{d,1}^2}{\sigma_{d,2}^2}\frac{n_d}{n}}.$$

If

$$\theta_d = \frac{\sigma_{d,1}^2}{\sigma_{d,2}^2},$$

we can see that when n_d tends to 0 (resp. $+\infty$), then $\alpha_d(n_d)$ tends to 1 (resp. 0). If $\sigma_{d,1}^2$ and $\sigma_{d,2}^2$ do not depend on d, then we obtain a simplification:

$$\alpha_d(n_d) \approx \frac{1}{1 + \theta\frac{n_d}{n}}.$$

6.2. EBLUP *and pseudo-*EBLUP *under a mixed model*

6.2.1. Nested error linear regression model

Let us now assume that we are interested in small area means. Henderson (1975); Battese, Harter and Fuller (1988); Prasad and Rao (1990); You and Rao (2002); Rao (2003) proposed a nested error linear regression model to estimate small area means. Using the Prasad and Rao (1990); You and Rao (2002); Rao (2003) notation, we can consider the mixed model approach

$$y_{dk} = \mathbf{x}_{dk}^{\top}\boldsymbol{\beta} + v_d + \varepsilon_{dk} \tag{6.2}$$

where $k = 1, \ldots, n_d$, $d = 1, \ldots, D$, v_d are independent centered normal variables with variances σ_v^2, ε_{dk} are independent centered normal variables with variances σ_ε^2. Moreover, the v_d are assumed to be independent from the ε_{dk}.

The mean for area A_d, denoted $\overline{Y}_d$, can be approximated by the parameter

$$\mu_d = \overline{\mathbf{X}}_d^\top \boldsymbol{\beta} + v_d$$

where

$$\overline{\mathbf{X}}_d = \frac{1}{N_d} \sum_{k=1}^{N_d} \mathbf{x}_{dk}.$$

You and Rao (2002) proposed a combination of the basic unit level model (6.2) with sample weights and obtained the following weighted area level model

$$\overline{y}_{dw} = \sum_{k=1}^{n_d} \frac{y_{dk}}{\pi_{dk} \sum_{l=1}^{n_d} \frac{1}{\pi_{dl}}} = \overline{\mathbf{x}}_{dw}^\top \boldsymbol{\beta} + v_d + \overline{\varepsilon}_{dw} \tag{6.3}$$

where

$$\mathbb{E}(\overline{\varepsilon}_{dw}) = 0$$

and

$$\mathrm{Var}(\overline{\varepsilon}_{dw}) = \sigma_\varepsilon^2 \sum_{k=1}^{n_d} (\pi_{dk} \sum_{l=1}^{n_d} \frac{1}{\pi_{dl}})^{-2} = \sigma_\varepsilon^2 \delta_{dw}.$$

6.2.2. EBLUP *under a mixed model*

When σ_v^2 and σ_ε^2 are known, it follows from Rao (1973); Henderson (1975); Battese, Harter and Fuller (1988); Prasad and Rao (1990); You and Rao (2002); Rao (2003) that the best linear unbiased predictor (BLUP) of μ is given by

$$\tilde{\mu}_d = \gamma_d \overline{y}_d + (\overline{\mathbf{X}}_d - \gamma_d \overline{\mathbf{x}}_d)^\top \tilde{\boldsymbol{\beta}}, \tag{6.4}$$

where $\tilde{\boldsymbol{\beta}}$ is the generalized least square estimator of $\boldsymbol{\beta}$,

$$\overline{y}_d = \frac{1}{n_d} \sum_{k=1}^{n_d} y_{dk},$$

$$\overline{\mathbf{x}}_d = \frac{1}{n_d} \sum_{k=1}^{n_d} \mathbf{x}_{dk}$$

and

$$\gamma_d = \left(1 + \frac{\sigma_\varepsilon^2}{\sigma_v^2 n_d}\right)^{-1}.$$

The variances σ_v^2 and σ_ε^2 can be estimated using two ordinary least squares regressions and the method of moments (see, for instance, Fuller and Battese, 1973; You and Rao, 2002; Rao, 2003).

