NRC Research Press

ARTICLE

Design-based properties of some small-area estimators in forest inventory with two-phase sampling

Daniel Mandallaz

Abstract: We consider the small-area estimation problem for forest inventories with two-phase sampling schemes. We propose an improvement to the synthetic estimator, when the true mean of the auxiliary variables over the small area is unknown and must be estimated, and likewise to the residual corrected small-area estimator. We derive the asymptotic design-based variances of these new estimators, the pseudo-synthetic and pseudo-small-area estimators, by also incorporating the design-based variance of the regression coefficients. We then propose a very simple mathematical device that transforms pseudo-small-area estimators into pseudo-synthetic estimators, which is very convenient for deriving asymptotic variances. The results are extended to cluster and two-stage sampling at the plot level. A case study and a simulation illustrate the theory.

Résumé : Nous considèrons le problème de l'estimation pour petits domaines dans le contexte d'inventaires forestiers en deux phases. Nous proposons une amélioration simple de l'estimateur synthétique quand la moyenne des variables auxiliaires dans le petit domaine doit être estimée en premier lieu, de même pour l'estimateur pour petit domaine basé sur les résidus. Nous calculons la variance sous le plan de sondage de ces nouveaux estimateurs en tenant compte de la variance des coefficients de régression. De plus, nous proposons un artifice mathématique qui permet de transformer un estimateur pour petit domaine en un estimateur synthétique, ce qui simplifie le calcul de la variance asymptotique. L'extension aux sondages par satellites et deux degrés au niveau de la placette est aussi traitée. Un exemple concret et une simulation illustrent la théorie.

1. Introduction

The motivation for this work is the increasing need to use national or regional inventories for local estimation to meet tighter budgetary constraints, which is only feasible under extensive use of auxiliary information provided, e.g., by remote sensing.

In this paper, we state only the most important results needed for practical use. The mathematical statistician will find the proofs together with further developments, in particular the equivalent g-weight formulation of the estimators, in Mandallaz (2012), which can be downloaded from the ETH library.

There is an immense literature in general survey sampling on the problem of small-area estimation (which also entails the so-called small-domain or subpopulation estimation problem). The book of Rao (2003) is an excellent but demanding reference essentially devoted to socioeconomical problems with finite populations; therefore, it is not directly applicable to many issues in the forest inventory context where the Monte Carlo or infinite population approach is now regarded as the best suitable framework for the classical dendrometrical quantities such as timber volume per surface area (see e.g., Mandallaz (2008)). In comparison, the small-area literature in the forest inventory context is of course smaller but still hardly manageable. The reader can consult Koehl et al. (2006, section 3.8) for a good specific review of small-area estimation in forest inventory. Broadly speaking, most of the procedures used in forest inventory can be classified as follows: (1) model-dependent (sometimes also called model-based) procedures, including kriging techniques (see, e.g., Lappi (2001), Mandallaz (1993, 2000), and Mandallaz (2008, chapters 7 and 8)) or mixed models with components of variances (see, e.g., Breidenbach and Nothdurft (2010) and Goerndt et al. (2011)); (2) nonparametric design-based procedures such as the nearest neighbors techniques (see, e.g., Breidenbach and Astrup (2012), McRoberts (2012), and Bafetta et al. (2011)); (3) the Bayesian approach (Finley et al. 2008); and (4) classical design-based regression estimators in two-phase sampling (Mandallaz 2008, section 6.3). From a mathematical point of view, the kriging techniques have the advantage of coherence because extensive forest inventories are performed with systematic grids. However, they are far more difficult to use. Most of the aforementioned techniques require the use of sophisticated software tools (e.g., maximum likelihood, numerical integration, re-sampling techniques, choice of tuning constants, etc.) with the associated "black-box" shortcomings.

In this paper, we investigate the properties of the classical regression estimators in the model-assisted framework (according to the terminology of Särndal et al. (2003)) in which prediction models are used to improve the efficiency but are not assumed to be correct as in the model-dependent approach. The validity of the statistical procedures is ensured by the randomization principle, i.e., we are still in the design-based inference framework (item 4 in the above classification), which has, of course, a definite advantage in official statistics. The Monte Carlo approach to forest inventory is used throughout. The paper presents several novelties: (1) the design-based variance of the regression coefficients is systematically taken into account, which leads to a new derivation of the celebrated g-weight technique (preliminary and to some extent heuristic results are given in Mandallaz (2008, chapter 7); (2) the standard synthetic and small-area estimators, which require the auxiliary information to be exhaustive or "wall-to-wall", are extended to pseudo-synthetic and pseudo-small-area estimators in which the auxiliary information is not exhaustive; (3) extending the model with the indicator variables of the small areas allows residual-based small-area estimates to transform into residual-free synthetic estimates, which greatly simplifies the calculation of the variance; (4) the results are generalized to cluster sampling and two-stage Poisson sampling at the plot level; and last but not least, (5) the techniques can be generalized to the case

Received 10 September 2012. Accepted 27 February 2013.

D. Mandallaz. Chair of Land Use Engineering, ETH Zurich, CH 8092 Zurich, Switzerland.
Corresponding author: Mandallaz (e-mail: daniel.mandallaz@env.ethz.ch).

Can. J. For. Res. **43**: 441–449 (2013) dx.doi.org/10.1139/cjfr-2012-0381

Published at www.nrcresearchpress.com/cjfr on 28 February 2013.

where one part of the auxiliary information is exhaustive and another part is not (work near completion). It must be emphasized that the proposed procedures can be easily implemented in standard software packages such as SAS or R: they require only the multiplication and inversion of matrices.

2. Methodology

Let us now define the sampling scheme. The first phase draws a large sample s_1 of n_1 points that are independently and uniformly distributed within the forest area F. At each point $x \in s_1$, auxiliary information is collected, very often coding information of a qualitative (e.g., following the interpretation of aerial photographs) or quantitative (e.g., timber volume estimates based on LIDAR measurements) nature. We shall assume that the auxiliary information at point x is described by the column vector $\mathbf{Z}(x) \in \Re^p$.

The second phase draws a small sample $s_2 \subset s_1$ of n_2 points from s_1 according to equal probability sampling without replacement. In the forested area F, we consider a well-defined population P of N trees with response variable Y_i, i = 1, 2, …, e.g., the timber volume. The objective is to estimate the overall spatial mean $\bar{Y} = \frac{1}{\lambda(F)}\sum_{i=1}^{N} Y_i$, where $\lambda(F)$ denotes the surface area of F (usually in hectares) and the mean over a small area $G \subset F$, defined as

$$[1] \qquad \bar{Y}_G = \frac{1}{\lambda(G)}\sum_{i=1}^{N} I_G(i)Y_i =: \frac{1}{\lambda(G)}\sum_{i\in G} Y_i$$

where the indicator variable $I_G(i)$ is 1 if tree i lies in G, and 0 otherwise.

For each point $x \in s_2$, trees are drawn from the population P with probabilities π_i, for instance, with concentric circles or angle count techniques. The set of trees selected at point x is denoted by $s_2(x)$. From each of the selected trees $i \in s_2(x)$, one determines Y_i. The indicator variable I_i is defined as

$$[2] \qquad I_i(x) = \begin{cases} 1 & \text{if } i \in s_2(x) \\ 0 & \text{if } i \notin s_2(x) \end{cases}$$

At each point $x \in s_2$, the terrestrial inventory provides the local density $Y(x)$:

$$[3] \qquad Y(x) = \frac{1}{\lambda(F)}\sum_{i=1}^{N} \frac{I_i(x)Y_i}{\pi_i} = \frac{1}{\lambda(F)}\sum_{i\in s_2(x)} \frac{Y_i}{\pi_i}$$

The term $\frac{1}{\lambda(F)\pi_i}$ is the tree extrapolation factor f_i with dimension per hectare. One must include possible boundary adjustments, $\lambda(F)\pi_i = \lambda(F \cap K_i)$, where K_i is the inclusion circle of tree i. In the infinite population or Monte Carlo approach, one samples the function $Y(x)$ for which the following important relation holds:

$$[4] \qquad \mathbb{E}_x(Y(x)) = \frac{1}{\lambda(F)}\int_F Y(x)\mathrm{d}x = \frac{1}{\lambda(F)}\sum_{i=1}^{N} Y_i = \bar{Y}$$

where $\mathbb{E}_x$ denotes the expectation with respect to a random point x uniformly distributed in F. This establishes the link between the infinite population (continuum) $\{x \in F|Y(x)\}$ and the finite population of trees $\{i = 1, 2, \ldots, N|Y_i\}$. The reader unfamiliar with the Monte Carlo approach can consult for a first perusal Mandallaz (2008) or Mandallaz and Ye (1999).

Boundary adjustments are usually performed only with respect to F and not with respect to the small area G. However, we shall assume that we also have

$$[5] \qquad \bar{Y}_G = \frac{1}{\lambda(G)}\int_G Y(x)\mathrm{d}x$$

The aforementioned randomization principle assumes that we have uniformly independently distributed points or clusters in the forested area F, whereas in practice, systematic grids are used. There is reasonable theoretical and empirical evidence that treating systematic grids as simple random samples is acceptable for point estimation and also for variance estimation (which will be, in most instances, slightly overestimated) for extensive forest inventories. From a mathematical point of view, the only correct and most efficient procedure, particularly for small-area estimation, is the geostatistical double-kriging technique (see Mandallaz (2008, chapter 7) for a brief introduction and Mandallaz (1993, 2000) for the mathematical aspects). However, double kriging is far more difficult to use than the relatively simple design-based procedures presented here and, as all model-dependent approaches, rests upon assumptions such as stationarity, which can be difficult to assess.

3. The model

We consider the linear model (the superscript t on vectors or matrices denotes the transposition operator)

$$[6] \qquad Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta} + R(x)$$

In the model-dependent approach, the point x is fixed and $R(x)$ is a random variable with zero mean and a given covariance structure. In the design-based approach, $Y(x)$, $\mathbf{Z}(x)$, and $R(x)$ are random variables because x is random and the true regression coefficient $\boldsymbol{\beta}$ is the theoretical least squares estimate minimizing

$$[7] \qquad \int_F R^2(x)\mathrm{d}x = \int_F (Y(x) - \boldsymbol{Z}^t(x)\boldsymbol{\beta})^2\mathrm{d}x$$

It satisfies the normal equation

$$[8] \qquad \left(\int_F \boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\mathrm{d}x\right)\boldsymbol{\beta} = \int_F Y(x)\boldsymbol{Z}(x)\mathrm{d}x$$

and the orthogonality relationship

$$[9] \qquad \int_F R(x)\boldsymbol{Z}(x)\mathrm{d}x = \mathbf{0}$$

We will assume that $\boldsymbol{Z}(x)$ contains the intercept term 1 or, more generally, that the intercept can be expressed as a linear combination of the component of $\boldsymbol{Z}(x)$, which ensures that the mean residual is zero, i.e.,

$$[10] \qquad \int_F R(x)\mathrm{d}x = 0$$

The important case of stratification amounts to taking $\boldsymbol{Z}^t(x) = (I_{F_1}(x), I_{F_2}(x), \ldots, I_{F_L}(x))$, where $F = \cup_{k=1}^{L} F_k$ and $I_{F_k}(x)$ is the zero–one indicator variable of the stratum k, F_k.

To simplify the notation let us set $\boldsymbol{A} = \mathbb{E}_x\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)$ and $\boldsymbol{U}(x) = Y(x)\boldsymbol{Z}(x)$. The normal equation then reads

$$[11] \qquad \boldsymbol{A\beta} = \mathbb{E}_x\boldsymbol{U}(x) := \boldsymbol{U}$$

Of course, only a sample-based normal equation is available, i.e.,

$$[12] \qquad \boldsymbol{A}_{s_2}\widehat{\boldsymbol{\beta}}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{U}(x) = \boldsymbol{U}_{s_2}$$

where we have set $\boldsymbol{A}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)$ and $\boldsymbol{U}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}Y(x)\boldsymbol{Z}(x)$. The theoretical and empirical regression vector parameters are

$$[13] \qquad \begin{aligned}\boldsymbol{\beta} &= \boldsymbol{A}^{-1}\boldsymbol{U}\\ \widehat{\boldsymbol{\beta}}_{s_2} &= \boldsymbol{A}_{s_2}^{-1}\boldsymbol{U}_{s_2}\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{s_2}$ is asymptotically design-unbiased for $\boldsymbol{\beta}$ by the law of large numbers (LNN). The design-based variance–covariance matrix of the regression coefficients is defined as

$$[14] \qquad \Sigma_{\hat{\boldsymbol{\beta}}_{s_2}} = \mathbb{E}(\widehat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta})^t$$

We define the empirical predictions and residuals by $\hat{Y}(x) = \boldsymbol{Z}^t(x)\widehat{\boldsymbol{\beta}}_{s_2}$ and $\hat{R}(x) = Y(x) - \hat{Y}(x)$, respectively. By using the Taylor linearization technique (see Mandallaz (2008, pp. 124–125) or Mandallaz (2012)), one obtains the asymptotically consistent estimate of the design-based variance of $\widehat{\boldsymbol{\beta}}_{s_2}$:

$$[15] \qquad \widehat{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}} := \boldsymbol{A}_{s_2}^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\right)\boldsymbol{A}_{s_2}^{-1}$$

Interestingly, this is precisely the robust estimate of the model-dependent covariance matrix discussed in a totally different context by Huber (1967) and Gregoire and Dyer (1989).

We have the sample orthogonality relations

$$[16] \qquad \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x)\boldsymbol{Z}(x) = \boldsymbol{0}, \quad \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x)\hat{Y}(x) = 0$$

in particular, the zero mean residual property $\frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x) = 0$.

4. The estimators

4.1. External models

If the model is external, i.e., not fitted with the inventory data at hand, the regression estimate is defined as

$$[17] \qquad \hat{Y}_{\text{reg}} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_0(x) + \frac{1}{n_2}\sum_{x\in s_2}R_0(x)$$

with the predictions $\hat{Y}_0(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0$ and the residuals $R_0(x) = Y(x) - \hat{Y}_0(x)$, where $\boldsymbol{\beta}_0$ is the given external regression coefficient, ideally obtained from another similar inventory. Note that in this case, the mean residual will not necessarily be zero. The estimator $\hat{Y}_{\text{reg}}$ has a very intuitive form: it is simply the mean of the predictions, corrected by the mean of the residuals. Unbiased estimates of the design-based variance of $\hat{Y}_{\text{reg}}$ are given in Mandallaz (2008, 2012).

The estimation for any small area $G \subset F$ is straightforward, indeed one restricts the samples of n_1 and n_2 points in F to the $n_{1,G}$ and $n_{2,G}$ points in G.

4.2. Internal models

In most applications, the model has to be fitted with the data provided by the current inventory. In this case, the model is said to be internal. In very large samples, one can treat an internal model as external and apply again the formulae given above, which obviously neglects the error in the regression coefficients. This is essentially the framework presented in Mandallaz (2008, chapter 5 and section 6.3). In the present paper, we show how one can take the design-based variance of the regression coefficients into account, albeit still in large samples.

The model-dependent estimator for the small area G is called the synthetic estimator and is given by

$$[18] \qquad \hat{Y}_{G,\text{synth}} = \frac{1}{\lambda(G)}\int_G \hat{Y}_{s_2}(x)\text{d}x = \frac{1}{\lambda(G)}\int_G \boldsymbol{Z}^t(x)\widehat{\boldsymbol{\beta}}_{s_2}\text{d}x = \overline{\boldsymbol{Z}}_G^t\widehat{\boldsymbol{\beta}}_{s_2}$$

where $\overline{\boldsymbol{Z}}_G = \frac{1}{\lambda(G)}\int_G \boldsymbol{Z}(x)\text{d}x$ is the true mean of the auxiliary vector over the small area G, which is available only if the first phase is exhaustive. Let us emphasize the fact that $\widehat{\boldsymbol{\beta}}_{s_2}$ is fitted with the full data set and not only with $\{Y(x), \boldsymbol{Z}(x)|x \in G\}$.

The synthetic estimator $\hat{Y}_{G,\text{synth}}$ has a design-based asymptotic bias approximately equal to $-\frac{1}{\lambda(G)}\int_G R(x)$, which is not zero unless $G = F$. Using eqs. 18 and 15, the estimated design-based variance of the synthetic estimator is

$$[19] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,\text{synth}}) = \overline{\boldsymbol{Z}}_G^t\hat{\Sigma}_{\widehat{\boldsymbol{\beta}}_{s_2}}\overline{\boldsymbol{Z}}_G$$

To compensate for the bias due to the nonvanishing mean residual over G, one considers the small-area estimator (Mandallaz 2008, p. 120):

$$[20] \qquad \hat{Y}_{G,\text{small}} = \hat{Y}_{G,\text{synth}} + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}(x)$$

where $s_{2,G} = s_2 \cap G$ and $n_{2,G} = \sum_{x\in s_2}I_G(x)$ is the number of points of s_2 falling within G. It is shown in Mandallaz (2012) that one has the following approximation:

$$[21] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,\text{small}}) \approx \hat{\mathbb{V}}(\hat{Y}_{G,\text{synth}}) + \frac{1}{n_{2,G}}\frac{1}{n_{2,G} - 1}\sum_{x\in s_{2,G}}(\hat{R}(x) - \hat{\bar{R}}_{2,G})^2$$

Note that by the zero mean residual property over F, one has $\hat{Y}_{F,\text{small}} = \hat{Y}_{F,\text{synth}}$.

If the first phase is nonexhaustive, i.e., $n_1 \neq \infty$, then one can replace the true mean $\overline{\boldsymbol{Z}}_G$ by its estimate in the large sample $\hat{\overline{\boldsymbol{Z}}}_{1,G} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}\boldsymbol{Z}(x)$, where $s_{1,G}$ is the set $s_1 \cap G$ of the $n_{1,G} = \sum_{x\in s_1}I_G(x)$ points of the large sample falling into the small area G. This gives the pseudo-synthetic estimator

$$[22] \qquad \hat{Y}_{G,\text{psynth}} = \hat{\overline{\boldsymbol{Z}}}{}_{1,G}^t\widehat{\boldsymbol{\beta}}_{s_2} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}\hat{Y}(x)$$

The variance of the pseudo-synthetic estimator is obviously larger than the variance of the synthetic estimator, as shown in Mandallaz (2012), and can be estimated by

$$[23] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,\text{psynth}}) = \hat{\overline{\boldsymbol{Z}}}{}_{1,G}^t\hat{\Sigma}_{\widehat{\boldsymbol{\beta}}_{s_2}}\hat{\overline{\boldsymbol{Z}}}_{1,G} + \widehat{\boldsymbol{\beta}}_{s_2}^t\hat{\Sigma}_{\hat{\overline{\boldsymbol{Z}}}_{1,G}}\widehat{\boldsymbol{\beta}}_{s_2}$$

where

$$[24] \qquad \hat{\Sigma}_{\hat{\overline{\boldsymbol{Z}}}_{1,G}} = \frac{1}{n_{1,G}(n_{1,G} - 1)}\sum_{x\in s_{1,G}}(\boldsymbol{Z}(x) - \hat{\overline{\boldsymbol{Z}}}_{1,G})(\boldsymbol{Z}(x) - \hat{\overline{\boldsymbol{Z}}}_{1,G})^t$$

Usually $\hat{Y}_{G,\text{psynth}}$ will have a small variance but at the cost of a potential bias.

To compensate for the potential bias of $\hat{Y}_{G,\mathrm{psynth}}$, we consider the pseudo-small-area estimator

$$[25] \qquad \hat{Y}_{G,\mathrm{psmall}} = \hat{Y}_{G,\mathrm{psynth}} + \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} \hat{R}(x)$$

which is asymptotically design-unbiased with estimated variance

$$[26] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,\mathrm{psmall}}) \approx \hat{\mathbb{V}}(\hat{Y}_{G,\mathrm{psynth}}) + \frac{1}{n_{2,G}} \frac{1}{n_{2,G}-1} \sum_{x \in s_{2,G}} (\hat{R}(x) - \bar{\hat{R}}_{2,G})^2$$

Because of the zero mean residual property over F, one also has $\hat{Y}_{F,\mathrm{psmall}} = \hat{Y}_{F,\mathrm{psynth}}$.

All of the estimators presented in this and the following sections can be written in the equivalent so-called g-weights form, given in Mandallaz (2012), which has attractive statistical properties (particularly convincing when $\boldsymbol{Z}(x)$ defines strata) and also computational advantages when the same $\boldsymbol{Z}(x)$ is used for different $Y(x)$.

In the next section, we present a simple reformulation of the problem that allows one to transform small-area estimators into synthetic estimators, which provides a simpler and more elegant mathematical framework.

4.3 Alternative estimators in extended model

Deriving the properties of $\hat{Y}_{G,\mathrm{small}}$ and $\hat{Y}_{G,\mathrm{psmall}}$ is complicated by the fact that $\int_G R(x)\mathrm{d}x \neq 0$. If we now extend the auxiliary information vector $\boldsymbol{Z}(x)$ to $\mathcal{Z}^t(x) = (Z^t(x), I_G(x)) \in \mathcal{R}^{(p+1)}$, the corresponding model reads

$$[27] \qquad Y(x) = \mathcal{Z}^t(x)\boldsymbol{\theta} + \mathcal{R}(x)$$

which leads to the normal equation for the extended parameter vector $\boldsymbol{\theta} \in \mathcal{R}^{(p+1)}$

$$[28] \qquad \left(\int_F \mathcal{Z}(x)\mathcal{Z}^t(x)\mathrm{d}x\right)\boldsymbol{\theta} =: \boldsymbol{\mathcal{A}}\boldsymbol{\theta} = \int_F Y(x)\mathcal{Z}(x)\mathrm{d}x$$

and the orthogonality relationship $\int_F \mathcal{R}(x)\mathcal{Z}(x)\mathrm{d}x = \boldsymbol{0}$. Because $I_F(x) \equiv 1$ is the intercept term (or linear combination of the components of $\boldsymbol{Z}(x)$) and $\mathcal{Z}(x)$ contains $I_G(x)$, we have the two zero mean residual properties

$$[29] \qquad \int_F \mathcal{R}(x)\mathrm{d}x = \int_G \mathcal{R}(x)\mathrm{d}x = 0$$

Let us define $\mathcal{A}_{s_2} = \frac{1}{n_2}\sum_{x \in s_2} \mathcal{Z}(x)\mathcal{Z}^t(x)$ and $\mathcal{U}_{s_2} = \frac{1}{n_2}\sum_{x \in s_2} Y(x)\mathcal{Z}(x)$. The estimated regression coefficient is then $\hat{\boldsymbol{\theta}}_{s_2} = \mathcal{A}_{s_2}^{-1}\mathcal{U}_{s_2}$.

In perfect analogy with eq. 15, the estimated covariance matrix is given by

$$[30] \qquad \hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{s_2}} = \mathcal{A}_{s_2}^{-1}\left(\frac{1}{n_2^2}\sum_{x \in s_2} \hat{\mathcal{R}}^2(x)\mathcal{Z}(x)\mathcal{Z}(x)^t\right)\mathcal{A}_{s_2}^{-1}$$

where we have set $\hat{\mathcal{R}}(x) = Y(x) - \mathcal{Z}^t(x)\hat{\boldsymbol{\theta}}_{s_2}$. If the first phase is exhaustive, we calculate the synthetic estimator in the extended model

$$[31] \qquad \hat{\hat{Y}}_{G,\mathrm{synth}} = \frac{1}{\lambda(G)}\int_G \mathcal{Z}^t(x)\hat{\boldsymbol{\theta}}_{s_2}\mathrm{d}x = \bar{\mathcal{Z}}_G^t\hat{\boldsymbol{\theta}}_{s_2}$$

It is shown in Mandallaz (2012) that

$$[32] \qquad \hat{\hat{Y}}_{G,\mathrm{synth}} = \bar{\boldsymbol{Z}}_G^t\hat{\boldsymbol{\beta}}_{s_2} + \frac{\alpha}{n_{2,G}}\sum_{x \in s_{2,G}} (Y(x) - \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2})$$

where

$$\alpha = \frac{\hat{p}_{2,G} - \hat{p}_{2,G}^2\bar{\boldsymbol{Z}}_G^t\boldsymbol{A}_{s_2}^{-1}\hat{\bar{\boldsymbol{Z}}}_{2,G}}{\hat{p}_{2,G} - \hat{p}_{2,G}^2\hat{\bar{\boldsymbol{Z}}}_{2,G}^t\boldsymbol{A}_{s_2}^{-1}\hat{\bar{\boldsymbol{Z}}}_{2,G}}$$

with $\hat{p}_{2,G} = \frac{n_{2,G}}{n_2}$. $\hat{\hat{Y}}_{G,\mathrm{synth}}$ and $\hat{Y}_{G,\mathrm{small}}$ are asymptotically equivalent because α tends to 1 in large samples.

By using eq. 19 and replacing $\boldsymbol{Z}(x)$ with $\mathcal{Z}(x)$, we obtain at once the asymptotic variance

$$[33] \qquad \hat{\mathbb{V}}(\hat{\hat{Y}}_{G,\mathrm{synth}}) = \bar{\mathcal{Z}}_G^t\hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{s_2}}\bar{\mathcal{Z}}_G$$

If the first phase is not exhaustive, we estimate the true mean of the extended auxiliary variables

$$[34] \qquad \hat{\bar{\mathcal{Z}}}_{1,G} = \frac{1}{n_{1,G}}\sum_{x \in s_{1,G}} \mathcal{Z}(x)$$

to get the pseudo-synthetic estimate in the extended model

$$[35] \qquad \hat{\hat{Y}}_{G,\mathrm{psynth}} = \hat{\bar{\mathcal{Z}}}_{1,G}^t\hat{\boldsymbol{\theta}}_{s_2}$$

By eq. 23, we get immediately the following consistent estimate of the design-based variance

$$[36] \qquad \hat{\mathbb{V}}(\hat{\hat{Y}}_{G,\mathrm{psynth}}) = \hat{\bar{\mathcal{Z}}}_{1,G}^t\hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{s_2}}\hat{\bar{\mathcal{Z}}}_{1,G} + \hat{\boldsymbol{\theta}}_{s_2}^t\hat{\Sigma}_{\hat{\bar{\mathcal{Z}}}_{1,G}}\hat{\boldsymbol{\theta}}_{s_2}$$

where

$$[37] \qquad \hat{\Sigma}_{\hat{\bar{\mathcal{Z}}}_{1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)}\sum_{x \in s_{1,G}} (\mathcal{Z}(x) - \hat{\bar{\mathcal{Z}}}_{1,G})(\mathcal{Z}(x) - \hat{\bar{\mathcal{Z}}}_{1,G})^t$$

Remarks

- For global estimation, the variance of $\hat{Y}_{F,\mathrm{psynth}} = \hat{Y}_{F,\mathrm{psmall}}$ depends on the n_2 residuals and n_1 predictions. By analogy with the model-dependent approach and as suggested by simulations, the confidence intervals should be based on the Student's t distribution with $n_2 - p$ degrees of freedom for $n_2 > 50$ and less degrees of freedom, say $n_2 - 2p$, for $n_2 \leq 50$.
- The variance of $\hat{Y}_{G,\mathrm{psmall}}$ depends crucially on the $n_{2,G} \geq 2$ residuals in the small area and is therefore potentially prone to outliers, whereas the variance of $\hat{\hat{Y}}_{G,\mathrm{psynth}}$ is based on all n_2 residuals occurring in $\hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{s_2}}$ and can therefore be expected to be more stable. Also, the design-based variance of the last $\boldsymbol{\theta}$ component associated with the indicator variable $I_G(x)$ is of order $n_{2,G}^{-1}$. By analogy with the model-dependent approach and as suggested by simulations, the confidence intervals should be based for both estimators on the Student's t distribution with $n_{2,G} - 1$ degrees of freedom.
- For a very small area H, the number of points $n_{2,H}$ may be too small or even zero. In such a case, one could imbed H in a somewhat larger small area $G \supset H$ with $n_{2,G}$ sufficiently large and consider the estimates $\hat{\hat{Y}}_{H,\mathrm{psynth}}$ in the extended model with

respect to G, which can be conjectured to have a smaller bias and a slightly larger variance than the direct estimates $\hat{\hat{Y}}_{G,\text{psynth}}$.

In the next section, we generalize the previous results to cluster sampling. The main ideas remain the same but the formulae are slightly more cumbersome due to the random cluster size.

5. Generalization to cluster sampling

Worldwide, most national inventories rely on cluster sampling to reduce the traveling costs. We follow the Monte Carlo description of cluster sampling as defined in Mandallaz (2008, section 5.5) or Mandallaz and Ye (1999). A cluster is identified by its origin x, uniformly distributed in $\tilde{F} \supset F$. The geometry of the cluster is given by M vectors e_1, …, e_M defining the random cluster $x_l = x + e_l$. $M(x) = \sum_{l=1}^{M} I_F(x_l)$ is the random number of points of the cluster falling into the forest area F. We define the local density at the cluster level by $Y_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l)Y(x_l)}{M(x)}$, likewise we set $Z_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l)\mathbf{Z}(x_l)}{M(x)}$. The set $\tilde{F}$ above can be mathematically defined as the smallest set $\{x \in \mathcal{R}^2 | M(x) \neq 0\}$. In the first phase, we have n_1 clusters identified by $x \in s_1$ and, in the second phase, n_2 clusters with $x \in s_2$, obtained by simple random sampling from s_1.

We shall use the model-assisted approach in which the regression coefficient $\boldsymbol{\beta}_c$ at the cluster level minimizes $\int_{x\in\tilde{F}} M(x)(Y_c(x) - \boldsymbol{\beta}^t\mathbf{Z}_c(x))^2 dx$ (see Mandallaz (2008, section 5.5) for details). An asymptotically design-unbiased estimate $\hat{\boldsymbol{\beta}}_{c,s_2}$ for $\boldsymbol{\beta}_c$ can be obtained by taking a sample copy of the resulting normal equations and one gets

$$[38] \qquad \begin{aligned} \hat{\boldsymbol{\beta}}_{c,s_2} &= \left(\frac{1}{n_2}\sum_{x\in s_2} M(x)\mathbf{Z}_c(x)\mathbf{Z}_c^t(x)\right)^{-1}\left(\frac{1}{n_2}\sum_{x\in s_2} M(x)Y_c(x)\mathbf{Z}_c(x)\right) \\ &:= \boldsymbol{A}_{c,s_2}^{-1}\left(\frac{1}{n_2}\sum_{x\in s_2} M(x)Y_c(x)\mathbf{Z}_c(x)\right) \end{aligned}$$

The empirical residuals at the cluster level are $\hat{R}_c(x) = Y_c(x) - \mathbf{Z}_c^t(x)\hat{\boldsymbol{\beta}}_{c,s_2}$, which satisfy the orthogonality relation $\sum_{x\in s_2} M(x)\hat{R}_c(x) \mathbf{Z}_c(x) = 0$ and, in particular, the zero mean residual property $\frac{\sum_{x\in s_2} M(x)\hat{R}_c(x)}{\sum_{x\in s_2} M(x)} = 0$.

Using mutatis mutandis the same arguments as in simple random sampling, we get the asymptotic robust design-based estimated variance–covariance matrix

$$[39] \qquad \hat{\Sigma}_{\hat{\boldsymbol{\beta}}_{c,s_2}} = \boldsymbol{A}_{c,s_2}^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2} M^2(x)\hat{R}_c^2(x)\mathbf{Z}_c(x)\mathbf{Z}_c^t(x)\right)\boldsymbol{A}_{c,s_2}^{-1}$$

In two-phase sampling, we estimate the mean of the auxiliary information over the small area G by

$$[40] \qquad \hat{\bar{\mathbf{Z}}}_{c,1,G} = \frac{\sum_{x\in s_{1,G}} M(x)\mathbf{Z}_c(x)}{\sum_{x\in s_{1,G}} M(x)}$$

with estimated covariance matrix (see Mandallaz (2008, section 4.3)).

$$[41] \qquad \hat{\Sigma}_{\hat{\bar{\mathbf{Z}}}_{c,1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)} \times \sum_{x\in s_{1,G}} \left(\frac{M(x)}{\overline{M}_{1,G}}\right)^2 (\mathbf{Z}_c(x) - \hat{\bar{\mathbf{Z}}}_{c,1,G})(\mathbf{Z}_c(x) - \hat{\bar{\mathbf{Z}}}_{c,1,G})^t$$

The pseudo-synthetic estimate is then

$$[42] \qquad \hat{Y}_{c,G,\text{psynth}} = \hat{\bar{\mathbf{Z}}}_{c,1,G}^t \hat{\boldsymbol{\beta}}_{c,s_2}$$

with variance

$$[43] \qquad \hat{\mathbb{V}}(\hat{Y}_{c,G,\text{psynth}}) = \hat{\bar{\mathbf{Z}}}_{c,1,G}^t \hat{\Sigma}_{\hat{\boldsymbol{\beta}}_{c,s_2}} \hat{\bar{\mathbf{Z}}}_{c,1,G} + \hat{\boldsymbol{\beta}}_{c,s_2}^t \hat{\Sigma}_{\hat{\bar{\mathbf{Z}}}_{c,1,G}} \hat{\boldsymbol{\beta}}_{c,s_2}$$

The above estimate is generally design-biased. Adjusting for the residuals, we get the small-area estimator

$$[44] \qquad \hat{Y}_{c,G,\text{psmall}} = \hat{Y}_{c,G,\text{psynth}} + \frac{\sum_{x\in s_{2,G}} M(x)\hat{R}_c(x)}{\sum_{x\in s_{2,G}} M(x)}$$

which is asymptotically design-unbiased, and an approximate estimate of its variance is

$$[45] \qquad \hat{\mathbb{V}}(\hat{Y}_{c,G,\text{psmall}}) = \hat{\mathbb{V}}(\hat{Y}_{c,G,\text{psynth}}) + \frac{1}{n_{2,g}(n_{2,G}-1)} \sum_{x\in s_{2,G}} \left(\frac{M(x)}{\overline{M}_{2,G}}\right)^2 (\hat{R}_c(x) - \bar{\hat{R}}_{2,G})^2$$

where $\overline{M}_{2,G} = \frac{1}{n_{2,G}} \sum_{x\in s_{2,G}} M(x)$ and $\bar{\hat{R}}_{2,G} = \frac{\sum_{x\in s_{2,G}} M(x)\hat{R}_c(x)}{\sum_{x\in s_{2,G}} M(x)}$.

Again, because of the zero mean residual over F, we have $\hat{Y}_{c,F,\text{psynth}} = \hat{Y}_{c,F,\text{psmall}}$.

When $n_1, n_{1,G} \rightarrow \infty$, we obtain $\hat{Y}_{c,G,\text{synth}}$ and $\hat{Y}_{c,G,\text{small}}$.

As in simple two-phase sampling, we can transform the above estimator into a synthetic estimator by considering the extended model $\mathcal{Z}_c^t(x) = (\mathbf{Z}_c^t(x), I_{c,G}(x)) \in \mathcal{R}^{(p+1)}$ with $I_{c,G}(x) = \frac{\sum_{l=1}^{M} I_G(x_l)}{M(x)}$. In extensive inventories, we can reasonably assume that all points of a cluster lying in the forest area F will belong to the same small area G so that in fact $I_{c,G} \equiv 1$ for all $x \in \tilde{G} = \{x | \sum_{l=1}^{M} I_G(x_l) > 0\}2$. One can use all of the previous results by simply replacing $\boldsymbol{Z}_c(x)$ with $\mathcal{Z}_c^t(x)$ and $\hat{\boldsymbol{\beta}}_{c,s_2}$ with $\hat{\boldsymbol{\theta}}_{c,s_2} \in \mathcal{R}^{(p+1)}$, say (see Mandallaz (2012) for details). We define the pseudo-synthetic estimator in the extended model according to

$$[46] \qquad \hat{\hat{Y}}_{c,\text{psynth}} = \hat{\bar{\mathcal{Z}}}_{c,1,G}^t \hat{\boldsymbol{\theta}}_{c,s_2}$$

with estimated variance

$$[47] \qquad \hat{\mathbb{V}}(\hat{\hat{Y}}_{c,G,\text{psynth}}) = \hat{\bar{\mathcal{Z}}}_{c,1,G}^t \hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{c,s_2}} \hat{\bar{\mathcal{Z}}}_{c,1,G} + \hat{\boldsymbol{\theta}}_{c,s_2}^t \hat{\Sigma}_{\hat{\bar{\mathcal{Z}}}_{c,1,G}} \hat{\boldsymbol{\theta}}_{c,s_2}$$

The synthetic estimator in the extended model corresponds formally to $n_1 = \infty$, i.e.,

$$[48] \qquad \hat{\hat{Y}}_{c,G,\text{synth}} = \bar{\mathcal{Z}}_G^t \hat{\boldsymbol{\theta}}_{c,s_2}$$

with estimated variance

$$[49] \qquad \hat{\mathbb{V}}(\hat{\hat{Y}}_{c,G,\text{synth}}) = \bar{\mathcal{Z}}_G^t \hat{\Sigma}_{\hat{\boldsymbol{\theta}}_{c,s_2}} \bar{\mathcal{Z}}_G$$

It is mathematically clear that one can generalize all of the previous results to the simultaneous estimation of $q \geq 2$ small areas by extending the model with q indicator variables (combined extended model). One can conjecture that the combined model will be less efficient, for any given small area, than the individual estimation and, on the other hand, that it will smooth out the residual pattern.

5.1 Generalization to two-stage sampling

In extensive forest inventories, costs to measure the response variable Y_i are high. For instance, a good determination of the volume may require that one records diameter at breast height (DBH), diameter at 7 m above the ground, and total height to utilize a three-way volume function. However, one could rely on a coarser, but cheaper, approximation of the volume based only on DBH. Nonetheless, it may be most sensible to assess those three parameters only on a subsample of trees. This procedure is used in the Swiss National Forest Inventory in which the subsample is selected by Poisson sampling. The reader is referred to Mandallaz (2008, 2012) or Mandallaz and Ye (1999) for a first perusal and to Mandallaz and Massey (2012) for recent developments. It suffices to say here that the local density is replaced by the so-called generalized local density for which all formulae given in the present paper remain valid: the second-stage variance is automatically taken into account.

6. Examples

6.1 Case study

We re-analyze the case study described with full details in chapter 8 of Mandallaz (2008). The inventoried area covers 218 ha. The values of the auxiliary information are classified on 16 combinations of the following stand variables:

1. developmental stage, with the four categories "pole stage" = 3, "young timber tree" = 4, "middle age timber tree" = 5, and "old timber tree" = 6;
2. degree of mixture, with the two categories "predominantly conifers" = 1 and "predominantly broadleaves" = 2; and
3. crown closure, with the two categories "dense" = 1 and "close" = 2.

These factors produced 4 × 2 × 2 = 16 possible stands, all of which were found on the study site.

The inventory utilized systematic cluster sampling. A cluster comprises five points: a central point, two points located 30 m east and west of the central point, and two more points located 40 m north and south of the central point. The first phase sets the central cluster point on a 120 m W–E × 75 m N–S rectangular grid (note that the clusters partially overlapped in the N–S direction). The second, terrestrial phase places the central point on a 1:4 subgrid of the first phase, i.e., on a 240 m W–E × 150 m N–S systematic rectangular grid. The terrestrial inventory was purely one stage using simple circular plots with 300 m² of horizontal surface area and an inventory threshold set at 12 cm DBH.

The auxiliary vector $\mathbf{Z}(x) \in \mathcal{R}^6$ corresponds to an additive ANOVA model (for the above three stand variables) with all its components in {0, 1, –1}.

We shall consider the following five small areas.

- G_1 (≈17 ha) was used for a full census. The condition that a cluster hitting the small area has all of its points in F within the small area is occasionally violated (i.e., $I_{c,G}(x) = \frac{\sum_{l=1}^{M} I_G(x_l)}{M(x)} < 1$ for some x), so that the extended model for G_1 is only approximately correct. The mean residual over the small area is not exactly zero. The true values for basal area and stem densities are known.
- G_2 (≈33 ha) is the easternmost part of the forest; small area $G_{21} \subset G_2$ (≈7 ha) is a small subset in the central part of G_2 chosen to have a small number (3) of complete clusters ($I_{c,G}(x) \equiv 1$) spread over many different stands.
- G_3 (≈46 ha) is the southernmost part of the forest.
- G_4 (≈55 ha) is the central part north of G_3.
- G_5 (≈84 ha) is the most western part north of G_3.

Table 1. Two-phase estimates for basal area.

	Sample sizes				
Domain	n_1:n_2	$n_1\overline{M}_1$:$n_2\overline{M}_2$	$\hat{Y}_{c,G,\mathrm{psynth}}$	$\hat{Y}_{c,G,\mathrm{psmall}}$	$\hat{\hat{Y}}_{c,G,\mathrm{psynth}}$
F	298:73	1203:298	31.34 [0.94]	31.34 [0.94] (0.91)	31.30 [0.92] 31.35 [0.93] 31.35 [0.94]
G_1 true = 29.60	29:8	92:19	30.28 [1.34]	23.99 [3.90] (3.68)	25.55 [3.79]
G_2	49:9	185:41	28.27 [1.40]	29.32 [2.52] (2.08)	29.31 [2.23]
$G_{21} \subset G_2$	17:3	39:15	25.55 [2.16]	29.52 [4.13] (3.53)	29.61 [2.95]
G_3	73:18	250:66	31.62 [1.47]	31.46 [2.33] (1.87)	31.46 [2.16]

Note: Standard errors given within square brackets refer to the g-weights model and those within parentheses refer to the external model. The mean residual for small area G_1 was –1.59 for the extended model instead of 0 because $I_{c,G}(x) \not\equiv 1$. The double-kriging estimates were 31.35 with an error of 0.71 for F and 29.46 with an error of 1.39 for G_1.

We have $F = G_2 \cup G_3 \cup G_4 \cup G_5$ and $I_{c,G_k} \equiv 1$ for k = 2, 3, 4, 5. Stand map of F and detailed maps of G_1 are given in Mandallaz (2008, chapter 8), and maps of G_2 to G_5 are given in Mandallaz (2012).

Tables 1 and 2 display the results for the basal area and the stem density for small areas G_1, G_2, G_{21}, and G_3.

The standard errors for $\hat{Y}_{c,G,\mathrm{psmall}}$ are given within parentheses when considering the internal model as an external one, i.e., by using the procedure defined in Mandallaz (2008, p. 121). The standard errors given within square brackets refer to the new variance estimates eqs. 45 and 47 or their equivalent g-weights versions.

The extended model for the $\hat{\hat{Y}}_{c,G_k,\mathrm{psynth}}$ contains only the indicator variable of the corresponding small area G_k. For this reason, the corresponding estimates for the entire domain F are given for G_1, G_2, and G_3 separately (in this order).

We also consider the joint estimation of F and small areas G_2, G_3, G_4, and G_5, which form a partition of F. The corresponding model $\mathcal{Z}(x)$ contains the four indicator variables $I_{G_k}(x)$ (k = 2, 3, 4, 5), the previous components $Z_l(x)$, l = 2, 3, 4, 5, 6, but no longer the intercept term $Z_1(x) \equiv 1$ (otherwise $\mathcal{A}_{s_2}$ would be singular because $Z_1(x)$ is a linear combination of the $I_{G_k}(x)$). The results for this estimator, denoted by $\hat{\hat{Y}}_{c,G,\mathrm{cpsynth}}$, are displayed in Table 3.

All of the calculations were performed with the linear algebra procedure proc iml of the statistical software package SAS.

6.1.1 *Discussion*

All point estimates were close to each other and do not differ significantly from each other. In the small area with full census, the pseudo-synthetic estimator was closer to the true values. As confirmed by simulations, this was due to the fact that the plots within this small area were in the lower tail of the distribution for basal area and stem density. The synthetic estimators always had the smallest standard errors but at the potential cost of a local bias.

For the classical small-area estimator, the standard errors based on the external model assumption were usually, but not always, smaller than their counterparts based on the g weights, but the differences were small, which is a reassuring result. The g-weights-based standard errors of the pseudo-synthetic estimators in the extended model for one single small area were usually smaller than their g-weights counterparts of the classical small-area estimator but generally still larger than under the external model assumption. The g-weights-based standard errors in the extended model with several small areas were comparable with

Table 2. Two-phase estimates for stem density.

	Sample sizes				
Domain	n_1:n_2	$n_1\overline{M}_1$:$n_2\overline{M}_2$	$\hat{Y}_{c,G,\text{psynth}}$	$\hat{Y}_{c,G,\text{psmall}}$	$\hat{\hat{Y}}_{c,G,\text{psynth}}$
F	298:73	1203:298	325.79 [12.80]	325.79 [12.80] (12.39)	325.62 [12.81] 325.88 [12.84] 325.72 [12.85]
G_1 true = 280.23	29:8	92:19	279.54 [22.65]	257.34 [45.81] (48.29)	258.20 [54.07]
G_2	49:9	185:41	400.49 [23.36]	406.47 [41.83] (43.49)	406.41 [36.22]
$G_{21} \subset G_2$	17:3	39:15	578.90 [35.48]	589.51 [85.68] (94.31)	589.74 [67.16]
G_3	73:18	250:66	279.75 [15.41]	282.46 [21.38] (16.56)	282.40 [20.14]

Note: Standard errors given within square brackets refer to the g-weights model and those within parentheses refer to the external model. The mean residual for small area G_1 was –1.00 for the extended model instead of 0 because $I_{c,G}(x)\not\equiv 1$. The double-kriging estimates were 325.84 with an error of 11.15 for F and 281.67 with an error of 29.61 for G_1.

Table 3. Two-phase combined estimates.

	Sample sizes			
Domain	n_1:n_2	$n_1\overline{M}_1$:$n_2\overline{M}_2$	Basal area $\hat{\hat{Y}}_{c,G,\text{cpsynth}}$	Stem density $\hat{\hat{Y}}_{c,G,\text{cpsynth}}$
F	298:73	1203:298	31.32 [0.93]	325.17 [12.62]
G_2	49:9	185:41	29.39 [2.23]	407.84 [39.04]
G_3	73:18	250:66	31.57 [2.09]	284.17 [18.09]
G_4	81:17	306:69	27.77 [1.99]	274.59 [23.49]
G_5	125:29	462:122	34.31 [1.24]	347.76 [16.61]

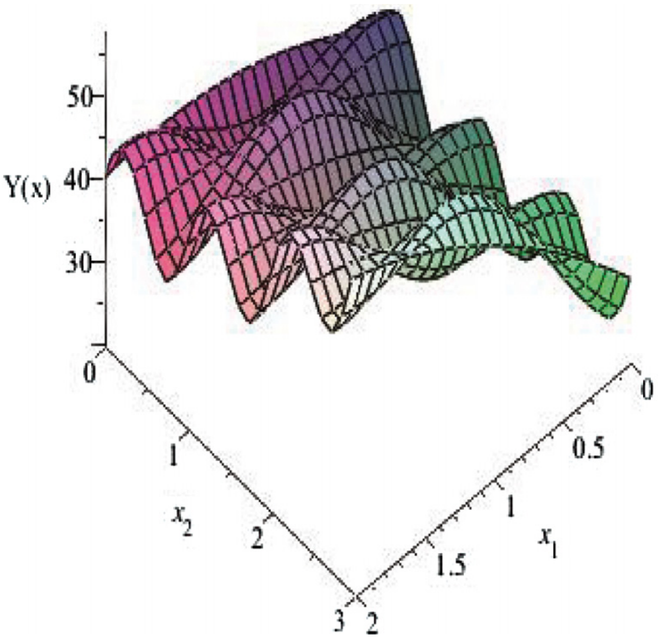
Note: Standard errors given within square brackets refer to the g-weights model. The classical estimates $\hat{Y}_{c,G,\text{psmall}}$ for G_4 were 27.78 (2.00) for basal area and 274.79 (24.12) for stem density. For G_5, the corresponding results were 34.33 (1.35) and 347.89 (17.49), respectively. As expected, on mathematical grounds, all extended models yielded zero ($<10^{-12}$) for the empirical means of the residuals over the entire domain and all small areas.

those derived specifically for one single small area. In this case study, the various design-based methods can be regarded as almost equivalent from a practical point of view (with a slight advantage to the new combined estimator in the extended model), which should be confirmed or eventually invalidated by further examples. The geostatistical double-kriging estimates were clearly better, but also far more work-intensive (several weeks with a specific software package as compared with a couple of days with SAS in terms of programming and model fitting).

6.2 A simulation example

To illustrate the theory and check empirically the validity of the various mathematical approximation used to derive the variance estimates, we present simulations performed on a purely artificial example. The local density $Y(x)$ is defined according to the following procedure. At point $x = (x_1, x_2)^t \in \mathbb{R}^2$, the auxiliary vector is defined as $\mathbf{Z}(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^t \in \mathbb{R}^6$. The true parameter is $\boldsymbol{\beta}_0 = (30, 13, -6, -4, 3, 2)^t \in \mathbb{R}^6$ and the local density over the domain $F = [0, 2] \times [0, 3]$ is given by the function

Fig. 1. Local density $Y(x)$.



[50] $$Y(x) = \mathbf{Z}^t(x)\boldsymbol{\beta}_0 + 6\cos(\pi x_1)\sin(2\pi x_2) := \hat{Y}_0(x) + R(x)$$

This function was obtained by fitting, after proper re-scaling, a full quadratic surface to the values for basal area (in m²/ha) resulting from a real inventory. The error term, however, is purely artificial in its sine–cosine structure. It can be checked that $\int_F R(x)\mathbf{Z}(x) = \mathbf{0}$, which is required for the model to be the best least squares fit (see eq. 9). The coefficient of determination is $R^2 = 0.82$.

The small area was defined as G = [0.3, 1.3] × [0.5, 2] ⊂ F with $\frac{\lambda(G)}{\lambda(F)} = \frac{1}{4}$. Figure 1 displays the local density. From a model-dependent point of view, it is clear that $Y(x)$ and $R(x)$ display a strong spatial correlation, which is, however, totally irrelevant within the design-based and model-assisted approach used in this paper. The inference is valid because x is uniformly random in F or G. Tables 4, 5, and 6 summarize the simulation results.

All of the simulations were performed with the linear algebra procedure proc iml of the statistical software package SAS, and the software Maple was used to calculate the true values given by integrals.

6.2.1. *Discussion*

Global estimation

Recall that, because of the zero mean residual over F, one has $\hat{Y}_{F,\text{psynth}} = \hat{Y}_{F,\text{psmall}}$ and $\hat{Y}_{F,\text{synth}} = \hat{Y}_{F,\text{small}}$.

The results can be summarized as follows.

- All point estimates are practically unbiased (even if the biases are statistically significant due to the huge sample size of 20 000 runs).
- The regression estimators substantially decrease the variance as compared with the one-phase sample mean.
- The empirical variances are in good agreement with their estimated counterparts, particularly for $n_2 \geq 50$ and for $\hat{Y}_{F,\text{psynth}}$.
- The distribution of all estimators is nicely bell-shaped but with heavier tails than the normal. The confidence limits based on the Student's t distribution with $n_2 - p$ degrees of freedom (as suggested by the model-dependent approach) are too small, particularly for $\hat{Y}_{F,\text{synth}}$ with $n_2 = 25$. Using the Student's t distribution with $n_2 - 2p$ degrees of freedom, one achieves the required 95% level.

Table 4. Estimates for the entire domain *F*.

	n_1:n_2		
	100:25	200:50	400:100
$\mathbb{E}^*(\hat{Y}_F)$	39.16	39.16	39.17
$\mathbb{V}^*(\hat{Y}_F)$	2.07	1.01	0.50
$\mathbb{E}^*(\hat{Y}_{F,\mathrm{psynth}})$	39.17	39.17	39.17
$\mathbb{V}^*(\hat{Y}_{F,\mathrm{psynth}})$	0.89	0.42	0.20
$\mathbb{E}^*(\widehat{\mathbb{V}}(\hat{Y}_{F,\mathrm{psynth}}))$	0.76	0.39	0.19
$\mathbb{E}^*(\widehat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{psynth}}))$	0.71	0.37	0.19
$\hat{P}_{F,\mathrm{psynth}}$ (%)	94.0	94.3	94.8
$\hat{P}_{F,\mathrm{psynth,ext}}$ (%)	93.0	93.9	94.5
$\mathbb{E}^*(\hat{Y}_{F,\mathrm{synth}})$	39.16	39.17	39.16
$\mathbb{V}^*(\hat{Y}_{F,\mathrm{synth}})$	0.50	0.21	0.10
$\mathbb{E}^*(\widehat{\mathbb{V}}(\hat{Y}_{F,\mathrm{synth}}))$	0.32	0.17	0.09
$\mathbb{E}^*(\widehat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{synth}}))$	0.26	0.16	0.08
$\hat{P}_{F,\mathrm{synth}}$ (%)	87.8	92.2	93.5
$\hat{P}_{F,\mathrm{synth,ext}}$ (%)	85.6	91.3	93.1

Note: The true mean values are $\bar{Y}_F = 39.17$ and $\bar{R}_F = 0$. The coefficient of determination is $R^2 = \frac{\mathbb{V}(\hat{Y}(x))}{\mathbb{V}(Y(x))} = 0.82$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ denote the empirical mean and variances, respectively, over the 20 000 runs. $\widehat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{psynth}})$ is the estimated variance under the external model assumption. $\hat{P}_{F,\mathrm{psynth}}$ are the empirical coverage probabilities for the 95% confidence intervals under the *t* distribution with $n_2 - p$ degrees of freedom, likewise $\hat{P}_{F,\mathrm{psynth,ext}}$ were calculated under the external model assumption.

Table 5. Estimates for the small area *G*.

	n_1:n_2		
	100:25	200:50	400:100
$\mathbb{E}^*(\hat{Y}_G)$	37.16	37.18	37.15
$\mathbb{V}^*(\hat{Y}_G)$	3.87	1.93	0.94
$\mathbb{E}^*(\widehat{\mathbb{V}}(\hat{Y}_G))$	3.89	1.92	0.93
$\mathbb{E}^*(\hat{\tilde{\hat{Y}}}_{G,\mathrm{psynth}})$	37.15	37.17	37.16
$\mathbb{V}^*(\hat{\tilde{\hat{Y}}}_{G,\mathrm{psynth}})$	2.16	1.05	0.49
$\mathbb{E}^*(\widehat{\mathbb{V}}(\hat{\tilde{\hat{Y}}}_{G,\mathrm{psynth}}))$	1.63	0.87	0.43
$\mathbb{E}^*$(CV) (%)	3.3	2.5	1.7
$\hat{P}$ (%)	94.8	93.9	93.7
$\mathbb{E}^*(\hat{Y}_{G,\mathrm{psmall}})$	37.12	37.16	37.16
$\mathbb{V}^*(\hat{Y}_{G,\mathrm{psmall}})$	2.10	1.04	0.49
$\mathbb{E}^*(\widehat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{G,\mathrm{psmall}}))$	2.03	1.01	0.49
$\mathbb{E}^*$(CV) (%)	3.7	2.7	1.8
$\hat{P}$ (%)	95.2	94.6	94.5
$\mathbb{E}^*(\alpha)$	1.08	1.03	1.01

Note: The true values are $\bar{Y}_G = 37.16$ and $\bar{R}_G = 0.66 \neq 0$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ are the empirical means and variances, respectively, based 20 000 runs, across all possible outcomes with $n_{2,G} \geq 3$. CV is the coefficient of variation of the point estimate and $\hat{P}$ is the empirical coverage probability for the 95% confidence intervals under the *t* distribution with $n_{2,G} - 1$ degrees of freedom. The coefficient α is defined by eq. 32 and tends to 1 as $n_2 \to \infty$.

- The external model assumption leads to a slight underestimation of the variance.
- On the whole the simulation confirms the asymptotic calculations of the variances.

Local estimation

One must keep in mind that the sample sizes $n_{1,G}$ and $n_{2,G}$ in the small area *G* are random variables. The empirical variances denoted by $\mathbb{V}^*$ in Tables 5 and 6 estimate, therefore, the unconditional variances, whereas the variance estimate $\widehat{\mathbb{V}}(\hat{Y}_G)$ for the sample mean $\hat{Y}_G$ and $\widehat{\mathbb{V}}_{\mathrm{ext}}(\cdot)$, under the external model assumptions, refer to the conditional variances, i.e., given $n_{1,G}$, $n_{2,G}$. The formulae based on the covariance matrices estimate directly the unconditional variances. Asymptotically, the conditional variances are consistent estimates of the unconditional ones. For each estimation technique, we give the empirical mean $\mathbb{E}^*$(CV) of the estimated coefficient of variation CV (or relative error) defined, in each run, as the square root of the estimated variance divided by the point estimate.

Table 6. Estimates for the small area *G*.

	n_1:n_2		
	100:25	200:50	400:100
$\mathbb{E}^*(\hat{\tilde{\hat{Y}}}_{G,\mathrm{synth}})$	37.15	37.17	37.16
$\mathbb{V}^*(\hat{\tilde{\hat{Y}}}_{G,\mathrm{synth}})$	1.63	0.77	0.35
$\mathbb{E}^*(\widehat{\mathbb{V}}(\hat{\tilde{\hat{Y}}}_{G,\mathrm{synth}}))$	1.23	0.65	0.34
$\mathbb{E}^*$(CV) (%)	2.8	2.2	1.6
$\hat{P}$ (%)	92.4	93.7	94.2
$\mathbb{E}^*(\hat{Y}_{G,\mathrm{small}})$	37.12	37.15	37.16
$\mathbb{V}^*(\hat{Y}_{G,\mathrm{small}})$	1.55	0.76	0.35
$\mathbb{E}^*(\widehat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{G,\mathrm{small}}))$	1.45	0.73	0.35
$\mathbb{E}^*$(CV) (%)	3.0	2.2	1.6
$\hat{P}$ (%)	92.8	94.2	94.3

Note: The true values are $\bar{Y}_G = 37.16$ and $\bar{R}_G = 0.66 \neq 0$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ are the empirical means and variances, respectively, based on 20 000 runs, across all possible outcomes with $n_{2,G} \geq 3$. CV is the coefficient of variation of the point estimate and $\hat{P}$ is the empirical coverage probability for the 95% confidence interval, under the *t* distribution with $n_{2,G} - 1$ degrees of freedom.

The results can be summarized as follows.

- All point estimates are practically unbiased (even if the biases are statistically significant due to the huge sample size of 20 000 runs).
- The regression estimators substantially decrease the variance as compared with the one-phase sample mean.
- As expected on theoretical grounds, coefficient α tends to 1 with increasing sample sizes.
- The agreement between empirical and estimated variances is, surprisingly, slightly better under the external model assumption, whereas the situation is reversed in the global estimation.
- The coefficient of variation are slightly larger under the external model assumptions and with small sample sizes.
- The unconditional distribution of all estimators is nicely bell-shaped but usually with heavier tails than the normal. Even when the distribution did not differ significantly from the normal, it was found that calculating the 95% confidence intervals accordingly was erroneous and one must rely instead on the Student's *t* distribution with at most $n_{2,G} - 1$ degrees of freedom.
- The empirical coverage probabilities of the 95% confidence intervals, calculated with the Student's *t* distribution on $n_{2,G} - 1$ degrees of freedom, were just below 95% for $\hat{\tilde{\hat{Y}}}_{G,\mathrm{psynth}}$ and $\hat{Y}_{G,\mathrm{psmall}}$ and around 93% for $\hat{\tilde{\hat{Y}}}_{G,\mathrm{synth}}$ and $\hat{Y}_{G,\mathrm{small}}$, with a slight advantage to the external model version.

Simulations were also performed with a smaller error term, $R(x) = 4\ (\cos \pi x_1)\sin(2\pi x_2)$, which leads to a higher coefficient of determination $R^2 = 0.91$. From a qualitative point of view, the results were very similar, of course with a higher variance reduction for the two-phase estimators.

On the whole, these simulations show that the various estimators have very similar performances and that the mathematical approximations are indeed quite good. For global estimation, the new estimators in the extended model perform better than the classical estimators based on the external model assumption, whereas for small-area estimation, the situation is not as clear-cut: very slight advantage to the new estimators in terms of relative errors and to the classical estimators in terms of coverage probability and agreement between empirical and estimated variances (the later finding being somewhat surprising). Of course, the simulation model used is extremely simple, with rather high R^2, and it would be premature to draw general conclusions.

7. Conclusions

In the case study, as well as in the simulations, the various design-based methods can be regarded as almost equivalent from a practical point of view, a reassuring result that should be confirmed or eventually invalidated by further studies.

The simulation example supports the adequacy of the asymptotic expansions and shows that the new estimators in the extended model can be slightly superior to the classical ones, particularly in small samples. The calculation of the 95% confidence interval should be based on the Student's t distribution with $n_2 - 2p$ degrees of freedom for global estimation and with $n_{2,G} - 1$ degrees of freedom for small-area estimation.

From a mathematical point of view, the technique of constructing the point and variance estimators in the model extended by the indicator variables of the small areas is the most elegant approach: it bypasses the residuals terms, allows for a straightforward calculation of the asymptotic variances that takes into account asymptotically the errors of the regression coefficients, and can be used for the simultaneous estimation of many small areas. Furthermore, these estimators are of the g-weights type, with known attractive properties. They are easy to implement in statistical software packages with standard linear algebra facilities, which is a definite advantage in the forest inventory context. Furthermore, as will be shown in a follow-up paper, they can be generalized to the case in which part of the auxiliary information is exhaustive, an important issue in view of the growing availability of high-tech remote sensing techniques such as LIDAR and related techniques.

Acknowledgements

I express my thanks to H. Heinimann (Chair of Land Use Engineering, ETH Zurich) for his support, to Alexander Massey (WSL Institute, Birmensdorf, Switzerland) for proofreading the manuscript, and to the reviewers for their comments and suggestions that contributed greatly to improving the original manuscript.

References

Bafetta, F., Corona, P., and Fattorini, L. 2011. Design-based diagnostics for k-NN estimators of forest resources. Can. J. For. Res. **41**(1): 59–72. doi:10.1139/X10-157.

Breidenbach, J., and Astrup, R. 2012. Small area estimation of forest attribute in the Norwegian National Forest Inventory. Eur. J. Forest Res. **131**: 1255–1267. doi:10.1007/s10342-012-0596-7.

Breidenbach, J., and Nothdurft, A. 2010. Comparison of nearest neighbours approaches for small area estimation of tree species-specific forest inventory attributes in central Europe using airborne laser scanner data. Eur. J. Forest Res. **129**: 833–846. doi:10.1007/s10342-010-0384-1.

Finley, A.O., Naerjee, S., and McRoberts, R. 2008. A Bayesian approch to multisource forest area estimation. Environ. Ecol. Stat. **15**: 241–258. doi:10.1007/s10651-007-0049-5.

Goerndt, M.E., Monleon, V.J., and Temesgen, H. 2011. A comparison of small-area estimation techniques to estimate selected stand attributes using LIDAR-derived auxiliary variables. Can. J. For. Res. **41**(6): 1189–1201. doi:10.1139/x11-033.

Gregoire, T., and Dyer, M. 1989. Model fitting under patterned heterogeineity of variance. Forest Sci. **35**: 105–125.

Huber, P.J. 1967. The behaviour of maximum likelihood estimates under nonstandard conditions. *In* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume 1. University of California Press, Berkeley, California. pp. 221–233.

Koehl, M., Magnussen, S., and Marchetti, M. 2006. Sampling methods, remote sensing and GIS multisource forest inventory. Springer, Berlin, Heidelberg.

Lappi, J. 2001. Forest inventory of small areas combining the calibration estimator and a spatial model. Can. J. For. Res. **31**(9): 1551–1560. doi:10.1139/x01-078.

Mandallaz, D. 1993. Geostatistical methods for double sampling schemes: applications to combined forest inventory. Habilitation thesis, Department of Environmental Systems Science, ETH Zurich, Technical report. Available from http://e-collection.library.ethz.ch.

Mandallaz, D. 2000. Estimation of the spatial covariance in universal kriging: application to forest inventory. Environ. Ecol. Stat. **7**: 263–284. doi:10.1023/A:1009619117138.

Mandallaz, D. 2008. Sampling techniques for forest inventories. Chapman and Hall, Boca Raton, Florida.

Mandallaz, D. 2012. Design-based properties of small-area estimators in forest inventory with two phase sampling. Department of Environmental Systems Science, ETH Zurich, Technical report. Available from http://e-collection.library.ethz.ch.

Mandallaz, D., and Massey, A. 2012. Comparison of estimators in one-phase two-stage Poisson sampling in forest inventories. Can. J. For. Res. **42**(12): 1865–1871. doi:10.1139/x2012-110.

Mandallaz, D., and Ye, R. 1999. Forest inventory with optimal two-phase, two-stage sampling schemes based on the anticipated variance. Can. J. For. Res. **29**(11): 1691–1708. doi:10.1139/x99-124.

McRoberts, R.E. 2012. Estimating forest attributes parameters for small areas using nearest neighbors techniques. Forest Ecol. Manag. **272**: 3–12. doi:10.1016/j.foreco.2011.06.039.

Rao, J. 2003. Small area estimation. John Wiley and Sons, Hoboken, New Jersey.

Särndal, C., Swenson, B., and Wretman, J. 2003. Model assisted survey sampling. Springer Series in Statistics, New York.