# Benchmarking a random intercept regression for small areas via additional columns and rows

R. Priam, N. Shlomo

**Abstract**

Statistical methods for small area generally suppose an underlying linear mixed regression model for the estimation of the population means per area of a variable of interest. The related estimates generally do not add up to the direct estimate of the total computed from the whole sample. A benchmarking method aims at adjusting the predictors and force their sum to be equal to the direct survey estimate. Sample data are supposed available for all the domains which are not empty. Special attention is given to the unit level small area model of Battese, Harter and Fuller (1988) because countries typically have business registers where covariate variables are available for the whole population. A method for benchmarking is considered herein with an explicit formulation. It is defined via an augmented linear system by adding a row and a column to a design matrix as a second step after the first step of the estimation of the variance components. Indeed if adding only one column for the area level case can be enough as seen in the litterature, adding also a row seems required for the unit level case. Empirical results based on simulated datasets are presented for validating our proposal and the interest of our approach.

**Index Terms**

Small area, linear random intercept regression model, benchmarking, augmented system.

## I. Introduction

In survey theory, a sample from the population is drawn for the construction of predictors of the whole population. The sampling is generally performed with the help of a set of variables which are known for every units for business surveys, this can be for instance the income or other variables which are collected exhaustively and are archived in a database called business register. Typically, only the targeted variable is not completely known while the covariates are all available. The population is also partitioned into several domains where one is interested to know the statistics for the corresponding sub-population. Domains can count very few individus and so the corresponding design-based estimators may have large variances due to the uncertainty associated. One classical solution in survey methods is to implement calibration approach with the help of auxiliary variables, like the ratio and the regression estimators. A more efficient solution is to suppose an underlying parametric model which leads to estimates of the targeted variable as a function of a known set of other variables. Pioneering research in this field is presented extensively in the book [1]. Following this idea, a regression function permits to *borrow the strength* between non empty areas in order to provide relevant candidate values with competitive properties for the mean estimation in each area of a target variable. The predictors are called synthetic in opposition to the estimations called design-based from survey theory.

Therefore, for the mean estimation at the level of the whole population, there are available two estimations for one given region considering the area inside the region: the prediction by summing the synthetic estimations per area with weights but also a direct estimation of the mean for the whole population by a design-based method from the complete sample. It is common to observe that these two estimators lead to different estimated values, this is a consequence of a lack in the chosen parametric model, outliers which induces errors at the more aggregated place, the complete or quasi emptiness of some areas. The difference comes then from the two ways to construct the estimation, one which relies on a model-based approach with a super-population hypothesis and one which is a design-based approach with a different sampling and modeling hypothesis. In fine, the less aggregated statistics are available at the level of larger regions or whole population, and for several reasons, it is important to have a perfect correspondence between the two approaches. This comes from the fact an enough large sample in the regions (or whole sample) induces a suitable estimator with a not too large variance, such as it can be preferred to a synthetic one. The final actors have also a better understanding and faith on the presented numbers when the correspondence is achieved. For all these reasons, some corrections of the small area estimates are needed in order to get the property true, this approach is called benchmarking and leads to benchmarked predictors.

In this document, we present a general overview for the benchmarking of small area predictor when the data come from a unit level regression model with a random effect for the intercept. The paper is organized as follows. Section 2 presents the general benchmarking equation after summarizing brievly the mean estimation via the linear model considered herein: the unit level small area model of Battese, Harter and Fuller (BHF) [2] which is a linear random intercept regression model. Section 3 develops further the proposed model which has been brievly introduced at the conference SAE'2011 [3], and lists some related existing estimators used in the experiments part. Section 5 ilustrates with a simulation studies for an empirical evaluation of the presented estimators. Section 5 discusses the contribution and ends with several perspectives.

