Report

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

Author(s):
Mandallaz, Daniel

Publication Date:
2012

Permanent Link:

https://doi.org/10.3929/ethz-a-007318974 →

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

Daniel Mandallaz [1]

Chair of Land Use Engineering

ETH Zurich

CH 8092 Zurich, Switzerland

[1]Tel. ++41(0)44 6323186 e-mail daniel.mandallaz@env.ethz.ch

Abstract

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

Résumé

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

1 Introduction

There is an extensive literature on the problem of small area estimation (or small domain estimation in general sampling). In this paper we shall investigate the properties of some estimators in the **model-assisted framework**, in which prediction models are used to improve the efficiency but are not assumed to be correct as in the **model-dependent approach**. The validity of the statistical procedures is ensured by the randomization principle: i.e. we are in the **design-based** inference framework, which has a definite advantage in official statistics. The reader is referred to (Koehl et al. (2006), section 3.8) for a good review of small-area estimation in forest inventory that presents alternative techniques, in particular Bayesian. Let us now define the sampling scheme.

The **first phase** draws a large sample s_1 of n_1 points that are independently and uniformly distributed within the forest area F. At each point $x \in s_1$ auxiliary information is collected, very often coding information of 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 $\mathbf{Z}(x) \in \Re^p$.

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

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

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

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

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

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

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

The term $\frac{1}{\lambda(F)\pi_i}$ is the tree extrapolation factor f_i with dimension ha^{-1}. One must include 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 $Y(x)$ (Mandallaz (2008)) for which the following important relation holds:

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

Where $\mathbb{E}_x$ denotes the expectation with respect to a random point x uniformly distributed in F. This establishes the link between the infinite population (continuum) $\{x \in F \mid Y(x)\}$ and the finite population of trees $\{i = 1, 2 \ldots N \mid Y_i\}$.

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

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

The afore mentioned randomization principle assume that we have uniformly independently distributed points or clusters in the forested area F, whereas in practice systematic grids are used. There is reasonable theoretical and empirical evidence that treating systematic grids as simple random samples is acceptable for point estimation and also for variance estimation (which will be in most instances slightly overestimated) for extensive forest inventories. From a mathematical point of view the only correct, and also most efficient, procedure, is the geostatistical Kriging technique (see Mandallaz (2008), chapter 7 for a brief introduction and further references), which, however, is difficult to use and not uncontroversial in some aspects (e.g. choice of spatial correlation models and stationarity assumptions).

2 The model

We consider the linear model (the upper script on vector or matrices denotes thereafter the transposition operator)

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

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

$$\int_F R^2(x)dx = \int_F (Y(x) - \boldsymbol{Z}^t(x)\boldsymbol{\beta})^2 dx$$

It satisfies the normal equation

$$[7] \qquad \Big(\int_F \boldsymbol{Z}(x)\boldsymbol{Z}^t(x)dx\Big)\boldsymbol{\beta} = \int_F Y(x)\boldsymbol{Z}(x)dx$$

and the orthogonality relationship

$$[8] \qquad \int_F R(x)\boldsymbol{Z}(x)dx = \mathbf{0}$$

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

$$\int_F R(x)dx = 0$$

The important case of stratification amounts to taking $\boldsymbol{Z}^t(x) = (I_{F_1}(x), I_{F_2}(x), \ldots I_{F_L}(x))$, where $F = \cup_{k=1}^{L} F_k$ and $I_{F_k}(x)$ is the zero-one indicator variable of the k-th stratum F_k. We emphasize the fact that in the design-based model-assisted approach the model [6] is not viewed as an adequate description of the complex stochastic process generating the $Y(x)$, but, more pragmatically, simply as a tool to reduce the variance of estimators of $\bar{Y}, \bar{Y}_G$. Of course, ideally, the model should capture qualitatively the main features of the underlying natural phenomenon.

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

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

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

$$\boldsymbol{A}_{s_2}\hat{\boldsymbol{\beta}}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{U}(x) = \boldsymbol{U}_{s_2}$$

where we have set

$$\boldsymbol{A}_{s_2} = \frac{1}{n_2} \sum_{x \in s_2} \boldsymbol{Z}(x) \boldsymbol{Z}^t(x)$$

and

$$\boldsymbol{U}_{s_2} = \frac{1}{n_2} \sum_{x \in s_2} Y(x) \boldsymbol{Z}(x)$$

The theoretical and empirical regression vector parameters are

$$\begin{aligned} \boldsymbol{\beta} &= \boldsymbol{A}^{-1} \boldsymbol{U} \\ [9] \qquad \hat{\boldsymbol{\beta}}_{s_2} &= \boldsymbol{A}_{s_2}^{-1} \boldsymbol{U}_{s_2} \end{aligned}$$

$\hat{\boldsymbol{\beta}}_{s_2}$ is asymptotically design-unbiased for $\boldsymbol{\beta}$. To calculate the design-based variance-covariance matrix of the regression coefficients we need

$$\mathbb{E}(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta})^t$$

we shall use the Taylor linearization technique. Let us consider the function $f(\cdot, \cdot)$ of an arbitrary (p, p) matrix $\boldsymbol{A}$ and an arbitrary $(p, 1)$ vector $\boldsymbol{U}$ defined by $f(\boldsymbol{A}, \boldsymbol{U}) = \boldsymbol{A}^{-1}\boldsymbol{U}$. We can write

$$\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta} = f(\boldsymbol{A}_{s_2}, \boldsymbol{U}_{s_2}) - f(\boldsymbol{A}, \boldsymbol{U})$$

which can be viewed as the differential of the function $f()$ at the point $\boldsymbol{P}_0 = (\boldsymbol{A}, \boldsymbol{U})$, which is the expected value of the random point $\boldsymbol{P}_{s_2} = (\boldsymbol{A}_{s_2}, \boldsymbol{U}_{s_2})$. The distances between the fixed and the random point are of the order $n_2^{-\frac{1}{2}}$ in design-probability (by the law of large numbers for $\boldsymbol{U}_{s_2}$ and $\boldsymbol{A}_{s_2}$ and the continuity of the inverse operation). The differential of $f(\cdot, \cdot)$ at $\boldsymbol{P}_0$ is, by the derivation rule for product

$$df = d(\boldsymbol{A}^{-1})\boldsymbol{U} + \boldsymbol{A}^{-1} d\boldsymbol{U}$$

Differentiating the identity $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ one gets

$$d(\boldsymbol{A}^{-1}) = -\boldsymbol{A}^{-1}(d\boldsymbol{A})\boldsymbol{A}^{-1}$$

and the following first-order Taylor expansion:

$$\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta} \approx -\boldsymbol{A}^{-1}(\boldsymbol{A}_{s_2} - \boldsymbol{A})\boldsymbol{A}^{-1}\boldsymbol{U} + \boldsymbol{A}^{-1}(\boldsymbol{U}_{s_2} - \boldsymbol{U})$$

Expanding this expression and substituting $\boldsymbol{A}^{-1}\boldsymbol{U} = \boldsymbol{\beta}$ we obtain the Taylor linearization

$$\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta} \approx \boldsymbol{A}^{-1}\Big(- \boldsymbol{A}_{s_2}\boldsymbol{\beta} + \frac{1}{n_2}\sum_{x \in s_2} \boldsymbol{U}(x)\Big)$$

which is, by definition, equal to

$$\boldsymbol{A}^{-1}\Big(- \frac{1}{n_2}\sum_{x \in s_2} \boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\boldsymbol{\beta} + \frac{1}{n_2}\sum_{x \in s_2} Y(x)\boldsymbol{Z}(x)\Big)$$

and consequently also to

$$\boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\sum_{x \in s_2}\big(Y(x) - \boldsymbol{Z}(x)^t\boldsymbol{\beta}\big)\boldsymbol{Z}(x)\Big) = \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\sum_{x \in s_2} R(x)\boldsymbol{Z}(x)\Big)$$

Thus, we finally arrive at

$$\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta} \approx \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\sum_{x \in s_2} R(x)\boldsymbol{Z}(x)\Big) \qquad [10]$$

Using [8] and the independence of the $R(x)\boldsymbol{Z}(x)$ one obtains the design-based variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{s_2}$

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}} \approx \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\mathbb{E}_x R^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\Big)\boldsymbol{A}^{-1} \qquad [11]$$

which can be estimated by replacing the theoretical residual $R(x)$ with their empirical counterparts $\hat{R}(x) = Y(x) - \hat{Y}(x)$, with $\hat{Y}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}$, and $\boldsymbol{A}$ with $\boldsymbol{A}_{s_2}$. We then get the **estimated design-based variance-covariance matrix** as

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} := \boldsymbol{A}_{s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\Big)\boldsymbol{A}_{s_2}^{-1} \quad [12]$$

Interestingly this is precisely the **robust estimate of the model-dependent covariance matrix** given in Gregoire and Dyer (1989) (see also Mandallaz (2008) p. 107). Setting $\hat{\sigma}^2 = \frac{\sum_{x\in s_2}\hat{R}^2(x)}{n_2}$ we get $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \approx \frac{\hat{\sigma^2}}{n_2}\boldsymbol{A}_{s_2}^{-1}$ whereas the model-dependent ordinary least squares theory gives the unbiased estimate of the covariance matrix as $(\frac{n_2}{n_2-p}\hat{\sigma}^2)\frac{1}{n_2}\boldsymbol{A}_{s_2}^{-1}$. The empirical residuals satisfy the sample orthogonality relation

$$\frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x)\boldsymbol{Z}(x) = \boldsymbol{0} \quad [13]$$

Theoretically one may use the exact matrix $\boldsymbol{A}$ if it is available or its estimate $\boldsymbol{A}_{s_1} = \frac{1}{n_1}\sum_{x\in s_1}\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)$ based on the large sample. However, the resulting point estimates are not always intuitively convincing and not optimal in the model-dependent framework. Beside, they are not available from the usual statistical software packages. For these reasons we shall only work with $\boldsymbol{A}_{s_2}$.

