

The R Package `forestinventory`: Design-Based Global and Small Area Estimations for Multi-Phase Forest Inventories

Andreas Hill

Alexander Massey

Abstract

Forest inventories provide reliable evidence-based information to assess the state and development of forests over time. They typically consist of a random sample of plot locations in the forest that are assessed individually by hand. Due to the high costs of these terrestrial campaigns, remote sensing information available in high quantity and low costs (e.g. LiDAR, stereophotogrammetry, etc.) is frequently incorporated in the estimation process in order to reduce inventory costs or improve estimation precision. With respect to this objective, the application of *multi-phase* forest inventory methods (e.g. double- and triple-sampling regression estimators) has proved to be efficient. While these methods have been successfully applied in practice, the availability of open-source software has been rare if not non-existent. The R package `forestinventory` provides a comprehensive set of global and small area regression estimators for multiphase forest inventories under simple and cluster sampling. The implemented methods have been demonstrated in various scientific studies ranging from small to large scale forest inventories, and can be used for post-stratification, regression and regression within strata. This article summarizes the mathematical theory of this family of design-based estimators and demonstrates their application in the R environment.

Keywords: forest inventory, design-based, infinite population approach, two- and three-phase sampling, regression estimators, small area estimation.

1. Introduction

In many countries, forest inventories have become an indispensable tool for evaluating the current state of forests as well as for tracking their development over time. They provide accurate quantitative information that can be used to define management actions and to adapt forest management strategies according to guidelines on national and international levels. As conducting a full census of all trees within any sizable forest area is clearly infeasible due to time and cost restrictions, forest inventories usually gather their information by means of statistical sampling methods. Typically this means that discrete sample locations (*sample plots*) are randomly chosen in the forest, making up the framework of a *terrestrial inventory*. This terrestrial sample data is then used to make estimates for the entire forested area and provide a measure of precision for those estimates in the form of confidence intervals. There is a broad range of literature describing the concepts and methods regarding the choice of different estimators under various sample designs (Gregoire and Valentine 2007; Köhl, Magnussen, and Marchetti 2006; Schreuder, Gregoire, and Wood 1993; Mandallaz 2008).

Terrestrial inventories have the benefit of being very flexible in the sense that they can be used to produce high quality estimates for a wide-variety of different forest attributes. However, they have the downside of being very expensive. Improving the precision of the estimates by increasing the number of sample plots essentially means that travel costs will rise as trained inventors are sent to more and more plot locations. This is why the number of terrestrial samples is often limited. Although national

inventories usually possess a sufficiently large terrestrial sample size to provide high estimation accuracies for larger areas, this is often not the case for smaller areas, such as forest management units. As a result, there has been an increasing need for alternative inventory methods that can maintain the same estimation precision at lower costs, or achieve higher estimation precision at identical costs (von Lüpke 2013). A method which has become particularly attractive is called *multi-phase sampling*. The core concept is to enlarge the sample size in order to gain higher estimation precision *without* enlarging the terrestrial sample size. This is done by using predictions of the terrestrial target variable at additional sample locations where the terrestrial information has not been gathered. These predictions are produced by regression models that use explanatory variables derived from auxiliary data, commonly in the form of spatially exhaustive remote sensing data in the inventory area. Regression estimators using this concept can consider either *one* additional sample of plot locations consisting of auxiliary information (two-phase or double-sampling) or *two* additional samples available in different sample sizes (three-phase or triple-sampling) (Gregoire and Valentine 2007; Saborowski, Marx, Nagel, and Böckmann 2010; Mandallaz 2013a,d; von Lüpke, Hansen, and Saborowski 2012). Their application to existing forest inventory systems have already showed their efficiency in terms of cost reduction and gain in estimation precision (Breidenbach and Astrup 2012; von Lüpke and Saborowski 2014; Mandallaz, Breschan, and Hill 2013; Magnussen, Mandallaz, Breidenbach, Lanz, and Ginzler 2014; Massey, Mandallaz, and Lanz 2014).

Despite the broad range of methodological literature and case studies, freely available software supporting the application of the proposed multi-phase sampling methods to forestry has been rare. Two exceptions are the R package **JoSAE** by Breidenbach (2015) that provides the GREG (Särndal, Swenson, and Wretman 2003) and EBLUP (Battese, Harter, and Fuller 1988) two-phase small area estimator for simple sampling derived under the *finite population approach*, and the package **maSAE** by Cullmann (2016) which provided an initial implementation of the extended synthetic estimator under two and three-phase sampling (Mandallaz *et al.* 2013; Mandallaz 2013d). However, a more comprehensive software package covering a larger variety of sample designs and estimators for forest inventories has not yet been available, which is the motivation behind the R package **forestinventory**. The package provides global and small area estimators for two-phase and three-phase forest inventories under simple and cluster sampling, which have been developed under the *infinite population approach* by Daniel Mandallaz at ETH Zurich between 2008 and 2017. The implemented methods have been demonstrated by case studies in Switzerland (Massey *et al.* 2014; Massey and Mandallaz 2015b; Mandallaz *et al.* 2013) and Germany (Hill, Mandallaz, Buddenbaum, Stoffels, and Langshausen 2017). The implemented estimators cover 32 inventory scenarios and can be used for post-stratification, regression and regression within strata (Massey 2015). The long-term objective of **forestinventory** is to make the broad range of estimators available to a large user community and to facilitate their application in science as well as operational forest management.

The objectives of this article are to **a)** establish the link between the mathematical description of the estimators and their implementation in our package, **b)** illustrate their application to real-world inventory data sets and **c)** address special cases and demonstrate how the package-functions handle them.

2. Methods and Structure of the Package

2.1. Forest Inventory in the Infinite Population Approach

Most forest inventories gather terrestrial information by sending field crews to randomly (or systematically) selected points in the forest area defined by GPS coordinates. The crew then defines a sample plot by measuring individual trees located within one or multiple constructed inclusion circles around the sample point x , and aggregating their characteristics (e.g. timber volumes) to local plot densities (e.g. the timber density in m^3/ha). The estimators implemented in **forestinventory** use the so called *infinite population approach* in order to bridge this inventory process to the mathematics behind the estimators. This approach assumes that the spatial distribution of the local density, $Y(x)$, over the forest area is determined by a fixed piecewise constant function, as visualized in Fig. 1. The *population total* of the target variable (e.g. the total timber volume of the forest) is mathematically equivalent to the integral of that density function, which is depicted in Fig. 1 as the volume under the density surface. From this perspective, the practical challenge is that the form of this function is unknown. Theoretically, we could get the total timber volume by observing the function value, i.e. the local density $Y(x)$, at each possible point x over the forest area and taking their sum. However, this is impossible because there is an infinite number of points in the forest area. Our strategy is thus to take a sample of points, s_2 , from an infinite population of possible points and use their associated local densities, i.e. $Y(x)$, to estimate the integral $Y = \frac{1}{\lambda(F)} \int_F Y(x) dx$ with $\hat{Y} = \frac{1}{n_2} \sum_{x \in s_2} Y(x)$. The total timber volume can then be obtained by multiplying \hat{Y} by the surface area of the forest, $\lambda(F)$. All estimators included in **forestinventory** rest upon this theoretical perspective. The key point in the *infinite population approach* is that a local density value $Y(x)$ is associated with the sample point x , which constitutes the sample unit, and not with the sample plot area. This has some theoretical advantages over the *finite population approach*, where the sample units are the actual plot areas usually assumed to be either circular or rectangular. This is mainly due to the impossibility of a perfect tessellation over an amorphous forest area by any choice of plot shape. Hence, the population in the finite approach is, strictly speaking, not well defined with respect to the forest area. The consideration of an underlying infinite population of sample points will also play an important role when incorporating auxiliary information in the frame of two- and three-phase estimation methods.

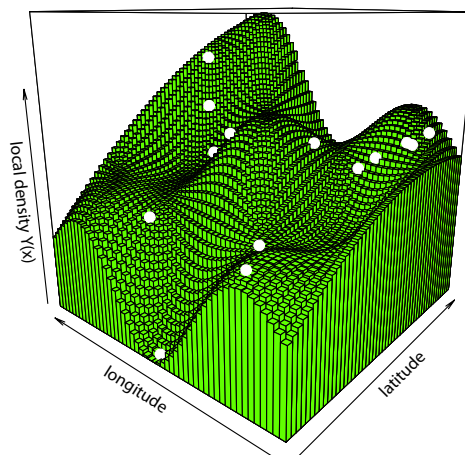


Figure 1: Artificial representation of a local density surface. The spatial distribution of a hypothetical density function for every point in a forested area is represented as a wavy piecewise constant green surface. Sample plots (white dots) inform the inventorist of the value of the density function at that point. Note that the plateaus of constant $Y(x)$ values here have the shape of squares whereas in reality they are likely to be formed by the intersection of circles around trees.

2.2. Two-Phase Sampling

The two-phase or double-sampling estimators use inventory information from **two** nested samples which are commonly referred to as *phases* (Fig. 2a). The first phase s_1 comprises n_1 sample locations that each provide a set of explanatory variables described by the column vector $\mathbf{Z}(x) = (z(x)_1, z(x)_2, \dots, z(x)_p)^t$ at each point $x \in s_1$. These explanatory variables are derived from auxiliary information that is available in high quantity within the forest area F . The second phase s_2 constitutes the terrestrial inventory conducted at n_2 subsamples of the large phase s_1 and provides the value of the target variable, i.e. the local density $Y(x)$ (e.g. the timber volume per hectare). For every $x \in s_1$, $\mathbf{Z}(x)$ is transformed into a prediction $\hat{Y}(x)$ of $Y(x)$ using the choice of some model, which in **forestinventory** is always a linear model fit in s_2 using ordinary least squares (OLS). The basic idea of this setup is to boost the sample size by providing a large sample of less precise but cheaper predictions of $Y(x)$ in s_1 and to correct any possible model bias, i.e. $\mathbb{E}(Y(x) - \hat{Y}(x))$, using the subsample of terrestrial inventory units where the value of $Y(x)$ is observed. In the design-based context, the two-phase estimator is typically unbiased regardless of the model used to produce the predictions. This property comes from the assumption that each phase's sample is selected via simple random sampling (see Section 2.5).

2.3. Three-Phase Sampling

Three-phase estimators extend the principle of two-phase sampling and use inventory information from **three** nested samples (*phases*) (Fig. 2a). The basic setup is that the explanatory variables calculated from the auxiliary information are available in two different frequencies. The phase s_0 provides a large number n_0 of auxiliary data, whereas the phase s_1 provides additional auxiliary data that are only available at n_1 subsamples of s_0 . The terrestrial information is then collected at a further subsample s_2 of s_1 . The motivation for three-phase sampling is that the additional set of explanatory variables available at s_1 , now denoted $\mathbf{Z}^{(1)}(x)$, adds considerable explanatory power to the set of variables available at all sample locations $x \in s_0$, denoted $\mathbf{Z}^{(0)}(x)$. From that it follows that we can define *two nested* regression models. The *full* set of predictor variables $\mathbf{Z}^t(x) = (\mathbf{Z}^{(0)t}(x), \mathbf{Z}^{(1)t}(x))$ can be used to calculate the predictions $\hat{Y}(x)$ of $Y(x)$ at all sample locations $x \in s_1$. The regression model applicable to the s_1 phase is thus referred to as the *full model*. Less accurate predictions, $\hat{Y}^{(0)}(x)$, can be produced at all the sample locations $x \in s_0$ using only the reduced set of explanatory variables $\mathbf{Z}^{(0)}(x)$. If there is a significant gain in model precision between the reduced and the full model and the sampling fraction between s_0 and s_1 is sufficiently large, the three-phase estimator normally leads to a further increase in estimation precision compared to the two-phase estimator.

2.4. Small Area Estimation

Small area estimation does not necessarily refer to small spatial areas but rather to areas that contain little or no terrestrial sample. To formulate this mathematically, we want to make an estimate for a subregion G of the entire inventory area F (Fig. 2b). As the sample size in the small area, $n_{2,G}$, is usually too small to provide sufficient estimation precision, multi-phase estimation can be efficient. However, $n_{2,G}$ may also be too small to justify fitting a separate regression model just for that area because the estimates produce undesirably large confidence intervals. The idea is then to *borrow strength* from the entire terrestrial sample s_2 of F to fit the model, and to apply this model to the small area. The potential bias of applying that model in G is then corrected for by using the empirical model residuals derived from that small area. If there are no terrestrial plots in G (i.e. $n_{2,G} = 0$), one cannot correct for a potential model bias in G and has to accept a potential bias in the estimator. These are called *synthetic* estimates and despite their potential bias, it is usually still possible to calculate their design-based variance.