## II. Mean estimation

A given population is partitioned into $m$ small areas or domains for the variable of interest $y$, and in each domain, a population mean exists, $\bar{Y}_i$. It is supposed that a sample of $n$ units is available from a population of $N$ units. The estimation of the mean in an area can be based on survey theory. In a design based approach, the sampling fractions are denoted $f_i = n_i/N_i$. The hypothesis is that the sample $s_n$ comes from a finite population of units $U_N$, and the usual mean estimation is $\bar{y}_i$. Here the design for the sampling procedure is a simple random one.

One is interested on estimating the population area means $\bar{Y}_i$ knowing $\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{in_i})'$ and the covariates at the level of each sub-population or (small) areas. In the case of small area, most of the sizes $n_i$ can be very small so that the variance is too high to provide enough relevant estimates of the mean values. Hence, an approach with a model is preferred. The sampled unit are drawn independently across small areas (by simple random sampling or SRS here) and they are supposed to have same BHF model than the population.

### A. Model based approach

The BHF model is a linear mixed model with a two-stage sampling.

*a) Model definition:* It is recall the usual way to estimate the mean in each area is briefly described. The model is written as follows:

$$y_{ij} = x'_{ij}\beta + u_i + e_{ij}\,,\; j = 1, \cdots, n_i\;,\; i = 1, \cdots, m\,.$$

Here, the model error is $e_{ij} \sim N(0, \sigma_e^2)$ while the random effects are $u_i \sim N(0, \sigma_u^2)$. The regression vector for the fixed effects is $\beta = (\beta_0, \beta_1, \cdots, \beta_p)' \in R^{p+1}$ for $p$ auxiliary variables. The design matrix of fixed effects for the covariates in each area is $X_i$. Note that the noise $e_{ij}$ and random effects $u_i$ are iid and independent.

For reader unfamiliar with the familly of linear mixed model, we recall that in a matricial format, the model can be written as follows:

$$\begin{array}{ccccccccc}
\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \vdots \\ \mathbf{y}_m \end{bmatrix} & = & \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \vdots \\ \mathbf{x}_m \end{bmatrix} \beta & + & \begin{bmatrix} \mathbf{1}_{n_1} & & & \\ & \mathbf{1}_{n_2} & & \\ & & \ddots & \\ & & & \ddots & \\ & & & & \mathbf{1}_{n_m} \end{bmatrix} u & + & \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \vdots \\ \mathbf{e}_m \end{bmatrix} \\
Y_s & = & X_s\beta & + & Z_s u & + & e_s\,.
\end{array}$$

Here, $\mathbf{1}_{n_i}$ is the vector of size $n_i$ with 1 for every components, and $u = (u_1, u_2, \cdots, u_m)'$, while $Y_s$, $X_s$ and $e_s$ are repectively is compound of the sample vectors $y_i = (y_{ij})'_{1 \le j \le n_i}$, $x_i = (x_{ij})'_{1 \le j \le n_i}$, and $e_i = (e_{ij})'_{1 \le j \le n_i}$ where $1 \le i \le m$. The dependance between the units are modeled with a Gaussian noise $Zu + e$ where $Z$ is for the groups or domains, such as finally, its variance is the diagonal matrix compound of the blocks $V_i = \sigma_u^2 \mathbf{1}_{n_i} \mathbf{1}'_{n_i} + \sigma_e^2 I_{n_i}$.

*b) Prediction:* The mean $\theta_i$ is the main quantity of interest in small area estimation as one wants to know the averages of $y$ per each area. Note that under the BHF model, an estimator of the population mean is:

$$\bar{Y}_i \approx \bar{X}_i \beta + u_i = \theta_i \, .$$

In practice this quantity is unknown for most of the area, and must be estimated from a sample. Here $\bar{X}_i$ is the mean vector of the known population covariates $(x_{ij})$ for the $i$-th area.

