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

revised version

Daniel Mandallaz, Andreas Hill, Alexander Massey

Deparment of Environmental Systems Science

Chair of Land Use Engineering

ETH Zurich

CH 8092 Zurich, Switzerland

January 2016

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.

Foreword

The first version of this technical report has been published in the e-collection in September 2012. Since then many technical reports on further developments have been published in the e-collection (Mandallaz (2013c,b); Mandallaz and Massey (2015); Mandallaz (2015); Massey (2015)) as well as papers in the Canadian Journal of Forest Research, CJFR, (Mandallaz (2013a); Mandallaz et al. (2013); Mandallaz (2014); Massey et al. (2014); Massey and Mandallaz (2015a,b). Recently, Dr. Alexander Massey and Andreas Hill, MSc, started writing an R-package implementing the various procedures discussed in the aforementioned publications. Writing this software revealed some minor notational ambiguities used in cluster sampling as well as small numerical discrepancies between the new case study results produced in R and the original ones obtained with the IML procedure of the SAS package. Detective work revealed that some SAS programs contained a minor error in a small-area variance formulae that caused the mean number of points per cluster in the small area to be taken from the second-phase instead of from the first-phase sample. We emphasize that the underlying concept, i.e. "borrowing strength" by estimating regression coefficients over an unrestricted large area to improve the calculation of point and variance estimates in a restricted small area, remains precisely the same and that all the original conclusions of this technical report, including all relevant implications on subsequent works, remain valid. This revised version improves the notation used for cluster sampling, presents updated tables with the corrected results for the case study and provides an enhanced discussion of the results. Some comments have been added to take into account results obtained since 2012. The mathematical derivations have also been carefully checked and, so far, confirmed.

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

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

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

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

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. Besides, 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 whose sample points are independent of the current inventory's sample points. 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]
$$\begin{aligned} \hat{Y}_{G,synth} &= \frac{1}{\lambda(G)} \int_G \hat{Y}_{s_2}(x) dx \\ &= \frac{1}{\lambda(G)} \int_G \boldsymbol{Z}^t(x) \hat{\boldsymbol{\beta}}_{s_2} dx = \bar{\boldsymbol{Z}}_G^t \hat{\boldsymbol{\beta}}_{s_2} \end{aligned}$$

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. $\hat{Y}_{G,synth}$ 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 $\hat{Y}_{G,synth}$ in the design-based inference framework.

First, let us note that $\hat{Y}_{G,synth}$ 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 $\hat{Y}_{G,synth}$ 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)

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

where $s_{2,G} = s_2 \cap G$ and $n_{2,G} = \sum_{x \in s_2} I_G(x)$ is the number of points of s_2 falling within G. It 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 $\hat{Y}_{G,synth}$ 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 $n_{1,G}$ it 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 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 $\hat{\bar{Y}}_{G,small}$ 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

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

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

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

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

to get the pseudo-synthetic estimate in the extended model

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

As in [45] we have

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

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

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

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

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 $\tilde{g}_{G,1}(x) = \hat{\tilde{\boldsymbol{Z}}}^t_{1,G} \boldsymbol{A}^{-1}_{s_2} \boldsymbol{Z}(x)$ 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 for any G (i.e. also for $G = F$) one has $\frac{1}{n_2} \sum_{x \in s_2} g_{G,1}(x) = \frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{G,1}(x) = 1$, i.e. the g-weights have means equal to 1.

4. When $G = F$ the estimator $\hat{Y}_{reg}$ and the sample mean $\frac{1}{n_2} \sum_{x \in s_2} Y(x)$ both converge towards the true value for an arbitrary $Y(x)$ 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, $\hat{\tilde{Y}}_{G,synth}$ and the sample mean $\frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} Y(x)$ converge towards the true value for an arbitrary $Y(x)$,