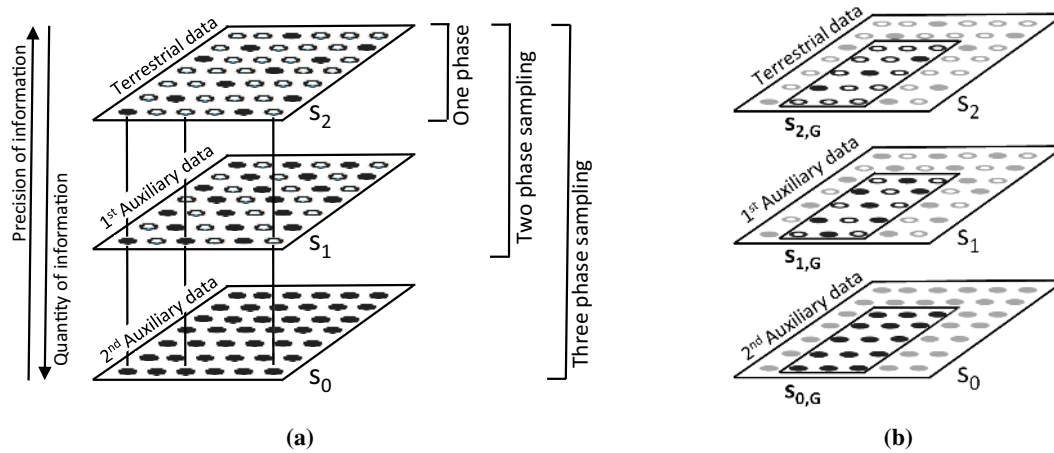


Figure 2: (a) Concept of multiphase sampling. The square represents the forest area for which an inventory is being conducted. The points denote the sample locations x . Filled points indicate *available* information. (b) Illustration of the Small Area Estimation problem

2.5. Design-Based vs. Model-Dependent Approach

The subject of model selection gets a lot of attention in the field of forest inventory. This is why it is important to understand that the mathematical interpretation of how a model is used to produce estimates is fundamentally different between the design-based and model-dependent approach. In the *model-dependent* (also known as *model-based*) framework, the sample locations x are fixed and the observation $Y(x)$ taken at location x is assumed to be a random variable as the forest is assumed to be the realization of a stochastic process. Although the model does not need to be fit from a probability sample, i.e. the sample locations could arbitrarily be chosen, the model should adequately describe the underlying stochastic process in order to efficiently ensure unbiased results. In practice this means that special attention must be made to ensure that the variable selection is appropriate to avoid overfitting, important variables are not omitted and all model assumptions are reasonably met through empirical verification. If a model is misspecified then estimation based on inference from that model may not be reliable. In the model-dependent framework one thus has to *trust* the model. In contrast, the *design-based* approach, on which all **forestinventory** estimators are based, rests upon the randomization of the sample locations x . While the sample locations x are independently and uniformly distributed in the forest, the forest itself and thus the values of the local densities $Y(x)$ at any location $x \in F$ are fixed and *not* the result of a stochastic process. A selected observation $Y(x)$ still remains a random variable, but solely due to the random sample mechanism. A consequence of this approach is that the estimation properties of design-based regression estimators (e.g. unbiasedness) typically hold regardless of the model that is chosen. The philosophy of the design-based approach is thus to use prediction models to *improve* the efficiency of the estimators without having to *rely* on their correct specification, which makes them very attractive to be used in official statistics. They are therefore also referred to as *model-assisted*. It should be noted that the randomization of sample locations upon which design-based inference depends, is in practice often replaced by systematic grids to minimize travelling costs in the terrestrial survey. However, there is reasonable evidence that softening this assumption is acceptable for point and variance estimation as long as the grid does not interact with periodic features in the forest structure (Mandallaz 2008). The variance will in most cases be slightly overestimated and lead to wider, more conservative confidence intervals (Mandallaz 2013a).

2.6. Package Structure

In the **forestinventory** package, estimators for two-phase and three-phase sampling are applied with the `twophase()` and `threephase()` functions. From these two overall function calls, various estimators for specific inventory scenarios under the chosen sampling design can be applied (Fig. 3). Choosing an estimator follows a tree-like structure which can serve the user as a guideline throughout this article as well as in future applications. The basic decision to make is whether an estimate and its variance should be computed for an entire inventory area (*global estimators*) or only for subregions of the entire inventory area (*small area estimators*). In the second case, the package offers three small area estimators that will in detail be described in the following sections. The estimators are available under *exhaustive* and *non-exhaustive* use of the auxiliary data. Additionally, the package also can calculate one-phase estimates solely based on terrestrial samples. All estimators are also available for *cluster* sampling, in which case a sample unit consists of multiple, spatially agglomerated samples. The following sections describe the mathematical details and the application of the multi-phase estimators implemented in the R package **forestinventory**. While Mandallaz (2008, 2013c,b, 2015) provides an extensive derivation of all estimators, we will provide the mathematical formulas that are actually implemented in the package. We will also restrict discussion to simple sampling, while the formulas for cluster sampling are available in the technical reports (Mandallaz, Hill, and Massey 2016; Mandallaz 2013c,b). A special case under cluster sampling is described in Section 6.

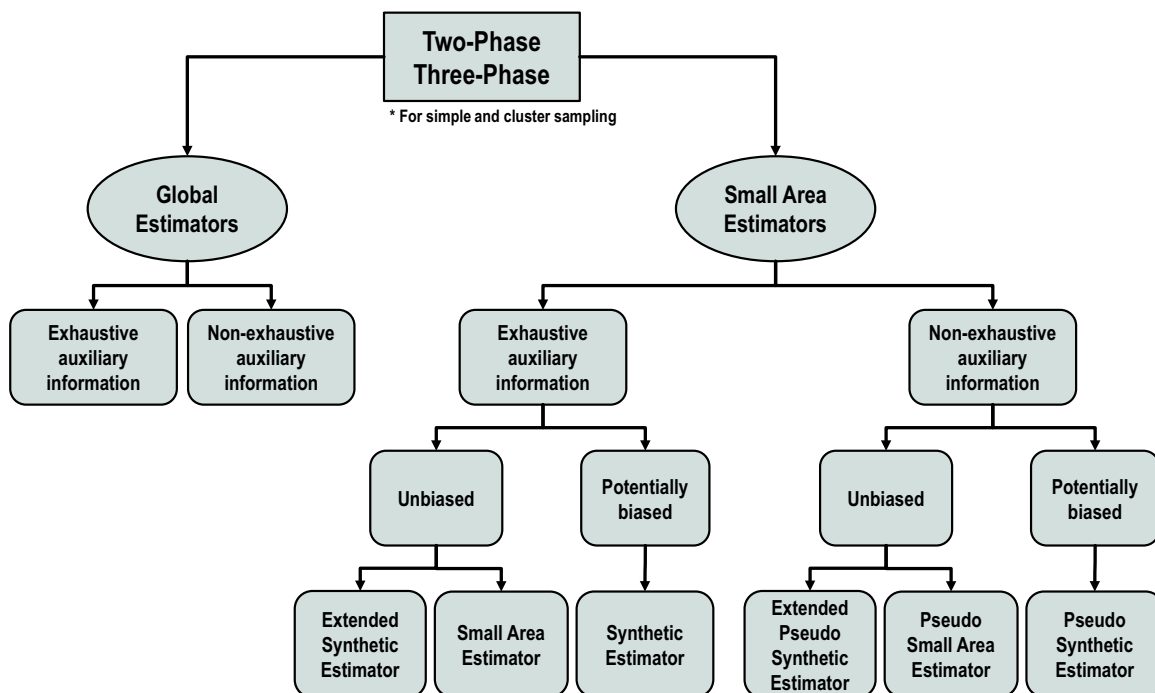


Figure 3: Structure of the multi-phase estimators in the R package **forestinventory**

3. Two-phase Estimators and their Application

3.1. Global Estimators

Mathematical Background

The vector of regression coefficients of the OLS regression model is found by using the following solution to the sample-based normal equation:

$$\hat{\boldsymbol{\beta}}_{s_2} = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}'(x) \right)^{-1} \left(\frac{1}{n_2} \sum_{x \in s_2} Y(x) \mathbf{Z}(x) \right) \quad (1)$$

The individual predictions can then be calculated as $\hat{Y}(x) = \mathbf{Z}'(x) \hat{\boldsymbol{\beta}}_{s_2}$ and the empirical model residuals, which are only available at all sample locations $x \in s_2$, are calculated as $\hat{R}(x) = Y(x) - \hat{Y}(x)$. Unless stated otherwise, **forestinventory** only uses *internal models* to calculate estimates. This means that the model fit, i.e. $\hat{\boldsymbol{\beta}}_{s_2}$, is derived from the current inventory data that are passed to the `twophase()` and `threephase()` functions. While virtually all inventors fit their models using the current inventory data, sometimes there is reason to use formulas derived from external models where the sample used to train the model is assumed to be taken from an independent source (Massey and Mandallaz 2015a). However, this usually occurs when using a model other than the OLS regression model and is beyond the scope of the package at this time.

The package provides the calculation of **point estimates** under exhaustive (*ex*) and non-exhaustive (*nex*) use of the auxiliary information, which means to respectively apply $\hat{\boldsymbol{\beta}}_{s_2}$ to $\bar{\mathbf{Z}}$, i.e. the *exact* spatial mean of $\mathbf{Z}(x)$, or to $\hat{\bar{\mathbf{Z}}}$, i.e. an *estimate* of the spatial mean of $\mathbf{Z}(x)$:

$$\hat{Y}_{reg2p,ex} = \bar{\mathbf{Z}}' \hat{\boldsymbol{\beta}}_{s_2} \quad (2a)$$

$$\hat{Y}_{reg2p,nex} = \hat{\bar{\mathbf{Z}}}' \hat{\boldsymbol{\beta}}_{s_2} \quad (2b)$$

Note that for *internal* linear models the mean of the empirical residuals $\frac{1}{n_2} \sum_{x \in s_2} \hat{R}(x)$ is zero by construction (*zero mean residual property*) which is why it does not appear in the point estimate. More explanation about how to obtain the auxiliary means is given in the next subsection.

The **forestinventory** package implements two kinds of variances for each of these point estimates: the *g-weight* formulation that accounts for the fact that our model is in fact *internal*, and the *external variance* formulation that assumes a true external regression model and thus neglects the uncertainty in the regression coefficients (Mandallaz *et al.* 2016).

The *g-weight* formulation is

$$\hat{V}(\hat{Y}_{reg2p,ex}) := \bar{\mathbf{Z}}' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \bar{\mathbf{Z}} \quad (3a)$$

$$\hat{V}(\hat{Y}_{reg2p,nex}) := \hat{\bar{\mathbf{Z}}}' \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\bar{\mathbf{Z}}} + \hat{\boldsymbol{\beta}}_{s_2}' \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathbf{Z}}}} \hat{\boldsymbol{\beta}}_{s_2} \quad (3b)$$

where the g-weight variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{s_2}$ is calculated as

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} := \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}'(x) \right)^{-1} \left(\frac{1}{n_2^2} \sum_{x \in s_2} \hat{R}^2(x) \mathbf{Z}(x) \mathbf{Z}'(x) \right) \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}'(x) \right)^{-1} \quad (4)$$

and the uncertainty caused by using the s_1 sample to estimate $\bar{\mathbf{Z}}$ by $\hat{\bar{\mathbf{Z}}}$ is accounted for by the variance-covariance matrix of the auxiliary vector $\hat{\bar{\mathbf{Z}}}$

$$\hat{\Sigma}_{\hat{\bar{\mathbf{Z}}}} = \frac{1}{n_1(n_1 - 1)} \sum_{x \in s_1} (\mathbf{Z}(x) - \hat{\bar{\mathbf{Z}}})(\mathbf{Z}(x) - \hat{\bar{\mathbf{Z}}})^t \quad (5)$$

The *external variance* formulation for linear regression models is

$$\begin{aligned} \hat{\mathbb{V}}_{ext}(\hat{Y}_{reg2p,ex}) &= \frac{1}{n_2} \hat{\mathbb{V}}_{s_2}(\hat{R}(x)) \\ \hat{\mathbb{V}}_{ext}(\hat{Y}_{reg2p,nex}) &= \frac{1}{n_1} \hat{\mathbb{V}}_{s_1}(\hat{Y}(x)) + \frac{1}{n_2} \hat{\mathbb{V}}_{s_2}(\hat{R}(x)) \end{aligned} \quad (6a)$$

where $\hat{\mathbb{V}}_{s_2}$ and $\hat{\mathbb{V}}_{s_1}$ indicate taking the sample variance over s_2 and s_1 respectively.

Note that when applied to internal linear regression models, the external variance is asymptotically unbiased and usually slightly smaller than the g-weight variance, where the uncertainty of the regression coefficients is accounted for by the variance-covariance matrix (Eq. 4). The external variances are provided in the package **forestinventory** in case the user wants to compare linear models to another model type where no g-weight formulation is possible, as is the case with non-parametric models like kNN.

Calculation of Explanatory Variables

We will now draw our attention to the calculation of the explanatory variables from the auxiliary data for both the non-exhaustive and exhaustive cases. Fig. 4b depicts how the **non-exhaustive** case often looks like in practice: a regular terrestrial grid s_2 is given by a terrestrial inventory (the points surrounded by dotted circles) and densified to a larger sample s_1 (the points). For every point x , each explanatory variable in the vector $\mathbf{Z}(x) = (z(x)_1, z(x)_2, \dots, z(x)_p)^t$ is calculated using a defined spatial extent of auxiliary information around that point called the *support* (the dark green square tiles). We emphasize that the value of the explanatory variables for $\mathbf{Z}(x)$ are associated with the sample point whereas the support is the spatial extent of the auxiliary information used to calculate those values. So far this is in perfect agreement with the presented theory of the non-exhaustive estimator, except for using regular grids rather than randomly placed sample points. The **forestinventory** package calculates the empirical mean of $\mathbf{Z}(x)$ automatically from the input data frame using $\hat{\bar{\mathbf{Z}}} = \frac{1}{n_1} \sum_{x \in s_1} \mathbf{Z}(x)$.

