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

Mandallaz Daniel

Department of Environmental Systems Science, ETH Zurich, CHN K74.1 CH 8092 Zurich, daniel.mandallaz@env.ethz.ch

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.

Key words: three-phase sampling, regression estimators, forest inventory, design-based Monte-Carlo approach

1 Introduction

The motivation for this work is due to the increasing need of using national or regional inventories for local estimation in order to meet tighter budgetary constraints, which is only feasible under extensive use of auxiliary information, provided e.g. by remote sensing (aerial photographs or LiDAR data). The small-area estimation problem is in this context of the utmost importance. The present paper is a terse summary of the results presented in Mandallaz (2014), which extends the so-called generalized regression estimator proposed in Mandallaz et al. (2013) to the case where 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 sub-sample of the null-phase, the first-phase, and the terrestrial inventory is performed on a sub-sample of the first phase, the second-phase. This set-up is particularly useful for national or regional inventories for two reasons: (1) the first component may not be available exhaustively (2) 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 three-phase estimators has also great potential in continuous forest inventory and has been successfully implemented in the Swiss National Inventory (**SNI**), which has moved from a periodic (every 10 years) to an annual survey (see the recent paper Massey et al. (2014) for details).

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 (2008, 2013a) for a first perusal and more bibliographical references. The interested reader will find the proofs of the results and further developments in the on-line technical report Mandallaz (2013b). Parts of the results (valid under the so-called external model assumption) have a very intuitive background and they can be easily implemented with standard statistical software packages while their performances are close to the g-weight procedures presented here (for which the R program maSAE is available from cran.r-project.org/web/packages/). Also, it is worth mentioning that at the present time there are no simple alternative **model-dependent** techniques for this three-phase set-up (relying e.g. on Kriging or mixed models).

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 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 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 $n_1 << n_0$ 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 $\boldsymbol{Z}^t(x) = (\boldsymbol{Z}^{(1)t}(x), \boldsymbol{Z}^{(2)t}(x)) \in \Re^{p+q}$ (the upper index 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.

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 (such as mean canopy height and eventually quantiles thereof), whereas $\boldsymbol{Z}^{(2)}(x) \in \Re^p$ 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 reason for introducing the null-phase sample s_0 is that the component $\boldsymbol{Z}^{(1)}(x)$ can be computationally prohibitive to calculate exhaustively in extensive forest inventories (see Mandallaz (2013a) for a case study with LiDAR data). In the afore mentioned 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 $\bar{Y} = \frac{1}{\lambda(F)} \sum_{i=1}^{N} Y_i$, 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 $I_i(x)$ 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 $Y(x)$

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

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 i-th tree. In the infinite population or Monte Carlo approach one samples the function $Y(x)$ and we have $\mathbb{E}_x(Y(x)) = \frac{1}{\lambda(F)} \int_F Y(x)dx = \frac{1}{\lambda(F)} \sum_{i=1}^{N} Y_i = \bar{Y}$, where $\mathbb{E}_x$ denotes the expectation with respect to a random point x uniformly distributed in F. 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 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

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

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

2. The reduced model M_1

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

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

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

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_k &= \Big(\frac{1}{n_k}\sum_{x\in s_k} \boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)t}(x)\Big)^{-1}\frac{1}{n_k}\sum_{x\in s_k} Y(x)\boldsymbol{Z}^{(1)}(x) \\ &:= (\boldsymbol{A}_k^{(1)})^{-1}\frac{1}{n_k}\sum_{x\in s_k} Y(x)\boldsymbol{Z}^{(1)}(x) =: (\boldsymbol{A}_k^{(1)})^{-1}\boldsymbol{U}_k^{(1)}, \quad k = 0,1,2 \end{aligned} \tag{4}$$

Likewise for the large large model we set

$$\begin{aligned} \hat{\boldsymbol{\beta}}_k &= \Big(\frac{1}{n_k}\sum_{x\in s_k} \boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\Big)^{-1}\frac{1}{n_k}\sum_{x\in s_k} Y(x)\boldsymbol{Z}(x) \\ &= \boldsymbol{A}_k^{-1}\frac{1}{n_k}\sum_{x\in s_k} Y(x)\boldsymbol{Z}(x) =: \boldsymbol{A}_k^{-1}\boldsymbol{U}_k, \quad k = 0,1,2 \end{aligned} \tag{5}$$

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 $\hat{Y}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_2$ and $\hat{Y}_1(x) = \boldsymbol{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}_2$.
Consistent estimates of the design-based covariance matrices are given by

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

with the empirical residuals $\hat{R}(x) = Y(x) - \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_2$ and $\hat{R}_1(x) = Y(x) - \boldsymbol{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}_2$. It is interesting to note that the above design-based covariance matrices are algebraically equivalent to the robust model-dependent covariance matrices discussed by Huber (1967). We define the three-phase generalized regression estimator as

$$\begin{aligned}
\hat{Y}_{F,g3reg} &= \frac{1}{n_0}\sum_{x\in s_0}\hat{Y}_1(x) + \frac{1}{n_1}\sum_{x\in s_1}(\hat{Y}(x) - \hat{Y}_1(x)) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x) - \hat{Y}(x)) \\
&= (\hat{\bar{\boldsymbol{Z}}}_0^{(1)} - \hat{\bar{\boldsymbol{Z}}}_1^{(1)})^t\hat{\boldsymbol{\alpha}}_2 + \hat{\bar{\boldsymbol{Z}}}_1^t\hat{\boldsymbol{\beta}}_2 \qquad (7)
\end{aligned}$$

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

It can be shown that the asymptotic design-based covariance matrices of $\hat{\mathbb{V}}_{0,1,2}(\hat{Y}_{F,g3reg})$ can be consistently estimated by

$$\hat{\mathbb{V}}_{0,1,2}(\hat{Y}_{F,g3reg}) = \hat{\boldsymbol{\alpha}}_2^t\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_0^{(1)}}\hat{\boldsymbol{\alpha}}_2 + \frac{n_2}{n_1}\hat{\bar{\boldsymbol{Z}}}_0^{(1)t}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_2}\hat{\bar{\boldsymbol{Z}}}_0^{(1)} + (1 - \frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_1^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_2}\hat{\bar{\boldsymbol{Z}}}_1^t \qquad (8)$$

where

$$\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_0^{(1)}} = \frac{1}{n_0}\frac{\sum_{x\in s_0}(\boldsymbol{Z}^{(1)}(x) - \hat{\bar{\boldsymbol{Z}}}_0^{(1)})(\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}}_0^{(1)})^t}{n_0 - 1} \qquad (9)$$

One can rewrite [8] in the celebrated g-weight form, computationally more suitable for practical implementation (see Mandallaz (2013b) for details) and which enjoys several nice statistical calibration properties.

It can be shown that[8] is asymptotically equivalent to the following so-called external variance estimate

$$\begin{aligned}
\hat{\mathbb{V}}_{ext}(\hat{Y}_{F,g3reg}) &= \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}_1(x) - \hat{\bar{Y}}_1)^2}{n_0 - 1} \\
&+ \frac{1}{n_1}\frac{1}{n_2 - 1}\sum_{x\in s_2}(\hat{R}_1(x) - \hat{\bar{R}}_1)^2 + (1 - \frac{n_2}{n_1})\frac{1}{n_2(n_2 - 1)}\sum_{x\in s_2}(\hat{R}(x) - \hat{\bar{R}})^2 \qquad (10)
\end{aligned}$$

where $\hat{\bar{Y}}_1 = \frac{1}{n_0}\sum_{x\in s_0}\hat{Y}_1(x)$, $\hat{\bar{R}}_1 = \frac{1}{n_2}\sum_{x\in s_2}\hat{R}_1(x) = 0$ and $\hat{\bar{R}} = \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x) = 0$.

For the limit case $n_0 = \infty$ one obtains the formulae discussed in Mandallaz et al. (2013).

5 Further results

The previous results can be easily adapted to the estimation problem for a small area $G \subset F$. One simply extends the reduced model with the indicator variable $I_G(x)$ of the small area G which ensures also zero mean residuals over G. Details are given in Mandallaz et al. (2013). Most national inventories rely on cluster sampling to reduce costs. Again, all the previous results can be extended to cluster sampling, the formulae are algebraically a bit more cumbersome due to the random number of plots per cluster falling in the forested area, see Mandallaz (2013b) for details. Finally, essentially the same results also hold if two-stage sampling is used at the plot level (as in the **SNI**, where a more accurate timber volume is obtained via further tree diameter and height measurements). One simply replace the local density $Y(x)$ by the so-called generalized local density $Y^*(x)$ (see Mandallaz and Massey (2012) for details).

References

Huber, P. J. (1967). The behaviour of maximum likelihood estimates under non-standard conditions. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistic*, **1**:pp. 221–233.

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Chapman and Hall, Boca Raton FL.

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

Mandallaz, D. (2013b). 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, http://e-collection.library.ethz.ch.

Mandallaz, D. (2014). A three-phase sampling extension of the generalized regression estimator with partially exhaustive information. *Can. J. For. Res.*, **44**:pp. 383–388.

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**:pp. 1023–1031.

Mandallaz, D. and Massey, A. (2012). Comparison of Estimators in One-Phase Two-Stage Poisson Sampling in Forest Inventories. *Can. J. For. Res.*, **42**:pp. 1865–1871.

Massey, A., Mandallaz, D., and Lanz, A. (2014). Integrating remote sensing and past inventoty data under the new annual design of the SWiss National Forest Inventory using three-phase design-based regression estimator. *Can. J. For. Res.*, p. accepted for publication.