The expression of $\tilde{\mu}_d$ in (6.4) can be re-written as

$$\tilde{\mu}_d = \gamma_d \left[\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^\top \tilde{\boldsymbol{\beta}}\right] + (1 - \gamma_d) \overline{\mathbf{X}}_d^\top \tilde{\boldsymbol{\beta}}.$$

The empirical best linear unbiased predictor estimator (EBLUP) of $\tilde{\mu}_d$, denoted $\hat{\mu}_d$, is then obtained by replacing σ_v^2 and σ_ε^2 by $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$ in the expression of γ_d

$$\hat{\mu}_d = \hat{\gamma}_d \left[\overline{y}_d + (\overline{\mathbf{X}}_d - \overline{\mathbf{x}}_d)^\top \hat{\boldsymbol{\beta}}\right] + (1-\hat{\gamma}_d)\overline{\mathbf{X}}_d^\top \hat{\boldsymbol{\beta}},$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$.

It follows that the expression of the total estimator is obtained by

$$\begin{aligned}
\hat{t}^d_{y,\text{EBLUP}} &= N_d \hat{\mu}_d, \\
&= \hat{\gamma}_d \left(N_d \overline{y}_d + (\mathbf{t}_x^d - N_d \overline{\mathbf{x}}_d)^\top \hat{\boldsymbol{\beta}}\right) + (1-\hat{\gamma}_d)\mathbf{t}_x^{d\top} \hat{\boldsymbol{\beta}}, \\
&= \hat{\gamma}_d \hat{t}^d_{y,\text{EBLUP,direct}} + (1-\hat{\gamma}_d)\hat{t}^d_{y,\text{EBLUP,synth}} \qquad (6.5)
\end{aligned}$$

*6.2.3. Pseudo-*EBLUP *under a mixed model*

The pseudo-BLUP estimator is obtained from a combination of the model (6.2) with the sample weights. From assuming that σ_v^2 and σ_ε^2 are known, the pseudo-BLUP estimator of μ_d from the aggregated model (6.3) is given by

$$\tilde{\mu}_{dw} = \gamma_{dw} \left[\overline{y}_{dw} + (\overline{\mathbf{X}}_{dw} - \overline{\mathbf{x}}_{dw})^\top \tilde{\boldsymbol{\beta}}_w\right] + (1-\gamma_{dw})\overline{\mathbf{X}}_{dw} \tilde{\boldsymbol{\beta}}_w, \qquad (6.6)$$

where

$$\gamma_{dw} = \left(1 + \frac{\sigma_\varepsilon^2 \delta_{dw}}{\sigma_v^2}\right)^{-1}$$

and

$$\tilde{\boldsymbol{\beta}}_w = \left[\sum_{d=1}^{D}\sum_{k=1}^{n_d} \frac{\mathbf{x}_{dk}}{\pi_{dk}} (\mathbf{x}_{dk} - \gamma_{dw}\overline{\mathbf{x}}_{dw})^\top\right]^{-1} \left[\sum_{d=1}^{D}\sum_{k=1}^{n_d} (\mathbf{x}_{dk} - \gamma_{dw}\overline{\mathbf{x}}_{dw}) \frac{y_{dk}}{\pi_{dk}}\right].$$

The pseudo-empirical best linear unbiased predictor estimator (pseudo-EBLUP) of $\tilde{\mu}_{dw}$, denoted $\hat{\mu}_{dw}$, is then obtained by replacing σ_v^2 and σ_ε^2 by $\hat{\sigma}_v^2$ and $\hat{\sigma}_\varepsilon^2$

$$\hat{\mu}_{dw} = \hat{\gamma}_{dw} \left[\overline{y}_{dw} + (\overline{\mathbf{X}}_{dw} - \overline{\mathbf{x}}_{dw})^\top \hat{\boldsymbol{\beta}}_w\right] + (1-\hat{\gamma}_{dw})\overline{\mathbf{X}}_{dw} \hat{\boldsymbol{\beta}}_w,$$

where $\hat{\boldsymbol{\beta}}_w = \tilde{\boldsymbol{\beta}}_w(\hat{\sigma}_u^2, \hat{\sigma}_\varepsilon^2)$.