The **exhaustive** case requires a closer look. In the infinite population approach, $\mathbf{Z}(x)$ refers to the sample point x and not the area around it. Deriving the *exact* spatial mean, $\bar{\mathbf{Z}} = \frac{1}{\lambda(F)} \int_F \mathbf{Z}(x) dx = (\frac{1}{\lambda(F)} \int_F z_1(x) dx, \dots, \frac{1}{\lambda(F)} \int_F z_p(x) dx)^t$, implies that we need to calculate the spatial mean of each component of $\mathbf{Z}(x)$ using all possible points in F . This is much like the situation we had with calculating the mean of the local density surface for $Y(x)$ in that we need to find the mean of $\mathbf{Z}(x)$ over an infinite number of sample points (i.e. $n_1 = \infty$). Although it is practically infeasible to assess $\mathbf{Z}(x)$ for every x , there are few cases where the exact mean can in fact be precisely calculated. The first case is when the explanatory variables are provided by polygon layers (e.g. map of development stages). In this case, one can calculate the exact mean as the area-weighted average of each categorical variable. The second case is when the exact mean can be calculated in one step, e.g. taking the mean of all height pixels of a raster canopy height model will perfectly equal the mean calculated by the use of an infinite number of supports (Mandallaz *et al.* 2013). However, for most types of explanatory variables we will try to get an approximation of $\bar{\mathbf{Z}}$ that is only negligibly different.

One implementation to approximate the exact mean $\bar{\mathbf{Z}}$ is shown in Fig. 4a, where the spatial arrangement of the supports (the dark green tiles) are tessellated to form a perfect partition over the inventory

area in order for all of the wall-to-wall auxiliary information to be exploited. It has to be noted that this setup would allow for a perfect calculation of the exact mean $\bar{\mathbf{Z}}$ in the *finite* population approach, i.e. deriving $\mathbf{Z}(x)$ for the finite population of supports that are considered the sampling units. While in the infinite population approach this implementation probably does not produce the true exact mean $\bar{\mathbf{Z}}$, n_1 is still expected to be reasonably large for the difference to be considered negligible as long as the size of the supports are not unreasonably large. However, the perfect tessellation implementation can also impose drawbacks. One is that a perfect tessellation by the supports strongly depends on the distance between the sample locations of s_1 and the support size. Since in practice the support size should ideally be chosen to achieve a best possible explanatory power of the regression model (thus minimizing the residual variation) a perfect tessellation might often not be feasible. In the infinite population frame, the supports are allowed to overlap if this seems necessary to acquire a sufficiently large sample n_1 to get a negligibly close approximation of $\bar{\mathbf{Z}}$. With this respect, the infinite population approach provides more flexibility than the finite approach.

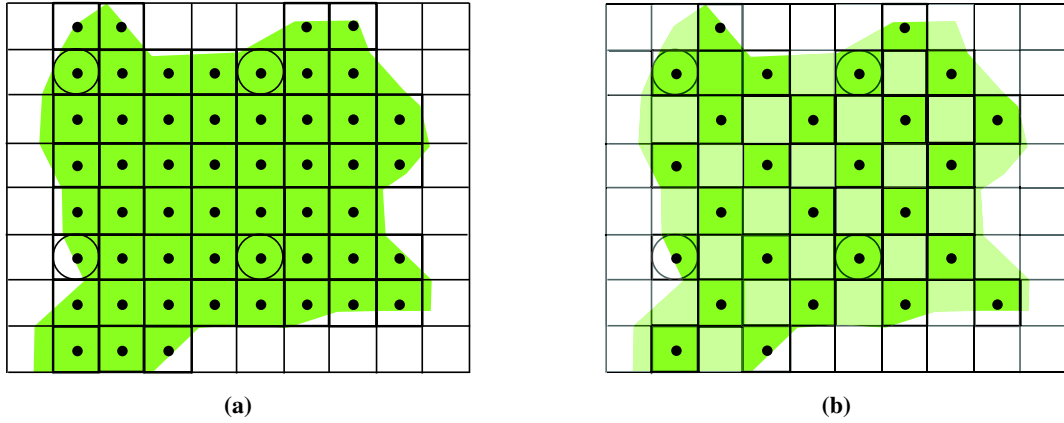


Figure 4: Concept of (a) exhaustive and (b) non-exhaustive calculation of explanatory variables including boundary adjustment at the support level. Auxiliary data are in both cases available over the entire inventory area marked by the large rectangle. A vector of explanatory variables $\mathbf{Z}(x)$ is calculated within the supports (small squares) at each sample location x (points) that falls into the forest area (green underlying polygon).

An extension to the so-far published estimators by Mandallaz is the consideration of a *boundary adjustment*. In forest inventories, the sample is often restricted to those sample locations located within the forest area. In case a consistent forest definition can be applied to both the s_2 and s_1 sample (e.g. by a polygon forest mask layer), it might be desired to restrict the calculation of the explanatory variables to the forest area within the given support (see Fig. 4). This method was suggested in Mandallaz *et al.* (2013) and led to an improvement in estimation precision. In order to ensure an unbiased calculation of either $\hat{\bar{\mathbf{Z}}}$ or $\bar{\mathbf{Z}}$, the respective means have then to be calculated as the *weighted* mean (Eq. 7) where the weight $w(x)$ is equal to the percentage of forested area within the support of sample location x .

$$\hat{\bar{\mathbf{Z}}} = \frac{\sum_{x \in s_1} w(x) \mathbf{Z}(x)}{\sum_{x \in s_1} w(x)} \quad (7)$$

Application

To demonstrate the use of the global two-phase estimators, we will use the grisons data set that comes with installing the package from the CRAN repository. The data set contains data from a simple (i.e. non-cluster) two-phase forest inventory conducted in 2007 that was used in Mandallaz *et al.* (2013) as a case study. The s_1 sample is comprised of 306 sample locations arranged on a

systematic grid containing auxiliary information in the form of LiDAR canopy height metrics (mean, stddev, max, q75). For a systematic subsample of 67 (s_2 sample), terrestrial information of the timber volume per hectare (tvol) on the sample plot level is provided from a terrestrial survey. We can load **forestinventory** and examine the grisons data set in the R environment as follows:

```
R> library(forestinventory)
R> data("grisons", package = "forestinventory")
R> head(grisons)
```

	phase_id_2p	boundary_weights	mean	stddev	max	q75	smallarea	tvol
1	2	1.0000000	9.301509	11.841066	40.86731	21.140011	C	107.80
2	1	1.0000000	12.156926	11.353946	39.79736	21.542512	A	NA
3	2	1.0000000	5.247305	5.738444	23.81775	9.530029	D	63.77
4	1	1.0000000	7.533882	9.332118	34.09998	13.022277	A	NA
5	1	0.6660761	6.105373	5.872390	23.32996	10.551788	B	NA
6	1	1.0000000	12.149097	10.161655	33.75732	20.969971	C	NA
7	2	1.0000000	6.377284	4.720547	17.96094	10.143495	D	154.10
8	1	1.0000000	1.247679	3.793701	22.71594	0.000000	B	NA
9	1	1.0000000	21.563423	7.493390	32.65515	27.811491	A	NA
10	2	1.0000000	13.547820	7.200165	36.14001	18.587038	A	256.15

Estimates can be made using the `onephase()`, `twophase()` or `threephase()` functions. The data frame inputted to these functions must have the structure where each row corresponds to a unique sample location and the columns specify the attributes associated to that respective sample location. Attributes that are missing, e.g. because they are associated with sample locations that were not selected in the subsample for the subsequent phase, should be designated as NA and the phase membership is encoded as numeric.

For global two-phase estimation, we have to specify

- the regression model (`formula`) as specified in the `lm()`-function (R Core Team 2017)
- the inputted data frame containing the inventory information (`data`)
- the list-object `phase_id` containing: the `phase.col` argument identifying the name of the column specifying membership to s_1 or s_2 , and the `terrgrid.id` argument specifying which numeric value indicates s_2 membership in that column
- the name of the column containing the weights $w(x)$ of the boundary adjustments (optional)

The **non-exhaustive estimator with boundary weight adjustment** can thus be applied as follows:

```
R> reg2p_nex <- twophase(formula=tvol ~ mean + stddev + max + q75,
+                        data=grisons,
+                        phase_id=list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                        boundary_weights = "boundary_weights")
```

The `twophase()` function creates an S3 object of class "twophase" with subclass "global". A readable **summary of the estimation results** can be obtained by passing this object to the `summary()` function, which automatically interprets what type of estimator was used and returns pertinent information such as the regression model formula, the point estimate (`estimate`), the g-weight and external variance (`g_variance` and `ext_variance`) as well as the sample sizes and the model R^2 :

```
R> summary(reg2p_nex)
```

Two-Phase global estimation

Call:

```
twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
         phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
         boundary_weights = "boundary_weights")
```

Method used:

Non-exhaustive global estimator

Regression Model:

```
tvol ~ mean + stddev + max + q75
```

Estimation results:

```
estimate ext_variance g_variance n1 n2 r.squared
383.5354      279.954    271.5057 306 67 0.6428771
```

'boundary_weight'- option was used to calculate weighted means of auxiliary variables

For practical use, one should normally always prefer the g-weight variance over the external variance. This is because when we use internal models, the regression coefficients actually depend on the terrestrial sample realized by the sampling design. In contrast to the external variance, the g-weight variance accounts for this sampling variability which results in more reliable point and variance estimates and also enjoys better statistical calibration properties (g-weights). The external and g-weight variances are asymptotically equivalent but the external variance is really only included here in case the user wants to compare to another estimator where no g-weight variance exists.

The **exhaustive estimator** can be applied by additionally passing a vector containing the *exact* means of the explanatory variables, i.e. \bar{Z} , to the optional argument `exhaustive`. This vector must be calculated beforehand in such a way that any desired boundary adjustment has already been applied. Note that the vector input to `exhaustive` must be in the same order that the `lm()`-function processes a `formula` object including the intercept term whose exact mean will always be 1. Particular caution must be taken if categorical variables are present because the `lm()`-function, which is internally used to set up the design-matrix, automatically creates dummy variables with one of the categories used as a reference. Using our `grisons` example, the correct order can always be extracted by the following R-code:

```
R> colnames(lm(formula = tvol ~ mean + stddev + max + q75, data = grisons, x = TRUE)$x)
```

The **exhaustive estimator** can be applied after defining the vector of exact means \bar{Z} taken from [Mandallaz *et al.* \(2013\)](#), denoted as `true.means.Z`:

```
R> true.means.Z <- c(1, 11.39, 8.84, 32.68, 18.03)
R> reg2p_ex <- twophase(formula = tvol ~ mean + stddev + max + q75,
+                       data = grisons,
+                       phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                       exhaustive = true.means.Z)
```

An alternative way to look at the estimation results without using the `summary()` is to query `reg2p_ex` directly:

```
R> reg2p_ex$estimation
```

```
estimate ext_variance g_variance n1 n2 r.squared
1 376.7426      202.5602    187.2787 Inf 67 0.6428771
```

Note that both variances of the exhaustive estimation are smaller than those of the non-exhaustive estimation. This is essentially because we eliminated one component of uncertainty by substituting the estimated means of the explanatory variables $\hat{\bar{\mathbf{Z}}}$ by their exact means $\bar{\mathbf{Z}}$.

3.2. Small Area Estimators

Mathematical Background

The **forestinventory** package provides three types of small area estimators each of which has an exhaustive and non-exhaustive form. We will use a different nomenclature for the non-exhaustive case in small area estimation since much of the existing literature shows preference for the label *pseudo* to indicate that the mean of the explanatory variables within the small area was based on a finite sample. The main idea for all these small area estimators is to calculate the regression coefficient vector $\hat{\boldsymbol{\beta}}_{s_2}$ and its variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}$ on the entire s_2 sample according to Eq. 1 and 4, and subsequently use that to make predictions for sample locations restricted to small area G .

We first introduce the **small area estimator** (*small*), which uses exhaustively computed explanatory variables, and its non-exhaustive version, the **pseudo small area estimator** (*psmall*).

$$\hat{Y}_{G,small,2p} = \bar{\mathbf{Z}}_G^t \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}(x) \quad (8a)$$

$$\hat{Y}_{G,psmall,2p} = \hat{\bar{\mathbf{Z}}}_G^t \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}(x) \quad (8b)$$

In the equations for the point estimates (Eq. 8a and 8b), we see that the globally derived regression coefficients are applied to the exhaustively or non-exhaustively calculated means of the explanatory variables ($\bar{\mathbf{Z}}_G$, $\hat{\bar{\mathbf{Z}}}_G$) which are now only based on the first-phase sample $s_{1,G}$ located within small area G . A potential bias of the regression model predictions in the small area G , due to fitting the regression model with data also outside of G , is then corrected by adding the mean of the empirical model residuals in G . This is called the *bias* or *residual correction* term.