which implies that $\tilde{g}_{G,1}(x)$ 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 simple random sampling one can derive non-parametric versions of the above procedures, i.e. kernel-based regression estimators or k nearest neighbors estimators (**knn**). However, they are more difficult to implement than the regression estimators described above, primarily because on must use re-sampling (bootstrap) techniques to obtain reliable variance estimators because analytical results are beyond reach and the classical external variance formula severely underestimate the true variance as shown by simulations. Also, the performances of the non-parametric estimators are not better than the regression estimators (for global and small-area estimation) if the coefficient of determination R^2 is sufficient large, say above 0.7, which is the case for timber volume assessment using modern remote sensing data as auxiliary information (e.g. canopy height parameters obtain via e.g. LiDAR). Details can be found in Mandallaz and Massey (2015), Massey and Mandallaz (2015a) and the PhD thesis Massey (2015). In the next section we generalize the previous results on regression estimators to cluster sampling, widely used in national inventories. The main ideas remain the same but the formulae are slightly more cumbersome due to the random cluster size. The non-parametric estimators are not yet available for cluster sampling.

4 Generalization to cluster sampling

We follow the description of cluster sampling as defined 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 (though 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

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

The small-area estimation set up is slightly more complicated because the points of a cluster with $M(x)\geq 0,\ x\in\tilde{F}$ can be spread over more than one small area. The notation used in the 2012 version of this technical report was not sufficiently explicit in this respect. This does not affect the regression coefficients, always based on the entire

sample in F (the borrowing strength philosophy), but the formulation "after restricting the samples to the small area of interest" is now more explicit in the notation. We define the following quantities:

$$\begin{aligned}
M_G(x) &= \sum_{l=1}^{L} I_G(x+e_l) \\
Y_{c,G}(x) &= \frac{\sum_{l=1}^{L} I_G(x+e_l) Y(x+e_l)}{M_G(x)} \\
\hat{\bar{Y}}_{c,2,G} &= \frac{\sum_{x\in s_{2,G}} M_G(x) Y_{c,G}(x)}{\sum_{x\in s_{2,G}} M_G(x)} \\
\boldsymbol{Z}_{c,G}(x) &= \frac{\sum_{l=1}^{L} I_G(x+e_l) \boldsymbol{Z}(x+e_l)}{M_G(x)} \\
\hat{Y}_{c,G}(x) &= \boldsymbol{Z}^t_{c,G}(x) \hat{\boldsymbol{\beta}}_{c,s_2} \\
\hat{R}_{c,G} &= Y_{c,G}(x) - \hat{Y}_{c,G}(x) \\
\hat{\bar{\boldsymbol{Z}}}_{c,1,G} &= \frac{\sum_{x\in s_{1,G}} M_G(x) \boldsymbol{Z}_{c,G}(x)}{\sum_{x\in s_{1,G}} M_G(x)}
\end{aligned} \qquad [57]$$

The estimated variance of the one-phase estimator is given by (Mandallaz (2008), section 4.3).

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_{c,2,G}) = \frac{1}{n_{2,G}(n_{2,G}-1))} \sum_{x\in s_{2,G}} \big(\frac{M_G(x)}{\bar{M}_{2,G}}\big)^2 (Y_{c,G}(x) - \hat{\bar{Y}}_{c,2,G})^2 \qquad [58]$$

where $\bar{M}_{2,G} = \frac{\sum_{x\in s_{2,G}} M_G(x)}{n_{2,G}}$.

The estimate of the covariance matrix of the auxiliary variables is given by

$$\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_{c,1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)} \sum_{x\in s_{1,G}} \big(\frac{M_G(x)}{\bar{M}_{1,G}}\big)^2 (\boldsymbol{Z}_{c,G}(x) - \hat{\bar{\boldsymbol{Z}}}_{c,1,G})(\boldsymbol{Z}_{c,G}(x) - \hat{\bar{\boldsymbol{Z}}}_{c,1,G})^t \qquad [59]$$

where $\bar{M}_{1,G} = \frac{\sum_{x\in s_{1,G}} M_G(x)}{n_{1,G}}$.

The **pseudo-synthetic estimate** is then

$$\begin{aligned} \hat{Y}_{c,G,psynth} &= \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G}\hat{\boldsymbol{\beta}}_{c,s_2} \\ &= \frac{1}{n_2}\sum_{x\in s_2} g_{c,1,G}(x)Y_c(x) \end{aligned} \quad [60]$$

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]

$$\hat{\mathbb{V}}(\hat{Y}_{c,G,psynth}) = \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}\hat{\bar{\boldsymbol{Z}}}_{c,1,G} + \hat{\boldsymbol{\beta}}^t_{c,s_2}\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,1,G}}\hat{\boldsymbol{\beta}}_{c,s_2} \quad [61]$$

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

$$\hat{Y}_{c,G,psmall} = \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G}\hat{\boldsymbol{\beta}}_{c,s_2} + \frac{\sum_{x\in s_{2,G}} M_G(x)\hat{R}_{c,G}(x)}{\sum_{x\in s_{2,G}} M_G(x)} \quad [62]$$

It is asymptotically design-unbiased and, by analogy with simple random sample (see [33]) its variance can be expected to be well approximated by

$$\begin{aligned} \hat{\mathbb{V}}(\hat{Y}_{c,G,psmall}) &= \hat{\bar{\boldsymbol{Z}}}^t_{c,1,G}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}\hat{\bar{\boldsymbol{Z}}}_{c,1,G} + \hat{\boldsymbol{\beta}}^t_{c,s_2}\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,1,G}}\hat{\boldsymbol{\beta}}_{c,s_2} \\ &+ \frac{1}{n_{2,G}(n_{2,G}-1)}\sum_{x\in s_{2,G}}\Big(\frac{M_G(x)}{\bar{M}_{2,G}}\Big)^2(\hat{R}_{c,G}(x) - \bar{\hat{R}}_{2,G})^2 \end{aligned} \quad [63]$$

where $\bar{M}_{2,G} = \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}} M_G(x)$ and $\bar{\hat{R}}_{2,G} = \frac{\sum_{x\in s_{2,G}} M_G(x)\hat{R}_{c,G}(x)}{\sum_{x\in s_{2,G}} M_G(x)}$. One can rewrite [63] 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

$I_{c,G}(x) = \frac{\sum_{l=1}^{M} I_G(x_l)}{M(x)} = \frac{M_G(x)}{M(x)}$. We shall assume that $M_G(x) \equiv M(x)$ for all $x \in \tilde{G} = \{x \mid \sum_{l=1}^{M} I_G(x_l) > 0\}$. If we have a partition of the forest in small areas $G_k, k = 1, 2, \ldots L$, then $\tilde{F} \subset \cup_{k=1}^{L} \tilde{G}_k$ but it is unlikely in general that the $\tilde{G}_k$ are disjoined. We can reasonably hope that in extensive forest inventory the surface area of points where $M_G(x) < M(x)$ is negligible. The theoretical normal equation reads

$$[64] \qquad \Big(\int_{\tilde{F}} M(x) \boldsymbol{\mathcal{Z}}_c(x) \boldsymbol{\mathcal{Z}}_c^t(x) dx \Big) \boldsymbol{\theta}_c =: \boldsymbol{\mathcal{A}}_c \boldsymbol{\theta}_c = \int_{\tilde{F}} M(x) Y_c(x) \boldsymbol{\mathcal{Z}}_c(x) dx$$

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

$$[65] \qquad \hat{\boldsymbol{\theta}}_{c,s_2} = \boldsymbol{\mathcal{A}}_{c,s_2}^{-1} \Big(\frac{1}{n_2} \sum_{x \in s_2} M(x) \boldsymbol{\mathcal{Z}}_c(x) Y(x) \Big)$$

with estimated design-based covariance matrix

$$[66] \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{c,s_2}} = \boldsymbol{\mathcal{A}}_{c,s_2}^{-1} \Big(\frac{1}{n_2^2} \sum_{\in s_2} M^2(x) \hat{\mathcal{R}}_c^2(x) \boldsymbol{\mathcal{Z}}_c(x) \boldsymbol{\mathcal{Z}}_c^t(x) \Big) \boldsymbol{\mathcal{A}}_{c,s_2}^{-1}$$

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

$$[67] \qquad \hat{\tilde{Y}}_{c,G,psynth} = \hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G}^t \hat{\boldsymbol{\theta}}_{c,s_2}$$

where $\hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G} = \frac{\sum_{x \in s_{1,G}} M_G(x) \boldsymbol{\mathcal{Z}}_{c,G}(x)}{\sum_{x \in s_{1,G}} M_G(x)}$ is the mean of the extended auxiliary vector over the

small area.