It also follows that the expression of the total estimator is obtained by

$$\begin{aligned}
\hat{t}^d_{y,\text{p-EBLUP}} &= N_d \hat{\mu}_{dw}, \\
&= \hat{\gamma}_{dw} \left(N_d \overline{y}_{dw} + (\mathbf{t}_x^d - N_d \overline{\mathbf{x}}_{dw})^\top \hat{\boldsymbol{\beta}}_w\right) + (1-\hat{\gamma}_{dw})\mathbf{t}_x^{d\top} \hat{\boldsymbol{\beta}}_w, \\
&= \hat{\gamma}_{dw} \hat{t}^d_{y,\text{p-EBLUP,direct}} + (1-\hat{\gamma}_{dw})\hat{t}^d_{y,\text{p-EBLUP,synth}} \qquad (6.7)
\end{aligned}$$

For a given area A_d, when the weights are calibrated with the known size N_d i.e.

$$\sum_{k=1}^{n_d} \frac{1}{\pi_{dk}} = N_d,$$

and when the unit level model (6.2) includes the intercept term, the estimator $\hat{t}^d_{y,\text{p-EBLUP}}$ satisfies the benchmarking property without any adjustment (see for instance You and Rao, 2002; Rao, 2003).

In the case that the weights are not calibrated with the known size N_d, a preliminary calibration on the weights can be done in order to obtain a coherent pseudo-EBLUP estimator. It follows that under a simple random sampling without replacement, a preliminary calibration on the weights leads to the equality $\gamma_d = \gamma_{dw}$. In fact, under a simple random sampling without replacement, the sampling weights are all equal: $1/\pi_{dk} = N/n$ for all k. Then, new weights calibrated on the known size N_d are obtained: $1/\pi_{dk,w} = N_d/n_d$. It follows that $\delta_{dw} = \sum_{k=1}^{n_d} (\frac{N_d/n_d}{\sum_{l=1}^{n_d} N_d/n_d})^2 = 1/n_d$ and then $\gamma_d = \gamma_{dw}$.

6.2.4. Composite form of the EBLUP *and the pseudo-*EBLUP

The EBLUP and the pseudo-EBLUP can be seen as composite estimators. They are weighted averages of a regression synthetic estimator and a pseudo-direct estimator. Similarly to the proposed method, when n_d tends to 0 (resp. $+\infty$), γ_d and γ_{dw} tend to 0 (resp. $+\infty$): when the size of an area is large enough, more weight is attached to the direct estimation part and vice versa. Then, an estimation of the tuning constant α_d can be obtained using an analogy with the parameters γ_d and γ_{dw} of the BLUP:

$$\hat{\alpha}_d^{\{1\}} = 1 - \hat{\gamma}_d, \tag{6.8}$$

or

$$\hat{\alpha}_d^{\{2\}} = \hat{\alpha}_{dw} = 1 - \hat{\gamma}_{dw}. \tag{6.9}$$

7. Simulation study

7.1. Simulated data with a mixed model

In order to test the performance of the proposed methodology, we ran a set of simulations.

7.1.1. Simulated population

A population of $N = 2,000$ units, with $D = 20$ disjoint areas of sizes $(N_1, \ldots, N_D)$, is created from a linear mixed model given in (6.2):

$$\begin{aligned}
\mathbf{x}_{dk} &= \begin{pmatrix} 1 & x_{dk} \end{pmatrix}^\top \text{ with } x_{dk} \overset{\text{iid}}{\sim} \mathcal{N}(20, \sigma_x^2 = 9),\\
\beta &= \begin{pmatrix} 12 & 0.4 \end{pmatrix}^\top,\\
v_d &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2 = 4),\\
\varepsilon_{dk} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 = 1).
\end{aligned}$$

The Figure 1 gives an overview of a generated population.