The package provides the *g-weight* variance for *small* and *psmall* respectively (Eq. 9a, 9b) as well as the *external variance* (Eq. 10a, 10b). Again note that all components are restricted to those available at the sample locations in the small area ($s_{1,G}$ and $s_{2,G}$), with exception of the regression coefficient components $\hat{\boldsymbol{\beta}}_{s_2}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}$.

$$\hat{V}(\hat{Y}_{G,small,2p}) := \bar{\mathbf{Z}}_G^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \bar{\mathbf{Z}}_G + \frac{1}{n_{2,G}} \hat{V}_{s_{2,G}}(\hat{R}(x)) \quad (9a)$$

$$\hat{V}(\hat{Y}_{G,psmall,2p}) := \hat{\bar{\mathbf{Z}}}_G^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\bar{\mathbf{Z}}}_G + \hat{\boldsymbol{\beta}}_{s_2}^t \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathbf{Z}}}_G} \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{V}_{s_{2,G}}(\hat{R}(x)) \quad (9b)$$

$$\hat{V}_{ext}(\hat{Y}_{G,small,2p}) := \frac{1}{n_{2,G}} \hat{V}_{s_{2,G}}(\hat{R}(x)) \quad (10a)$$

$$\hat{V}_{ext}(\hat{Y}_{G,psmall,2p}) := \frac{1}{n_{1,G}} \hat{V}_{s_{2,G}}(Y(x)) + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right) \frac{1}{n_{2,G}} \hat{V}_{s_{2,G}}(\hat{R}(x)) \quad (10b)$$

where $\hat{\mathbb{V}}_{s_{2,G}}$ indicates taking the sample variance over $s_{2,G}$. If boundary adjustment is applied, the simple mean of the explanatory variable over the small area $\hat{\mathbf{Z}}_G = \frac{1}{n_{1,G}} \sum_{x \in s_{1,G}} \mathbf{Z}(x)$ is replaced by its weighted version $\hat{\mathbf{Z}}_G = \frac{\sum_{x \in s_{1,G}} w(x) \mathbf{Z}(x)}{\sum_{x \in s_{1,G}} w(x)}$, and likewise for exhaustively used auxiliary information.

The **synthetic estimator** (*synth*) and **pseudo synthetic estimator** (*psynth*) are commonly applied when no terrestrial sample is available within the small area G (i.e. $n_{2,G} = 0$). In this case, the point estimates (Eq. 11a and 11b) are based *only* on the predictions generated by applying the globally derived regression model to the auxiliary vectors $\bar{\mathbf{Z}}_G$ and $\hat{\mathbf{Z}}_G$ respectively. However, the bias correction using the observed residuals $\hat{R}(x)$ is not applied as was the case in the small and pseudo small area estimator (Eq. 8a and 8b). Thus, the (pseudo) synthetic estimator has a potentially unobservable design-based bias. Also note that the residual variation can no longer be considered in the g-weight variance (Eq. 11c and 11d). Therefore, the synthetic estimators will usually have a smaller variance than estimators incorporating the regression model uncertainties, but at the cost of a potential bias. Due to the absence of available residuals in G , there is also no external variance form for the synthetic and pseudo synthetic estimator.

$$\hat{Y}_{G,\text{synth},2p} = \bar{\mathbf{Z}}_G^t \hat{\boldsymbol{\beta}}_{s_2} \quad (11a)$$

$$\hat{Y}_{G,\text{psynth},2p} = \hat{\mathbf{Z}}_G^t \hat{\boldsymbol{\beta}}_{s_2} \quad (11b)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,\text{synth},2p}) = \bar{\mathbf{Z}}_G^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \bar{\mathbf{Z}}_G \quad (11c)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,\text{psynth},2p}) = \hat{\mathbf{Z}}_G^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\mathbf{Z}}_G + \hat{\boldsymbol{\beta}}_{s_2}^t \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_G} \hat{\boldsymbol{\beta}}_{s_2} \quad (11d)$$

where the variance-covariance matrix of the auxiliary vector $\hat{\mathbf{Z}}_G$ is estimated by

$$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_G} = \frac{1}{n_{1,G}(n_{1,G} - 1)} \sum_{x \in s_{1,G}} (\mathbf{Z}(x) - \hat{\mathbf{Z}}_G)(\mathbf{Z}(x) - \hat{\mathbf{Z}}_G)^t \quad (12)$$

The synthetic estimators, *synth* and *psynth*, have attractively compact formulas but come with the downside of potential bias in their point estimates which can make the variances seem deceptively optimistic. The *small* and *psmall* estimators overcome this issue by using a bias correction term, i.e. $\frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} \hat{R}(x)$. The motivation behind the **extended synthetic** and **extended pseudo synthetic estimator** (*extsynth* and *extpsynth*) is to use the same mathematically elegant formulas of the (pseudo) synthetic estimators while at the same time ensuring that the empirical residuals of the prediction model in the entire area F and the small area G are by construction both zero at the same time. This is accomplished by extending the vector of auxiliary information $\mathbf{Z}(x)$ by a binary categorical indicator variable $I_G(x)$ which takes the value 1 if the sample location x lies inside the target small area G and is otherwise set to 0. Recalling that linear models fitted using OLS have zero mean residual property by construction also if categorical variables are used, this leads to unbiased point estimates. The new *extended* auxiliary vector thus becomes $\mathbf{Z}^t(x) = (\mathbf{Z}^t(x), I_G(x))$ and can be used to replace its non-extended counterpart $\mathbf{Z}^t(x)$ wherever it is used in Eq. 11 and 12. Note that the package functions internally extend the data set by the indicator variable if the *extsynth* or *extpsynth* estimator is called.

Not every equation needs to be re-written here, but to give an example of the notational change, the regression coefficient under extended model approach becomes

$$\hat{\boldsymbol{\theta}}_{s_2} = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}^t(x) \right)^{-1} \left(\frac{1}{n_2} \sum_{x \in s_2} Y(x) \mathbf{Z}(x) \right) \quad (13)$$

The point estimates and their g-weight variances can then be re-written as

$$\hat{Y}_{G,extsynth,2p} = \bar{\mathbf{Z}}_G^t \hat{\boldsymbol{\theta}}_{s_2} \quad (14a)$$

$$\hat{Y}_{G,extpsynth,2p} = \hat{\mathbf{Z}}_G^t \hat{\boldsymbol{\theta}}_{s_2} \quad (14b)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,extsynth,2p}) = \bar{\mathbf{Z}}_G^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}} \bar{\mathbf{Z}}_G \quad (14c)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,extpsynth,2p}) = \hat{\mathbf{Z}}_G^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}} \hat{\mathbf{Z}}_G + \hat{\boldsymbol{\theta}}_{s_2}^t \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_G} \hat{\boldsymbol{\theta}}_{s_2} \quad (14d)$$

While the formulas look similar to the synthetic estimators, note that a decomposition of $\hat{\boldsymbol{\theta}}_{s_2}$ reveals that the residual correction term is now included in the regression coefficient $\hat{\boldsymbol{\theta}}_{s_2}$ (Mandallaz *et al.* 2016) and thus the estimates are asymptotically design-unbiased.

The package also provides the external variance for both the *extended* synthetic and *extended* pseudo synthetic estimator. Note that neither the extended model approach nor external variance estimates are possible in the absence of terrestrial samples and thus model residuals in G , which is precisely when one must rely on the (pseudo) synthetic estimates. The external variance forms of *extsynth* and *extpsynth* are

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,extsynth,2p}) = \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_2,G}(\hat{\mathbb{R}}(x)) \quad (15a)$$

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,extpsynth,2p}) = \frac{1}{n_{1,G}} \hat{\mathbb{V}}_{s_2,G}(Y(x)) + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right) \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_2,G}(\hat{\mathbb{R}}(x)) \quad (15b)$$

where $\hat{\mathbb{R}}(x)$ are the empirical residuals under the extended auxiliary vector.

To summarize, the (pseudo) synthetic estimator can be applied whether terrestrial inventory sample is found in the small area or not, but has a deceptively small g-weight variance due to its potential bias. When terrestrial sample is observed in the small area, we can produce (asymptotically) design-unbiased estimates and variances using either *small* or *psmall* which remove this bias explicitly with a mean residual term, or more elegantly with *extsynth* or *extpsynth* which simply use the same synthetic formulas while including an indicator variable for the small area in the model formula to remove the bias by construction in OLS.

Application

Small area estimates in the **forestinventory** package can be applied by specifying the optional argument `small_area`. The input data set has to include an additional column of class `factor` that describes the small area membership of the sample location represented by that row. The argument `small_area` requires a `list`-object that comprises

- the name of the column specifying the small area of each observation (`sa.col`)
- a vector specifying the small area(s) for which estimations are desired (`areas`)
- the argument `unbiased` that controls which of the three available estimators is applied

In order to apply the **pseudo small area estimator** (*psmall*) with boundary adjustment, we set `unbiased=TRUE` as well as the optional argument `psmall=TRUE`:

```
R> psmall_2p <- twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
+                       phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                       boundary_weights = "boundary_weights",
+                       small_area = list(sa.col = "smallarea", areas = c("A", "B")),
+                       unbiased = TRUE),
+                       psmall = TRUE)
R> summary(psmall_2p)
```

Two-phase small area estimation

Call:

```
twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
  phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
  small_area = list(sa.col = "smallarea", areas = c("A", "B")),
  unbiased = TRUE), boundary_weights = "boundary_weights",
  psmall = TRUE)
```

Method used:

Pseudo small area estimator

Regression Model:

$tvol \sim mean + stddev + max + q75$

Estimation results:

	area estimate	ext_variance	g_variance	n1	n2	n1G	n2G	r.squared
A	393.9713	1009.034	1308.117	306	67	94	19	0.6428771
B	419.6416	1214.035	1259.472	306	67	81	17	0.6428771

'boundary_weight'- option was used to calculate weighted means of auxiliary variables

The small area functions all return an S3 object of class "twophase" with subclass "smallarea". In addition to global estimation, the estimation object now comprises the estimates and variances for all small areas (column area). We can view the sample sizes by looking into the object itself

```
R> psmall_2p$samplesizes
```

\$A

	n1G	n2G	n1	n2
plots	94	19	306	67

\$B

	n1G	n2G	n1	n2
plots	81	17	306	67

The **extended pseudo synthetic estimator** (*extpsynth*) can be applied by setting `unbiased=TRUE` and leaving the optional argument `psmall` to its default value of `FALSE`:

```
R> extpsynth_2p <- twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
+                           phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                           boundary_weights = "boundary_weights",
+                           small_area = list(sa.col = "smallarea", areas = c("A", "B")),
+                           unbiased = TRUE)
R> extpsynth_2p$estimation
```

	area estimate	ext_variance	g_variance	n1	n2	n1G	n2G	r.squared
1	A 391.9356	995.5602	1017.633	306	67	94	19	0.6526503
2	B 419.7231	1214.6053	1019.191	306	67	81	17	0.6428854

The **forestinventory** package automatically includes the indicator variable for the small area behind the scenes so there is no need for the user to implement it. Notice that the R^2 's (r.squared) under the *extpsynth* estimator vary between the small areas, while they are identical under the *psmall* estimator. This is because under the *extpsynth* estimator, the regression model is recalculated for each small area estimation after adding the indicator variable for the respective small area in the globally derived design matrix. In case of the *psmall* estimator, the regression model stays the same for each small area estimation. Although the results of both estimators should always be close to each other, we recommend applying both estimators and compare the results afterwards in order to reveal unsuspected patterns in the data, particularly in the case of cluster sampling (see Section 6).

Setting the argument `unbiased=FALSE` applies the **pseudo synthetic estimator** to the selected small areas. Note that in the *grisons* data set, all small areas possess much more than the suggested minimum number of terrestrial observations (a rule of thumb is that $n_{2,G} \geq 6$) required to produce reliable design-unbiased estimates. Hence, choosing to use *psynth* is probably not desirable and is just applied here for demonstration purposes. In this case the residual correction will not be applied.

```
R> psynth_2p <- twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
+                         phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                         boundary_weights = "boundary_weights",
+                         small_area = list(sa.col = "smallarea", areas = c("A", "B")),
+                         unbiased = FALSE)
R> psynth_2p$estimation
```

	area	estimate	ext_variance	g_variance	n1	n2	n1G	n2G	r.squared
1	A	421.8863	NA	546.8651	306	67	94	19	0.6428771
2	B	418.7399	NA	566.3361	306	67	81	17	0.6428771

We see here that the *psynth* variances are almost only half the variances of the *psmall* and *extended psynth* estimators. However, *psmall* and *extended psynth* are design unbiased and their variances reflect the fact that they account for potential bias of the regression model predictions. The g-weight variance of *psynth* completely neglects a potential bias and as a result risks severely overstating the estimation precision.

The **exhaustive versions** of the small area estimators (Eq. 8a, 9a, 10a, 11a, 11c) are specified via the optional argument `exhaustive`. Its application requires that we know the exact means of all explanatory variables within the small area(s) of interest. In contrast to the *global* estimators, the exact means have now to be delivered in the form of a `data.frame`, where each row corresponds to a small area, and each column specifies the exact mean of the respective explanatory variable. Note that likewise the case of global estimation, the order of the explanatory variables in the data frame has to match the order in which they appear in the design matrix defined by the `lm()`-function in R. In order to tell R which row describes which small area, the row names have to match the respective names of the small areas specified in the `areas` argument.

For the *grisons* data set, the exact means of the explanatory variables for the small areas used in Mandallaz *et al.* (2013) are thus defined by

```
R> colnames(lm(formula = tvol ~ mean + stddev + max + q75, data = grisons, x = TRUE)$x)

R> true.means.Z.G <- data.frame(Intercept = rep(1, 4),
+                               mean = c(12.85, 12.21, 9.33, 10.45),
+                               stddev = c(9.31, 9.47, 7.90, 8.36),
+                               max = c(34.92, 35.36, 28.81, 30.22),
+                               q75 = c(19.77, 19.16, 15.40, 16.91))
R> rownames(true.means.Z.G) <- c("A", "B", "C", "D")
```



```
R> true.means.Z.G
```

	Intercept	mean	stddev	max	q75
A	1	12.85	9.31	34.92	19.77
B	1	12.21	9.47	35.36	19.16
C	1	9.33	7.90	28.81	15.40
D	1	10.45	8.36	30.22	16.91

The **extended synthetic estimator** (*extsynth*) can then be applied by

```
R> extsynth_2p <- twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
+                           phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                           small_area = list(sa.col = "smallarea", areas = c("A", "B")),
+                           unbiased = TRUE),
+                           exhaustive = true.means.Z.G)
R> extsynth_2p$estimation
```

	area	estimate	ext_variance	g_variance	n1	n2	n1G	n2G	r.squared
1	A	372.6930	744.3658	696.5739	Inf	67	Inf	19	0.6526503
2	B	387.5116	693.8576	708.1105	Inf	67	Inf	17	0.6428854

Just as in the global case, we see that the variance has again been significantly decreased by substituting the *exact* auxiliary means and both first phase sample sizes are now infinity. Note that the function extracts the required exact means for small area "A" and "B" from the complete set of exact means defined in `true.means.Z.G`.

