Mathematical details of two-phase/two-stage and three-phase/two-stage regression estimators in forest inventories Design-based Monte Carlo approach

Author(s):
Mandallaz, Daniel

Publication Date:
2015-05

Permanent Link:
https://doi.org/10.3929/ethz-a-010441361 →

Mathematical details of two-phase/two-stage and three-phase/two-stage regression estimators in forest inventories: design-based Monte Carlo approach

Daniel Mandallaz [1]

Department of Environmental Systems Science

Chair of Land Use Engineering

ETH Zurich

CH 8092 Zurich, Switzerland

Mai 2015

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

Foreword

This technical report discusses some mathematical details of regression estimators under two or three-phase sampling with two-stage sampling of trees at the plot level. Discussions with colleagues and graduate students convinced me that the mathematical derivations presented (and often simply sketched) in previous work deserved clarification. This is particularly true, because of space restriction, for the papers Mandallaz (2013a), Mandallaz et al. (2013), Mandallaz (2014) and to a lesser degree for the the book Mandallaz (2008) and the technical reports (Mandallaz (2012), Mandallaz (2013c), Mandallaz (2013b)). The great physicist and Nobel prize laureate Richard Feynman once said, after a life-time frequentation of mathematicians, that he finally understood what the word obvious really means, namely proved. Of course, the level he was referring to is stratospheric in comparison with the down-to-earth properties of regression estimators in forest inventory. Nevertheless, I thought it would not be harmful, a few months before my retirement, to clarify some theoretical issues in survey sampling, in particular the differences between design-based model-assisted versus model-dependent inference, which are not always well understood by practitioners, as I can now judge after 25 years experience as a reviewer and author. It turns out that all the previously published results are indeed correct, at least asymptotically. The main point is that one can use mutatis mutandis the two-phase/three-phase one-stage formulae also with two-stage sampling at the plot level. The second-stage variance resulting from Poisson sampling at the plot level is so to speak automatically taken into account. This holds in particular for the easy to use so-called external variances. In contrast to previous work the notation is more pedantic, which is the price to pay to be as clear and accurate as possible.

1 Introduction

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 spatial mean over F** is defined as

[1] $$\bar{Y}_F = \frac{1}{\lambda(F)} \sum_{i=1}^{N} Y_i =: \frac{1}{\lambda(F)} \sum_{i \in F} Y_i$$

Let us first recall the definitions of local density and generalized local density. For a random point x uniformly distributed in F trees are selected at x from the population $\mathcal{P}$ for instance with concentric circles or angle count techniques. Each tree has its associated inclusion circle K_i centered on the tree. The set of trees selected at point x is denoted by $s_2(x)$. The first-stage inclusion probabilities are $\pi_i = \frac{\lambda(K_i \cap F}{\lambda(F)}$. For each of the selected trees $i \in s_2(x)$ one determines Y_i. The indicator variable I_i is defined as

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

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

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

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 at the forest edge: $\lambda(F)\pi_i = \lambda(F \cap K_i)$. In the infinite population approach (now better known as Monte Carlo approach) one samples the function $Y(x)$ (Mandallaz (2008)) for which the following important relation holds:

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

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 $\{i = 1, 2 \ldots N \mid Y_i\}$. The infinite populations and super-populations models were introduced in my PhD thesis (Mandallaz (1991)) to solve the obvious contradiction that a domain of the plane cannot be viewed (i.e. partitioned) as a finite population of single or concentric circles (even infinitely many with the angle count). Furthermore, the local density $Y(x)$ as defined above is a regionalized variable in the sense of Matheron (Matheron (1965)), who developed the purely model-dependent geostatistical Kriging techniques with primarily the estimation of mining or oil resources in mind (where one has to estimate 3-dimensional integrals). The application of geostatistical methods in forest inventory was therefore quite natural in the Monte Carlo approach and led in particular to the model-dependent Double Kriging procedure discussed in my habilitation thesis (Mandallaz (1993)).

Readers totally unfamiliar with concepts such as Monte Carlo approach, design-based versus model-dependent inference (unfortunately sometimes also called model-based inference) should maybe consult for a first perusal the paper of Gregoire (1998) on general survey sampling and the text book of Mandallaz (2008) (chapter 4, 6 and 7) in the context of Monte Carlo approach to forest inventory.

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 D_7 at $7m$ above ground and total height H 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 and species alone. 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.

From each of the selected trees $i \in s_2(x)$ one determines coarse approximation $\hat{Y}_i$ (e.g. based on species and DBH) of the pseudo exact value Y_i (pseudo exact in the sense that is is a better approximation of the unknown true bole volume based on e.g. species, DBH, D_7 and H). 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 determines the (pseudo) exact variable Y_i. The second stage indicator variable is

[5]
$$J_i(x) = \begin{cases} 1 \text{ if } i \in s_3(x) \\ 0 \text{ if } i \notin s_3(x) \end{cases}$$

The residual s at the tree level is denoted by $R_i = Y_i - \hat{Y}_i$ which is known only for trees $i \in s_3(x)$. The **generalized local density** $Y^*(x)$ is defined according to

[6]
$$\begin{aligned} Y^*(x) &= \frac{1}{\lambda(F)} \left(\sum_{i=1}^{N} \frac{I_i(x)\hat{Y}_i}{\pi_i} + \sum_{i=1}^{N} \frac{I_i(x)J_i(x)R_i}{\pi_i p_i} \right) \\ &= \frac{1}{\lambda(F)} \left(\sum_{i \in s_2(x)} \frac{\hat{Y}_i}{\pi_i} + \sum_{i \in s_3(x)} \frac{R_i}{\pi_i p_i} \right) \end{aligned}$$

where the p_i are the conditional inclusion probabilities for the the second stage sampling, i.e. $p_i = \mathbb{P}(J_i(x) = 1 \mid I_i(x) = 1)$. The second term is an Horwitz-Thomson estimator and the generalized local density can be viewed as a difference estimator (Särndal et al. (2003), section 6.3). The expected value and the variance under Poisson sampling at point

x are easily found to be

[7]
$$\begin{aligned}\mathbb{E}_{Poisson}(Y^*(x)) &= Y(x)\\ \mathbb{V}_{Poisson}(Y^*(x)) &= \frac{1}{\lambda^2(F)}\Big(\sum_{i\in s_2(x)}\frac{R_i^2(1-p_i)}{\pi_i^2 p_i}\Big) =: V(x)\end{aligned}$$

If necessary $V(x)$ can be estimated unbiasedly by

[8]
$$\hat{V}(x) = \frac{1}{\lambda^2(F)}\Big(\sum_{i\in s_3(x)}\frac{R_i^2(1-p_i)}{\pi_i^2 p_i^2}\Big)$$

2 Two-phase sampling

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 and denoted by $\mathbf{Z}(x) \in \mathbb{R}^p$. 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, or simple random sampling SRS**. The first phase randomization procedure will be indicated by the subscript 1, the second phase by the subscripts 2 or x and the second-stage procedure at a given point x by the subscript 3. Note that the points $x \in s_2$ inherit from s_1 the property of being i.i.d. uniformly distributed in F. One has directly from the definition of Poisson sampling and the decomposition rule

$$\mathbb{V}_{2,3}Y^*(x) = \mathbb{V}_2\mathbb{E}_{3|2}Y^*(x) + \mathbb{E}_2\mathbb{V}_{3|2}(Y^*(x))$$

the following properties

$$\begin{aligned}
\mathbb{E}_{3|2}Y^*(x) &= Y(x) \\
\mathbb{E}_{2,3}Y^*(x) &= \mathbb{E}_x(Y(x)) = \bar{Y}_F \\
\mathbb{V}_{2,3}(Y^*(x)) &= \mathbb{V}_x(Y(x)) + \mathbb{E}_x(V(x)) \\
&= \mathbb{V}_x(Y(x)) + \frac{1}{\lambda^2(F)}\Big(\sum_{i=1}^{N}\frac{(R_i^2(1-p_i)}{\pi_i p_i}\Big)
\end{aligned} \quad [9]$$

