DISS. ETH NO. .....

Integration of Small Area Estimation Procedures of Timber Volume Resources in Large Scale Forest Inventories

A dissertation submitted to the
ETH ZURICH

for the degree of
Doctor of Sciences

presented by

Andreas Christian Hill
MSc. of Sciences

born 15 January 1986
citizen of Germany

accepted on the recommendation of

Prof. Dr. H.R. Heinimann, examiner
PD Dr. D. Mandallaz, supervisor, co-examiner
Prof. Dr. J. Saborowski, co-examiner
Dr. Johannes Breidenbach, co-examiner

Zurich, 2018

Contents

Summary **vii**

Zusammenfassung **ix**

1 General Introduction **1**
1.1 Short history of forest inventory . . . 2
1.2 Thesis objective and structure . . . 3

2 The *R* Package forestinventory: design-based global and small area estimations for multi-phase forest inventories **8**
2.1 Introduction . . . 10
2.2 Methods and structure of the package . . . 12
2.3 Two-phase estimators and their application . . . 16
2.4 Three-phase estimators and their application . . . 27
2.5 Calculation of confidence intervals . . . 33
2.6 Special cases and scenarios . . . 34
2.7 Analysis and visualization . . . 38
2.8 Future plans . . . 41

3 Combining canopy height and tree species map information for large scale timber volume estimations under strong heterogeneity of auxiliary data and variable sample plot sizes **43**
3.1 Introduction . . . 45
3.2 Materials and Methods . . . 47
3.3 Results . . . 54
3.4 Discussion . . . 61
3.5 Conclusion . . . 64

4 A double-sampling extension of the German National Forest Inventory for design-based small area estimation on forest district levels **66**
4.1 Introduction . . . 68
4.2 Terrestrial sampling design of the German NFI . . . 70
4.3 Double sampling in the infinite population approach . . . 70
4.4 Estimators . . . 71
4.5 Case study . . . 76
4.6 Results . . . 83
4.7 Discussion . . . 87
4.8 Conclusion . . . 89

5 Accuracy assessment of timber volume maps using forest inventory data and LiDAR canopy height models **92**
5.1 Introduction . . . 94
5.2 Materials and Methods . . . 96

5.3 Results 105
5.4 Discussion 109
5.5 Conclusion 111

6 Synthesis **113**
6.1 Main findings 114
6.2 Limitations and criticisms 115
6.3 Conclusion and implications for future work 116

Bibliography **119**

Curriculum Vitae **128**

List of Figures

2.1 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. 12

2.2 (a) Concept of multi-phase 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. 14

2.3 Structure of the multi-phase estimators in the *R* package **forestinventory**. 15

2.4 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 $\boldsymbol{Z}(x)$ is calculated within the supports (small squares) at each sample location x (points) that falls into the forest area (green underlying polygon). 18

3.1 Spatial distribution of the BWI3 cluster samples over Rhineland-Palatinate 48

3.2 Separate ALS acquisitions in Rhineland-Palatinate over the years. The colors also indicate the quality of the data: *light*: low point densities ($0.04/m^2$), *dark*: high point densities ($>4/m^2$). Blue semitransparent layer: state and communal forest area. 50

3.3 Identification (a) and visualization (b) of potential support radii used for calculating the predictor variables on plot level based on ECDF of maximum limiting distances of all BWI3 sample locations in RLP. 53

3.4 Classification accuracy for the main tree species of a sample location *before* and *after* calibration: *top)* overall accuracies. *bottom)* user's accuracies. *ind*: plot individual support sizes. 55

3.5 10-fold RMSE_{cv}[%] and adjusted R^2 realized under various support choices for the CHM and *treespecies* explanatory variables 57

3.6 Effect on the adjusted R^2 when substituting the actual main tree species with the predicted main tree species of a sample plot. Each point in the graph represents the timber volume regression model under different supports and threshold settings. The *dotted* line tracks the the model with the highest adjusted R^2 under the use of the error-free *treespecies* variable. Semitransparent colours for the data points are used to visualize overlap. 58

3.7 Visualization of the timber volume prediction function (*final regression model*) on sample plot level for each main plot tree species and ALS acquisition year. For visualization purposes, the predictor variable *stddev* was set to its average value within the respective *treespecies* and *ALSyear* categories. The terrestrially observed timber volume values are plotted in the background. 59

3.8 R^2-values of the final regression model, submodel 1 and submodel 2 achieved *within* the ALS acquisition year strata. 61

4.1 *Left*: Study area with delineated FA forest management units. *Right*: Example for each of the three management units (from top to bottom): FA, FR and forest stand unit overlayed with the extended double-sampling cluster design. *Green*: Forest stand polygon layer defining the forest area of this study. 77
4.2 Left: CHM (top) and tree species classification map (bottom) available on the federal state level. Right: Magnified illustration of the supports used to derive the explanatory variables from the auxiliary data. 80
4.3 Cumulative distribution of estimation errors under SRS, PSMALL, EXTPSYNTH and the PSYNTH estimator. *Left*: Results for the 45 FA units. *Right*: Results for the 388 (SRS), 321 (PSMALL, EXTPSYNTH) and 403 (PSYNTH) FR units. . . . 84
4.4 *Left*: Comparison of the g-weight variance between the PSMALL and the EXTPSYNTH estimator for the 321 FR units. *Right*: Difference in g-weight variance between the PSMALL and the EXTPSYNTH estimator in dependence of the terrestrial data ($n_{2,G}$) in the FR unit. 85
4.5 Cumulative distribution of variance reduction by the PSMALL and EXTPSYNTH compared to the SRS estimator for the 45 FA and 321 FR units. 86
4.6 Share of the overall variance by the residual term of the PSMALL estimator for various small area sample sizes. Points are scaled by the overall percentage reduction/increase of the variance compared to SRS. 87

5.1 Study site, including the distribution of the 67 field plots (regional forest inventory) that are part of the forest area derived from the TLM3D (Topographic Landscape Model) data (with approval of swisstopo JA100120/JD100042). 97
5.2 Terrestrial observed timber volume on plot level for all 67 sample plots of the study area (histogram bandwidth = 50 m^3/ha; origin = 0 m^3/ha). 98
5.3 Conceptual model illustrating the workflow of producing the timber volume map and assessing its accuracy. 99
5.4 Schematic design of the timber volume map. 104
5.5 Model-predicted timber volume on plot level against the observed timber volume of the 67 field plots; the distribution of the predictions and observations are also indicated on the respective axis. 105
5.6 Overall accuracies with 95% confidence intervals and corresponding kappa coefficients for a pre-defined number of classes, calculated for constant and locally-adapted class widths. 106
5.7 Volume map with model-predictions on a continuous scale for the entire study area covering 2000 hectares of forest with a spatial resolution of 25 meters; the subareas show the classified volume map using a constant class width of 200 m^3/ha (upper) and locally-optimized class widths for five classes (lower). 108

List of Tables

3.1 Descriptive statistics of the forest observed on NFI sample plots located within communal and state forest area (n=5791). 49
3.2 Accuracy metrics for submodels of final OLS regression model. p gives the number of parameters for each model. Interaction terms are indicated by ':'. 59
3.3 R^2, RMSE and RMSE% of final regression model within ALS acquisition year strata ($ALSyear$). $Area_{ALSyear}$: Area covered by ALS acquisition given in km^2. n: number of validation data. 60

4.1 Descriptive statistics of the forest observed on NFI sample plots located within communal and state forest area (n=5791). 78
4.2 Sample size for each phase in entire study area. $n_{\{1,2\},plots}$: number of plots. $n_{\{1,2\}}$: number of clusters. TSPEC: tree species map information. 79
4.3 Classification accuracies of the *treespecies* variable before and after calibration. n_{ref}: number of terrestrial reference plots. n_{class}: number of classified plots. 81
4.4 Model fit metrics for the two OLS regression models on the cluster level. Interaction terms are indicated by ':'. () give the respective values on the plot level. 82
4.5 R^2, RMSE and RMSE% on the cluster level of the full regression model within ALS acquisition year strata ($ALSyear$). $Area_{ALSyear}$: Area covered by ALS acquisition given in km^2. n: sample size of validation data. () give the respective values on the plot level. 82
4.6 Descriptive summary of point estimates and estimation errors on the two forest district levels. N_u: number of evaluated small area units. 83
4.7 Descriptive summary of variance reduction compared to SRS and relative efficiencies on the two forest district levels. N_u: number of evaluated small area units. 86

5.1 Summary statistics of the timber volume observed at the 67 terrestrial sample plots (given in m^2/ha). 98
5.2 Summary statistics of the timber volume observed at the 67 terrestrial sample plots (given in m^2/ha). 99
5.3 Summary statistics of the CHM metrics calculated at the 67 terrestrial sample plots (in meters). 101
5.4 Summary statistics of the regression model. 105
5.5 Constant class width of 200 m^3/ha (five timber volume classes); OAA (± 95% confidence interval) (%): 55.22 (42.58, 67.4); kappa: 0.36; square sum of class width: 200'000. 107
5.6 Optimized class width for five timber volume classes; OAA (± 95% confidence interval) (%): 64.18 (51.53, 75.53); kappa: 0.54; square sum of class width: 176'600. Class widths smaller than 200 m^3/ha are indicated by *. 107

Acknowledgements

Summary

The objective of this cumulative thesis was to contribute methods for incorporating auxiliary information, such as remote sensing data, in existing forest inventories. The overall focus of these methods was to increase the value of existing large scale inventories, which are usually characterised by low sampling frequencies, on small-scale management levels. This research question was investigated in two main studies.

The first study presented in this thesis focused on the application of latest design-based regression estimators for small area estimation. The objective of the study was to develop a double-sampling procedure for the German National Forest Inventory (NFI) and evaluate whether the implied combination of remote sensing data and the terrestrial NFI data can provide sufficient estimation precision on two small-scale management levels. The approach was applied by the example of timber volume estimation within two forest district levels in the German federal state of Rhineland-Palatinate. The work on this study is presented in three subsequent chapters: the first chapter gives a comprehensive review of model-assisted design-based small area regression estimators and demonstrates their implementation in a statistical software package that was developed for the purpose of this study. The second chapter addresses challenges that occurred during the identification of a suitable timber-volume regression model to be used in the estimators, i.e. 1) heterogeneity of the remote sensing data due to quality variations and time gaps to the terrestrial survey, 2) derivation of explanatory variables under angle count sampling, and 3) incorporation of tree species map information. The third chapter illustrates the implementation of the small area estimation procedure and evaluate its potential by comparing the estimation precisions to those derived under one-phase sampling, i.e. exclusively using the terrestrial NFI data available within each management unit. The major results were the following: it was demonstrated that on both management levels, the suggested double-sampling procedure was able to substantially reduce the estimation error compared to the standard one-phase approach. Additionally, post-stratification according to variables that reflect quality variations or temporal asynchronicity in the auxiliary information turned out to be an effective means to improve the precision of OLS regression models. The additional stratification to tree species led to a further improvement of model performance and suggests the use of such information in forthcoming inventories.

The second study was carried out in a model-dependent framework and investigated an alternative approach of deriving pixel-wise confidence intervals for forest attribute maps. The core approach was to use available inventory data for map validation by applying concepts of accuracy assessment usually known from evaluation of categorical classification results. The definition of intervals within the range of map predictions and their respective reference data were used to calculate the user's accuracies as the probability for each interval to reflect the ground truth. In this framework, an optimization algorithm was developed to automatically identify intervals that lead to optimal classification schemes. The method was tested on a timber volume map for a mountainous study site in Switzerland using data from the regional forest inventory as validation data. The results from the study suggested that resulting class accuracies can turn out to be substantially lower than overall model accuracy metrics such as R^2 or root mean square error might suggest. The validation method thus provides an additional accuracy metric that can be easily calculated and improves the information about the accuracy provided by forest attribute maps.

Zusammenfassung

Das Ziel dieser kumulativen Dissertation war der Beitrag zu Methoden für die Einbeziehung von Hilfsinformationen, wie beispielsweise Fernerkundungsdaten, in bereits bestehende Waldinventuren. Das übergeordnete Ziel war es dabei, den Wert von bestehenden Grossrauminventuren auf kleinräumigen Managementebenen zu erhöhen. Dies wurde in zwei Studien untersucht.

Die erste Studie, welche in dieser Dissertation vorgestellt wird, befasste sich mit der Anwendung design-basierter Kleingebietsschätzer. Das Ziel der Studie war es, ein zwei-phasiges Stichproben-Konzept für die Deutsche Nationalinventur zu entwickeln und zu evaluieren, ob eine entsprechende Kombination von Fernerkundungsdaten und terrestrischen Inventurdaten ausreichende Schätzgenauigkeiten auf zwei kleinräumigen Managementebenen ermöglicht.

Chapter 1

General Introduction

"Science never solves a problem without creating ten more."

— George Bernard Shaw

1.1 Short history of forest inventory

The first known conductions of forest inventories date back to the 14th and 15th century. They looked quite different from today's inventories and exclusively comprised a visual inspection that was carried out by riding or walking through the forests. These inspections were a means to acquire a representative impression about the state of large forest areas as well as to determine spatial units for harvesting (Zöhrer, 1980). Whereas these inspections constituted a response to already increasing wood shortages, the first reference to sustainable forest management in literature is only found two centuries later in the book *Silviculatura Oeconomica* written by von Carlowitz (1713), who suggested concepts that besides reforestation mostly targeted at a 'continuous and sustainable' use of wood. However, these concepts required updated, profound and reliable information about the forested areas.

While the first use of sample plots to gather representative information about the state and development of forests date back to Hartig (1795), major advancements of sample-based inventories came up at the beginning of the 20th century together with the development of statistical sampling methods. In North America, the first sample-based inventories (so-called *timber cruises*) were conducted around 1930. The surveys were initially performed by visual assessments and later by a full census of all trees along systematically arranged lines. The idea of these so-called *strip-sampling* inventories was to gather information only for a small percentage of the forest, and this information was then subsequently extrapolated to the entire forest area. In addition, the surveyed strips were also used to provide forest attribute maps. These strip sampling techniques were further developed by Goodspeed (1934) and Langballe & Fogh (1938) who both proposed to collect information only within sample plots aligned with the strip lines (line-plot sampling) in substitution for a complete census within the stripes as a means to reduce costs and to make the cruises more efficient.

In addition to improving the efficiency of sampling techniques, it also became of high importance to reliably quantify the estimation errors associated to estimated forest attributes. Solutions were developed in two mathematical frameworks of inference that rely on either randomization of the sampling process or sampling from an underlying stochastic process and are today known as the *design-based* and *model-dependent* approach. Randomization of the sample plot locations in order to allow for a valid estimate of the sampling error was first recommended by Hasel (1938). A defining advancement in design-based survey sampling constituted the concept of unequal probability sampling by Hansen & Hurwitz (1943), who showed that using inclusion probabilities proportional to the value of the target attribute (so-called *probability proportional to size* or *PPS* sampling) could substantially increase estimation precision. An unbiased estimator for unequal probability sampling was then contributed by Horvitz & Thompson (1952). The concept of *PPS* sampling is implemented in most of today's forest inventories by the use of concentric sample plots. A method which perfectly realizes PPS sampling is the angle-count sampling (ACS) technique introduced by Bitterlich in 1947 (Bitterlich, 1984), and it was Grosenbaugh (1958) who related ACS to the probabilistic sampling theory. A further important development was the reformulation of the design-based estimation frame, in particular the Horwitz-Thompson estimator, within the *infinite population approach* by Mandallaz (2008), which provided a much better definition of the underlying population for forest inventories compared to design-based survey methods for finite populations as applied in official statistics (Särndal et al., 2003, e.g.).

The huge advancements in the conduction of national forest inventories in the Nordic countries around 1920 considerably contributed to the development of estimators in the model-dependent framework, since these inventories used the concept of systematic strip sampling. Especially the variance estimators for systematic sampling by Matérn (1947, 1960) were used to quantify the estimation precisions (Kangas & Maltamo, 2006). Even while Matérns variance estimators relied on modelling spatial trends, it was Mandallaz (1993) who first applied geospatial kriging techniques in the field of forest inventory.

1.2 Thesis objective and structure

The objective of this thesis was to contribute methodologies to the recent developments in combining existing forest inventory data from field surveys with auxiliary data derived from remote sensing data. The particular focus was to investigate the potential increase in value of already existing large scale terrestrial forest inventories to be used on small-scale management levels. The main question of this thesis to be addressed was: what estimation accuracies can be realized using forest inventory data on much smaller spatial levels as their sampling intensities have originally been designed for? This question was investigated in two main studies: The first study (study 1) constitutes the main part of this thesis and concentrated on exploring the capabilities and performances of design-based multi-phase regression estimators in the service of small area estimation. The second study (study 2) investigated a new approach for evaluating the estimation accuracies of forest attribute maps, which are considered to be a special case of small area estimation. The following subsections will give a more detailed introduction to each of the studies.

1.2.1 Study 1: Design-based small area estimation

This study constituted the major work of the thesis and had the objective to develop and evaluate a double-sampling estimation procedure for the German National Forest Inventory (German NFI). The particular objective of the study was to investigate whether the use of the German NFI data can provide acceptable estimation precision on two forest district levels when incorporated in small area estimation procedures. Similar studies have been conducted in Norway (Breidenbach & Astrup, 2012) and Switzerland (Magnussen et al., 2014; Steinmann et al., 2013), but no extensive study had yet been available for Germany. The results from this study were considered to provide valuable evidence whether a double-sampling extension of the German NFI might be a cost-saving alternative to a regional terrestrial forest district inventory (FDI). It was thus a prerequisite to gather information over a sufficiently large number of small area units in order to allow for reliable conclusions. For this reason, the study was conducted in the German federal state Rhineland-Palatinate (RLP) where the German NFI was extended to a double-sampling design and three types of small area regression estimators were applied in order to derive point and variance estimates of mean standing timber volume on two forest district levels comprising 45 and 405 units respectively.

In this framework, we decided to particularly explore the performances of design-based regression estimators in the infinite population approach. Methods for this family of estimators have considerably been contributed to by the works of Mandallaz (2008, 2013a,d) and Mandallaz et al. (2013), and applications of the suggested regression estimators for global estimations have been intensively investigated by Massey (2015). Thus, our study was also a continuation in the application of these estimators for the special case of small area estimation. The design-based double-sampling estimators suggested by Mandallaz were also favored for the following reasons: First, the estimators are explicitly formulated for cluster sampling designs such as applied in the German NFI, which has not yet been the case for frequently used model-dependent estimators. Second, the design-based frame considerably relaxes requirements on the regression model which seemed appropriate facing severe quality restrictions in the auxiliary data of the study area. Third, the estimators provide the asymptotically unbiased g-weight variance estimator which a) accounts for the design-dependency of the regression coefficients on the sample under the commonly applied *internal model approach*, and b) is also robust to heteroscedasticity of model residuals.

The conduction of this study was divided into 3 work packages that each addressed major milestones towards the overall study objective. The work and results of these work packages are respectively presented in chapter 2, chapter 3 and chapter 4. In the following, we will give an introduction and some additional background information to each of these chapters.

Work Package 1: Software implementation

With respect to the study objective, work package 1 addressed the need of a robust and flexible software implementation of the design-based regression estimators that could handle large inventory data sets and process a large number of small area estimations at once. Whereas several of the estimators suggested by Mandallaz had been applied in simulations and real-world case studies (Mandallaz, 2013a; Mandallaz et al., 2013; Mandallaz, 2013d; Massey et al., 2014; Massey & Mandallaz, 2015a,b), there had yet not been an unified and consistent implementation of the estimators in the same software environment. The work on this study was thus taken as an opportunity to implement the full range of these regression estimators in the statistical software *R* (R Core Team, 2017). The implementation procedure comprised three steps in general: First, a comprehensive review of the regression estimators published by Mandallaz; second, the completion of yet missing estimators for three-phase small area estimation; and third, the actual implementation of the estimators in *R*. The latter seemed to be the software of choice, as it currently constitutes one of the most intensively used statistical software and also provides interfaces to data base systems in which inventory and geodata are commonly stored. A review of existing software for multi-stage and multi-phase estimation revealed that in comparison to official statistics, applications particularly suited for forest inventories have been rare. Exceptions are the *R* package **JoSAE** by Breidenbach (2015) and the **maSAE** package by Cullmann (2016). However, a more comprehensive software package covering a larger variety of sample designs and estimators - particularly in the design-based infinite population framework - had not yet been available. In order to address this lack between availability and recent interest in such methods, we also made our software package freely available (*R* package **forestinventory**) that can be installed from the CRAN server (`https://CRAN.R-project.org/package=forestinventory`). Chapter 2 describes the implementation and the application of the two-phase and three-phase estimators in *R* and provides a comprehensive review of the design-based regression estimators for global and small area estimation published by Mandallaz (2008, 2013a,d) and Mandallaz et al. (2013). The availability of the software package in combination with its comprehensive documentation also had the objective to support the transparency and the reproducibility of the methods applied in this thesis.

Work Package 2: Processing of auxiliary data and model building

The objective of work package 2 was to find a suitable ordinary least square (OLS) regression model to be used as *internal model* in the small area regression estimators. In order to apply the estimators to all management units in RLP, the regression model had to allow for predicting the standing timber volume of a German NFI sample plot at any location over the federal state forest area. This also imposed the restriction on the auxiliary data to be available at the federal state level. A similar large-scale study with the same model purpose had recently been published by Maack et al. (2016) for the German federal state of Baden-Württemberg. Likewise in this study, we also derived explanatory variables from country-wide airborne laser scanning (ALS) data, which were in our case however characterized by severe quality variations as well as time lags of up to 10 years between the ALS acquisitions and the terrestrial survey date. The objective of our study thus was to specifically address techniques to improve the performance of ordinary least square regression models under such restricting conditions. Additionally, the study also explored the use of tree species information derived from a country-wide tree species classification map as additional explanatory data. The integration of tree species information in timber volume prediction models has often been stated as some of the most promising but often missing and thus not well investigated information (Koch, 2010; White et al., 2016). In this context, one yet existing gap of knowledge also concerned the effect of species misclassifications, i.e. errors in the explanatory variables, on the precision of the regression model. We addressed this question by proposing a calibration technique for removing a potential bias in the regression coefficients caused by such misclassifications. An

additional challenge that further increased the complexity of the model selection procedure was the identification of optimal extraction areas (*supports*) for the explanatory variables under varying plot sizes due to the angle count sampling technique applied in the German NFI. The overall question of the study was whether the frame of an OLS regression model provided enough flexibility to cope with the mentioned challenges in the data set. Besides these modeling-specific aspects, the work on this study comprised the integration and storage of both the terrestrial NFI data and the remote sensing data in a PostgreSQL database using a PostGIS extension. The latter allowed for a georelational storage and query of both data sources and provided fast computation of explanatory variables for large data sets.

Work Package 3: Small area estimation

Work package 3 comprised the actual application of the regression estimators for small area estimation and built upon a synthesis of the methods developed in work package 1 and 2. The aim of this study was to finally investigate which accuracies can be realized for timber volume estimation on small scale forest management units when using the German NFI data in the implemented small area regression estimators (work package 1). This first comprised an extension of the existing NFI sample grid in the study area (RLP) to a double-sampling cluster design, and the derivation of the explanatory variables used in the regression model (work package 2) at each sample location. Three types of design-based small area regression estimators were then applied to derive point and variance estimates of mean standing timber volume within 45 and 405 forest districts (*Forstämter* and *Forestreviere*). The small area estimators we considered were the *pseudo-small*, *extended pseudo-synthetic* and the *pseudo-synthetic* design-based small area estimator for cluster sampling suggested by Mandallaz (2013a); Mandallaz et al. (2013). An evaluation of the error distribution of these estimators on both small area levels served as a first means to quantify the estimation accuracies realizable under each estimator. The estimation results of the multi-phase estimators were also compared to the one-phase estimator for cluster sampling that exclusively uses the terrestrially observed data available within a small area unit in order to specify the gain in efficiency provided by the suggested double-sampling procedure. The results of our evaluations were subsequently used to discuss the potential of the suggested design-based regression estimators for future applications with respect to alternative auxiliary data and transferability to change estimation.