As in [61] the estimated variance is given by

$$[68] \qquad \hat{\mathbb{V}}(\hat{\tilde{Y}}_{c,G,psynth}) = \hat{\tilde{\boldsymbol{\mathcal{Z}}}}^t_{c,1,G} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{c,s_2}} \hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G} + \hat{\boldsymbol{\theta}}^t_{c,s_2} \hat{\Sigma}_{\hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G}} \hat{\boldsymbol{\theta}}_{c,s_2}$$

where

$$[69] \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)} \sum_{x \in s_{1,G}} \left(\frac{M_G(x)}{\bar{M}_{1,G}}\right)^2 (\boldsymbol{\mathcal{Z}}_{c,G}(x) - \hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G})(\boldsymbol{\mathcal{Z}}_{c,G}(x) - \hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G})^t$$

Remark:

Even if the condition $M_G(x) \equiv M(x)$ is fulfilled in $s_{2,G}$ we cannot expect $\hat{\tilde{Y}}_{c,G,psynth}$ to be asymptotically unbiased (although the sample copy of the normal equation converges towards the theoretical normal equation this does not guarantee zero mean residuals over G as we have seen). However the bias can be expected to be negligible.

Obviously a decomposition similar to [50] will hold so that $\hat{\tilde{Y}}_{c,psynth}$ and $\hat{Y}_{c,G,psmall}$ in [62,67] will be close to each other in large samples.

Defining the g-weights at the cluster level as

$$[70] \qquad \tilde{g}_{c,1,G}(x) = \hat{\tilde{\boldsymbol{\mathcal{Z}}}}^t_{c,1,G} \boldsymbol{\mathcal{A}}^{-1}_{c,s_2} M(x) \boldsymbol{\mathcal{Z}}_c(x)$$

we obtain as usual

$$[71] \qquad \hat{\tilde{Y}}_{c,G,psynth} = \frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{c,1,G}(x) Y_c(x)$$

and

$$[72] \qquad \hat{\mathbb{V}}(\hat{\tilde{Y}}_{c,G,psynth}) = \frac{1}{n_2^2} \sum_{x \in s_2} \tilde{g}^2_{c,1,G}(x) \hat{\mathcal{R}}^2(x) + \hat{\boldsymbol{\theta}}^t_{c,s_2} \hat{\Sigma}_{\hat{\tilde{\boldsymbol{\mathcal{Z}}}}_{c,1,G}} \hat{\boldsymbol{\theta}}_{c,s_2}$$

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

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

with estimated variance

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

with obvious modification for the g-weights

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

Properties of the g-weights:

1. We have $\frac{1}{n_2} \sum_{x \in s_2} g_{c,1,G}(x) \boldsymbol{Z}_c(x) = \hat{\bar{\boldsymbol{Z}}}_{c,1,G}$ and $\frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{c,1,G}(x) \boldsymbol{Z}_c(x) = \hat{\bar{\boldsymbol{Z}}}_{1,G}$.
2. By considering $Y_c(x) \equiv 1$ one gets $\frac{1}{n_2} \sum_{x \in s_2} g_{c,1,G}(x) = \frac{1}{n_2} \sum_{x \in s_2} \tilde{g}_{c,1,G}(x) = 1$, i.e. the g-weights have means equal to 1.
3. When $G = F$ the estimator $\hat{Y}_{c,F,psynth}$ and the sample mean $\frac{\sum_{x \in s_2} M(x) Y(x)}{\sum_{x \in s_2} M(x)}$ both converge towards the true value over F. Hence, for large n_2, $g_{c,1,F}(x) \approx \frac{M(x)}{\bar{M}_2}$ with $\bar{M}_2 = \frac{1}{n_2} \sum_{x \in s_2} M(x)$. Likewise for $\hat{\bar{Y}}_{G,psynth}$ and the sample mean over G, $\hat{\bar{Y}}_{c,2,G} = \frac{\sum_{x \in s_{2,G}} M_G(x) Y_{c,G}(x)}{\sum_{x \in s_{2,G}} M_G(x)}$. Hence, for large n_2 we get $\tilde{g}_{c,1,G}(x) \approx 0$ for $x \notin \tilde{G}$ (negative values are possible) and to $\tilde{g}_{c,1,G}(x) \approx \frac{M_G(x)}{\bar{M}_{2,G}}$ for $x \in \tilde{G}$, where $\bar{M}_{2,G} = \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} M_G(x)$.