2.1 The model

In the **design-based approach** we set

$$Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta} + R_{\boldsymbol{\beta}}(x) =: \hat{Y}_{\boldsymbol{\beta}}(x) + R_{\boldsymbol{\beta}}(x) \quad [10]$$

the quantities $Y(x), \boldsymbol{Z}(x), R_{\boldsymbol{\beta}}(x)$ are random because the point x is random and uniformly distributed in F. Equation [10] can be viewed, for any fixed given $\boldsymbol{\beta}$, as simply defining the residual term $R_{\boldsymbol{\beta}}(x)$. **The true regression coefficient $\boldsymbol{\beta}_0$ is by definition the theoretical least squares estimate**

$$[11] \qquad \boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta}} \int_F (Y(x) - \boldsymbol{Z}^t(x)\boldsymbol{\beta})^2 dx$$

It satisfies the normal equation

$$[12] \qquad \Big(\int_F \boldsymbol{Z}(x)\boldsymbol{Z}^t(x)dx\Big)\boldsymbol{\beta}_0 = \int_F Y(x)\boldsymbol{Z}(x)dx$$

and the orthogonality relationship

$$[13] \qquad \int_F R_0(x)\boldsymbol{Z}(x)dx = \mathbf{0}$$

where $R_0(x) = R_{\boldsymbol{\beta}_0} = Y(x) - \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0 =: Y(x) - \hat{Y}_0(x)$. 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 the following properties

$$[14] \qquad \frac{1}{\lambda(F)}\int_F R_0(x)dx = 0 \text{ and } \frac{1}{\lambda(F)}\int_F \hat{Y}_0(x) = \frac{1}{\lambda(F)}\int_F Y(x)dx = \bar{Y}$$

In the **model-dependent approach** one postulates a model

$$[15] \qquad Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta} + R(x)$$

where **for a fixed point** x the quantities $Y(x)$ and $R(x)$ are viewed random variables, $\boldsymbol{R}(x)$ with zero mean and a given covariance structure. The vector of auxiliary $\boldsymbol{Z}(x)$ can have for a given x non-random or random components. We emphasize the fact that in the design-based model-assisted approach the model [10] is not viewed as the true complex stochastic process generating the $Y(x)$, but, more pragmatically, **simply as a tool to get better estimators of $\bar{Y}$ than the ordinary sample mean (which corresponds to the model $\boldsymbol{Z}(x)$ containing only the intercept and therefore no auxiliary information at all**. Of course, ideally, the model should capture qualitatively the main features of the underlying natural phenomenon. Hence, one could say that in the design-based approach the forest under investigation is a **fixed entity observed at random points**, whereas in the model-dependent set-up the actual forest is viewed as **the realization of a complex stochastic process observed at given fixed points**. That these two totally different approaches yield usually (but not always, particularly for small area estimation) similar results in practice is in some sense a miracle.

To simplify the notation we set 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

$$\boldsymbol{A}\boldsymbol{\beta}_0 = \mathbb{E}_x \boldsymbol{U}(x) := \boldsymbol{U}$$

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

$$\begin{aligned}\boldsymbol{\beta}_0 &= \boldsymbol{A}^{-1}\boldsymbol{U}\\ \hat{\boldsymbol{\beta}}_{s_2} &= \boldsymbol{A}_{s_2}^{-1}\boldsymbol{U}_{s_2}\end{aligned} \tag{16}$$

$\hat{\boldsymbol{\beta}}_{s_2}$ is asymptotically design-unbiased for $\boldsymbol{\beta}_0$. It can be shown that the asymptotic design-base variance-covariance matric of $\hat{\boldsymbol{\beta}}_{s_2}$ is given by

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}} = \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\mathbb{E}_x R_0^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\Big)\boldsymbol{A}^{-1} \tag{17}$$

which can be estimated by replacing the theoretical residuals $R_0(x) = Y(x) - \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0$ with their empirical counterparts $\hat{R}_{s_2}(x) = Y(x) - \hat{Y}_{s_2}(x)$, with $\hat{Y}_{s_2}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}$, and $\boldsymbol{A}$ with $\boldsymbol{A}_{s_2}$. We then get the **estimated design-based variance-covariance matrix** as

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} := \boldsymbol{A}_{s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}_{s_2}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}(x)^t\Big)\boldsymbol{A}_{s_2}^{-1} \tag{18}$$

Interestingly this is precisely the **robust estimate of the model-dependent covariance matrix** proposed by Huber (1967) and also discussed by Gregoire and Dyer (1989) in a forestry context (see also Mandallaz (2008), pp. 123-125 for a proof).

With two-stage sampling one replaces the true local density $Y(x)$ by its estimate $Y^*(x)$ and considers

$$\hat{\boldsymbol{\beta}}_{s_2}^* = \boldsymbol{A}_{s_2}^{-1}\boldsymbol{U}_{s_2}^* \tag{19}$$

where we have set $\boldsymbol{U}^*_{s_2} = \frac{1}{n_2}\sum_{x\in s_2} Y^*(x)\boldsymbol{Z}(x)$. One has $\mathbb{E}_{3|2}(\hat{\boldsymbol{\beta}}^*_{s_2}) = \hat{\boldsymbol{\beta}}_{s_2}$.

We can now derive the design-based variance covariance matrix of $\hat{\boldsymbol{\beta}}^*_{s_2}$. One has

$$\begin{aligned}
\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^*_{s_2}} &= \mathbb{E}_{2,3}\big((\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2})(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2})^t\big) + \mathbb{E}_{2,3}(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}_0)^t \\
&+ \mathbb{E}_{2,3}\big(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2})(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}_0) + \mathbb{E}_{2,3}\big(\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2}) \\
&= \mathbb{E}_2\mathbb{V}_{3|2}(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2}) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{s_2}} + 0 + 0 \\
&= \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{s_2}} + \mathbb{E}_2\boldsymbol{A}_2^{-1}(\frac{1}{n_2^2}\sum_{x\in s_2} V(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x))\boldsymbol{A}_2^{-1} \\
&\approx \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{s_2}} + \boldsymbol{A}^{-1}(\frac{1}{n_2}\mathbb{E}_x V(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x))\boldsymbol{A}^{-1} \\
&= \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\mathbb{E}_x\big((V(x) + R^2(x))\big)\boldsymbol{Z}(x)\boldsymbol{Z}(x)\big)^t\Big)\boldsymbol{A}^{-1}
\end{aligned}$$

where we have use the independence of Poisson sampling at different points, equations [17] and the identity

$$\hat{\boldsymbol{\beta}}^*_{s_2} - \boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2} + \hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{\beta}_0$$

Hence, we have

[20]
$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}^*_{s_2}} = \boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\mathbb{E}_x\big((V(x) + R^2(x))\big)\boldsymbol{Z}(x)\boldsymbol{Z}(x)\big)^t\Big)\boldsymbol{A}^{-1}$$

The main point is that the variance of $\hat{\boldsymbol{\beta}}^*_{s_2}$ decrease as n_2^{-1}, like the variance of $\hat{\boldsymbol{\beta}}_{s_2}$, so that $\hat{\boldsymbol{\beta}}^*_{s_2}$ is also an asymptotically unbiased and consistent estimator of $\boldsymbol{\beta}_0$. The true prediction $\hat{Y}_0 = \boldsymbol{Z}(x)^t\boldsymbol{\beta}_0$ is of course not available and we have the following possibilities

[21]
$$\begin{aligned}
\hat{Y}_{\boldsymbol{\beta}}(x) &= \boldsymbol{Z}^t(x)\boldsymbol{\beta} \\
\hat{Y}_{s_2}(x) &= \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2} = \hat{Y}_0(x) + O_p\big(\frac{1}{\sqrt{n_2}}\big) \\
\hat{Y}^*_{s_2}(x) &= \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}^*_{s_2} = \hat{Y}_0(x) + O_p\big(\frac{1}{\sqrt{n_2}}\big)
\end{aligned}$$

where we have set $\hat{Y}_0(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0$.

Therefore, we can treat all the above predictions asymptotically as if they were depending on $\boldsymbol{Z}(x)$ only and not on the sample, which is precisely the definition of the external variance assumption.

Setting now $\hat{R}^*_{s_2}(x) = Y^*(x) - \hat{Y}^*_{s_2}(x) \approx Y^*(x) - \hat{Y}_0(x)$ we get first $\mathbb{E}_{3|x}\hat{R}^*_{s_2}(x) = \hat{R}_{s_2}(x)$, $\mathbb{V}_{3|2}\hat{R}^*_{s_2}(x) = \mathbb{V}_{3|2}Y^*(x) = V(x)$ and finally $\mathbb{E}_{3|x}(R^*_{s_2}(x))^2 = \hat{R}^2_{s_2}(x) + V(x)$. From this we have the following consistent estimator of the design-based variance covariance matrix of $\hat{\boldsymbol{\beta}}^*_{s_2}$

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

2.2 The two-phase one-stage regression estimators

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

[23] $$\hat{Y}_{reg} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_{\boldsymbol{\beta}}(x) + \frac{1}{n_2}\sum_{x\in s_2}R_{\boldsymbol{\beta}}(x)$$

with the predictions $\hat{Y}_{\boldsymbol{\beta}}(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}$ and the residuals $R_{\boldsymbol{\beta}}(x) = Y(x) - \hat{Y}_{\boldsymbol{\beta}}(x)$, where $\boldsymbol{\beta}$ 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

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

to obtain

$$[25] \qquad \mathbb{V}(\hat{Y}_{reg}) = \frac{1}{n_1}\mathbb{V}(Y(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R_{\boldsymbol{\beta}}(x))$$

which can be unbiasedly estimated with

$$[26] \qquad \hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(Y(x)-\bar{Y}_2)^2 + (1-\frac{n_2}{n_1})\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}(R_{\boldsymbol{\beta}}(x)-\bar{R}_{\boldsymbol{\beta},2})^2$$

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

If the model is **internal** the regression coefficients are estimated with the inventory data at hand to obtain first $\hat{\boldsymbol{\beta}}_{s_2}$, second the predictions $\hat{Y}_{s_2}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}$ and residuals $\hat{R}_{s_2}(x) = Y(x) - \hat{Y}_{s_2}(x)$ and finally the two-phase one-stage regression estimator

$$[27] \qquad \hat{Y}_{reg} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_{s_2}(x) + \frac{1}{n_2}\sum_{x\in s_2}\hat{R}_{s_2}(x)$$

It can be shown (Mandallaz (2013a) and Mandallaz (2012) for details) that asymptotically ($n_2 \to \infty$) that one can treat an internal model as an external one, so that we have the consistent variance estimate

$$[28] \qquad \hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(Y(x)-\bar{Y}_{s_2})^2 + (1-\frac{n_2}{n_1})\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}(\hat{R}_{s_2}(x)-\bar{\hat{R}}_{s_2})^2$$

where we have set $\bar{Y}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2} Y(x)$ and $\bar{\hat{R}}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\hat{R}_{s_2}(x)$. Because predictions and residuals are orthogonal by construction we also have

$$[29] \qquad \hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(\hat{Y}_{s_2}(x)-\bar{\hat{Y}}_{s_2})^2 + \frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}(\hat{R}_{s_2}(x)-\bar{\hat{R}}_{s_2})^2$$

where $=\bar{\hat{Y}}_{s_2} = \frac{1}{n_2}\sum_{x\in s_2}\hat{Y}_{s_2}(x)$. The first term contains the empirical variance of the predictions in the small sample s_2, which could be replace by the empirical variance in the large sample s_1. Furthermore, by construction the mean of the residuals in the second term is zero by construction so that one can also use

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

where $\bar{\hat{Y}}_{s_1} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_{s_2}(x)$. This version is more in line with the variance obtained via the design-based covariance matrix derived in (Mandallaz (2012) eqn. [28]) and which reads

[31] $$\mathbb{V}(\hat{Y}_{reg}) = \bar{\boldsymbol{Z}}^t\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}}\bar{\boldsymbol{Z}} + \boldsymbol{\beta}^t\Sigma_{\hat{\bar{\boldsymbol{Z}}}_1}\boldsymbol{\beta}$$

which yields the asymptotically consistent estimate of the variance

[32] $$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \hat{\bar{\boldsymbol{Z}}}_1^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_1 + \hat{\boldsymbol{\beta}}_{s_2}^t\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_1}\hat{\boldsymbol{\beta}}_{s_2}$$

with the estimated variance-covariance matrix of the $\boldsymbol{Z}(x)$ given by

[33] $$\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_1} = \frac{1}{n_1(n_1-1)}\sum_{x\in s_1}(\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}}_1)(\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}}_1)^t$$

with $\hat{\bar{\boldsymbol{Z}}}_1 = \frac{1}{n_1}\sum_{x\in s_1}\boldsymbol{Z}(x)$.
The **g-weights** are defined as

[34] $$g_2(x) = \hat{\bar{\boldsymbol{Z}}}_1^t\boldsymbol{A}_{s_2}^{-1}\boldsymbol{Z}(x)$$

They satisfy the **calibration properties**

[35] $$\frac{1}{n_2}\sum_{x\in s_2} g_2(x)\boldsymbol{Z}(x) = \hat{\hat{\boldsymbol{Z}}}_1$$

In particular $\frac{1}{n_2}\sum_{x\in s_2} g_2(x) = 1$. The regression estimator can be rewritten, because of the zero mean residuals (i.e. $\frac{1}{n_2}\sum_{x\in s_2} \hat{R}_{s_2}(x) = 0$) as the weighted mean

[36] $$\hat{Y}_{reg} = \frac{1}{n_2}\sum_{x\in s_2} g_2(x)Y(x)$$

and also as

[37] $$\hat{Y}_{reg} = \frac{1}{n_2}\sum_{x\in s_2} \tilde{g}_2(x)Y(x)$$

with the g-weights

[38] $$\tilde{g}_2(x) = 1 + (\hat{\hat{\boldsymbol{Z}}}_1 - \hat{\hat{\boldsymbol{Z}}}_2)^t \boldsymbol{A}_{s_2}^{-1}\boldsymbol{Z}(x)$$

Hence, for all $Y(x)$, we have

$$\frac{1}{n_2}\sum_{x\in s_2} g_2(x)Y(x) = \frac{1}{n_2}\sum_{x\in s_2} \tilde{g}_2(x)Y(x)$$

Since the g-weights depend only on $\boldsymbol{Z}(x)$ and not $Y(x)$ we must have

[39] $$g_2(x) = \tilde{g}_2(x)$$

Because $\hat{\hat{\boldsymbol{Z}}}_1 - \hat{\hat{\boldsymbol{Z}}}_2$ tends to zero almost surely and in quadratic mean, we see that $\tilde{g}_2(x) = g_2(x) = 1 + O_p(\frac{1}{\sqrt{n_2}})$. We have thus proved the **fundamental property that the g-weights are asymptotically equal to** 1.

Furthermore the estimated variance can be expressed as

$$\begin{aligned}\hat{\mathbb{V}}(\hat{Y}_{reg}) &= \frac{1}{n_2^2}\sum_{x\in s_2} g_2^2(x)\hat{R}^2_{s_2}(x) + \hat{\boldsymbol{\beta}}^t_{s_2}\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_1}\hat{\boldsymbol{\beta}}_{s_2} \\ &= \frac{1}{n_2^2}\sum_{x\in s_2} g_2^2(x)\hat{R}^2_{s_2}(x) + \frac{1}{n_1(n_1-1)}\sum_{x\in s_1}(\hat{Y}_{s_2}(x) - \bar{\hat{Y}}_1)^2 \qquad [40]\end{aligned}$$

with $\bar{\hat{Y}}_1 = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}_{s_2}(x)$ The second term in the last equation is the variance of the predictions over F. Since the g-weights tend to 1 asymptotically [30] and [40] are asymptotically equivalent.

Remarks:

- The notation $g_2(x)$ for g-weights can be misleading in the sense that they do not depent solely on x, and consequently $\boldsymbol{Z}(x)$ but also on the sample s_1 (because of $\boldsymbol{A}_2$ and s_1 (because of $\hat{\bar{\boldsymbol{Z}}}_1$. Hence, a better notation would be $g_{s_1,s_2}(x)$, which is cumbersome to use in complicated formulae.

- If the first phase is exhaustive, i.e. $n_1 = \infty$, we have $g_2(x) = \bar{\boldsymbol{Z}}^t\boldsymbol{A}^{-1}_{s_2}\boldsymbol{Z}(x)$. With $Y(x) = \boldsymbol{Z}^t(x)\boldsymbol{\beta}_0 + R_0(x)$ and $\bar{Y}_F = \frac{1}{\lambda(F)}\int_F \hat{Y}_0(x)dx = \bar{\boldsymbol{Z}}^t\boldsymbol{\beta}_0$ we get, using [35]

$$\begin{aligned}\hat{Y}_{reg} - \bar{Y}_F &= \frac{1}{n_2}\sum_{x\in s_2} g_2(x)Y(x) - \bar{\boldsymbol{Z}}^t\boldsymbol{\beta}_0 \\ &= \frac{1}{n_2}\sum_{x\in s_2} g_2(x)(\boldsymbol{Z}^t(x)\boldsymbol{\beta}_0 + R_0(x)) - \bar{\boldsymbol{Z}}^t(x)\boldsymbol{\beta}_0 \\ &= \bar{\boldsymbol{Z}}^t\boldsymbol{\beta}_0 + \frac{1}{n_2}\sum_{x\in s_2} g_2(x)R_0(x) - \bar{\boldsymbol{Z}}^t\boldsymbol{\beta}_0 \\ &= \sum_{x\in s_2} g_2(x)R_0(x)\end{aligned}$$

  We also have $\mathbb{E}_x g_2(x)R_0(x) \approx \bar{\boldsymbol{Z}}^t\boldsymbol{A}^{-1}\mathbb{E}_x\boldsymbol{Z}(x)R_0(x) = 0$ because of the orthogonality relationship. Comparing with [40] we see that for calculating the variance we can consider the $g_2(x)R_0(x)$ and then the $g_2(x)\hat{R}(x)$ as i.i.d with zero mean.

2.3 The two-phase two-stage regression estimator

We will now show that treating an internal model as if it were external is also asymptotically correct in the two-phase two-stage procedures. This was up to certain point taken for granted in Mandallaz (2008),section 5.2, and subsequent work, where the predictions were considered to depend only on $\boldsymbol{Z}(x)$, which is only asymptotically correct.

The predictions are now calculated according to $\hat{Y}^*_{s_2}(x) = \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}^*_{s_2}$ and the two-phase two-stage regression estimator reads now

$$\begin{aligned} \hat{Y}^*_{reg} &= \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}^*_{s_2}(x) + \frac{1}{n_2}\sum_{x\in s_2}\hat{R}^*_{s_2}(x) \\ &= \hat{\hat{\boldsymbol{Z}}}_1\hat{\boldsymbol{\beta}}^*_{s_2} \end{aligned} \qquad [41]$$

with $\hat{R}^*_{s_2}(x) = Y^*_{s_2}(x) - \hat{Y}^*_{s_2}(x)$. Obviously one has $\mathbb{E}_{3|2,1}\hat{Y}^*_{reg} = \hat{Y}_{reg}$. Using the variance decomposition

$$\mathbb{V}_{1,2,3}(\hat{Y}^*_{reg}) = \mathbb{V}_{1,2}\mathbb{E}_{3|2,1}(\hat{Y}^*_{reg}) + \mathbb{E}_{1,2}\mathbb{V}_{3|2,1}(\hat{Y}^*_{reg}) \qquad [42]$$

we get

$$\mathbb{V}_{1,2,3}(\hat{Y}^*_{reg}) = \mathbb{V}_{1,2}(\hat{Y}_{reg}) + \mathbb{E}_{1,2}\mathbb{V}_{3|2,1}(\hat{\hat{\boldsymbol{Z}}}_1\hat{\boldsymbol{\beta}}^*_{s_2}) \qquad [43]$$

The first term is given by [31] and for the third we need the conditional variance-covariance matrix $\Sigma_{\hat{\boldsymbol{\beta}}^*_{s_2}|2,1}$, which is defined as

$$\Sigma_{\hat{\boldsymbol{\beta}}^*_{s_2}|2,1} = \mathbb{E}_{3|2,1}(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2})(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2})^t$$

Because Poisson sampling is performed independently at points $x \in s_2$ we have with [19] and [7] $\mathbb{V}_{3|2,1} \sum_{x \in s_2} Y^*(x) = \sum_{x \in s_2} V(x)$

$$\mathbb{V}_{3|2,1}\Big(\sum_{x \in s_2} Y^*(x)\Big) = \sum_{x \in s_2} V(x)$$

This gives after some algebra

$$\Sigma_{\hat{\boldsymbol{\beta}}^*_{s_2}|2,1} = \boldsymbol{A}_{s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x \in s_2} V(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\Big)\boldsymbol{A}_{s_2}^{-1} \qquad [44]$$

Using [43], [31], [17] and collecting the pieces we obtain obtain the asymptotic variance

$$\begin{aligned}
\mathbb{V}_{1,2,3}(\hat{Y}^*_{reg}) &= \frac{1}{n_1}\mathbb{V}(\hat{Y}_0(x)) + \bar{\boldsymbol{Z}}^t\boldsymbol{A}^{-1}\Big(\frac{1}{n_2}\mathbb{E}_x(R_0^2(x) + V(x))\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\Big)\boldsymbol{A}^{-1}\bar{\boldsymbol{Z}} \\
&= \frac{1}{n_1}\mathbb{V}(\hat{Y}_0(x)) + \bar{\boldsymbol{Z}}^t\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*_{s_2}}\bar{\boldsymbol{Z}} \qquad [45]
\end{aligned}$$

and consequently the following consistent estimate of the variance

$$\begin{aligned}
\hat{\mathbb{V}}(\hat{Y}^*_{reg}) &= \hat{\boldsymbol{\beta}^*}^t_{s_2}\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_1}\hat{\boldsymbol{\beta}^*}_{s_2} + \hat{\bar{Z}}_1^t\boldsymbol{A}_{s_2}^{-1}\big(\frac{1}{n_2^2}\sum_{x \in s_2}(\hat{R}^*(x))^2\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\big)\boldsymbol{A}_{s_2}^{-1}\hat{\bar{\boldsymbol{Z}}}_1 \\
&= \frac{1}{n_1}\frac{1}{n_1 - 1}\sum_{x \in s_2}(\hat{Y}^*(x) - \bar{\hat{Y}}^*_1)^2 + \frac{1}{n_2^2}\sum_{x \in s_2} g_2^2(x)(\hat{R}^*(x))^2 \qquad [46]
\end{aligned}$$

which is algebraically the same as [35] except for the $*$. Since the $g_2^2(x)$ tend to 1 we have proved that the external variance estimate is also equivalent to the consistent g-weight variance in two-phase two-stage sampling.

In the model-dependent geostatistical set-up the situation is more complicated, even in simple one-phase inventory, and one has to use Kriging with measurement error instead (see section 7.3 in Mandallaz (2008)).

For timber volume assessment in the the Swiss National Inventory the second stage variance is remarkably small (around 3% in comparison to the overall variance (i.e. $\frac{\mathbb{E}V(x)}{\mathbb{V}(Y^*(x))} \approx 0.03$. However,the slight increase in variance is more than counterbalanced by drastic savings in measurement costs (Mandallaz and Massey (2012) and Massey (2011) for details).

Remarks:

1. One-phase two-stage sampling is the special case with $\boldsymbol{Z}(x)$ consisting only of the intercept, which leads to the results given in (Mandallaz (2008), pp.70-71)

$$\begin{aligned}
\hat{Y}^* &= \frac{1}{n_2}\sum_{x\in s_2} Y^*(x) \\
\mathbb{V}(\hat{Y}^*) &= \frac{1}{n_2}\mathbb{V}_x(Y(x)) + \frac{1}{n_2}\mathbb{E}_x V(x) \\
\hat{\mathbb{V}}(\hat{Y}^*) &= \frac{1}{n_2(n_2-1)}\sum_{x\in s_2}(Y^*(x)-\hat{Y}^*)^2 \\
\mathbb{E}_{2,3}(\hat{\mathbb{V}}(\hat{Y}^*) &= \mathbb{V}(\hat{Y}^*)
\end{aligned}$$

   These results are exact and not only asymptotically correct in contrast to the general case.

2. From a theoretical point of view the g-weights variance estimates have better statistical properties. The simple post-stratification example is particularly instructive in this respect: the external variance estimate does not use the strata weights estimated from the large sample and the variance is not inversely proportional to the number of s_2-points in the strata, in contrast to the g-weight variance estimate (see Mandallaz (2008), pp. 84-86). Also, it is closer, under regularity conditions, to the model-dependent variance if the model is correct (see Mandallaz (2008), pp. 114-116).

3. In practice one uses primarily systematic grids, for which no design-based variance estimate exists. In this case the only mathematically absolutely correct approach

is to use the geostatistical Kriging procedures: treating systematic grids as random samples will usually overestimate the variance, particularly in one-phase sampling and for small-area estimation. The situation is better in two-phase sampling because the variance of the residuals is generally the dominating term, with a much shorter range of the spatial correlation than the range of the observations $Y(x)$. For global estimation the two-phase regression estimator and its geostatistical counterpart, the Double-Kriging estimator, seem to give very similar results (both for point and variance estimates). With two-stage sampling one has to use Kriging with measurement errors (see Mandallaz (2008), chapter 7, for an introduction to geostatistical methods in forest inventory and Mandallaz (1993) for a detailed mathematical treatment). It is fair to say that geostatistical techniques are far more difficult to implement than the design-based procedures and, in my opinion, they are not suitable for routine work in forest inventory. They are undoubtedly very successful in oil and mining resource assessment, where the financial stakes are completely different. However, the increasing demand for small-area estimation and the availability of free geostatistical software packages within the R-project (20 years ago the software packages, essentially used in the mining and oil industry, were extremely costly) with good interfaces to GIS software will certainly promote geostatistical techniques in forest inventories.

3 Three-phase sampling

3.1 Background

The **null phase** draws a very large sample s_0 of n_0 points $x_i \in s_0$ $(i = 1, 2 \ldots n_0)$ that are independently and uniformly distributed within the forest area F. At each of those points auxiliary information is collected, very often coding information of qualitative nature (e.g. following the interpretation of aerial photographs) or quantitative (e.g. timber volume estimates based on LiDAR measurements). We shall assume that the auxiliary information at point x is described by the column vector $\boldsymbol{Z}^{(1)}(x) \in \Re^p$. The case $n_0 = \infty$, i.e. $\boldsymbol{Z}^{(1)}(x)$ is **exhaustive**, has been investigated in Mandallaz et al. (2013). The **first phase** draws a large sample $s_1 \subset s_0$ of $n_1 << n_0$ points by simple random sampling in s_0. Note that the points $x \in s_1$ are also uniformly independently distributed in F. For each point in the first phase a further component $\boldsymbol{Z}^{(2)}(x) \in \Re^q$ of the auxiliary information is available and hence also the vector $\boldsymbol{Z}^t(x) = (\boldsymbol{Z}^{(1)t}(x), \boldsymbol{Z}^{(2)t}(x)) \in \Re^{p+q}$ (the upper index t denotes the transposition operator). The **second phase** draws a small sample $s_2 \subset s_1$ of n_2 points from s_1 by simple random sampling and consists of the terrestrial inventory. Note that we have used the terms null, first and second phases instead of first, second and third phases simply to ensure compatibility with the terminology used in previous work, in which the null-phase was exhaustive.

To set the stage the component $\boldsymbol{Z}^{(1)}(x) \in \Re^p$ can be based on the interpretation of aerial photographs or on simple characteristics of the canopy height obtained from LiDAR data (such as mean canopy height and eventually quantiles thereof), whereas $\boldsymbol{Z}^{(2)}(x) \in \Re^p$ is based on other computationally intensive characteristics of the canopy requiring individual tree detection (e.g. tree species ortree volume prediction based on tree height). The reason for introducing the null-phase sample s_0 is that the component $\boldsymbol{Z}^{(1)}(x)$ can be computationally prohibitive to calculate exhaustively in extensive forest inventories (see

Mandallaz (2013a) for a case study with LiDAR data).

3.2 The models

We shall work with the following linear models

1. The large model M

   [47] $$Y(x)=\boldsymbol{Z}^t(x)\boldsymbol{\beta}+R_{\boldsymbol{\beta}}(x)=\boldsymbol{Z}^{(1)t}(x)\boldsymbol{\beta}^{(1)}+\boldsymbol{Z}^{(2)t}(x)\boldsymbol{\beta}^{(2)}+R_{\boldsymbol{\beta}}(x)=:\hat{Y}_{\boldsymbol{\beta}}(x)+R_{\boldsymbol{\beta}}(x)$$

   with $\boldsymbol{\beta}^t=(\boldsymbol{\beta}^{(1)t},\boldsymbol{\beta}^{(2)t})$ and the predictions $\hat{Y}_{\boldsymbol{\beta}}(x)=\boldsymbol{Z}^t(x)\boldsymbol{\beta}$ and residuals $R_{\boldsymbol{\beta}}=Y(x)-\hat{Y}_{\boldsymbol{\beta}}(x)$.

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

   The theoretical regression parameter $\boldsymbol{\beta}_0$ minimizes $\int_F(Y(x)-\boldsymbol{Z}^t(x)\boldsymbol{\beta})^2dx$. It satisfies the normal equation $\big(\int_F\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)dx\big)\boldsymbol{\beta}_0=\int_F Y(x)\boldsymbol{Z}(x)dx$ and the orthogonality relationship $\int_F R_{\boldsymbol{\beta}_0}(x)\boldsymbol{Z}(x)dx=\mathbf{0}$, in particular the zero mean residual property $\frac{1}{\lambda(F)}\int_F R_{\boldsymbol{\beta}_0}(x)dx=0$.

2. The reduced model M_1

   [48] $$Y(x)=\boldsymbol{Z}^{(1)t}(x)\boldsymbol{\alpha}+R_{1,\boldsymbol{\alpha}}(x)=:\hat{Y}_{1,\boldsymbol{\alpha}}(x)+R_{1,\boldsymbol{\alpha}}(x)$$

   The theoretical regression parameter $\boldsymbol{\alpha}_0$ minimizes $\int_F(Y(x)-\boldsymbol{Z}^{(1)t}(x)\boldsymbol{\alpha})^2dx$. It satisfies the normal equation $\big(\int_F\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)t}(x)dx\big)\boldsymbol{\alpha}_0=\int_F Y(x)\boldsymbol{Z}^{(1)}(x)dx$ and the orthogonality relationship $\int_F R_{1,\boldsymbol{\alpha}_0}(x)\boldsymbol{Z}^{(1)}(x)dx=\mathbf{0}$, in particular the zero mean residual property $\frac{1}{\lambda(F)}\int_F R_{1,\boldsymbol{\alpha}_0}(x)dx=0$.

3.3 The three-phase one-stage generalized regression estimator

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

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_{s_k} &= \Big(\frac{1}{n_k}\sum_{x\in s_k}\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)t}(x)\Big)^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}^{(1)}(x)\\ &:= (\boldsymbol{A}_k^{(1)})^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}^{(1)}(x) =: (\boldsymbol{A}_k^{(1)})^{-1}\boldsymbol{U}_k^{(1)}, \quad k=0,1,2\end{aligned} \tag{[49]}$$

Likewise for the large large model we set

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{s_k} &= \Big(\frac{1}{n_k}\sum_{x\in s_k}\boldsymbol{Z}(x)\boldsymbol{Z}^{t}(x)\Big)^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}(x)\\ &= \boldsymbol{A}_k^{-1}\frac{1}{n_k}\sum_{x\in s_k}Y(x)\boldsymbol{Z}(x) =: \boldsymbol{A}_k^{-1}\boldsymbol{U}_k, \quad k=0,1,2\end{aligned} \tag{[50]}$$

Note that only $\hat{\boldsymbol{\alpha}}_{s_2}$ and $\hat{\boldsymbol{\beta}}_{s_2}$ are observable, because $Y(x)$ is only available at $x \in s_2$, and that in general the vector consisting of the first p components of $\hat{\boldsymbol{\beta}}_{s_2}$ is not equal to $\hat{\boldsymbol{\alpha}}_{s_2}$. The empirical predictions and residuals are defined, in a slightly simplified notation, as

$$\begin{aligned}\hat{Y}(x) &= \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}_{s_2}\\ \hat{Y}_1(x) &= \boldsymbol{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}_{s_2}\\ \hat{R}(x) &= Y(x)-\hat{Y}(x)\\ \hat{R}_1(x) &= Y(x)-\hat{Y}_1(x)\end{aligned}$$

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

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} &= \boldsymbol{A}_2^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\Big)\boldsymbol{A}_2^{-1} \\ \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} &= (\boldsymbol{A}_1^{(1)})^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}_1^2(x)\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)t}(x)\Big)(\boldsymbol{A}_1^{(1)})^{-1} \end{aligned} \tag{51}$$

Note that the matrix $\boldsymbol{A}_1^{(1)}$ is used for $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}$ instead of $\boldsymbol{A}_2^{(1)}$, both being asymptotically equivalent. This is necessary in order to have compatibility with the g-weight version of the estimated variance of the regression estimators discussed below. We need the following notation for the various means of the auxiliary vectors

$$\begin{aligned} \bar{\boldsymbol{Z}}^{(1)} &= \frac{1}{\lambda(F)}\int_F \boldsymbol{Z}^{(1)}(x)dx \\ \hat{\bar{\boldsymbol{Z}}}_0^{(1)} &= \frac{1}{n_0}\sum_{x\in s_0}\boldsymbol{Z}^{(1)}(x) \\ \hat{\bar{\boldsymbol{Z}}}_1^{(1)} &= \frac{1}{n_1}\sum_{x\in s_1}\boldsymbol{Z}^{(1)}(x) \\ \hat{\bar{\boldsymbol{Z}}}_k &= \frac{1}{n_k}\sum_{x\in s_k}\boldsymbol{Z}(x),\ k=1,2 \end{aligned} \tag{52}$$

The g-weights are defined as

$$\begin{aligned} g_1^{(1)}(x) &= \hat{\bar{\boldsymbol{Z}}}_0^{t(1)}(\boldsymbol{A}_1^{(1)})^{-1}\boldsymbol{Z}^{(1)}(x) \\ g_2(x) &= \hat{\bar{\boldsymbol{Z}}}_1^t\boldsymbol{A}_2^{-1}\boldsymbol{Z}(x) \end{aligned} \tag{53}$$

Note that $\boldsymbol{A}_1^{(1)}$ instead of $\boldsymbol{A}_2^{(1)}$ is used in the definition of $g_1^{(1)}(x)$. The g-weights satisfy the calibration properties

[54]
$$\begin{aligned}\frac{1}{n_1}\sum_{x\in s_1} g_1^{(1)}(x)\boldsymbol{Z}^{(1)}(x) &= \hat{\bar{\boldsymbol{Z}}}_0^{(1)} \\ \frac{1}{n_2}\sum_{x\in s_2} g_2(x)\boldsymbol{Z}(x) &= \hat{\bar{\boldsymbol{Z}}}_1\end{aligned}$$

The mean of the g-weights over s_1 and s_2 are equal to 1. We have already proved that the $g_2(x)$ tend to 1. With an obvious modification of the proof it can be shown that this also holds for $g_1^{(1)}(x)$.

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

[55]
$$\begin{aligned}\hat{Y}_{greg} &= \frac{1}{\lambda(F)}\int_F \hat{Y}_1(x)dx + \frac{1}{n_1}\sum_{x\in s_1}(\hat{Y}(x)-\hat{Y}_1(x)) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x)-\hat{Y}(x)) \\ &= (\bar{\boldsymbol{Z}}^{(1)} - \hat{\bar{\boldsymbol{Z}}}_1^{(1)})^t\hat{\boldsymbol{\alpha}}_{s_2} + \hat{\bar{\boldsymbol{Z}}}_1^t\hat{\boldsymbol{\beta}}_{s_2}\end{aligned}$$

Here, $\boldsymbol{Z}^{(1)}(x)$ is exhaustive, i.e. known at all $x \in F$, though in practice it suffices to know the true mean $\bar{\boldsymbol{Z}}^{(1)}$.

Using the results given in Mandallaz et al. (2013) (see Mandallaz (2013b) for the proofs) one has under the **external model assumption** the variance estimate

[56]
$$\hat{\mathbb{V}}(\hat{Y}_{greg}) = \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}\hat{R}_1^2(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)$$

Using the estimated design-based variance-covariance matrices one obtains alternatively the consistent estimate

[57]
$$\hat{\mathbb{V}}(\hat{Y}_{greg}) = \frac{n_2}{n_1}\bar{\boldsymbol{Z}}^{(1)t}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}^{(1)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_1^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_1$$

which can be rewritten in the g-weight form

[58] $$\hat{\mathbb{V}}(\hat{Y}_{greg}) = \frac{1}{n_1 n_2}\sum_{x\in s_2}(g_1^{(1)}(x))^2\hat{R}_1^2(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2^2}\sum_{x\in s_2} g_2^2(x)\hat{R}^2(x)$$

Because the g-weights tend to 1 the variance estimates [56] and [58] are asymptotically equivalent.

If $\boldsymbol{Z}^{(1)}(x)$ is no longer exhaustive we replace the first term in [55] by its sample mean in the null phase sample and we define the new three-phase estimator as

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

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

Note that the third terms in [55] and [59] vanish for internal models because by assumption the intercept term is contained in $\boldsymbol{Z}^{(1)}$ (or it can be expressed as a linear combination of its components) so that the residuals sum up to zero over s_2.

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

[60] $$\begin{aligned}\mathbb{V}_{0,1,2}(\hat{Y}_{g3reg}) &= \mathbb{E}_0\mathbb{E}_{1|0}\mathbb{V}_{2|0,1}(\hat{Y}_{F,g3reg}) + \mathbb{E}_0\mathbb{V}_{1|0}\mathbb{E}_{2|0,1}(\hat{Y}_{F,g3reg}) + \mathbb{V}_0\mathbb{E}_{1|0}\mathbb{E}_{2|0,1}(\hat{Y}_{F,g3reg}) \\ &= \frac{1}{n_0}\mathbb{V}_x(Y(x)) + (1-\frac{n_1}{n_0})\frac{1}{n_1}\mathbb{V}_x(R_1(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}_x(R(x))\end{aligned}$$

The above equation holds for any fixed given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. At the true value $\boldsymbol{\alpha}_0$ the predictions $\hat{Y}_1(x)$ and the residuals $R_1(x)$ are orthogonal in the design-based sense, which implies $\mathbb{V}_x(Y(x)) = \mathbb{V}_x(\hat{Y}_1(x)) + \mathbb{V}_x(R_1(x))$ and the following external variance

[61] $$\mathbb{V}_{0,1,2}(\hat{Y}_{g3reg}) = \frac{1}{n_0}\mathbb{V}_x(\hat{Y}_1(x)) + \frac{1}{n_1}\mathbb{V}_x(R_1(x)) + (1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R(x))$$

As the empirical residuals based on $\hat{\boldsymbol{\alpha}}_{s_2}$ and $\hat{\boldsymbol{\beta}}_{s_2}$ have zero mean we get the external variance estimate

$$[62] \quad \hat{\mathbb{V}}_{ext}(\hat{Y}_{g3reg}) = \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}_1(x)-\hat{\bar{Y}}_1)^2}{n_0-1} + \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}\hat{R}_1^2(x) + (1-\frac{n_2}{n_1})\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)$$

One can also derive the asymptotic theoretical variance via the design-based covariance matrices, which gives

$$[63] \qquad \mathbb{V}_{0,1,2}(\hat{Y}_{g3reg}) = \frac{1}{n_0}\mathbb{V}(\hat{Y}_{1,\boldsymbol{\alpha}}(x)) + \frac{n_2}{n_1}\bar{\boldsymbol{Z}}^{(1)t}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}^{(1)} + (1-\frac{n_2}{n_1})\bar{\boldsymbol{Z}}^t\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}}\bar{\boldsymbol{Z}}^t$$

which can be consistently estimated by

$$[64] \qquad \hat{\mathbb{V}}_{0,1,2}(\hat{Y}_{g3reg}) = \hat{\boldsymbol{\alpha}}_{s_2}^t\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_0^{(1)}}\hat{\boldsymbol{\alpha}}_{s_2} + \frac{n_2}{n_1}\hat{\bar{\boldsymbol{Z}}}_0^{(1)t}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_0^{(1)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_1^t\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_1^t$$

with the design-based covariance matrix of the first component

$$[65] \qquad \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_0^{(1)}} = \frac{1}{n_0}\frac{\sum_{x\in s_0}(\boldsymbol{Z}^{(1)}(x)-\hat{\bar{\boldsymbol{Z}}}_0^{(1)})(\boldsymbol{Z}(x)-\hat{\bar{\boldsymbol{Z}}}_0^{(1)})^t}{n_0-1}$$

The proof of [63] is rather intricate and given in Mandallaz (2013b). One can rewrite [64] as

$$[66] \qquad \begin{aligned} \hat{\mathbb{V}}(\hat{Y}_{g3reg}) &= \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}_1(x)-\hat{\bar{Y}}_1)^2}{n_0-1} \\ &+ \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}(g_1^{(1)}(x))^2\hat{R}_1^2(x) + (1-\frac{n_2}{n_1})\frac{1}{n_2^2}\sum_{x\in s_2}g_2^2(x)\hat{R}^2(x) \end{aligned}$$

where $\hat{\bar{Y}}_1 = \frac{1}{n_0}\sum_{x\in s_0}\hat{Y}_1(x)$.

Because the g-weights tend to 1 [66] is asymptotically equivalent to the external variance estimate given in [62].

3.4 The three-phase two-stage generalized regression estimator

For the point estimate one simply replaces $Y(x)$ by $Y^*(x)$ in [50] to obtain $\hat{\boldsymbol{\alpha}}^*_{s_2}$, $\hat{\boldsymbol{\beta}}^*_{s_2}$. The predictions and residuals are defined as

$$
\begin{aligned}
\hat{Y}^*(x) &= \boldsymbol{Z}^t(x)\hat{\boldsymbol{\beta}}^*_{s_2} \\
\hat{Y}_1^*(x) &= \boldsymbol{Z}^{(1)t}(x)\hat{\boldsymbol{\alpha}}^*_{s_2} \\
\hat{R}^*(x) &= Y^*(x) - \hat{Y}^*(x) \\
\hat{R}_1^*(x) &= Y^*(x) - \hat{Y}_1^*(x)
\end{aligned}
$$

The tree-phase two-stage regression estimator reads

$$
\begin{aligned}
\hat{Y}^*_{g3reg} &= \frac{1}{n_0}\sum_{x\in s_0}\hat{Y}_1^*(x) + \frac{1}{n_1}\sum_{x\in s_1}(\hat{Y}^*(x) - \hat{Y}_1^*(x)) + \frac{1}{n_2}\sum_{x\in s_2}(Y^*(x) - \hat{Y}^*(x)) \\
\text{[67]} \qquad &= (\hat{\hat{\boldsymbol{Z}}}^{(1)}_0 - \hat{\hat{\boldsymbol{Z}}}^{(1)}_1)^t\hat{\boldsymbol{\alpha}}^*_{s_2} + \hat{\hat{\boldsymbol{Z}}}^t_1\hat{\boldsymbol{\beta}}^*_{s_2}
\end{aligned}
$$

Since $\hat{\boldsymbol{\alpha}}^*_{s_2} = \boldsymbol{\alpha}_0 + O_p(\frac{1}{n_2})$ and $\hat{\boldsymbol{\alpha}}^*_{s_2} = \boldsymbol{\beta}_0 + O_p(\frac{1}{n_2})$ the external variance assumption is asymptotically valid. To calculate the overall variance we use the standard decomposition and [60]

$$
\begin{aligned}
\mathbb{V}_{0,1,2,3}(\hat{Y}^*_{g3reg}) &= \mathbb{V}_{0,1,2}\mathbb{E}_{3|0,1,2}(\hat{Y}^*_{g3reg}) + \mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg}) \\
&= \mathbb{V}_{0,1,2}(\hat{Y}_{g3reg}) + \frac{1}{n_2}\mathbb{E}_x(V(x)) \\
\text{[68]} \qquad &= \frac{1}{n_0}\mathbb{V}_x(\hat{Y}_1(x)) + \frac{1}{n_1}\mathbb{V}_x(R_1(x)) + (1 - \frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R(x)) + \frac{1}{n_2}\mathbb{E}_x(V(x))
\end{aligned}
$$

To obtain an estimate of the external variance we adapt [62]

$$\begin{aligned} \hat{\mathbb{V}}_{ext}(\hat{Y}^*_{g3reg}) &= \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}^*_1(x)-\hat{\bar{\hat{Y}}}{}^*_1)^2}{n_0-1} \\ &+ \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}(\hat{R}^*_1(x))^2+(1-\frac{n_2}{n_1})\frac{1}{n_2^2}\sum_{x\in s_2}(\hat{R}^*(x))^2 \end{aligned} \tag{[69]}$$

As in two-phase sampling we use the asymptotically valid relations

$$\begin{aligned} \mathbb{E}_{3|0,1,2}(\hat{R}^*(x))^2 &= R^2(x)+V(x) \\ \mathbb{E}_{3|0,1,2}(\hat{R}^*_1(x))^2 &= R^2_1(x)+V(x) \end{aligned} \tag{[70]}$$

Hence, we get for the expected value of the external variance estimate

$$\begin{aligned} \mathbb{E}_{0,1,2,3}\hat{\mathbb{V}}_{ext}(\hat{Y}^*_{g3reg}) &= \frac{1}{n_0}\mathbb{V}(\hat{Y}_1(x))+\frac{1}{n_1}\mathbb{V}(R_1(x))+(1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R(x)) \\ &+ \frac{1}{n_1}\mathbb{E}_xV(x)+(1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{E}_xV(x) \\ &= \frac{1}{n_0}\mathbb{V}(\hat{Y}_1(x))+\frac{1}{n_1}\mathbb{V}(R_1(x))+(1-\frac{n_2}{n_1})\frac{1}{n_2}\mathbb{V}(R(x))+\frac{1}{n_2}\mathbb{E}_xV(x) \\ &= \mathbb{V}_{ext}(\hat{Y}_{g3reg})+\frac{1}{n_2}\mathbb{E}_xV(x) \end{aligned} \tag{[71]}$$

which is precisely [68]. Hence, the external variance estimate [] is asymptotically unbiased. We can also derive a g-weight version of the variance of $\hat{Y}^*_{g3reg}$ by using

$$\begin{aligned} \mathbb{V}(\hat{Y}^*_{g3reg}) &= \mathbb{V}_{0,1,2}\mathbb{E}_{3|0,1,2}(\hat{Y}^*_{g3reg})+\mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg}) \\ &= \mathbb{V}_{0,1,2}(\hat{Y}_{g3reg})+\mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg}) \\ &= \frac{1}{n_0}\mathbb{V}(\hat{Y}_{1,\boldsymbol{\alpha}_2}(x))+\frac{n_2}{n_1}\bar{\boldsymbol{Z}}^{(1)t}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}^{(1)}+(1-\frac{n_2}{n_1})\bar{\boldsymbol{Z}}^t\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}}\bar{\boldsymbol{Z}}^t+\mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg}) \end{aligned}$$