3 The estimators

3.1 External models

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

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

with the predictions $\hat{Y}_0(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta_0}$ and the residuals $R_0(x) = Y(x) - \hat{Y}_0(x)$, where $\boldsymbol{\beta_0}$ is the given external regression coefficient, ideally obtained from another similar inventory. Note that in this case the mean residual will not necessarily be zero. To calculate the variance one uses the decomposition

[15] $$\mathbb{V}_{1,2}(\hat{Y}_{reg}) = \mathbb{V}_1\mathbb{E}_{2|1}(\hat{Y}_{reg}) + \mathbb{E}_1\mathbb{V}_{2|1}(\hat{Y}_{reg})$$

to obtain

[16] $$\mathbb{V}(\hat{Y}_{reg}) = \frac{1}{n_1}\mathbb{V}(Y(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R_0(x))$$

which can be unbiasedly estimated with

[17] $$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(Y(x)-\bar{Y}_2)^2 + (1-\frac{n_2}{n_1})\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}(R_0(x)-\bar{R}_{0,2})^2$$

where $\bar{Y}_2 = \frac{1}{n_2}\sum_{x\in s_2} Y(x)$ and $\bar{R}_{0,2} = \frac{1}{n_2}\sum_{x\in s_2} R_0(x)$.

The estimation for any small area $G \subset F$ is straightforward, indeed one simply restricts the samples of n_1 and n_2 points in F to the $n_{1,G}$ and $n_{2,G}$ points in G and apply the above formulae to obtain an unbiased estimate of the conditional variance (i.e. given $n_{1,G}$ and

$n_{2,G}$, which are realizations random variables because in our set-up only n_1 and n_2 are fixed).

3.2 Internal models

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

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

$$[18] \qquad \begin{aligned} \hat{Y}_{G,synth} &= \frac{1}{\lambda(G)} \int_G \hat{Y}_{s_2}(x) dx \\ &= \frac{1}{\lambda(G)} \int_G \boldsymbol{Z}^t(x) \hat{\boldsymbol{\beta}}_{s_2} dx = \bar{\boldsymbol{Z}}_G^t \hat{\boldsymbol{\beta}}_{s_2} \end{aligned}$$

where $\bar{\boldsymbol{Z}}_G = \frac{1}{\lambda(G)} \int_G \boldsymbol{Z}(x) dx$ is the true mean of the auxiliary vector over the small area G, which is available only if the first phase is exhaustive. $\hat{Y}_{G,synt}$ is unbiased under the model, but not optimal as it does not the take the model-dependent spatial correlation of the $Y(x)$ into account. Let us emphasize the fact that the model, i.e. $\hat{\boldsymbol{\beta}}_{s_2}$, is fitted with the full data set and not only with $\{Y(x), \boldsymbol{Z}(x) \mid x \in G\}$.

In this paper we shall investigate the properties of $\hat{Y}_{G,synt}$ in the design-based inference framework.

First, let us note that $\hat{Y}_{G,synt}$ is a design-based consistent sample copy of

$$\frac{1}{\lambda(G)}\int_G \hat{Y}(x)dx = \frac{1}{\lambda(G)}\int_G (Y(x) - R(x))dx = \bar{Y}_G - \frac{1}{\lambda(G)}\int_G R(x)dx$$

Consequently, the synthetic estimator $\hat{Y}_{G,synt}$ has a design-based asymptotic bias equal to $-\frac{1}{\lambda(G)}\int_G R(x)$, which is not zero unless $G = F$ (we have zero mean residual over the entire domain, see [8]) or, which is unlikely, zero mean residual over the small area of interest. Using [18] and [12] **the estimated design-based variance of the synthetic estimator** is

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

We define the g-weights as

$$g_G(x) = \bar{\boldsymbol{Z}}_G^t \boldsymbol{A}_{s_2}^{-1} \boldsymbol{Z}(x) \quad [20]$$

It is easily checked that one can rewrite the point estimate and its estimated variance as

$$\begin{aligned} \hat{Y}_{G,synth} &= \frac{1}{n_2}\sum_{x\in s_2} g_G(x)Y(x) \\ \hat{\mathbb{V}}(\hat{Y}_{G,synth}) &= \frac{1}{n_2^2}\sum_{x\in s_2} g_G^2(x)\hat{R}^2(x) \end{aligned} \quad [21]$$

where the $\hat{R}(x) = Y(x) - \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}$ are the empirical residuals. In the above the special case $G = F$ is possible. The g-weights enjoy several attractive statistical properties (see Särndal et al. (2003) for the aspects in general sampling theory and Mandallaz (2008) for their Monte-Carlo counterparts in forest inventory).

To compensate for the bias due to the non vanishing mean residual over G one considers

the **small-area estimator** (Mandallaz (2008) p.120)

$$[22] \qquad \hat{Y}_{G,small} = \hat{Y}_{G,synt} + \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} \hat{R}(x)$$

where $s_{2,G} = s_2 \cap G$ and $n_{2,G} = \sum_{x \in s_2} I_G(x)$ is the number of points of s_2 falling within G. It can be shown (Mandallaz (2008)) that $\hat{Y}_{G,small}$ is asymptotically design-unbiased with estimated design-based variance given by

$$[23] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,small}) = \frac{1}{n_{2,G}} \frac{1}{n_{2,G}-1} \sum_{x \in s_{2,G}} (\hat{R}(x) - \bar{\hat{R}}_{2,G})^2$$

where

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

is the estimated mean residual over a small area. The above variance estimate neglects the variance of $\hat{\boldsymbol{\beta}}_{s_2}$ and is therefore valid only if n_2 is very large and $n_2 >> n_{2,G}$. To have better insight we use the expansion [10] to obtain

$$[24] \qquad \hat{Y}_{G,small} - \bar{Y}_G = \bar{\boldsymbol{Z}}_G^t \boldsymbol{A}^{-1} \Big(\frac{1}{n_2} \sum_{x \in s_2} R(x) \boldsymbol{Z}(x) \Big) + \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} R(x) - \bar{R}_G$$

which leads to the variance

$$[25] \quad \mathbb{E}(\hat{Y}_{G,small} - \bar{Y}_G)^2 = \bar{\boldsymbol{Z}}_G^t \boldsymbol{A}^{-1} \Big(\frac{1}{n_2} \mathbb{E} R^2(x) \boldsymbol{Z}(x) \boldsymbol{Z}(x)^t \Big) \boldsymbol{A}^{-1} \bar{\boldsymbol{Z}}_G + \mathbb{V}\big(\frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} R(x) \big) + C$$

where the cross-product term C is given by

$$2 \bar{\boldsymbol{Z}}_G^t \boldsymbol{A}^{-1} \mathbb{E}\big((\frac{1}{n_2} \sum_{x \in s_2} R(x) \boldsymbol{Z}(x)) (\frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} R(x) - \bar{R}_G) \big)$$

Both terms of the product tend to zero at rate $(n_2^{-\frac{1}{2}})$, which unfortunately is of the same order as the first two terms. However, using the fact that the $R(x)$, $\boldsymbol{Z}(x)$ are independent of $R(y)$, $\boldsymbol{Z}(y)$ for $x \neq y$ we obtain after tedious but simple calculations

$$C = \frac{1}{n_2} \bar{\boldsymbol{Z}}_G^t \boldsymbol{A}^{-1} \Big(\mathbb{E}_{x \in G} R^2(x) \boldsymbol{Z}(x) - \bar{R}_G \mathbb{E}_{x \in G} R(x) \boldsymbol{Z}(x) \Big)$$

which we can reasonably assume to be negligible. The above arguments suggest therefore the following estimate of the design-based variance of the small-area estimator with exhaustive first phase

$$\begin{aligned} [26] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,small}) &= \bar{\boldsymbol{Z}}_G^t \boldsymbol{A}_{s_2}^{-1} \Big(\frac{1}{n_2^2} \sum_{x \in s_2} \hat{R}^2(x) \boldsymbol{Z}(x) \boldsymbol{Z}(x)^t \Big) \boldsymbol{A}_{s_2}^{-1} \bar{\boldsymbol{Z}}_G \\ &+ \frac{1}{n_{2,G}} \frac{1}{n_{2,G} - 1} \sum_{x \in s_{2,G}} (\hat{R}(x) - \bar{\hat{R}}_{2,G})^2 \end{aligned}$$

Comparing with [17] (after restricting the samples to G) we see that **treating an internal model as an external model** (i.e. ignoring the variability of $\hat{\boldsymbol{\beta}}_{s_2}$) **will underestimate the variance of the small area estimate**. The first term in [26] reflects the uncertainty in the regression coefficients.