Remarks:

With simple random sampling 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. The same essentially holds in cluster sampling, with a slight advantage maybe for $\hat{Y}_{c,G,psmall}$,with the external variance, in terms of simplicity and validity, since one cannot guarantee in general zero mean residuals over G for the pseudo-synthetic estimator $\hat{\hat{Y}}_{c,G,psynth}$ in the extended model. It is certainly good practice to calculate the different small-area estimators and their estimated variances, particularly for cluster sampling. Large differences could reveal unsuspected patterns in the data (or eventually even programming errors).

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.

To calculate confidence intervals we recommend to use the Student's T distribution on n_2 d.f. for F and $n_{2,G} - 1$ d.f. for any small area G. If $n_{2,G}$ is to small one can use the pseudo-synthetic estimator $\hat{Y}_{c,G,psynth}$, which will yield short confidence intervals, at the cost of a potential bias. In this case it might instructive to embed G in a larger domain G^* with n_{2,G^*} large enough and to calculate the pseudo-synthetic estimate for G in the model extended with the indicator variable of G^*.

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

[76]
$$J_i(x) = \begin{cases} 1 \text{ if } i \in s_3(x) \\ 0 \text{ if } i \notin s_3(x) \end{cases}$$

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

$$\begin{aligned} Y^*(x) &= \frac{1}{\lambda(F)} \left(\sum_{i=1}^{N} \frac{I_i(x) Y_i^*}{\pi_i} + \sum_{i=1}^{N} \frac{I_i(x) J_i(x) R_i}{\pi_i p_i} \right) \\ [77] \qquad &= \frac{1}{\lambda(F)} \left(\sum_{i \in s_2(x)} \frac{Y_i^*}{\pi_i} + \sum_{i \in s_3(x)} \frac{R_i}{\pi_i p_i} \right) \end{aligned}$$

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. More mathematical details are given in Mandallaz (2015).

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.

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

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

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

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

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

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

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

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

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

with estimated asymptotic design-based variance

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

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 [80] 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 $(\frac{n_{21}}{n_2 p_1}\hat{\bar{Y}}_1, \frac{n_{22}}{n_2 p_2}\hat{\bar{Y}}_2, \dots \frac{n_{2L}}{n_2 p_L}\hat{\bar{Y}}_L)^t$ 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 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 obtained from the manual interpretation of aerial photographs:

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. The coefficient of determination R^2 (in the classical model-dependent approach, i.e. without taking the cluster structure into account) was satisfactory for the stem density (0.53) but not for the basal area (0.19). This case study was performed in 1989 and is the only data available for cluster sampling in Switzerland (the reason for using cluster sampling was to provide good estimates of the variograms for the geostatistical Kriging procedures, see Mandallaz (2008) and Mandallaz (1993) for details). It should be emphasized that with modern remote sensing techniques (e.g. LiDAR or digital aerial photographs) one can obtain much better R^2's, e.g. around 0.70 or more for timber volume (mean canopy height being the most important explanatory variable).

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_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 $I_{c,G_k}(x) \equiv 1$ 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 3 display the results for the basal area and the stem density for the small areas G_1, G_2, G_3, G_4 and G_5.

