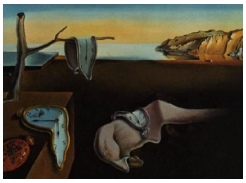

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

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

Andreas Hill
ETH Zürich

Alexander Massey

Abstract

Forest inventories provide reliable and evidence-based information to assess the state and development of forests over time. The information is collected at discrete sample locations distributed over the forest area by means of statistical sampling methods. This sample is then used to provide estimates of a target variable for a given spatial unit. Due to the high costs of the terrestrial campaigns, there is a need for alternative methods that can provide a) the same estimation precision at lower costs or b) higher estimation precision at identical costs. With respect to this objective, the application of double- and triple-sampling regression estimators (so-called *multiphase* forest inventory methods) has proved to be efficient. The core concept of multiphase methods is to combine the terrestrial sample with a much larger sample of target variable *predictions* based on *auxiliary information* that is available in high quantity and low costs. Whereas these methods have been successfully applied in practice, the availability of repective open-source software has been rare up to 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 covering small to large scale forest inventories, and can be used for double- and triple sampling for stratification, regression and regression within strata. This article provides a bridge from the mathematical summary of the estimators to their implementation and application in R.

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

1. Introdution

In many countries, forest inventories have become an integral instrument to judge the current state as well as the development of forests within reoccurring time periods. They provide quantitative information when it comes to define management actions and to adapt forest management strategies according to guidelines on national- and international level. Whereas the most straightforward approach of collecting the required information would be a full census of all trees within the forest area of interest, this remains only feasible for spatially small areas and is even then scarcely done in practice due to time-

and cost restrictions. For this reason, forest inventories gather their information of interest by means of statistical sampling methods, i.e. information is gathered only at discrete sample locations (*sample plots*) in the forest in the framework of a *terrestrial inventory*. This sample is then used to provide estimations of a target variable for a given spatial unit, and there is a broad range of concepts and methods regarding the choice of the sample design and respective estimators (Gregoire and Valentine 2007; Köhl, Magnussen, and Marchetti 2006; Schreuder, Gregoire, and Wood 1993; Mandallaz 2008). The information of terrestrial samples is assumed to be very precise, and increasing the precision of the estimates could primarily be achieved by increasing the terrestrial sample size. However, since collecting the terrestrial information is very time consuming and expensive, the number of terrestrial samples is usually limited. Whereas in national inventories the terrestrial sample size is still sufficient to provide high estimation accuracies on the national scale, small sample sizes on smaller spatial scales such as forest management units often cause high estimation errors that hamper using the inventory information.

Developments in recent years thus showed 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 so called *multiphase 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 auxiliary information (two-phase or double-sampling) or *two* additional samples of auxiliary information available in different sample sizes (three-phase or triple-sampling) (Gregoire and Valentine 2007; Saborowski, Marx, Nagel, and Böckmann 2010; Mandallaz 2013a,b; 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 these promising developments, a standard application of two and three-phase sampling methods in forest practice has been hampered by a lack of available software. One exception is the R package **JoSAE** by Breidenbach (2015) that provides the GREG (Särndal, Swensson, and Wretman 2003) and EBLUP (Battese, Harter, and Fuller 1988) two-phase small area estimator for simple sampling derived under the finite population approach. However, a more comprehensive software package covering a larger variety of sampling designs and estimators applicable to forest inventories has up to now been missing. Our motivation has been to addressed this lack by 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 2015; Mandallaz *et al.* 2013) and Germany (Hill, Mandallaz, Buddenbaum, Stoffels, and Langshausen 2017). The package comprises two- and three phase sampling regression estimators for global and small area estimations under simple and cluster sampling design and thus cover 32 inventory scenarios in total. The estimators can be used for 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 particular objective of this article is a) to establish the link between the mathematical description

of the estimators and their implementation in our package, b) to illustrate their application by the respective functions in our package to real-world inventory scenarios and c) to highlight special cases, i.e. rare inventory scenarios, and demonstrate how the package-functions deal with such situations (including error-checking functions and data adjustments).

2. Methods and Structure of the Package

2.1. Two-Phase Sampling

The two-phase or double-sampling estimators use inventory information from **two** nested samples which are commonly reffered to as *phases* (figure 1a). The first phase s_1 comprises n_1 sample locations that provide a set of explanatory variables described by the column vector $\boldsymbol{Z}(x) \in \Re^p$ 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 $Y(x)$ such as the timber volume density per hectare. The set of explanatory variables at each sample location $x \in s_1$ is now transformed into a prediction $\hat{Y}(x)$ of $Y(x)$ by the application of an oridinary least square (OLS) regression model. In the R package **forestinventory**, estimators for two-phase sampling can be applied by the `twophase`-function. Their implementation and application is described in detail in section 3 and 4.