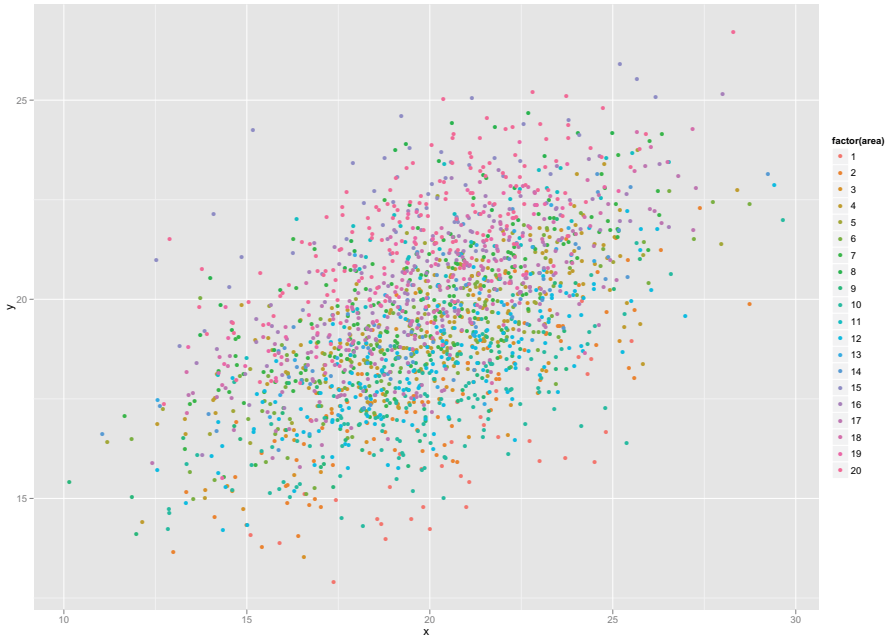

FIG 1. *Generated population from the linear mixed model given in 7.1.1: the different colors indicate the areas).*

7.1.2. Precision comparison between the weights splitting estimator and some classical small area estimators

From the generated population, $B = 10,000$ samples of size $n = 200$ are drawn by a simple random sampling without replacement. Within areas, the EBLUP estimator, the pseudo-EBLUP estimator and the proposed estimator (with their respective direct and synthetic components) are computed to estimate the totals t_y^d, for $d = 1, \ldots, 20$. The relative root mean square error (%RRMSE) is used to quantify the performance of the estimators.

For a given estimator $\hat{t}_y^d$, the $\%\text{RRMSE}_d$ is obtained as follows

$$\%\text{RRMSE}_d = 100 \times \frac{\sqrt{\text{MSE}_d}}{t_y^d}, \tag{7.1}$$

where MSE_d is the sum of the square of the bias and the variance

$$\text{MSE}_d = \left(\frac{1}{B}\sum_{b=1}^{B}\hat{t}_{y,b}^d - t_y^d\right)^2 + \frac{N-n}{N-1}\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{t}_{y,b}^d - \frac{1}{B}\sum_{b=1}^{B}\hat{t}_{y,b}^d\right)^2.$$

Table 1 shows a comparison between the proposed method and the EBLUP estimator. Table 2 shows a comparison between the proposed method and the

TABLE 1
Computed %RRMSE of the EBLUP *estimator, of the proposed estimator, and their respective component estimators, from 10,000 drawn samples of size 200 by a simple random sampling without replacement from the generated population (see Figure 1)*

			EBLUP components			Weights Splitting method		
Area	$\overline{n}_d$	$\overline{\hat{\alpha}}_d$	Direct	Synthetic	EBLUP	Calibration	Extra	Composite
1	3.45	0.14	5.26	24.68	5.95	49.14	24.10	9.70
2	10.03	0.04	1.94	10.62	1.92	27.72	10.09	2.96
3	5.66	0.07	2.68	8.65	2.55	37.40	8.14	3.24
4	17.40	0.02	1.10	1.34	1.07	20.50	0.67	0.98
5	4.69	0.09	2.58	1.48	2.32	41.11	1.52	1.91
6	12.53	0.03	1.29	3.60	1.26	25.07	3.03	1.31
7	18.67	0.02	1.09	2.03	1.07	19.91	2.20	1.06
8	2.81	0.17	3.94	11.88	4.02	53.72	12.22	5.53
9	7.28	0.05	2.02	7.82	1.96	32.72	7.31	2.62
10	15.01	0.02	1.41	12.35	1.41	22.41	11.83	2.61
11	4.41	0.10	2.20	7.38	2.14	43.12	7.73	3.05
12	18.15	0.02	1.14	7.11	1.12	20.45	6.59	1.59
13	3.51	0.14	2.97	1.10	2.57	48.92	0.71	2.08
14	7.76	0.05	1.76	1.11	1.67	31.96	0.65	1.46
15	4.88	0.09	2.46	14.11	2.63	40.50	14.44	5.08
16	12.40	0.03	1.28	5.36	1.25	24.86	5.69	1.60
17	16.82	0.02	1.22	1.62	1.19	21.13	0.95	1.12
18	11.60	0.03	1.25	3.65	1.22	25.65	3.95	1.39
19	6.48	0.06	1.84	5.57	1.76	35.13	5.89	2.19
20	16.46	0.02	1.08	10.76	1.10	21.53	11.11	2.23

pseudo-EBLUP estimator. For each simulation run, each size n_d (for $d = 1, \ldots, D$) is not fixed because the sampling design is a simple random sampling without replacement from the overall population. That is why, the second columns of Table 1 and Table 2 represent the means $\overline{n}_d$ (for $d = 1, \ldots, D$) of each area size through the $B = 10,000$ drawings. The tuning constants $\hat{\alpha}_d$ and $\hat{\alpha}_{dw}$ also are re-estimated in each simulation run. That also is why, the third columns of Table 1 and Table 2 represent the means $\overline{\hat{\alpha}}_d$ (resp. $\overline{\hat{\alpha}}_{wd}$) (for $d = 1, \ldots, D$) of each area tuning constant through the $B = 10,000$ drawings. As discussed in Section 6.2.3, the equalities $\hat{\alpha}_d = \hat{\alpha}_{wd}$ and $\overline{\hat{\alpha}}_d = \overline{\hat{\alpha}}_{wd}$ are obtained because the sampling design is a simple random sampling without replacement from the overall population.

In Table 1, the direct estimation component of the EBLUP estimator appears to be generally better than the global calibration which is the direct estimation part of the weights splitting method. This can be explained by the fact that the direct estimation component of the EBLUP is a regression estimator built from the parameter $\hat{\beta}$ of the synthetic estimation component of the EBLUP. The synthetic part of the EBLUP and the extra-contribution part seem to have equivalent performance. In spite of the weakness of the direct estimation part of the proposed composite estimator relative to the direct estimation part of the EBLUP, the two composite estimators perform equivalently thanks to the matrix calibration computed during the weights splitting method procedure. The result shown in Table 1 is quite difficult to interpret, because the weight-splitting estimator incorporates sampling weights and is benchmarked, while the EBLUP estimator does not.

TABLE 2
*Computed %RRMSE of the pseudo-*EBLUP *estimator, of the proposed estimator, and their respective component estimators, from 10,000 drawn samples of size 200 by a simple random sampling without replacement from the generated population (see Figure 1)*

			pseudo-EBLUP components			Weights Splitting method		
Area	$\overline{n}_d$	$\overline{\hat{\alpha}}_{dw}$	Direct	Synthetic	pseudo-EBLUP	Calibration	Extra	Composite
1	3.45	0.14	5.25	24.58	5.97	49.14	24.10	9.70
2	10.03	0.04	1.94	10.49	1.92	27.72	10.09	2.96
3	5.66	0.07	2.67	8.51	2.54	37.40	8.14	3.24
4	17.40	0.02	1.10	0.86	1.07	20.50	0.67	0.98
5	4.69	0.09	2.58	1.16	2.30	41.11	1.52	1.91
6	12.53	0.03	1.29	3.38	1.26	25.07	3.03	1.31
7	18.67	0.02	1.09	1.83	1.07	19.91	2.20	1.06
8	2.81	0.17	3.94	11.90	4.03	53.72	12.22	5.53
9	7.28	0.05	2.02	7.68	1.96	32.72	7.31	2.62
10	15.01	0.02	1.41	12.23	1.41	22.41	11.83	2.61
11	4.41	0.10	2.20	7.38	2.14	43.12	7.73	3.05
12	18.15	0.02	1.14	6.97	1.12	20.45	6.59	1.59
13	3.51	0.14	2.97	0.49	2.55	48.92	0.71	2.08
14	7.76	0.05	1.76	0.49	1.66	31.96	0.65	1.46
15	4.88	0.09	2.47	14.14	2.64	40.50	14.44	5.08
16	12.40	0.03	1.28	5.33	1.25	24.86	5.69	1.60
17	16.82	0.02	1.22	1.22	1.19	21.13	0.95	1.12
18	11.60	0.03	1.25	3.58	1.22	25.65	3.95	1.39
19	6.48	0.06	1.84	5.55	1.76	35.13	5.89	2.19
20	16.46	0.02	1.08	10.78	1.10	21.53	11.11	2.23

Table 2 seems to be fairer in terms of comparison because both the pseudo-EBLUP estimator and the proposed estimator respect the benchmarking property and use sampling weights. As previously, whereas the synthetic part of the pseudo-EBLUP and of the proposed method seem to perform equivalently, the direct estimation component of the pseudo-EBLUP performs better than the direct estimation part of the weights splitting method. This also can be explained by the construction of the direct estimation of the pseudo-EBLUP with the parameter $\hat{\beta}_w$ which is derived from the synthetic part of the pseudo-EBLUP. Despite this disadvantage, when the calibration and the extra-contribution parts of the weights splitting method are mixed and are re-calibrated on the rows (benchmarking constraints) and on the columns (calibration constraints on the areas), the pseudo-EBLUP estimator and the weights splitting estimator appear to perform equivalently. These obtained similar efficiencies can also partly be explained by the same weight attached to the synthetic component of the pseudo-EBLUP and attached to the extra-contribution component of the proposed composite estimator.

7.1.3. Resampling procedure to estimate the variance of the weights splitting estimator

From the generated artificial population (see Figure 1), a resampling procedure can be performed to estimate the variance of the weights splitting estimator. Given a sample drawn from the generated population, B samples are drawn from

this initial sample (for instance $B = 500$). The B weights splitting estimators are computed from these B new samples. The variance by bootstrap is obtained from computing the empirical variance of the B weights splitting estimators. The considered sampling design is always a simple random sampling without replacement.

In order to test the efficiency of the variance estimation by bootstrap, the Algorithm 1 is considered.

Algorithm 1 Resampling procedure for testing the efficiency of the variance estimation by bootstrap

Consider the generated artificial population (see Figure 1)
Consider $B^* = B^{**} = 500$ the number of iterations
Consider $D = 20$ the number of areas
Consider $n = 200$ the sample size
for $b^* = 1$ **to** B^* **do**
  Draw a sample s^{b^*} of size n by a simple random sampling without replacement from the generated population
  Compute the weights splitting estimators $(\hat{t}^1_{y,c})^{b^*}, \ldots, (\hat{t}^d_{y,c})^{b^*}, \ldots, (\hat{t}^D_{y,c})^{b^*}$
  for $b^{**} = 1$ **to** B^{**} **do**
    Draw a sample $s^{b^*}_{b^{**}}$ of size $n = 200$ by a simple random sampling with replacement from s^{b^*}
    Compute the weights splitting estimators $(\hat{t}^1_{y,c})^{b^*}_{b^{**}}, \ldots, (\hat{t}^d_{y,c})^{b^*}_{b^{**}}, \ldots, (\hat{t}^D_{y,c})^{b^*}_{b^{**}}$
  end for
  Compute $(\hat{\sigma}^{b^*}_1, \ldots, \hat{\sigma}^{b^*}_d, \ldots, \hat{\sigma}^{b^*}_D)$, the empirical variances of the B^{**} obtained weights splitting estimators $(\hat{t}^1_{y,c})^{b^*}_{b^{**}}, \ldots, (\hat{t}^d_{y,c})^{b^*}_{b^{**}}, \ldots, (\hat{t}^D_{y,c})^{b^*}_{b^{**}}$
end for
Compute the empirical variances of the B^* weights splitting estimators $(\hat{t}^1_{y,c})^{b^*}, \ldots, (\hat{t}^d_{y,c})^{b^*}, \ldots, (\hat{t}^D_{y,c})^{b^*}$
Compute the empirical means of the B^* variances $(\hat{\sigma}^{b^*}_1, \ldots, \hat{\sigma}^{b^*}_d, \ldots, \hat{\sigma}^{b^*}_D)$ (bootstrap variance).

In the Algorithm 1, each sample s^{b^*} drawn from the generated population leads to weights splitting estimators $(\hat{t}^d_{y,c})^{b^*}$ with bootstrap variances $\hat{\sigma}^{b^*}_d$ for $d = 1 \ldots D$, where

$$\begin{aligned} \hat{\sigma}^{b^*}_d &= \mathrm{Var}\{(\hat{t}^d_{y,c})^{b^*}_{b^{**}}\} \\ &= \frac{N-n}{N-1}\frac{1}{B^{**}-1}\sum_{b^{**}=1}^{B^{**}}\left((\hat{t}^d_{y,c})^{b^*}_{b^{**}} - \frac{1}{B^{**}}\sum_{b^{**}=1}^{B^{**}}(\hat{t}^d_{y,c})^{b^*}_{b^{**}}\right)^2. \end{aligned}$$

For $d = 1 \ldots D$, for $b^* = 1 \ldots B^*$, consider

$$\mathrm{Var}\{(\hat{t}^d_{y,c})^{b^*}\} = \frac{N-n}{N-1}\frac{1}{B^*-1}\sum_{b^*=1}^{B^*}\left((\hat{t}^d_{y,c})^{b^*} - \frac{1}{B^*}\sum_{b^*=1}^{B^*}(\hat{t}^d_{y,c})^{b^*}\right)^2$$

the simulated variance of the weights splitting estimator and consider

$$\tilde{\mathbb{E}}_{sim}\{\hat{\sigma}^{b^*}_d\} = \frac{1}{B^*}\sum_{b^*=1}^{B^*}\hat{\sigma}^{b^*}_d$$

TABLE 3
Variance of the weights splitting estimators vs *Mean of the bootstrap variances of the weights splitting estimators*

Area	$\mathrm{Var}\{(\hat{t}^d_{y,c})^{b^*}\}$	$\tilde{\mathbb{E}}_{sim}\{\hat{\sigma}^{b^*}_d\}$
1	1264.0	989.3
2	1566.2	1550.0
3	612.8	570.0
4	1021.0	1163.5
5	324.6	274.9
6	778.4	810.6
7	1288.8	1573.6
8	574.2	432.9
9	713.8	699.0
10	2555.2	2768.5
11	505.2	409.1
12	2088.3	1981.9
13	218.0	164.5
14	485.1	472.0
15	1442.2	1264.0
16	1134.5	1290.5
17	1275.3	1383.5
18	813.8	837.5
19	528.9	543.1
20	3438.0	3849.4

TABLE 4
Numbers of times (%) the true total values t^d_y *(for* $d = 1 \dots D$*), from the generated population (see Figure 1), lie within the* 95% *confidence intervals built with bootstrap variances*

Area	N_d	Numbers of times (%)
1	34	75
2	100	84
3	57	78
4	174	93
5	47	84
6	125	91
7	187	93
8	28	65
9	73	85
10	150	87
11	44	75
12	182	87
13	35	88
14	78	91
15	49	73
16	124	89
17	168	94
18	116	87
19	65	86
20	164	87

the simulated expectation of the weights splitting bootstrap variance estimator. Table 3 gives a comparison between these two quantities. It shows that the two quantities are very closed.

Once the bootstrap variances computed, construction of confidence intervals can be processed. Table 4 reports the numbers of times (%) the true total values

t_y^d (for $d = 1 \dots D$), from the generated population, lie within the 95% confidence intervals built with bootstrap variances. Table 4 shows that given an area, the number of times the true total value lies within the 95 % confidence interval increases with the size of the area.

7.2. Data with county crop areas

We compare the pseudo-EBLUP and the weights splitting procedures applied to a real data given by Battese, Harter and Fuller (1988) and taken up by You and Rao (2002); Rao (2003). These authors wanted to estimate the mean of hectares of corn per segment for $D = 12$ counties in north-central Iowa. Each county is divided into area segments and the areas under corn are in the area segments. The authors used a sample s of $n = 36$ segments and assumed simple random sampling within areas. The population is assumed to follow the linear mixed model given in (6.2) where the function of interest is the number of hectares of corn per segment per county, and the auxiliary information is the number of pixels seen as corn and as soybeans.

Table 5 reports the pseudo-EBLUP and weights splitting estimates of hectares of corn with their respective coefficients of variation. The variances are obtained by a resampling procedure of 1,000 iterations (see Algorithm 2). The two es-

TABLE 5
Estimated hectares of corn with coefficients of variation

County	n_d	pseudo-EBLUP		Weights Splitting Method	
		Estimate	c.v. (%)	Estimate	c.v. (%)
		Corn			
Cerro Gordo	1	120.5	2.6	121.8	3.0
Hamilton	1	125.3	2.7	122.7	3.2
Worth	1	106.3	7.8	108.3	6.4
Humboldt	2	107.3	8.4	111.1	5.7
Franklin	3	143.8	4.5	142.8	5.3
Pocahontas	3	111.5	5.0	111.8	6.0
Winnebago	3	112.1	5.5	113.8	4.6
Wright	3	121.3	3.8	120.2	3.3
Webster	4	115.1	3.3	114.7	4.5
Hancock	5	124.5	3.4	124.2	3.1
Kossuth	5	106.6	3.2	109.3	3.8
Hardin	5	143.5	3.3	141.0	3.0