If the first-phase is non-exhaustive, i.e. $n_1 \neq \infty$, then one can replace the true mean $\bar{\boldsymbol{Z}}_G$ by its estimate in the large sample

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

where $s_{1,G}$ is the set $s_1 \cap G$ of the $n_{1,G} = \sum_{x \in s_1} I_g(x)$ points of the large sample falling into the small area G. This gives the **pseudo-synthetic estimator**

$$[27] \qquad \hat{Y}_{G,psynth} = \hat{\bar{\boldsymbol{Z}}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} = \frac{1}{n_{1,G}} \sum_{x \in s_{1,G}} \hat{Y}(x)$$

which is clearly asymptotically equivalent to $\hat{Y}_{G,synt}$ as $n_1 \to \infty$ and its design-based expected value tends to $\bar{\boldsymbol{Z}}_G^t\boldsymbol{\beta}$. To calculate the asymptotic variance we use the decomposition (actually the first order Taylor expansion)

$$\Delta = \hat{\bar{\boldsymbol{Z}}}_{1,G}^t\hat{\boldsymbol{\beta}}_{s_2} - \bar{\boldsymbol{Z}}_G^t\boldsymbol{\beta} = \hat{\bar{\boldsymbol{Z}}}_{1,G}^t(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}) + (\hat{\bar{\boldsymbol{Z}}}_{1,G}^t - \bar{\boldsymbol{Z}}_G)^t\boldsymbol{\beta}$$

Asymptotically we get

$$\mathbb{E}\Delta^2 = \bar{\boldsymbol{Z}}_G^t\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}}\bar{\boldsymbol{Z}}_G + \boldsymbol{\beta}^t\boldsymbol{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{1,G}}\boldsymbol{\beta} + 2\mathbb{E}\Big((\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta})^t\hat{\bar{\boldsymbol{Z}}}_{1,G}(\hat{\bar{\boldsymbol{Z}}}_{1,G} - \bar{\boldsymbol{Z}}_G)^t\boldsymbol{\beta}\Big)$$

where $\boldsymbol{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{1,G}}$ is the covariance matrix of $\hat{\bar{\boldsymbol{Z}}}_{1,G}$. The first two terms are of order n_2^{-1} and n_1^{-1} respectively. For the third term we note that $\mathbb{E}(\hat{\bar{\boldsymbol{Z}}}_{1,G}(\hat{\bar{\boldsymbol{Z}}}_{1,G} - \bar{\boldsymbol{Z}}_G)^t)$ is equal to the covariance matrix $\boldsymbol{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{1,G}}$ and therefore of order n_1^{-1} and that $\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}$ is of order $n_2^{-\frac{1}{2}}$. The last term is therefore of smaller order than the first two which leads to the following asymptotic design-based estimate of variance

[28] $$\hat{\mathbb{V}}(\hat{Y}_{G,psynth}) := \hat{\bar{\boldsymbol{Z}}}_{1,G}^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{1,G} + \hat{\boldsymbol{\beta}}_{s_2}^t\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{1,G}}\hat{\boldsymbol{\beta}}_{s_2}$$

The variance-covariance matrix of the auxiliary vector $\hat{\bar{\boldsymbol{Z}}}_G$ is estimated by

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

Usually $\hat{Y}_{G,psynt}$ will have a small variance but at the cost of a potential bias. We can rewrite [27] and [28] with the g-weights

[30] $$\begin{aligned} g_{G,1}(x) &= \hat{\bar{\boldsymbol{Z}}}_{1,G}^t\boldsymbol{A}_{s_2}^{-1}\boldsymbol{Z}(x) \\ \hat{Y}_{G,psynth} &= \frac{1}{n_2}\sum_{x\in s_2} g_{G,1}(x)Y(x) \end{aligned}$$

and after some algebra we get

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}_{G,psynth}) &= \frac{1}{n_2^2}\sum_{x\in s_2} g_{G,1}^2(x)\hat{R}^2(x) + \hat{\boldsymbol{\beta}}_{s_2}^t \hat{\Sigma}_{\hat{\boldsymbol{Z}}_{1,G}} \hat{\boldsymbol{\beta}}_{s_2} \\ [31] \qquad &= \frac{1}{n_2^2}\sum_{x\in s_2} g_{G,1}^2(x)\hat{R}^2(x) + \frac{1}{n_{1,G}(n_{1,G}-1)}\sum_{x\in s_{1,G}} (\hat{Y}(x) - \bar{\hat{Y}}_{1,G})^2\end{aligned}$$

where $\bar{\hat{Y}}_{1,G} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}} \hat{Y}(x)$. The second term in the last equation is the variance of the predictions over G.

The **pseudo small-area estimator**

$$[32] \qquad \hat{Y}_{G,psmall} = \hat{Y}_{G,psynth} + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}} \hat{R}(x)$$

is asymptotically design-unbiased and intuitively its variance can be expected to be well approximated by

$$[33] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,psmall}) = \hat{\bar{\boldsymbol{Z}}}_{1,G}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\bar{\boldsymbol{Z}}}_{1,G} + \hat{\boldsymbol{\beta}}_{s_2}^t \hat{\Sigma}_{\hat{\boldsymbol{Z}}_{1,G}} \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_{2,G}} (\hat{R}(x) - \bar{\hat{R}}_{2,G})^2$$

A tedious formal proof can be given by using [15], [25] and [28].

Using the same arguments as in [31] we also have

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}_{G,psmall}) &= \frac{1}{n_2^2}\sum_{x\in s_2} g_{G,1}^2(x)\hat{R}^2(x) + \frac{1}{n_{1,G}(n_{1,G}-1)}\sum_{x\in s_{1,G}} (\hat{Y}(x) - \bar{\hat{Y}}_{1,G})^2 \\ [34] \qquad &+ \frac{1}{n_{2,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_{2,G}} (\hat{R}(x) - \bar{\hat{R}}_{2,G})^2\end{aligned}$$

This should be compared with the external version [17]

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}_{G,psmall}) &= \frac{1}{n_{1,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_{2,G}}(Y(x)-\bar{Y}_{2,G})^2 \\ &+ (1-\frac{n_{2,G}}{n_{1,G}})\frac{1}{n_{2,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_{2,G}}(\hat{R}(x)-\bar{\hat{R}}_{2,G})^2\end{aligned} \quad [35]$$

For very large $n_{1,G}$ is is clear that the external version will underestimate the variance as it neglects the first term in [34], which albeit is also small for large n_2.
The special case $G = F$ deserves special attention: because of the zero mean residual we have $\hat{Y}_{F,psmall} = \hat{Y}_{F,psynt} = \hat{Y}_{reg}$ and [28,34] lead to the estimated variance

$$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_2^2}\sum_{x\in s_2} g_{F,1}^2(x)\hat{R}^2(x) + \frac{1}{n_1(n_1-1)}\sum_{x\in s_1}(\hat{Y}(x)-\bar{\hat{Y}}_1)^2 \quad [36]$$

with $\bar{\hat{Y}}_1 = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}(x)$. The external version [35] gives

$$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(Y(x)-\bar{Y}_2)^2 + (1-\frac{n_2}{n_1})\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}\hat{R}^2(x) \quad [37]$$

Writing $Y(x) = \hat{Y}(x) + \hat{R}(x)$ an using [13] we can rewrite [37] as

$$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(\hat{Y}(x)-\bar{\hat{Y}}_2)^2 + \frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}\hat{R}^2(x) \quad [38]$$

For $G = F$ the $g_{F,s_1}^2(x)$ are asymptotically equal to 1 (see Mandallaz (2008) p. 113 and the properties of the g-weights discussed below) so that both versions [36] and [38] are asymptotically equivalent. However, version [36] estimates the variance of the predictions in the large sample, which is better, and it rests upon the g-weights for the residual part, which it is known to have better conditional properties (see Mandallaz (2008) p. 84 and section 6.1 for the important special case of stratification).

In the next section we present a simple reformulation of the problem that allows one to transform small-area estimators into synthetic estimators, which offers a great mathematical advantage.