2.2. Three-Phase Sampling

The three-phase or triple-sampling estimators extend the principle of two-phase sampling and use inventory information from **three** nested samples (*phases*). The basic assumption is that auxiliary information is available in two different sample sizes.

A first phase of auxiliary information (e.g. taken from remote sensing data) is used to generate model predictions based on multiple linear regression using the method of ordinary least squares. A subsample of the first phase comprises a second phase which contains further auxiliary information that produces another set of model predictions. A further subsample produces a third final phase based on terrestrial observations (i.e. the local densities of the ground truth) and is used to correct for bias in the design-based sense. The estimation method is available for simple and cluster sampling and includes the special case where the first phase is based on an exhaustive sample (i.e. a census). Small-area applications are supported for synthetic estimation as well as two varieties of bias-corrected estimators: the traditional small-area estimator and an asymptotically equivalent version derived under Mandallaz's extended model approach.

Explain three phase sampling, giving refrences to existing literature and studies

2.3. Small Area Estimation

Explain:

- What is the difference between *global* and *small area* estimation?
- explain the structure of the package (graphic) and state that we will only concentrate on specific cases ...

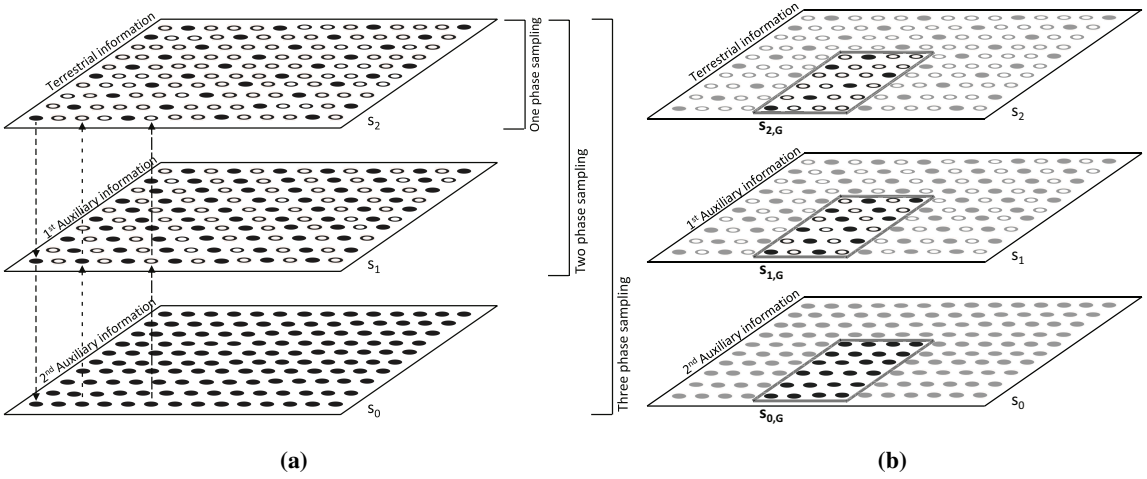

Figure 1: (a) Concept of multiphase sampling. The square represents the forest area for which an inventory is being conducted. The points denote the sample locations x. Filled points indicate *available* information. **(b)** Illustration of the Small Area Estimation problem

3. Global Estimators and their Application

3.1. Double Sampling (Two-Phase) Estimators

Mathematical Background

Mention:

- What is the difference between model-dependent and design-based? (bias-correction, dont have to believe in model predictions.
- explain the structure of the package (graphic) and state that we will only concentrate on specific cases ...

The regression coefficients of the OLS regression model are found by solving the sample-based normal equation. In case of **simple sampling**, the vector of regression coefficients are derived as

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

The design-based variance-covariance matrix of the regression coefficients is then calculated as

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

with the empirical residuals, i.e. the regression model residuals, available at all sample location $x \in s_2$ being

$$\hat{R}(x) = Y(x) - \hat{Y}(x) \tag{3}$$

The **point estimate** for simpe sampling is calculated according to equation 4. Note that this form results under particular case where the regression coefficients are derived using the data from the current inventory (**internal** regression model). In this case, the mean of the residuals $\frac{1}{n_2}\sum_{x\in s_2}R(x)$ is

