NRC Research Press
ARTICLE

New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation

Daniel Mandallaz, Jochen Breschan, and Andreas Hill

Abstract: We consider two-phase sampling schemes where one component of the auxiliary information is known in every point ("wall-to-wall") and a second component is available only in the large sample of the first phase, whereas the second phase yields a subsample with the terrestrial inventory. This setup is of growing interest in forest inventory thanks to the recent advances in remote sensing, in particular, the availability of LiDAR data. We propose a new two-phase regression estimator for global and local estimation and derive its asymptotic design-based variance. The new estimator performs better than the classical regression estimator. Furthermore, it can be generalized to cluster sampling and two-stage tree sampling within plots. Simulations and a case study with LiDAR data illustrate the theory.

Résumé : Cet article propose un nouvel estimateur pour les inventaires forestiers utilisant des plans de sondage à deux phases pour lesquels l'information auxiliaire consiste en une première composante exhaustive connue en chaque point et une seconde composante connue seulement aux points du grand échantillon de la première phase. La deuxième phase consiste en un sous-échantillon des points de la première phase dans lesquels l'inventaire terrestre est effectué. Ce contexte est appelé à jour un rôle croissant grâce aux développements récents dans l'aquisition de données par télédétection. Nous proposons un nouvel estimateur par regression, aussi bien pour l'estimation globale que locale, et nous donnons sa variance asymptotique sous le plan de sondage. Le nouvel estimateur peut être adapté aux inventaire par satellites et avec tirage double des arbres au niveau de la placette terrestre. Un exemple utilisant des données LiDAR et des simulations illustrent la théorie. Le nouvel estimateur a de meilleures performance que l'estimateur de régression classique.

1. Introduction

The motivation for this work is due to the increasing need for using 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. The small-area estimation problem is of the utmost importance in this context. The fundamental idea (known as "borrowing strength") is rather simple: models are fitted globally and applied locally, albeit with minor modifications. This paper introduces the so-called generalized regression estimator, which is a further development of the two-phase regression estimator presented in Mandallaz (2013*a*): it adapts and improves the estimator when part of the auxiliary information is exhaustive or "wall-to-wall". The methodology and terminology rests upon the design-based Monte Carlo approach to sampling theory for forest inventory. The reader unfamiliar with this topic should first consult Mandallaz (2013*a*) for a short bibliographical review and a thorough but terse analysis of the regression estimator in a new setting, particularly useful in the small-area estimation context. Mandallaz (2008), chapters 4 and 5, is recommended for a first perusal, as well as Mandallaz (2012) for a detailed analysis of the standard two-phase regression estimators.

In this paper, we only state the most important results needed for practical use. The statistician will find the proofs together with further developments in Mandallaz (2013*b*). It must be emphasized that the proposed procedures can be easily implemented in standard software packages like SAS or R.

The methods presented here are far easier to implement but yield larger standard errors than the model-dependent geostatistical double or universal block-kriging techniques (particularly for small-area estimation and with long-range spatial correlation; see Mandallaz (2008), chapter 7) that are used in numerous applications with georeferenced data (e.g., in mining, petrology, and soil sciences).

2. Methodology

The *first phase* draws a large sample s_1 of n_1 points $x_i \in s_1$ ($i = 1, 2, \ldots n_1$) that are independently and uniformly distributed within the forest area F. At each of those points auxiliary information is collected, very often coding information of qualitative nature (e.g., following the interpretation of aerial photographs) or quantitative (e.g., timber volume estimates based on LiDAR measurements). We shall assume that the auxiliary information at point x is described by the row vector $\mathbf{Z}^t(x) = (\mathbf{Z}^{(1)t}(x), \mathbf{Z}^{(2)t}(x)) \in \Re^{p+q}$ (the upper index t denotes the transposition operator). The first component $\mathbf{Z}^{(1)}(x) \in \Re^p$ of this vector is known at all points $x \in F$; it is the *exhaustive* part of the auxiliary information, e.g., it could be given by thematic maps. The second component $\mathbf{Z}^{(2)}(x) \in \Re^q$ is known only at points $x \in s_1$.

Received 8 May 2013. Accepted 7 August 2013.

D. Mandallaz, J. Breschan, and A. Hill. Chair of Land Use Engineering, Department of Environmental Systems Science, ETH Zurich, CH 8092 Zurich, Switzerland.

Corresponding author: Daniel Mandallaz (e-mail: daniel.mandallaz@env.ethz.ch).

Can. J. For. Res. **43**: 1023–1031 (2013) dx.doi.org/10.1139/cjfr-2013-0181

Published at www.nrcresearchpress.com/cjfr on 22 October 2013.

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 $\mathcal{P}$ of N trees with response variable Y_i, $i = 1, 2,\ldots$, e.g., the timber volume. The objective is to estimate the 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 ha). For each point $x \in s_2$, trees are drawn from the population $\mathcal{P}$ 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

$$I_i(x) = \begin{cases} 1 \text{ if } i \notin s_2(x) \\ 0 \text{ if } i \in s_2(x) \end{cases} \tag{1}$$

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

$$Y(x) = \frac{1}{\lambda(F)}\sum_{i=1}^{N}\frac{I_i(x)Y_i}{\pi_i} = \frac{1}{\lambda(F)}\sum_{i \in s_2(x)}\frac{Y_i}{\pi_i} \tag{2}$$

The term $\frac{1}{\lambda(F)\pi_i}$ is the tree extrapolation factor f_i with dimension ha−1. Because of possible boundary adjustments $\lambda(F)\pi_i = \lambda(F \cap K_i)$, where K_i is the inclusion circle of the ith tree. In the infinite population or Monte Carlo approach, one samples the function $Y(x)$ and we have $\mathbb{E}_x(Y(x)) = \frac{1}{\lambda(F)}\int_F Y(x)\mathrm{d}x = \frac{1}{\lambda(F)}\sum_{i=1}^{N} Y_i = \bar{Y}$, where $\mathbb{E}_x$ denotes the expectation with respect to a random point x uniformly distributed in F.

3. The models

We shall work with the following linear models (see Mandallaz (2013*b*) for more details):

1. The large model *M*

$$Y(x) = \mathbf{Z}^t(x)\boldsymbol{\beta} + R(x) = \mathbf{Z}^{(1)t}(x)\boldsymbol{\beta}^{(1)} + \mathbf{Z}^{(2)t}(x)\boldsymbol{\beta}^{(2)} + R(x) \tag{3}$$

with $\boldsymbol{\beta}^t = (\boldsymbol{\beta}^{(1)t}, \boldsymbol{\beta}^{(2)t})$. The intercept term is contained in $\mathbf{Z}^{(1)}(x)$ or a linear combination of the components of $\mathbf{Z}^{(1)}(x)$ is a constant equal to 1.

The theoretical regression parameter $\boldsymbol{\beta}$ minimizes $\int_F (Y(x) - \mathbf{Z}^t(x)\boldsymbol{\beta})^2\mathrm{d}x$, and it satisfies the normal equation $(\int_F \mathbf{Z}(x)\mathbf{Z}^t(x)\mathrm{d}x)\boldsymbol{\beta} = \int_F Y(x)\mathbf{Z}(x)\mathrm{d}x$ and the orthogonality relationship $\int_F R(x)Z(x)\mathrm{d}x = \mathbf{0}$, in particular, the zero mean residual property $\frac{1}{\lambda(F)}\int_F R(x)\mathrm{d}x = 0$.

2. The reduced model M_1

$$Y(x) = \mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha} + R_1(x) \tag{4}$$

The theoretical regression parameter $\boldsymbol{\alpha}$ minimizes $\int_F (Y(x) - \mathbf{Z}^{(1)t}(x)\boldsymbol{\alpha})^2\mathrm{d}x$. It satisfies the normal equation $(\int_F \mathbf{Z}^{(1)}(x)Z^{(1)t}(x)\mathrm{d}x)\boldsymbol{\alpha} = \int_F Y(x)\mathbf{Z}^{(1)}(x)\mathrm{d}x$ and the orthogonality relationship $\int_F R_1(x)Z^{(1)}(x)\mathrm{d}x = 0$, in particular, the zero mean residual property $\frac{1}{\lambda(F)}\int_F R_1(x)\mathrm{d}x = 0$.