4. Three-phase Estimators and their Application

4.1. Global Estimators

Mathematical Background

Solving the sample-based normal equations, the vector of regression coefficients $\hat{\alpha}_{s_2}$ for the *reduced model*, i.e. using the *reduced set of explanatory variables* $\mathbf{Z}^{(0)}(x)$ available at $x \in s_0$, and likewise the vector of regression coefficients $\hat{\beta}_{s_2}$ for the *full model*, i.e. using the *full set of explanatory variables* $\mathbf{Z}^t(x) = (\mathbf{Z}^{(0)t}(x), \mathbf{Z}^{(1)t}(x))$ available only at a subset $x \in s_1 \subset s_0$, are derived as

$$\hat{\alpha}_{s_2} = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}^{(0)}(x) \mathbf{Z}^{(0)t}(x) \right)^{-1} \frac{1}{n_2} \sum_{x \in s_2} Y(x) \mathbf{Z}^{(0)}(x) \quad (16a)$$

$$\hat{\beta}_{s_2} = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}^t(x) \right)^{-1} \frac{1}{n_2} \sum_{x \in s_2} Y(x) \mathbf{Z}(x) \quad (16b)$$

The package allows for the calculation of **point estimates** under exhaustive and non-exhaustive use of the auxiliary information in the s_0 phase. Fitting the model using s_2 (i.e. *internally*) ensures the zero mean residual property over s_2 .

$$\begin{aligned} \hat{Y}_{reg3p,ex} &= \frac{1}{\lambda(F)} \int_F \mathbf{Z}^{(0)t}(x) \hat{\alpha}_{s_2} + \frac{1}{n_1} \sum_{x \in s_1} (\mathbf{Z}^t(x) \hat{\beta}_{s_2} - \mathbf{Z}^{(0)t}(x) \hat{\alpha}_{s_2}) + \frac{1}{n_2} \sum_{x \in s_2} (Y(x) - \mathbf{Z}^t(x) \hat{\beta}_{s_2}) \\ &= (\bar{\mathbf{Z}}_0^{(0)} - \hat{\mathbf{Z}}_1^{(0)})^t \hat{\alpha}_{s_2} + \hat{\mathbf{Z}}_1^t \hat{\beta}_{s_2} \end{aligned} \quad (17a)$$

$$\begin{aligned} \hat{Y}_{reg3p,nex} &= \frac{1}{n_0} \sum_{x \in s_0} \mathbf{Z}^{(0)t}(x) \hat{\alpha}_{s_2} + \frac{1}{n_1} \sum_{x \in s_1} (\mathbf{Z}^t(x) \hat{\beta}_{s_2} - \mathbf{Z}^{(0)t}(x) \hat{\alpha}_{s_2}) + \frac{1}{n_2} \sum_{x \in s_2} (Y(x) - \mathbf{Z}^t(x) \hat{\beta}_{s_2}) \\ &= (\hat{\mathbf{Z}}_0^{(0)} - \hat{\mathbf{Z}}_1^{(0)})^t \hat{\alpha}_{s_2} + \hat{\mathbf{Z}}_1^t \hat{\beta}_{s_2} \end{aligned} \quad (17b)$$

Intuitively, the three phase estimator is simply the mean of the predictions using the reduced model, corrected by the mean difference between the reduced model predictions and the more accurate full model predictions, corrected by the mean difference between the ground truth and the full model predictions. For the compact version of the formula in the *non-exhaustive case*, the *estimated* means of $\mathbf{Z}^{(0)}(x)$ over both the s_0 and s_1 phase, as well as the *estimated* mean of $\mathbf{Z}(x)$ over the s_1 phase are calculated according to Eq. 18. If the exact mean over s_0 is known, the estimated mean $\hat{\mathbf{Z}}_0^{(0)}$ can simply be replaced by the exact mean $\bar{\mathbf{Z}}_0^{(0)}$. Note that in case of applied boundary adjustment (Section 3), the simple mean is again replaced by the weighted mean analogous to Eq. 7.

$$\hat{\mathbf{Z}}_0^{(0)} = \frac{1}{n_0} \sum_{x \in s_0} \mathbf{Z}^{(0)}(x), \quad \hat{\mathbf{Z}}_1^{(0)} = \frac{1}{n_1} \sum_{x \in s_1} \mathbf{Z}^{(0)}(x), \quad \hat{\mathbf{Z}}_1 = \frac{1}{n_1} \sum_{x \in s_1} \mathbf{Z}(x) \quad (18)$$

The package again provides the g-weight and external variances. The **g-weight variance** formulation is

$$\hat{V}(\hat{Y}_{reg3p,ex}) = \frac{n_2}{n_1} \bar{\mathbf{Z}}_0^{(0)t} \hat{\Sigma}_{\hat{\alpha}_{s_2}} \bar{\mathbf{Z}}_0^{(0)} + \left(1 - \frac{n_2}{n_1}\right) \hat{\mathbf{Z}}_1^t \hat{\Sigma}_{\hat{\beta}_{s_2}} \hat{\mathbf{Z}}_1 \quad (19a)$$

$$\hat{V}(\hat{Y}_{reg3p,nex}) = \hat{\alpha}_{s_2}^t \hat{\Sigma}_{\hat{\alpha}_{s_2}} \hat{\alpha}_{s_2} + \frac{n_2}{n_1} \hat{\mathbf{Z}}_0^{(0)t} \hat{\Sigma}_{\hat{\alpha}_{s_2}} \hat{\mathbf{Z}}_0^{(0)} + \left(1 - \frac{n_2}{n_1}\right) \hat{\mathbf{Z}}_1^t \hat{\Sigma}_{\hat{\beta}_{s_2}} \hat{\mathbf{Z}}_1 \quad (19b)$$

with the variance-covariance matrix of $\hat{\mathbf{Z}}_0^{(0)}$ and the variance-covariance matrices of the regression coefficients $\hat{\boldsymbol{\alpha}}_{s_2}$ and $\hat{\boldsymbol{\beta}}_{s_2}$:

$$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_0^{(0)}} = \frac{1}{n_0(n_0 - 1)} \sum_{x \in s_0} (\mathbf{Z}^{(0)}(x) - \hat{\mathbf{Z}}_0^{(0)})(\mathbf{Z}^{(0)}(x) - \hat{\mathbf{Z}}_0^{(0)})^t \quad (20a)$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}^{(0)}(x) \mathbf{Z}^{(0)t}(x) \right)^{-1} \left(\frac{1}{n_2} \sum_{x \in s_2} \hat{R}^{(0)2}(x) \mathbf{Z}^{(0)}(x) \mathbf{Z}^{(0)t}(x) \right) \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}^{(0)}(x) \mathbf{Z}^{(0)t}(x) \right)^{-1} \quad (20b)$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} = \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}^t(x) \right)^{-1} \left(\frac{1}{n_2} \sum_{x \in s_2} \hat{R}^2(x) \mathbf{Z}(x) \mathbf{Z}^t(x) \right) \left(\frac{1}{n_2} \sum_{x \in s_2} \mathbf{Z}(x) \mathbf{Z}^t(x) \right)^{-1} \quad (20c)$$

Note that $\hat{R}(x) = Y(x) - \mathbf{Z}^t(x) \hat{\boldsymbol{\beta}}_{s_2}$ denotes the empirical residuals of the full model, whereas $\hat{R}^{(0)}(x) = Y(x) - \mathbf{Z}^{(0)t} \hat{\boldsymbol{\alpha}}_{s_2}$ denotes the empirical residuals of the reduced model. The **external variance** form under linear regression models is defined as

$$\hat{V}_{ext}(\hat{Y}_{reg3p,ex}) = \frac{1}{n_1} \hat{V}_{s_2}(\hat{R}^{(0)}(x)) + \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \hat{V}_{s_2}(\hat{R}(x)) \quad (21a)$$

$$\hat{V}_{ext}(\hat{Y}_{reg3p,nex}) = \frac{1}{n_0} \hat{V}_{s_0}(\hat{Y}^{(0)}(x)) + \frac{1}{n_1} \hat{V}_{s_2}(\hat{R}^{(0)}(x)) + \left(1 - \frac{n_2}{n_1}\right) \frac{1}{n_2} \hat{V}_{s_2}(\hat{R}(x)) \quad (21b)$$

where \hat{V}_{s_0} indicates taking the sample variance over s_0 .

Application

In order to demonstrate the three-phase estimators in the package, we created an artificial three-phase scenario by recoding the phase indicators in the `grisons` data set (column `phase_id_3p`) according to the terminology used in this article (0 for s_0 , 1 for s_1 , 2 for s_2). We now assume that the mean canopy height (mean) is available at all 306 sample locations $x \in s_0$, whereas we have the explanatory variables `stddev`, `max` and `q75` only at 128 subsamples s_1 of s_0 . At 40 further subsamples s_2 we have the observations $Y(x)$ from the field inventory. Based on this setup, we can now define the reduced and full regression model formulas to be used in the `threephase()` function (note that the models are nested):

```
R> formula.rm <- tvol ~ mean # reduced model applied to s0 phase
R> formula.fm <- tvol ~ mean + stddev + max + q75 # full model applied to s1 phase
```

Compared to the `twophase()`-function, we now have to specify two regression models, i.e. the nested reduced (`formula.s0`) and full (`formula.s1`) regression model. In addition, we also have to specify the indication of the s_1 phase (`s1.id`) in the argument `phase_id` (note that `forestinventory` implicitly assumes that all other rows in the input data set belong to s_0). The global three-phase estimation can thus be applied by

```
R> reg3p_nex <- threephase(formula.s0 = formula.rm,
+                          formula.s1 = formula.fm, data = grisons,
+                          phase_id = list(phase.col="phase_id_3p", s1.id = 1,
+                                          terrgrid.id = 2),
+                          boundary_weights = "boundary_weights")
```

```
R> summary(reg3p_nex)
```

Three-phase global estimation

Call:

```
threephase(formula.s0 = formula.rm, formula.s1 = formula.fm,
  data = grisons, phase_id = list(phase.col = "phase_id_3p",
  s1.id = 1, terrgrid.id = 2), boundary_weights = "boundary_weights")
```

Method used:

Non-exhaustive global estimator

Full Regression Model:

```
tvol ~ mean + stddev + max + q75
```

Reduced Regression Model:

```
tvol ~ mean
```

Estimation results:

estimate	ext_variance	g_variance	n0	n1	n2	r.squared_reduced	r.squared_full
372.0896	454.4064	451.3626	306	128	40	0.527363	0.7166608

'boundary_weight'- option was used to calculate weighted means of auxiliary variables

The `summary()` of a `threephase()`-function now recalls *both* regression model formulas and also gives the R^2 for both the reduced (`r.squared_reduced`) and the full (`r.squared_full`) models. We are told that including `stddev`, `max` and `q75` yields a 20 % improvement in R^2 . When comparing to using only `mean` under a two-phase approach, we would see a considerable reduction in variance by the three-phase extension.

4.2. Small Area Estimators

Mathematical Background

The three two-phase small area estimators described in Section 3.2 can also be extended to the three-phase scenario. The general principle thereby stays the same, i.e. the regression coefficients of the reduced and full model and their variance-covariance matrices are calculated on the entire s_2 sample according to Eq. 16a, 16b, 20b and 20c, and are subsequently used to make predictions for sample locations restricted to small area G .