*c) Estimator EBLUP:* The estimation of the mean in an area is based on the theory of mixed model. When it is supposed that the variance components $\alpha = (\sigma_u^2, \sigma_e^2)'$ are known, the usual estimator of the mean in each area is called BLUP and denoted $\tilde{\theta}_i$ for the best linear unbiased predictor of $\theta_i$. This supposes that it is unbiased and its variance is minimum among all the linear mean estimators considering the covariates and the available sample. The variance matrices $V_i$ leads to the regression coefficients which are solution of an usual generalized regression model written:

$$\tilde{\beta}(\alpha) \;\; = \;\; \left(\sum_i x_i' x_i - \gamma_i \bar{x}_i' \bar{x}_i\right)^{-1} \left(\sum_i x_i' y_i - \gamma_i \bar{x}_i' \bar{y}_i\right) . \tag{1}$$

This leads to the BLUP (best linear unbiased estimator) for the small area mean $\theta_i$ as follows:

$$\theta_i(\alpha) = \gamma_i \bar{y}_i + \left(\bar{X}_i - \gamma_i \bar{x}\right) \tilde{\beta} \, . \tag{2}$$

Note that is an area was empty, the estimator would be $\bar{X}_i \tilde{\beta}$ because an usual estimate of the random effect is not available. Here the quantite involves are the random effect and the shrinkage factor:

$$\begin{aligned} u_i &= \gamma_i \left(\bar{y}_i - \bar{x}_i \tilde{\beta}\right) && (3) \\ \gamma_i &= \sigma_u^2 / (\sigma_u^2 + \sigma_e^2 / n_i) \, . && (4) \end{aligned}$$

In practice, the variance components are estimated by restricted maximum likelihood (ML), by maximum likelihood maximum procedure (REML) or by the fitting of constants method. When replacing $\sigma_u^2$ and $\sigma_e^2$ by the estimates $\hat{\sigma}_u^2$ and $\hat{\sigma}_e^2$, in $V_i$, $u_i$, and $\gamma_i$, ones gets their estimated versions $\hat{V}_i$, $\hat{u}_i$ and $\hat{\gamma}_i$. Moreover, an estimator for the vector of regression coefficients, $\hat{\beta} = \tilde{\beta}(\hat{\alpha})$ with $\hat{\alpha} = (\hat{\sigma}_u^2, \hat{\sigma}_e^2)'$, is obtained as in (1) by replacing the true variances for the random effects and the noises by their estimates. Finally, having these estimates, the Empirical EBLUP or EBLUP of the small area mean, $\hat{\theta}_i = \theta_i(\hat{\alpha})$ is obtained as the BLUP by replacing $\tilde{\beta}$ by $\hat{\beta}$ and $\gamma_i$ by $\hat{\gamma}_i$. The estimation of the random effect $\hat{u}_i$ is similarly written than $u_i$ by replacing the parameters involved by their estimators such that for the estimated shrinking factor $\hat{\gamma}_i$. When the sampling faction are not too small, it can be used an alternative version $\hat{\theta}_i^f$ which uses only the non observed population for the averaged covariates. The sample mean of the explicative variables is denoted $\bar{x}_i = {n_i}^{-1} \sum_j^{n_i} x_{ij}$. In the following, the sample size $n_i$ are supposed different to zero, eventually equal to one, as usually.

The estimator $\hat{\theta}_i$ may be improved by benchmarking when the synthetic population mean estimator obtained from the model is made equal to the estimator from the usual sampling theory.

### *B. Benchmarking*

For statistical studies in survey methodology it is required that the means obtained for each area from a model-based approach sum up with the mean estimate from a design-based approach. An estimator of the population total having the benchmarking properties is such as:

$$\tilde{T}_y = \sum_i N_i \, \hat{\theta}_i = \hat{T}_y \, .$$

Here the quantity which is approximated is the unknown population total $T_y$ while $\hat{T}_y$ is a design-based estimator of the unknown true total $T_y$, for instance $N \bar{y}_s$ with $\bar{y}_s$ the usual design based estimator of the mean for the whole sample. See for instance [4] for a general introduction to benchmarking in small areas. In the next section, it is reviewed several approaches proposed previously in the literature for the unit level model and a new proposed method is discussed.

## III. Method via supplementary column and row

For the property of benchmarking it is proposed to add a row and a column to the design matrix of the GLS when the regression coefficients are computed in order to force the matching between the estimate of population total from the BHF model and the design-based total one.