Let us emphasize the fact that in this paper we consider only the properties of estimators in the *design-based* paradigm and that we do not assume these models to be correct in the sense of *model-dependent* inference (see Mandallaz (2013*a*)).

4. The generalized regression estimator

We consider the following design-based least squares estimators of the regression coefficients of the reduced model, which are solutions of sample copies of the normal equations.

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_k &= \left(\frac{1}{n_k}\sum_{x \in s_k}\mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)}(x)^t\right)^{-1}\frac{1}{n_k}\sum_{x \in s_k} Y(x)\mathbf{Z}^{(1)}(x) \\ &:= \left(\boldsymbol{A}_k^{(1)}\right)^{-1}\frac{1}{n_k}\sum_{x \in s_k} Y(x)\mathbf{Z}^{(1)}(x) \quad k = 1, 2 \end{aligned} \tag{5}$$

Likewise for the large large model we set

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k &= \left(\frac{1}{n_k}\sum_{x \in s_k}\mathbf{Z}(x)\mathbf{Z}(x)^t\right)^{-1}\frac{1}{n_k}\sum_{x \in s_k} Y(x)\mathbf{Z}(x) \\ &:= \boldsymbol{A}_k^{-1}\frac{1}{n_k}\sum_{x \in s_k} Y(x)\mathbf{Z}(x) \end{aligned} \tag{6}$$

Note that only $\hat{\boldsymbol{\alpha}}_2$ and $\hat{\boldsymbol{\beta}}_2$ are observable because $Y(x)$ is only available at $x \in s_2$, and that in general the vector consisting of the first p components of $\hat{\boldsymbol{\beta}}_2$ is not equal to $\hat{\boldsymbol{\alpha}}_2$ (they are if the corresponding explanatory variables are orthogonal in the classical least squares sense).

The large model yields the predictions $\hat{Y}(x) = \mathbf{Z}^t(x)\hat{\boldsymbol{\beta}}_2$ and the reduced model yields the predictions $\hat{Y}_1(x) = \mathbf{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}_2$.

The *generalized regression estimate* is defined as

$$\hat{Y}_{\text{greg}} = \frac{1}{\lambda(F)}\int_F \hat{Y}_1(x)\mathrm{d}x + \frac{1}{n_1}\sum_{x \in s_1}(\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_2}\sum_{x \in s_2}(Y(x) - \hat{Y}(x)) \tag{7}$$

This estimator is the Monte Carlo version of Särndal's regression estimator for two-phase sampling in a finite population (see Särndal et al. (2003), eq. 9.7.20). It is clear by the law of large numbers that $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\alpha}}_2$ are asymptotically design-unbiased estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. This implies at once that $\mathbb{E}_{1,2}\hat{Y}_{\text{greg}} = \mathbb{E}_1\mathbb{E}_{2|1}\hat{Y}_{\text{greg}} \approx \bar{Y}$ (with $\mathbb{E}_{2|1}$ denoting the conditional expectation of the second phase given the first phase, i.e., simple random sampling without replacement in the finite population of the points in s_1), and $\mathbb{E}_1$ denotes the expectation with respect to uniformly distribution points of the first phase (i.e., to $\mathbb{E}_x$). The generalized regression estimate is therefore asymptotically design-unbiased.

To understand the usefulness of $\hat{Y}_{\text{greg}}$ we shall assume for the time being that the model is *external*, i.e., not fitted by the inventory data, and that the regression coefficients have given fixed values. This is equivalent to neglecting the design-based variance of the regression coefficients. Using the well-known variance decomposition (e.g., see Mandallaz (2008), Appendix B, or Särndal et al. (2003), p. 136)

$$\mathbb{V}(\hat{Y}_{\text{greg}}) = \mathbb{V}_1\mathbb{E}_{2|1}(\hat{Y}_{\text{greg}}) + \mathbb{E}_1\mathbb{V}_{2|1}(\hat{Y}_{\text{greg}}) \tag{8}$$

we get the design-based variance as

$$\mathbb{V}(\hat{Y}_{\text{greg}}) = \frac{1}{n_1}\mathbb{V}_x(R_1(x)) + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\mathbb{V}_x(R(x)) \tag{9}$$

The variances $\mathbb{V}_x(\cdot)$ are calculated under the uniform distribution in F of the random point x. This should be compared with the standard result for the variance of the regression estimator $\hat{Y}_{\text{reg}}$ under the large model

$$\hat{Y}_{\mathrm{reg}} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}(x) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x) - \hat{Y}(x)) \tag{10}$$

whose theoretical variance is given by

$$\mathbb{V}(\hat{Y}_{\mathrm{reg}}) = \frac{1}{n_1}\mathbb{V}_x(Y(x)) + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\mathbb{V}_x(R(x)) \tag{11}$$

Thus, by also using the exhaustive information, the variance of the observations in eq. (11) is replaced by the variance of the residuals under the reduced model, a very nice and intuitive result indeed.

We now give an alternative definition of $\hat{Y}_{\mathrm{greg}}$ based on the estimated regression coefficients, which is essential to derive the design-based variance with *internal models*, i.e., fitted with the inventory data at hand, and for future generalization to the small-area estimation problem. To this end we need the following mean values:

$$\begin{aligned}
\bar{\mathbf{Z}}^{(1)} &= \frac{1}{\lambda(F)}\int_F \mathbf{Z}^{(1)}(x)\mathrm{d}x\\
\hat{\bar{\mathbf{Z}}}^{(1)}_1 &= \frac{1}{n_1}\sum_{x\in s_1}\mathbf{Z}^{(1)}(x)\\
\hat{\bar{\mathbf{Z}}}_k &= \frac{1}{n_k}\sum_{x\in s_k}\mathbf{Z}(x) \qquad k = 1, 2
\end{aligned} \tag{12}$$

The regression estimate can be rewritten as

$$\begin{aligned}
\hat{Y}_{\mathrm{greg}} &= \left(\bar{\mathbf{Z}}^{(1)} - \hat{\bar{\mathbf{Z}}}^{(1)}_1\right)^t\hat{\boldsymbol{\alpha}}_2 + \left(\hat{\bar{\mathbf{Z}}}_1 - \hat{\bar{\mathbf{Z}}}_2\right)^t\hat{\boldsymbol{\beta}}_2 + \frac{1}{n_2}\sum_{x\in s_2}Y(x)\\
&= \left(\bar{\mathbf{Z}}^{(1)} - \hat{\bar{\mathbf{Z}}}^{(1)}_1\right)^t\hat{\boldsymbol{\alpha}}_2 + \hat{\bar{\mathbf{Z}}}^t_1\hat{\boldsymbol{\beta}}_2
\end{aligned} \tag{13}$$

The last equation follows from the fact that the sum of the residuals is zero by construction. Note that it suffices to know the integral of $\mathbf{Z}^{(1)}(x)$ and $\mathbf{Z}^{(1)}(x)$ for all $x \in s_1$ and not necessarily the values at all points $x \in F$.

5. Variance estimates

To obtain a simple estimate of the variance in eq. (9), we can treat the internal model as an external one and replace the theoretical residuals by their empirical versions $\hat{R}_1(x) = Y(x) - \hat{Y}_1(x) = Y(x) - \mathbf{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}_2$ and $\hat{R}(x) = Y(x) - \hat{Y}(x) = Y(x) - \mathbf{Z}^t(x)\hat{\boldsymbol{\beta}}_2$, which have zero means, to obtain

$$\hat{\mathbb{V}}(\hat{Y}_{\mathrm{greg}}) = \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}\hat{R}^2_1(x) + \frac{1}{n_2}\left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\sum_{x\in s_2}\hat{R}^2(x) \tag{14}$$

To derive variance estimates with better statistical properties, we shall use the g-weights technique (for details see Mandallaz (2008), section 6.2 for the Monte Carlo approach, and Särndal et al. (2003), sections 6.5 and 6.6 for finite populations). The g-weights are defined by

$$\begin{aligned}
g_2(x) &= 1 + \left(\hat{\bar{\mathbf{Z}}}_1 - \hat{\bar{\mathbf{Z}}}_2\right)^t\mathbf{A}^{-1}_2\mathbf{Z}(x) = \hat{\bar{\mathbf{Z}}}^t_1\mathbf{A}^{-1}_2\mathbf{Z}(x)\\
g^{(1)}_1(x) &= 1 + \left(\bar{\mathbf{Z}}^{(1)} - \hat{\bar{\mathbf{Z}}}^{(1)}_1\right)^t\left(\mathbf{A}^{(1)}_1\right)^{-1}\mathbf{Z}^{(1)}(x) = \bar{\mathbf{Z}}^{(1)t}\left(\mathbf{A}^{(1)}_1\right)^{-1}\mathbf{Z}^{(1)}(x)
\end{aligned} \tag{15}$$

That the two versions of the g-weights are equivalent is a consequence of the zero residual sum for any local density. Straightforward algebra leads to the following important calibration properties:

$$\begin{aligned}
\frac{1}{n_2}\sum_{x\in s_2}g_2(x)\mathbf{Z}(x) &= \hat{\bar{\mathbf{Z}}}_1\\
\frac{1}{n_1}\sum_{x\in s_1}g^{(1)}_1(x)\mathbf{Z}^{(1)}(x) &= \bar{\mathbf{Z}}^{(1)}
\end{aligned} \tag{16}$$

Intuitively, because the g-weights provide perfect estimates for the means of the auxiliary variables, they must perform well for the response variables if the models are adequate. One can express the design-based variance of $\hat{Y}_{\mathrm{greg}}$ with the g-weights. To this end, we need the following asymptotically consistent estimates of the design-based variance of the regression coefficients:

$$\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_2} = \mathbf{A}^{-1}_2\left(\frac{1}{n^2_2}\sum_{x\in s_2}\hat{R}^2(x)\mathbf{Z}(x)\mathbf{Z}^t(x)\right)\mathbf{A}^{-1}_2 \tag{17}$$

and

$$\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}_2} = \left(\mathbf{A}^{(1)}_1\right)^{-1}\left(\frac{1}{n^2_2}\sum_{x\in s_2}\hat{R}^2_1(x)\mathbf{Z}^{(1)}(x)\mathbf{Z}^{(1)t}(x)\right)\left(\mathbf{A}^{(1)}_1\right)^{-1} \tag{18}$$

Proofs can be found in Mandallaz (2008) (section 6.4) and Mandallaz (2012, 2013*b*). These estimates have also been discussed in a totally different context, i.e., in the model-dependent least squares theory under nonstandard conditions, by Huber (1967) and Gregoire and Dyer (1989); they are sometimes called *robust covariance matrices*.

It can be shown that

$$\hat{\mathbb{V}}(\hat{Y}_{\mathrm{greg}}) = \frac{n_2}{n_1}\bar{\mathbf{Z}}^{(1)t}\hat{\Sigma}_{\hat{\boldsymbol{\alpha}}_2}\bar{\mathbf{Z}}^{(1)} + \left(1 - \frac{n_2}{n_1}\right)\hat{\bar{\mathbf{Z}}}^t_1\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_2}\hat{\bar{\mathbf{Z}}}_1 \tag{19}$$

is a consistent estimate of the design-based variance of $\hat{Y}_{\mathrm{greg}}$ (see Mandallaz (2013*b*) for the proof). It can be confirmed that this is the same as

$$\begin{aligned}
\hat{\mathbb{V}}(\hat{Y}_{\mathrm{greg}}) = \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}\left(g^{(1)}_1(x)\right)^2\hat{R}^2_1(x)\\
+ \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n^2_2}\sum_{x\in s_2}g^2_2(x)\hat{R}^2(x)
\end{aligned} \tag{20}$$

which is the perfect Monte Carlo analogy of eq. 9.7.22 in Särndal et al. (2003), result 9.7.1. However, once again, the Monte Carlo approach is much better suited to the needs of forest inventory than the finite population framework. It suffices to compare the complexity of 9.7.1 with the simplicity of eq. (19) to be convinced. Furthermore, the interesting links of the design-based covariance matrices of the regression coefficients in eq. (19) to their robust model-dependent counterparts is absent in 9.7.1. Finally, the generalization of the above results to small-area estimation and to cluster sampling is rather straightforward in the Monte Carlo approach, as we shall see, whereas it would be extremely cumbersome in the finite population approach.

Note that in eq. (18) one could use $\mathbf{A}^{(1)}_2$ instead of $\mathbf{A}^{(1)}_1$ to obtain an asymptotically equivalent variance estimate, but the second calibration properties in eq. (16) would be violated.

For a better intuitive understanding, we note that

$$\hat{Y}_{\mathrm{greg}} \approx \bar{\mathbf{Z}}^{(1)t}\hat{\boldsymbol{\beta}}^{(1)}_2 + \hat{\bar{\mathbf{Z}}}^{(2)t}_1\hat{\boldsymbol{\beta}}^{(2)}_2 \tag{21}$$

with equality if the components $\mathbf{Z}^{(1)}(x)$ and $\mathbf{Z}^{(2)}(x)$ are orthogonal, otherwise, we have only asymptotic equivalence (see Mandallaz (2013*b*) for details). This is intuitively very appealing: the exhaustive component $\mathbf{Z}^{(1)}(x)$ occurs with its known true mean and the nonexhaustive component $\mathbf{Z}^{(2)}(x)$ occurs with its estimated mean from the large sample, as compared with the classical two-phase estimator $\hat{Y}_{\mathrm{reg}} = \hat{\bar{\mathbf{Z}}}_1^{(1)t}\hat{\boldsymbol{\beta}}_2^{(1)} + \hat{\bar{\mathbf{Z}}}_1^{(2)t}\hat{\boldsymbol{\beta}}_2^{(2)}$.

The reader can consult Mandallaz (2013*b*) to get a better insight into the qualitative aspects of Y_{greg} in the important special case of poststratification.

6. Generalized small-area estimators

We consider a small area $G \subset F$ and we want to estimate

$$\bar{Y}_G = \frac{1}{\lambda(G)}\sum_{i=1}^{N} I_G(i)Y_i = \frac{1}{\lambda(G)}\int_G Y(x)\mathrm{d}x$$

where $I_G(i) = 1$ if the ith tree is in G; otherwise, $I_G(i) = 0$. Strictly speaking, the last equality holds only if boundary adjustments are performed in G, whereas they are in most instances only performed with respect to F. We shall need the following notation: $s_{1,G} = s_1 \cap G$, $s_{2,G} = s_2 \cap G$, $n_{k,G} = \sum_{s\in s_2} I_G(x)\quad k = 1{,}2$. The simplest solution is to restrict the samples to G, i.e., to consider *the generalized small-area estimator*

$$\text{(22)}\qquad \hat{Y}_{G,\mathrm{greg}} = \frac{1}{\lambda(G)}\int_G \hat{Y}_1(x)\mathrm{d}x + \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}(\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}(Y(x) - \hat{Y}(x))$$

and treat the internal model as an external one to obtain the variance estimate

$$\text{(23)}\qquad \hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{G,\mathrm{greg}}) = \frac{1}{n_{1,G}(n_{2,G}-1)}\sum_{x\in s_{2,G}}(\hat{R}_1(x) - \hat{\bar{R}}_{1,G})^2 + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right)\frac{1}{n_{2,G}(n_{2,G}-1)}\sum_{x\in s_2}(\hat{R}(x) - \hat{\bar{R}}_G)^2$$

where $\hat{\bar{R}}_{1,G} = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}_1(x)$ and $\hat{\bar{R}}_G = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}(x)$. This variance estimate neglects the uncertainty of the regression coefficients, but there is empirical evidence that this is acceptable in large samples (see Mandallaz (2012) for examples with $\hat{Y}_{\mathrm{reg}}$). We can rewrite $\hat{Y}_{G,\mathrm{greg}}$ as

$$\text{(24)}\qquad \hat{Y}_{G,\mathrm{greg}} = \left(\bar{\mathbf{Z}}_G^{(1)} - \hat{\bar{\mathbf{Z}}}_{1,G}^{(1)}\right)^t\hat{\boldsymbol{\alpha}}_2 + \hat{\bar{\mathbf{Z}}}_{1,G}^t\hat{\boldsymbol{\beta}}_2 + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}(x)$$

where we have set

$$\bar{\mathbf{Z}}_G^{(1)} = \frac{1}{\lambda(G)}\int_G \mathbf{Z}^{(1)}(x)\mathrm{d}x, \quad \hat{\bar{\mathbf{Z}}}_{1,G} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}\mathbf{Z}(x) = \left(\hat{\bar{\mathbf{Z}}}_{1,G}^{(1)t}, \hat{\bar{\mathbf{Z}}}_{1,G}^{(2)t}\right)^t$$