The increase in variance due to two-stage sampling is therefore $\mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg})$. Recalling that the Poisson sampling schemes at different points are independent we get for the extra term

$$
\begin{aligned}
\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg}) &= \mathbb{E}_{3|0,1,2}\Big((\hat{\bar{\boldsymbol{Z}}}^{(1)}_0 - \hat{\bar{\boldsymbol{Z}}}^{(1)}_1)^t(\hat{\boldsymbol{\alpha}}^*_{s_2} - \hat{\boldsymbol{\alpha}}_{s_2}) + \hat{\bar{\boldsymbol{Z}}}^{(1)}_1(\hat{\boldsymbol{\beta}}^*_{s_2} - \hat{\boldsymbol{\beta}}_{s_2})\Big)^2 \\
&= (\hat{\bar{\boldsymbol{Z}}}^{(1)}_0 - \hat{\bar{\boldsymbol{Z}}}^{(1)}_1)^t(\boldsymbol{A}^{(1)}_2)^{-1}\big(\frac{1}{n_2^2}\sum_{x\in s_2} V(x)\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}^{(1)}(x)\big)(\boldsymbol{A}^{(1)}_2)^{-1}(\hat{\bar{\boldsymbol{Z}}}^{(1)}_0 - \hat{\bar{\boldsymbol{Z}}}^{(1)}_1) \\
&+ 2(\hat{\bar{\boldsymbol{Z}}}^{(1)}_0 - \hat{\bar{\boldsymbol{Z}}}^{(1)}_1)^t(\boldsymbol{A}^{(1)}_2)^{-1}\big(\frac{1}{n_2^2}\sum_{x\in s_2} V(x)\boldsymbol{Z}^{(1)}(x)\boldsymbol{Z}(x)\big)\boldsymbol{A}_2^{-1}\hat{\bar{\boldsymbol{Z}}}_1 \\
&+ \hat{\bar{\boldsymbol{Z}}}_1\boldsymbol{A}_2^{-1}\big(\frac{1}{n_2^2}\sum_{x\in s_2} V(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x)\big)\boldsymbol{A}_2^{-1}\hat{\bar{\boldsymbol{Z}}}_1
\end{aligned}
$$

Taking the expectation $\mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg})$ we see that the first and second term are of smaller order than the third, so that we get asymptotically for the increase in variance

[72] $$\mathbb{E}_{0,1,2}\mathbb{V}_{3|0,1,2}(\hat{Y}^*_{g3reg}) = \bar{\boldsymbol{Z}}^t\boldsymbol{A}^{-1}\big(\frac{1}{n_2}\mathbb{E}_x(V(x)\boldsymbol{Z}(x)\boldsymbol{Z}^t(x))\big)\boldsymbol{A}^{-1}\bar{\boldsymbol{Z}}$$

Writing its estimate with the g-weight $g(x)$ tending to 1 we see that the extra term is asymptotically equal to $\frac{1}{n_2}\mathbb{E}_xV(x)$, which is the same as the value obtained with the external variance in [71].

Intuitively it is appealing to use [63] with the matrices $\boldsymbol{\Sigma}_{\boldsymbol{\alpha}^*_2}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}^*_2}$ to get the variance estimate

$$
\begin{aligned}
\hat{\mathbb{V}}_{0,1,2,3}(\hat{Y}^*_{g3reg}) &= \hat{\boldsymbol{\alpha}}^{*t}_2\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}^{(1)}_0}\hat{\boldsymbol{\alpha}}^*_{s_2} + \frac{n_2}{n_1}\hat{\bar{\boldsymbol{Z}}}^{(1)t}_0\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}^*_{s_2}}\hat{\bar{\boldsymbol{Z}}}^{(1)}_0 + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}^t_1\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}^*_{s_2}}\hat{\bar{\boldsymbol{Z}}}^t_1 \\
&= \frac{1}{n_0}\frac{\sum_{x\in s_0}(\hat{Y}^*_1(x) - \hat{\bar{Y}}^*_1)^2}{n_0-1} \\
&+ \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}(g^{(1)}_1(x))^2(\hat{R}^*_1(x))^2 + (1-\frac{n_2}{n_1})\frac{1}{n_2^2}\sum_{x\in s_2} g_2^2(x)(\hat{R}^*(x))^2 \qquad [73]
\end{aligned}
$$

Using [70] we get

$$\begin{aligned}
\mathbb{E}_{1,2,3}\big(\hat{\mathbb{V}}(\hat{Y}^*_{g3reg})\big) &= \frac{1}{n_0}\mathbb{V}(\hat{Y}_1(x)) \\
&+ \mathbb{E}_{1,2}\Big(\frac{1}{n_1 n_2}\sum_{x\in s_2}(g_1^{(1)}(x))^2(R_1^2(x)+V(x))\Big) \\
&+ \mathbb{E}_{1,2}\Big(\frac{1}{n_2^2}\sum_{x\in s_2} g_2^2(x)(R^2(x)+V(x))\Big) \\
&- \mathbb{E}_{1,2}\Big(\frac{1}{n_1 n_2}\sum_{x\in s_2} g_2^2(x)(R^2(x)+V(x))\Big)
\end{aligned} \tag*{[74]}$$

Grouping the terms and since the g-weights tend to 1 we get

$$\begin{aligned}
\mathbb{E}_{1,2,3}\big(\hat{\mathbb{V}}(\hat{Y}^*_{g3reg})\big) &= \frac{1}{n_0}\mathbb{V}(\hat{Y}_1(x)) \\
&+ \mathbb{E}_{1,2}\frac{1}{n_1 n_2}\sum_{x\in s_2}((g_1^{(1)}(x))^2 - g_2^2(x))V(x) \\
&+ \frac{n_2}{n_1}\bar{\boldsymbol{Z}}^{(1)t}\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}^{(1)} + (1-\frac{n_2}{n_1})\bar{\boldsymbol{Z}}^t\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_{s_2}}\bar{\boldsymbol{Z}}^t \\
&+ \frac{1}{n_2}\mathbb{E}_x(V(x)) \\
&= \mathbb{V}(\hat{Y}_{g3reg}) + \frac{1}{n_2}\mathbb{E}_x(V(x))
\end{aligned} \tag*{[75]}$$

We have proved that in three-phase two-stage sampling the external variance estimate [3.4] and the g-weight variance estimate [73] are asymptotically consistent and equivalent. **One can obtain the results for two-phase two-stage and three-phase two-stage sampling by replacing in all formulae the local density $Y(x)$ by the general local density $Y^*(x)$ in all the corresponding formulae valid for two-phase one-stage and three-phase one-stage sampling schemes.**

The remarks on two-phase systematic sampling remain a fortiori valid for three-phase systematic sampling: the impact of the long range spatial correlation of the $Y(x)$ or $\hat{Y}(x)$ being smaller than under two-phase sampling ($\frac{1}{n_0} << \frac{1}{n_1}$), and because the residuals $R(x)$,

$R_1(x)$ have a spatial correlation usually very small in comparison to F, likewise for the better statistical properties of the g-weight variance estimates (see Mandallaz (2013b) pp. 24-29 for a stratification example with exhaustive first-phase).

4 Concluding remarks

It is clear that the same results will apply to the estimation of a small area $G \subset F$: for the external variance this is trivially achieved by restricting the samples to the small area, whereas in the g-weight approach one extends the reduced model by the indicator variable of G (thus ensuring zero mean residuals over G) and replaces the empirical means of the auxiliary variables over F by their means over $G \subset F$. The resulting g-weights $g^{(1)}(x)$, respectively $g_2(x)$, tend to 0 for $x \notin G$ and to $\frac{n_1}{n_{1,G}}$, respectively $\frac{n_2}{n_{2,G}}$, for $x \in G$.

The results extend trivially to cluster sampling if the number of plots per cluster is constant. The ideas are the same but the algebra is slightly more cumbersome if this not the case (see Mandallaz (2008) section 4.5 to get the flavor of the required modifications).

Formal proof for two/three phase one-stage sampling schemes are given in the technical reports Mandallaz (2012), Mandallaz (2013c) and Mandallaz (2013b) and it should be clear from the arguments given in this technical report that the same results will also hold in two/three phase two-stage cluster sampling: details are left as an exercise for the reader. Interestingly, Feynman did not comment this lame excuse, also over-used in mathematical books or papers!

Current work is devoted to the development of geostatistical Kriging versions of the three-phase design-based estimators.

References

Gregoire, T. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Can.J.For.Res.*, **28**:pp. 1429–1447.

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

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

Mandallaz, D. (1991). *A Unified Approach to Sampling Theory for Forest Inventory Based on Infinite Population Models*. Ph.D. thesis, ETH Zurich, Chair of Forest Inventory and Planning, http://e-collection.library.ethz.ch/.

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.ethb.ethz.ch.

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

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

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

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

Massey, A. (2011). Comparison of Estimators in One-Phase Two-Stage Poisson Sampling in Forest Inventories. Technical report, ETH Zurich, Department of Mathematics, MSc Thesis, http://e-collection.library.ethz.ch.

Matheron, G. (1965). *Les Variable Régionalisées et leur Estimation.* Masson, Paris.

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