1.2.2 Study 2: Mapping

Forest attribute maps provide an area-wide overview of important information such as development stages, tree species or growing conditions and have thus always been of high interest for forest practitioners. For a long time, such maps were exclusively produced by hand and required expensive field visits and visual inspection of aerial photography. This amount of work also hampered a frequent updating of the maps. However, the production of maps (*mapping*) covering large areas has lately been substantially supported by the availability of exhaustive remote sensing data in combination with modeling techniques (Brosofske et al., 2014). Especially maps of predicted forest attributes in high resolution are considered to support the spatially precise allocation of management operations such as harvesting. An example is the use of rasterized timber volume prediction maps for the optimal allocation of cable roads in the frame of harvesting in steep slope mountainous terrains (Bont & Heinimann, 2012; Bont et al., 2015).

Despite the advantages of providing such high-resolution predictions maps, one should however have in mind that the predictions are often made for considerably small spatial units (map pixels). In most cases, the map pixels match the extent of an inventory sample plot on which the prediction model has been calibrated. One can thus interpret mapping as an extreme case of *small area*

estimation. Design-based double-sampling estimators for small area estimation (study 1) provide closed-form variance formulas that allow for quantifying the estimation precision for every small area unit individually by its estimation error and confidence interval. However, these concepts of quantifying the uncertainty cannot be transferred to mapping approaches in the particular case of a small area unit (i.e. map pixel) corresponding to the size of a sample plot. It is thus necessary that similar efforts than for model building are invested in methods that reliably quantify the resulting map accuracies. A common way to characterize the map accuracy is to use metrics such as the coefficient of determination (R^2) or cross-validated root mean square error (RMSE), which rather address the overall prediction performance than the accuracy of individual predictions. For this reason, the specification of confidence intervals on pixel level has been stated as an important contribution to map accuracy assessment (McRoberts, 2010).

In case of continuous response variables such as standing timber volume, the application of linear regression models allows for providing a confidence region for each prediction (i.e. pixel) based on the *prediction interval* (Fahrmeir et al., 2013, pp.136–139). The objective of the study presented in Chapter 3 was to investigate an alternative approach of deriving pixel-wise confidence intervals by applying well-known concepts of accuracy assessment for categorical classification results (Congalton & Green, 2008). The core of the suggested approach was the definition of intervals within the range of terrestrial data and their respective model predictions, and subsequently calculate the *user's accuracy* for each of those intervals. The calculated users' accuracies can then be regarded as the confidence levels for the chosen intervals. In this framework, we demonstrated an optimization algorithm (heuristic search method) that - given a pre-defined number of intervals - automatically identifies the interval boundaries with respect to the best possible classification accuracies. The motivation for the development of this method was twofold: first, to provide a map user with the possibility to evaluate the map detail, i.e. number of intervals/classes, in dependence of the realizable prediction accuracies; and second, to allow for identifying intervals of the response variable for which the map produces considerably high or low prediction accuracies.

The suggested methods were applied in a mountainous study site in the canton of Grisons (Switzerland). We used the regional forest district inventory data in combination with data from an airborne laser scanning acquisition to produce a map of the standing timber volume on sample plot level, on which we subsequently applied the developed accuracy assessment. The setup of this study also addressed the overall question of the thesis, i.e. what prediction accuracies can be realized on small spatial scales when using forest inventory data that are only available in comparatively low sampling frequencies.

Chapter 2

The *R* Package forestinventory: design-based global and small area estimations for multi-phase forest inventories

Andreas Hill[1], Alexander Massey[1]

[1]ETH Zürich
Department of Environmental Systems Science,Universitätstrasse 16, 8092 Zürich, Switzerland

Submitted to:
Journal of Statistical Software (accepted).

– Alexander Massey is co-author of the R package *forestinventory* and supported writing of the manuscript.

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 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 multi-phase 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.

2.1 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 & Valentine, 2007; Köhl et al., 2006; Schreuder et al., 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 inventorists 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 (two-phase or double-sampling) or two additional samples available in different sample sizes (three-phase or triple-sampling) (Gregoire & Valentine,

2007; Saborowski et al., 2010; Mandallaz, 2013a,d; von Lüpke et al., 2012). (Gregoire & Valentine, 2007; Saborowski et al., 2010; Mandallaz, 2013a,d; von Lüpke et al., 2012). Their application to existing forest inventory systems have already showed their efficiency in terms of cost reduction and gain in estimation precision (Breidenbach & Astrup, 2012; von Lüpke & Saborowski, 2014; Mandallaz et al., 2013; Magnussen & Tomppo, 2014; Massey et al., 2014).

Multi-stage and multi-phase estimation has already been implemented in commercial as well as open-source software, such as the survey sampling procedures in *SAS* (*SAS* Institute Inc., 2015) and the **survey** package in *R* (Lumley, 2016). However, both are targeted towards list-sampling as it is applied in official statistics. Available software providing multi-phase sampling methods better suited for forest inventories has been rare. Two exceptions are the *R* package **JoSAE** by Breidenbach (2015) and the **maSAE** package by Cullmann (2016). 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 & Mandallaz, 2015b; Mandallaz et al., 2013) and Germany (Hill et al., 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.

2.2 Methods and structure of the package

2.2.1 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 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. 2.1. 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. 2.1 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 $Y(x)$ 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.

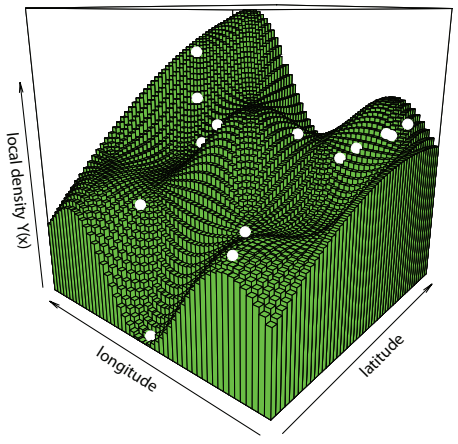


Figure 2.1: Artificial representation of a local density surface. The spatial distribution of a hypothetical density function for every point in a forested area is represented as a wavy piecewise constant green surface. Sample plots (white dots) inform the inventorist of the value of the density function 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.

2.2.2 Two-phase sampling

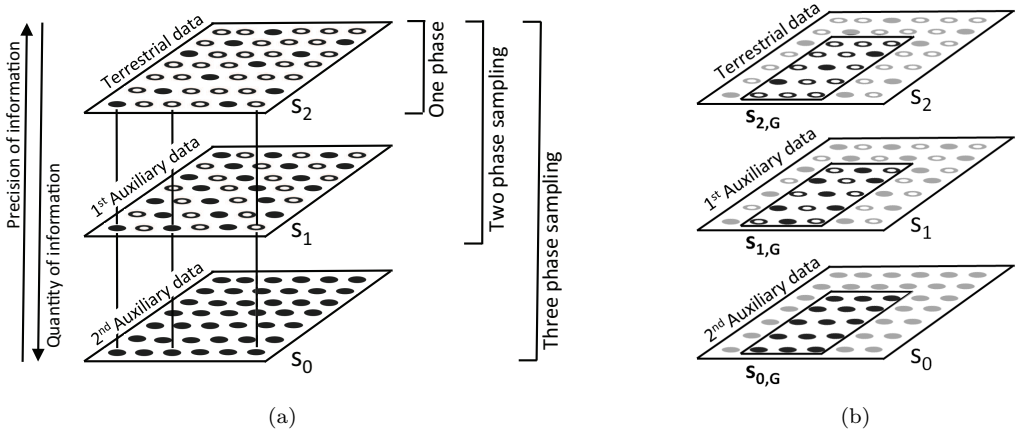
The two-phase or double-sampling estimators use inventory information from two nested samples which are commonly referred to as phases (Fig. 2.2a). The first phase s_1 comprises n_1 sample locations that each provide a set of explanatory variables described by the column vector $\boldsymbol{Z}(x) = (z(x)_1, z(x)_2, ..., z(x)_p)^\top$ 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 $x \in s_1$, $\boldsymbol{Z}(x)$ 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 2.2.5).

2.2.3 Three-phase sampling

Three-phase estimators extend the principle of two-phase sampling and use inventory information from three nested samples (phases) (Fig. 2.2a). 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 $\boldsymbol{Z}^{(1)}(x)$, adds considerable explanatory power to the set of variables available at all sample locations $x \in s_0$, denoted $\boldsymbol{Z}^{(0)}(x)$. From that it follows that we can define two nested regression models. The full set of predictor variables $\boldsymbol{Z}^\top(x) = (\boldsymbol{Z}^{(0)\top}(x), \boldsymbol{Z}^{(1)\top}(x))$ 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 $\boldsymbol{Z}^{(0)}(x)$. 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.

2.2.4 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. 2.2b). 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.



(a) (b)

Figure 2.2: (a) Concept of multi-phase 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.

2.2.5 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 density surface 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).

2.2.6 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. 2.3). 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 can also 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 et al., 2016; Mandallaz, 2013c,b). A special case under cluster sampling is described in Section 2.6.

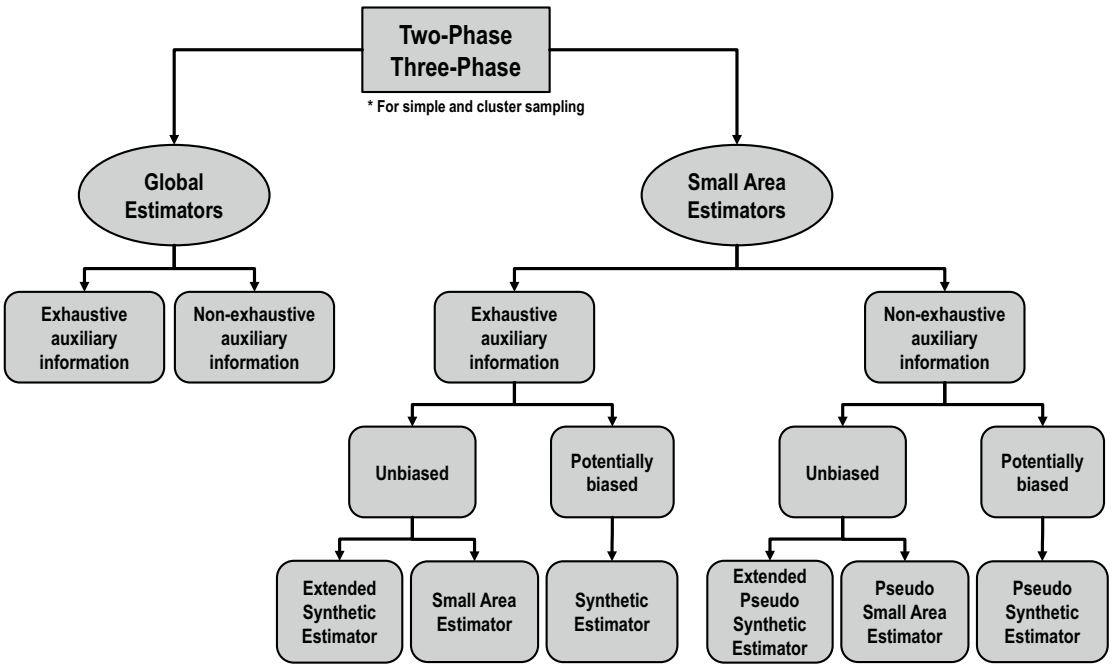


Figure 2.3: Structure of the multi-phase estimators in the *R* package **forestinventory**.

2.3 Two-phase estimators and their application

2.3.1 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:

$$\hat{\boldsymbol{\beta}}_{s_2} = \Big(\frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big)^{-1}\Big(\frac{1}{n_2}\sum_{x\in s_2}Y(x)\boldsymbol{Z}(x)\Big) \tag{2.1}$$

The individual predictions can then be calculated as $\hat{Y}(x) = \boldsymbol{Z}^\top(x)\hat{\boldsymbol{\beta}}_{s_2}$ 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 inventorists 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 & 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 $\bar{\boldsymbol{Z}}$, i.e., the exact spatial mean of $\boldsymbol{Z}(x)$, or to $\hat{\bar{\boldsymbol{Z}}}$, i.e., an estimate of the spatial mean of $\boldsymbol{Z}(x)$:

$$\hat{Y}_{reg2p,EX} = \bar{\boldsymbol{Z}}^\top\hat{\boldsymbol{\beta}}_{s_2} \tag{2.2a}$$

$$\hat{Y}_{reg2p,NEX} = \hat{\bar{\boldsymbol{Z}}}^\top\hat{\boldsymbol{\beta}}_{s_2} \tag{2.2b}$$

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 obain 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

$$\hat{\mathbb{V}}(\hat{Y}_{reg2p,EX}) := \bar{\boldsymbol{Z}}^\top\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\bar{\boldsymbol{Z}} \tag{2.3a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{reg2p,NEX}) := \hat{\bar{\boldsymbol{Z}}}^\top\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}} + \hat{\boldsymbol{\beta}}_{s_2}^\top\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}}\hat{\boldsymbol{\beta}}_{s_2} \tag{2.3b}$$

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

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} := \Big(\frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big)^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big)\Big(\frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big)^{-1} \tag{2.4}$$

and the uncertainty caused by using the s_1 sample to estimate $\bar{\boldsymbol{Z}}$ by $\hat{\bar{\boldsymbol{Z}}}$ is accounted for by the

variance-covariance matrix of the auxiliary vector $\hat{\bar{\boldsymbol{Z}}}$

$$\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}} = \frac{1}{n_1(n_1-1)} \sum_{x \in s_1} (\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}})(\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}})^\top \tag{2.5}$$

The external variance formulation for linear regression models is

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{reg2p,EX}) = \frac{1}{n_2} \hat{\mathbb{V}}_{s_2}(\hat{R}(x)) \tag{2.6a}$$
$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{reg2p,NEX}) = \frac{1}{n_1} \hat{\mathbb{V}}_{s_1}(\hat{Y}(x)) + \frac{1}{n_2} \hat{\mathbb{V}}_{s_2}(\hat{R}(x))$$

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. 2.4). 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.

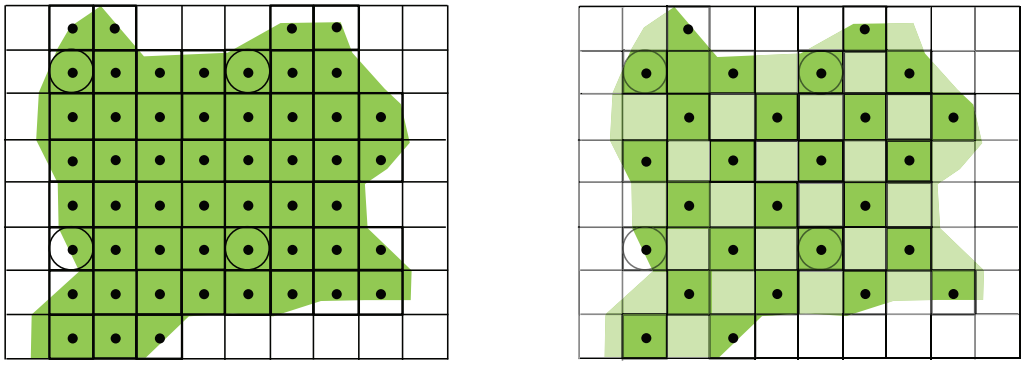
Calculation of explanatory variables

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. 2.4b 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 $\boldsymbol{Z}(x) = (z(x)_1, z(x)_2, ..., z(x)_p)^\top$ 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 $\boldsymbol{Z}(x)$ 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 $\boldsymbol{Z}(x)$ automatically from the input data frame using $\hat{\bar{\boldsymbol{Z}}} = \frac{1}{n_1} \sum_{x \in s_1} \boldsymbol{Z}(x)$.

The exhaustive case requires a closer look. In the infinite population approach, $\boldsymbol{Z}(x)$ refers to the sample point x and not the area around it. Deriving the exact spatial mean, $\bar{\boldsymbol{Z}} = \frac{1}{\lambda(F)} \int_F \boldsymbol{Z}(x)dx = (\frac{1}{\lambda(F)} \int_F z_1(x)dx, ..., \frac{1}{\lambda(F)} \int_F z_p(x)dx)^\top$, implies that we need to calculate the spatial mean of each component of $\boldsymbol{Z}(x)$ 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 $\boldsymbol{Z}(x)$ over an infinite number of sample points (i.e., $n_1 = \infty$). Although it is practically infeasible to assess $\boldsymbol{Z}(x)$ 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 $\bar{\boldsymbol{Z}}$ that is only negligibly different.

One implementation to approximate the exact mean $\bar{\boldsymbol{Z}}$ is shown in Fig. 2.4a, 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. Is has to be noted that this setup would allow for a perfect calculation of the exact mean $\bar{\boldsymbol{Z}}$ in the finite population approach, i.e., deriving $\boldsymbol{Z}(x)$ 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 $\bar{\boldsymbol{Z}}$, 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 $\bar{\boldsymbol{Z}}$. With this respect, the infinite population approach provides more flexibility than the finite approach.



(a) (b)

Figure 2.4: 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 $\boldsymbol{Z}(x)$ is calculated within the supports (small squares) at each sample location x (points) that falls into the forest area (green underlying polygon).

Boundary adjustment

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. 2.4). 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 $\hat{\bar{\boldsymbol{Z}}}$ or $\bar{\boldsymbol{Z}}$, the respective means have then to be calculated as the weighted mean (Eq. 2.7) where the weight $w(x)$ is equal to the percentage of forested area within the support of sample location x.

$$\hat{\bar{\boldsymbol{Z}}} = \frac{\sum_{x \in s_1} w(x)\boldsymbol{Z}(x)}{\sum_{x \in s_1} w(x)} \tag{2.7}$$

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 airborne laserscanning (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	1.00	9.30	11.84	40.87	21.14	C	107.80
2	1	1.00	12.16	11.35	39.80	21.54	A	NA
3	2	1.00	5.25	5.74	23.82	9.53	D	63.77
4	1	1.00	7.53	9.33	34.10	13.02	A	NA
5	1	0.67	6.11	5.87	23.33	10.55	B	NA
6	1	1.00	12.15	10.16	33.76	20.97	C	NA
7	2	1.00	6.38	4.72	17.96	10.14	D	154.10
8	1	1.00	1.25	3.79	22.72	0.00	B	NA
9	1	1.00	21.56	7.49	32.66	27.81	A	NA
10	2	1.00	13.55	7.20	36.14	18.59	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. Note that **forestinventory** implicitly assumes that all rows not indcated as s_2 belong to the s_1 phase.
- 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., $\bar{\boldsymbol{Z}}$, 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 $\bar{\boldsymbol{Z}}$ 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 $\hat{\bar{\boldsymbol{Z}}}$ by their exact means $\bar{\boldsymbol{Z}}$.

2.3.2 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. 2.1 and 2.4, 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).

$$\hat{Y}_{G,SMALL,2p} = \bar{\boldsymbol{Z}}_G^\top \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}(x) \tag{2.8a}$$

$$\hat{Y}_{G,PSMALL,2p} = \hat{\bar{\boldsymbol{Z}}}_G^\top \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}(x) \tag{2.8b}$$

In the equations for the point estimates (Eq. 2.8a and 2.8b), we see that the globally derived regression coefficients are applied to the exhaustively or non-exhaustively calculated means of the explanatory variables ($\bar{\boldsymbol{Z}}_G$, $\hat{\bar{\boldsymbol{Z}}}_G$) 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. 2.9a, 2.9b) as well as the external variance (Eq. 2.10a, 2.10b). 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 $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}$.

$$\hat{\mathbb{V}}(\hat{Y}_{G,SMALL,2p}) := \bar{\boldsymbol{Z}}_G^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \bar{\boldsymbol{Z}}_G + \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.9a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,PSMALL,2p}) := \hat{\bar{\boldsymbol{Z}}}_G^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\bar{\boldsymbol{Z}}}_G + \hat{\boldsymbol{\beta}}_{s_2}^\top \hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_G} \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.9b}$$

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,SMALL,2p}) := \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.10a}$$

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,PSMALL,2p}) := \frac{1}{n_{1,G}} \hat{\mathbb{V}}_{s_{2,G}}(Y(x)) + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right) \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.10b}$$

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 vector over the small area $\hat{\bar{\boldsymbol{Z}}}_G = \frac{1}{n_{1,G}} \sum_{x \in s_{1,G}} \boldsymbol{Z}(x)$

is replaced by its weighted version $\hat{\bar{\boldsymbol{Z}}}_G = \frac{\sum_{x \in s_{1,G}} w(x)\boldsymbol{Z}(x)}{\sum_{x \in s_{1,G}} w(x)}$, 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. 2.11a and 2.11b) are based only on the predictions generated by applying the globally derived regression model to the auxiliary vectors $\bar{\boldsymbol{Z}}_G$ and $\hat{\bar{\boldsymbol{Z}}}_G$ 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. 2.8a and 2.8b). 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. 2.11c and 2.11d). 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.

$$\hat{Y}_{G,SYNTH,2p} = \bar{\boldsymbol{Z}}_G^\top \hat{\boldsymbol{\beta}}_{s_2} \tag{2.11a}$$

$$\hat{Y}_{G,PSYNTH,2p} = \hat{\bar{\boldsymbol{Z}}}_G^\top \hat{\boldsymbol{\beta}}_{s_2} \tag{2.11b}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,SYNTH,2p}) = \bar{\boldsymbol{Z}}_G^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \bar{\boldsymbol{Z}}_G \tag{2.11c}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,PSYNTH,2p}) = \hat{\bar{\boldsymbol{Z}}}_G^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} \hat{\bar{\boldsymbol{Z}}}_G + \hat{\boldsymbol{\beta}}_{s_2}^\top \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_G} \hat{\boldsymbol{\beta}}_{s_2} \tag{2.11d}$$

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

$$\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_G} = \frac{1}{n_{1,G}(n_{1,G}-1)} \sum_{x \in s_{1,G}} (\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}}_G)(\boldsymbol{Z}(x) - \hat{\bar{\boldsymbol{Z}}}_G)^\top \tag{2.12}$$

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 ensuring that the mean of the empirical prediction model residuals 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 $\boldsymbol{Z}(x)$ 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 $\mathbb{Z}^\top(x) = (\boldsymbol{Z}^\top(x), I_G^\top(x))$ and can be used to replace its non-extended counterpart $\boldsymbol{Z}^\top(x)$ whereever it is used in Eq. 2.11 and 4.15. 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

$$\hat{\boldsymbol{\theta}}_{s_2} = \Big(\frac{1}{n_2} \sum_{x \in s_2} \mathbb{Z}(x)\mathbb{Z}^\top(x)\Big)^{-1} \Big(\frac{1}{n_2} \sum_{x \in s_2} Y(x)\mathbb{Z}(x)\Big) \tag{2.13}$$

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

$$\hat{Y}_{G,EXTSYNTH,2p} = \bar{\mathbf{Z}}_G^\top \hat{\boldsymbol{\theta}}_{s_2} \tag{2.14a}$$

$$\hat{Y}_{G,EXTPSYNTH,2p} = \hat{\bar{\mathbf{Z}}}_G^\top \hat{\boldsymbol{\theta}}_{s_2} \tag{2.14b}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,EXTSYNTH,2p}) = \bar{\mathbf{Z}}_G^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}} \bar{\mathbf{Z}}_G \tag{2.14c}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,EXTPSYNTH,2p}) = \hat{\bar{\mathbf{Z}}}_G^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{s_2}} \hat{\bar{\mathbf{Z}}}_G + \hat{\boldsymbol{\theta}}_{s_2}^\top \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathbf{Z}}}_G} \hat{\boldsymbol{\theta}}_{s_2} \tag{2.14d}$$

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

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,EXTSYNTH,2p}) = \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2,G}}(\hat{\mathbb{R}}(x)) \tag{2.15a}$$

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,EXTPSYNTH,2p}) = \frac{1}{n_{1,G}} \hat{\mathbb{V}}_{s_{2,G}}(Y(x)) + \Big(1 - \frac{n_{2,G}}{n_{1,G}}\Big) \frac{1}{n_{2,G}} \hat{\mathbb{V}}_{s_{2_G}}(\hat{\mathbb{R}}(x)) \tag{2.15b}$$

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

To summarize, the synthetic estimators SYNTH and PSYNTH 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 specifiying 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),
+    small_area = list(sa.col = "smallarea", areas = c("A", "B"),
+    unbiased = TRUE), psmall = TRUE, boundary_weights = "boundary_weights")

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 ~ 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),
+    small_area = list(sa.col = "smallarea", areas = c("A", "B"),
+    unbiased = TRUE), boundary_weights = "boundary_weights")
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^2s (`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 2.6).