3.3 Alternative estimators in extended model

The main difficulty stems from the fact that $\int_G R(x)dx \neq 0$. If we now extend the auxiliary information vector $\boldsymbol{Z}(x)$ to $\boldsymbol{\mathcal{Z}}^t(x) = (\boldsymbol{Z}^t(x), I_G(x)) \in \mathcal{R}^{(p+1)}$, the corresponding model reads

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

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

$$\Big(\int_F \boldsymbol{\mathcal{Z}}(x)\boldsymbol{\mathcal{Z}}^t(x)dx\Big)\boldsymbol{\theta} =: \boldsymbol{\mathcal{A}}\boldsymbol{\theta} = \int_F Y(x)\boldsymbol{\mathcal{Z}}(x)dx$$

and the orthogonality relationship

$$\int_F \mathcal{R}(x)\boldsymbol{\mathcal{Z}}(x)dx = \mathbf{0}$$

Since $I_F(x) \equiv 1$ is the intercept term (or linear combination of the components of $\boldsymbol{Z}(x)$) and $\boldsymbol{\mathcal{Z}}(x)$ contains $I_G(x)$ **we have the two zero mean residual properties**

$$\int_F \mathcal{R}(x)dx = \int_G \mathcal{R}(x)dx = 0$$

Hence, by including the $0, 1$ indicator variable of the small area G into the model, we enforce zero mean residual over F and G. Note also that G must be a proper subset of F, otherwise $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{A}}_{s_2}$ are singular. In practice near-singularity could cause numerical

problems, so that the small area G must indeed be small with respect to F. Simple calculations yield the following block structure for $\boldsymbol{\mathcal{A}}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{\mathcal{Z}}(x)\boldsymbol{\mathcal{Z}}^t(x)$

$$[40] \qquad \boldsymbol{\mathcal{A}}_{s_2} = \begin{bmatrix} \boldsymbol{A}_{s_2} & \hat{p}_{2,G}\hat{\bar{\boldsymbol{Z}}}_{2,G} \\ \hat{p}_{2,G}\hat{\bar{\boldsymbol{Z}}}^t_{2,G} & \hat{p}_{2,G} \end{bmatrix}$$

where we have set $\hat{p}_{2,G} = \frac{n_{2,G}}{n_2}$, $\hat{\bar{\boldsymbol{Z}}}_{2,G} = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\boldsymbol{Z}(x)$. Using formulae for the inversion of partitioned matrices (see e.g. Searle (1971) p. 27 and Tian and Takane (2009) for useful generalizations) one obtains

$$[41] \qquad \boldsymbol{\mathcal{A}}^{-1}_{s_2} = \begin{bmatrix} \boldsymbol{A}^{-1}_{s_2} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} + \frac{1}{\gamma}\begin{bmatrix} \hat{p}^2_{2,G}\boldsymbol{A}^{-1}_{s_2}\hat{\bar{\boldsymbol{Z}}}_{2,G}\hat{\bar{\boldsymbol{Z}}}^t_{2,G}\boldsymbol{A}^{-1}_{s_2} & -\hat{p}_{2,G}\boldsymbol{A}^{-1}_{s_2}\hat{\bar{\boldsymbol{Z}}}_{2,G} \\ -\hat{p}_{2,G}\hat{\bar{\boldsymbol{Z}}}^t_{2,G}\boldsymbol{A}^{-1}_{s_2} & 1 \end{bmatrix}$$

with $\gamma = \hat{p}_{2,G} - \hat{p}^2_{2,G}\hat{\bar{\boldsymbol{Z}}}^t_{2,G}\boldsymbol{A}^{-1}_{s_2}\hat{\bar{\boldsymbol{Z}}}_{2,G}$.

We need

$$\frac{1}{n_2}\sum_{x\in s_2}Y(x)\boldsymbol{\mathcal{Z}}(x) = (\frac{1}{n_2}\sum_{x\in s_2}Y(x)\boldsymbol{Z}^t(x), \hat{p}_{2,G}\hat{\bar{Y}}_G)^t = ((\boldsymbol{A}^{-1}_{s_2}\hat{\boldsymbol{\beta}}_{s_2})^t, \hat{p}_{2,G}\hat{\bar{Y}}_G)^t$$

where $\hat{\bar{Y}}_G = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}Y(x)$. This leads after some algebra to the following relationship between the regressions coefficients

$$[42] \qquad \hat{\boldsymbol{\theta}}_{s_2} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_{s_2} \\ 0 \end{bmatrix} + \frac{1}{\gamma}\begin{bmatrix} -\hat{p}^2_{2,G}(\hat{\bar{Y}}_G - \hat{\bar{\boldsymbol{Z}}}^t_{2,G}\hat{\boldsymbol{\beta}}_{s_2})\boldsymbol{A}^{-1}_{s_2}\hat{\bar{\boldsymbol{Z}}}_{2,G} \\ \hat{p}_{2,G}(\hat{\bar{Y}}_G - \hat{\bar{\boldsymbol{Z}}}^t_{2,G}\hat{\boldsymbol{\beta}}_{s_2}) \end{bmatrix}$$

Note that the term

$$\hat{\bar{Y}}_G - \hat{\bar{\boldsymbol{Z}}}^t_{2,G}\hat{\boldsymbol{\beta}}_{s_2} = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}(Y(x) - \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2})$$

is precisely the mean residual over the small area. Hence, the last component of $\hat{\boldsymbol{\theta}}_{s_2}$ is essentially the residual term. We see that the original regression coefficient $\hat{\boldsymbol{\beta}}_{s_2}$ is corrected in the extended model by the residual term and that the impact of this correction tends to zero as the small area gets smaller with respect to F, a very intuitive result indeed. One obtains a very similar but not identical result by least squares minimization under the constraint of zero mean residual over the small area (see Searle (1971), pp 112-113). In perfect analogy with [12] the estimated covariance matrix is given by

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}} = \boldsymbol{\mathcal{A}}_{s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{\mathcal{R}}^2(x)\boldsymbol{\mathcal{Z}}(x)\boldsymbol{\mathcal{Z}}(x)^t\Big)\boldsymbol{\mathcal{A}}_{s_2}^{-1} \tag{43}$$

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

$$\hat{\bar{Y}}_{G,synth} = \frac{1}{\lambda(G)}\int_G \boldsymbol{\mathcal{Z}}^t(x)\hat{\boldsymbol{\theta}}_{s_2}dx = \bar{\boldsymbol{\mathcal{Z}}}_G^t\hat{\boldsymbol{\theta}}_{s_2} \tag{44}$$

With $\bar{\boldsymbol{\mathcal{Z}}}_G^t = (\bar{\boldsymbol{Z}}_G^t, 1)$ and some algebra one finally obtains

$$\hat{\bar{Y}}_{G,synth} = \bar{\boldsymbol{Z}}_G^t\hat{\boldsymbol{\beta}}_{s_2} + \frac{\alpha}{n_{2,G}}\sum_{x\in s_{2,G}}\big(Y(x) - \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}\big) \tag{45}$$

where we have set

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

Clearly $\hat{\bar{Y}}_{G,synth}$ and $\hat{Y}_{G,small}$ are asymptotically equivalent because α tends to 1 in large samples. Note that $\alpha = 1$ if the sample is exactly balanced, i.e. if $\hat{\bar{\boldsymbol{Z}}}_{2,G} = \bar{\boldsymbol{Z}}_G$.
By using [19] and replacing $\boldsymbol{Z}(x)$ with $\boldsymbol{\mathcal{Z}}(x)$ we obtain at once the asymptotic variance

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_{G,synth}) = \bar{\boldsymbol{\mathcal{Z}}}_G^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}}\bar{\boldsymbol{\mathcal{Z}}}_G \tag{46}$$

One can rewrite $\hat{\tilde{Y}}_{G,synth}$ in terms of g-weights in the extended model as in [20] and [21] with

$$\tilde{g}_G(x) = \bar{\boldsymbol{\mathcal{Z}}}_G^t \boldsymbol{\mathcal{A}}_{s_2}^{-1} \boldsymbol{\mathcal{Z}}(x)$$

$$\begin{aligned} \hat{\tilde{Y}}_{G,synth} &= \frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_G(x) Y(x) \\ \hat{\mathbb{V}}(\hat{\tilde{Y}}_{G,synth}) &= \frac{1}{n_2^2} \sum_{x \in s_2} \tilde{g}_G^2(x) \hat{\mathcal{R}}^2(x) \end{aligned} \quad [47]$$

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

$$\hat{\bar{\boldsymbol{\mathcal{Z}}}}_{1,G} = \frac{1}{n_{1,G}} \sum_{x \in s_{1,G}} \boldsymbol{\mathcal{Z}}(x) \quad [48]$$

to get the pseudo-synthetic estimate in the extended model

$$\hat{\tilde{Y}}_{G,psynth} = \hat{\bar{\boldsymbol{\mathcal{Z}}}}_{1,G}^t \hat{\boldsymbol{\theta}}_{s_2} \quad [49]$$

As in [45] we have

$$\hat{\tilde{Y}}_{G,psynth} = \hat{\bar{\boldsymbol{Z}}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} + \frac{\alpha_1}{n_{2,G}} \sum_{x \in s_{2,G}} \left(Y(x) - \boldsymbol{Z}^t(x) \hat{\boldsymbol{\beta}}_{s_2} \right) \quad [50]$$

where we have set $\alpha_1 = \frac{\hat{p}_{2,G} - \hat{p}_{2,G}^2 \hat{\bar{\boldsymbol{Z}}}_{1,G}^t \boldsymbol{A}_{s_2}^{-1} \hat{\bar{\boldsymbol{Z}}}_{2,G}}{\hat{p}_{2,G} - \hat{p}_{2,G}^2 \hat{\bar{\boldsymbol{Z}}}_{2,G}^t pmbA_{s_2}^{-1} \hat{\bar{\boldsymbol{Z}}}_{2,G}}$.

By [28] we get immediately the following consistent estimate of the design-based variance

$$\hat{\mathbb{V}}(\hat{\tilde{Y}}_{G,psynth}) = \hat{\bar{\boldsymbol{\mathcal{Z}}}}_{1,G}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}} \hat{\bar{\boldsymbol{\mathcal{Z}}}}_{1,G} + \hat{\boldsymbol{\theta}}_{s_2}^t \hat{\Sigma}_{\hat{\bar{\boldsymbol{\mathcal{Z}}}}_{1,G}} \hat{\boldsymbol{\theta}}_{s_2} \quad [51]$$

The variance-covariance matrix of $\hat{\tilde{\boldsymbol{Z}}}_{1,G}$ can be estimated as usual by

$$[52] \qquad \hat{\Sigma}_{\hat{\tilde{\boldsymbol{Z}}}_{1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)} \sum_{x \in s_{1,G}} (\boldsymbol{Z}(x) - \hat{\tilde{\boldsymbol{Z}}}_{1,G})(\boldsymbol{Z}(x) - \hat{\tilde{\boldsymbol{Z}}}_{1,G})^t$$

Again, one can rewrite the above expression with the g-weights $\tilde{g}_{G,1}(x) = \hat{\tilde{\boldsymbol{Z}}}^t_{1,G} \boldsymbol{\mathcal{A}}^{-1}_{s_2} \boldsymbol{Z}(x)$ namely

$$[53] \qquad \hat{\tilde{Y}}_{G,psynth} = \frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{G,1}(x) Y(x)$$

$$[54] \qquad \hat{\mathbb{V}}(\hat{\tilde{Y}}_{G,psynth}) = \frac{1}{n_2^2} \sum_{x \in s_2} \tilde{g}^2_{G,1}(x) \hat{\mathcal{R}}^2(x) + \hat{\boldsymbol{\theta}}^t_{s_2} \hat{\Sigma}_{\hat{\tilde{\boldsymbol{Z}}}_{1,G}} \hat{\boldsymbol{\theta}}_{s_2}$$

Properties of the g-weights:

1. The g-weights enjoy the calibration properties $\frac{1}{n_2} \sum_{x \in s_2} g_{G,1}(x) \boldsymbol{Z}(x) = \hat{\tilde{\boldsymbol{Z}}}_{1,G}$ and $\frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{G,1}(x) \boldsymbol{Z}(x) = \hat{\tilde{\boldsymbol{Z}}}_{1,G}$. The proof is immediate by transposing the equalities and by the very definition of the g-weights.

2. The fact that one can assume the g-weights depend only on the point x and not on the whole sample s_2 when calculating variances is fully justified by the Taylor expansion leading to the robust design-based covariances.

3. By considering formally the trivial constant local density $Y(x) \equiv 1$ and solving the normal equations one sees that $\frac{1}{n_2} \sum_{x \in s_2} g_{G,1}(x) = \frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{G,1}(x) = 1$, i.e. the g-weights have means equal to 1.

4. When $G = F$ the estimator $\hat{Y}_{reg}$ is asymptotically equivalent to the sample mean over F, i.e. to $\frac{1}{n_2} \sum_{x \in s_2} Y(x)$. This must hold for an arbitrary density $Y(x)$ and therefore one gets $\lim_{n_2 \to \infty} g_{F,1}(x) = 1 > 0$. This is not true for a proper subset $G \subset F$. In this case, $\hat{\tilde{Y}}_{G,synt}$ is asymptotically equivalent to the sample mean over

the small domain, i.e. to $\hat{\tilde{Y}}_G$. Hence, $\tilde{g}_{G,1}(x)$ will tend to 0 for $x \notin G$ (negative values are possible) and to $\frac{n_2}{n_{2,G}}$ for $x \in G$.

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

4 Generalization to cluster sampling

We follow the description of cluster sampling as define in Mandallaz (2008) (especially section 5.5). A cluster is identified by its origin x, uniformly distributed in $\tilde{F} \supset F$. The geometry of the cluster is given by M vectors $e_1, \dots e_M$ defining the random cluster $x_l = x + e_l$. $M(x) = \sum_{l=1}^{M} I_F(x_l)$ is the random number of points of the cluster falling into the forest area F. We define the local density at the cluster level by $Y_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l) Y(x_l)}{M(x)}$, likewise we set $\boldsymbol{Z}_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l) \boldsymbol{Z}(x_l)}{M(x)}$. The set $\tilde{F}$ above can be mathematically defined as the smallest set $\{x \in \mathcal{R}^2 \mid M(x) \neq 0\}$. In the first phase we have n_1 clusters identified by $x \in s_1$ and in the second phase n_2 clusters with $x \in s_2$, obtained by simple random sampling from s_1.

We shall use the model-based approach, in which the regression coefficient $\boldsymbol{\beta}_c$ at the cluster level minimizes

$$\mathbb{E}_{x \in \tilde{F}} M(x)(Y_c(x) - \boldsymbol{\beta}^t \boldsymbol{Z}_c(x))^2$$

In the pure design-based approach the weights will be $M^2(x)$ but this leads to non-zero mean residual (thought close zero in practice), and the definitions of the regression estimator and of the normal equation are slightly different (see Mandallaz (2008), section 5.5 for details). The choice of $M(x)$ rather than $M^2(x)$ as weights is suggested by the model-dependent approach. When $Y_c(x)$ is the mean of the $M(x)$ observations, its variance can be expected to be inversely proportional to $M(x)$. This procedure leads to the normal

equation

$$\Big(\mathbb{E}_{x\in\tilde{F}}M(x)\boldsymbol{Z_c}(x)\boldsymbol{Z_c}(x)^t\Big)\boldsymbol{\beta}_c = \mathbb{E}_{x\in\tilde{F}}M(x)Y_c(x)\boldsymbol{Z_c}(x)$$

and to $\mathbb{E}_{x\in\tilde{F}}M(x)R_c(x)=0$. An asymptotically design-unbiased estimate $\hat{\boldsymbol{\beta}}_{c,s_2}$ for $\boldsymbol{\beta}_c$ can be obtained by taking a sample copy of the above equation, i.e.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{c,s_2} &= \Big(\frac{1}{n_2}\sum_{x\in s_2}M(x)\boldsymbol{Z}_c(x)\boldsymbol{Z}_c^t(x)\Big)^{-1}\Big(\frac{1}{n_2}\sum_{x\in s_2}M(x)Y_c(x)\boldsymbol{Z}_c(x)\Big)\\ &:= \boldsymbol{A}_{c,s_2}^{-1}\Big(\frac{1}{n_2}\sum_{x\in s_2}M(x)Y_c(x)\boldsymbol{Z}_c(x)\Big) \qquad [55]\end{aligned}$$

The empirical residuals at the cluster level are

$$\hat{R}_c(x) = Y_c(x) - \boldsymbol{Z}_c^t(x)\hat{\boldsymbol{\beta}}_{c,s_2}$$

which satisfy the orthogonality relation

$$\sum_{x\in s_2}M(x)\hat{R}_c(x)\boldsymbol{Z}_c(x) = 0$$

and in particular the zero mean residual property

$$\frac{\sum_{x\in s_2}M(x)\hat{R}_c(x)}{\sum_{x\in s_2}M(x)} = 0$$

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

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}} = \boldsymbol{A}_{c,s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}M^2(x)\hat{R}_c^2(x)\boldsymbol{Z}_c(x)\boldsymbol{Z}_c^t(x)\Big)\boldsymbol{A}_{c,s_2}^{-1} \qquad [56]$$

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

$$\hat{\bar{\boldsymbol{Z}}}_{c,1,G} = \frac{\sum_{x \in s_{1,G}} M(x) \boldsymbol{Z}_c(x)}{\sum_{x \in S_{1,G}} M(x)}$$

with estimated covariance matrix

$$[57] \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_{c,1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)} \sum_{x \in s_{1,G}} \left(\frac{M(x)}{\bar{M}_{1,G}}\right)^2 (\boldsymbol{Z}_c(x) - \hat{\bar{\boldsymbol{Z}}}_{c,1,G})(\boldsymbol{Z}_c(x) - \hat{\bar{\boldsymbol{Z}}}_{c,1,G})^t$$

according to (Mandallaz (2008), section 4.3). The **pseudo-synthetic estimate** is then

$$[58] \qquad \begin{aligned} \hat{Y}_{c,G,psynth} &= \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G} \hat{\boldsymbol{\beta}}_{c,s_2} \\ &= \frac{1}{n_2} \sum_{x \in s_2} g_{c,1,G}(x) Y_c(x) \end{aligned}$$

with the g-weights $g_{c,1,G}(x) = \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G} \boldsymbol{A}^{-1}_{c,s_2} M(x) \boldsymbol{Z}_c(x)$. The estimated variance is as in [28]

$$[59] \qquad \hat{\mathbb{V}}(\hat{Y}_{c,G,psynth}) = \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}} \hat{\bar{\boldsymbol{Z}}}_{c,1,G} + \hat{\boldsymbol{\beta}}^t_{c,s_2} \hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,1,G}} \hat{\boldsymbol{\beta}}_{c,s_2}$$

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

$$[60] \qquad \hat{Y}_{c,G,psmall} = \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}} + \frac{\sum_{x \in s_{2,G}} M(x) \hat{R}_c(x)}{\sum_{x \in s_{2,G}} M(x)}$$

It is asymptotically design-unbiased and intuitively its variance can be expected to be approximated by

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}_{c,G,psmall}) &= \hat{\tilde{\boldsymbol{Z}}}^t_{c,1,G}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}\hat{\tilde{\boldsymbol{Z}}}_{c,1,G} + \hat{\boldsymbol{\beta}}^t_{c,s_2}\hat{\Sigma}_{\hat{\tilde{\boldsymbol{Z}}}_{c,1,G}}\hat{\boldsymbol{\beta}}_{c,s_2} \\ &+ \frac{1}{n_{2,g}(n_{2,G}-1)}\sum_{x\in s_{2,G}}\big(\frac{M(x)}{\bar{M}_{2,G}}\big)^2(\hat{R}_c(x)-\bar{\hat{R}}_{2,G})^2 \end{aligned} \quad [61]$$

where $\bar{M}_{2,G} = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}} M(x)$ and $\bar{\hat{R}}_{2,G} = \frac{\sum_{x\in s_{2,G}} M(x)\hat{R}_c(x)}{\sum_{x\in s_{2,G}} M(x)}$. One can rewrite [61] with g-weights and predictions as in [34].

As in simple two-phase sampling we can transform the above estimator into a synthetic estimator by considering the extended model $\boldsymbol{\mathcal{Z}}^t_c(x) = (\boldsymbol{Z}^t_c(x), I_{c,G}(x)) \in \mathcal{R}^{(p+1)}$ with $I_{c,G}(x) = \frac{\sum_{l=1}^M I_G(x_l)}{M(x)}$. In extensive inventories we can reasonably assume that all the points of a cluster lying in the forest area F will belong to the same small area G so that in fact $I_{c,G}(x) \equiv 1$ for all $x \in \tilde{G} = \{x \mid \sum_{l=1}^M I_G(x_l) > 0\}$. The theoretical normal equation reads

$$\Big(\int_{\tilde{F}} M(x)\boldsymbol{\mathcal{Z}}_c(x)\boldsymbol{\mathcal{Z}}^t_c(x)dx\Big)\boldsymbol{\theta}_c =: \boldsymbol{\mathcal{A}}_c\boldsymbol{\theta}_c = \int_{\tilde{F}} M(x)Y_c(x)\boldsymbol{\mathcal{Z}}_c(x)dx \quad [62]$$

which satisfy by construction the two zero mean residuals properties $\int_{\tilde{F}} M(x)\mathcal{R}_c(x)dx = \int_{\tilde{G}} M(x)\mathcal{R}_c(x)dx = 0$. The second equality will only hold approximately if $I_{c,G}(x) < 1$ for some x in $\tilde{G}$.

With $\boldsymbol{\mathcal{A}}_{c,s_2} = \frac{1}{n_2}\sum_{x\in s_2} M(x)\boldsymbol{\mathcal{Z}}_c(x)\boldsymbol{\mathcal{Z}}^t_c(x)$ we obtain the estimate of the regression coefficients at the cluster level

$$\hat{\boldsymbol{\theta}}_{c,s_2} = \boldsymbol{\mathcal{A}}^{-1}_{c,s_2}\Big(\frac{1}{n_2}\sum_{x\in s_2} M(x)\boldsymbol{\mathcal{Z}}_c(x)Y(x)\Big) \quad [63]$$

with estimated design-based covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{c,s_2}} = \boldsymbol{\mathcal{A}}^{-1}_{c,s_2}\Big(\frac{1}{n_2^2}\sum_{\in s_2} M^2(x)\hat{\mathcal{R}}^2_c(x)\boldsymbol{\mathcal{Z}}_c(x)\boldsymbol{\mathcal{Z}}^t_c(x)\Big)\boldsymbol{\mathcal{A}}^{-1}_{c,s_2} \quad [64]$$

where the $\hat{\mathcal{R}}_c(x) = Y_c(x) - \boldsymbol{\mathcal{Z}}_c^t(x)\hat{\boldsymbol{\theta}}_{c,s_2}$ are the empirical residuals at the cluster level with respect to the extended model. We define the pseudo-synthetic estimator in the extended model according to

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

where $\hat{\bar{\boldsymbol{\mathcal{Z}}}}_{c,1,G} = \frac{\sum_{x\in s_{1,G}} M(x)\boldsymbol{\mathcal{Z}}_c(x)}{\sum_{x\in s_{1,G}} M(x)}$ is the mean of the extended auxiliary vector over the small area.

Obviously a decomposition similar to [50] will hold so that $\hat{\bar{Y}}_{c,psynth}$ and $\hat{Y}_{c,G,small}$ in [60,65] are asymptotically equivalent.

As in [59] the estimated variance is given by

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

where

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

Defining the g-weights at the cluster level as

[68] $$\tilde{g}_{G,c,1}(x) = \hat{\bar{\boldsymbol{\mathcal{Z}}}}_{c,1,G}^t \boldsymbol{\mathcal{A}}_{c,s_2}^{-1} M(x)\boldsymbol{\mathcal{Z}}_c(x)$$

we obtain as usual

[69] $$\hat{\bar{Y}}_{c,G,psynth} = \frac{1}{n_2} \sum_{x\in s_2} \tilde{g}_{G,c,1}(x) Y_c(x)$$

and

$$\hat{\mathbb{V}}(\hat{\tilde{Y}}_{c,G,psynth}) = \frac{1}{n_2^2}\sum_{x\in s_2}\tilde{g}^2_{G,c,1}(x)\hat{\mathcal{R}}^2(x) + \hat{\boldsymbol{\theta}}^t_{c,s_2}\hat{\Sigma}_{\hat{\tilde{\boldsymbol{Z}}}_{c,1,G}}\hat{\boldsymbol{\theta}}_{c,s_2} \quad [70]$$

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

$$\hat{\tilde{Y}}_{c,G,synth} = \bar{\boldsymbol{Z}}^t_G\hat{\boldsymbol{\theta}}_{c,s_2} \quad [71]$$

with estimated variance

$$\hat{\mathbb{V}}(\hat{\tilde{Y}}_{c,G,synth}) = \bar{\boldsymbol{Z}}^t_G\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{c,s_2}}\bar{\boldsymbol{Z}}_G \quad [72]$$

with obvious modification for the g-weights

$$\begin{aligned}
\tilde{g}_{G,c}(x) &= \bar{\boldsymbol{Z}}^t_G\boldsymbol{A}^{-1}_{c,s_2}M(x)\boldsymbol{Z}_c(x) \\
\hat{\tilde{Y}}_{c,G,synth} &= \frac{1}{n_2}\sum_{x\in s_2}\tilde{g}_{G,c}(x)Y_c(x) \\
\hat{\mathbb{V}}(\hat{\tilde{Y}}_{c,G,synth}) &= \frac{1}{n_2^2}\sum_{x\in s_2}\tilde{g}^2_{G,c}(x)\hat{\mathcal{R}}^2(x) \qquad [73]
\end{aligned}$$

Properties of the g-weights:

1. We have $\frac{1}{n_2}\sum_{x\in s_2} g_{c,G,1}(x)\boldsymbol{Z}_c(x) = \hat{\tilde{\boldsymbol{Z}}}_{c,1,G}$ and $\frac{1}{n_2}\sum_{x\in s_2}\tilde{g}_{c,G,1}(x)\boldsymbol{Z}_c(x) = \hat{\tilde{\boldsymbol{Z}}}_{1,G}$.

2. By considering $Y_c(x) \equiv 1$ one gets $\frac{1}{n_2}\sum_{x\in s_2} g_{c,G,1}(x) = \frac{1}{n_2}\sum_{x\in s_2}\tilde{g}_{c,G,1}(x) = 1$, i.e. the g-weights have means equal to 1.

3. When $G = F$ the estimator $\hat{Y}_{c,reg}$ is asymptotically equivalent to the sample mean over F, i.e. to $\frac{\sum_{x\in s_2} M(x)_c Y(x)}{\sum_{x\in s_2} M(x)}$. Thus, for large n_2, $g_{c,F,1}(x) \approx \frac{M(x)}{\bar{M}_2}$ with $\bar{M}_2 = \frac{1}{n_2}\sum_{x\in s_2} M(x)$. Likewise, $\hat{\tilde{Y}}_{G,synt}$ is asymptotically equivalent to the sample mean over the small domain, i.e. to $\bar{Y}_{c,G,2} = \frac{\sum_{x\in s_{2,G}} M(x)_c Y(x)}{\sum_{x\in s_{2,G}} M(x)}$. Hence, for large n_2 we get

$\tilde{g}_{c,G,1}(x) \approx 0$ for $x \notin \tilde{G}$ (negative values are possible) and to $\tilde{g}_{c,G,1}(x) \approx \frac{M(x)}{\bar{M}_{2,G}}$ for $x \in \tilde{G}$, where $\bar{M}_{2,G} = \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} M(x)$.

The construction of design-unbiased small-area estimators as synthetic or pseudo-synthetic estimators in the extended model containing the indicator variable of the small area of interest is mathematically more convenient than the classical approach. The mathematical approximation of the variances is also more satisfactory than simply treating the internal model as an external one and it can be formulated within the g-weight technique, which is known to offer several theoretical advantages.