The essential difference to $\hat{Y}_{\mathrm{greg}} = \hat{Y}_{F,\mathrm{greg}}$ is that the mean residual term in eq. (24) does no longer vanish in general, which makes the calculation of the variance very difficult. To bypass this difficulty, we use the technique presented in Mandallaz (2012) by extending the model with the indicator variable $I_G(x)$ of the small area G, **which insures zero mean residual over F and G.** We can include $I_G(x)$ in $\mathbf{Z}^{(1)}(x)$ or $\mathbf{Z}^{(2)}(x)$. It seems more natural to include it in the first component so that the zero mean residual properties will hold for both the reduced and the large model. Also, it is reasonable to assume that the perimeter and, consequently, the surface area of G are known. We consider, therefore, the following extended models with auxiliary vectors: $\mathcal{Z}^t(x) = (\mathcal{Z}^{(1)t}(x), \mathcal{Z}^{(2)t}(x))$, where $\mathcal{Z}^{(1)t}(x) = (\mathbf{Z}^{(1)t}(x), I_G^t(x))$ and $\mathcal{Z}^{(2)t}(x) = \mathbf{Z}^{(2)t}(x)$. To have a uniform notation throughout, we also change the notation for the second component, i.e., we will use $\mathcal{Z}^{(2)}(x)$ instead of $\mathbf{Z}^{(2)}(x)$ in this section. Therefore, we have the following setup:

1. The large extended model M

$$\text{(25)}\qquad Y(x) = \mathcal{Z}(x)^t\boldsymbol{\theta} + \mathcal{R}(x) = \mathcal{Z}^{(1)t}(x)\boldsymbol{\theta}^{(1)} + \mathcal{Z}^{(2)t}(x)\boldsymbol{\theta}^{(2)} + \mathcal{R}(x)$$

with $\boldsymbol{\theta}^t = (\boldsymbol{\theta}^{(1)t}, \boldsymbol{\theta}^{(2)t})$. The intercept term is contained in $\mathcal{Z}^{(1)}(x)$ or it is a linear combination of its components.

The theoretical regression parameter $\boldsymbol{\theta}$ minimizes $\int_F(Y(x) - \mathcal{Z}^t(x)\boldsymbol{\theta})^2\mathrm{d}x$. It satisfies the normal equation $(\int_F \mathcal{Z}(x)\mathcal{Z}^t(x)\mathrm{d}x)\boldsymbol{\theta} = \int_F(Y(x)\mathcal{Z}(x)\mathrm{d}x)$, the orthogonality relationship $\int_F \mathcal{R}(x)\mathcal{Z}(x)\mathrm{d}x = \mathbf{0}$ and, in particular, the zero mean residual properties $\frac{1}{\lambda(F)}\int_F \mathcal{R}(x)\mathrm{d}x = \frac{1}{\lambda(G)}\int_G \mathcal{R}(x)\mathrm{d}x = 0$.

2. The reduced extended model M_1

$$\text{(26)}\qquad Y(x) = \mathcal{Z}^{(1)t}(x)\boldsymbol{\gamma} + \mathcal{R}_1(x)$$

The theoretical regression parameter $\boldsymbol{\gamma}$ minimizes $\int_F(Y(x) - \mathcal{Z}^{(1)t}(x)\boldsymbol{\gamma})^2\mathrm{d}x$. It satisfies the normal equation $(\int_F \mathcal{Z}^{(1)}(x)\mathcal{Z}^{(1)t}(x)\mathrm{d}x)\boldsymbol{\gamma} = \int_F Y(x)\mathcal{Z}^{(1)}(x)\mathrm{d}x$, the orthogonality relationship $\int_G R_1(x)\mathcal{Z}^{(1)}(x)\mathrm{d}x = 0$, and, in particular, the zero mean residual properties $\frac{1}{\lambda(F)}\int_F \mathcal{R}_1(x)\mathrm{d}x = \frac{1}{\lambda(G)}\int_G \mathcal{R}_1(x)\mathrm{d}x = 0$.

We can obviously apply mutatis mutandis on all of the previous results. The estimated regression coefficients are

$$\text{(27)}\qquad \begin{aligned}\hat{\boldsymbol{\gamma}}_2 &= \left(\frac{1}{n_2}\sum_{x\in s_2}\mathcal{Z}^{(1)}(x)\mathcal{Z}^{(1)t}(x)\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}^{(1)}(x)\\ &:= \left(\mathcal{A}_2^{(1)}\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}^{(1)}(x)\end{aligned}$$

and

$$\text{(28)}\qquad \begin{aligned}\hat{\boldsymbol{\theta}}_2 &= \left(\frac{1}{n_2}\sum_{x\in s_2}\mathcal{Z}(x)\mathcal{Z}^t(x)\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}(x)\\ &:= \mathcal{A}_2^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}(x)\end{aligned}$$

According to eqs. (17) and (18), the estimated covariance matrices are

$$\text{(29)}\qquad \hat{\Sigma}_{\hat{\boldsymbol{\theta}}_2} = \mathcal{A}_2^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{\mathcal{R}}^2(x)\mathcal{Z}(x)\mathcal{Z}^t(x)\right)\mathcal{A}_2^{-1}$$

and

$$\hat{\Sigma}_{\hat{\boldsymbol{\gamma}}_2} = \left(\mathcal{A}_1^{(1)}\right)^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{\mathcal{R}}_1^2(x)\mathcal{Z}(x)\mathcal{Z}^t(x)\right)\left(\mathcal{A}_1^{(1)}\right)^{-1}$$

where $\hat{\mathcal{R}}(x) = Y(x) - \mathcal{Z}^t(x)\hat{\boldsymbol{\theta}}_2$ and $\hat{\mathcal{R}}_1(x) = Y(x) - \mathcal{Z}^{(1)t}(x)\hat{\boldsymbol{\gamma}}_2$ are the residuals. Because the sum of the residuals over $s_{2,G}$ is now zero, we can write the new small-area estimator $\hat{\tilde{Y}}_{G,\mathrm{greg}}$ as in eq. (13)

$$(30) \qquad \hat{\hat{Y}}_{G,\mathrm{greg}} = \left(\bar{\mathcal{Z}}_G^{(1)} - \hat{\bar{\mathcal{Z}}}_{1,G}^{(1)}\right)^t \hat{\boldsymbol{\gamma}}_2 + \hat{\bar{\mathcal{Z}}}_{1,G}^t \hat{\boldsymbol{\theta}}_2$$

where we have set

$$\bar{\mathcal{Z}}_G^{(1)} = \frac{1}{\lambda(G)} \int_G \mathcal{Z}^{(1)}(x)\mathrm{d}x, \quad \hat{\bar{\mathcal{Z}}}_{1,G}^{(1)} = \frac{1}{n_{1,G}} \sum_{x \in s_{1,G}} \mathcal{Z}^{(1)}(x),$$
$$\hat{\bar{\mathcal{Z}}}_{1,G} = \frac{1}{n_{1,G}} \sum_{x \in s_{1,G}} \mathcal{Z}(x)$$

To get an estimate of the design-based variance we use mutatis mutandis for eq. (19)

$$(31) \qquad \hat{\mathbb{V}}(\hat{\hat{Y}}_{G,\mathrm{greg}}) = \frac{n_2}{n_1} \bar{\mathcal{Z}}_G^{(1)t} \hat{\Sigma}_{\hat{\boldsymbol{\gamma}}_2} \bar{\mathcal{Z}}_G^{(1)} + \left(1 - \frac{n_2}{n_1}\right) \hat{\bar{\mathcal{Z}}}_{1,G}^t \hat{\Sigma}_{\hat{\boldsymbol{\theta}}_2} \hat{\bar{\mathcal{Z}}}_{1,G}$$