The external variances for $\hat{Y}_{c,G,psmall}$ are given by the equations

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

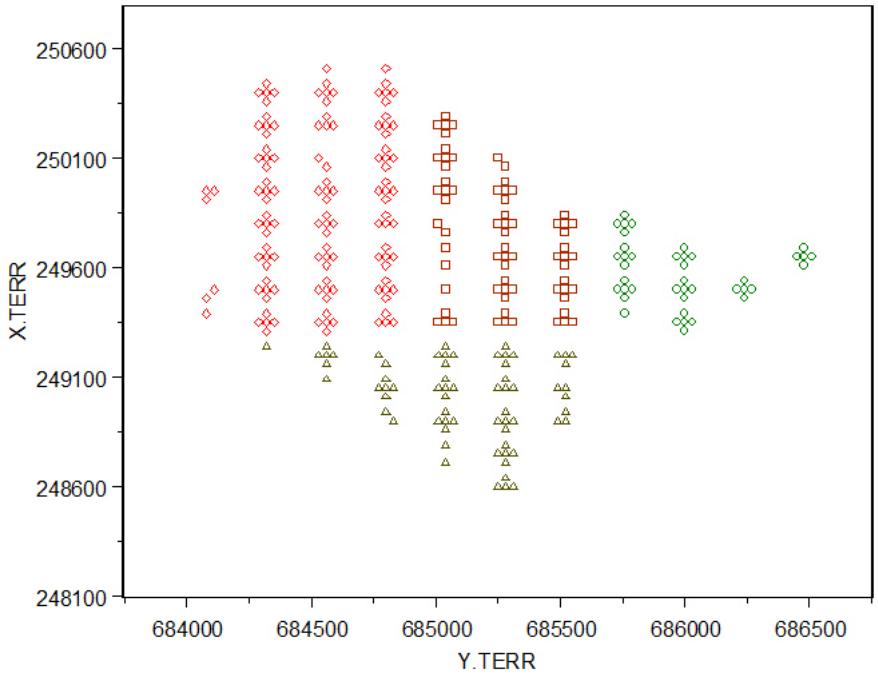
and

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

The resulting standard errors are given in $(-)$.

Tables 1 (basal area) and 3 (density of stems) summarize the results for the sample mean $\hat{\bar{Y}}_G$, the pseudo-synthetic estimator $\hat{Y}_{c,G,psynth}$, the pseudo small-area estimator $\hat{Y}_{c,G,psmall}$

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



and the pseudo-synthetic estimator in the extended model $\hat{\tilde{Y}}_{c,G,psynth}$ for F and the small areas G_1, G_2, G_3, G_4, G_5.

The extended model for the $\hat{\tilde{Y}}_{c,G_k,psynth}$ contains only the indicator variable of the corresponding small area G_k. For this reason the corresponding estimates for the entire domain F depend on the selected G_k and they are given separately in Tables 2 and 4.

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 variables $I_{G_k}(x)$ $(k = 2, 3, 4, 5)$, the previous components $Z_l(x), l = 2, 3, 4, 5, 6$ but no longer the intercept term $Z_1(x) \equiv 1$ (otherwise $\boldsymbol{\mathcal{A}}_{s_2}$ would be singular because $Z_1(x)$ is a linear combination of the $I_{G_k}(x)$). The results for this estimator, denoted by $\hat{\tilde{Y}}_{c,G,cpsynth}$, are

displayed in Table 5.

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

7 Discussion and conclusions

The results for the basal area are disappointing, except maybe for global estimation, in the sense that the regression estimators, with the exception of the pseudo-synthetic estimator $\hat{Y}_{c,G,psynth}$, frequently have a larger error than the sample mean. This is of course due to the very low coefficient of determination $R^2 = 0.19$ and we shall therefore not pursue the discussion for basal area, simply noting that all point estimates were close to each other, with the exception of the small area G_1 with the full census, for which the pseudo-synthetic estimator was closer to the true values (also for the density of stems). 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 both basal area and the density of stem.

As expected for the stem density (with an acceptable $R^2 = 0.53$) the two-phase estimators perform much better than the sample mean. For the classical small-area estimator $\hat{Y}_{c,G,psmall}$ the standard errors based on the external model assumption were smaller than their counterparts based on the g-weights (i.e. on the design-based covariance matrix of the regression coefficients) for F, G_3, G_4, G_5 and slightly larger for G_1, G_2, but the differences were small, a reassuring result. The point estimates of $\hat{Y}_{c,G,psmall}$ were very close to the point estimates of the pseudo-synthetic estimator $\hat{\hat{Y}}_{c,G,psynth}$ (in the extended model with the single indicator variable of the small area of interest), which has slightly smaller g-weights standard errors, except for F and G_1. From a practical point of view $\hat{Y}_{c,G,psmall}$ and $\hat{\hat{Y}}_{c,G,psynth}$ are essentially equivalent. The estimators $\hat{\hat{Y}}_{c,G,psynth}$, $\hat{Y}_{c,G,psmall}$ and the combined estimator $\hat{\hat{Y}}_{c,G,cpsynth}$ (model without intercept but extended with the 4 indicator variables of G_2, G_3, G_4, G_5) were also practically equivalent for F, G_2, G_3, G_4, G_5, the

combined estimator having slightly smaller g-weights standard errors.