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

5 Generalization to two-stage sampling

In many applications costs to measure the response variable Y_i are high. For instance, a good determination of the volume may require that one records DBH, as well as the diameter at $7m$ above ground and total height in order to utilize a three-way volume function. However, one could rely on a coarser, but cheaper, approximation of the volume based only on DBH. Nonetheless, it may be most sensible to assess those three parameters only on a sub-sample of trees. We now briefly formalize this simple idea, which is used in the Swiss National Forest Inventory. The reader is referred to (Mandallaz (2008), section 4.4, 4.5, 5.4 and 9.5) for details. For each point $x \in s_2$ trees are drawn with probabilities π_i. The set of selected trees is denoted by $s_2(x)$. From each of the selected trees $i \in s_2(x)$

one gets an approximation Y_i^* of the exact value Y_i. From the finite set $s_2(x)$ one draws a sub-sample $s_3(x) \subset s_2(x)$ of trees by Poisson sampling. For each tree $i \in s_3(x)$ one then measures the exact variable Y_i. Let us now define the second stage indicator variable

[74]
$$J_i(x) = \begin{cases} 1 \text{ if } i \in s_3(x) \\ 0 \text{ if } i \notin s_3(x) \end{cases}$$

To construct a good point estimate, we must have **the residual** $R_i = Y_i - Y_i^*$ which is known only for trees $i \in s_3(x)$. The **generalized local density** $Y^*(x)$ is defined according to

$$\begin{aligned} Y^*(x) &= \frac{1}{\lambda(F)} \left(\sum_{i=1}^{N} \frac{I_i(x) Y_i^*}{\pi_i} + \sum_{i=1}^{N} \frac{I_i(x) J_i(x) R_i}{\pi_i p_i} \right) \\ &= \frac{1}{\lambda(F)} \left(\sum_{i \in s_2(x)} \frac{Y_i^*}{\pi_i} + \sum_{i \in s_3(x)} \frac{R_i}{\pi_i p_i} \right) \end{aligned} \quad [75]$$

where the p_i are the conditional inclusion probabilities for the the second stage sampling, i.e. $p_i = \mathbb{P}(J_i(x) = 1 \mid I_i(x) = 1)$. It follows from general principles presented in (Mandallaz (2008), sections 4.4 and 4.5) that one can use all the previous results by replacing everywhere the exact local densities $Y(x)$, or $Y(x_l)$ in cluster sampling, by the corresponding generalized local densities $Y^*(x)$ or $Y^*(x_l)$. The second-stage variance is automatically taken into account.

6 Examples

6.1 Post-stratification

We consider the important special case of post-stratification, which illustrates the main issues. We consider a forested area F partitioned in L strata F_k, i.e. $F = \cup_{k=1}^{L} F_k$ and

a small area $G \subset F$, we set $G_k = G \cap F_k$. Note some G_k might be the empty set. The auxiliary vector is defined by the indicator variables of the L strata, i.e.

$$\boldsymbol{Z}^t(x) = (I_{F_1}(x), I_{F_2}(x), \ldots I_{F_L}(x))$$

where $I_{F_k}(x) = 1$ if $x \in F_k$ otherwise $I_{F_k}(x) = 0$. Note that condition [8] is fulfilled. Straightforward calculations lead to the (L, L) diagonal matrix $\boldsymbol{A}_{s_2} = \frac{1}{n_2} diag(n_{2,k})$ where $n_{2,k} = \sum_{x \in s_2} I_{F_k}(x)$. This leads to the obvious regression estimate

$$\hat{\boldsymbol{\beta}}_{s_2} = (\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_L)^t$$

with the empirical strata means $\hat{\beta}_k = \frac{1}{n_{2,k}} \sum_{x \in s_2 \cap F_k} Y(x) = \hat{\bar{Y}}_k$. After some elementary algebra the estimated variance-covariance matrix is found to be the diagonal (L, L) matrix

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} = diag(\frac{s_k^2}{n_{2,k}})$$

where $s_k^2 = \frac{1}{n_{2,k}} \sum_{x \in F_k} \hat{R}^2(x)$ with $\hat{R}(x) = Y(x) - \hat{\beta}_k$ for $x \in F_k$.
One obtains for the empirical mean of the auxiliary vector over the small area

$$\hat{\bar{\boldsymbol{Z}}}_{1,G} = (\hat{p}_{1,G}, \hat{p}_{2,G}, \ldots, \hat{p}_{L,G})^t = \hat{\boldsymbol{p}}_{1,G}$$

where

$$\hat{p}_{k,G} = \frac{\sum_{x \in s_1} I_{G_k}(x)}{\sum_{x \in s_1} I_G(x)} := \frac{n_{1k,G}}{n_{1,G}}$$

are the proportions of the strata surfaces areas within the small area as estimated from the large sample. Conditionally on $n_{1,G}$ the $n_{1k,G}$ follow the multinomial distribution with cell probabilities given by the vector $\boldsymbol{p}_G^t = (\frac{\lambda(G_1)}{\lambda(G)}, \frac{\lambda(G_2)}{\lambda(G)}, \ldots \frac{\lambda(G_L)}{\lambda(G)})$. In this case the

estimated variance-covariance matrix is known to be given by

$$[76] \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{p}}_{1,G}} = \frac{1}{n_{1,G}} \begin{bmatrix} \hat{p}_{1,G}(1-\hat{p}_{1,G}) & \hat{p}_{1,G}\hat{p}_{2,G} & \dots & \hat{p}_{1,G}\hat{p}_{1,L} \\ \hat{p}_{1,G}\hat{p}_{2,G} & \hat{p}_{2,G}(1-\hat{p}_{2,G}) & \dots & \hat{p}_{2,G}\hat{p}_{1,L} \\ \dots & \dots & \dots & \dots \\ \hat{p}_{1,G}\hat{p}_{1,L} & \hat{p}_{2,G}\hat{p}_{1,L} & \dots & \hat{p}_{1,L}(1-\hat{p}_{1,L}) \end{bmatrix}$$

Note that the same is obtained by using [52] after replacing $n_{1,G}-1$ by $n_{1,G}$. Simple algebra leads then to the pseudo-synthetic estimate

$$[77] \qquad \hat{Y}_{G,psynth} = \hat{\boldsymbol{p}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} = \sum_{k=1}^{L} \hat{p}_{k,G}\hat{\beta}_k$$

with estimated asymptotic design-based variance

$$[78] \qquad \hat{\mathbb{V}}(\hat{Y}_{G,psynth}) = \sum_{k=1}^{L} \hat{p}_{k,G}^2 \frac{s_k^2}{n_{2,k}} + \frac{1}{n_{1,G}} \sum_{k=1}^{L} \hat{p}_{k,G}\big(\hat{\beta}_k - \hat{Y}_{G,psynt}\big)^2$$

When $n_1 = \infty$ and $G = F$ this is precisely the exact conditional variance estimate, i.e. given the n_k. The g-weights are $g_{s_2}(x) = p_k \frac{n_2}{n_{2,k}}$ for $x \in F_k$, where $p_k = \frac{\lambda(F_k)}{\lambda(F)}$. Thus, for $n_1 < \infty$ the overall variance will depend on the variances within strata and on the variance between strata, which is given by the second term. Note also that for $G = F$ in [78] the strata weights are estimated from the large sample whereas this is not the case for the external model approach [17], which illustrates perfectly the better conditional properties of the g-weights technique (see Mandallaz (2008), p.84).

Remarks

If we assume the $\lambda(F_k)$ and therefore $\boldsymbol{A}$ to be known, the estimator

$$\hat{\boldsymbol{\beta}}_0 = \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\sum_{x\in s_2} Y(x)\boldsymbol{Z}(x)\Big)$$

is easily found to be $(\frac{n_{21}}{n_2 p_1}\hat{\bar{Y}}_1, \frac{n_{22}}{n_2 p_2}\hat{\bar{Y}}_2, \ldots \frac{n_{2L}}{n_2 p_L}\hat{\bar{Y}}_L)^t$ and yields $\hat{Y}_{synth} = \frac{1}{n_2}\sum_{\in s_2} Y(x)$, which is unbiased but useless. If we use $\boldsymbol{A}_{s_2}$ to estimate $\boldsymbol{\beta}$ we get as shown above $\hat{\beta}_k = \hat{\bar{Y}}_k$, which is very intuitive, and if we use $\boldsymbol{A}^{-1}$ instead of $\boldsymbol{A}_{s_2}^{-1}$ in the Taylor approximation for the variance, we obtain $\hat{\mathbb{V}}(\hat{Y}_{synth}) = \sum_{k=1}^{L} \frac{n_{2k} s_k^2}{n_2^2}$ instead of $\sum_{k=1}^{L} p_k^2 \frac{s_k^2}{n_{2k}}$, the later is of course much better from a conditional point of view (even if both estimates are asymptotically equivalent). This examples illustrates why it is better to work with $\boldsymbol{A}_{s_2}^{-1}$ throughout.

It can be easily checked that the original small-area estimator is given by

$$\hat{Y}_{G,psmall} = \sum_{k=1}^{L} \hat{p}_{k,G}\hat{\bar{Y}}_k + \sum_{k=1}^{L} \frac{n_{2k,G}}{n_{2,G}}(\hat{\bar{Y}}_{k,G} - \hat{\bar{Y}}_k)$$

where $n_{2k,G} = \sum_{x \in s_2} I_{G_k}(x)$ and $\hat{\bar{Y}}_{k,G} = \frac{1}{n_{2k,G}}\sum_{x \in s_2 \cap G_k} Y(x)$. Thus, the residual term will have an impact if the strata means within the small area differ from the strata means within the entire domain, which is intuitively clear.

The formulae for the variances are very cumbersome and not really informative, likewise for $\hat{\bar{Y}}_{G,psynth}$ in the extended model.