For completeness, we also consider the pseudo-synthetic regression estimator (see Mandallaz (2013*a*))

$$(32) \qquad \hat{Y}_{\mathrm{psynth},G,\mathrm{reg}} = \hat{\bar{\mathbf{Z}}}_{1,G}^t \hat{\boldsymbol{\beta}}_2$$

with estimated variance

$$(33) \qquad \hat{\mathbb{V}}(\hat{Y}_{\mathrm{psynth},G,\mathrm{reg}}) = \hat{\bar{\mathbf{Z}}}_{1,G}^t \hat{\Sigma}_{\hat{\boldsymbol{\beta}}_2} \hat{\bar{\mathbf{Z}}}_{1,G} + \hat{\boldsymbol{\beta}}_2^t \hat{\Sigma}_{\hat{\bar{Z}}_{1,G}} \hat{\boldsymbol{\beta}}_2$$

where

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

and the pseudo-synthetic generalized regression estimator

$$(34) \qquad \hat{Y}_{\mathrm{psynth},G,\mathrm{greg}} = \left(\bar{\mathbf{Z}}_G^{(1)} - \hat{\bar{\mathbf{Z}}}_{1,G}^{(1)}\right)^t \hat{\boldsymbol{\alpha}}_2 + \hat{\bar{\mathbf{Z}}}_{1,G}^t \hat{\boldsymbol{\beta}}_2$$

with estimated variance

$$(35) \qquad \hat{\mathbb{V}}(\hat{Y}_{\mathrm{psynth},G,\mathrm{greg}}) = \frac{n_2}{n_1} \bar{\mathbf{Z}}_G^{(1)t} \hat{\Sigma}_{\hat{\boldsymbol{\alpha}}_2} \bar{\mathbf{Z}}_G^{(1)} + \left(1 - \frac{n_2}{n_1}\right) \hat{\bar{\mathbf{Z}}}_{1,G}^t \hat{\Sigma}_{\hat{\boldsymbol{\beta}}_2} \hat{\bar{\mathbf{Z}}}_{1,G}$$

The pseudo-synthetic estimators have a smaller variance but at the cost of potential bias because the mean residual over the small area G is not zero in general.

For a small area H with say $n_{2,G} < 5$, one can imbed H in a G with sufficiently large $n_{2,G}$ and calculate the pseudo-synthetic estimators (i.e., by using accordingly $\bar{\mathcal{Z}}_H^{(1)}$, $\hat{\bar{\mathcal{Z}}}_{1,H}^{(1)}$, $\hat{\bar{\mathcal{Z}}}_{1,H}$, $\hat{\boldsymbol{\gamma}}_2$, and $\hat{\boldsymbol{\theta}}_2$, in the previous formulae) in the model $\mathcal{Z}(x)$ extended with $I_G(x)$ instead of $I_H(x)$. The bias can be expected to be smaller than the bias of the synthetic estimators with respect to the model without $I_G(x)$.

7. Generalization to cluster and two-stage sampling

Worldwide, most national inventories rely on cluster sampling to reduce traveling costs, and two-stage sampling at the plot level is often used to increase the accuracy of timber volume (e.g., timber volume based on DBH and species is determined for all trees in $s_2(x)$ and a subset $s_3(x) \subset s_2(x)$ of trees is selected by Poisson sampling to get a better approximation of the volume based on, e.g., species, DBH, diameter at 7 m aboveground and tree height). All the previous results can be easily generalized to cluster sampling and also to two-stage sampling (the local density $Y(x)$ being simply replaced by the so-called generalized local density $Y^*(x)$). The reader will find all the details and examples in Mandallaz (2013*a*, 2013*b*) and Mandallaz and Massey (2012).

8. Examples

8.1. Simulation

To illustrate the theory and empirically check the validity of the various mathematical approximations used to derive the variance estimates, we present simulations performed on an example used in Mandallaz (2013*a*) to evaluate the performances of various estimators, in particular $\hat{Y}_{\mathrm{reg}}$.

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

$$(36) \qquad Y(x) = Y_0(x) + R(x)$$

where $Y_0(x) = \mathbf{Z}^t(x)\boldsymbol{\beta}_0$ and $R(x) = 6\cos(\pi x_1)\sin(2\pi x_2)$. We have chosen $\mathbf{Z}^{(1)}(x) = (1, x_1, x_2)^t$ and $\mathbf{Z}^{(2)}(x) = (x_1^2, x_1x_2, x_2^2)^t$. The exact least squares regression coefficient vector for the reduced model is found to be $\boldsymbol{\alpha}_0 = (25.17, 9.53, 3.00)^t$. Note that the components of $\boldsymbol{\alpha}_0$ are different from the first three components of $\boldsymbol{\beta}_0$. The coefficient of determination is $R^2 = 0.82$ for the full model $\mathbf{Z}(x)$ and $R^2 = 0.72$ for the reduced model $\mathbf{Z}^{(1)}(x)$. It is instructive to note that the reduced model ignores some important features of the local density $Y(x)$, albeit its R^2 is not much smaller than the R^2 of the full, and, in this case by definition correct, model. This can be best observed by looking at the plot of the local density given in Mandallaz (2013*a*).

The small area was defined as $G = [0.3,1.3] \times [0.5,2] \subset F$ with $\frac{\lambda(G)}{\lambda(F)} = \frac{1}{4}$.

The calculations for the small area under the so-called external model assumptions refer to $\hat{Y}_{G,\mathrm{greg}}$ and $\hat{\mathbb{V}}_{\mathrm{ext}}$ given in eqs. (23) and (24). Tables 1 and 2 summarize the simulation results that complete those given in Mandallaz (2013*a*).

All 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.

8.1.1. Discussion of the simulation results

Global estimation

Recall that, because of the zero mean residual over F, one has $\hat{\hat{Y}}_{F,\mathrm{greg}} = \hat{Y}_{F,\mathrm{greg}}$ (i.e., the model does not need to be extended because the indicator variable $I_F(x)$ is the intercept term).

- All point estimates are practically unbiased (even if the bias can be statistically significant owing to the huge sample size of 20 000 runs).
- The generalized regression estimator $\hat{Y}_{F,\mathrm{greg}}$ (with $n_1 < \infty$) substantially decreases the variance as compared with the standard regression estimator $\hat{Y}_{F,\mathrm{reg}}$ and comes even very close to $\hat{Y}_{F,\mathrm{reg}}$ with $n_1 < \infty$.
- The empirical variances are in good agreement with their estimated counterparts, particularly for $n_2 \geq 50$.
- The distribution of all estimators is bell-shaped but with heavier tails than the normal. The confidence limits based on a Student's distribution with $n_2 - p$ df (as suggested by the model-dependent approach) are too small, particularly for $n_2 = 25$. Using a Student's distribution with $n_2 - 2p$ df, one almost

Table 1. Simulation results for: $F = [0,2] \times [0,3]$.

	n_1:n_2		
Parameter	100:25	200:50	400:100
$\mathbb{E}^*(\hat{Y}_{F,\mathrm{greg}})$	39.17	39.16	39.17
$\mathbb{V}^*(\hat{Y}_{F,\mathrm{greg}})$	0.54	0.23	0.11
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{greg}}))$	0.36	0.19	0.10
$\mathbb{E}^*(\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{greg}}))$	0.32	0.18	0.10
$\hat{P}$	89.7	92.9	94.2
$\hat{P}_{\mathrm{ext}}$	88.9	92.3	93.8
$\mathbb{E}^*(\hat{Y}_{F,\mathrm{reg}})$	39.17	39.16	39.17
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{reg}}))$	0.76	0.39	0.19
	∞:25	∞:50	∞:100
$\mathbb{E}^*(\hat{Y}_{F,\mathrm{reg}})$	39.17	39.17	39.17
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{reg}}))$	0.32	0.17	0.09

Note: The true value is $\bar{Y}_F = 39.17$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ denote the empirical means and variances based on 20 000 runs. $\hat{Y}_{F,\mathrm{reg}}$ is the standard regression estimator for $\hat{Y}_{F,\mathrm{psynth}}$ discussed in Mandallaz (2013*a*), which for $n_1 = \infty$ is the same as $\hat{Y}_{F,\mathrm{synth}}$ and corresponds to the known $\bar{\mathbf{Z}}_F$. $\hat{P}$ is the empirical coverage probability in percent of the 95% confidence interval based on the Student's distribution with $n_2 - p$ df. $\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{greg}})$ and $\hat{P}_{\mathrm{ext}}$ refer to the calculations under the external model assumption.