Source: LANDSAT data from Table 1 in Battese, Harter and Fuller (1988, p. 29)

Algorithm 2 Resampling procedure for the variance estimation by bootstrap

Consider $B = 1,000$ the number of iterations
Consider $n = 36$ the sample size
for $b = 1$ **to** B **do**
  Draw a sample s^b of size n by a simple random sampling with replacement within areas from the sample s (the area sample sizes are fixed and are the same as those in s).
  Compute the pseudo-EBLUP and weights splitting estimators from the sample s^b.
end for
Compute the variances of the B obtained pseudo-EBLUP and weights splitting estimators with considering the finite population correction factor.

timators automatically respect the benchmarking property. Indeed, we have $\sum_{d=1}^{D} N_d \hat{\mu}_{dw} = \hat{\mathbf{t}}_{y,\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x,\pi})^\top \hat{\boldsymbol{\beta}}_w = 815016,3$ and $\sum_{d=1}^{D} \hat{t}_{y,c}^{d} = \hat{t}_{y,w} = 817087,2$. Table 5 gives quite similar estimates with similar efficiencies.

8. Concluding remarks

The simulations show that, even when the data are really generated by a mixed model, the proposed estimator does not seem worse than the Eblup and pseudo-Eblup estimators. The proposed estimators offer several advantages. The regional estimates are coherent with the overall estimates. The estimator takes the sampling weights into account. The estimator can be written as a weighting system and can thus be applied on any variable of interest.

As for all the composite estimators, the proposed estimator clearly states that the parameters α_d (for $d = 1 \dots D$) depend on the variable of interest. In the case of a set of variables of interest belonging to a specified theme, the constant α_d can be chosen as the mean of the tuning constants obtained from each variable of interest.

The next step will consist of computing a variance estimator of the composite estimator. Since the proposed estimator is obtained by successive matrix calibrations, the computation of this variance is complex, which is thus a challenging objective. If a closed form of variance estimator is intractable, resampling procedures for simple random sampling as seen in Section 7.1.3 can be considered; resampling procedures for sampling design with unequal inclusion probabilities can be based on Antal and Tillé (2011) methodology.

The proposed method is not particularly robust for the resistance to outliers. It could be an interesting further topic of research.

Acknowledgments

The authors would like to thank an anonymous referee for useful comments and suggestions.

References

Antal, E. and Tillé, Y. (2011). A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population. *Journal of the American Statistical Association* **106** 534–543. MR2847968

Arora, H. R. and Brackstone, G. J. (1977). An investigation of the properties of raking ratio estimator: I. With simple random sampling. *Survey Methodology* **3** 62–83.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association* **83** 28–36.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** 376–282. MR1173804

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. MR0548019

Fuller, W. A. and Battese, G. E. (1973). Transformation for estimation of linear models with nested error structure. *Journal of the American Statistical Association* **68** 626–632. MR0359188

Gonzalez, M. E. (1973). Use and Evaluation of Synthetic Estimates. In *Proceedings of the Social Statistics Section* 33–36. American Statistical Society.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrika* **31** 423–447.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47** 663–685. MR0053460

Ireland, C. T. and Kullback, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188. MR0229329

Prasad, N. G. N. and Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small Area Estimators. *Journal of the American Statistical Association* **85** 163–171. MR1137362

Prasad, N. G. N. and Rao, J. N. K. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology* **25** 67–72.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. John Wiley, New York. MR0346957

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New-York. MR1953089

Särndal, C.-E., Swensson, B. and Wretman, J. H. (1992). *Model Assisted Survey Sampling*. Springer. MR1140409

Ugarte, M., Militino, A. and Goicoa, T. (2009). Benchmarked estimates in small areas using linear mixed models with restrictions. *Test* **18** 342–364. 10.1007/s11749-008-0094-x. MR2520341

You, Y. and Rao, J. N. K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics* **30** 431–439. MR1944372

You, Y. and Rao, J. N. K. (2003). Pseudo hierarchical Bayes small area estimation combining unit level models and survey weights. *Journal of Statistical Planning and Inference* **111** 197–208. MR1955881