The pseudo-synthetic estimator $\hat{Y}_{c,G,psynth}$ yields point estimates close to those of $\hat{Y}_{c,G,psmall}$, $\hat{\tilde{Y}}_{c,G,psynth}$, $\hat{\tilde{Y}}_{c,G,cpsynth}$ (recall that $\hat{Y}_{c,F,psmall} = \hat{Y}_{c,F,psynth}$) for F, G_2, G_3, G_4, G_5 but, of course, with much smaller g-weights based standard errors, at a cost of a potential bias. For G_1 the point estimate $\hat{Y}_{c,G_1,psynth}$ was very close to the true value (for the density of stem and the basal area), precisely because the clusters hitting G_1 were, as afore mentioned, not "representative" for G_1. This can happen!

From a mathematical point of view the g-weights technique in the models extended by the indicator variables of the small areas is without any doubts the most elegant approach in simple random sampling: it bypasses the residuals terms and allows for a straightforward calculation of the asymptotic variance that takes into account the errors of the regression coefficients. The same is likely to hold in large sample also with cluster sampling. The results from the case study show that the differences between the asymptotically unbiased estimators are very small from a practical point of the view, and that a potentially biased estimator can by chance be closer to the truth.

References

Gregoire, T. and Dyer, M. (1989). Model fitting under patterned heterogeneity 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. (1993). Geostatistical methods for double sampling schemes:Applications to combined forest inventory. Technical report, ETH Zurich, Department of Environmental Systems Science, habilitation thesis, http://e-collection.linrary.ethz.ch.

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. (2013c). Regression estimators in forest inventories with two-phase sampling and partially exhaustive information with application to small-area estimation. 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. (2015). Mathematical details on two-phase two-stage and three-phase two-stage regression estimators in forest inventories: designed-based Monte-Carlo approach. Technical report, ETH Zurich, Department of Environmental Systems Science, 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**:pp. 1023–1031.

Mandallaz, D. and Massey, A. (2015). Regression and non-parametric estimators for two-phase forest inventories in the design-based Monte-Carlo approach. Technical report, ETH Zurich, Department of Environmental Systems Science, http://e-collection.library.ethz.ch.

Massey, A. (2015). *Multiphase estimation procedures for forest inventories under the design-based Monte-Carlo approach*. Ph.D. thesis, ETH Zurich, Chair of Forest Inventory and Planning, http://e-collection.library.ethz.ch/.

Massey, A. and Mandallaz, D. (2015a). Comparison of classical, kernel-based and nearest neighbors regression estimators: designed-based Monte-Carlo approach in the context of two-phase forest inventories. *Can. J. For. Res.*, **45**(10):pp. 1480–1488.

Massey, A. and Mandallaz, D. (2015b). Designed-based regresession estimation of net change for forest inventories. *Can. J. For. Res.*, **45**(10):pp. 1775–1784.

Massey, A., Mandallaz, D., and Lanz, A. (2014). Integrating remote sensing and past inventory data under the new annual design of the Swiss National Forest Inventory using three-phase design-based regression estimation. *Can. J. For. Res.*, **44**(10):pp. 1177–1186.