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 desireable 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),
+    small_area = list(sa.col = "smallarea", areas = c("A", "B"),
+    unbiased = FALSE), boundary_weights = "boundary_weights")

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 EXTPSYNTH estimator. However, PSMALL and EXTPSYNTH 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. 2.8a, 2.9a, 2.10a, 2.11a, 2.11c) 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`.

2.4 Three-phase estimators and their application

2.4.1 Global estimators

Mathematical background

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

$$\hat{\boldsymbol{\alpha}}_{s_2} = \Big(\frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{Z}^{(0)}(x)\boldsymbol{Z}^{(0)\top}(x)\Big)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\boldsymbol{Z}^{(0)}(x) \tag{2.16a}$$

$$\hat{\boldsymbol{\beta}}_{s_2} = \Big(\frac{1}{n_2}\sum_{x\in s_2}\boldsymbol{Z}(x)\boldsymbol{Z}^{\top}(x)\Big)^{-1}\frac{1}{n_2}\sum_{x\in s_2}Y(x)\boldsymbol{Z}(x) \tag{2.16b}$$

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.

$$\begin{aligned}\hat{Y}_{reg3p,EX} &= \frac{1}{\lambda(F)}\int_F \boldsymbol{Z}^{(0)\top}(x)\hat{\boldsymbol{\alpha}}_{s_2} + \frac{1}{n_1}\sum_{x\in s_1}(\boldsymbol{Z}^{\top}(x)\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{Z}^{(0)\top}(x)\hat{\boldsymbol{\alpha}}_{s_2}) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x) - \boldsymbol{Z}^{\top}(x)\hat{\boldsymbol{\beta}}_{s_2}) \\ &= (\bar{\boldsymbol{Z}}_0^{(0)} - \hat{\bar{\boldsymbol{Z}}}_1^{(0)})^{\top}\hat{\boldsymbol{\alpha}}_{s_2} + \hat{\bar{\boldsymbol{Z}}}_1^{\top}\hat{\boldsymbol{\beta}}_{s_2} \end{aligned} \tag{2.17a}$$

$$\begin{aligned}\hat{Y}_{reg3p,NEX} &= \frac{1}{n_0}\sum_{x\in s_0}\boldsymbol{Z}^{(0)\top}(x)\hat{\boldsymbol{\alpha}}_{s_2} + \frac{1}{n_1}\sum_{x\in s_1}(\boldsymbol{Z}^{\top}(x)\hat{\boldsymbol{\beta}}_{s_2} - \boldsymbol{Z}^{(0)\top}(x)\hat{\boldsymbol{\alpha}}_{s_2}) + \frac{1}{n_2}\sum_{x\in s_2}(Y(x) - \boldsymbol{Z}^{\top}(x)\hat{\boldsymbol{\beta}}_{s_2}) \\ &= (\hat{\bar{\boldsymbol{Z}}}_0^{(0)} - \hat{\bar{\boldsymbol{Z}}}_1^{(0)})^{\top}\hat{\boldsymbol{\alpha}}_{s_2} + \hat{\bar{\boldsymbol{Z}}}_1^{\top}\hat{\boldsymbol{\beta}}_{s_2} \end{aligned} \tag{2.17b}$$

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 $\boldsymbol{Z}^{(0)}(x)$ over both the s_0 and s_1 phase, as well as the estimated mean of $\boldsymbol{Z}(x)$ over the s_1 phase are calculated according to Eq. 2.18. If the exact mean over s_0 is known, the estimated mean $\hat{\bar{\boldsymbol{Z}}}_0^{(0)}$ can simply be replaced by the exact mean $\bar{\boldsymbol{Z}}_0^{(0)}$. Note that in case of applied boundary adjustment (Section 2.3), the simple mean is again replaced by the weighted mean analogous to Eq. 2.7.

$$\hat{\bar{\boldsymbol{Z}}}_0^{(0)} = \frac{1}{n_0}\sum_{x\in s_0}\boldsymbol{Z}^{(0)}(x), \quad \hat{\bar{\boldsymbol{Z}}}_1^{(0)} = \frac{1}{n_1}\sum_{x\in s_1}\boldsymbol{Z}^{(0)}(x), \quad \hat{\bar{\boldsymbol{Z}}}_1 = \frac{1}{n_1}\sum_{x\in s_1}\boldsymbol{Z}(x) \tag{2.18}$$

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

$$\hat{\mathbb{V}}(\hat{Y}_{reg3p,EX}) = \frac{n_2}{n_1}\bar{\boldsymbol{Z}}^{(0)\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}^{(0)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_1^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_1 \tag{2.19a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{reg3p,NEX}) = \hat{\boldsymbol{\alpha}}_{s_2}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_0^{(0)}}\hat{\boldsymbol{\alpha}}_{s_2} + \frac{n_2}{n_1}\hat{\bar{\boldsymbol{Z}}}_0^{(0)\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_0^{(0)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_1^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_1 \tag{2.19b}$$

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

$$\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_0^{(0)}} = \frac{1}{n_0(n_0-1)} \sum_{x \in s_0} (\boldsymbol{Z}^{(0)}(x) - \hat{\bar{\boldsymbol{Z}}}_0^{(0)})(\boldsymbol{Z}^{(0)}(x) - \hat{\bar{\boldsymbol{Z}}}_0^{(0)})^\top \tag{2.20a}$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}} = \Big(\frac{1}{n_2} \sum_{x \in s_2} \boldsymbol{Z}^{(0)}(x)\boldsymbol{Z}^{(0)\top}(x)\Big)^{-1} \Big(\frac{1}{n_2^2} \sum_{x \in s_2} \hat{R}^{(0)2}(x)\boldsymbol{Z}^{(0)}(x)\boldsymbol{Z}^{(0)\top}(x)\Big) \Big(\frac{1}{n_2} \sum_{x \in s_2} \boldsymbol{Z}^{(0)}(x)\boldsymbol{Z}^{(0)\top}(x)\Big)^{-1} \tag{2.20b}$$

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}} = \Big(\frac{1}{n_2} \sum_{x \in s_2} \boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big)^{-1} \Big(\frac{1}{n_2^2} \sum_{x \in s_2} \hat{R}^2(x)\boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big) \Big(\frac{1}{n_2} \sum_{x \in s_2} \boldsymbol{Z}(x)\boldsymbol{Z}^\top(x)\Big)^{-1} \tag{2.20c}$$

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

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{reg3p,EX}) = \frac{1}{n_1}\hat{\mathbb{V}}_{s_2}(\hat{R}^{(0)}(x)) + (1 - \frac{n_2}{n_1})\frac{1}{n_2}\hat{\mathbb{V}}_{s_2}(\hat{R}(x)) \tag{2.21a}$$

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{reg3p,NEX}) = \frac{1}{n_0}\hat{\mathbb{V}}_{s_0}(\hat{Y}^{(0)}(x)) + \frac{1}{n_1}\hat{\mathbb{V}}_{s_2}(\hat{R}^{(0)}(x)) + (1 - \frac{n_2}{n_1})\frac{1}{n_2}\hat{\mathbb{V}}_{s_2}(\hat{R}(x)) \tag{2.21b}$$

where $\hat{\mathbb{V}}_{s_0}$ 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
R> formula.fm <- tvol ~ mean + stddev + max + q75
```

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
 372.0896     454.4064   451.3626 306 128 40          0.527363
 r.squared_full
      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.

2.4.2 Small area estimators

Mathematical background

The three two-phase small area estimators described in Section 2.3.2 can also be extended to the three-phase scenario. The general principle thereby stays the same, i.e., the regession coefficients of the reduced and full model and their variance-covariance matrices are calculated on the entire s_2 sample according to Eq. 2.16a, 2.16b, 2.20b and 2.20c, and are subsequently used to make predictions for sample locations restricted to small area G.

The unbiased point estimates of the SMALL and PSMALL estimator 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., $\hat{R}_G(x) = Y_G(x) - \boldsymbol{Z}_G^\top(x)\hat{\boldsymbol{\beta}}_{s_2}$, to the point estimate. The difference between the mean $\hat{\bar{\boldsymbol{Z}}}_{1,G}^{(0)}$ and the more precise or exact mean $\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)}$ and $\bar{\boldsymbol{Z}}_{0,G}^{(0)}$ is again considered as a correction term likewise in the global estimation (Eq. 2.17).

$$\hat{Y}_{G,SMALL,3p} = (\bar{\boldsymbol{Z}}_{0,G}^{(0)} - \hat{\bar{\boldsymbol{Z}}}_{1,G}^{(0)})^\top \hat{\boldsymbol{\alpha}}_{s_2} + \hat{\bar{\boldsymbol{Z}}}_{1,G}^\top \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}_G(x) \tag{2.22a}$$

$$\hat{Y}_{G,PSMALL,3p} = (\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)} - \hat{\bar{\boldsymbol{Z}}}_{1,G}^{(0)})^\top \hat{\boldsymbol{\alpha}}_{s_2} + \hat{\bar{\boldsymbol{Z}}}_{1,G}^\top \hat{\boldsymbol{\beta}}_{s_2} + \frac{1}{n_{2,G}} \hat{R}_G(x) \tag{2.22b}$$

The g-weight variance is then calculated as

$$\hat{\mathbb{V}}(\hat{Y}_{G,SMALL,3p}) = \frac{n_2}{n_1}\bar{\boldsymbol{Z}}_{0,G}^{(0)\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}_{0,G}^{(0)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_{1,G}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{1,G} + \frac{1}{n_{2,G}}\hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.23a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,PSMALL,3p}) = \hat{\boldsymbol{\alpha}}_{s_2}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)}}\hat{\boldsymbol{\alpha}}_{s_2} + \frac{n_2}{n_1}\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_{1,G}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{1,G} + \frac{1}{n_{2,G}}\hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.23b}$$

with the variance-covariance matrix

$$\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)}} = \frac{1}{n_{0,G}(n_{0,G}-1)}\sum_{x\in s_{0,G}}(\boldsymbol{Z}^{(0)}(x)-\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)})(\boldsymbol{Z}^{(0)}(x)-\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)})^{\top} \tag{2.24}$$

The external variance is defined as:

$$\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,SMALL,3p}) = \frac{1}{n_{1,G}}\hat{\mathbb{V}}_{s_{2,G}}(\hat{R}^{(0)}(x)) + (1-\frac{n_{2,G}}{n_{1,G}})\frac{1}{n_{2,G}}\hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x)) \tag{2.25a}$$

$$\begin{aligned}\hat{\mathbb{V}}_{ext}(\hat{Y}_{G,PSMALL,3p}) = {} & \frac{1}{n_{0,G}}\hat{\mathbb{V}}_{s_{2,G}}(Y(x)) + (1-\frac{n_{1,G}}{n_{0,G}})\frac{1}{n_{1,G}}\hat{\mathbb{V}}_{s_{2,G}}(\hat{R}^{(0)}(x)) \\ & + (1-\frac{n_{2,G}}{n_{1,G}})\frac{1}{n_{2,G}}\hat{\mathbb{V}}_{s_{2,G}}(\hat{R}(x))\end{aligned} \tag{2.25b}$$

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

The synthetic (SYNTH) and pseudo synthetic (PSYNTH) estimator 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. 2.26a and 2.26b) and g-weight variance (Eq. 2.26c and 2.26d) 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.

$$\hat{Y}_{G,SYNTH,3p} = (\bar{\boldsymbol{Z}}_{0,G}^{(0)} - \hat{\bar{\boldsymbol{Z}}}_{1,G}^{(0)})^{\top}\hat{\boldsymbol{\alpha}}_2 + \hat{\bar{\boldsymbol{Z}}}_{1,G}^{\top}\hat{\boldsymbol{\beta}}_{s_2} \tag{2.26a}$$

$$\hat{Y}_{G,PSYNTH,3p} = (\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)} - \hat{\bar{\boldsymbol{Z}}}_{1,G}^{(0)})^{\top}\hat{\boldsymbol{\alpha}}_2 + \hat{\bar{\boldsymbol{Z}}}_{1,G}^{\top}\hat{\boldsymbol{\beta}}_{s_2} \tag{2.26b}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,SYNTH,3p}) = \frac{n_2}{n_1}\bar{\boldsymbol{Z}}_{0,G}^{(0)\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\bar{\boldsymbol{Z}}_{0,G}^{(0)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_{1,G}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{1,G} \tag{2.26c}$$

$$\hat{\mathbb{V}}(\hat{Y}_{G,PSYNTH,3p}) = \hat{\boldsymbol{\alpha}}_2^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)}}\hat{\boldsymbol{\alpha}}_2 + \frac{n_2}{n_1}\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\alpha}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{0,G}^{(0)} + (1-\frac{n_2}{n_1})\hat{\bar{\boldsymbol{Z}}}_{1,G}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{s_2}}\hat{\bar{\boldsymbol{Z}}}_{1,G} \tag{2.26d}$$

The extended synthetic (EXTSYNTH) and extended pseudo synthetic (EXTPSYNTH) estimator 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 $\mathbb{Z}^{\top}(x) = (\mathbb{Z}^{(0)\top}(x), \boldsymbol{Z}^{(1)\top}(x))$, where $\mathbb{Z}^{(0)\top}(x) = (\boldsymbol{Z}^{(0)\top}(x), I_G^{\top}(x))$. 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 $\boldsymbol{Z}$ with $\mathbb{Z}$ (and likewise $\boldsymbol{Z}^{(0)}$ with $\mathbb{Z}^{(0)}$) into Eq. 2.16, 2.20, 2.25 and

2.26. Just as in the two-phase case, the resulting point estimates are now unbiased and have an associated g-weight variance that accounts for the variability of the regression coefficients resulting from the random sample s_2.

Application

We will demonstrate the use of three-phase small area estimation in the package **forestinventory** by applying the EXTSYNTH and SYNTH estimator to the `grisons` data set. The setup is thus exactly the same as in the example for global three-phase estimation (Section 2.4.1). However, this time will use the exact auxiliary mean of the mean canopy height variable (`mean`) and assume that we do not know the exact means of the remaining explanatory variables `stddev`, `max` and `q75`. We thus first define the true means for each small area just as we did in the `twophase()` example (Section 2.3.2):

```
R> truemeans.G <- data.frame(Intercept = rep(1, 4),
+    mean = c(12.85, 12.21, 9.33, 10.45))
R> rownames(truemeans.G) <- c("A", "B", "C", "D")
```

Three-phase small area estimation in the package can in general be applied by additionally specifying the `small_area` list argument. The exhaustive estimators can be called by optionally passing a `data.frame` containing the exact auxiliary means to the `exhaustive` argument. The EXTSYNTH estimator can be applied by setting the argument `unbiased` to `TRUE` (default):

```
R> extsynth_3p <- threephase(formula.rm, formula.fm, 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"), unbiased = TRUE),
+    exhaustive = truemeans.G, boundary_weights = "boundary_weights")

R> extsynth_3p$estimation

  area estimate ext_variance g_variance  n0  n1 n2 n0G n1G n2G
1    A 382.6405     1642.055   1518.741 Inf 128 40 Inf  38  12
2    B 368.9013     1501.211   1530.576 Inf 128 40 Inf  34  11
  r.squared_reduced r.squared_full
1         0.5454824      0.7242913
2         0.5354637      0.7171512
```

The SYNTH estimator can be applied by changing the argument `unbiased` to `FALSE`, which causes the function to not apply a bias correction in the respective small area.

```
R> synth_3p <- threephase(formula.rm, formula.fm, 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"), unbiased = FALSE),
+    exhaustive = truemeans.G, boundary_weights = "boundary_weights")

R> synth_3p$estimation

  area estimate ext_variance g_variance  n0  n1 n2 n0G n1G n2G
1    A 409.3390           NA   410.7529 Inf 128 40 Inf  38  12
2    B 375.4608           NA   461.8250 Inf 128 40 Inf  34  11
  r.squared_reduced r.squared_full
1          0.527363      0.7166608
2          0.527363      0.7166608
```

We see that the `threephase()`-function returns the sample sizes in the entire inventory area as well as within each small area. The value `Inf` for `n0G` indicates that the explanatory variables at s_0 sample locations used in the reduced model were in our case derived exhaustively. If we compare

the two results, we see that the SYNTH estimation again yields a much smaller variance than the EXTSYNTH estimation, but at the cost of a potential bias.

We can also analyse how the exhaustive derivation of `mean` performed compared to the case where `mean` is non-exhaustively available but at a very large s_0 phase with $n_{0,G} >> n_{1,G}$. To do this, we additionally compute the EXTPSYNTH estimates. As we can see, the exhaustive derivation of `mean` only yielded a slightly smaller variance.

```
R> extpsynth_3p <- threephase(formula.rm, formula.fm, 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"), unbiased = TRUE),
+    boundary_weights = "boundary_weights")
```

```
R> extpsynth_3p$estimation
```

```
  area estimate ext_variance g_variance  n0  n1 n2 n0G n1G n2G
1    A 395.1882     1901.211   1858.204 306 128 40  94  38  12
2    B 389.8329     1846.995   1816.655 306 128 40  81  34  11
  r.squared_reduced r.squared_full
1         0.5454824      0.7242913
2         0.5354637      0.7171512
```

2.5 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 repreatedly 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 obtained using the Student's t distribution defined as

One-phase estimation:

$$CI_{1-\alpha}(\hat{Y}) = \left[\hat{Y} - t_{n_2-1,1-\frac{\alpha}{2}}\sqrt{\hat{\mathbb{V}}(\hat{Y})}, \hat{Y} + t_{n_2-1,1-\frac{\alpha}{2}}\sqrt{\hat{\mathbb{V}}(\hat{Y})}\right] \tag{2.27}$$

Two-phase and three-phase global estimation:

$$CI_{1-\alpha}(\hat{Y}) = \left[\hat{Y} - t_{n_2-p,1-\frac{\alpha}{2}}\sqrt{\hat{\mathbb{V}}(\hat{Y})}, \hat{Y} + t_{n_2-p,1-\frac{\alpha}{2}}\sqrt{\hat{\mathbb{V}}(\hat{Y})}\right] \tag{2.28}$$

Two-phase and three-phase small area estimation:

$$CI_{1-\alpha}(\hat{Y}) = \left[\hat{Y} - t_{n_{2,G}-1,1-\frac{\alpha}{2}}\sqrt{\hat{\mathbb{V}}(\hat{Y})}, \hat{Y} + t_{n_{2,G}-1,1-\frac{\alpha}{2}}\sqrt{\hat{\mathbb{V}}(\hat{Y})}\right] \tag{2.29}$$

where $\hat{Y}$ is the point estimate, $\hat{\mathbb{V}}(\hat{Y})$ 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). Note that in case of synthetic estimations (SYNTH, PSYNTH), the degrees of freedom are $n_2 - p$ as is the case for global estimation. 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 2.3.2) 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
```

2.6 Special cases and scenarios

2.6.1 Post-stratification

A special case of multi-phase regression estimation 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> twophase(formula = tvol ~ 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> twophase(formula=tvol ~ 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 design-based estimators included in **forestinventory** can be decreased by reducing the sum of squared 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.

2.6.2 Small area estimation under cluster sampling

As mentioned in Section 2.2.6, 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 $x_2, ..., x_M$ are created close to the cluster origin x by adding a fixed set of spatial vectors $e_2, ..., e_M$ 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 **forstinventory** 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 2.3.2 and 2.4.2). In this case, it is adviseable to use the (pseudo) small area estimator (SMALL or 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> psmall.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> psmall.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.

2.6.3 Violation of nesting in sample design

As explained in Section 2.2, 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. 2.2a) 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> 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> 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> 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 $s_1 \backslash s_2$ 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> 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.
```

2.7 Analysis and visualization

2.7.1 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 EXTPSYNTH and PSYNTH estimator 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), boundary_weights = "boundary_weights")

R> psynth_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 = FALSE), boundary_weights = "boundary_weights")

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 2 1 2  ...
 $ method           : Factor w/ 3 levels "onephase","twophase",..: 1 3 3 3 2 2 2  ...
 $ estimator        : Factor w/ 3 levels "onephase","psynth extended",..: 1 2 2 3 ...
 $ vartype          : Factor w/ 3 levels "ext_variance",..: 3 1 2 2 1 2 2 3 1 2 ...
 $ estimate         : num [1:28] 410 395 395 422 392 ...
 $ variance         : num [1:28] 1987 1901 1858 726 996 ...
 $ std              : num [1:28] 44.6 43.6 43.1 26.9 31.6 ...
 $ error            : num [1:28] 10.86 11.03 10.91 6.39 8.05 ...
 $ 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] 317 299 300 367 326 ...
 $ ci_upper         : num [1:28] 504 491 490 476 458 ...
 - 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 devided by the point estimate:

$$error_{[\%]} = \frac{\sqrt{\hat{\mathbb{V}}(\hat{Y})}}{\hat{Y}} \cdot 100 \tag{2.30}$$

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      1017.6327   twophase psynth extended 48.8 1.952686
2    B     3175.068      1019.1913   twophase psynth extended 67.9 3.115281
3    C     1180.853       763.0731 threephase psynth extended 35.4 1.547496
4    D     2290.652      1110.2454   twophase psynth extended 51.5 2.063194
```

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:

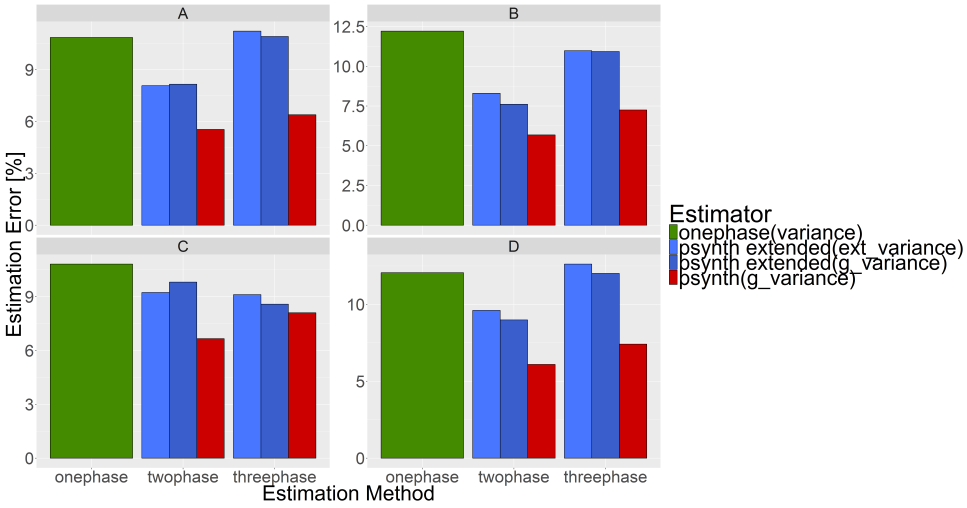
$$rel.eff = \frac{\hat{\mathbb{V}}_{onephase}(\hat{Y})}{\hat{\mathbb{V}}_{multiphase}(\hat{Y})} \tag{2.31}$$

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 EXTPSYNTH 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.

2.7.2 Visualization

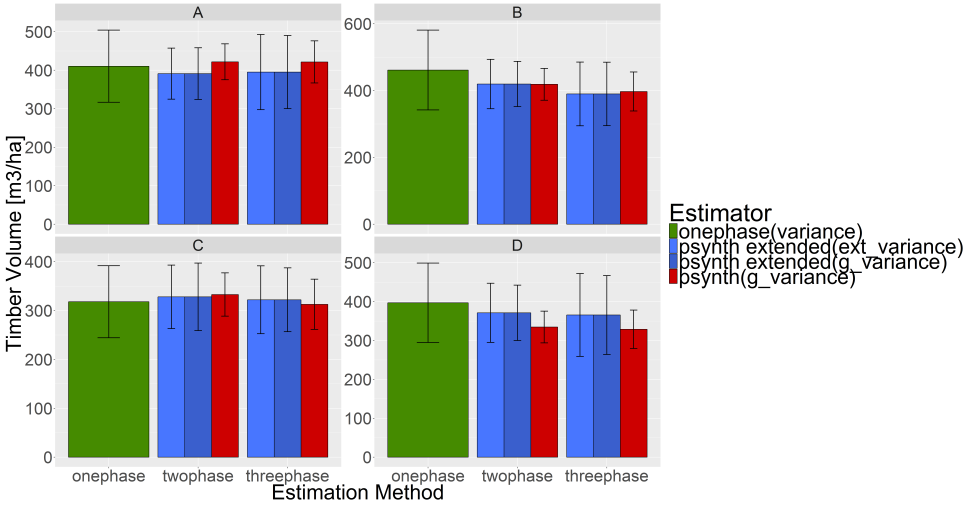
The **forestinventory** package also provides a `S3` generic plot method based on the **ggplot2** package (Wickham, 2009) to visualize the estimation results in two ways: 1) the point estimates with overlayed confidence intervals, and 2) the estimation errors. Both plots can be obtained by passing the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



2.8 Future plans

The **forestinventory** package currently provides a fairly well-rounded toolkit for forestry inventorists 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 & 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.

Chapter 3

Combining canopy height and tree species map information for large scale timber volume estimations under strong heterogeneity of auxiliary data and variable sample plot sizes

Andreas Hill[1], Henning Buddenbaum[2], Daniel Mandallaz[1]

[1]ETH Zürich
Department of Environmental Systems Science, Universitätstrasse 16, 8092 Zürich, Switzerland

[2]Trier University
Environmental Remote Sensing and Geoinformatics Department, Behringstrasse 21, 54286 Trier, Germany

Submitted to:
European Journal of Forest Research (in review).

– Henning Buddenbaum processed the airborne Laserscanning data and supported writing the manuscript.

– Daniel Mandallaz supported the statistical data analysis.

Abstract

A timber-volume regression model applicable to the state and communal forest area of the federal German state of Rhineland-Palatinate is identified using a combination of airborne laser scanning (ALS)-derived metrics and information from a satellite-based tree species classification map available on the federal state level. As is common in many forest inventory datasets, strong heterogeneity in the ALS data due to different acquisition dates and misclassifications in the tree species classification map had noticeable effects on the regression model's performance. This article specifically addresses techniques that improve the performance of ordinary least square regression models under such restricting conditions. We introduce a calibration technique to neutralize the effect of misclassifications in the tree species variable that originally caused a residual inflation of 0.05 in adjusted R^2. Incorporating the calibrated tree species information improved the model accuracy by up to 0.07 in adjusted R^2 and suggests the use of such information in forthcoming inventories. We also found that including ALS quality information as categorical variables within the regression model considerably mitigates issues with time lags between the ALS and terrestrial data acquisition and ALS quality variations (increase of 0.09 in adjusted R^2). The model achieved an adjusted R^2 of 0.48 and a cross-validated root mean square error ($RMSE_{cv}$) of 46.7% under incorporation of the tree species and ALS quality information, and was thus improved by 0.12 in adjusted R^2 (5% in $RMSE_{cv}$) compared to the simple model only containing ALS height metrics (adjusted R^2=0.36, $RMSE_{cv}$=51.7%).

3.1 Introduction

Forest inventory methods are the primary tools used to assess the current state and development of forests over time. They provide reliable evidence-based information that is used to define and identify management actions as well as to adapt forest management strategies to both national and international guidelines. Two methods that have become particularly attractive are so-called *double-sampling* (Mandallaz, 2008, Ch. 5) and *mapping* (Brosofske et al., 2014) procedures. The core concept of these methods is to use predictions of the terrestrial target variable at additional sample locations where the terrestrial information has not been gathered. These predictions are produced by models that use explanatory variables derived from *auxiliary data*, commonly in the form of spatially exhaustive remote sensing data in the inventory area. Especially models to predict timber volume based on airborne laser scanning (ALS) have been extensively investigated for a long time (NÃ¦sset, 1997). The specific scope of double-sampling is to enlarge the terrestrial sample size by a much larger sample of predictions of the target variable in order to gain higher estimation precision without performing additional expensive terrestrial measurements. Model-dependent and design-based regression estimators are used in a broad range of double sampling concepts and methods (Gregoire & Valentine, 2007; Köhl et al., 2006; Schreuder et al., 1993; Saborowski et al., 2010; Mandallaz, 2013a,d) and have been applied to existing inventory systems (Breidenbach & Astrup, 2012; von Lüpke & Saborowski, 2014; Mandallaz et al., 2013; Magnussen & Tomppo, 2014; Massey et al., 2014). While double-sampling methods provide reliable estimates for a given spatial unit, e.g. a forest district, they do not provide information about the spatial distribution of the estimated quantity within this area. For this reason, the same modeling technique used in double-sampling procedures has also been intensively used to produce exhaustive prediction maps that provide pixelwise estimations of a target variable in high spatial resolution (Bohlin et al., 2017; Latifi et al., 2010; Tonolli et al., 2011; Hill et al., 2014; Nink et al., 2015).

To allow for an area-wide application of the prediction model, both double sampling and mapping methods require that the remote sensing data are available over the entire inventory area. This is usually not a limiting factor in *small-scale* applications. In the optimal case, the remote

sensing data are in principle collected in accordance to the specific study objective. Quality standards that have often been addressed are that *a)* the remote sensing data should be acquired close to or even at the time of the terrestrial inventory in order to ensure best possible comparability between the target variable on the ground and the remote sensing derived variables (McRoberts et al., 2015); *b)* the remote sensing technology and its spectral and spatial resolution should be chosen according to the modelling purpose (Köhl et al., 2006); and *c)* the variation in quality of the remote sensing data over the inventory area should be minimized in order to avoid artificial noise in the data (NÃ¦sset, 2014). Despite the increasing availability and decreasing costs of remote sensing data (White et al., 2016), these quality standards of the remote sensing data can often not be guaranteed for *large-scale* applications (Maack et al., 2016), and trade-offs must be accepted (Jakubowski et al., 2013). The prime objective is then to produce the best possible prediction model given the restrictions imposed by the available remote sensing information. The exploration of scarcely used remote sensing products and the optimization of prediction models under severe quality restrictions in the remote sensing data are thus one of the challenges in large-scale model-supported inventory applications.

Among the still rarely used remote sensing data in large scale applications, the integration of tree species information in prediction models - especially for timber volume estimation - has been stated as some of the most promising but often missing information (Koch, 2010; White et al., 2016). As timber volume estimations on the single tree level in forest inventories are often based on species-specific biomass and volume equations (Husmann et al., 2017; Zianis et al., 2005), the application of species-specific models is expected to be a key factor for improving estimation precision (White et al., 2016). This has been supported by studies from Breidenbach et al. (2008) who achieved a substantial improvement in accuracy of their timber volume prediction model when including a variable estimating the deciduous proportion derived from leaf-off ALS data. Similar gains in model performance were also reported by Straub et al. (2009) and Latifi et al. (2012) who used broadleaf and coniferous information based on color infrared orthophotos as a categorical explanatory variable. However, studies that explore the use of more species-specific information (i.e. a further discrimination of tree species) as explanatory variables have been rare. Further investigations are thus necessary especially in countries whose forests are characterized by a larger variety of tree species that may also occur in mixed and uneven-aged stands (McRoberts et al., 2010). The area-wide tree species information in most studies was obtained from satellite and airborne remote sensing sensors based on automatic classification methods. Whereas the presence of misclassifications has already been addressed (Latifi et al., 2012), an issue that has so far been neglected is how misclassifications actually affect the prediction model (Gustafson, 2003).

A frequently encountered problem in large scale forest inventories is the lack of temporal synchronicity between the remote sensing acquisition and the terrestrial survey. As a result, the available remote sensing data often exhibit notable time-lags with respect to the date of the terrestrial inventory. This has often been addressed as a major drawback, especially for the application of design-based change estimation (Massey & Mandallaz, 2015b).

Our study is embedded in the current implementation of design-based regression estimators (Mandallaz, 2013a; Mandallaz et al., 2013; Mandallaz, 2013d) for estimating the standing timber volume within the state and communal forest management units over the entire state of Rhineland-Palatinate (RLP, Germany). With respect to this overall objective, the aim of this study was to derive an ordinary least square (OLS) regression model to generate predictions of the standing timber volume associated with a sample location of the Third German National Forest Inventory (BWI3) over the entire state and communal forest area (6155 km^2). A merged ALS dataset from different acquisition years and a satellite-based tree species classification map for the five main tree species in RLP was available for the entire inventory area and consequently used to derive predictor variables. The major limiting factors for using these data in a regression analysis are **(i)** variation in the ALS data quality as well as time-lags of up to 10 years between the ALS acquisitions and

the terrestrial survey, **(ii)** misclassifications in the tree species classification map and **(iii)** the ambiguous choice of a suitable extraction area (*support*) for all remote sensing information under angle count sampling in the terrestrial survey (variable sample plot sizes). For this reason, we address the following specific research questions:

1. How can tree species map information be optimally used within a regression model that predicts timber volume? What effects do misclassifications have on the predictions and how can these effects be minimized?

2. What are the effects of quality restrictions and substantial time lags between the ALS- and terrestrial data acquisition on the regression model and how can these effects be mitigated?

3. Does support size influence model accuracy? What is the optimal support size and what are the determining factors?

3.2 Materials and Methods

3.2.1 Study Area

The German federal state Rhineland-Palatinate (RLP) is located in the western part of Germany and borders Luxembourg, France and Belgium (Fig. 3.1). With 42.3% (appr. 8400 km^2) of the entire state area (19850 km^2) covered by forest, RLP is one of the two states with the highest forest coverage among all federal states of Germany (Thünen-Institut, 2014). The forest area of RLP is divided into three ownership classes, i.e. state forest (27%), communal forest (46%) and privately owned forest (27%). The most frequent tree species in RLP are European beech (*Fagus sylvatica*, 21.8%), oak (*Quercus petrea* and *Quercus robur*, 20.2%), Norway spruce (*Picea abies*, 19.5%), Scots pine (*Pinus sylvestris*, 9.9%), Douglas fir (*Pseudozuga menziesii*, 6.4%), European larch (*Larix decidua*, 2.4%) and Silver fir (*Abies alba*, 0.7%). The share of broadleaf tree species is 58.7%. The forests of RLP further exhibit heterogeneous structures (Thünen-Institut, 2014): around 82% of the forest area in RLP are mixed forest stands (i.e. at least two different tree species occur in the same stand) and 69% of the forest area exhibit a multi-layered vertical structure. While the average tree age is around 80 years, most of the forest area (20%) is occupied by trees between 40 and 60 years of age, whereas 27% of the trees are older than 100 years. Spatially variable climate conditions have a strong influence on the local growth dynamics as well as tree species composition and create a large variety of forest structures, ranging from characteristic oak coppices (Moselle valley), pure spruce, beech and Scots pine forests (e.g. Hunsrück and Palatinate forest) to mixed forests comprising variable proportions of oak, larch, spruce, Scots pine and beech. Accordingly, RLP has been divided into 16 bioclimatic growing regions that form homogeneous areas with respect to the afore mentioned characteristics (Gauer & Aldinger, 2005).

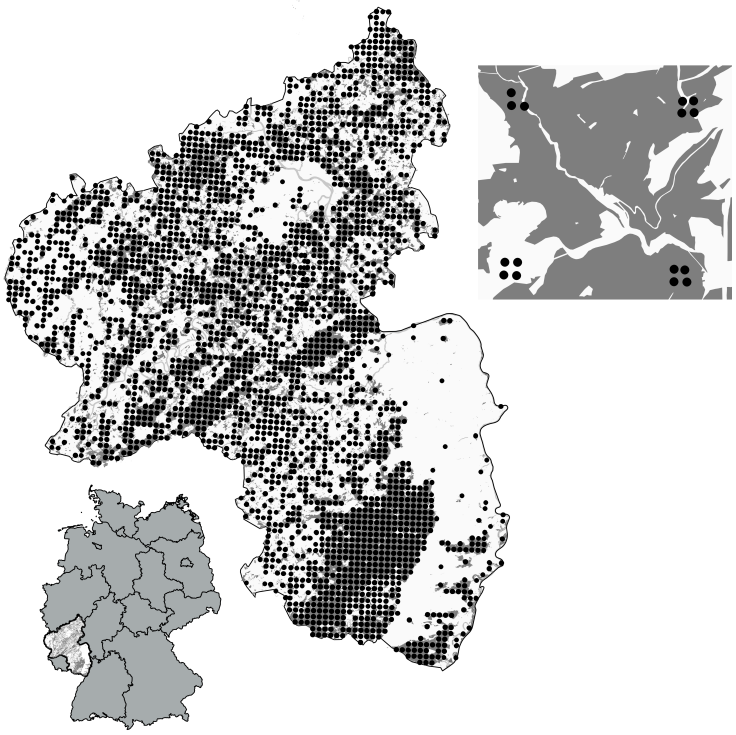


Figure 3.1: Spatial distribution of the BWI3 cluster samples over Rhineland-Palatinate

3.2.2 Terrestrial Inventory Data

The German National Forest Inventory (NFI) is carried out over the entire forest area of Germany in reoccurring time periods of 10 years. The most recent inventory (BWI3) has been conducted in the years 2011 and 2012. In this framework, Rhineland-Palatinate is covered by a 2x2 km grid that defines the sample locations for the terrestrial survey. A sample unit consists of four sample locations (also referred to as *sample plots*) that are arranged in squares (so called *clusters*) with a side length of 150 metres (Fig. 3.1). The number of plots per cluster can however vary between 1 and 4 depending on forest/non-forest decisions on the plot level (Bundesministerium für Ernährung, 2011). In the field survey of the BWI3, sample trees for timber volume estimations are selected according to the angle count sampling technique (Bitterlich, 1984), using a basal area factor (BAF) of 4 that is respectively adjusted for boundary effects at the forest border (Bundesministerium für Ernährung, 2011). A further selection criterion for a tree to be recorded is a diameter at breast height (dbh) of at least 7 cm. This sampling technique was applied to 8092 sample plots (2810 clusters) in RLP, resulting in the collection of 56561 sample trees for which the dbh, the tree diameter at 7 m ($D7$) and the tree species were recorded for all trees. Tree height measurements were conducted only for a subset of all sample trees and used to predict the height for the remaining sample. During the last inventory, all plot center positions were remeasured with differential global positioning system (DGPS) technique. Knowledge about the exact plot positions were considered crucial to provide optimal comparability between the terrestrial observations and the information derived from the auxiliary data. A detailed analysis by Lamprecht et al. (2017) indicated that horizontal DGPS errors do not exceed 8 meters for 80% of all plots in RLP. For 162 plots, the DGPS coordinates were replaced by their former target coordinates due to missing or implausible values. In order to derive a volume estimation for each sample tree, the BWI3 estimates a taper curve for each sample tree by calibrating the random effects term of linear mixed-effects taper models with

the set of diameters and corresponding height measurements taken from the respective sample tree (Kublin et al., 2013). The integration of the derived taper curves consequently lead to a volume prediction for each sample tree. Since the overall objective of the study was to subsequently use the identified regression model for design-based timber volume estimations of state and communal forest management units, we already restricted the sample plots used for modeling to the state and communal forest area (73% of the entire forest area of RLP). This provides the advantage that when the regression model is used as an *internal model* in design-based estimators, the model predictions hold the assumption on the residuals to be zero on average over the state and communal forest area by construction of OLS technique (Mandallaz, 2013a; Mandallaz et al., 2013; Mandallaz, 2013d). The dataset of this study hence comprised 5791 plots (2055 clusters). For this sample, the timber volume density per hectare on plot level, $Y(x)$, was calculated according to the formula of one-phase one-stage sampling (Mandallaz, 2008, Ch. 4.2). The timber volume density per hectare on plot level was used as the response variable in the regression analysis.

Table 3.1: Descriptive statistics of the forest observed on NFI sample plots located within communal and state forest area (n=5791).

Variable	Mean	SD	Maximum
Timber Volume (m^3/ha)	300.86	195.55	1375.31
Mean DBH (mm)	354.90	137.22	1123.20
Mean height (dm)	239.60	72.43	497.43
Mean stem density per hectare	101.00	114.01	1010.31

3.2.3 Auxiliary Data

ALS Canopy Height Model

Between 2003 and 2013, the topographic survey institution of RLP acquired airborne laser scanning (ALS) data over the entire state of RLP at leaf-off condition (Fig. 3.2). The objective of this campaign was to derive a countrywide digital terrain and surface model based on the acquired ALS point clouds. During the extended acquisition period, airborne laser scanning technology and data quality evolved significantly. The tiles recorded in 2002 and 2003 have a rather poor quality with about only 0.04 points per m^2 , while more recently acquired datasets contained about 5 points per m^2 . The data was delivered as two separate datasets comprising the Vegetation First Pulse (VEF) and Ground (GRD) points. All point clouds were stored as three-column (easting, northing, and height above sea level) ASCII files in tiles of 1 km^2. In order to create a surface model (DSM) in a given raster resolution, the highest point of the combined VEF and GRD dataset was identified in each raster cell and saved as a thinned surface point cloud. For the elevation model (DEM), the mean of all GRD points in the cell was calculated, and the result was saved as a thinned ground point cloud. The thinned point clouds were then aggregated to larger tiles and interpolated to raster images using a Delauney interpolation in the Matlab software (Mathworks, 2017). The resulting DSM and DEM raster sets were then subtracted from each other to calculate a canopy height model (CHM) in raster format, providing discrete information about the canopy surface height of the entire forest area of RLP in a spatial resolution of 5 meters. The thinning process led to much smaller datasets that could be processed in larger tiles and considerably lowered processing times compared to the original dense point clouds. Since the data was recorded in leaf-off condition, the original point clouds contained many returns from within the crowns of deciduous trees. The thinned dataset provided the advantage that those measurements did not skew the vegetation height estimate in the final CHM.

As explanatory variables, the mean canopy height (*meanheight*) and the standard deviation (*stddev*) were calculated as the mean and standard deviation of all raster values within a predefined circle (i.e. *support* of the explanatory variable, see Section 3.2.4) around each sample plot center. In order to correct for edge effects at the forest border, each support area was previously intersected with the state and communal forest area, which was defined by a polygon mask provided by the forest service (Fig. 3.3b). Restricting the support area and thus the evaluation of the auxiliary data to the forest area is a means to optimize the coherence between explanatory variables computed at the forest boundary and the corresponding terrestrial response variable (Mandallaz et al., 2013). The tree height is one prominent predictor variable in the taper functions of the BWI3 that are used to calculate a timber volume value for each sample tree (Kublin, 2003; Kublin et al., 2013). A visual inspection of the tree volumes of all sample trees collected in the BWI3 within RLP against their tree heights also revealed the characteristic shape of an allometric relationship between these variables (Online Resource 1). It was hypothesized that this relationship on single-tree level is also apparent on the aggregated level of a sample plot and cluster, and can be used within the frame of regression modeling.

The strength of correlation between *meanheight* and timber volume on plot level was expected to show high variation according to the mentioned time-lag up to 10 years between ALS acquisition and terrestrial survey. The quality of the height information was also expected to vary according to changing sensor technologies and different point densities used over the years. For these reasons, the ALS acquisition year (*ALSyear*) for each sample plot was considered as a potential categorical explanatory variable to explain the variation in the data introduced by these factors. For this purpose, the acquisition year *2008* was further divided into *2008* and *2008_1*. In the latter, the data quality turned out to be very poor due to sensor failures during the acquisition. Additionally, the years *2006* and *2007* as well as *2012* and *2013* were pooled in order to increase the number of observations per factor level for modelling reasons. As a result, the *ALSyear* variable comprised nine categories (*2002*, *2003*, *2007*, *2008*, *2008_1*, *2009*, *2010*, *2011* and *2012*).

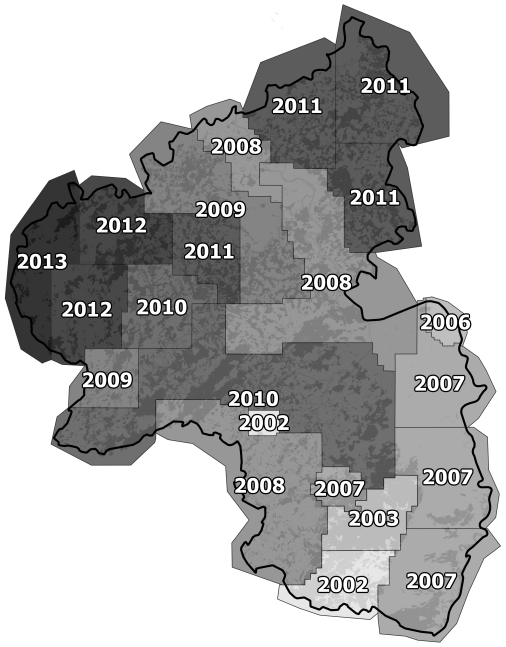


Figure 3.2: Separate ALS acquisitions in Rhineland-Palatinate over the years. The colors also indicate the quality of the data: *light*: low point densities ($0.04/m^2$), *dark*: high point densities ($>4/m^2$). Blue semitransparent layer: state and communal forest area.

Tree Species Classification Map

A countrywide satellite-based classification map of the five main tree species (European beech, Sessile and Pedunculate oak, Norway spruce, Douglas fir, Scots pine) described in Stoffels et al. (2015) was used to derive tree species information on sample plot level. The classified tree species map has a grid size of 5 meters and predicts five of the seven tree species that are used in the BWI3 taper functions (Kublin et al., 2013) to calculate the timber volume of a sample tree. Due to unavailable satellite data for the classification, the tree species map excluded one patch with an area of 415 km^2 in the south-west part of RLP, and two further patches with an area of 76 km^2 and 100 km^2 in the northern part (Stoffels et al., 2015). The tree species information was consequently missing for 411 (7%) of the 5791 sample locations.

Prediction of main plot tree species

A visual inspection of all BWI3 sample trees of RLP suggested that a stratification of the relation between tree height and timber volume according to these seven tree species may provide a considerable reduction in variation within the tree species groups (Online Resource 1). This led to the hypothesis that this tree species specific signal might also be apparent on sample plot and cluster level and can consequently be used to increase the accuracy of the prediction model. Based on the tree species classification map, the main tree species of each sample plot was calculated as an additional categorical explanatory variable (*treespecies*) with six categories following a similar approach as Latifi et al. (2012): one of the five tree species was assigned as the main plot tree species if its proportion within the edge-corrected support around the sample location exceeded a predefined threshold. If this threshold was either not exceeded by any of the five tree species or exceeded by several tree species sharing the same proportion, the respective sample plot was assigned the category 'Mixed'. We hypothesized that the choice of the threshold-value might have an influence on the resulting classification accuracy and the regression model accuracy (Section 3.2.5). We thus investigated the application of 5 threshold settings, i.e. 0%, 50%, 60%, 80% and 100%.

Calibration

Our analyses revealed that the prediction of the main tree species for a sample plot can be subject to misclassifications (Section 3.3.1). Errors in the explanatory variables of linear regression models can however lead to a bias of the regression coefficients in the direction of zero due to an artificial introduction of noise (Carroll et al., 2006, Ch. 3). This can cause an inflation of the residual variance and a consequent decrease of the model accuracy (Magnussen et al., 2010a). In case of classification, the impacts of misclassifications on the model properties are even harder to predict (Gustafson, 2003, Ch. 3). While errors in the explanatory variables do not affect the unbiasedness of the estimators in the design-based framework, a reduction or elimination of the classification errors could provide an improvement of the regression model accuracy and thereby potentially lead to smaller prediction and estimation errors. We therefore addressed the effect of misclassifications in the *treespecies* variable categories as well as means to correct these errors.

We transferred the concept of *regression calibration* as known from classical measurement error statistics (Carroll et al., 2006) to the problem of misclassifications in the *treespecies* variable. In regression calibration, one considers an error-prone explanatory variable W that can be measured in high quantity, whereas X constitutes the same but error-free variable whose determination is however very expensive. In order to yield a corrected or less error-prone version of W, one can define a calibration model $f_{calmod}(X, W)$ that predicts X as a function of W. After calibration on a training set, $f_{calmod}()$ can then be applied to any observed W and yields the corrected, less error-prone variable W_{calib}. Using W_{calib} instead of W in the regression model then asymptotically

provides an unbiased estimate of the regression coefficients and thus corrects for the attenuation to zero.

We transferred this concept by using a random forest algorithm (Breiman, 2001) as calibration model. We considered the main tree species of the sample trees at each plot location x as the error-free variable $treespecies_{terr}$, that would also yield the highest model accuracies when used as predictor variable. The objective of the calibration model was thus to provide an improved classification accuracy of each predicted main plot tree species category with respect to $treespecies_{terr}$. The calibration model was considered to correct for potential systematic misclassifications and thus minimize the effect of misclassifications on the regression model when substituting the uncalibrated with the calibrated *treespecies* variable. The random forest algorithm is a machine learning algorithm that grows a large number of decorrelated classification trees by considering only a subset of all provided predictor variables for each split. In the case of classification, new data are thus predicted by aggregating the predictions of all trees using a majority vote. We calibrated the random forest algorithm (f_{RF}) with a set of p predictor variables that comprised the initial prediction of the main plot tree species (*treespecies*), the mean canopy height (*meanheight*) and standard deviation (*stddev*) derived from the CHM, the proportion of coniferous trees estimated from the tree species classification map ($prop.conif$) and the bioclimatic growing region (wgb) at the sample location (Eq. 3.1). Using explanatory variables of the timber volume regression model in the calibration model provided the advantage of reduced data storage compared to computing alternative variables for calibration. The calibration model was implemented using the random forest algorithm (Liaw & Wiener, 2002) in the statistical software R (R Core Team, 2017). The algorithm was grown with 2000 trees, considering $\sqrt{p} \approx 3$ of the predictors for each split.

$$\begin{aligned} treespecies_{terr}(x) = f_{RF}(&treespecies, meanheight, \\ &stddev, prop.conif, wgb) \end{aligned} \tag{3.1}$$

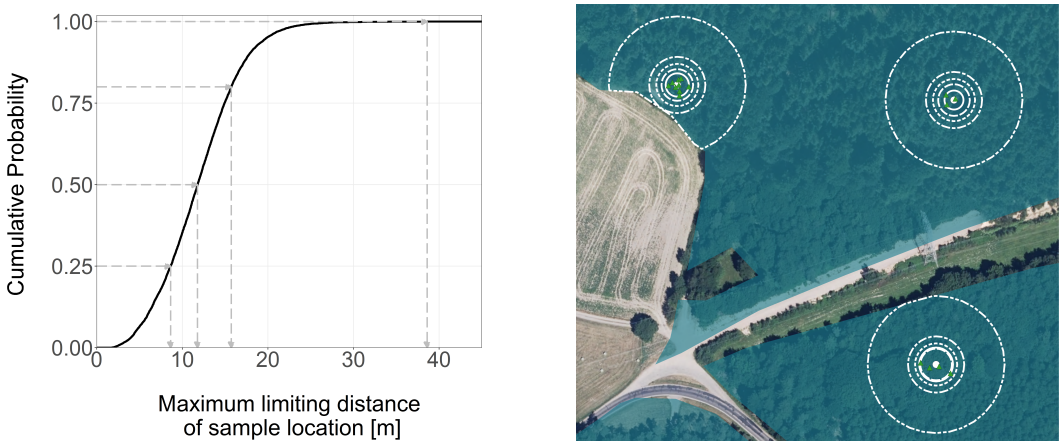
The calibration model was subsequently applied to the entire dataset. We then investigated the effect on the regression model performance (regression coefficients, model accuracy) when substituting the calibrated (less error-prone) for the uncalibrated (most error-prone) variable, and likewise for the actual (error-free) main plot tree species derived from the sampled trees of the respective sample plot under identical threshold settings.

3.2.4 Choice of Support under Angle Count Sampling

One characteristic of angle count sampling applied in the BWI3 is that a sample plot does not have a fixed radius in which trees are selected (*fixed-radius plot*), but each tree generates an individual radius from the plot center depending on its diameter at breast height (*variable-radius plot*). This tree-individual radius is known as the *limiting distance* from the plot center where the tree would still be included in the sample. A consequence of the absence of a fixed plot radius is the question about the optimal support (Hollaus et al., 2007), i.e. the spatial extent around the plot center in which the auxiliary information is evaluated and transformed into an explanatory variable. It has widely been hypothesized that the best relationship between the target variable on the ground and any explanatory variable derived from the auxiliary information is obtained if the support is spatially identical to the sample plot extent. In case of angle count sampling, an individual extent for each sample plot can be approximated by regarding the maximum limiting distances of its sample trees as the outer plot radius. However, many design-based applications under double-sampling do not allow for a between-plot change of the support for a specific explanatory variable (Mandallaz, 2013d,a).

For this reason, the task is to find a unique support for each auxiliary information that leads to

the best overall model accuracy. Deo et al. (2016) conducted extensive analysis to identify optimal supports for modelling standing timber volume for *variable-radius plot* designs in conifer forests. They analysed 24 different radii (i.e. circular supports) in which they extracted 57 metrics from a ALS derived point cloud with an average point density of 18 pulses per square meter. They successively evaluated the prediction performance of each support size by using the ALS metrics in a random forest algorithm and comparing the resulting model accuracies. In order to identify the best-performing supports for our explanatory variables, we followed a similar approach. The explanatory variables were calculated using *individual* (i.e. plot-varying) supports (*ind*), i.e. an individual support radius was used for each plot according to the maximum limiting distance of all sample trees associated to the respective sample plot. We then compared the model accuracies achieved by the individual supports against the model accuracies from a set of $fixed$ (i.e. non plot-varying) supports. The extents of the fixed supports were chosen from the cumulative distribution function (ECDF) of the maximum limiting distances of all 5791 sample plots of the analysed forest area (Fig. 3.3a). We considered the 25^{th} ($q25$, 9 meters), 50^{th} ($q50$, 12 meters), 80^{th} ($q80$, 15 meters) and the 100^{th} ($q100$, 38 meters) percentiles, resulting in support diameters of 18, 24, 30 and 76 meters (Fig. 3.3). While in this study we also used circular supports to extract the auxiliary information, also other support-shapes are possible (e.g. rectangles, hexagons). We also want to emphasize that the use of different support sizes for each explanatory variable is perfectly valid in the infinite population framework of design-based estimators (Mandallaz, 2013d,a).



(a) (b)

Figure 3.3: Identification (a) and visualization (b) of potential support radii used for calculating the predictor variables on plot level based on ECDF of maximum limiting distances of all BWI3 sample locations in RLP.

3.2.5 Model Building and Evaluation

In order to judge the quality of the *treespecies* variable, the user's accuracy for each classified species category and the overall accuracy of the classification scheme was calculated based on the confusion matrix (Congalton & Green, 2008). As reference data, we calculated the actual main plot tree species by applying the respective threshold to the sample trees of each sample plot. The classification accuracy was evaluated for all support sizes for both the calibrated and the uncalibrated *treespecies* variables. The measures of the regression model accuracy using both CHM- and *treespecies* variables were defined as the 10-fold cross-validated root mean square error (RMSE_{cv}, equation 3.2) and the adjusted coefficient of determination (adjusted R^2) of the multiple linear regression model defined in equation 3.3. Additionally, we considered the interaction terms

meanheight:*treespecies*, $meanheight^2$:*treespecies*, *meanheight*:*ALSyear*, *stddev*:*ALSyear* and *meanheight*:*stddev* and performed a variable selection based on the Akaike Information Criterion (AIC) (Akaike, 2011) in order to minimize the number of variables in the model. Due to a pronounced unbalanced design in the *treespecies*-*ALSyear* strata (Online Resource 2), no interaction between *treespecies* and *ALSyear* was possible. We evaluated the model for all support combinations, considering the use of individual support sizes for each auxiliary information, using both the calibrated and the uncalibrated *treespecies* variable. The calibration model (Section 4.5.4) for the *treespecies* variable was recalculated for each respective support-threshold setting. 206 sample plots included no sample trees and the timber volume density $Y(x)$ was thus set to *zero*. These *zero-plots* were removed from the modeling dataset since they acted as leverage points in cases where the ALS height metrics were recorded long before the terrestrial survey. Together with the missing tree species information (Section 4.5.4), the modeling dataset s was limited to n=5171 observations.

$$RMSE = \sqrt{\frac{\sum_{x \in s} \left(\hat{Y}(x) - Y(x)\right)^2}{n}} \tag{3.2a}$$

$$RMSE\% = \frac{RMSE}{\frac{1}{n}\sum_{x \in s} Y(x)} \tag{3.2b}$$

$$\begin{aligned} Y(x) = &\beta_0 + \beta_1 \cdot meanheight + \beta_2 \cdot meanheight^2 + \\ &\beta_3 \cdot stddev + \\ &\beta_4 \cdot ALSyear_1 + ... + \beta_{12} \cdot ALSyear_9 + \\ &\beta_{13} \cdot treespecies_1 + ... + \beta_{18} \cdot treespecies_6 + \varepsilon(x) \end{aligned} \tag{3.3}$$

3.3 Results

3.3.1 Classification Accuracies

Effect of Support Size and Threshold

Evaluating the uncalibrated tree species predictions revealed a dependency of the classification accuracies on both the applied threshold and the support size. Firstly, increasing the threshold led to a decrease in the user's accuracies for most of the tree species independent of the support choice (Fig. 3.4). The reason for this is that raising the threshold to higher values leads to a higher probability for the reference class than for the predicted class to be assigned as class 'Mixed'. This is due to the distinct difference in the spatial resolution between the reference and prediction data: the rather coarse spatial resolution of the tree species raster map causes the predicted class to remain classified as one of the five tree species much longer than the reference data, which consist of individual sample trees of a sample location. This effect is amplified by high thresholds. The probability of the predicted class to also be classified as 'Mixed' can however be raised by increasing the number of raster cells to be evaluated. For this reason, the user's accuracies improve when using larger support sizes, and this effect is most pronounced under high thresholds. This scale-threshold dependency of the user's accuracy particularly affects tree species that most commonly occur in mixed forest stands in Rhineland-Palatinate, i.e. *Scots pine*, *oak* and *beech*. The user's accuracies for tree species that are mostly prominent in pure forest stands (*spruce*, *Douglas fir*) logically turned out to be much more robust to changes in the thresholds and support sizes. Among the uncalibrated tree species predictions, *beech* and *spruce* produced the best predictions achieving UAs of up to 70% and 80%. Although the predictions for *Douglas fir* and *Scots pine* generally

performed less well than *beech* and *spruce*, similar UAs can be produced by adjusting the threshold and support choices. UAs for *oak* never performed better than 50%.

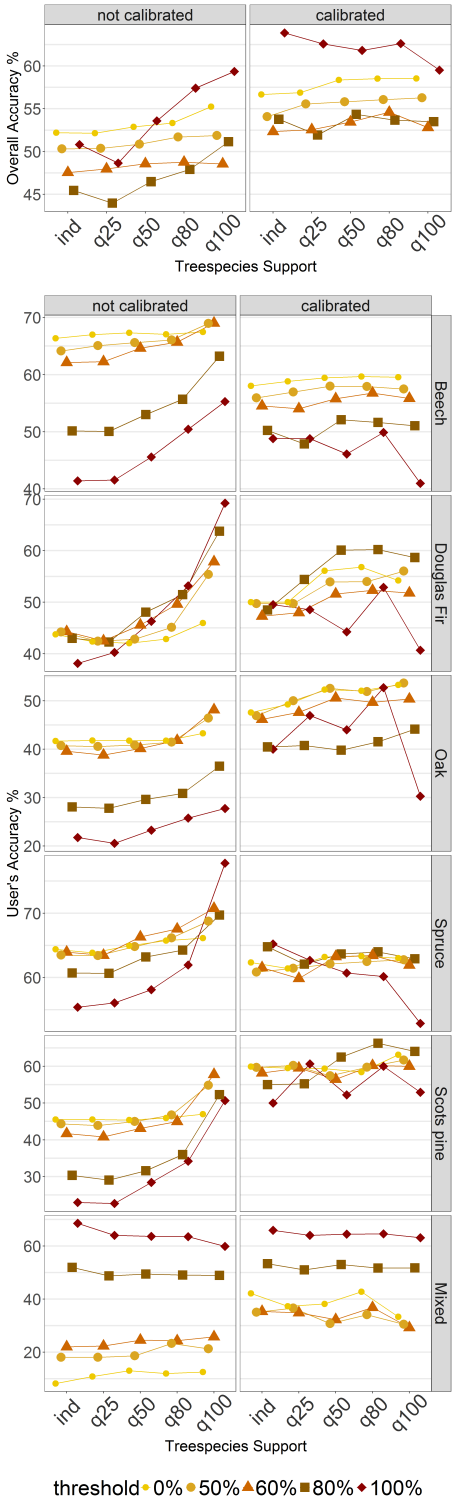


Figure 3.4: Classification accuracy for the main tree species of a sample location *before* and *after* calibration: *top)* overall accuracies. *bottom)* user's accuracies. *ind*: plot individual support sizes.

Calibration

Calibration substantially diminished the effect of the scale-threshold dependency for the five tree species and also increased the UAs for *Scots pine* and *oak*. Whereas the UAs for *beech* and *spruce* were found to be slightly lower after calibration, the overall accuracy under each support choice was always considerably increased by calibrating the tree species prediction (Fig. 3.4). With respect to the calculated random forest models, the initial tree species prediction (*treespecies*) and the information about the growing region (*wgb*) turned out to be the most valuable information, followed by the estimated proportion of coniferous trees (*prop.conif*) and the mean canopy height (*meanheight*).

3.3.2 Regression Model Accuracies

Effect of Support Size and Threshold

Figure 3.5 shows the accuracies of the regression model (equation 3.3) achieved under all possible combinations of support sizes for the auxiliary data. The stepwise selection procedure always included all considered single and interaction terms. In terms of adjusted R^2 and RMSE_{cv}, the analysis revealed that the choice of the CHM support size controls the overall level of the model's accuracy. The information about the main plot tree species can then be used to further improve the model fit under suitable *treespecies* support and threshold settings. When using the uncalibrated *treespecies* variable, an increase of the *treespecies* support size causes an increase in the model performance if low thresholds are used, whereas high thresholds (80%, 100%) cause a decrease in the model performance. This threshold-dependency could be removed by calibrating the *treespecies* variable. The highest adjusted R^2 and the lowest RMSE_{cv} were realized using the $q50$ support for both the CHM and calibrated *treespecies* variables in combination with a *treespecies* threshold of 100%, resulting in (adjusted R^2 of 0.48 and RMSE_{cv} of 136.62 m^2/ha (43.8%). However, various support and threshold combinations for the CHM and *treespecies* variables can be used to yield almost identical RMSE_{cv} and adjusted R^2 values. A detailed table of the model accuracies is given in Online Resource 4.

Effect of Misclassifications

We accessed the magnitude of the misclassification effect for all models that were analysed in Section 3.3.2, i.e. for all possible support and threshold combinations for the CHM and *treespecies* predictor variables. We first compared the adjusted R^2 of each model when using the uncalibrated *treespecies* variable against the adjusted R^2 using the actual, i.e. error-free variable. We then did the same comparison for the model using the calibrated treespecies predictor variable. Figure 3.6 provides a visualization of this comparison. Note that only the model with the predicted tree species variables can be applied to additional sample locations where no terrestrial survey has been carried out. As expected, the highest adjusted R^2 for every evaluated model was always achieved using the error-free tree species variable, whereas the missclassifications in the tree species variable led to a systematic decrease of the model accuracy. The calibration of the initially predicted main plot tree species using the random forest classification algorithm (Section 4.5.4) turned out to not only improve the classification accuracies (Section 3.3.1), but also to considerably decrease the effect of the missclassifications on the regression model predictions and accuracy. Figure 3.6 (*right*) shows that the adjusted R^2 under the actual and the calibrated predicted tree species variable are in general much closer to, and in many cases even on the identity line. The differentiation into two distinct point clouds results from the poor model performance under support size $q100$ for the CHM variables (i.e. the lower point cloud). Whereas the misclassifications in the uncalibrated *treespecies* variable led to a residual inflation of 0.01 - 0.05 in adjusted R^2, it was only between

0 and 0.01 after calibration. Further analysis revealed that when using the calibrated *treespecies* variable, the regression coefficients were almost identical to the ones received using the actual main plot tree species.

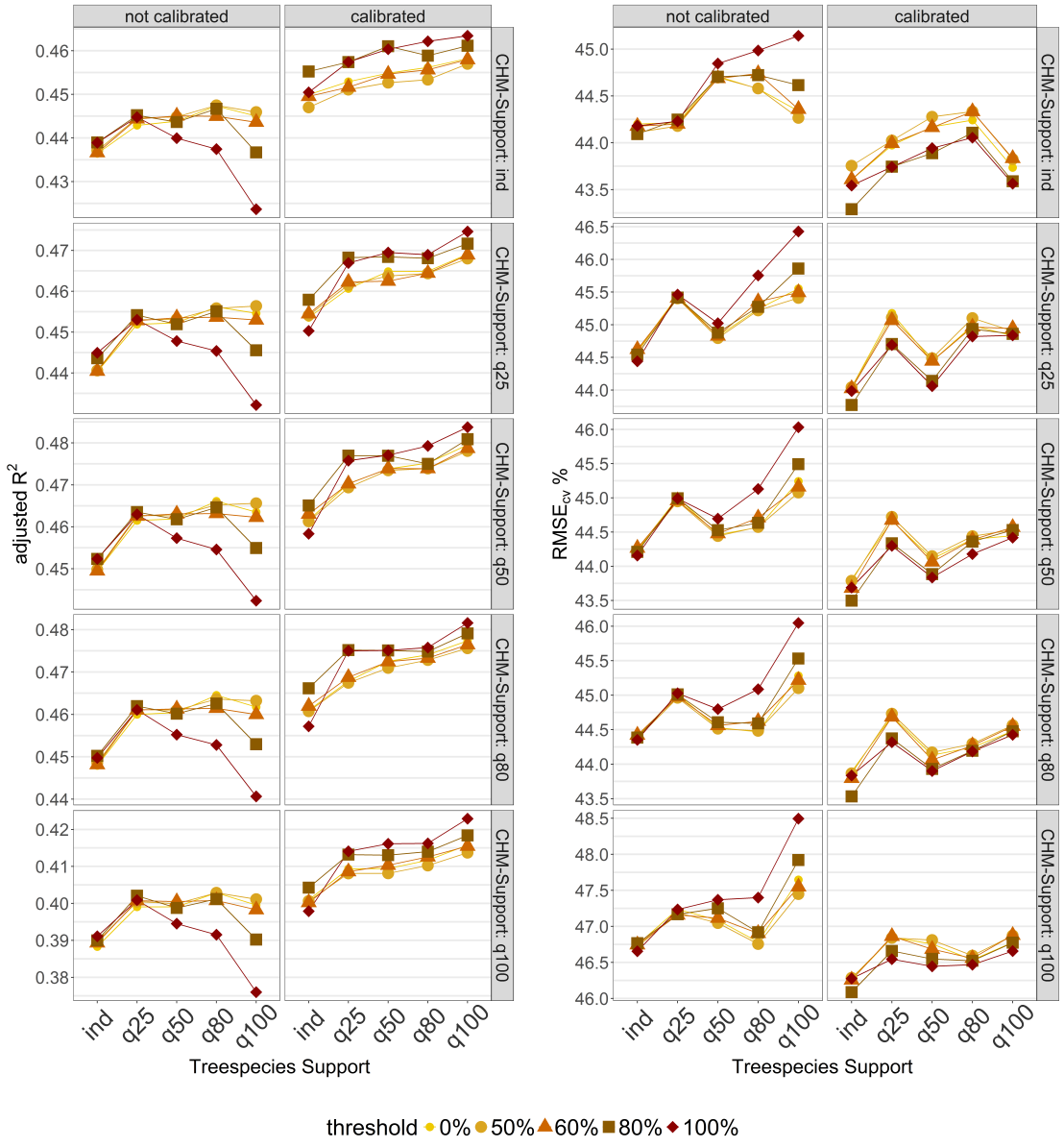


Figure 3.5: 10-fold RMSE_{cv} [%] and adjusted R^2 realized under various support choices for the CHM and *treespecies* explanatory variables

3.3.3 Final Regression Model

In order to address research questions 1 and 2 (i.e. the gain in model accuracy by tree species information and effect of heterogeneity in the ALS data), we investigated the model properties in more detail. For this purpose, we decided to use the best found model that was achieved under the support settings of $q50$ for both auxiliary data with a threshold of 100% for the tree species

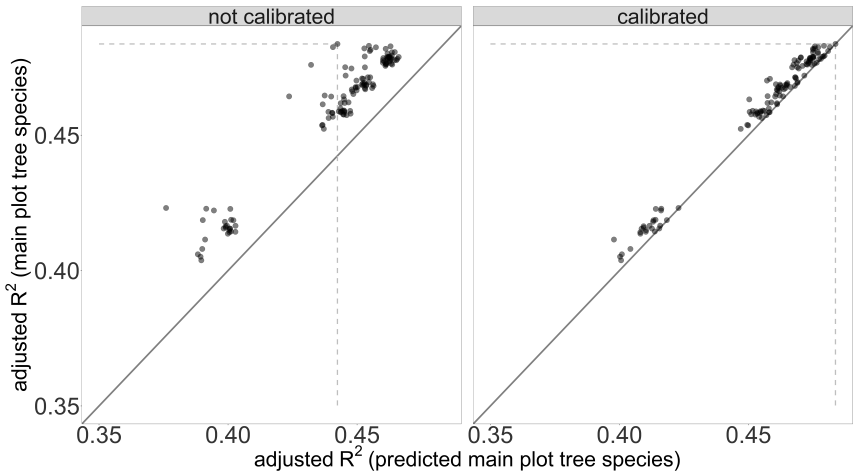


Figure 3.6: Effect on the adjusted R^2 when substituting the actual main tree species with the predicted main tree species of a sample plot. Each point in the graph represents the timber volume regression model under different supports and threshold settings. The *dotted* line tracks the the model with the highest adjusted R^2 under the use of the error-free *treespecies* variable. Semitransparent colours for the data points are used to visualize overlap.

variable as the regression model of choice. The reason for inspecting this model was that *a)* the model provided the highest adjusted R^2 among all validated models while reducing the data handling complexity for upcoming applications (i.e. identical support sizes for all remote sensing data) and *b)* the calibration neutralized the effects of misclassifications on the model predictions. The interaction term between $meanheight^2$ and *treespecies* (i.e. considering separate curvatures for each tree species) turned out not to have a significant influence on the model accuracy and was dropped, resulting in an adjusted R^2 of 0.48 and a slightly increased RMSE_{cv} of 140.62 m^2/ha (46.7%). The final model thus comprised 39 parameters (regression coefficients), i.e. the intercept, 3 main effects for continuous variables, 13 main effects for categorical variables and 22 interaction parameters (Table 4.4).

We also conducted an analysis for detecting influential data points or outliers for the final regression model. We here considered the commonly applied criteria of leverages and Cook's Distance as amongst others described in Fahrmeir et al. (2013, p. 160-167). The critical threshold of $2p/n$ (i.e. twice the average of the hat matrix' diagonal entries) was exceeded by 10% of the observations. However, only 3% of these leverage points were assigned to studentized residuals with values > 1 or < -1. Removing these observations from the dataset and refitting the model led to an adjusted R^2 of 0.49 compared to 0.48 when including them. Additionally, Cook's Distance values D_i did not exceed a value of 0.019, and were thus far apart from the commonly used critical threshold of $D_i > 0.5$ that indicate a considerably change of the regression model results when omitting them. We thus decided not to remove any observations from the modelling dataset. We thus decided not to remove any observations from the modelling dataset.

Interpretation of Final Regression Model

Figure 3.7 provides a visualisation of the timber volume predictions separated by the calibrated tree species and the ALS acquisition years. Sample plots classified as *oak* and *Scots pine* revealed to have an almost identical relationship (nearly identical slopes) for the mean canopy height - timber volume relationship. They only differ by a marginally higher intercept for *Scots pine* plots, meaning that given the same mean canopy height a sample plot dominated by *Scots pine* yields a marginally higher timber volume on the plot level than a plot dominated by *oak*. *Beech*-dominated

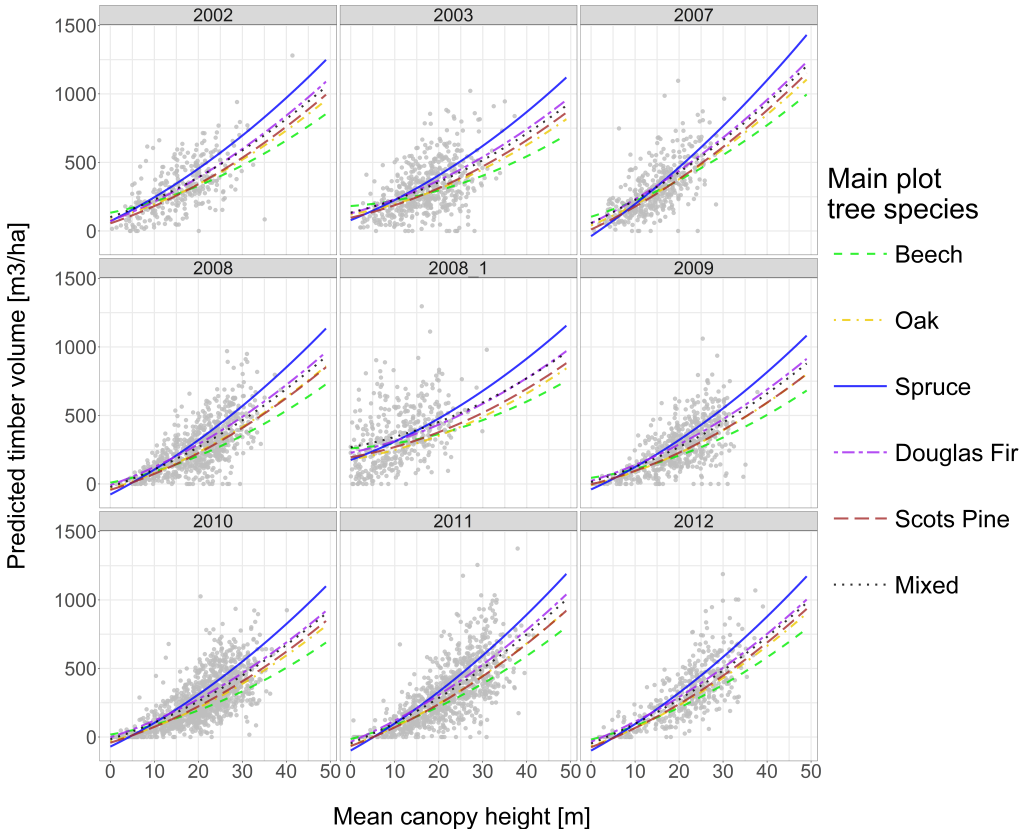


Figure 3.7: Visualization of the timber volume prediction function (*final regression model*) on sample plot level for each main plot tree species and ALS acquisition year. For visualization purposes, the predictor variable *stddev* was set to its average value within the respective *treespecies* and *ALSyear* categories. The terrestrially observed timber volume values are plotted in the background.

Table 3.2: Accuracy metrics for submodels of final OLS regression model. p gives the number of parameters for each model. Interaction terms are indicated by ':'.

model terms	model	p	R^2_{adj}	$RMSE_{cv}$	$RMSE_{cv}$%
meanheight + stddev + meanheight^2 + treespecies + ALSyear + meanheight:treespecies + meanheight:ALSyear + meanheight:stddev + stddev:ALSyear	final model	39	0.48	140.62	46.69
meanheight + stddev + meanheight^2 + meanheight:stddev	submodel 1	5	0.36	155.54	51.65
meanheight + stddev + meanheight^2 + ALSyear + meanheight:ALSyear + meanheight:stddev + stddev:ALSyear	submodel 2	29	0.45	145.62	48.35
meanheight + stddev + meanheight^2 + treespecies + meanheight:treespecies + meanheight:stddev	submodel 3	15	0.40	150.32	49.92

sample plots tend to achieve a higher timber volume than *oak* and *Scots pine* for canopy heights below 20 meters, but realize the lowest timber volumes for canopy heights above 20 metres. Sample plots dominated by any of the remaining coniferous tree species (*Douglas fir*, *spruce*) revealed to have higher slopes than broadleaf classified plots. This indicates that given the same mean canopy height, sample plots dominated by *Douglas fir* and *spruce* yield higher timber volume values than broadleaf- or *Scots pine* dominated sample plots, and this difference becomes more pronounced with increasing mean canopy heights. Within the group of coniferous-dominated sample plots, *spruce* turned out to have the highest slope, thereby yielding the highest timber volume values for mean canopy heights above 15 meters. An undesired characteristic of the model is that the predicted timber volume can in some cases ($< 1\%$) take negative values for low canopy heights (e.g. for *spruce*-dominated plots with *meanheight* below 5 meters and *stddev* of 4 meters). However, we chose not to use a log-transformation of the response variable. Doing so would have prevented the subsequent calculation of the g-weight variance of the design-based estimators (Mandallaz, 2013a; Mandallaz et al., 2013), which is only possible for response variables on the original scale. The g-weight variance provides the benefit of a better variance estimate for internal models by considering the dependency of the regression coefficients on the realized sample. The rare occurrence of negative predictions were however not considered to have an influence on subsequent design-based estimates when averaging multiple predictions within given spatial domains.

Effect of Time-Lags and Heterogeneity in ALS Data

Incorporating the ALS acquisition year as a categorical variable (*ALSyear*) in the regression model substantially accounted for the variability in the data introduced by *a)* the time-lags between ALS acquisition and terrestrial survey, and *b)* variation in ALS data quality which are due to sensor- and post processing techniques (Table 4.4). Whereas the adjusted R^2 for the regression model without considering the ALS acquisition year as additional predictor variable (*submodel 1*) was 0.36, it could already been increased to 0.40 by including the tree species variable (*submodel 2*). A further stratification by the ALS acquisition year increased the adjusted R^2 of *submodel 1* from 0.36 to 0.45, and the adjusted R^2 of *submodel 3* from 0.40 to 0.48.

We further analysed the model residuals within each ALS acquisition year (within-group variation) for the final model and nested submodels. It turned out that the R^2 values vary distinctly between the ALS acquisition year strata (Table 4.5). More precisely, the within-group R^2 can be higher and lower than the overall R^2 of the respective model. Figure 3.8 shows that a stratification according to the ALS acquisition years (submodel 2) can already increase the R^2 in most acquisition year strata, compared to the basic model using only the ALS height metrics as predictor variables (submodel 1). In the ALS acquisition year stratum 2007, the increase in R^2 even reached 0.08.

Table 3.3: R^2, RMSE and RMSE% of final regression model within ALS acquisition year strata (*ALSyear*). $Area_{ALSyear}$: Area covered by ALS acquisition given in km^2. n: number of validation data.

ALSyear	$Area_{ALSyear}$	R^2	RMSE	RMSE%	n
2012	2807	0.61	135.84	44.87	408
2011	4361	0.57	146.21	48.29	883
2010	4182	0.51	120.90	39.93	1171
2009	2100	0.42	133.42	44.07	559
2008	2968	0.48	130.38	43.06	701
2008_1	2116	0.33	175.43	57.94	394
2007	3498	0.46	136.47	45.08	418
2003	602	0.27	154.48	51.02	529
2002	775	0.44	141.55	46.75	314

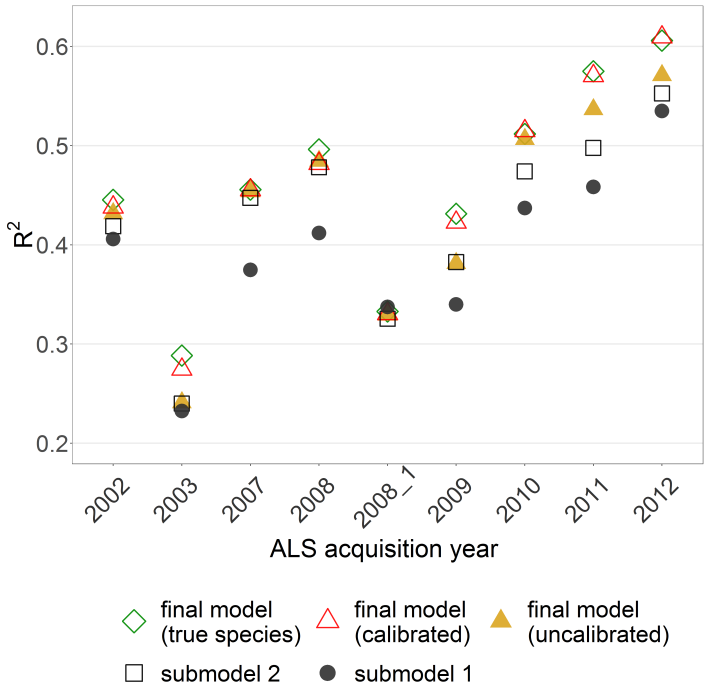


Figure 3.8: R^2-values of the final regression model, submodel 1 and submodel 2 achieved *within* the ALS acquisition year strata.

Added Value of Tree Species Map Information

Introducing the predicted main tree species of a sample plot as an additional categorical variable to submodel 2 yielded a further increase in the adjusted R^2 of 0.03 (Table 4.4). However, the improvement was even more pronounced in ALS acquisition years close or identical to the year of the terrestrial inventory (Fig. 3.8). We observed an increase of 0.06 in R^2 for ALS acquisition year 2012, and of 0.07 for ALS acquisition year 2011. The analysis illustrated once more that misclassifications in the tree species variable generally reduce model accuracy compared to using error-free tree species information. The residual inflations caused by the misclassifications in the uncalibrated *treespecies* variable within the *ALSyear* strata were up to 0.05 in R^2. However, the calibration was able to substantially decrease or even remove the effects of misclassifications on the model accuracy in all ALS acquisition year strata.

3.4 Discussion

3.4.1 Stratification according to ALS Acquisition Years and Tree Species

Incorporating the main tree species of a sample location in the timber volume regression model increased the model accuracy and revealed strong evidence for the existence of a tree species specific behavior concerning timber volume on the plot level. This result seems reasonable regarding the species specific taper functions on single-tree level applied in the BWI3 (Kublin, 2003; Kublin et al., 2013). These findings also agree with those of Latifi et al. (2012) who found an almost identical improvement in RMSE of 2% when stratifying to broadleaf and coniferous tree species. The overall RMSE of their model was however 10% smaller than in our study. The overall RMSE of

their model was however 10% smaller than in our study. This might be due to a more heterogeneous dataset of much smaller sample size in the cited study, but also because the temporal alignment between the auxiliary data acquisition and the terrestrial survey was much better than in our case. Additionally, the number of different tree species present in their dataset was lower than in our case and only comprised Scots pine, European beech and oak. The individual effects of spruce and Douglas fir indicated by our model also support the findings of Breidenbach et al. (2008), who found a higher percentage of coniferous trees in a sample plot to increase the timber volume predictions. This was not true for Scots pine and oak whose effects turned out to be very similar for our dataset. However, in our study the stratification according to the ALS acquisition years severely limited the flexibility of species-specific prediction functions and model interpretability. In particular, using the ALS acquisition years as categorical variables led to highly unbalanced datasets when stratifying according to the main plot tree species. This prevented the use of further stratification variables such as bioclimatic growing regions due to confounding effects and consequent singularities in the design matrices. Using the ALS acquisition years as categorical variables also implied an artificial increase in the number of parameters in the OLS regression model, which was however not regarded as critical with respect to overfitting issues due to the high amount of observations used for fitting the regression coefficients (Draper & Smith, 2014, Ch. 15.1). A stratification to the ALS acquisition years however proved to be an effective means in accounting for the artificially introduced noise in the data caused by quality variations and the large time-lags between the remote sensing and terrestrial data. It allowed for a model accuracy that was very close to those reported by Maack et al. (2016) who conducted a very similar study in the German federal state of Baden-Württemberg. Model accuracies were also particularly higher in ALS acquisition year strata in which the data showed considerably less noise or were closer to the date of the terrestrial survey. This effect was significantly reduced or even removed when merging several ALS acquisition year strata. Promising steps with respect to more up-to-date canopy height information have already been made, as the topographic survey institution of RLP is currently processing a canopy height model from aerial imagery acquisitions for 2011 and 2012 covering the entire federal state. These aerial photography acquisitions will in the future be conducted in a two-year period, allowing to derive up-to-date canopy height information in the framework of future forest inventories. For a smaller study area, Kirchhoefer et al. (2017) have already demonstrated that similar model accuracies for German NFI data can be achieved using imagery-based canopy height models.

Incorporating the calibrated tree species information further improved the model accuracy by 0.03 in adjusted R^2. Compared to the simple model only containing ALS height metrics, including the ALS quality and calibrated tree species information increased the adjusted R^2 by 0.12 in total. A differentiated evaluation of the final regression model revealed that the highest R^2-values were achieved within ALS acquisitions year strata close or identical with the year of the terrestrial survey, showing differences of up to 0.3 between the R^2s. Also the gain in R^2 by including the tree species information was largest (i.e. 0.07) in combination with ALS information acquired in the year of the terrestrial inventory. These insights were particularly interesting with respect to the further use of the regression model for small area estimations. Small area estimators generally gain modeling strength by defining the prediction model *globally* (i.e. using all data in the inventory area), and then applying the so-derived prediction model to a subset of observations located within the area of interest (Mandallaz, 2013a). Consequently, the proposed stratification technique in the prediction model is expected to yield a gain in model accuracy and a reduction of the small area estimation errors if the small area domain mostly includes data from strata that have high within-strata model accuracies. Findings of Breidenbach et al. (2008) indicated that a further increase in model accuracies could possibly be achieved when incorporating these categorical variables as random rather than fixed-effects in linear mixed-effects models (Pinheiro & Bates, 2000). The reason we did not apply this family of models was that small area regression estimators subsequently applied

in RLP (Mandallaz, 2013a; Mandallaz et al., 2013) require the internal models to be fitted by OLS technique.

3.4.2 Calibration of Tree Species Map Information

The accuracy assessment of the initially derived main plot species from the classification map revealed the presence of misclassifications that led to a decrease in model accuracy. This is in agreement with the potential effects of erroneous explanatory variables discussed in Carroll et al. (2006) and Gustafson (2003), i.e. an increase of variability (noise) in the data that can increase the amount of unexplainable variance and thereby reduce the model accuracy. One reason for the misclassifications were that the classification algorithm of Stoffels et al. (2015) was exclusively trained in pure stands with the objective to predict the *dominant tree species* of a forest stand. Thus, our requirements on the classification map differed considerably from the ones imposed by Stoffels et al. (2015) and have to be considered as far more difficult to meet. Firstly, the reference data used in the accuracy assessment also included understory trees that were recorded in the BWI3 sample. Secondly, determining an exact spatial validation unit for a sample location (support) is not possible due to the properties of angle count sampling (Section 3.2.4). Thirdly, distinct discrepancies in the spatial scale between the reference data and the classification map severely hamper exact predictions of the main plot tree species especially in mixed forest stands. The latter issue caused a pronounced dependency of the user's accuracy on the support and threshold choice, particularly for tree species that most commonly occur in mixed forest structures, i.e. *Scots pine* (91%), *oak* (90%) and *beech* (85%) (Thünen-Institut, 2014). With respect to this set-up, the application of our calibration method proved to be of high value. It led to an increase in the classification accuracies, particularly for those tree species that performed worse in the uncalibrated setup, and thereby successfully minimized and even removed the deleterious effect of misclassifications on model accuracy and regression coefficients. Whereas the extensive analysis in our study deepened the understanding of the afore mentioned scale-effects, an alternative method for future applications could be to use map-derived percentages of each tree species as predictor variables in the random forest algorithm in order to directly predict the terrestrially observed main plot tree species.

3.4.3 Choice of Support under Angle Count Sampling

The validation of different support sizes underlined that the support choice can impact the accuracy of a prediction model, and thus confirmed the findings of Deo et al. (2016). In the present study, differences in the model accuracies however turned out to be small for most support choices. An exception was the choice of the $q100$ support for the CHM derived variables (38 meter radius), where the model accuracy was considerably worse than under the optimal settings. Contrary to our hypothesis, the use of plot-individual supports did not yield the best prediction performance overall. Kirchhoefer et al. (2017) recently came to the same result when they transferred the angle-count sampling technique to a pixel-wise selection method of the auxiliary data that resembles the sample tree selection even more precisely. In their study, the application of fixed support sizes did also not perform worse than under variable supports. We consider two plausible reasons for the joint findings: first, the determination of an exact spatial extent that can be transferred to auxiliary data extraction remains technically infeasible under angle count sampling. Thus, angle count sampling does not seem to be adequate when linking inventory information with remote sensing data. Secondly, inaccuracies in the DGPS-measurements of the plot center locations as reported by Lamprecht et al. (2017) may have an increased impact on the model accuracy the more exact the auxiliary data derivation spatially corresponds to those of the field survey. However, the extensive analysis carried out in our study also indicated that the optimal support size does not

only depend on the spatial extent of the field plots, but also on the spatial resolution of the remote sensing data as well as the context in which the derived information is used in the prediction model. In the case of transforming the tree species information map into a suitable categorical predictor variable, the use of a large support size of 38 meter radius turned out to yield the best model accuracy. However, only few sample locations in the study area were actually characterized by limiting circles of that particular size. An analysis to find the best support settings therefore seems to be advisable prior to further applications of design-based or model-dependent inventory methods so as not to lose model accuracy by unsuitable support choices. The concept of the demonstrated analysis method for identifying suitable supports can be transferred to any kind of auxiliary information, predictor variable and prediction model.

3.5 Conclusion

We draw three major conclusions from our study: (1) our analyses strongly indicated that the acquisition of auxiliary data close to the date of the terrestrial survey is a key factor to achieve good model accuracies. Particularly for large-scale inventory applications, this requirement is often difficult to meet. In such cases, we consider that the proposed method of including quality information about the auxiliary data in a prediction model can be an effective technique for improving the prediction accuracy. Ongoing studies investigate whether this modelling technique can also lead to smaller estimation errors of design-based estimators. (2) Our study also indicated that the relationship between the field measured timber volume and remote-sensing derived height information is tree species specific. We expect that using the tree species information in a timber volume model would even lead to higher prediction accuracies when combined with explanatory variables that can further explain the variation within each tree species group, such as bioclimatic growing conditions, soil properties and stand density on the plot level. (3) We consider the demonstrated calibration technique to be a valuable method for future studies where an external tree species map (i.e. the map was not created for the specific study objective) is used in prediction models. The application of a calibration model can also be transferred to any error-prone explanatory variable and be a simple means to clean the data set from noise and thus increase the model accuracy.

Acknowledgements

We want to express our gratitude to Prof. H. Heinimann (Chair of Land Use Engineering, ETH Zurich) for supporting this study. We want to explicitly thank Dr. Johannes Stoffels from the Environmental Sensing and Geoinformatics Group of Trier University for providing the tree species classification map as well as for constructive discussions when it came to interpreting the results. Special gratitude is owed to the State Forest Service of Rhineland-Palatinate, in particular Dr. Joachim Langshausen, Jürgen Dietz and Claus-Andreas Lessander, for collaboration and providing the forest inventory and geodata. We also want to thank Kai Husmann and Christoph Fischer from the Northwest German Forest Research Institution Göttingen for their advice in processing the terrestrial inventory data.

Chapter 4

A double-sampling extension of the German National Forest Inventory for design-based small area estimation on forest district levels

Andreas Hill[1], Daniel Mandallaz[1], Joachim Langshausen[2]

[1]ETH Zürich
Department of Environmental Systems Science, Universitätstrasse 16, 8092 Zürich, Switzerland
[2] State Forest Service Rhineland-Palatinate
Office for Forest Planning, Rhein-Mosel-Strasse 7-9, 56281 Emmelshausen, Germany

Submitted to:

– Daniel Mandallaz developed the design-based estimators. He also supervised the statistical analysis and writing of the manuscript.

– Joachim Langshausen supervised the study on the part of the State Forest Service Rhineland-Palatinate and supported writing of the manuscript.

Abstract

The German National Forest Inventory consists of a systematic grid of permanent sample plots and provides a reliable evidence-based assessment of the state and the development of Germany's forests on national and federal state level in a 10 year interval. However, the data have yet been scarcely used for estimation on smaller management levels such as forest districts due to insufficient sample sizes within the area of interests and the implied large estimation errors. In this study, we present a double-sampling extension to the existing German National Forest Inventory (NFI) that allows for the application of recently developed design-based small area regression estimators. We illustrate the implementation of the estimation procedure and evaluate its potential by the example of timber volume estimation on two small scale management levels (45 and 405 forest district units respectively) in the federal German state of Rhineland-Palatinate. An airborne laserscanning (ALS) derived canopy height model and a tree species classification map based on satellite data were used as auxiliary data in an ordinary least square regression model to produce the timber volume predictions. The results support that the suggested double-sampling procedure can substantially increase estimation precision on both management levels: the two-phase estimators were able to reduce the variance of the SRS estimator by 43% and 25% on average for the two management levels respectively.

4.1 Introduction

The German National Forest Inventory (NFI) provides reliable evidence-based and accurate information of the current state and the development of Germany's forest over time. The NFI thereby has the responsibility to satisfy various information needs including reporting to public and state forestry administrations, wood-based industries and the public on the national level, as well as to the Food and Agriculture Organization of the United Nations (FAO) and to the United Nations Framework Convention on Climate Change (UNFCC) on the international level (Polley et al., 2010). The current design of the German NFI rests solely upon a terrestrial cluster inventory that is carried out at sample locations systematically distributed over the entire forest state area of Germany. In order to cover a large area of 114'191 ha (Thünen-Institut, 2014), the sample size has been specifically chosen to satisfy high estimation accuracies for forest attributes on the national and federal state levels. However, sample sizes often drop dramatically when entering spatial units below the federal state level. This is particularly true for forest management levels such as forest districts for which the estimation uncertainties turn out to be unacceptably large due to the very limited number of sample plots within these units. For this reason, the German NFI data have not yet been extensively incorporated into operational planning on forest district management levels. In most German federal states, management strategies are thus still based on expert judgments from time-consuming standwise inventories (SFI), which are prone to systematic deviations Kuliešis et al. (2016) and do not provide any measure of uncertainty.

Some German federal states, such as Lower Saxony, have approached this problem by establishing a regional Forest District Inventory (FDI) with a much higher sampling density than used by the NFI in order to scientifically base their regional management strategies on quantitative and accurate information (Böckmann et al., 1998). However, such FDIs are cost-intensive and, facing increasing restrictions in budget and staff resources, there has been a need for more cost-efficient inventory methods (von Lüpke, 2013). One method which has proven to be efficient is double- or two-phase sampling (Särndal et al., 2003; Gregoire & Valentine, 2007; Köhl et al., 2006; Mandallaz, 2008). Double-sampling incorporates less expensive auxiliary information and can be used to either increase estimation precision under a fixed terrestrial sample size, or maintain estimation precision under reduced terrestrial sample size. Double-sampling procedures have already been used for

stratification in the FDI of Lower Saxony (Saborowski et al., 2010), and Grafström et al. (2017) illustrated how to use the auxiliary information to determine optimised balanced terrestrial sample designs. Recent studies have extended double-sampling to triple-sampling estimation methods using auxiliary information derived at two different sampling intensities. An example can be found in von Lüpke et al. (2012) who illustrated an extension of the existing two-phase FDI of Lower Saxony to a three-phase design that uses updates of past inventory data as additional auxiliary information and allows for a significant reduction of the terrestrial sample size in intermediate inventories. Another example is Massey et al. (2014) who developed a triple-sampling extension based on the ideas of Mandallaz (2013d) for the Swiss NFI that can significantly reduce the increase in estimation uncertainty caused by the new annual inventory design.

Two-phase and three-phase samplings techniques have also been applied to small area estimation (SAE). SAE techniques address the situation where the number of samples within a subunit, or small area (SA), of the entire sampling frame is too small to provide reliable estimates for that unit. A broad range of SA estimators used in forest inventories (Köhl et al., 2006) originally comes from official statistics. One such method that is commonly applied is known as indirect estimation (Rao, 2015), where statistical models are used to convert auxiliary information into predictions of the target variable that is rarely or not observed in the small area. These models are trained using data from outside the small area in order to "borrow strength" from areas where information is available. Of numerous applications of SAE in forestry (Breidenbach & Astrup, 2012; Goerndt et al., 2011; Steinmann et al., 2013; Mandallaz et al., 2013), most use unit-level models, i.e. the inventory plot is the unit of the response variable in the training data used for the model fit. Such unit-level models have been intensively investigated for timber volume estimation using various remote sensing auxiliary data (Koch, 2010; NÃ¦sset, 2014). Other studies have investigated area-level models, where the auxiliary information is only provided on the SA level (Magnussen et al., 2017). Some studies have illustrated that even NFI data derived under low sampling densities can still be used to provide acceptable precision of small area estimates on much smaller management levels. One example is Breidenbach & Astrup (2012) who used data from the Norwegian NFI to make small area estimation for standing timber volume for 14 municipalities where the number of NFI samples within these areas were between 1 and 35. The estimation errors under the applied model-dependent and design-based small area estimators turned to be markedly smaller than under the standard one-phase estimator. Another example is Magnussen & Tomppo (2014) who recently used the Swiss NFI data to estimate timber volume within 108 Swiss forest districts with sample sizes between 9 and 206. Similar studies using German NFI data for small area estimation have been lacking.

The aim of this study was to investigate whether the German NFI data can provide acceptable estimation precision on two forest district levels using the latest small area estimation procedures. We therefore conducted a study in the German federal state Rhineland-Palatinate where we extended the German NFI to a double-sampling design and applied three types of design-based small area regression estimators in order to derive point and variance estimates of mean standing timber volume for 45 and 405 forest districts respectively. The SA-estimators we considered were the *pseudo-small*, *extended pseudo-synthetic* and the *pseudo-synthetic* design-based small area estimator suggested by Mandallaz (2013a) and Mandallaz et al. (2013). Auxiliary data consisted of a canopy height model (CHM) obtained from a countrywide airborne laser scanning (ALS) and a tree species classification map to be used for regression within tree species strata. The estimation precisions were compared to those obtained by the standard one-phase estimator for cluster sampling under simple random sampling. The chosen double-sampling estimators were selected for several reasons: **(i)** the design-based framework relaxes dependencies on the regression model assumptions which seemed appropriate facing severe quality restrictions in the ALS data; **(ii)** the estimators can be used with *non-exhaustive*, i.e. non wall-to-wall, auxiliary information; **(iii)** all estimators are explicitly formulated for cluster sampling which has not yet been the case for fre-

quently used model-dependent estimators; and **(iv)** the asymptotically unbiased g-weight variance accounts for estimating the regression coefficients on the same sample used for estimation (*internal model approach*) and is also robust under heteroscedasticity of the model residuals. The results from this study were considered to provide valuable information whether the suggested procedure might be a cost-saving alternative to a regional FDI.

4.2 Terrestrial sampling design of the German NFI

The German NFI is a periodic inventory that is carried out every 10 years over the entire forest area of Germany. The most recent inventory (BWI3) was conducted in 2011 and 2012. While information was originally gathered on a systematic 4x4 km grid, some federal states such as Rhineland-Palatinate have switched to a densified 2x2 km grid. The German NFI uses a cluster sampling design, which means that a sample unit consists of at most four sample locations (also referred to as *sample plots*) that are arranged in a square, called *cluster*, with a side length of 150 metres. The number of plots per cluster can vary between 1 and 4 depending on forest/non-forest decisions by the field crews on the individual plot level (Bundesministerium für Ernährung, 2011). In the field survey of the BWI3, sample trees for timber volume estimation are selected according to the angle count sampling technique (Bitterlich, 1984), using a basal area factor (BAF) of 4 that is respectively adjusted for sample trees at the forest boundary by a geometric intersection of the boundary transect with the individual tree's inclusion circle (Bundesministerium für Ernährung, 2011). A further inventory threshold for a tree to be recorded is a diameter at breast height (DBH) of at least 7 cm. For each sample tree that is selected by this procedure, the DBH, the absolute tree height, the tree diameter at 7 m ($D7$) and the tree species is measured and used to estimate the volume at the tree level. These volume estimates are based on the application of tree species specific taper curves that are adjusted to the set of diameters and corresponding height measurements taken from the respective sample tree (Kublin et al., 2013).

4.3 Double sampling in the infinite population approach

The estimators used in this study have been proposed by (Mandallaz, 2013a; Mandallaz et al., 2013) and derive their mathematical properties under the so-called infinite population approach. Therefore, we shall first provide a short introduction into this general estimation framework. We start by assuming that the population P of trees $i \in 1, 2, ..., N$ within a forest of interest F is exactly defined, and each tree i has a response variable Y_i (e.g. its timber volume) that can be used to define the population mean Y (e.g., the average timber volume per unit area) over F. Since a full census of all tree population individuals is almost never feasible, Y has to be estimated based on a sample. In the infinite population approach this sample is a set of points or locations x distributed independently and uniformly over the set of all possible points in F. Each point x has an associated local density $Y(x)$ (e.g., the timber volume per unit area) whose spatial distribution is given by a fixed (i.e. non stochastic) piecewise constant function. The population mean Y is mathematically equivalent to the integral of the local density function surface divided by the surface area of F, $\lambda(F)$, i.e. $Y = \frac{1}{N}\sum_{i=1}^{N} Y_i = \frac{1}{\lambda(F)}\int_F Y(x)dx$, and thus the population mean Y corresponds to a spatial mean. Since the actual local density function is unobserved in its entirety, one estimates Y by taking a sample s_2 consisting of n_2 points and measuring each of their respective local densities. This sampling procedure is often referred to as *one-phase sampling* (OPS) and s_2 is referred to as the terrestrial inventory. In contrast to the one-phase approach, *two-phase* or *double-sampling* procedures use information from two nested samples (phases). Practically speaking, the terrestrial inventory s_2 is embedded in a large phase s_1 comprising n_1 sample locations that each provide a set of explanatory variables described by the column vector $\boldsymbol{Z}(x) = (z(x)_1, z(x)_2, ..., z(x)_p)^\top$ at each

point $x \in s_1$. These explanatory variables are derived from auxiliary information that is available in high quantity within the forest F. For every $x \in s_1$, $\boldsymbol{Z}(x)$ is transformed into a prediction $\hat{Y}(x)$ of $Y(x)$ using the choice of some prediction model. The basic idea of this method 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 this context, it is also important to note that the response and auxiliary variables are assumed to be error-free and the resulting errors for the point estimates reflect only the uncertainty due to sampling.

4.4 Estimators

4.4.1 Design-based one-phase estimator for cluster sampling (SRS)

The one-phase estimator for cluster sampling (SRS) constitutes the *status quo* that is currently applied under the existing one-phase sampling design of the German NFI in order to obtain point and variance estimates for the mean timber volume of a given estimation unit. In order to provide all estimators in the infinite population framework and ensure a consistent terminology with the two-phase estimators in Section 4.4.2, we will introduce the SRS estimator that is applied in the BWI3 algorithms (Schmitz et al., 2008) in the form given in Mandallaz (2008); Mandallaz et al. (2016).

In order to calculate the local density $Y_c(x)$ at the cluster level, a cluster is defined as consisting of M sample locations (in the BWI3, we have $M = 4$) where $M-1$ sample locations $x_2, ..., x_M$ are created close to the cluster origin x_1 by adding a fixed set of spatial vectors $e_2, ..., e_M$ to x_1. The actual number of plots per cluster, $M(x)$, is a random variable due to the uniform distribution of x_l $(l = 1, ..., M)$ in the forest F and to the forest/non-forest decision for each sample location x_l:

$$M(x) = \sum_{l=1}^{M} I_F(x_l) \quad \text{where} \quad I_F(x_l) = \begin{cases} 1 \text{ if } x_l \in F \\ 0 \text{ if } x_l \notin F \end{cases} \tag{4.1}$$

The local density on cluster level $Y_c(x)$, which is in our case the timber volume per hectare, is then defined as the average of the individual sample plot densities $Y(x_l)$:

$$Y_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l) Y(x_l)}{M(x)} \tag{4.2}$$

The local density $Y(x_l)$ on individual sample plot level was calculated according to the description in Mandallaz (2008), which can be rewritten for angle-count sampling technique applied in the BWI3. The general form of $Y(x)$ in Mandallaz (2008) is given as the Horwitz-Thompson estimator

$$Y(x_l) = \sum_{i \in s_2(x_l)} \frac{Y_i}{\pi_i \lambda(F)} \tag{4.3}$$

where Y_i is in our case the timber volume of the tree i recorded at sample location x in m^3. Each tree has an inclusion probability π_i that is well defined as the proportion of its inclusion circle area $\lambda(K_i)$ within the forest area $\lambda(F)$, i.e. via their geometric intersection:

$$\pi_i = \frac{\lambda(K_i \cap F)}{\lambda(F)} \tag{4.4}$$

The radius R_i of the tree's inclusion circle K_i is given by $R_i = DBH_i / cf_{i,corr}$ (also referred to as *limiting distance*), where $cf_{i,corr}$ is the original counting factor cf corrected for potential

boundary effects at the forest border. In case of angle-count sampling, we can rewrite π_i as

$$\pi_i = \frac{G_i}{cf_{i,corr}\lambda(F)} \tag{4.5}$$

since the intersection area $\lambda(K_i \cap F)/\lambda(F)$ can be expressed using the trees basal area G_i (in m^2) and the corrected counting factor:

$$\lambda(K_i \cap F) = \frac{G_i}{cf_{i,corr}} \quad \text{where} \quad cf_{i,corr} = cf\frac{\lambda(K_i)}{\lambda(K_i \cap F)} \tag{4.6}$$

Eq. 4.5 in Eq. 4.3 yields the rewritten form of $Y(x_l)$ for angle count sampling that conforms to the definition used in the BWI3 algorithms (Schmitz et al., 2008):

$$Y(x_l) = \sum_{i \in s_2(x_l)} \frac{cf_{i,corr}Y_i}{G_i} = \sum_{i \in s_2(x_l)} nha_i Y_i \tag{4.7}$$

where nha_i is the number of trees per hectare represented by tree i. The local densities on cluster level can then be used to derive the estimated spatial mean $\hat{\bar{Y}}_c$ and its estimated variance $\hat{\mathbb{V}}(\hat{\bar{Y}}_c)$ for any given spatial unit for which $n_2 \geq 2$ (n_2 denoting the number of clusters):

$$\hat{\bar{Y}}_c = \frac{\sum_{x \in s_2} M(x)Y_c(x)}{\sum_{x \in s_2} M(x)} \tag{4.8a}$$

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_c) = \frac{1}{n_2(n_2-1)} \sum_{x \in s_2} \Big(\frac{M(x)}{\bar{M}_2}\Big)^2 (Y_c(x) - \hat{\bar{Y}}_c)^2 \tag{4.8b}$$

with $\bar{M}_2 = \frac{\sum_{x \in s_2} M(x)}{n_2}$.

4.4.2 Design-based small area regression estimators for cluster sampling

All three considered small area estimators use ordinary least square (OLS) regression models to produce predictions of the local density $Y_c(x)$ directly on the cluster level c. We consider the internal model approach, where the estimators take into account that the regression coefficients on the cluster level were fitted using the same sample used for estimation. To apply this to small area estimation, the vector of estimated regression coefficients on the cluster level is found by "borrowing strength" from the entire terrestrial sample s_2 of the current inventory:

$$\hat{\boldsymbol{\beta}}_{c,s_2} = \boldsymbol{A}_{c,s_2}^{-1}\Big(\frac{1}{n_2}\sum_{x \in s_2} M(x)Y_c(x)\boldsymbol{Z}_c(x)\Big) \tag{4.9a}$$

$$\boldsymbol{A}_{c,s_2} = \frac{1}{n_2}\sum_{x \in s_2} M(x)\boldsymbol{Z}_c(x)\boldsymbol{Z}_c^\top(x) \tag{4.9b}$$

$\boldsymbol{Z}_c(x)$ is the vector of explanatory variables on the cluster level, which is calculated as the weighted average of the explanatory variables $\boldsymbol{Z}(x_l)$ on the individual plot levels $x_1, ..., x_l$ (Eq.4.10). The weight $w(x_l)$ is the proportion of the support-area within the forest F used to derive the explanatory variables from the raw auxiliary information.

$$\boldsymbol{Z}_c(x) = \frac{\sum_{l=1}^{M} I_F(x_l)w(x_l)\boldsymbol{Z}(x_l)}{\sum_{l=1}^{M} I_F(x_l)w(x_l)} \tag{4.10}$$

The estimated design-based variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}$ accounts for the fact that the regression model is internal and reflects the sampling variability that occurs when estimating the regression coefficients on the realized sample s_2. It is defined as

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}} = \boldsymbol{A}_{c,s_2}^{-1}\Big(\frac{1}{n_2^2}\sum_{x\in s_2} M^2(x)\hat{R}_c^2(x)\boldsymbol{Z}_c(x)\boldsymbol{Z}_c^\top(x)\Big)\boldsymbol{A}_{c,s_2}^{-1} \tag{4.11}$$

with

$$\hat{R}_c = Y_c(x) - \boldsymbol{Z}_c^\top(x)\hat{\boldsymbol{\beta}}_{c,s_2} = Y_c(x) - \hat{Y}_c(x) \tag{4.12}$$

being the empirical model residuals at the cluster level, which by construction of OLS satisfy the important zero mean residual property, i.e. $\frac{\sum_{x\in s_2} M(x)\hat{R}_c(x)}{\sum_{x\in s_2} M(x)} = 0$.

In the following, we will give a short description of each small area estimator and refer to Mandallaz (2013a); Mandallaz et al. (2016, 2013) if the reader requires additional details or proofs. The estimators have also been implemented in the R-package *forestinventory* (Hill & Massey, 2017) which was used to compute all estimates in this study.

Pseudo Small Area Estimator (PSMALL)

All point information used for small area estimation is now restricted to that available at the sample locations $s_{1,G}$ or $s_{2,G}$ in the small area G, with exception of $\hat{\boldsymbol{\beta}}_{c,s_2}$ and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}$ which are always based on the entire sample s_2. We thus first define the following quantities on the small area level:

$$\hat{\bar{\boldsymbol{Z}}}_{c,G} = \frac{\sum_{x\in s_{1,G}} M_G(x)\boldsymbol{Z}_{c,G}(x)}{\sum_{x\in s_{1,G}} M_G(x)} \qquad \text{where} \quad \boldsymbol{Z}_{c,G}(x) = \frac{\sum_{l=1}^{L} I_G(x_l)\boldsymbol{Z}(x_l)}{M_G(x)} \tag{4.13a}$$

$$Y_{c,G}(x) = \frac{\sum_{l=1}^{L} I_G(x_l)Y(x_l)}{M_G(x)} \qquad \text{and} \quad \hat{Y}_{c,G}(x) = \hat{\bar{\boldsymbol{Z}}}_{c,G}^\top\hat{\boldsymbol{\beta}}_{c,s_2} \tag{4.13b}$$

$$\bar{\hat{R}}_{2,G} = \frac{\sum_{x\in s_{2,G}} M_G(x)\hat{R}_{c,G}(x)}{\sum_{x\in s_{2,G}} M_G(x)} \qquad \text{where} \quad \hat{R}_{c,G}(x) = Y_{c,G}(x) - \hat{Y}_{c,G}(x) \tag{4.13c}$$

Note that the restriction to G, i.e. $I_G(x_l) = \{0, 1\}$, is made on the individual sample plot level x_l, and $M_G(x) = \sum_{l=1}^{L} I_G(x_l)$ thus is the number of sample plots per cluster within the small area. The asymptotically design-unbiased point estimate of *PSMALL* is then defined according to Eq. 4.14a. The first term estimates the small area population mean of G by applying the globally derived regression coefficients to the small area cluster means of the explanatory variables $\hat{\bar{\boldsymbol{Z}}}_{c,G}$. The second term then corrects for a potential bias of the regression model predictions in the small area G by adding the mean of the empirical residuals $\bar{\hat{R}}_{2,G}$ in G. This correction is necessary because the zero mean residual property that holds in F is not guaranteed to hold in small area G under this construction.

$$\hat{Y}_{c,G,PSMALL} = \hat{\bar{\boldsymbol{Z}}}_{c,G}^{\top}\hat{\boldsymbol{\beta}}_{c,s_2} + \bar{\hat{R}}_{2,G} \tag{4.14a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{c,G,PSMALL}) = \hat{\bar{\boldsymbol{Z}}}_{c,G}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}\hat{\bar{\boldsymbol{Z}}}_{c,G} + \hat{\boldsymbol{\beta}}_{c,s_2}^{\top}\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,G}}\hat{\boldsymbol{\beta}}_{c,s_2} + \frac{1}{n_{2,G}(n_{2,G}-1)}\sum_{x\in s_{2,G}}\Big(\frac{M_G(x)}{\bar{M}_{2,G}}\Big)^2(\hat{R}_{c,G}(x) - \bar{\hat{R}}_{2,G})^2 \tag{4.14b}$$

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

The variance-covariance matrix of the auxiliary vector $\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,G}}$ is thereby defined as

$$\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)}\sum_{x\in s_{1,G}}\big(\frac{M_G(x)}{\bar{M}_{1,G}}\big)^2(\boldsymbol{Z}_{c,G}(x) - \hat{\bar{\boldsymbol{Z}}}_{c,G})(\boldsymbol{Z}_{c,G}(x) - \hat{\bar{\boldsymbol{Z}}}_{c,G})^{\top} \tag{4.15}$$

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

The estimated design-based variance of $\hat{Y}_{c,G,PSMALL}$ is given by Eq. 4.14b. Basically, the first term constitutes the variance introduced by the uncertainty in the regression coefficients, whereas the second term expresses the variance caused by estimating the exact auxiliary mean in G using a non-exhaustive sample $s_{1,G}$. The third term is the variance of the model residuals and thus accounts for the inaccuracies of the model predictions. Note that the first term can also be rewritten using g-weights (Mandallaz et al., 2016, pg.14) which ensures some beneficial calibration of the auxiliary variables to the first-phase sample.

Pseudo Synthetic Estimator (PSYNTH)

The PSYNTH estimator is commonly applied when no terrestrial sample is available within the small area G (i.e. $n_{2,G} = 0$). The point estimate (Eq. 4.16a) is thus only based on the predictions generated by applying the globally derived regression coefficients to the small area cluster means of the explanatory variables $\hat{\bar{\boldsymbol{Z}}}_{c,G}$. Note that the bias correction term using the empirical residuals (Eq. 4.14a) can no longer be applied. The PSYNTH estimator thus has a potential unobservable design-based bias.

$$\hat{Y}_{c,G,PSYNTH} = \hat{\bar{\boldsymbol{Z}}}_{c,G}^{\top}\hat{\boldsymbol{\beta}}_{c,s_2} \tag{4.16a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{c,G,PSYNTH}) = \hat{\bar{\boldsymbol{Z}}}_{c,G}^{\top}\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_{c,s_2}}\hat{\bar{\boldsymbol{Z}}}_{c,G} + \hat{\boldsymbol{\beta}}_{c,s_2}^{\top}\hat{\Sigma}_{\hat{\bar{\boldsymbol{Z}}}_{c,G}}\hat{\boldsymbol{\beta}}_{c,s_2} \tag{4.16b}$$

The contribution to the variance by the model residuals in small area G can also no longer be considered (Eq. 4.16b). As a result, the synthetic estimator will usually have a smaller variance than estimators that consider the model residuals, but at the cost of a potential bias. Note that the PSYNTH estimator is still design-based, but one purely has to rely on the validity of the regression model within the small area as it is the case in the model-dependent framework.

Extended Pseudo Synthetic Estimator (EXTPSYNTH)

The EXTPSYNTH estimator (Eq. 4.17) has been proposed by Mandallaz (2013a) as a transformed version of the PSMALL estimator that has the form of the PSYNTH estimator but remains asymptotically design unbiased. It has the advantage that the mean of the empirical model residuals of the OLS regression model for the entire area F and the small area G are by construction both zero at the same time, i.e. $\hat{\bar{R}}_c = \hat{\bar{R}}_{c,G} = 0$. This is realized by *extending* the auxiliary vector $\boldsymbol{Z}_c(x)$ by the indicator variable $I_{c,G}$ which takes the value 1 if the entire cluster lies within the small area G and 0 if the entire cluster is outside G, i.e. $I_{c,G}(x) = \frac{M_G(x)}{M(x)}$. The extended auxiliary vector thus becomes $\mathbb{Z}_c^\top(x) = (\boldsymbol{Z}_c^\top(x), I_{c,G}(x))$ and the new regression coefficient using $\mathbb{Z}_c(x)$ instead of $\boldsymbol{Z}_c(x)$ in Eq. 4.9 is denoted as $\hat{\boldsymbol{\theta}}_{s_2}$. All remaining components are calculated by plugging in $\mathbb{Z}_c(x)$ in Eq. 4.13. A decomposition of $\hat{\boldsymbol{\theta}}_{s_2}$ reveals that the residual correction term is now included in the regression coefficient $\hat{\boldsymbol{\theta}}_{s_2}$.

$$\hat{Y}_{c,G,EXTPSYNTH} = \hat{\bar{\mathbb{Z}}}_{c,G}^\top \hat{\boldsymbol{\theta}}_{c,s_2} \tag{4.17a}$$

$$\hat{\mathbb{V}}(\hat{Y}_{c,G,EXTPSYNTH}) = \hat{\bar{\mathbb{Z}}}_{c,G}^\top \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\theta}}_{c,s_2}} \hat{\bar{\mathbb{Z}}}_{c,G} + \hat{\boldsymbol{\theta}}_{c,s_2}^\top \hat{\boldsymbol{\Sigma}}_{\hat{\bar{\mathbb{Z}}}_{c,G}} \hat{\boldsymbol{\theta}}_{c,s_2} \tag{4.17b}$$

However, it is important to note that $\hat{\bar{R}}_{c,G} = 0$ under the extended regression model only holds if the sample plots $x_1, ..., x_l$ of a cluster are *all* either inside our outside the small area, i.e. $M_G(x) \equiv M(x)$, and thus $I_{c,G}(x) = \frac{M_G(x)}{M(x)}$ can only take the values 1 or 0. Mandallaz et al. (2016) assumed that the effects on the estimates should be negligible as the number of occasions where $M_G(x) < M(x)$ was considered to be small in practical implementations. It was thus a further objective of this study to investigate the actual occurrences and effects of this phenomenon by comparing the estimates of EXTPSYNTH to those of PSMALL.