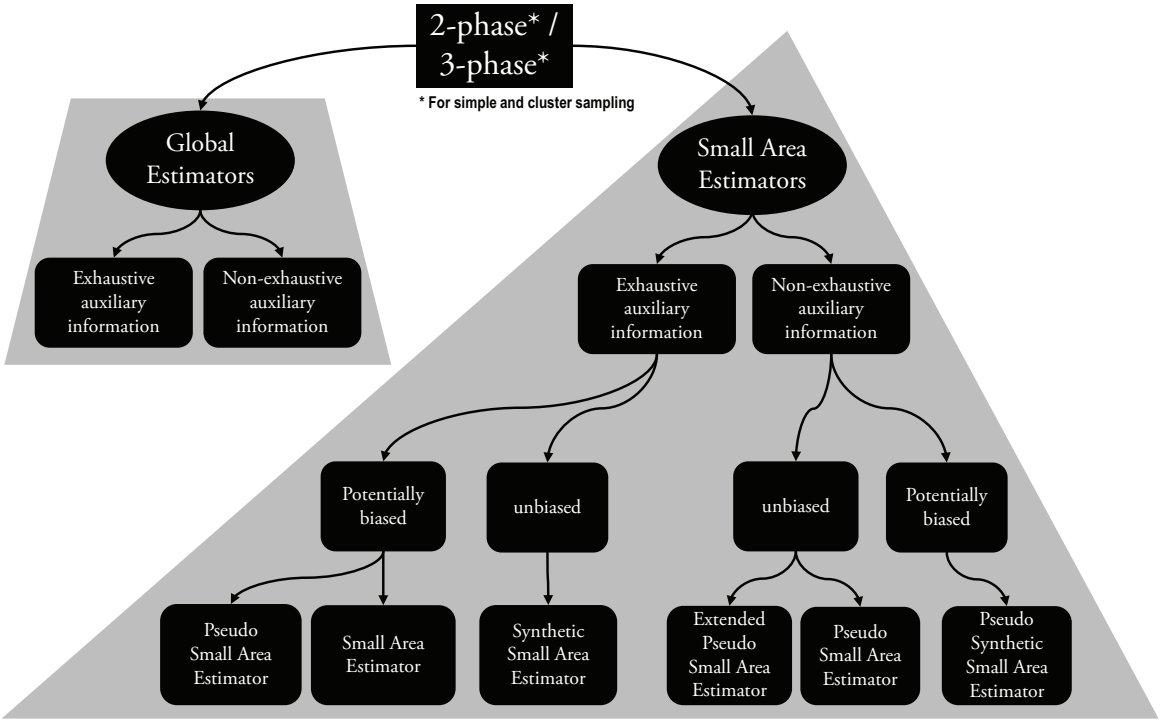

Figure 2: Structure of the multiphase estimators in the R package **forestinventory**

zero by definition and does not have to be added as is necessary for **external** models. Since in the package **forestinventory** only allows for internal models, the equation simplifies to

$$\hat{Y}_{reg} = \hat{\bar{\boldsymbol{Z}}}^t \hat{\boldsymbol{\beta}}_{s_2} \tag{4}$$

The estimation precision of the point estimate is specified by the estimated **design-based variance** as given in equation 5. Note that this is mathematically identical to the **g-weight** formulation of the design-based variance given in (Mandallaz, Hill, and Massey 2016). The package **forestinventory** additionally provides the **external** variance (equation 6). Note that the external variance neglects the uncertainty in the regression coefficients and is thus usually slightly lower than the design-based variance, where this uncertainty is considered by the variance-covariance matrix of the regression coefficients.

$$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \hat{\bar{\boldsymbol{Z}}}^t \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\bar{\boldsymbol{Z}}} \tag{5}$$

$$\hat{\mathbb{V}}(\hat{Y}_{reg}) = \frac{1}{n_1} \frac{1}{n_2 - 1} \sum_{x \in s_2} (Y(x) - \bar{Y}_2)^2 + (1 - \frac{n_2}{n_1}) \frac{1}{n_2} \frac{1}{n_2 - 1} \sum_{x \in s_2} (R(x) - \bar{R})^2 \tag{6}$$

Application

3.2. Triple Sampling (Three-Phase) Estimators

Mathematical Background

Application

4. Small Area Estimators and their Application

4.1. Double Sampling (Two-Phase) Estimators

4.2. Triple Sampling (Three-Phase) Estimators

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.

Gregoire TG, Valentine HT (2007). *Sampling strategies for natural resources and the environment*. CRC Press.

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). "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, 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 (2015). "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.

Saborowski J, Marx A, Nagel J, Böckmann T (2010). "Double sampling for stratification in periodic inventoriesâĂŤInfinite 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.

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
URL: http://www.lue.ethz.ch/people/hilla

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

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd