NRC Research Press

NOTE

A three-phase sampling extension of the generalized regression estimator with partially exhaustive information

Daniel Mandallaz

Abstract: We consider three-phase sampling schemes in which one component of the auxiliary information is known in the very large sample of the so-called null phase and the second component is available only in the large sample of the first phase, whereas the second phase provides the terrestrial inventory data. We extend to three-phase sampling the generalized regression estimator that applies when the null phase is exhaustive, for global and local estimation, and derive its asymptotic design-based variance. The new three-phase regression estimator is particularly useful for reducing substantially the computing time required to treat exhaustively very large data sets generated by modern remote sensing technology such as LiDAR.

Résumé : Nous proposons un nouvel estimateur pour les inventaires forestiers utilisant des plans de sondage à trois phases pour lesquels l'information auxiliaire consiste en une première composante connue en chaque point de la phase « nulle » et une seconde composante connue seulement en chaque point de la première phase, alors que la deuxième phase consiste en l'inventaire terrestre. Nous proposons une nouvelle version de l'estimateur par regression, aussi bien pour l'estimation globale que locale, et nous donnons la variance asymptotique sous le plan de sondage. Le nouvel estimateur est particulièrement utile pour réduire substantiellement le temps de calcul requis pour un traitement exhaustif de très grandes bases de données obtenues par les moyens modernes de télédétection tels que LiDAR.

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 by, e.g., remote sensing (aerial photographs or LiDAR data). The small-area estimation problem is, in this context, of the utmost importance. This paper adapts the so-called generalized regression estimator proposed in Mandallaz et al. (2013) to the case in which the first component of the auxiliary information is no longer exhaustive but is provided by a very large sample, the null phase. The second component is available on a subsample of the null phase, the first phase, and the terrestrial inventory is performed on a subsample of the first phase, the second phase. This setup is particularly useful for national or regional inventories for two reasons: (*i*) the first component may not be available exhaustively, and (*ii*) even if it were, it may be computationally prohibitive for some of its variables, particularly those based on sophisticated algorithms requiring single tree identification (as in Mandallaz et al. 2013). The "big data" issue resulting from fast developments in remote sensing technology is a challenge even for powerful computing resources, and the three-phase regression estimators will certainly play an important role in this context.

The methodology and terminology of the present paper 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 new approach to regression estimators particularly useful in the small-area estimation context. Chapters 4 and 5 in Mandallaz (2008) are recommended for a first perusal. The interested reader will find the proofs of the results, rather technical, together with further developments in the on-line technical report Mandallaz (2013*b*). Fortunately, parts of the results (valid under the so-called external model assumption) have a very intuitive background and can be easily implemented with standard statistical software packages, while their performances are close to those of the more advanced g-weight procedures, which require some matrix algebra programming. Also, it is worth mentioning that at the present time, there are no simple alternative model-dependent techniques for this three-phase setup (relying on, e.g., Kriging or mixed models).

The present work generalizes in many respects results (g-weight variances and small-area estimation) given in von Lüpke et al. (2012), where the first component is generated by a categorical variable defining strata. In a slightly different context, Fattorini et al. (2006) considers the case in which both components are generated by stratification and the null phase is performed with unaligned systematic sampling. Current work deals with the implementation of the three-phase estimators in up-dating procedures for the Swiss National Inventory (SNI), which has moved from a periodic (every 10 years) to an annual survey (up-dating was also the issue in von Lüpke et al. (2012)).

2. Methodology

The null phase draws a very large sample s_0 of n_0 points, $x_i \in s_0$ ($i = 1, 2, \ldots, n_0$), 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 (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 $\boldsymbol{Z}^{(1)}(x) \in \Re^p$. The case $n_0 = \infty$, i.e., $\boldsymbol{Z}^{(1)}(x)$ is exhaustive, has been investigated in Mandallaz et al. (2013). The first phase draws a large sample $s_1 \subset s_0$ of $n_1 \ll n_0$ points by simple random sampling in s_0. Note that the points $x \in s_1$ are also uniformly independently distributed in F. For each point in the first phase, a further component $\boldsymbol{Z}^{(2)}(x) \in \Re^q$ of the auxiliary information is available and hence also the vector

Received 1 November 2013. Accepted 13 December 2013.

D. Mandallaz. ETH Zurich, Department of Environmental Systems Science, Chair of Land Use Engineering, CHN K74.1, CH 8092 Zurich, Switzerland.
E-mail for correspondence: daniel.mandallaz@env.ethz.ch.

$\boldsymbol{Z}^t(x) = (\boldsymbol{Z}^{(1)t}(x), \boldsymbol{Z}^{(2)t}(x)) \in \Re^{p+q}$ (the superscript t denotes the transposition operator). The second phase draws a small sample $s_2 \subset s_1$ of n_2 points from s_1 by simple random sampling and consists of the terrestrial inventory. Note that we have used the terms null, first, and second phases instead of first, second, and third phases simply to ensure compatibility with the terminology used in previous work in which the null-phase was exhaustive.

To set the stage, the component $\boldsymbol{Z}^{(1)}(x) \in \Re^p$ can be based on the interpretation of aerial photographs or on simple characteristics of the canopy height obtained from LiDAR data (e.g., mean canopy height and eventually quantiles thereof), whereas $Z^{(2)}(x) \in \Re^q$ is based on other computationally intensive characteristics of the canopy requiring individual tree detection (e.g., tree species or tree volume prediction based on tree height). The null-phase sample s_0 is introduced because the component $\boldsymbol{Z}^{(1)}(x)$ can be computationally prohibitive to calculate exhaustively in extensive forest inventories (see Mandallaz (2013*a*) for a case study with LiDAR data). In the aforementioned continuous annual SNI $\boldsymbol{Z}^{(1)}(x)$ from the null phase is based on data obtained from aerial photographs and on simple stratification criteria, $\boldsymbol{Z}^{(2)}(x)$ from the first phase is based on updates of previous terrestrial inventory plots, and the second phase provides the annual local density $Y(x)$ defined below.

In the forested area F, we consider a well-defined population $\mathcal{P}$ of N trees with response variable Y_i, $i = 1, 2, \ldots, N$, e.g., the timber volume. The objective is to estimate the spatial mean $\overline{Y} = \frac{1}{\lambda(F)}\sum_{i=1}^{N} Y_i$, where $\lambda(F)$ denotes the surface area of F (usually in hectares). 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 $I_i(x)$ is equal to 1 if $i \in s_2(x)$, otherwise $I_i(x) = 0$. At each point $x \in s_2$, the terrestrial inventory provides the local density $Y(x)$:

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

The term $\frac{1}{\lambda(F)\pi_i}$ is the tree extrapolation factor f_i with dimension ha^{-1}. Because of 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 $Y(x)$, and we have $\mathbb{E}_x(Y(x)) = \frac{1}{\lambda(F)}\int_F Y(x)\mathrm{d}x = \frac{1}{\lambda(F)}\sum_{i=1}^{N} Y_i = \overline{Y}$, where $\mathbb{E}_x$ denotes the expectation with respect to a random point x uniformly distributed in F. In practice, one uses embedded systematic grids, which in most instances can be treated as random samples for global estimation, whereas for local estimation, the design-based error can be expected to be slightly larger than the model-dependent error obtained by the more sophisticated double kriging techniques (the geostatistical version of the standard two-phase regression estimator; Mandallaz 2008, chapter 8). It can be safely conjectured that the same will hold true in the present context.

3. The models

We shall work with the following linear models.

1. The large model M

$$(2) \qquad Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta} + R(x) = \boldsymbol{Z}^{(1)t}(x)\boldsymbol{\beta}^{(1)} + \boldsymbol{Z}^{(2)t}(x)\boldsymbol{\beta}^{(2)} + R(x) =: \hat{Y}(x) + R(x)$$

with $\boldsymbol{\beta}^t = (\boldsymbol{\beta}^{(1)t}, \boldsymbol{\beta}^{(2)t})$ and the theoretical predictions $\hat{Y}(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}$. The intercept term is contained in $\boldsymbol{Z}^{(1)}(x)$ or a linear combination of the components of $\boldsymbol{Z}^{(1)}(x)$ is constant equal to 1. The theoretical regression parameter $\boldsymbol{\beta}$ minimizes $\int_F (Y(x) - \boldsymbol{Z}^{(1)t}(x)\boldsymbol{\beta})^2\mathrm{d}x$; it satisfies the normal equation $(\int_F \boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\mathrm{d}x)\boldsymbol{\beta} = \int_F Y(x)\boldsymbol{Z}(x)\mathrm{d}x$ and the orthogonality relationship $\int_F R(x)\boldsymbol{Z}(x)\mathrm{d}x = \boldsymbol{0}$, 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

$$(3) \qquad Y(x) = \boldsymbol{Z}^{(1)t}(x)\boldsymbol{\alpha} + R_1(x) =: \hat{Y}_1(x) + R_1(x)$$

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

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 the above models to be correct in the sense of model-dependent inference (see Mandallaz (2013*a*)).

4. The three-phase 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:

$$(4) \qquad \hat{\boldsymbol{\alpha}}_k = \left(\frac{1}{n_k}\sum_{x\in s_k}\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)t}(x)\right)^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}^{(1)}(x) := \left(\boldsymbol{A}_k^{(1)}\right)^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}^{(1)}(x) =: \left(\boldsymbol{A}_k^{(1)}\right)^{-1}\boldsymbol{U}_k^{(1)}, \quad k = 0, 1, 2$$

Likewise for the large model, we set

$$(5) \qquad \hat{\boldsymbol{\beta}}_k = \left(\frac{1}{n_k}\sum_{x\in s_k}\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\right)^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}(x) = \boldsymbol{A}_k^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}(x) =: \boldsymbol{A}_k^{-1}\boldsymbol{U}_k, \quad k = 0, 1, 2$$

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 $\hat{\boldsymbol{\alpha}}_2$.

For simplicity, we shall use the same notation for the theoretical and empirical predictions of both models, i.e., we set $\hat{Y}(x) = Z^t(x)\hat{\boldsymbol{\beta}}_2$ and $\hat{Y}_1(x) = Z^{(1)t}(x)\hat{\boldsymbol{\alpha}}_2$.

Consistent estimates of the design-based covariance matrices are given by

$$(6) \qquad \begin{aligned}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_2} &= \boldsymbol{A}_2^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\right)\boldsymbol{A}_2^{-1}\\ \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_2} &= \left(\boldsymbol{A}_1^{(1)}\right)^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}_1^2(x)\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)t}(x)\right)\left(\boldsymbol{A}_1^{(1)}\right)^{-1}\end{aligned}$$

with the empirical residuals $\hat{R}(x) = Y(x) - \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_2$ and $\hat{R}_1(x) = Y(x) - \boldsymbol{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}_2$ (see Mandallaz (2013*a*) and Mandallaz (2008, pp. 124–125) for details).

The generalized regression estimate discussed by Mandallaz et al. (2013) is defined as

$$\hat{Y}_{F,\mathrm{greg}} = \frac{1}{\lambda(F)}\int_F \hat{Y}_1(x)\mathrm{d}x + \frac{1}{n_1}\sum_{x\in s_1}(\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x) - \hat{Y}(x)) = \left(\overline{Z}^{(1)} - \hat{\overline{Z}}_1^{(1)}\right)^t\hat{\boldsymbol{\alpha}}_2 + \hat{\overline{Z}}_1^t\hat{\boldsymbol{\beta}}_2 \tag{7}$$

where

$$\overline{Z}^{(1)} = \frac{1}{\lambda(F)}\int_F Z^{(1)}(x)\mathrm{d}x, \quad \hat{\overline{Z}}_1^{(1)} = \frac{1}{n_1}\sum_{x\in s_1} Z^{(1)}(x), \quad \hat{\overline{Z}}_k = \frac{1}{n_k}\sum_{x\in s_k} Z(x), \quad k = 1, 2 \tag{8}$$

Here, $Z^{(1)}(x)$ is exhaustive, i.e., known at all $x \in F$. $\hat{Y}_{F,\mathrm{greg}}$ is the sum of three components: the true mean of the predictions of the reduced model (null phase), the mean of the differences between the predictions of the large and reduced models (first phase), and the mean of the residuals of the large model (second phase).

Under the external model assumption, the design-based variance is given by

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

Furthermore,

$$\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{greg}}) = \frac{n_2}{n_1}\overline{Z}^{(1)t}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_2}\overline{Z}^{(1)} + \left(1 - \frac{n_2}{n_1}\right)\hat{\overline{Z}}_1^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_2}\hat{\overline{Z}}_1 \tag{10}$$

is a consistent estimate of the design-based variance of $\hat{Y}_{F,\mathrm{greg}}$, which can be rewritten in g-weight form (see Mandallaz et al. (2013)). Under the external model assumptions, it is straightforward to obtain approximate estimates of the design-based variances of $\hat{Y}_{F,\mathrm{greg}}$ by using the standard variance estimates in eq. 9 after replacing the theoretical residuals by the empirical ones. The generalized regression estimator $\hat{Y}_{G,\mathrm{greg}}$ for any small area $G \subset F$ has been discussed in Mandallaz et al. (2013).

If $Z^{(1)}(x)$ is no longer exhaustive, we replace the first term in eq. 7 by its sample mean in the null-phase sample and we define the new three-phase estimator as

$$\hat{Y}_{F,\mathrm{g3reg}} = \frac{1}{n_0}\sum_{x\in s_0}\hat{Y}_1(x) + \frac{1}{n_1}\sum_{x\in s_1}(\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x) - \hat{Y}(x)) = \left(\hat{\overline{Z}}_0^{(1)} - \hat{\overline{Z}}_1^{(1)}\right)^t\hat{\boldsymbol{\alpha}}_2 + \hat{\overline{Z}}_1^t\hat{\boldsymbol{\beta}}_2 \tag{11}$$

where $\hat{\overline{Z}}_0^{(1)} = \frac{1}{n_0}\sum_{x\in s_0} Z^{(1)}(x)$.

Note that the third terms in eqs. 7 and 11 vanish for internal models because, by assumption, the intercept term is contained in $Z^{(1)}$ (or it can be expressed as a linear combination of its components) so that the residuals sum up to zero over s_2.

By using the properties of conditional expectations and variances (see Mandallaz (2008), appendix B), we see that $\hat{Y}_{F,\mathrm{g3reg}}$ is asymptotically design unbiased and that, under the external model assumption, one has

$$\mathbb{V}_{0,1,2}(\hat{Y}_{F,\mathrm{g3reg}}) = \mathbb{E}_0\mathbb{E}_{1|0}\mathbb{V}_{2|0,1}(\hat{Y}_{F,\mathrm{g3reg}}) + \mathbb{E}_0\mathbb{V}_{1|0}\mathbb{E}_{2|0,1}(\hat{Y}_{F,\mathrm{g3reg}}) + \mathbb{V}_0\mathbb{E}_{1|0}\mathbb{E}_{2|0,1}(\hat{Y}_{F,\mathrm{g3reg}}) = \frac{1}{n_0}\mathbb{V}_x(Y(x)) + \left(1 - \frac{n_1}{n_0}\right)\frac{1}{n_1}\mathbb{V}_x(R_1(x)) + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\mathbb{V}_x(R(x)) \tag{12}$$

As the predictions $\hat{Y}_1(x)$ and the residuals $R_1(x)$ are orthogonal in the design-based sense then, because $\mathbb{V}_x(Y(x)) = \mathbb{V}_x(\hat{Y}_1(x)) + \mathbb{V}_x(R_1(x))$, one obtains

$$\mathbb{V}_{0,1,2}(\hat{Y}_{F,\mathrm{g3reg}}) = \frac{1}{n_0}\mathbb{V}_x(\hat{Y}_1(x)) + \frac{1}{n_1}\mathbb{V}_x(R_1(x)) + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2}\mathbb{V}_x(R(x)) \tag{13}$$

Alternatively, one can use the design-based covariance matrices in eq. 6 to obtain the consistent g-weight like variance estimate

$$\hat{\mathbb{V}}_{0,1,2}(\hat{Y}_{F,\mathrm{g3reg}}) = \hat{\boldsymbol{\alpha}}_2^t\hat{\boldsymbol{\Sigma}}_{\hat{\overline{Z}}_0^{(1)}}\hat{\boldsymbol{\alpha}}_2 + \frac{n_2}{n_1}\hat{\overline{Z}}_0^{(1)t}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_2}\hat{\overline{Z}}_0^{(1)} + \left(1 - \frac{n_2}{n_1}\right)\hat{\overline{Z}}_1^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_2}\hat{\overline{Z}}_1^t \tag{14}$$

where

$$\hat{\boldsymbol{\Sigma}}_{\hat{\overline{Z}}_0^{(1)}} = \frac{1}{n_0}\frac{\sum_{x\in s_0}\left(Z^{(1)}(x) - \hat{\overline{Z}}_0^{(1)}\right)\left(Z^{(1)}(x) - \hat{\overline{Z}}_0^{(1)}\right)^t}{n_0 - 1} \tag{15}$$

One can rewrite eq. 14 in a g-weight form, computationally more suitable for practical implementation (see Mandallaz (2013*b*) for details). Equation 14 is asymptotically equivalent to the following variance estimate obtained from the empirical version of eq. 13:

$$\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{g3reg}}) = \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}_1(x) - \hat{\overline{Y}}_1)^2}{n_0 - 1} + \frac{1}{n_1}\frac{1}{n_2 - 1}\times\sum_{x\in s_2}(\hat{R}_1(x) - \hat{\overline{R}}_1)^2 + \left(1 - \frac{n_2}{n_1}\right)\frac{1}{n_2(n_2 - 1)}\sum_{x\in s_2}(\hat{R}(x) - \hat{\overline{R}})^2 \tag{16}$$

where $\hat{\overline{Y}}_1 = \frac{1}{n_0}\sum_{x\in s_0}\hat{Y}_1(x)$, $\hat{\overline{R}}_1 = \frac{1}{n_2}\sum_{x\in s_2}\hat{R}_1(x) = 0$ and $\hat{\overline{R}} = \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x) = 0$. The proof is rather intricate and is given in Mandallaz (2013*b*).

By comparing with eq. 10, we see that in large samples, we have

$$\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{g3reg}}) - \hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{greg}}) = \hat{\boldsymbol{\alpha}}_2^t\hat{\boldsymbol{\Sigma}}_{\hat{\overline{Z}}_0^{(1)}}\hat{\boldsymbol{\alpha}}_2 = \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}_1(x) - \hat{\overline{Y}}_1)^2}{n_0 - 1} \tag{17}$$

Hence, the extra term in the variance due to three-phase sampling is given by the variance of the predictions of the reduced model in the very large null-phase sample, which can be made very small while cutting down significantly the computing time.

5. Application to small-area estimation

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

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

where $I_G(i) = 1$ if tree i is in G, otherwise $I_G(i) = 0$. Strictly speaking, the last equality holds 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: $s_{0,G} = s_0 \cap G$, $s_{1,G} = s_1 \cap G$, $s_{2,G} = s_2 \cap G$, $n_{k,G} = \sum_{s\in s_k} I_G(x)$, $k = 0, 1, 2$.

The simplest procedure is to restrict the samples to G, i.e., to consider the generalized small-area estimator:

$$\hat{Y}_{G,\mathrm{g3reg}} = \frac{1}{n_{0,G}}\sum_{x\in s_{0,G}}\hat{Y}_1(x) + \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}(\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}(Y(x) - \hat{Y}(x)) \tag{18}$$

Under the external model assumption and further mild regularity conditions (see Mandallaz (2013*b*)), one can derive the following variance estimate (similar to eq. 16):

$$\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{G,\mathrm{g3reg}}) = \frac{1}{n_{0,G}(n_{0,G}-1)}\sum_{x\in s_{0,G}}(\hat{Y}_1(x) - \hat{\bar{Y}}_{1,G})^2 + \frac{1}{n_{1,G}(n_{2,G}-1)}\sum_{x\in s_{2,G}}(\hat{R}_1(x) - \hat{\bar{R}}_{1,G})^2 + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right)\frac{1}{n_{2,G}(n_{2,G}-1)}\sum_{x\in s_{2,G}}(\hat{R}(x) - \hat{\bar{R}}_G)^2 \tag{19}$$

with

$$\hat{\bar{Y}}_{1,G} = \frac{1}{n_{0,G}}\sum_{x\in s_{0,G}}\hat{Y}_1(x), \hat{\bar{R}}_{1,G} = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}_1(x), \hat{\bar{R}}_G = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}(x)$$

The main difficulty to obtain a better variance estimate is that the residual third term in eq. 18 is no longer zero in general. The most elegant way to bypass this difficulty is by extending the model with the indicator variable $I_G(x)$ of the small area G. As the procedure is discussed in details in Mandallaz et al. (2013) and Mandallaz (2013*b*), we can here briefly sketch the procedure.

We have the extended models with auxiliary vectors: $\mathcal{Z}^t(x) = (\mathcal{Z}^{(1)t}(x), \mathcal{Z}^{(2)t}(x))$, where $\mathcal{Z}^{(1)t}(x) = (Z^{(1)t}(x), I^t_G(x))$ and $\mathcal{Z}^{(2)t}(x) = Z^{(2)t}(x)$.

1. The large extended model M

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

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.

2. The reduced extended model M_1

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

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

$$\hat{\boldsymbol{\gamma}}_2 = \left(\frac{1}{n_2}\sum_{x\in s_2}\mathcal{Z}^{(1)}(x)\mathcal{Z}^{(1)t}(x)\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}^{(1)}(x) := \left(\mathcal{A}_2^{(1)}\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}^{(1)}(x) \tag{20}$$

and

$$\hat{\boldsymbol{\theta}}_2 = \left(\frac{1}{n_2}\sum_{x\in s_2}\mathcal{Z}(x)\mathcal{Z}^t(x)\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}(x) := \mathcal{A}_2^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\mathcal{Z}(x) \tag{21}$$

The estimated covariance matrices are

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_2} = \mathcal{A}_2^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{\mathcal{R}}^2(x)\mathcal{Z}(x)\mathcal{Z}^t(x)\right)\mathcal{A}_2^{-1}$$
$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\gamma}}_2} = \left(\mathcal{A}_1^{(1)}\right)^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{\mathcal{R}}_1^2(x)\mathcal{Z}^{(1)}(x)\mathcal{Z}^{(1)t}(x)\right)\left(\mathcal{A}_1^{(1)}\right)^{-1} \tag{22}$$

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 $\hat{\bar{Y}}_{G,\mathrm{g3reg}}$ as in eq. 11, i.e.,

$$\hat{\bar{Y}}_{G,\mathrm{g3reg}} = \left(\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G} - \hat{\bar{\mathcal{Z}}}^{(1)}_{1,G}\right)^t\hat{\boldsymbol{\gamma}}_2 + \hat{\bar{\mathcal{Z}}}^t_{1,G}\hat{\boldsymbol{\theta}}_2 \tag{23}$$

where we have set

$$\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G} = \frac{1}{n_{0,G}}\sum_{x\in s_{0,G}}\mathcal{Z}^{(1)}(x),\ \hat{\bar{\mathcal{Z}}}^{(1)}_{1,G} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}\mathcal{Z}^{(1)}(x),\quad \hat{\bar{\mathcal{Z}}}_{1,G} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}\mathcal{Z}(x)$$

To get an estimate of the design-based variance, we use mutatis mutandis eq. 14 and obtain

$$\hat{\mathbb{V}}_{0,1,2}(\hat{\bar{Y}}_{G,\mathrm{g3reg}}) = \hat{\boldsymbol{\gamma}}_2^t\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G}}\hat{\boldsymbol{\gamma}}_2 + \frac{n_2}{n_1}\hat{\bar{\mathcal{Z}}}^{(1)t}_{0,G}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\gamma}}_2}\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G} + \left(1 - \frac{n_2}{n_1}\right)\hat{\bar{\mathcal{Z}}}^t_{1,G}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_2}\hat{\bar{\mathcal{Z}}}^t_{1,G} \tag{24}$$

where

$$\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G}} = \frac{1}{n_{0,G}}\frac{1}{n_{0,G}-1}\sum_{x\in s_{0,G}}\left(\mathcal{Z}^{(1)}(x) - \hat{\bar{\mathcal{Z}}}^{(1)}_{0,G}\right)\left(\mathcal{Z}^{(1)}(x) - \hat{\bar{\mathcal{Z}}}^{(1)}_{0,G}\right)^t \tag{25}$$

is the empirical covariance matrix of $\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G}$ over G.

The first term in eq. 24 can be rewritten as the empirical variance of the predictions $\hat{\bar{Y}}_1(x) = \mathcal{Z}^{(1)t}\hat{\boldsymbol{\gamma}}_2$, that is

$$\hat{\boldsymbol{\gamma}}_2^t\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathcal{Z}}}^{(1)}_{0,G}}\hat{\boldsymbol{\gamma}}_2 = \frac{1}{n_{0,G}-1}\frac{1}{n_{0,G}}\sum_{x\in s_{0,G}}(\hat{\bar{Y}}_1(x) - \hat{\bar{\bar{Y}}}_{1,G})^2$$

with $\hat{\bar{\bar{Y}}}_{1,G} = \frac{1}{n_{0,G}}\sum_{x\in s_{0,G}}\hat{\bar{Y}}_1(x)$.

It is shown in Mandallaz (2013*b*) that the external variance estimate eq. 19 and the g-weight variance estimate eq. 24 are asymptotically equivalent, which is not obvious at first glance.

For a very small area H, the number of points $n_{2,H}$ may be to small or even zero. In such a case, one could imbed H in a somewhat larger area $G \supset H$ with $n_{2,G}$ sufficiently large and consider the synthetic estimator in the extended model with respect to G (see Mandallaz (2013*b*) for details).

As shown in Mandallaz et al. (2013), it is straightforward to consider simultaneously a limited number of small areas by extending the model with as many small-area indicator variables. For a large number, say S, of small areas, one should rather use S different models each extended by the corresponding indicator variable (to avoid singular or near singular matrices) or rely on the external model approach for areas with $n_{2,G}$ sufficiently large (say at least 3 to 4). In any case, with a large number S of confidence intervals, the problem of overall coverage probability (essentially

the multiple testing issue) arises. The synthetic estimators will of course have shorter 95% confidence intervals that could be adjusted by Bonferonni's inequality (using, e.g., 99% level for 10 small areas will ensure that the overall coverage probability is at least 90%), but at the cost of potential biases. In mining, the double kriging procedure is frequently used to produce two maps: one for the small-area point estimates and one for the corresponding Kriging errors, but the problem of overall coverage probability remains, independently of the technique used.

6. Generalization to cluster sampling and two-stage sampling

The generalization to cluster sampling (widely used in extensive forest inventories) and to two-stage Poisson sampling of trees at the plot level (as used in the Swiss National Inventory to obtain more accurate estimates of timber volume) is straightforward; the reader can consult Mandallaz (2013*a*) for an overview and Mandallaz (2013*b*) for full details. Qualitatively, the same results hold, the increase of variance is equal to the variance of the predictions from the reduced model over the small area considered, and the second-stage variance is automatically accounted for.

7. Examples

To illustrate the theory and check empirically the validity of the various mathematical approximations used to derive the variance estimates, we present simulations performed on a purely artificial example already discussed in Mandallaz (2013*a*) and Mandallaz et al. (2013).

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 $\boldsymbol{Z}(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)^t \in \mathbb{R}^6$. 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

$$Y(x) = Y_0(x) + R(x) \tag{26}$$

where $Y_0(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0$ and $R(x) = 6\cos(\pi x_1)\sin(2\pi x_2)$. We have chosen $\boldsymbol{Z}^{(1)}(x) = (1, x_1, x_2)^t$ and $\boldsymbol{Z}^{(2)}(x) = (x_1^2, x_1x_2, x_2^2)^t$. The exact least square regression coefficient vector for the reduced model is found to be $\boldsymbol{\alpha}_0 = (25.17, 9.53, 3.00)^t$. Note that the large model is, in this case, the true model (which is, of course, never the case in practice!) and that the components of $\boldsymbol{\alpha}_0$ of the reduced model are different from the first three components of $\boldsymbol{\beta}_0$. The coefficient of determination is $R^2 = 0.82$ for the full model $\boldsymbol{Z}(x)$ and $R^2 = 0.72$ for the reduced model $\boldsymbol{Z}^{(1)}(x)$. 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 $\hat{Y}_{G,\mathrm{g3reg}}$ and $\hat{\mathbb{V}}_{\mathrm{ext}}$ given in eqs. 18 and 19. Tables 1 and 2 summarize the simulation results, which complete those given in Mandallaz (2013*a*) and Mandallaz et al. (2013).

All of the simulations were performed with the linear algebra procedure proc iml of the statistical software package SAS (version 9.2; www.sas.com), and the software Maple (version 14; www.maplesoft.com) was used to calculate the true values given by integrals.

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

	n_0:n_1:n_2		
	400:100:25	800:200:50	1600:400:100
$\mathbb{E}^*(\hat{Y}_{F,\mathrm{g3reg}})$	39.17	39.16	39.17
$\mathbb{V}^*(\hat{Y}_{F,\mathrm{g3reg}})$	0.63	0.27	0.13
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{g3reg}}))$	0.45	0.24	0.13
$\mathbb{E}^*(\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{F,\mathrm{g3reg}}))$	0.43	0.23	0.12
$\hat{P}$	90.9	93.6	94.4
$\hat{P}_{\mathrm{ext}}$	91.4	93.4	94.3
Δ_F^2	1.30	1.25	1.23

Note: The true value is $\overline{Y}_F = 39.17$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ denote the empirical means and variances, respectively, based on 20 000 runs. $\hat{P}$ is the empirical coverage probability (in %) of the 95% confidence intervals based on the g-weight variance and the Student distribution with $n_2 - p$ degrees of freedom. $\hat{\mathbb{V}}_{\mathrm{ext}}(\cdot)$ and $\hat{P}_{\mathrm{ext}}$ refer to the calculations based on the external variance. $\Delta_F^2 = \mathbb{E}^*\left(\frac{\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{g3reg}})}{\hat{\mathbb{V}}(\hat{Y}_{F,\mathrm{greg}})}\right)$ reflects the increase of variance due to three-phase sampling.

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

	n_0:n_1:n_2		
	400:100:25	800:200:50	1600:400:100
$\mathbb{E}^*(\hat{\tilde{Y}}_{G,\mathrm{g3reg}})$	37.16	37.15	37.16
$\mathbb{V}^*(\hat{\tilde{Y}}_{G,\mathrm{g3reg}})$	1.89	0.86	0.41
$\mathbb{E}^*(\hat{\mathbb{V}}(\hat{\tilde{Y}}_{G,\mathrm{g3reg}}))$	1.40	0.76	0.38
$\hat{P}$	94.2	93.8	94.5
$\mathbb{E}^*(\hat{Y}_{G,\mathrm{g3reg}})$	37.13	37.14	37.16
$\mathbb{V}^*(\hat{Y}_{G,\mathrm{g3reg}}))$	1.82	0.82	0.41
$\mathbb{E}^*(\hat{\mathbb{V}}_{\mathrm{ext}}(\hat{Y}_{G,\mathrm{g3reg}}))$	1.71	0.82	0.40
$\hat{P}_{\mathrm{ext}}$	94.8	94.4	94.7
Δ_G^2	1.12	1.08	1.07

Note: The true value is $\overline{Y}_G = 37.16$. $\mathbb{E}^*(\cdot)$ and $\mathbb{V}^*(\cdot)$ denote the empirical means and variances, respectively, based on 20 000 runs. $\hat{P}$ and $\hat{P}_{\mathrm{ext}}$ are the empirical coverage probabilities (in %) of the 95% confidence intervals based on the Student distribution with $n_{2,G} - 1$ degrees of freedom. $\Delta_G^2 = \mathbb{E}^*\left(\frac{\hat{\mathbb{V}}(\hat{Y}_{G,\mathrm{g3reg}})}{\hat{\mathbb{V}}(\hat{Y}_{G,\mathrm{greg}})}\right)$ reflects the increase of variance due to three-phase sampling.

Discussion

1. All point estimates are practically unbiased (even if the bias can be statistically significant due to the huge sample size of 20 000 runs).
2. The empirical variances are in good agreement with their estimated counterparts, particularly for $n_2 \geq 50$.
3. The confidence limits for F based on the Student distribution with $n_2 - p$ degrees of freedom (df) are to small, particularly for $n_2 = 25$. One could replace n_2^2 by $n_2(n_2 - p)$ in the middle terms of eq. 6, as suggested by ordinary least squares theory, to achieve the required 95% level by using the Student distribution on $n_2 - p$ df. In most practical instances, n_2 will be so large that one can calculate the 95% confidence intervals according to the normal distribution.
4. The empirical coverage probabilities for the small area G, based on the Student distribution with $n_{2,G} - 1$ df, are very close to the nominal value. The g-weight based confidence intervals are shorter than those based on the external variance.
5. The simulation confirms the asymptotic calculations of the variances.
6. The asymptotic increase of variance due to three-phase sampling

$$\Delta_F^2 = \frac{\mathbb{V}(\hat{Y}_{F,\mathrm{g3reg}})}{\mathbb{V}(\hat{Y}_{F,\mathrm{greg}})}$$

can be calculated by using eq. 13. One obtains $\Delta_F^2 = 1.22$ for n_0:n_1:n_2 = 16:4:1 as chosen for the simulation. In this example, the null phase accounts for 18% of the variance, the first phase accounts for 28%, and the second-phase accounts for 54%.
7. The increase of variance due to three-phase sampling is smaller and almost negligible from a practical point of view

for small-area estimation. The corresponding asymptotic value of $\Delta_G^2 = \frac{\mathbb{V}(\hat{Y}_{G,\text{g3reg}})}{\mathbb{V}(\hat{Y}_{G,\text{greg}})}$ is 1.07.

Very similar results were obtained by reanalyzing the case study presented in Mandallaz et al. (2013): the null phase now being a 125 m × 125 m grid instead of wall-to-wall, the first-phase being a 250 m × 250 m grid, and the second phase being a 500 m × 500 m grid. The reader will find the description of the two components $\mathbf{Z}^{(1)}(x)$ and $\mathbf{Z}^{(2)}(x)$, as well as the model used in the aforementioned reference. As expected, the three-phase point estimates do not differ significantly from their partially exhaustive counterparts. For instance, the increase in variance for the combined estimators (i.e., including the indicator variables of the four small areas) were as follows: (1) Δ_F^2 = 1.08 and Δ_G^2 = 1.11, 1.09, 1.05, and 1.04 for the small areas G_1, G_2, G_3, and G_4, respectively, when the only exhaustive component is the mean canopy height $Z_1(x)$; (2) Δ_F^2 = 1.10 and Δ_G^2 = 1.14, 1.12, 1.06, and 1.05, with the four exhaustive components $Z_k(x)$, k = 1, 2, 3, 4 (table 7 in Mandallaz et al. 2013). In the first case, the true mean $\overline{Z}_1$ is not computationally prohibitive because it can be obtained, up to negligible boundary effects, in a "one-shot" procedure. In the second case, the computing time can be reduced from approximately 2 days to less than 2 h with the available PC (Dell E6420 with an Intel i5-2520M processor, 2.5 GHz, 4 GB RAM, 64-bit operating system).

8. Conclusions

If the wall-to-wall component of the auxiliary information is computationally prohibitive, the three-phase generalized regression estimator $\hat{Y}_{F,\text{g3reg}}$ (and its small-area version $\hat{Y}_{G,\text{g3reg}}$) is an excellent alternative that can drastically reduce the computing time at the cost of a small increase in the variance as compared with the partially exhaustive estimators $\hat{Y}_{\text{greg}}$. The procedures based on the external model assumption can be directly implemented with standard statistical procedures available in the software packages R (www.r-project.org) or SAS, whereas their g-weight counterparts, which have slightly better performances for small sample sizes, can be easily implemented with standard linear algebra facilities. The new three-phase regression estimator is clearly better that the classical two-phase regression estimator. The new estimator is of great potential usefulness in conjunction with very large LiDAR data sets. Work near completion shows that the new estimator has also good performances in the continuous forest inventory context.

Acknowledgements

I express my thanks to Prof. H. Heinimann (Chair of Land Use Engineering, ETH Zurich) for his support, as well as to the two reviewers and to the Associate Editor for their corrections, comments, and suggestions, which contributed to improve the original manuscript.

References

Fattorini, L., Marcheselli, M., and Pisani, C. 2006. A three-phase sampling strategy for large-scale multiresource forest inventories. J. Agric. Biol. Environ. Stat. **11**: 296–316. doi:10.1198/108571106X130548.

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

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 three-phase sampling and two multivariate components of auxiliary information. Technical report, ETH Zurich, Department of Environmental Systems Science. Available from http://e-collection.library.ethz.ch.

Mandallaz, D., Breschan, J., and Hill, A. 2013. New regression estimators in forest inventory with two-phase sampling and partially exhaustive information: a design-based Monte Carlo approach with applications to small-area estimation. Can. J. For. Res. **43**(11): 1023–1031. doi:10.1139/cjfr-2013-0181.

von Lüpke, N., Hansen, J., and Saborowski, J. 2012. A three-phase sampling procedure for continuous forest inventory with partial re-measurement and updating of terrestrial sample plots. Eur. J. For. Res. **131**: 1979–1990. doi:10.1007/s10342-012-0648-z.