Table 2. Simulation results for $G = [0.3,1.3] \times [0.5,2] \subset F$

	n_1:n_2		
Parameter	100:25	200:50	400:100
$\mathbb{E}^*(\hat{\hat{Y}}_{G,\mathrm{greg}})$	37.16	37.15	37.17
$\mathbb{V}^*(\hat{\hat{Y}}_{G,\mathrm{greg}})$	1.69	0.81	0.38
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{\hat{Y}}_{G,\mathrm{greg}}))$	1.29	0.71	0.36
$\hat{P}$	93.0	93.6	94.3
$\mathbb{E}^*(\hat{Y}_{G,\mathrm{greg}})$	37.13	37.14	37.17
$\mathbb{V}^*(\hat{Y}_{G,\mathrm{greg}})$	1.63	0.80	0.38
$\mathbb{E}^*(\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{G,\mathrm{greg}}))$	1.55	0.78	0.37
$\hat{P}_{\mathrm{ext}}$	94.0	94.2	94.7
$\mathbb{E}^*(\hat{\hat{Y}}_{G,\mathrm{reg}})$	37.15	37.17	37.16
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{\hat{Y}}_{G,\mathrm{reg}}))$	1.63	0.87	0.43
	∞:25	∞:50	∞:100
$\mathbb{E}^*(\hat{\hat{Y}}_{G,\mathrm{reg}})$	37.15	37.17	37.16
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{\hat{Y}}_{G,\mathrm{reg}}))$	1.23	0.65	0.34

Note: The true value is $\bar{Y}_G = 37.16$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ denote the empirical means and variances based on 20 000 runs. $\hat{P}$ and $\hat{P}_{\mathrm{ext}}$ are the empirical coverage probabilities (in percent) of the 95% confidence intervals based on the Student's distribution with $n_{2,G} - 1$ df. $\hat{\hat{Y}}_{G,\mathrm{reg}}$ is the regression estimator for $\hat{\hat{Y}}_{G,\mathrm{psynth}}$, the pseudo-synthetic estimator in the extended model, discussed in Mandallaz (2013*a*)), which for $n_1 = \infty$ is the same as $\hat{\hat{Y}}_{G,\mathrm{synth}}$ and corresponds to the known $\bar{\mathbf{Z}}_G$.

achieves the required 95% level. In most practical instances, n_2 will be so large that one can calculate the 95% confidence intervals according to the normal distribution.

- The external model assumption leads to a slight underestimation of the variance.
- 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 1 and 2, therefore, estimate the unconditional variances, whereas the variance estimates under the external model assumptions refer to the conditional variances, i.e., given $n_{1,G}$, $n_{2,G}$.

- All point estimates are practically unbiased (even if the bias can be statistically significant owing to the huge sample size of 20 000 runs).
- The generalized regression estimator $\hat{\hat{Y}}_{G,\mathrm{greg}}$ (with $n_1 < \infty$) substantially decreases the variance as compared with the standard regression estimator $\hat{\hat{Y}}_{G,\mathrm{reg}}$ and comes even close to $\hat{\hat{Y}}_{G,\mathrm{reg}}$ with $n_1 < \infty$.
- The agreement between empirical and estimated variances is satisfactory for $n_2 \geq 50$ ($\mathbb{E}(n_{2,G}) \geq 12.5$) and slightly better under the external model assumptions.
- The empirical coverage probabilities of the confidence intervals based on a Student's distribution with $n_{2,G} - 1$ df were close to the nominal values. The external model assumptions yield slightly wider confidence intervals.
- These simulations show that the generalized regression estimator can perform much better than the standard regression estimator and that the various asymptotic derivations for the variance estimates are adequate. As already noted in Mandallaz (2013*a*), the simple external model version performs very well in moderately large samples. Again, extending the model with the indicator variables of the small areas allows for a mathematically elegant solution. Of course, the simulation model used is very simple, with continuous explanatory variables and high R^2, and it would be premature to draw general conclusions. As shown in Mandallaz (2013*b*) in the case of poststratification, the g-weight-based variances have more appealing theoretical properties than those based on the external model assumptions. It can be conjectured that this will also hold in more complex situations with discrete explanatory variables.

8.2. Case study

We briefly present results obtained for a forest inventory at the enterprise level. The study object is located in eastern Switzerland, and the forests are located on steep slopes at an altitude of 1300–1700 m a.s.l. The rationale for this choice was threefold. First, it essentially consists of one tree species (*Picea abies* (L.) Karst.). Second, there is a considerable amount of structural variability, both horizontally and vertically. And third, high-resolution LiDAR data at leaf on status (TRIMBLE Harrier 68 scanner, >5 points per square metre) were available. The reader will find the technical details in Heinimann and Breschan (2012) and Hill (2013). After the extraction of a crown height model (CHM), it is possible to obtain several statistical parameters of the CHM, such as mean and various quantiles or measures of dispersion. By using sophisticated algorithms for tree top recognition (see Hyyppä et al. (2001) and Morsdorf et al. (2004)) it is possible to identify the locations and heights of the dominant trees and, with regression models (log-volume on log-height) based on the Swiss National Inventory data, also to estimate their volumes. The terrestrial second phase of the inventory is a simplified one-stage version of the Swiss National Inventory (SNI; with two concentric circles of 200 and 500 m² and volume function based on DBH alone; for details on SNI, see Mandallaz (2008), chapter 10), using a grid of 500 m × 500 m. The first phase of the inventory is based on squares $Q(x)$ of 637 m² circumscribed to the 500 m² circle at point $x \in s_2$ and lying on a 250 m × 250 m grid. We consider in detail the case in which the exhaustive auxiliary vector $\mathbf{Z}^{(1)}(x)$ at point x consists of the component $Z_1(x)$, the mean canopy height (in m) together with the intercept term $Z_0(x) \equiv 1$ term, and the small-area indicator variables $I_{G_k}(x)$. The nonexhaustive components $\mathbf{Z}^{(2)}(x)$ retained for the analysis are $Z_2(x)$, the maximal canopy height

Table 3. Characteristics of the auxiliary variables by areas.

	Area				
Auxillary variable	F	G_1	G_2	G_3	G_4
$\lambda(G)$	1974.5	586.6	531.9	425.6	430.4
$\frac{\lambda(G)}{\lambda(F)}$	1.00	0.30	0.27	0.21	0.22
$\hat{\bar{Z}}_1$	11.56	13.29	13.02	9.19	9.60
$\bar{Z}_1$	11.39	12.85	12.21	9.33	10.45
$\hat{\bar{Z}}_2$	32.77	35.66	35.31	28.23	29.95
$\hat{\bar{Z}}_3$	18.62	20.93	20.45	15.34	16.26
$\hat{\bar{Z}}_4$	9.05	9.86	9.78	7.57	8.27
$\hat{\bar{Z}}_5$	476.55	592.12	554.23	325.47	366.03
$\hat{\bar{Z}}_6$	378.40	362.91	352.12	395.00	416.69

Note: $\lambda(G)$ is the surface area in hectares of any domain G. $Z_1(x)$ is the mean canopy height (in m) at point x, with empirical and true means $\hat{\bar{Z}}_1$ and $\bar{Z}_1$ (exhaustive information). $Z_2(x)$ is the maximal canopy height (in m) at point x, $Z_3(x)$ is the 75% quantile of the canopy height at point x (in m), $Z_4(x)$ is the standard deviation of the canopy height at x (in m), with empirical weighted means in the large sample $\hat{\bar{Z}}_k$, $k = 2, 3, 4$ (nonexhaustive). $Z_5(x)$ is the LiDAR-estimated volume density at point x (in m^3·ha^{-1}), $Z_6(x)$ is the LiDAR-estimated density of stems at x, with empirical weighted means in the large sample $\hat{\bar{Z}}_k$, $k = 5, 6$ (nonexhaustive). The other exhaustive auxiliary variables are either the constant intercept term or the indicator variables of the small areas.

Table 4. Results for the entire domain F: $n_1 = 306$, $n_2 = 67$.

Estimator	Point estimate	Standard error
$\hat{Y}_F$	399.43	23.82
$\hat{Y}_{F,\mathrm{reg}}$	385.74	16.98
$\hat{Y}_{\mathrm{ext},F,\mathrm{reg}}$	385.74	16.45
$\hat{\tilde{Y}}^{(1)}_{F,\mathrm{reg}}$	384.18	17.03
$\hat{\tilde{Y}}^{(2)}_{F,\mathrm{reg}}$	385.62	16.98
$\hat{\tilde{Y}}^{(3)}_{F,\mathrm{reg}}$	386.07	16.95
$\hat{\tilde{Y}}^{(4)}_{F,\mathrm{reg}}$	384.08	16.86
$\hat{\tilde{Y}}^{(c)}_{F,\mathrm{reg}}$	383.29	16.95
$\hat{Y}_{F,\mathrm{greg}}$	381.65	14.73
$\hat{Y}_{\mathrm{ext},F,\mathrm{greg}}$	381.65	14.40
$\hat{\tilde{Y}}^{(1)}_{F,\mathrm{greg}}$	380.57	14.61
$\hat{\tilde{Y}}^{(2)}_{F,\mathrm{greg}}$	384.10	14.69
$\hat{\tilde{Y}}^{(3)}_{F,\mathrm{greg}}$	382.04	14.70
$\hat{\tilde{Y}}^{(4)}_{F,\mathrm{greg}}$	380.17	14.57
$\hat{\tilde{Y}}^{(c)}_{F,\mathrm{greg}}$	379.90	14.52

Note: $\hat{Y}_F$ is the sample mean. $\hat{Y}_{F,\mathrm{reg}} = \hat{Y}_{\mathrm{ext},F,\mathrm{reg}}$ is the regression estimator (eq. (10)). Because of zero mean residual over F, the external model assumption leads to the same point estimate but with a different variance, likewise for the generalized regression estimator, $\hat{Y}_{F,\mathrm{greg}}$ (eq. (7)). $\hat{\tilde{Y}}^{(k)}_{F,\mathrm{reg}}$ are the regression estimators in the extended model with a single indicator variable, $I_{G_k}(x)$ ($k = 1, 2, 3, 4$); likewise for the generalized regression estimators, $\hat{\tilde{Y}}^{(k)}_{F,\mathrm{greg}}$ $\hat{\tilde{Y}}^{(c)}_{F,\mathrm{reg}}$ is the regression estimator with all four indicator variables, $I_{G_k}(x)$, but without an intercept, likewise for the generalized regression estimator, $\hat{\tilde{Y}}^{(c)}_{F,\mathrm{greg}}$.

Table 5. Results for the small areas.

	Small area			
Parameter	G_1	G_2	G_3	G_4
$n_{1,G}$:$n_{2,G}$	94:19	81:17	66:15	65:16
$\hat{Y}_G$	410.40 (44.58)	461.44 (56.35)	318.00 (34.36)	396.85 (47.86)
$\hat{\tilde{Y}}^{(k)}_{G,\mathrm{reg}}$	395.12 (30.85)	427.06 (34.02)	315.22 (33.68)	370.79 (33.20)
$\hat{\tilde{Y}}^{(c)}_{G,\mathrm{reg}}$	394.59 (30.92)	428.05 (33.78)	323.49 (35.17)	371.16 (33.12)
$\hat{Y}_{\mathrm{ext},G,\mathrm{reg}}$	398.29 (28.80)	426.48 (35.01)	315.17 (31.66)	368.08 (36.01)
$\hat{\tilde{Y}}^{(k)}_{G,\mathrm{greg}}$	384.07 (24.70)	408.41 (29.98)	318.30 (30.23)	391.24 (30.62)
$\hat{\tilde{Y}}^{(c)}_{G,\mathrm{greg}}$	383.77 (24.71)	408.21 (29.43)	326.76 (31.45)	392.05 (30.42)
$\hat{Y}_{\mathrm{ext},G,\mathrm{greg}}$	387.95 (25.02)	407.50 (29.19)	318.30 (30.93)	388.06 (31.24)

Note: Standard errors of the point estimates are given in parentheses. $\hat{Y}_G$ is the sample mean in the small area. $\hat{\tilde{Y}}^{(k)}_{G,\mathrm{reg}}$ is the regression estimator for small-area G_k in the extended model with the single indicator variable, $I_{G_k}(x)$, $k = 1, 2, 3, 4$; $\hat{\tilde{Y}}^{(c)}_{G,\mathrm{reg}}$ is the regression estimator with all four indicator variables $I_{G_k}(x)$, but without the intercept and likewise for the generalized regression estimators $\hat{\tilde{Y}}^{(k)}_{F,\mathrm{greg}}$. $\hat{Y}_{\mathrm{ext},G,\mathrm{greg}}$ and $\hat{Y}_{\mathrm{ext},G,\mathrm{greg}}$ are the regression and generalized regression estimators, respectively, for the small area G_k in the external model approach.

(in m); $Z_3(x)$, the 75% quantile; $Z_4(x)$, the standard deviation; calculated with data in $Q(x) \cap F$ and, therefore, adjusted for boundary effects at the forest edge. Further nonexhaustive components are $Z_5(x)$, the LiDAR-estimated volume density (in m^3·ha^{-1}); and $Z_6(x)$, the LiDAR-estimated density of stems obtained from single tree recognition; $Z_5(x)$ and $Z_6(x)$ are based on the 500 m^2 circle at x and adjusted for boundary effects at the forest edge. The true means of $Z_1(x)$ over the forested area F or any small area $G \subset F$ are calculated as the weighted means over the squares $Q(x)$ with center point x in F, the weights being proportional to $\lambda(Q(x) \cap F)$. To ensure asymptotic consistency for $n_1 \rightarrow \infty$, the same procedure is used to calculate the empirical means of $Z_k(x)$, $l = 2, 3, 4, 5, 6$ over F and G in the large sample $s_1 \supset s_2$ (for the intercept, $\bar{Z}_0 = 1$ directly). The terrestrial local density $Y(x)$ for timber volume at $x \in s_2$ performs boundary adjustment at the tree level. Note that one cannot obtain the true mean of $\boldsymbol{Z}^{(1)}(x)$ by using circles because a tessellation of the plane with circles is obviously impossible. Mathematically, the vector function $\boldsymbol{Z}(x)$ is assumed to be stepwise constant over the $Q(x) \cap F$ (31 622 squares have a nonvoid intersection with F). It must be emphasized that the calculations for $Z_5(x)$ and $Z_6(x)$ are computationally intensive as compared with $Z_k(x)$, $k = 2, 3, 4$; they have been implemented in the MATLAB software. The statistical calculations have been programmed within the iml procedure of the statistical software package SAS. The reduced model with the exhaustive variable $Z_1(x)$ alone yields $R^2 = 0.50$ ($Z_5(x)$ alone yields $R_2 = 0.52$), whereas the full model gives $R^2 = 0.67$. The choice of this model is based on forward and backward selection techniques and goodness-of-fit criteria (R^2, C_p, AIC), but also, and even primarily, on the interpretability of the results. Because of the collinearity between many auxiliary variables, other choices are possible, leading essentially to the same R^2. In our opinion, the final choice is not the responsibility of the statistical algorithm but of the subject scientist. All the auxiliary variables except for $Z_6(x)$ are significant in the model-dependent sense. $Z_6(x)$ was retained because it is known from the SNI data to play a role. Another practical difficulty is that the nominal GPS coordinates of the terrestrial and LiDAR plots were possibly not equal to their actual values (i.e., the points denoted by x in $Y(x)$ and $\boldsymbol{Z}(x)$ are not necessarily equal in practice), with a possible discrepancy of up to 10 m (Steinmann et al. (2011)), essentially because of the difficult topographical conditions. This can potentially reduce the coefficients of determination. It is possible to slightly improve the fit by using automatic matching algorithms, aimed at identifying a point u in the neighborhood of point x that optimizes the similarity of sample tree patterns and volume density between terrestrial and LiDAR-derived data (Hill (2013)), but we prefer to present only the results based on the raw data. It can be expected that the GPS accuracy will be improved in the near future. In any case, the LiDAR-based explanatory variables yield a substantial improvement as compared with the regression estimators using the

Table 6. Pseudo-synthetic estimators for the small areas.

	Small area			
Estimator	G_1	G_2	G_3	G_4
$\hat{Y}_{\text{psynth},G,\text{reg}}$	425.96 (24.27)	421.67 (23.69)	319.87 (24.31)	334.59 (22.81)
$\hat{Y}_{\text{psynth},G,\text{greg}}$	415.61 (15.49)	402.69 (14.98)	323.01 (17.00)	354.57 (16.91)

Note: Point estimates and standard errors (in parentheses) were calculated according to eqs. (32)–(35).

Table 7. Results with $Z_k(x) = 1, 2, 3, 4$ exhaustive.

	Area				
Estimator	F	G_1	G_2	G_3	G_4
$\hat{Y}^{(c)}_{\text{greg}}$	377.47 (14.00)	377.02 (24.00)	397.43 (26.83)	330.67 (31.16)	403.99 (30.25)

standard Swiss stand maps obtained by the interpretation of aerial photographs, with R^2 in the range 0.3–0.4 (Mandallaz (2008), chapter 10).

The fitted models were

- For the full model

$$\hat{Y}(x) = 322.57 + 52.55Z_1(x) - 19.24Z_2(x) - 33.04Z_3(x) + 71.06Z_4(x) + 0.19Z_5(x) - 0.09Z_6(x)$$

- For the reduced model

$$\hat{Y}_1(x) = 116.16 + 23.46Z_1(x)$$

The entire domain F was partitioned into four approximately equal small areas (G_1, G_2, G_3, and G_4) in a north–south direction. These regression coefficients can vary significantly, but not substantially, after the introduction of one or more small-area indicator variables.

Table 3 displays the main characteristics of F, G_1,…G_4, and Tables 4, 5, and 6 summarize the results. In this example, the auxiliary components $Z_k(x)$, $k = 2, 3, 4$ were also available in their exhaustive versions, and it is possible to calculate the generalized regression estimator with the reduced model $\mathbf{Z}^{(1)} = (Z_1(x), Z_2(x), Z_3(x), Z_4(x))^t$. Table 7 summarizes the results for the combined estimator (i.e., with all four small-area indicator variables). Results for the other estimators were very similar. The variance reduction from the sample mean to the regression estimator is substantial (particularly for small-area estimation), and it can be further improved, but not as dramatically, by using the generalized regression estimator. The mean canopy height appears to be the exhaustive variable of choice, as it is not very computing intensive and alone it yields most of the supplementary variance reduction.

8.2.1. *Discussion of the case study*

Clearly, the regression estimator substantially reduces the error as compared with the sample mean, a reduction further enhanced by the generalized regression estimators. The introduction of one or more small-area indicator variables has little impact, neither on the point estimates nor on the errors. Again, the external model approach is remarkably close to the more sophisticated procedures, probably because the sample sizes (around 17 terrestrial observations) in the small areas were large enough. The goodness of fit in the small area G_3 is far below average (0.3 as compared with 0.7 in the other small areas), so that in this case the regression estimators do not perform much better than the sample mean. A possible explanation is that the forest is more homogenous in G_3: $Z_4(x)$ is smaller and the slopes of the simple linear regression of $Y(x)$ on $Z_1(x)$ or $Z_5(x)$ alone are much smaller than in the other small areas and only borderline significant (in other words we have an interaction between the small areas and the most important auxiliary variables).

We also considered smaller areas (with 5–10 observations) and the differences between the various versions of the estimators were more important (for the point estimates as well as the errors), with some evidence that the external approach yields larger errors, as also suggested by the simulation results.

The various versions of the regression or generalized regression small-area estimates are close to each other, except for the pseudo-synthetic estimators, for which the differences, although not significant, can be more than one standard error, especially in G_4. The standard errors of the pseudo-synthetic estimators are, of course, substantially smaller, but at the cost of potential bias. The generalized regression estimator with the maximum possible number of exhaustive components yields a further very small variance reduction (except in G_3), but is more computing intensive.

One reviewer raised the question as to the additivity of the estimates, i.e., whether one has

$$(37) \qquad \hat{Y}_F = \sum_{k=1}^{4} \frac{\lambda(G_k)}{\lambda(F)} \hat{Y}_{G_k} =: \hat{Y}_{\text{weighted}}$$

where $\hat{Y}_F$ and $\hat{Y}_{G_k}$ are any of the estimators considered here. The answer is that we have only approximate additivity (this will also hold true for the geostatistical point estimates). The relative absolute difference $\Delta = \frac{|\hat{Y}_F - \hat{Y}_{\text{weighted}}|}{\hat{Y}_F}$ is highest for the external estimators (1.21%) and lowest for the combined generalized estimators (0.008%), as expected on algebraic grounds, since the regression coefficients do not depend on the particular domains in this case.

The external model approach is easier to implement as it requires only predictions and residuals that can be obtained from standard packages such as SAS or R; this can be an advantage particularly for complex $\mathbf{Z}(x)$ (incorporating, e.g., interaction or nested effects). Also, one could use robust estimation procedures for linear models. One should keep in mind that the resulting errors of the regression coefficients given by SAS or R are model-dependent and not design-based, although the differences were small in the present case. From a mathematical point of view, the methods presented here are more coherent and also necessary if one has to rely on synthetic estimators.

The computing-intensive LiDAR-based estimators yield a substantial improvement as compared with the classical procedures relying on stand maps. Ongoing research investigates these techniques at the national and regional levels.

9. Conclusion

The generalized regression estimator that makes use of the partially exhaustive information is a further improvement over the classical regression estimator, as shown by the simulations and the case study. From a mathematical point of view, the introduction of models extended by the indicator variables of small areas is very appealing and yields excellent asymptotic properties of the resulting point and error estimates. It is also reassuring to see that the naive external model approach, which can be readily implemented with existing software packages, also performs very well if the samples are sufficiently large. For very small areas, one must rely on the techniques presented in this paper, in particular for the error of the synthetic estimators. The new estimators appear to be particularly promising in the context of LiDAR-based auxiliary information.

Acknowledgements

We express our thanks to the reviewers for their comments and suggestions, which greatly improved the manuscript; to Professor H. Heinimann (Chair of Land Use Engineering, ETH Zurich) for his support; and to the Forest Service of the Canton of Grisons (Switzerland) and the Remote Sensing Laboratories of the University of Zurich for providing the data used in the case study.

References

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

Heinimann, H., and Breschan, J. 2012. Pre-harvest assessment based on LiDAR data. Croat. J. For. Eng. **33**: 169–180.

Hill, A. 2013. Comparison of small area estimators in forest inventories using airborne laser-scanning data. Technical report, ETH Zurich, Department of Environmental Systems Science and University of Göttingen, M.Sc. thesis. Available from http://e-collection.library.ethz.ch.

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, Berkeley, California, 21 June to 18 July 1965 and 27 December 1965 to 7 January 1966. Vol 1: Statistics. University of California Press, Berkeley, Calif. pp. 221–233.

Hyyppä, J., Kelle, O., Lehikoinen, M., and Inkinen, M. 2001. A segmentation-based method to retrieve stem volume estimates from 3-D height models produced by laser scanner. IEEE Trans. Geosci. Remote Sens. **39**: 969–975. doi:10.1109/36.921414.

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

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

Mandallaz, D. 2013*a*. Design-based properties of some small-area estimators in forest inventory with two-phase sampling. Can. J. For. Res. **43**(5): 441–449. doi:10.1139/cjfr-2012-0381.

Mandallaz, D. 2013*b*. Regression estimators in forest inventories with two-phase sampling and partially exhaustive information with application to small-area estimation. ETH Zurich, Department of Environmental Systems Science, Tech. rep. 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**(10): 1865–1871. doi:10.1139/x2012-110.

Morsdorf, F., Meier, E., Kötz, B., Itten, K.I., Dobbertine, M., and Allgöwer, B. 2004. LiDAR-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. Remote Sens. Environ. **92**: 353–362. doi:10.1016/j.rse.2004.05.013.

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

Steinmann, K., Ginzler, C., and Lanz, A. 2011. Kombination von Landesforstinventar- und Fernerkundungsdaten für Kleingebietsschätzungen. Schweiz. Z. Forstwes. **162**: 290–299. doi:10.3188/szf.2011.0290.