6.2 Case study

We reanalyze the case study described with full details in Chapter 9 of Mandallaz (2008). The inventoried area covered $218ha$. The auxiliary information is based on 16 stands defined by the following qualitative variables:

1. **Developmental stage**

   This entails four categories "pole stage=3," "young timber tree=4," "middle age timber tree=5," and "old timber tree=6." These were assigned according to the dominant diameter.

2. **Degree of mixture**

This variable was simplified to the categories of "predominantly conifers=1" and "predominantly broadleaves=2."

3. **Crown closure**

   This variable was based on canopy density, defined as the proportion of the entire ground surface within the stand that was covered by the tree crowns. It was simplified to the categories of "dense=1" and "close=2."

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

The inventory utilized systematic cluster sampling. The cluster comprises five points: central point, two points each established 30 m east or west of the central point; two other points each established 40 m either north or south of the central point.

The first phase sets the central cluster point on a 120 m W-E by 75 m N-S rectangular grid (note that the clusters partially overlapped in the N-S direction). The second, terrestrial phase, place the central point on a 1:4 sub-grid of the first phase, i.e. on a 240 m W-E by 150 m N-S systematic rectangular grid. The terrestrial inventory was purely one-stage with simple circular plots of $300m^2$ horizontal surface area, and an inventory threshold set at 12cm DBH.

We use the following linear model with the vector $\boldsymbol{Z}(x)$:

- $Z_1(x) \equiv 1$ intercept term

- $Z_2(x) = 1$ if x lies in Development Stage 3 and $Z_2(x) = 0$ otherwise
  $Z_3(x) = 1$ if x lies in Development Stage 4 and $Z_3(x) = 0$ otherwise
  $Z_4(x) = 1$ if x lies in Development Stage 5 and $Z_4(x) = 0$ otherwise
  $Z_2(x) = Z_3(x) = Z_4(x) = -1$ if x lies in Development Stage 6

- $Z_5(x) = 1$ if x lies in a coniferous stand and $Z_5(x) = -1$ otherwise

- $Z_6(x) = 1$ if x lies in a dense stand and $Z_6(x) = -1$ otherwise

Hence, we have an additive ANOVA model with 7 parameters, as compared with 16 parameters for the full stratification model.

We shall consider 5 small areas:

- Small area G_1 ($\approx 17ha$) was used for a full census. The condition that a cluster hitting the small area has all its points in F within the small area is occasionally violated (i.e. $I_{c,G}(x) = \frac{\sum_{l=1}^{M} I_G(x_l)}{M(x)} < 1$ for some x), so that the extended model for G_1 is only approximately correct. The mean residual over the small area is not exactly zero. The true values for basal area and stem densities are known.

- Small area G_2 ($\approx 33ha$) is the most eastern part of the forest.

- Small area $G_{21} \subset G_2$ ($\approx 7ha$) is a small subset in the central part of G_2 chosen to have a small number (3) of complete clusters ($I_{c,G}(x) \equiv 1$) spread over many different stands.

- Small area G_3 ($\approx 46ha$) is the most southern part of the forest.

- Small area G_4 ($\approx 55ha$) is the central part north of G_3.

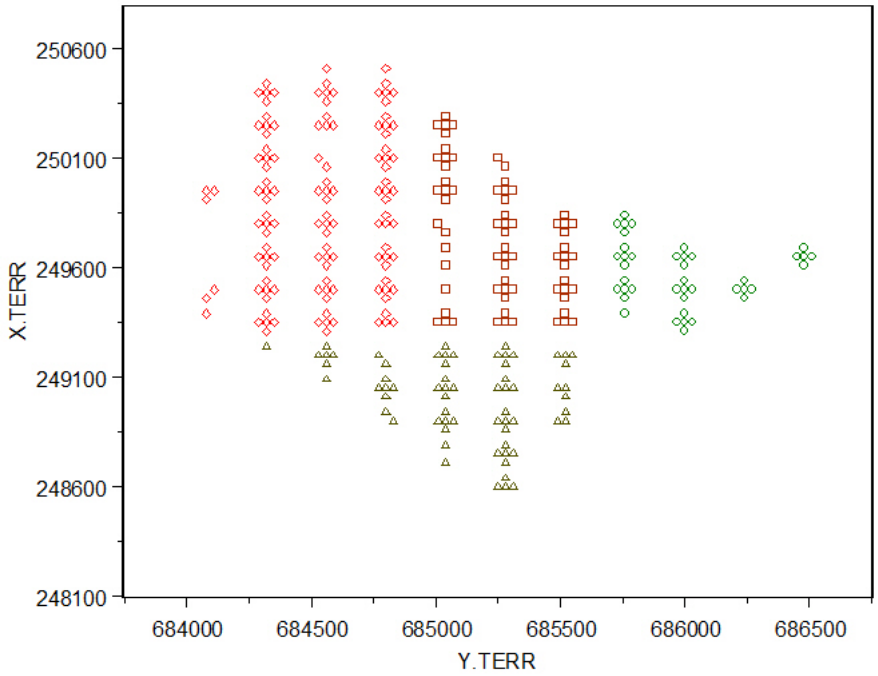
- Small area G_5 ($\approx 84ha$) is the most western part north of of G_3.

We have $F = G_2 \cup G_3 \cup G_4 \cup G_5$ and $I_{c,G_k} \equiv 1$ for $k = 2, 3, 4, 5$. Fig. 1 displays the terrestrial plots according to the domains $G_2 - G_5$. Stand map of F and detailed maps of G_1 are given in Mandallaz (2008) Chapter 8.

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

The standard error for $\hat{Y}_{c,G,small}$ are given within $(-)$ when considering the internal model

Figure 1: **Location of terrestrial plots in $G_2 - G_5$**



as an external one, i.e. by using the formulae (see Mandallaz (2008) p. 87):

$$\hat{Y}_{c,G,psmall} = \frac{\sum_{x\in s_{1,G}} M(x)\hat{Y}_c(x)}{\sum_{x\in s_{1,G}} M(x)} + \frac{\sum_{x\in s_{2,G}} M(x)\hat{R}_c(x)}{\sum_{x\in s_{2,G}} M(x)}$$

$$\begin{aligned}\widehat{\mathbb{V}}(\widehat{Y}_{c,G,psmall}) &= \left(1-\frac{n_{2,G}}{n_{1,G}}\right)\frac{1}{n_{2,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_{2,G}}\left(\frac{M(x)}{\bar{M}_{2,G}}\right)^2(\hat{R}_c(x)-\hat{\bar{R}}_{2,G})^2 \\ &+ \frac{1}{n_{1,G}}\frac{1}{(n_{2,G}-1)}\sum_{x\in s_{2,G}}\left(\frac{M(x)}{\bar{M}_{2,G}}\right)^2(Y_c(x)-\hat{Y}_{2,G})^2\end{aligned}$$

The standard errors given in [−] refer to the various Taylor expansions given in [59], [61], [66] or their equivalent g-weights versions.

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

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

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

7 Discussion and conclusions

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

For the classical small-area estimator the standard errors based on the external model assumption were usually, but not always, smaller than their counterparts based on the g-weights (equivalent to the Taylor asymptotic expansions), but the differences were small,

a reassuring result. The g-weights based standard errors of the synthetic estimators in the extended model for one single small area were usually smaller than their g-weights counterparts of the classical small-area estimator but generally still larger than under the external model assumption. The g-weights based standard errors in the extended model with several small areas were comparable to those derived specifically for one single small area. In this case study the various methods can be regarded as practically almost equivalent, which should be confirmed or eventually invalidated by further examples.

From a mathematical point of view the Taylor based g-weights technique in the models extended by the indicator variables of the small areas is without any doubts the most elegant approach: it bypasses the residuals terms and allows for a straightforward calculation of the asymptotic variances that takes into account the errors of the regression coefficients.

References

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

Koehl, M., Magnussen, S., and Marchetti, M. (2006). *Sampling Methods, Remote Sensing and GIS Multisource Forest Inventory*. Springer, Berlin Heidelberg.

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

Särndal, C., Swenson, B., and Wretman, J. (2003). *Model Assisted Survey Sampling*. Springer Series in Statistics, New York.

Searle, S. (1971). *Linear Models*. John Wiley, New York.

Tian, Y. and Takane, Y. (2009). The inverse of any two-by-two non singular partitioned matrix and three matrix inverse completion problems. *Computers and Mathematics with Application*, **57**:pp. 1294–1304.

Table 1: **Two-phase estimates for basal area**

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

Standard errors: [−] (Taylor, g-weights) and (−) (external model)

Remark:

The mean residual for small area G_1 was -1.59 for the extended model instead of 0 because $I_{c,G}(x) \not\equiv 1$.

Table 2: **Two-phase estimates for stem density**

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

Standard errors: [−] (Taylor, g-weights) and (−) (external model)

Remark:

The mean residual for small area G_1 was -1.00 for the extended model instead of 0 because $I_{c,G}(x) \not\equiv 1$.

Table 3: **Two-phase combined estimates**

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

Standard errors: [−] (Taylor, g-weights).

Remarks:

The classical estimates $\hat{Y}_{c,G,psmall}$ for G_4 were: 27.78(2.00) for basal area and 274.79(24.12) for stem density. For G_5 the corresponding results were 34.33(1.35) and 347.89(17.49).

As expected on mathematical grounds all extended models yielded empirical means of the residuals over the entire domain and over one or many small areas which were equal to zero ($< 10^{-12}$).