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: **Results for basal area**

Domain	$n_1 : n_2$ $n_1\bar{M}_1 : n_2\bar{M}_2$	$\hat{\bar{Y}}_G$	$\hat{Y}_{c,G,psynth}$	$\hat{Y}_{c,G,psmall}$	$\hat{\bar{Y}}_{c,G,psynth}$
F	298 : 73 1203 : 298	31.90 (1.08)	31.34 [0.94] (0.89)	31.34 [0.94] (0.89)	see Table 2
G_1 true=29.60	29 : 8 92 : 19	24.54 (3.59)	30.28 [1.37]	23.99 [3.96] (3.68)	25.55 [3.76]
G_2	49 : 9 185 : 41	30.69 (2.05)	28.27 [1.47]	29.32 [2.55] (2.08)	29.31 [2.33]
G_3	73 : 18 250 : 66	32.32 (2.05)	31.62 [1.54]	31.46 [2.37] (1.87)	31.46 [2.18]
G_4	81 : 17 306 : 69	27.63 (2.15)	29.72 [1.15]	27.78 [2.27] (2.00)	27.77 [1.98]
G_5	125 : 29 462 : 122	34.49 (1.78)	33.50 [1.00]	34.33 [1.55] (1.35)	34.33 [1.30]

Design-based standard errors: [−] (g-weights based) and (−) (based on external or sample variance)

For G_1 the mean residual of $\hat{\bar{Y}}_{c,G,psynth}$ was -1.59 and 0, as expected, for all other domains.

Table 2: **Results basal area for F with extended models**

$\hat{\hat{Y}}^{(1)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(2)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(3)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(4)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(5)}_{c,F,psynth}$
31.30	31.36	31.35	31.30	31.33
[0.92]	[0.93]	[0.94]	[0.92]	[0.94]

$\hat{\hat{Y}}^{(k)}_{c,F,psynth}$ is the pseudo-synthetic estimate for F in the model extended with the indicator variable of domain G_k, $k = 1, 2, 3, 4, 5$. The g-weights based standard errors are given in $[-]$.

Table 3: **Results for stem density**

Domain	$n_1 : n_2$ $n_1\bar{M}_1 : n_2\bar{M}_2$	$\hat{\bar{Y}}_G$	$\hat{Y}_{c,G,psynth}$	$\hat{Y}_{c,G,psmall}$	$\hat{\bar{Y}}_{c,G,psynth}$
F	298 : 73 1203 : 298	321.03 (18.44)	325.79 [12.80] (12.09)	325.79 [12.80] (12.09)	see Table 4
G_1 true=280.23	29 : 8 92 : 19	245.61 (65.51)	279.54 [27.26]	257.34 [48.26] (48.29)	258.20 [49.56]
G_2	49 : 9 185 : 41	393.50 (79.69)	400.49 [30.95]	406.47 [42.96] (43.49)	406.41 [39.36]
G_3	73 : 18 250 : 66	275.25 (20.98)	279.75 [17.16]	282.46 [22.68] (16.56)	282.40 [20.77]
G_4	81 : 17 306 : 69	284.06 (39.51)	307.85 [19.69]	274.79 [26.64] (24.12)	274.53 [23.55]
G_5	125 : 29 462 : 122	342.35 (24.50)	332.66 [15.33]	347.89 [21.26] (17.49)	347.91 [18.65]

Design-based standard errors: [−] (g-weights based) and (−) (based on external or sample variance)

For G_1 the mean residual of $\hat{\bar{Y}}_{c,G,psynth}$ was -0.99 and 0, as expected, for all other domains.

Table 4: **Results for stem density for F with extended models**

$\hat{\hat{Y}}^{(1)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(2)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(3)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(4)}_{c,F,psynth}$	$\hat{\hat{Y}}^{(5)}_{c,F,psynth}$
325.62	325.88	325.72	325.15	325.56
[12.81]	[12.84]	[12.85]	[12.67]	[12.62]

$\hat{\hat{Y}}^{(k)}_{c,F,psynth}$ is the pseudo-synthetic estimate for F in the model extended with the indicator variable of domain G_k, $k = 1, 2, 3, 4, 5$. The g-weights based standard errors are given in $[-]$.

Table 5: **Two-phase combined estimates**

Domain	**sample sizes** $n_1 : n_2$ $n_1\bar{M}_1 : n_2\bar{M}_2$	**basal area** $\hat{\hat{Y}}_{c,G,cpsynth}$	**stem density** $\hat{\hat{Y}}_{c,G,cpsynth}$
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]

The g-weights based standard errors are given in $[-]$. 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 ($< 10^{-12}$).