The unbiased **point estimates** of the **small** (*small*) and **pseudo small area estimator** (*psmall*) are calculated by applying the globally derived reduced and full regression model coefficients to the small area means of the explanatory variables, and then corrected for a potential model bias in G by adding the small area mean of the full model residuals, i.e. $\hat{R}_G(x) = Y_G(x) - \mathbf{Z}'_G(x)\hat{\boldsymbol{\beta}}_{s_2}$, to the point estimate. The difference between the mean $\hat{\mathbf{Z}}_{1,G}^{(0)}$ and the more precise or exact mean $\hat{\mathbf{Z}}_{0,G}^{(0)}$ and $\bar{\mathbf{Z}}_{0,G}^{(0)}$ is again considered as a correction term likewise in the global estimation (Eq. 17).

$$\hat{Y}_{G,small,3p} = (\bar{\mathbf{Z}}_{0,G}^{(0)} - \hat{\mathbf{Z}}_{1,G}^{(0)})^t \hat{\boldsymbol{\alpha}}_{s_2} + \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}_G(x) \quad (22a)$$

$$\hat{Y}_{G,psmall,3p} = (\hat{\mathbf{Z}}_{0,G}^{(0)} - \hat{\mathbf{Z}}_{1,G}^{(0)})^t \hat{\boldsymbol{\alpha}}_{s_2} + \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}_G(x) \quad (22b)$$

The **g-weight variance** is then calculated as

$$\hat{\mathbb{V}}(\hat{Y}_{G,small,3p}) = \frac{n_2}{n_1} \bar{\mathbf{Z}}_{0,G}^{(0)t} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} \bar{\mathbf{Z}}_{0,G}^{(0)} + (1 - \frac{n_2}{n_1}) \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\mathbf{Z}}_{1,G} + \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \quad (23a)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,psmall,3p}) = \hat{\boldsymbol{\alpha}}_{s_2}^t \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_{0,G}^{(0)}} \hat{\boldsymbol{\alpha}}_{s_2} + \frac{n_2}{n_1} \hat{\mathbf{Z}}_{0,G}^{(0)t} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} \hat{\mathbf{Z}}_{0,G}^{(0)} + (1 - \frac{n_2}{n_1}) \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\mathbf{Z}}_{1,G} + \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \quad (23b)$$

with the variance-covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_{0,G}^{(0)}} = \frac{1}{n_{0,G}(n_{0,G} - 1)} \sum_{x \in s_{0,G}} (\mathbf{Z}^{(0)}(x) - \hat{\mathbf{Z}}_{0,G}^{(0)}) (\mathbf{Z}^{(0)}(x) - \hat{\mathbf{Z}}_{0,G}^{(0)})^t \quad (24)$$

The **external variance** is defined as:

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,small,3p}) = \frac{1}{n_{1,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}^{(0)}(x)) + (1 - \frac{n_{2,G}}{n_{1,G}}) \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \quad (25a)$$

$$\begin{aligned} \hat{\mathbb{V}}_{ext}(\hat{Y}_{G,psmall,3p}) &= \frac{1}{n_{0,G}} \hat{\mathbb{V}}_{s_{2,G}}(Y(x)) + (1 - \frac{n_{1,G}}{n_{0,G}}) \frac{1}{n_{1,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}^{(0)}(x)) \\ &+ (1 - \frac{n_{2,G}}{n_{1,G}}) \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \end{aligned} \quad (25b)$$

where $\hat{R}^{(0)}(x) = Y(x) - \hat{Y}^{(0)}(x)$ with $\hat{Y}^{(0)}(x) = \mathbf{Z}^{(0)t}(x) \hat{\boldsymbol{\alpha}}_{s_2}$.

The **synthetic** (*synth*) and **pseudo synthetic estimator** (*psynth*) can be applied if no terrestrial samples are available in the small area, i.e. $n_{2,G} = 0$. Consequently, the residual correction and the residual variation term of the full model can no longer be applied and drops from the point estimate (Eq. 26a and 26b) and g-weight variance (Eq. 26c and 26d) formulas. The point estimates are again potentially biased since $\frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} \hat{R}(x) = 0$ for the full model residuals can not be ensured within small area G . Also the variance will be small but to the cost of ignoring the model uncertainties. Note that there is again no external variance formula for the synthetic and pseudo synthetic estimation.

$$\hat{Y}_{G,synth,3p} = (\bar{\mathbf{Z}}_{0,G}^{(0)} - \hat{\mathbf{Z}}_{1,G}^{(0)})^t \hat{\boldsymbol{\alpha}}_2 + \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} \quad (26a)$$

$$\hat{Y}_{G,psynth,3p} = (\hat{\mathbf{Z}}_{0,G}^{(0)} - \hat{\mathbf{Z}}_{1,G}^{(0)})^t \hat{\boldsymbol{\alpha}}_2 + \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\beta}}_{s_2} \quad (26b)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,synth,3p}) = \frac{n_2}{n_1} \bar{\mathbf{Z}}_{0,G}^{(0)t} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} \bar{\mathbf{Z}}_{0,G}^{(0)} + (1 - \frac{n_2}{n_1}) \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\mathbf{Z}}_{1,G} \quad (26c)$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,psynth,3p}) = \hat{\boldsymbol{\alpha}}_2^t \hat{\boldsymbol{\Sigma}}_{\hat{\mathbf{Z}}_{0,G}^{(0)}} \hat{\boldsymbol{\alpha}}_2 + \frac{n_2}{n_1} \hat{\mathbf{Z}}_{0,G}^{(0)t} \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} \hat{\mathbf{Z}}_{0,G}^{(0)} + (1 - \frac{n_2}{n_1}) \hat{\mathbf{Z}}_{1,G}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\mathbf{Z}}_{1,G} \quad (26d)$$

The **extended synthetic** (*extsynth*) and **extended pseudo synthetic estimator** (*extpsynth*) ensures that the residuals of the full model over both the entire inventory area F and the small area G are zero at the same time, i.e. $\frac{1}{n_2} \sum_{x \in s_2} \hat{R}(x) = \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} \hat{R}(x) = 0$. This is again realized by extending the vector of explanatory variables by a binary categorical indicator variable $I_G(x)$ which takes the value 1 if the observation lies inside the small area G and is otherwise set to 0. The extended auxiliary vector is thus defined as $\mathbf{Z}^t(x) = (\mathbf{Z}^{(0)t}(x), \mathbf{Z}^{(1)t}(x))$, where $\mathbf{Z}^{(0)t}(x) = (\mathbf{Z}^{(0)t}(x), I_G(x))$. In other words, when the *extended* option is chosen, **forestinventory** automatically adds the binary indicator variable for the desired small area for *all* observations in the input data frame (i.e. s_0). The regression coefficients, point estimates and variance estimates are calculated by replacing \mathbf{Z} with \mathbf{Z} (and likewise $\mathbf{Z}^{(0)}$ with $\mathbf{Z}^{(0)}$) into Eq. 16, 20, 25 and 26. Just as in the two-phase case, the resulting point estimates are now

5. Calculation of Confidence Intervals

Converting the estimated variance into a 95% *confidence interval* (CI) allows for a more practical interpretation of a point estimate's precision. The correct interpretation of a CI is not that there is a 95% probability that it contains the true value. In the design-based context, the true value of the population parameter we are trying to estimate, albeit unknown, is fixed and the sample is randomly generated under the sample design. Theoretically, if we were to repeatedly conduct the inventory using the same estimation method, estimator and auxiliary information under newly drawn random samples and calculate the 95% CI from each sample, then 95% of the CIs are expected to contain the true population parameter. The confidence level $1 - \alpha$ (e.g. 95%) is thus the expected frequency or proportion of possible confidence intervals to contain the unknown population parameter under resampling and is therefore often also referred to as *coverage rate*. The CI is also linked to hypothesis testing in that its associated point estimate is considered statistically different from any given value that lies outside the CI boundaries.

Based on the central limit theorem it can be assumed that under hypothetical repeated sampling the point estimates will asymptotically follow a normal distribution. However, on the recommendation of Mandallaz (2013a), better confidence intervals can be obtained using the Student's t distribution (Mandallaz (2013a)) defined as

One-Phase Estimation

$$CI_{1-\alpha}(\hat{Y}) = \left[\hat{Y} - t_{n_2-1, 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})}, \hat{Y} + t_{n_2-1, 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})} \right] \quad (27)$$

Two-Phase and Three-Phase Global Estimation

$$CI_{1-\alpha}(\hat{Y}) = \left[\hat{Y} - t_{n_2-p, 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})}, \hat{Y} + t_{n_2-p, 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})} \right] \quad (28)$$

Two-Phase and Three-Phase Small Area Estimation

$$CI_{1-\alpha}(\hat{Y}) = \left[\hat{Y} - t_{n_{2,G}-1, 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})}, \hat{Y} + t_{n_{2,G}-1, 1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{Y})} \right] \quad (29)$$

where \hat{Y} is the point estimate, $\hat{V}(\hat{Y})$ is the estimated variance, $1 - \alpha$ is the confidence level and p constitutes the number of parameters used in the (full) regression model. In case of cluster-sampling, $n_{2,G}$ is the number of terrestrial clusters (a cluster constitutes the sample unit and comprises multiple sample plots). In **forestinventory**, the confidence intervals for all estimation methods and estimators can be computed by the S3 generic method `confint()`, which requires an estimation object created by either the `onephase()`, `twophase()` or `threephase()` function. For example, the 95% confidence interval for the small area estimates by the `extpsynth` estimator (Section 3.2) are calculated by:

```
R> confint(extpsynth_2p)
```

```
95% Confidence Intervals for twophase small area estimation
```

	area	estimate	ci_lower_ext	ci_upper_ext	ci_lower_g	ci_upper_g
1	A	391.9356	325.6463	458.2250	324.9155	458.9558
2	B	419.7231	345.8418	493.6043	352.0456	487.4006

6. Special Cases and Scenarios

6.1. Post-Stratification

A special case of multi-phase regression estimators is *post-stratification*, which can further be divided into the cases of *multi-phase sampling for stratification* and *multi-phase sampling for regression within strata*. Both imply the use of one or more categorical variables in the regression model(s), leading to classical ANOVA and ANCOVA models.

To demonstrate *post-stratification*, we first create an artificial categorical variable *development stage* (stage) by clustering the mean canopy heights of the *grisons* data set into 3 height classes:

```
R> grisons$stage <- as.factor(kmeans(grisons$mean, centers = 3)$cluster)
```

Two-phase sampling for stratification is applied if the model only contains categorical variables, in this case the factor variable *stage*. Linear regression models only fitted with categorical variables produce ANOVA models, which when used in multi-phase regression estimators, is equivalent to post-stratification. For our example, this means that the model predictions are simply the means of the terrestrial response values within each development stage (within-strata means).

```
R> # two-phase sampling for stratification
R> twophase(formula=tvoll ~ stage,
+          data=grisons,
+          phase_id=list(phase.col = "phase_id_2p", terrgrid.id = 2),
+          boundary_weights = "boundary_weights")
```

Two-phase sampling for regression within strata implies the combination of continuous and categorical variables within the model (i.e. we have an ANCOVA model). If an interaction term is not present between categorical and continuous variables, the regression lines within the strata will have the same slope but different intercepts. If an interaction term is present, both the intercept and the slope are allowed to vary within the strata. Note that one can actually use the entire range of OLS regression techniques in the multi-phase estimators, including higher order terms and transformations of the explanatory variables, which makes them very flexible.

```
R> # two-phase sampling for regression within strata:
R> twophase(formula=tvoll ~ mean + stddev + max + q75 + stage,
+          data=grisons,
+          phase_id=list(phase.col = "phase_id_2p", terrgrid.id = 2),
+          boundary_weights = "boundary_weights")
```

The variance of all model-assisted estimators included in **forestinventory** can be decreased by reducing the sum of square residuals of the regression model. In case of post-stratification, this particularly implies minimizing the within strata residual square sum. Also, for post-stratification, the *g*-weight variance should be trusted over the external variance because it has the advantage that the strata weights are estimated from the large sample rather than the terrestrial sample s_2 .

6.2. Small Area Estimation under Cluster Sampling

As mentioned in Section 2.6, *cluster sampling* is a special case of sample designs where the sample consists of more than one spatially agglomerated sample points. One randomly places the sample location x in the inventory area as in the simple sampling design, but then $M - 1$ additional sample locations x_2, \dots, x_M are created close to the cluster origin x by adding a fixed set of spatial vectors

e_2, \dots, e_M to x . The idea of cluster sampling is to increase the amount of information without increasing the travel costs of the terrestrial campaign. However, the information gathered at all sub-locations of a cluster is then averaged on the cluster level, and this average value then references exactly one point, i.e. the cluster origin x . Without going into too much mathematical detail, the estimators under simple sampling are thus extended in a way that all parameters (local density, mean vector of explanatory variables, mean model residuals) have to be calculated as the weighted cluster means with $M(x)$ being the cluster weights. Whereas the geometric form and the number of sample locations per cluster M is fixed (i.e. defined by the inventorist), the actual number of points $M(x)$ falling into the forest area F at sample location x is random because the cluster origin x is random. The **forestinventory** package identifies clusters via a unique cluster ID that is assigned to a column in the input data set. Its column name is passed to the argument `cluster` in the `twophase()` and `threephase()` function calls.

For small area applications, the scenario might occur where the points of a cluster at sample locations x spread over more than one small area, i.e. only a subset $M_G(x) < M(x)$ is included in the small area of interest. In this case, the zero mean residual property within the small area, $\frac{\sum_{x \in s_{2,G}} M(x) \hat{R}_c(x)}{\sum_{x \in s_{2,G}} M(x)} = 0$, is no longer guaranteed when using the **extended** and **pseudo extended synthetic estimator** (see *extsynth* and *extpsynth* in Sections 3.2 and 4.2). In this case, it is advisable to use the **(pseudo) small area estimator** (*psmall*) where the zero mean residual property is still ensured.

In order to keep track of such cases, **forestinventory** tells the user to do so by returning a warning message:

```
R> extpsynth.clust <- twophase(formula = basal ~ stade + couver + melange, data=zberg,
+                             phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                             cluster = "cluster",
+                             small_area = list(sa.col = "ismallold", areas = c("1")),
+                             unbiased = TRUE)
```

Warning message:

At least one terrestrial cluster not entirely included within small area 1.

Zero mean residual assumption for small area maybe violated.

Check mean_Rc_x_hat_G and consider alternative estimator 'psmall'

```
R> extpsmall.clust <- twophase(formula = basal ~ stade + couver + melange, data=zberg,
+                             phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                             cluster = "cluster",
+                             small_area = list(sa.col = "ismallold", areas = c("1")),
+                             unbiased = TRUE),
+                             psmall = TRUE)
```

```
R> extpsynth.clust$estimation
```

```
area estimate ext_variance g_variance n1 n2 n1G n2G r.squared
1 1 25.54748 14.03806 14.16853 298 73 29 8 0.205741
```

```
R> extpsmall.clust$estimation
```

```
area estimate ext_variance g_variance n1 n2 n1G n2G r.squared
1 1 23.98581 16.30509 15.69473 298 73 29 8 0.1873795
```

Comparing the *extpsynth* and *psmall* estimates, we see that in this particular case the point estimates are close and more important, the external as well as the g-weight variances only differ marginally. This can be taken as evidence that the violation of the zero mean residual property can here be expected to have negligible consequences.

6.3. Violation of Nesting in Sample Design

As explained in Section 2, a basic prerequisite for the application of multi-phase estimators is that the sample phases (s_0, s_1, s_2) are *nested* in each other. The correct nesting thereby concerns the spatial arrangement of the sample phases (Fig. 2a) as well as the availability of terrestrial and auxiliary information per phase and sample location. For the latter, **forestinventory** runs validity checks in the background, provides warning and error messages and, if possible, applies first-aid adjustments to the inventory data set to prevent the calculations from failing. We will demonstrate possible nesting violations by applying the global three-phase estimator to the *grisons* and *zberg* data sets.

Violation 1

Based on the nesting rule, $s_2 \in s_1 \in s_0$, each s_2 and s_1 sample location *must* have all explanatory variables available that are used in the full (and thus reduced) regression model. If e.g. an s_2 and/or s_1 point misses a variable which is used in the full *and* reduced model (in this case *mean*), the function will **delete** this sample point from the dataset and produce the following messages:

```
R> ## delete "mean" value from an s2- (i.e. s1- and s0-) sample point:
R> grisons[which(grisons$phase_id_3p==2)[1], "mean"] <- NA

R> threephase(formula.s0 = tvol ~ mean,
+             formula.s1 = tvol ~ mean + stddev + max + q75,
+             data = grisons,
+             phase_id = list(phase.col="phase_id_3p", s1.id = 1, terrgrid.id = 2),
+             boundary_weights = "boundary_weights")
Warning messages:
1: In threephase(formula.s0 = tvol ~ mean, formula.s1 = tvol ~ mean + :
  Sample design not nested: for 1 terrestrial plots at least one auxiliary parameter
  of the first phase (s1) is missing
2: In threephase(formula.s0 = tvol ~ mean, formula.s1 = tvol ~ mean + :
  Sample design not nested: for 1 terrestrial plots at least one auxiliary parameter
  of the zero phase (s0) is missing
3: In threephase(formula.s0 = tvol ~ mean, formula.s1 = tvol ~ mean + :
  1 rows deleted due to missingness in the set of auxiliary parameters for the zero phase (s0)
  (1 terrestrial plots affected by deletion)
```

Violation 2

However, if an s_2 and/or s_1 point is missing a variable which is only used in the full regression model (in this example *q75*), the function will **recode** the phase indicator of that point to s_0 , since the point still provides the required information for the reduced model. If this concerns an s_2 sample location, the associated value of the response variable can no longer be used.

```
R> ## delete "q75" value from an s2-sample point:
R> grisons[which(grisons$phase_id_3p==2)[1], "q75"] <- NA

R> threephase(formula.s0 = tvol ~ mean,
+             formula.s1 = tvol ~ mean + stddev + max + q75,
+             data = grisons,
+             phase_id = list(phase.col="phase_id_3p", s1.id = 1, terrgrid.id = 2),
+             boundary_weights = "boundary_weights")
Warning messages:
1: In threephase(formula.s0 = tvol ~ mean, formula.s1 = tvol ~ mean + :
```

```

Sample design not nested: for 1 terrestrial plots at least one auxiliary parameter
of the first phase (s1) is missing
2: In threephase(formula.s0 = tvol ~ mean, formula.s1 = tvol ~ mean + :
  Changed the phase_id for 1 rows to the zero phase (s0) due to missingness in the set of
  auxiliary parameters for the first phase (s1) (1 terrestrial information no longer usable
  by this change)

```

Violation 3

If an s_0 point misses at least one of the explanatory variables used in the reduced model, the sample locations are **deleted** from the data set.

```

R> ## delete "mean" value from an s0-sample point:
R> grisons[which(grisons$phase_id_3p==0)[1], "mean"] <- NA

R> threephase(formula.s0 = tvol ~ mean,
+             formula.s1 = tvol ~ mean + stddev + max + q75,
+             data = grisons,
+             phase_id = list(phase.col="phase_id_3p", s1.id = 1, terrgrid.id = 2),
+             boundary_weights = "boundary_weights")
Warning message:
In threephase(formula.s0 = tvol ~ mean, formula.s1 = tvol ~ mean + :
  1 rows deleted due to missingness in the set of auxiliary parameters for the zero phase (s0)
  (0 terrestrial plots affected by deletion)

```

Note that all the automatic data adjustments (deletion, recoding) have to be accepted with caution. Recapitulating, the unbiasedness of estimators in the design-based framework is based on the uniform and independent randomization of the sample locations. This means that every possible location within the forest area F , as well as pairs of locations, have inclusion and joint inclusion probabilities greater than zero. Whereas this is already violated in practice by the use of regular grids, one can still expect that these grids do not exclude specific forest structures. If any information should be missing at the sample locations, one should clarify the reason for this and make sure that the information can reasonably be assumed to be *completely missing at random*.

Violation 4

If a categorical variable is used in the regression model(s) and the terrestrial sample s_2 is considerably small compared to the s_1 phase, it might occur that a category is only present in the $s_1 \setminus s_2$ sample, and thus missing in the s_2 sample. Therefore, an internal regression model cannot be calculated and the function **stops** with the following error message:

```

R> ## delete s2-points with "stade"-level '300'
R> zberg <- zberg[-which(zberg.n$phase_id_2p == 2 & zberg.n$stade=="300"), ]

R> twophase(formula = basal ~ stade + couver + melange,
+           data = zberg,
+           phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+           cluster = "cluster")
Error in check.mandatoryInputs(formula, data, phase_id) :
  Level '300' of factor variable 'stade' existing in s1(s0)- but not in s2 sample.
  Calculation of coefficient not feasible.

```

7. Analysis and Visualization

7.1. Analysis

We often want to compare the results and performances of different estimation methods and estimators for a given global or small area inventory, which can be easily accomplished in **forestinventory** using the `estTable()` function. This function restructures the results from the `onephase()`, `twophase()` and `threephase()` objects and merges them into one single data set that provides the basis for further analysis. For demonstration purposes, we will first recalculate the one-phase estimator as well as the two-phase and three-phase *extended pseudo synthetic* and the *pseudo synthetic* estimators for the `grisons` data set:

```
R> op <- onephase(formula = tvol~1, data = grisons,
+               phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+               area = list(sa.col = "smallarea", areas = c("A", "B", "C", "D")))

R> extpsynth_2p <- twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
+                          phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                          small_area = list(sa.col = "smallarea",
+                                             areas = c("A", "B", "C", "D"),
+                                             unbiased = TRUE))

R> psynth_2p <- twophase(formula = tvol ~ mean + stddev + max + q75, data = grisons,
+                       phase_id = list(phase.col = "phase_id_2p", terrgrid.id = 2),
+                       boundary_weights = "boundary_weights",
+                       small_area = list(sa.col = "smallarea",
+                                         areas = c("A", "B", "C", "D"),
+                                         unbiased = FALSE))

R> extpsynth_3p <- threephase(formula.s0 = tvol ~ mean,
+                            formula.s1 = tvol ~ mean + stddev + max + q75, data = grisons,
+                            phase_id = list(phase.col = "phase_id_3p", s1.id = 1, terrgrid.id = 2),
+                            small_area=list(sa.col = "smallarea", areas = c("A", "B", "C", "D"),
+                                           unbiased = TRUE),
+                            boundary_weights = "boundary_weights")

R> psynth_3p <- threephase(formula.s0 = tvol ~ mean,
+                          formula.s1 = tvol ~ mean + stddev + max + q75, data = grisons,
+                          phase_id = list(phase.col = "phase_id_3p", s1.id = 1, terrgrid.id = 2),
+                          small_area = list(sa.col = "smallarea",
+                                             areas = c("A", "B", "C", "D"),
+                                             unbiased = FALSE),
+                          boundary_weights = "boundary_weights")
```

We can then aggregate all estimation objects in a list and pass it to the `estTable()`-function:

```
R> grisons.sae.table <- estTable(est.list = list(op, extpsynth_2p, psynth_2p,
+                                             extpsynth_3p, psynth_3p),
+                              sae = TRUE,
+                              vartypes = c("variance", "g_variance", "ext_variance"))
```

The function merges the estimation results and returns a list object with the subclasses "esttable" "smallarea". The `vartypes` argument can be used to restrict the `estTable()` output

to certain types of variances. If one prefers the `data.frame` format for further analysis, this can easily be done using `as.data.frame(grisons.sae.table)`. Note however that **forestinventory** provides several S3 generic methods specifically for the class "esttable".

The structure of an `esttable` object is very similar to the objects created by the small area estimation functions of the package. However, the point estimates and variances from all estimation objects passed to `estTable()` have been stored in one single column (`estimate` and `variance`) and can be distinguished by the variables `method`, `estimator` and `vartype` which specify the estimation method (one, two or three-phase), the estimator and the type of variance that was applied (`g_` for *g-weight* and `ext_` for *external* variance). By default, the confidence intervals are also added.

```
R> str(grisons.sae.table)
```

```
List of 20
 $ area          : chr [1:28] "A" "A" "A" "A" ...
 $ domain        : Factor w/ 2 levels "global","smallarea": 1 2 2 2 2 2 1 2 2 ...
 $ method        : Factor w/ 3 levels "onephase","twophase",...: 1 3 3 3 2 2 1 3 3 ...
 $ estimator      : Factor w/ 3 levels "onephase","psynth extended",...: 1 2 2 3 2 2 3 1 2 2 ...
 $ vartype        : Factor w/ 3 levels "ext_variance",...: 3 1 2 2 1 2 2 3 1 2 ...
 $ estimate       : num [1:28] 410 395 395 422 391 ...
 $ variance       : num [1:28] 1987 1968 1858 726 996 ...
 $ std            : num [1:28] 44.6 44.4 43.1 26.9 31.6 ...
 $ error          : num [1:28] 10.86 11.22 10.91 6.39 8.07 ...
 $ n2             : num [1:28] 19 40 40 40 67 67 67 17 40 40 ...
 $ n2G            : num [1:28] NA 12 12 12 19 19 19 NA 11 11 ...
 $ n1             : num [1:28] NA 128 128 128 306 306 306 NA 128 128 ...
 $ n1G            : num [1:28] NA 38 38 38 94 94 94 NA 34 34 ...
 $ n0             : int [1:28] NA 306 306 306 NA NA NA NA 306 306 ...
 $ n0G            : int [1:28] NA 94 94 94 NA NA NA NA 81 81 ...
 $ r.squared      : num [1:28] NA NA NA NA 0.653 ...
 $ r.squared_reduced: num [1:28] NA 0.545 0.545 0.527 NA ...
 $ r.squared_full  : num [1:28] NA 0.724 0.724 0.717 NA ...
 $ ci_lower       : num [1:28] 317 298 300 367 325 ...
 $ ci_upper       : num [1:28] 504 493 490 476 457 ...
 - attr(*, "row.names")= int [1:28] 1 2 3 4 5 6 7 8 9 10 ...
 - attr(*, "class")= chr [1:3] "list" "esttable" "smallarea"
```

Note that `estTable()` also returns the *estimation error* (`error`) that is defined as the *standard error* divided by the *point estimate*:

$$error_{[\%]} = \frac{\sqrt{\hat{V}(\hat{Y})}}{\hat{Y}} \cdot 100 \quad (30)$$

As multi-phase estimation techniques are primary intended to increase estimation precision, the function `mphase.gain()` can be applied to quantify the potential benefit of a multi-phase global or small area estimate compared to its respective one-phase estimate. The function takes an `esttable` object as input and returns a summary of which multi-phase method and estimator performed best using the precision from the one-phase estimator as a *baseline*. If the `esttable` object contains more than one multi-phase estimation object, `mphase.gain()` identifies the one with the smallest variance and compares it to the onephase estimation. The argument `pref.vartype` can be used to define what type of variance (*g-weight* or *external*) should be used for the comparison. Synthetic estimates (*synth* and *psynth* estimator) are not considered for the comparison under the default setting (`exclude.synth = TRUE`) since they usually have a much smaller variance at the cost of a potential bias.

```
R> mphase.gain(grisons.sae.table, pref.vartype = "g_variance")
```

	area	var_onephase	var_multiphase	method	estimator	gain	rel.eff
1	A	1987.117	1016.9557	twophase	psynth extended	48.8	1.953986
2	B	3175.068	1019.2698	twophase	psynth extended	67.9	3.115041
3	C	1180.853	763.0731	threephase	psynth extended	35.4	1.547496
4	D	2290.652	1112.7346	twophase	psynth extended	51.4	2.058579

The function call returns a data frame containing the one-phase variance (`var_onephase`) and the variance of the best performing multi-phase estimator (`var_multiphase`). The multi-phase estimation procedure is again specified in the `method` and `estimator` column. The last two columns quantify the potential benefit of the multi-phase estimation. The `gain` is the reduction (if its value is positive) in variance when applying the multi-phase as alternative to the one-phase estimation. For example, it is indicated that the *two-phase extended psynth* estimation procedure for small area "B" leads to a 67.9 % reduction in variance compared to the one-phase procedure. The column `rel.eff` specifies the *relative efficiency* which is defined as the ratio between the one-phase variance and the multi-phase variance:

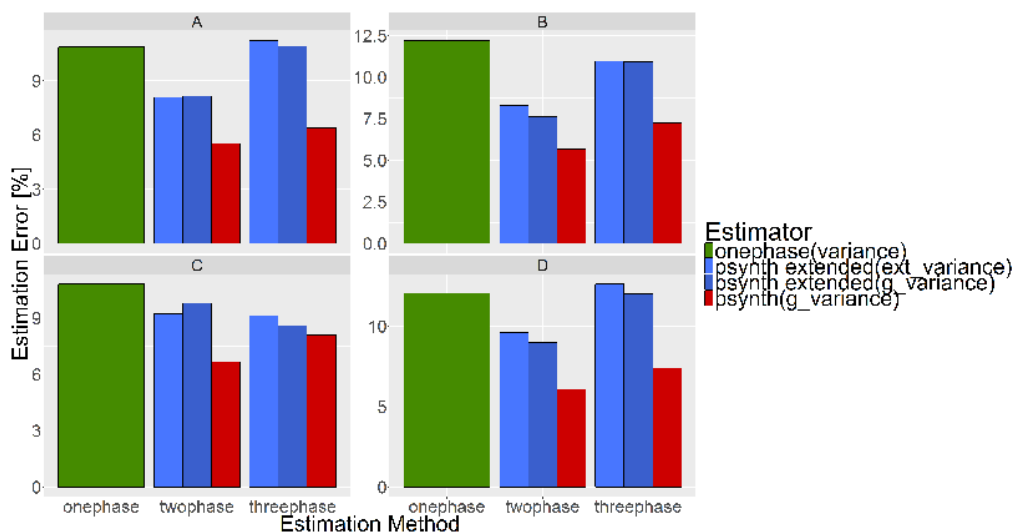
$$rel.eff[\%] = \frac{\hat{V}_{onephase}(\hat{Y})}{\hat{V}_{multiphase}(\hat{Y})} \cdot 100 \quad (31)$$

The *relative efficiency* can be interpreted as the relative sample size of the one-phase estimator needed to achieve the variance of the multi-phase estimator. For small area "B" we can thus see that we would have to increase the terrestrial sample size by factor 3 in the one-phase approach in order to get the same estimation precision as the *two-phase extended psynth* estimator. If the average costs for a terrestrial sample plot survey are known, the relative efficiency can thus be a simple means of quantifying the financial benefit of using multi-phase estimation for forest inventories.

7.2. Visualization

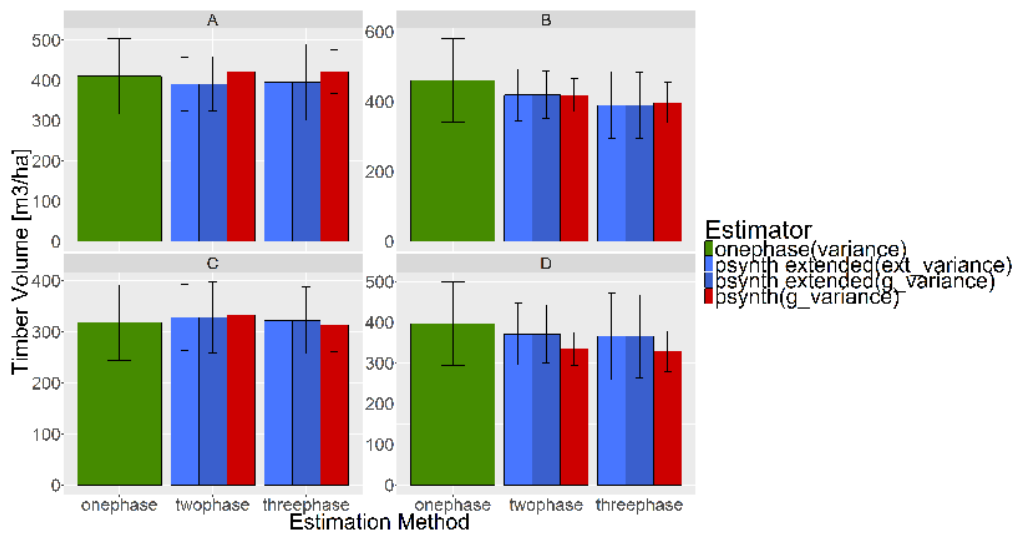
The `forestinventory` package also provides a S3 generic plot method based on the `ggplot2` package (Wickham 2009) to visualize the estimation results in two ways: 1) the *point estimates* with overlaid confidence intervals, and 2) the *estimation errors*. Both plots can be obtained by passing the `esstable` object to the `plot()` function.

```
R> plot(grisons.sae.table, ncol = 2)
```



Whereas the estimation errors are plotted by default, the point estimates and confidence intervals are returned when setting the argument `yvar = "estimate"`. Note that the graphics can arbitrarily be extended by additional **ggplot2** parameterizations.

```
R> plot(grisons.sae.table, ncol = 2, yvar = "estimate") +
+   ylab("Timber Volume [m3/ha]")
```



8. Future Plans

The **forestinventory** package currently provides a fairly well-rounded toolkit for forestry inventors to integrate auxiliary information into their estimates using the model-assisted methods under the design-based approach. Although 32 combinations of inventory scenarios, estimators and sample designs are covered, there are still potential improvements planned for the future. As this is an open-source project, everyone is encouraged to give feedback and/or make contributions on the package's development page on GitHub (Hill 2017). Currently planned extensions include:

- Implement parallel procedures for efficiently calculating many small areas.
- Allow functions to accept objects of class `data.table` from the **data.table** package (Dowle and Srinivasan 2017) to improve memory efficiency.
- Enable the user to choose other types of models than linear regressions fitted with OLS.

Acknowledgements

We want to express our gratitude to Prof. H. Heinimann (Chair of Land Use Engineering, ETH Zurich) for supporting this study and providing the possibility of working on the package. We also want to thank Daniel Mandallaz for his support in completing the range of the already published estimators in the frame of the three-phase small area estimators, as well as many helpful discussions and advice throughout the implementation of our package. Our thanks also go to Meinrad Abegg for proofreading the manuscript, and to the Amt für Wald und Naturgefahren of the Swiss canton of grisons for providing the example data.

References

- Battese GE, Harter RM, Fuller WA (1988). “An error-components model for prediction of county crop areas using survey and satellite data.” *Journal of the American Statistical Association*, **83**(401), 28–36. doi:10.1080/01621459.1988.10478561.
- Breidenbach J (2015). *JoSAE: Functions for some Unit-Level Small Area Estimators and their Variances*. R package version 0.2.3, URL <https://CRAN.R-project.org/package=JoSAE>.
- Breidenbach J, Astrup R (2012). “Small area estimation of forest attributes in the Norwegian National Forest Inventory.” *European Journal of Forest Research*, **131**(4), 1255–1267. doi:10.1007/s10342-012-0596-7.
- Cullmann AD (2016). *maSAE: Mandallaz’ Model-Assisted Small Area Estimators*. R package version 0.1-5, URL <https://CRAN.R-project.org/package=maSAE>.
- Dowle M, Srinivasan A (2017). *data.table: Extension of ‘data.frame’*. R package version 1.10.4-1, URL <https://CRAN.R-project.org/package=data.table>.
- Gregoire TG, Valentine HT (2007). *Sampling strategies for natural resources and the environment*. CRC Press.
- Hill A (2017). *forestinventory*. GitHub repository, URL <https://github.com/AndreasChristianHill/forestinventory>.
- Hill A, Mandallaz D, Buddenbaum H, Stoffels J, Langshausen J (2017). “Implementation of design-based small area estimations on forest district level in Rhineland-Palatinate by combining remote sensing data with data of the Third National German Inventory.” Third International Workshop on Forest Inventory Statistics, Freiburg.
- Köhl M, Magnussen SS, Marchetti M (2006). *Sampling methods, remote sensing and GIS multiresource forest inventory*. Springer Science & Business Media.
- Magnussen S, Mandallaz D, Breidenbach J, Lanz A, Ginzler C (2014). “National forest inventories in the service of small area estimation of stem volume.” *Canadian Journal of Forest Research*, **44**(9), 1079–1090. doi:10.1139/cjfr-2013-0448. URL <https://doi.org/10.1139/cjfr-2013-0448>.
- Mandallaz D (2008). *Sampling techniques for forest inventories*. CRC Press. doi:10.1201/9781584889779. URL <https://doi.org/10.1201/9781584889779>.
- Mandallaz D (2013a). “Design-based properties of some small-area estimators in forest inventory with two-phase sampling.” *Canadian Journal of Forest Research*, **43**(5), 441–449. doi:10.1139/cjfr-2012-0381. URL <https://doi.org/10.1139/cjfr-2012-0381>.
- Mandallaz D (2013b). “Regression estimators in forest inventories with three-phase sampling and two multivariate components of auxiliary information.” *Technical report*, Department of Environmental Systems Science, ETH Zurich. doi:10.3929/ethz-a-009990020. URL <https://doi.org/10.3929/ethz-a-009990020>.
- Mandallaz D (2013c). “Regression estimators in forest inventories with two-phase sampling and partially exhaustive information with applications to small-area estimation.” *Technical report*, Department of Environmental Systems Science, ETH Zurich. doi:10.3929/ethz-a-007623322. URL <https://doi.org/10.3929/ethz-a-007623322>.

- Mandallaz D (2013d). “A three-phase sampling extension of the generalized regression estimator with partially exhaustive information.” *Canadian Journal of Forest Research*, **44**(4), 383–388. doi:10.1139/cjfr-2013-0449. URL <https://doi.org/10.1139/cjfr-2013-0449>.
- Mandallaz D (2015). “Mathematical details of two-phase/two-stage and three-phase/two-stage regression estimators in forest inventories. Design-based Monte Carlo approach.” *Technical report*, Department of Environmental Systems Science, ETH Zurich. doi:10.3929/ethz-a-010441361. URL <https://doi.org/10.3929/ethz-a-010441361>.
- Mandallaz D, Breschan J, Hill A (2013). “New regression estimators in forest inventories with two-phase sampling and partially exhaustive information: a design-based monte carlo approach with applications to small-area estimation.” *Canadian Journal of Forest Research*, **43**(11), 1023–1031. doi:10.1139/cjfr-2013-0181. URL <https://doi.org/10.1139/cjfr-2013-0181>.
- Mandallaz D, Hill A, Massey A (2016). “Design-based properties of some small-area estimators in forest inventory with two-phase sampling - revised version.” *Technical report*, Department of Environmental Systems Science, ETH Zurich. doi:10.3929/ethz-a-010579388. URL <https://doi.org/10.3929/ethz-a-010579388>.
- Massey A, Mandallaz D (2015a). “Comparison of classical, kernel-based, and nearest neighbors regression estimators using the design-based Monte Carlo approach for two-phase forest inventories.” *Canadian Journal of Forest Research*, **45**(11), 1480–1488. doi:10.1139/cjfr-2015-0151. URL <https://doi.org/10.1139/cjfr-2015-0151>.
- Massey A, Mandallaz D (2015b). “Design-based regression estimation of net change for forest inventories.” *Canadian Journal of Forest Research*, **45**(12), 1775–1784. doi:10.1139/cjfr-2015-0266. <https://doi.org/10.1139/cjfr-2015-0266>, URL <https://doi.org/10.1139/cjfr-2015-0266>.
- Massey A, Mandallaz D, 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.” *Canadian Journal of Forest Research*, **44**(10), 1177–1186. doi:10.1139/cjfr-2014-0152. URL <https://doi.org/10.1139/cjfr-2014-0152>.
- Massey AF (2015). *Multiphase estimation procedures for forest inventories under the design-based Monte Carlo approach*. Ph.D. thesis, ETH Zurich. doi:10.3929/ethz-a-010536381. URL <https://doi.org/10.3929/ethz-a-010536381>.
- R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Saborowski J, Marx A, Nagel J, Böckmann T (2010). “Double sampling for stratification in periodic inventories-Infinite population approach.” *Forest ecology and management*, **260**(10), 1886–1895. doi:10.1016/j.foreco.2010.08.035. URL <https://doi.org/10.1016/j.foreco.2010.08.035>.
- Särndal CE, Swensson B, Wretman J (2003). *Model assisted survey sampling*. Springer Science & Business Media.
- Schreuder HT, Gregoire TG, Wood GB (1993). *Sampling methods for multiresource forest inventory*. John Wiley & Sons.
- von Lüpke N (2013). *Approaches for the optimisation of double sampling for stratification in repeated forest inventories*. Ph.D. thesis, University of Göttingen.

- von Lüpke N, Hansen J, Saborowski J (2012). “A three-phase sampling procedure for continuous forest inventory with partial re-measurement and updating of terrestrial sample plots.” *European Journal of Forest Research*, **131**(6), 1979–1990. ISSN 1612-4677. doi:10.1007/s10342-012-0648-z.
- von Lüpke N, Saborowski J (2014). “Combining double sampling for stratification and cluster sampling to a three-level sampling design for continuous forest inventories.” *European journal of forest research*, **133**(1), 89–100. doi:10.1007/s10342-013-0743-9. URL <http://dx.doi.org/10.1007/s10342-013-0743-9>.
- Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-0-387-98140-6. URL <http://ggplot2.org>.

Affiliation:

Andreas Hill
Department of Environmental Systems Science
Chair of Landuse Engineering
ETH Zürich
Universitätstrasse 16
8092 Zürich, Switzerland
E-mail: andreas.hill@usys.ethz.ch
Telephone: +41/44/632 32 36

Alexander Massey
Department of Environmental Systems Science
Chair of Landuse Engineering
ETH Zürich
Universitätstrasse 16
8092 Zürich, Switzerland
E-mail: afmass@gmail.com