### A. Matricial expression

We get the system with the noise denoted $v_s$ for the samples coming after marginalizing out the random effects:

$$\begin{aligned}\underbrace{\begin{pmatrix} Y_s \\ y_+ \end{pmatrix}}_{Y_{sa}} &= \begin{pmatrix} X_s & w'_a \\ x'_a & w_+ \end{pmatrix}\begin{pmatrix} \beta_{aw} \\ \alpha_{aw} \end{pmatrix} + \begin{pmatrix} v_{sw} \\ v_+ \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} X_{sw} \\ x'_{aw} \end{pmatrix}}_{X_{aw}}\underbrace{\begin{pmatrix} \beta_{sw} \\ \alpha_{aw} \end{pmatrix}}_{\beta_{aw}} + \underbrace{\begin{pmatrix} v_{sw} \\ v_+ \end{pmatrix}}_{v_{aw}} .\end{aligned} \tag{5}$$

It is supposed that the auxiliary variables of the population are completed by a new component with the column:

$$w'_a = (w_{a1}, w_{a2}, \cdots, w_{an}) = (\tilde{w}_{ij}) \text{ where } \tilde{w}_{ij} = f_i^{-1} - 1 ,$$

such as the new $(ij)^{\text{th}}$ row of the new design matrix for the explicative variables with for an *horizontal augmentation*, is:

$$x_{ijw} = (x'_{ij}, \tilde{w}_{ij})' .$$

It is then also added a $(n+1)^{\text{th}}$ row to the system for a *vertical augmentation*, with $x_{aw} = (x_a, w_+)'$ in the design matrix, and $y_+$ in $Y_s$. Let have then $\bar{X}_{irw}$ which comes with the same column augmentation than the samples by adding the quantity $f_i^{-1} - 1$ at the right of the vector $\bar{X}_{ir}$. Let also have $\bar{x}_{iw}$ for the sample mean of the augmented covariates $x_{ijw}$ defined just above per area. Then it is clear that the following quantities are suitable:

$$\begin{aligned} x_a &= \sum_{i=1}^{m}(N_i - n_i)\left\{-\bar{X}_{ir} + (2\hat{\gamma}_i - 1)\bar{x}_i\right\} , \\ y_+ &= \sum_{i=1}^{m}\left\{(2\hat{\gamma}_i - 1)(N_i - n_i) + n_i\left(1 - \frac{N}{n}\right)\right\}\bar{y}_i , \\ w_+ &= 2\sum_{i=1}^{m}\left\{(\hat{\gamma}_i - 1)\frac{(N_i - n_i)^2}{n_i}\right\} . \end{aligned}$$

The new variance matrix of the noise in the clusters $v_{aw}$ is $\hat{V}_{aw} = diag(\hat{V}, w_+\hat{\sigma}_e^2)$. It is added only a new diagonal element for the variance for the new row in the system, note that it is negative but this has not been a problem in the simulations. A negative variance term may be a limit of the proposal for the understanding of the way how the constraint work but has not been problematic for the practical experiments considered in the dedicated section, with the estimator $\hat{\beta}_{aw}$ evaluated as the previous vector $\hat{\beta}$ for the non augmented model, but with the new sytem by changing $X_s$ into $X_{aw}$, $Y_s$ into $Y_{sa}$ and $\hat{V}$ into $\hat{V}_{aw}$. With this solution for the new system, the new random effects $\hat{u}_{iw}$ are also written as the usual estimator but for the extended matrix of the covariates. The resulting estimator is as follows:

$$\hat{\theta}_i^{PSW} = f_i\bar{y}_i + (1 - f_i)\left\{\bar{X}_{irw}\hat{\beta}_{aw} + \hat{u}_{iw}\right\} .$$

The benchmarking of the estimators of the mean inside each area can be justified via the orthogonality of the residuals with the last column of the system. When $x_{iw}$ is the augmented matrix $x_i$ with a column having all same elements equal to $w_i = f_i^{-1} - 1$ and $w_{ai} = w_i\mathbf{1}_{n_i}$, considering the orthogonality with the new column of the residuals it is obtained the following benchmarking equation because:

$$w'_a V^{-1}(Y_s - X_{sw}\hat{\beta}_{aw}) + w_+(w_+\hat{\sigma}_e^2)^{-1}(y_+ - x'_{aw}\hat{\beta}_{aw}) \;\; = \;\; 0 .$$

This results into the wanted result, with the sum of the design estimates which is equal to the sum ot the synthetic estimates per area. Note that the new regression coefficients may be written as a function of the original eblup, as explained just above.

### B. Explicit expression

Let denote $s_a = w_+\hat{\sigma}_e^2$ and $h_a = x'_{aw}\hat{V}_s^{-1}x_{aw}$. A more explicit formula for the regression coefficients in the proposed benchmarked mean estimator may be written as follows:

$$\hat{\beta}_{aw} \;=\; \hat{\beta}_w + \frac{1}{s_a + h_a}\hat{C}_s^{-1}x_{aw}\left(y_+ - x'_{aw}\hat{\beta}_w\right).$$

Here, the vector of regression coefficients $\hat{\beta}_w$ is obtained for only one new column added to the system, and is different from $\hat{\beta}_{sw}$ which required the additional row to be computed. An explicit expression of $\hat{\beta}_w$ as a function of the former vector $\hat{\beta}$ for the regular eblup, with $\hat{C}_s = (X'_s\hat{V}^{-1}X_s)$, may be as follows:

$$\hat{\beta}_w \;=\; \begin{pmatrix} \hat{\beta} \\ 0 \end{pmatrix} + \begin{pmatrix} \frac{1}{k}\hat{C}_s^{-1}bb'\hat{\beta} - \frac{d}{k}\hat{C}_s^{-1}b \\ -\frac{1}{k}b'\hat{\beta} + \frac{d}{k} \end{pmatrix}$$

Where,

$$\begin{aligned}
d = w'V_s^{-1}Y_s \quad &= \quad \tfrac{N}{\hat{\sigma}_e^2} \times \textstyle\sum_{i=1}^m \tfrac{N_i}{N}(1 - f_i)(1 - \hat{\gamma}_i)\bar{y}_i \\
b = w'\hat{V}_s^{-1}X_s \quad &= \quad \tfrac{N}{\hat{\sigma}_e^2}\textstyle\sum_{i=1}^m \tfrac{N_i}{N}(1 - f_i)(1 - \hat{\gamma}_i)\bar{x}_i \\
c = w'\hat{V}_s^{-1}w \quad &= \quad \hat{\sigma}_e^{-2}\textstyle\sum_i \tfrac{(N_i - n_i)^2}{n_i}(1 - \hat{\gamma}_i) \\
k = c - b'C_s^{-1}b \quad &= \quad \tfrac{N^2}{\hat{\sigma}_e^2}\textstyle\sum_{i=1}^m \tfrac{N_i^2}{N^2}\tfrac{(1-f_i)^2}{n_i}(1 - \hat{\gamma}_i) - b'\hat{C}_s^{-1}b
\end{aligned}$$

$$\begin{aligned}
\tfrac{1}{k}\hat{C}_s^{-1}b(b'\hat{\beta} - d) \quad &= \quad -\tfrac{1}{k\hat{\sigma}_e^2}\hat{C}_s^{-1}b \times \textstyle\sum_{i=1}^m(N_i - n_i)(1 - \hat{\gamma}_i)(\bar{y}_i - \bar{x}'_i\hat{\beta}) \\
-\tfrac{1}{k}b'\hat{\beta} + \tfrac{d}{k} \quad &= \quad \tfrac{1}{k\hat{\sigma}_e^2}\textstyle\sum_{i=1}^m(N_i - n_i)(1 - \hat{\gamma}_i)(\bar{y}_i - \bar{x}'_i\hat{\beta})
\end{aligned}$$

These explicit expressions may induce that the component for the additional column might be small at least for enough large values of the $n_i$ as $\bar{y}_i \approx \bar{x}'_i\hat{\beta}$, and not too small value of $\hat{\sigma}_e^2$.

### C. Alternative restricted predictors

If the BHF model is directly benchmarked at the level of the parameter inference for the regression coefficients before evaluation of the random effects, the new estimators is written similarly but with a new vector of coefficients. Approaches can be separated into the internal or external ones when they alter the model for inducing that the weighted sum with the synthetic estimators agree with the constraints, or when they alter the estimates coming without benchmarking for a correction of the values and insure the same constraints.

- In [5], Pfeffermann and Barnard have introduced a restricted version of a linear system (see [6]) for the estimation of the eblup. The later was shown to lead exactly to the same solution for the regression coefficients and the random effects. See also [3] for an expression when the sampling fractions are non negligible and denoted $\hat{\theta}_i^{PB}$ hereafter. When the variance components are estimated, for the $i^{\text{th}}$ area. It has been shown to be unbiased, and extended in [7], [8] for business surveys by considering also several regions which aggregated small areas. Note that constraining directly the profile likelihood may be also possible.
- In the ratio estimator, the benchmarking is corrected proportionally to each synthetic estimator by adding multiplicative factor obtained by divinding the sample total estimate with the one from the synthetic mean area predictors, say $\hat{\theta}_i^{RT} = \hat{T}_y\tilde{T}_y^{-1}\hat{\theta}_i$. Despite its drawback to correct with a same factor the area mean estimates, it is widely used.
- As explained in [9], another form of benchmarking is the additive one which is defined by looking for a new estimator which minimizes the distance with the unbenchmarked one under the restriction for the total. This leads to an additive correction, where the different additional quantities can have diverse orders of magnitude on the contrary to the ratio case. In [2], a benchmarked estimator is given at the last part of the article which considers the variance for free quantities $a_i$. It can be choosen an estimate of the mean square error of the un-benchmarked eblup, denoted $\hat{\theta}_i^{VAR}$ hereafter.

## IV. Empirical study

The empirical study is divided into the checking of the approach via a simulation exercice and an application of the estimator to one dataset. The simulation exercise with the population described in the previous sub-section was carried out using R for windows. The parameters for the variance components are estimated by the fitting of constant method at a first step while the regression parameters come from the generalized least-squares problem under the constraints.

### A. Indicators

A main purpose of the small area approach is to construct estimators with the propery of unbiasness as most of possible and with a reduced variance in comparison to the result with design-based methods. The benchmarking is often described in the litterature as able to reduce and eventual bias for reason of misspecification of the model for instance. Therefore we evaluate the bias and also the variability of the estimations per area in the simulation. The indicators considered herein are usual, they are the average absolute relative bias (AARB) and the average relative mean square error (ARMSE):

$$\begin{aligned} AARB &= m^{-1}\sum_{i=1}^{m}\frac{1}{B}\sum_{b=1}^{B}\left|\hat{\theta}_i^{(b)}/\bar{Y}_i^{(b)}-1\right| \\ ARMSE &= m^{-1}\sum_{i}^{m}\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_i^{(b)}/\bar{Y}_i^{(b)}-1\right)^2 . \end{aligned}$$

Such estimators are averaging over all the small area the obtained absolute values of the bias per area, and also their values raised to the square for a variance principle. The idea is to measure if the obtained estimates from the sample are really near the true values from the population, the main purpose to achieve, and by prefering also a relative distance there. Note that an indicator as the proportion of times when the altered estimator is greater than the eblup could give an idea of the sign of the eventual bias is called $BIASG$, while the estimator of the cumulated bias $BIASR$ is written as AARB without the absolute value.

### B. Model based simulation

To study the estimators, synthetic populations are generated and a sample is used for the parameters estimation. The model of population is the BHF one, this is written with a two-sampling scheme. Here, the vector of covariates with first component for the intercept, $x_{ij}$ are generated from random variables and fixed during the whole section. The values for the number of units per area in the population, and the sample, and the values for the true regression coefficients are listed in the table above:

$$\begin{aligned} y_{ij} &= \mu_{ij} + e_{ij} \\ \mu_{ij} &= x_{ij}'\beta + u_i \, . \end{aligned} \qquad \begin{aligned} \alpha &= (4.0, 0.2) \\ \beta &= (1.00, 0.25, 0.35, 0.45) \\ N_i &\in (100\mathbf{1}_{10}', 250\mathbf{1}_{10}', 500\mathbf{1}_{10}') \\ n_i &\in (\mathbf{1}_5', \mathbf{31}_5', \mathbf{51}_5', \mathbf{51}_5', 10\mathbf{1}_5', 15\mathbf{1}_5', 20\mathbf{1}_5') \end{aligned}$$

TABLE I

Descriptive for the populations and samples generated.

Note that $p = 3$, $N = 8500$ and $n = 270$. The resulting sampling fractions are $n_i/N_i \in (1\%, 2\%, 3\%, 4\%)$, while the true shrinking factors $\gamma_i$ are all different from one. A population is generated $B = 200$ times for the $m = 30$ areas.

It must be noticed that the sampling is model based, and not design-based. In this setting, one unique population of covariates is generated such that the set of units is fixed for the population. A unique sample of units $s$ is drawn by simple random sampling, and a numerical procedure commonly found in the literature permits to evaluate the variability of the small area predictors. Precisely, B populations are generated while the sample of units is kept unchanged, while the whole covariates remain also. Only the noises at the level of the units and the level of the random effects are realy updated.

The obtained results are summarized in the following table with percents for the three first rows:

| | $\hat{\theta}_i^f$ | $\hat{\theta}_i^{PB}$ | $\hat{\theta}_i^{VAR}$ | $\hat{\theta}_i^{RT}$ | $\hat{\theta}_i^{PSW}$ |
|---|---|---|---|---|---|
| BIASG | 0.50 | 0.49 | 0.48 | 0.49 | 0.50 |
| BIASR | $-3.1\,10^{-7}$ | $-6.72\,10^{-6}$ | $-1.27\,10^{-5}$ | $-2.63\,10^{-6}$ | $1.12\,10^{-5}$ |
| AARB | $1.07\,10^{-3}$ | $2.00\,10^{-3}$ | $3.23\,10^{-3}$ | $1.42\,10^{-3}$ | $4.35\,10^{-3}$ |
| ARMSE | $7.0\,10^{-8}$ | $1.8\,10^{-7}$ | $7.7\,10^{-7}$ | $7.0\,10^{-8}$ | $1.58\,10^{-7}$ |

The statistics considered in the simulation are for evaluation of the bias, the precision, and the variability. Finally, for this dataset, the synthetic estimators behave very similarly. Note that for empty areas, this may be different because the random effects are corrected differently.

### C. *Results for real data*

An example of obtained predictors is presented in the table below for the five synthetic model-based estimators of the mean and the design-based one.

| Area | $\bar{Y}_i$ | $\bar{y}_i$ | $\hat{\theta}_i^f$ | $\hat{\theta}_i^{PB}$ | $\hat{\theta}_i^{VAR}$ | $\hat{\theta}_i^{RT}$ | $\hat{\theta}_i^{PSW}$ |
|---|---|---|---|---|---|---|---|
| 1 | 910.530 | 910.971 | 910.461 | 910.424 | 910.354 | 910.421 | 910.646 |
| 2 | 115.470 | 115.698 | 115.441 | 115.404 | 115.334 | 115.436 | 115.457 |
| 3 | 128.838 | 130.078 | 128.903 | 128.866 | 128.796 | 128.898 | 128.863 |
| 4 | 1138.748 | 1138.530 | 1138.743 | 1138.704 | 1138.635 | 1138.693 | 1138.458 |
| 5 | 728.990 | 727.418 | 729.146 | 729.108 | 729.039 | 729.114 | 729.030 |
| 6 | 717.874 | 716.840 | 717.868 | 717.830 | 717.760 | 717.837 | 717.578 |
| 7 | 600.181 | 599.380 | 600.260 | 600.222 | 600.153 | 600.234 | 599.969 |
| 8 | 1158.484 | 1158.231 | 1158.429 | 1158.391 | 1158.322 | 1158.378 | 1158.351 |
| 9 | 55.291 | 55.042 | 55.346 | 55.308 | 55.239 | 55.344 | 55.251 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 24 | 1285.844 | 1286.076 | 1285.918 | 1285.881 | 1285.897 | 1285.862 | 1285.945 |
| 25 | 316.396 | 316.290 | 316.420 | 316.382 | 316.398 | 316.406 | 316.404 |
| 26 | 1554.418 | 1554.137 | 1554.405 | 1554.367 | 1554.383 | 1554.337 | 1554.292 |
| 27 | 1551.462 | 1551.563 | 1551.464 | 1551.427 | 1551.443 | 1551.397 | 1551.439 |
| 28 | 1430.383 | 1430.619 | 1430.383 | 1430.344 | 1430.361 | 1430.320 | 1430.272 |
| 29 | 158.516 | 158.223 | 158.514 | 158.476 | 158.493 | 158.507 | 158.450 |
| 30 | 994.399 | 994.557 | 994.442 | 994.404 | 994.421 | 994.399 | 994.427 |

The four ones which are benchmarked have their weigthed sum $\sum_i N_i \hat{\theta}_i$ which is equal to the total $N\bar{y}_s = 7385504$, while the true total is 7385717. Note that the estimator of the total from the eblup is not equal to the estimated sum as expected. On this dataset, the values are not very different actually, as the $N_i$ are large, only the decimals of the estimators need to be updated finally.

## V. Conclusion and discussion

Herein, benchmarking the linear nested regression model is developped via a new augmented system. The approach is simply adding a row and a column to the design matrix. An explicit expression for the new regression coefficients is also provided in closed-form. The estimator is compared with several ones from the litterature with several datasets. Note that augmenting directly the system considered just before the the Pfeffermann-Barnard method is not considered hereing as internal benchmarking is the main objective. A future perspective is the variance for the estimator which may be derived from the existing litterature. Another perspective is to find alternative augmentation schemes by adding several columns instead of just one for instance or for benchmarking with aggregated areas.

# References

[1] JNK Rao, *Small Area Estimation*, Wiley series in survey methodology, John Wiley and Sons, New York, 2003.

[2] George E. Battese, Rachel M. Harter, and Wayne A. Fuller, "An error-components model for prediction of county crop areas using survey and satellite data", *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 28–36, 1988.

[3] R. Priam and N. Shlomo, "Calibration of small area estimates in business surveys", in *Small area estimation*, 2011.

[4] D. Pfeffermann, "New important developments in small area estimation", Report, Southampton, GB, Southampton Statistical Sciences Research Institute, 2010.

[5] D. Pfeffermann and C.H. Barnard, "Some new estimators for small-area means with application to the assessment of farmland values", *Journal of Business & Economic Statistics*, pp. 73–84, 1991.

[6] D. Pfeffermann, "On extensions of the gauss-markov theorem to the case of stochastic regression coefficients", *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 46, no. 1, pp. pp. 139–148, 1984.

[7] M. D. Ugarte, A. F. Militino, and T. Goicoa, "Adjusting economic estimates in business surveys", *Journal of Applied Statistics*, vol. 35, no. 11, pp. 1253–1265, 2008.

[8] M.D. Ugarte, A.F. Militino, and T. Goicoa, "Benchmarked estimates in small areas using linear mixed models with restrictions", *Test*, vol. 18, pp. 342–364, 2009.

[9] J. Wang, W. A. Fuller, and Y. Qu, "Small area estimation under a restriction", *Survey Methodology*, vol. 34, pp. 29–36, 2008.