4.4.3 Measures of estimation accuracy

The estimation precision was quantified by the estimation error, which is the ratio of the standard error and the point estimate:

$$error_{[\%]} = \frac{\sqrt{\hat{\mathbb{V}}(\hat{Y})}}{\hat{Y}} * 100 \tag{4.18}$$

We further calculated the 95% confidence interval for each estimate for visualization purposes. The confidence intervals can also be used heuristically for hypothesis testing to determine whether the point estimates of the three estimators for a given small area are statistically different. The confidence intervals for the SRS estimator can be obtained as:

$$CI_{1-\alpha}(\hat{\bar{Y}}_c) = \hat{\bar{Y}}_c \pm t_{n_2-1,1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\hat{\bar{Y}}_c)} \tag{4.19}$$

The confidence intervals for the PSMALL and EXTPSYNTH estimates are calculated as:

$$CI_{1-\alpha}(\hat{Y}_{c,G,EXTPSYNTH}) = \hat{Y}_{c,G,EXTPSYNTH} \pm t_{n_{2,G}-1,1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\hat{Y}_{c,G,EXTPSYNTH})} \tag{4.20a}$$

$$CI_{1-\alpha}(\hat{Y}_{c,G,PSMALL}) = \hat{Y}_{c,G,PSMALL} \pm t_{n_{2,G}-1,1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\hat{Y}_{c,G,PSMALL})} \tag{4.20b}$$

For the PSYNTH estimates, the confidence intervals are

$$CI_{1-\alpha}(\hat{Y}_{c,G,PSYNTH}) = \hat{Y}_{c,G,PSYNTH} \pm t_{n_2-p,1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\hat{Y}_{c,G,PSYNTH})} \tag{4.21}$$

with p being the number of parameters used in the regression model including the intercept term.

In order to address the potential benefits of the small area estimators compared with the SRS approach, we calculated the *relative efficiency* (Eq. 4.22) which can be interpreted as the relative sample size under SRS needed to achieve the variance under the double-sampling (DS) estimators.

$$rel.eff = \frac{\hat{\mathbb{V}}(\hat{Y}_{SRS})}{\hat{\mathbb{V}}(\hat{Y}_{DS})} \tag{4.22}$$

4.5 Case study

4.5.1 Study area and small area units

The German federal state Rhineland-Palatinate (*RLP*) is located in the western part of Germany and borders Luxembourg, France and Belgium. With 42.3% (appr. 8400 km^2) of the entire state area (19850 km^2) covered by forest, RLP is one of the two states with the highest forest coverage among all federal states of Germany (Thünen-Institut, 2014). The forests of RLP are further characterised by a pronounced diversity in bioclimatic growing conditions that have strong influence on the local growth dynamics as well as tree species composition (Gauer & Aldinger, 2005) and are further characterised by large variety of forest structures ranging from characteristic oak coppices (Moselle valley), pure spruce, beech and scots pine forests (i.a. Hunsrück and Palatinate forest) up to mixed forests comprising variable proportions of oak, larch, spruce, Scots pine and beech. Around 82% of the forest area in RLP are mixed forest stands and 69% of the forest area exhibit a multi-layered vertical structure. The forest area of RLP are divided into 3 ownership classes, i.e. state forest (27%), communal forest (46%) and privately owned forest (27%). The forest service of RLP has the legal mandate to sustainably manage the state and communal forest area (73% of the entire forest area), including forest planning, harvesting and the sale of wood (LWaldG, 2000). For this reason, the entire forest area has been spatially organised in 3 main hierarchical management units (Figure 5.1). On the upper level, RLP has been divided into 45 Forstämter (*FA*), which are further divided into a total number of 405 Forstreviere (*FR*). The next level are the forest stands (104'184 in total) for which expert judgements are conducted by SFIs in a 5 to 10 year period in order to set up management strategies for the upcoming 10 years. The FAs and FRs constituted the SA units for which design-based small area estimations of the mean standing timber volume were calculated by incorporating the available terrestrial inventory data of the BWI3 in the estimators described in Section 4.4. The average area of the SA units was 43'777 ha on the FA-level, and 4624 ha on the FR level.

4.5.2 Terrestrial sample

Rhineland-Palatinate (RLP) is covered by a 2x2 km inventory grid of the German NFI. In the last inventory (BWI3) conducted in the year 2013, timber volume information was derived for 2810 clusters (8092 plots) in the field survey. The local timber volume density on the plot and cluster level for this sample was consequently calculated according to Section 4.4.1. In the framework of this survey, the plot center coordinates were re-measured with the differential global satellite navigation system (DGPS) technique. Knowledge about the exact plot positions were considered crucial to provide optimal comparability between the terrestrial observations and the information derived from the auxiliary information. A comparison of the DGPS coordinates with the so-far

used target coordinates revealed that 90% of all horizontal deviations lay in the range of 25 meters. A detailed analysis of horizontal DGPS errors in RLP by Lamprecht et al. (2017) indicated that 80% of the plots should not exceed horizontal DGPS errors of 8 meters. For 162 plots, the DGPS coordinates were replaced by their target coordinates due to missingness or implausible values. The terrestrial sample size $n_{2,G}$ within the FA units was 46 clusters on average and ranged between 11 and 64. Within the FR units, $n_{2,G}$ was considerably smaller with an average of 5 clusters and a range between 0 and 13.

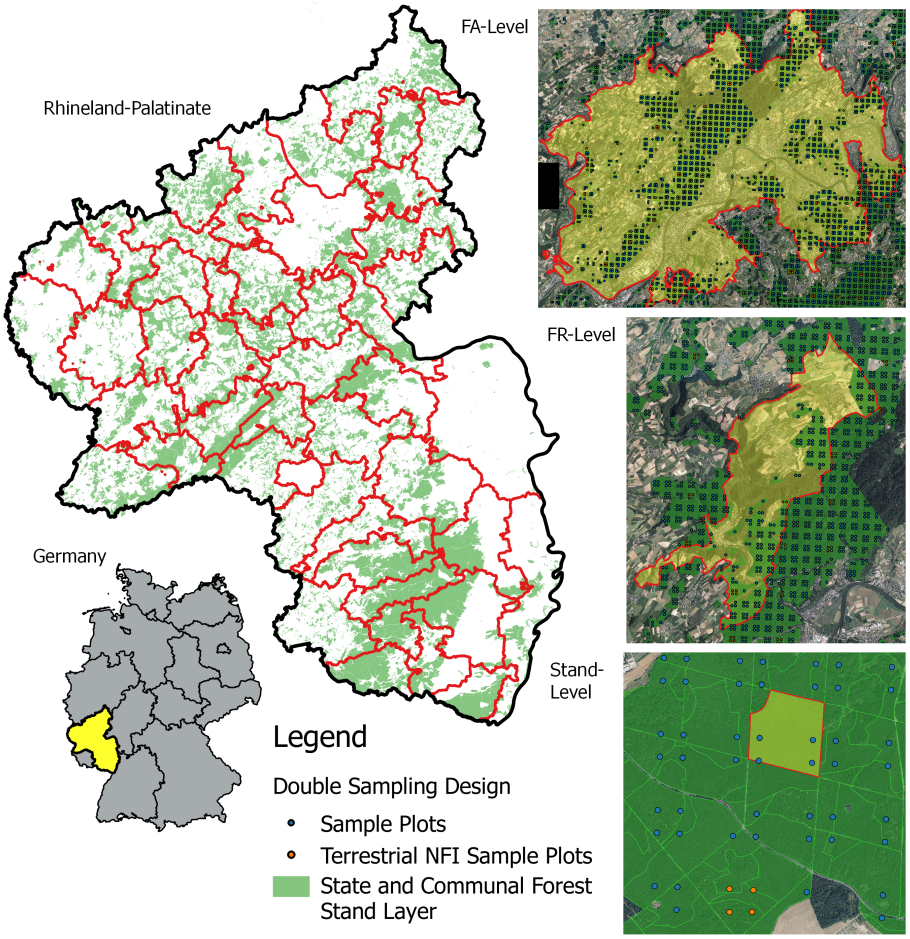


Figure 4.1: *Left*: Study area with delineated FA forest management units. *Right*: Example for each of the three management units (from top to bottom): FA, FR and forest stand unit overlayed with the extended double-sampling cluster design. *Green*: Forest stand polygon layer defining the forest area of this study.

4.5.3 Extension to double-sampling design

In order to apply the small area estimators (Section 4.4.2), the existing NFI design was extended to a double-sampling design by densifying the existing systematic 2x2 km grid to a grid size of 500x500 m that constituted the large first phase s_1 in accordance to Section 4.3 (Figure 4.1, *right*). The existing terrestrial phase s_2 was consequently integrated by replacing the target coordinates of the respective s_1 clusters by the terrestrially measured DGPS coordinates. For our study, we restricted the sampling frame to the communal and state forest. The forest/non-forest decision for each plot was thereby made by a spatial intersection of the plot center coordinates with a

polygon layer of the communal and state forest stands provided by the forest service. Using this stand layer provided the advantage to consistently apply the same forest/non-forest definition to the entire sample s_1 in order to decide about excluding or including a plot in the sampling frame. The terrestrial sample size n_2 was thus reduced to 2055 clusters (5791 plots). Table 4.1 provides a short descriptive summary about the volume densities and the main attributes of the NFI plots located in the state and communal forest sampling frame. The densification led to an average sample size $n_{1,G}$ of 759 clusters (range: 246 – 1022) in the FA units, and 88 clusters (range: 1 – 194) in the FR units.

Table 4.1: Descriptive statistics of the forest observed on NFI sample plots located within communal and state forest area (n=5791).

Variable	Mean	SD	Maximum
Timber Volume (m^3/ha)	300.86	195.55	1375.31
Mean DBH (mm)	354.90	137.22	1123.20
Mean height (dm)	239.60	72.43	497.43
Mean stem density per hectare	101.00	114.01	1010.31

4.5.4 Auxiliary data

LiDAR canopy height model

A prerequisite for the application of the suggested two-phase small area estimators is the identification of suitable auxiliary data available over the entire study area. From 2003 to 2013, the topographic survey institution of RLP conducted an airborne laserscanning acquisition over the entire federal state during leaf-off conditions in order to derive a countrywide digital terrain model (DTM) as well as a digital surface model (DSM). For this study, the recorded ALS data was used to create a canopy height model (CHM) in raster format, providing discrete information about the canopy surface height of the forest area in a spatial resolution of 5 meters (Fig. 4.2, *top*). The CHM was calculated as the difference between the digital terrain model and the digital surface model that were derived by a Delauney interpolation of the ground and first ALS pulses respectively. A more detailed description of the procedure can be found in Hill et al. (2018). The CHM provided the most valuable information to be used in the OLS regression model for predicting the timber volume on the plot and cluster level. However, it should be noted that the prolonged acquisition period of the ALS campaign led to the possibility of poor temporal alignment with the BWI3 survey, sometimes up to 10 years. In addition, the quality of the CHM varied substantially as ALS technology evolved over the years. For example, the ALS acquisitions recorded in 2002 and 2003 exhibited particularly poor quality with about only 0.04 point per m^2, whereas more recent datasets contained more than 5 points per m^2. Furthermore, CHM information was not available at 16 sample locations due to sensor failures. These plots were deleted from the sampling frame and treated as missing at random. This assumption was considered to be reasonable as the respective sample locations did not exclude specific forest structures.

Tree species map

Additional auxiliary data was derived from a countrywide satellite-based classification map predicting the five main tree species (Stoffels et al., 2015), i.e. European beech, Sessile and Pedunculate oak, Norway spruce, Douglas fir and Scots pine (Fig. 4.2, *bottom*). The tree species map has a grid size of 5x5 m and was calculated from 22 bi-temporal satellite images (SPOT5 and RapidEye)

using a spatially adaptive classification algorithm (Stoffels et al., 2012). As timber volume estimation on the tree level is often based on species-specific biomass and volume equations, the use of tree species information has often been stated as a key factor for improving the precision of timber volume estimates (White et al., 2016). In this respect, incorporating the tree species map was particularly attractive as it predicts five of the seven tree species that are used in the BWI3 taper functions (Kublin et al., 2013) to calculate the timber volume of a sample tree. However, due to unavailable satellite data, the tree species map excluded one large patch with an area of 415 km^2 in the south-west part of RLP covering an entire FA unit consisting of 10 FR units. In 9 additional FR units, the tree species information was also missing for a subset of the sample locations due to two additional patches with areas of 76 km^2 and 100 km^2 respectively in the northern part of RLP. For these 19 FR units, small area estimation was thus restricted to using only the available CHM information in the regression model. Thus, 411 of 5791 sample locations (approximately 7%) used to fit the regression model were affected by missing tree species information. A summary of the sample sizes and missing auxiliary data for both the CHM and the tree species map is provided in Table 4.2.

Table 4.2: Sample size for each phase in entire study area. $n_{\{1,2\},plots}$: number of plots. $n_{\{1,2\}}$: number of clusters. TSPEC: tree species map information.

Sampling frame	$n_{1,plot}$	n_1	$n_{2,plot}$	n_2
communal and state forest	96'854	33'365	5791	2055
missing CHM	18	10	0	0
missing TSPEC	7060	3587	414	385
missing CHM *and* TSPEC	3	2	0	0
missing CHM *or* TSPEC	7075	3595	414	385

4.5.5 Calculation of the explanatory variables

Canopy height model

The continuous explanatory variables derived from the CHM were the mean canopy height (*meanheight*) and the standard deviation (*stddev*). The quantities were calculated by evaluating the raster values around each sample location within a circle with a predefined radius of 12 meters, i.e. the support. In order to correct for edge effects at the forest border, the intersection of each support area to the state and communal forest area was determined using a polygon mask provided by the state forest service. The percentage of the support within the forest layer was used as the weight $w(x_l)$ introduced in Eq. 4.10 in order to derive the weighted mean of the explanatory variables on the cluster level. Neglecting the support adjustment would deteriorate the coherence between explanatory variables computed at the forest boundary and the corresponding local density that already includes a potential boundary adjustment, thus introducing unnecessary noise to the model. The boundary adjustment to the support also makes the sampling frame more consistent for the different data sources (Section 4.5.3).

The ALS acquisition year (*ALSyear*) was added as a categorical variable in order to account for the time lag with the terrestrial survey as well as to help explain the heterogeneity in the data introduced by the varying ALS quality. In 2008, a sensor error produced particularly poor ALS quality so the year was divided accordingly into two factor levels, denoted *2008_1* and *2008*. Furthermore, in order to increase the number of observations per factor level the years 2006 and 2007 were pooled together and the same was done for 2012 and 2013. The result was nine factor levels denoted as *2002*, *2003*, *2007*, *2008_1*, *2008*, *2009*, *2010*, *2011* and *2012*.

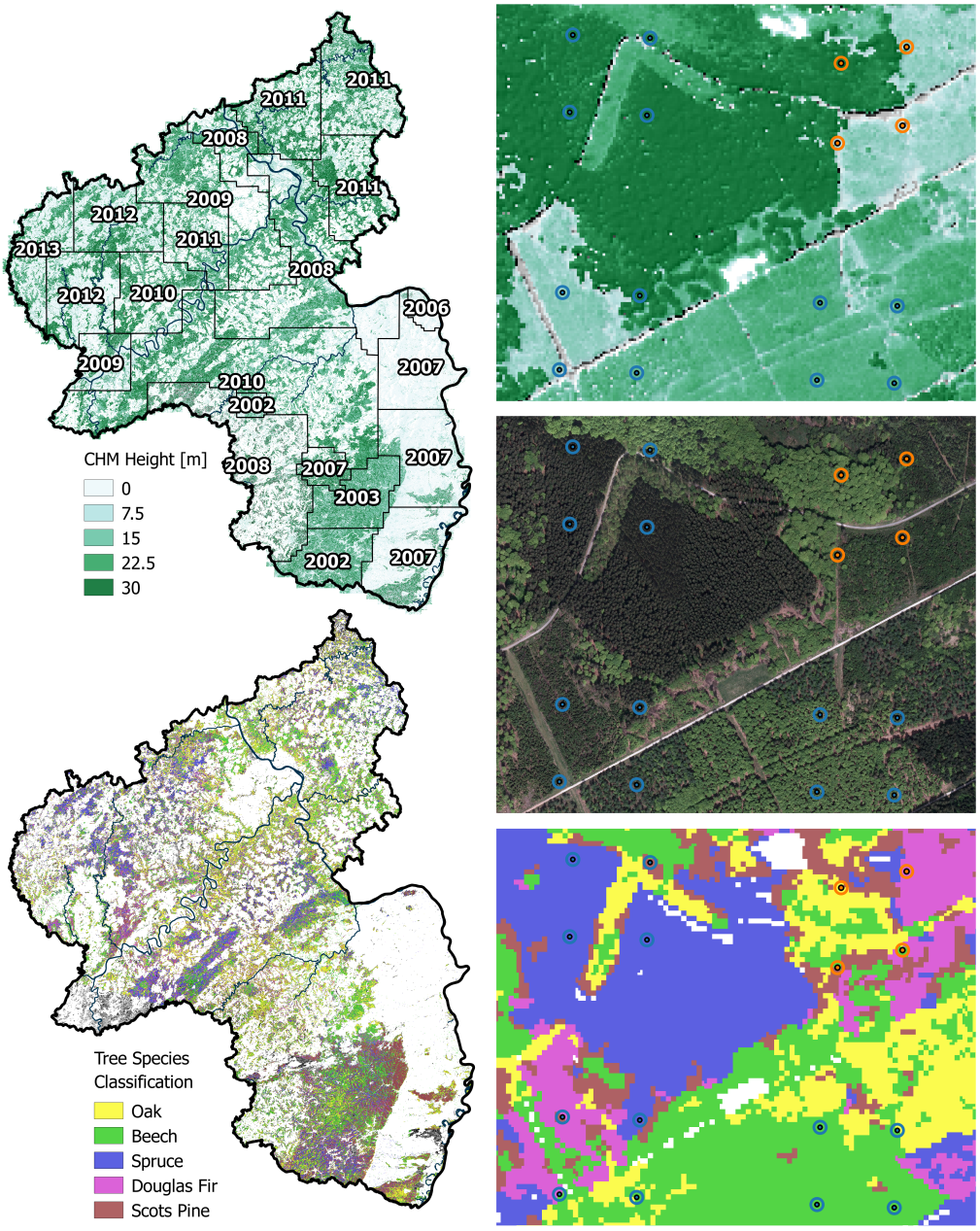


Figure 4.2: Left: CHM (top) and tree species classification map (bottom) available on the federal state level. Right: Magnified illustration of the supports used to derive the explanatory variables from the auxiliary data.

Tree species map

The tree species map was used to predict the main tree species at each sample plot which served as an additional categorical variable called *treespecies*. This involved two consecutive processing steps. In the first step, one of the five tree species was assigned to a sample location if 100% of the raster values within the edge-corrected support were classified as that species. Otherwise, the sample location was assigned the value 'mixed'. Likewise for the CHM variables, the support radius was 12 meters although the use of different support sizes for each explanatory variable would be in agreement with the two-phase estimators presented in Section 4.4.2. When using the *treespecies*

variable in a regression model, the support size and the percentage threshold parameters had to be optimized in order to minimize the variance within each level which subsequently leads to improved model precision. A detailed analysis and description of the optimal parameter processing for the explanatory variables of the present data set is provided in Hill et al. (2018). In a second step, the *treespecies* variable was also passed through a calibration model in order to reduce the effects of misclassification errors on the regression model coefficients and to increase model accuracy. The calibration model consisted of a decision tree from a random forest algorithm (Breiman, 2001) that was trained to predict the actual main plot tree species (known for all terrestrial plots) based on available auxiliary variables. These variables were the predicted *treespecies* variable, the mean canopy height and standard deviation of the CHM, as well as the proportion of coniferous trees estimated from the classification map and the growing region derived from a polygon map. The algorithm was grown with 2000 trees considering 3 of the predictors for each split. We thus applied this calibration model to the *treespecies* variable derived at all sample locations s_1. Table 4.3 gives the classification accuracies (Congalton & Green, 2008) of the *treespecies* variable after calibration.

Table 4.3: Classification accuracies of the *treespecies* variable before and after calibration. n_{ref}: number of terrestrial reference plots. n_{class}: number of classified plots.

Main plot species	Producer's accuracy[%]	User's accuracy[%]	n_{ref}	n_{class}
Beech	22.31	47.02	883	419
Douglas Fir	24.78	48.72	230	117
Oak	11.07	48.48	289	66
Spruce	53.15	61.13	651	566
Scots Pine	22.91	46.07	179	89
Mixed	84.49	64.53	3152	4127
	Overall accuracy: 61.96%		5384	5384

4.5.6 Regression Model

The model selection process for this study required a substantial time commitment due to sophisticated challenges such as: a) the heterogeneity of the remote sensing data, b) the identification of the optimal support sizes under angle count sampling, and c) the incorporation of tree species information. Here, only a summary of the extensive analysis that was performed is provided but the reader can refer to Hill et al. (2018) if more details are desired.

The model with highest adjusted R^2 and lowest RMSE was achieved using *meanheight*, *meanheight*^{2}, *stddev*, *ALSyear* and *treespecies* as main effects, and including interaction terms between *meanheight* and *ALSyear*, *stddev* and *ALSyear*, *meanheight* and *stddev*, and *meanheight* and *treespecies*. Summary information about the adjusted R^2, RMSE and RMSE% of the selected models is provided in Table 4.4. As the two-phase estimators described in Section 4.4.2 derive and apply the regression coefficients and the residuals on the aggregated cluster level, we re-evaluated the model as used in the estimators on the cluster level (formulas given in Appendix) and found improved model fits compared to the plot level (adjusted R^2 of 0.59 and RMSE of 101.61 m^3/ha and 33.6%). The stratification by the ALS acquisition year substantially improved the model fit, indicating that it is an effective means in accounting for the noise in the data caused by ALS quality variations and time-gaps between the ALS and the terrestrial survey. However, the stratification led to a highly unbalanced data set when a further *treespecies* stratification was included. For this reason, a individual species modeling within each *ALSyear* stratum remained infeasible, but might have further improved the model fit. An additional evaluation of the model's performance within each ALS acquisition year stratum revealed that the quality of the model fit substantially varied

between the strata (Table 4.5). In particular, values above the overall adjusted R^2 were higher in ALS acquisition years close to the terrestrial survey date compared to years with larger time gaps.

As described in Section 4.5.4, the information of the tree species classification map was missing within 1 FA and 19 FR units. For these small area units, we applied the regression model without the *treespecies* variable (Table 4.4, reduced model). However, the adjusted R^2s of the full and reduced model were found to be very similar on both the plot and cluster level. This implied that the variance reduction of the reduced model when applied to the two-phase estimators would likely be comparable to that of the full model, which is why a joint evaluation of the estimation results was performed (Section 5.3).

Table 4.4: Model fit metrics for the two OLS regression models on the cluster level. Interaction terms are indicated by ':'. () give the respective values on the plot level.

model terms	model	R^2_{adj}	RMSE	RMSE%
meanheight + stddev + meanheight^2 + treespecies + ALSyear + meanheight:treespecies + meanheight:ALSyear + meanheight:stddev + stddev:ALSyear	full model	0.58 (0.48)	90.11 (139.22)	29.76 (45.98)
meanheight + stddev + meanheight^2 + ALSyear + meanheight:ALSyear + meanheight:stddev + stddev:ALSyear	reduced model	0.55 (0.45)	95.23 (144.13)	31.65 (47.60)

Table 4.5: R^2, RMSE and RMSE% on the cluster level of the full regression model within ALS acquisition year strata (*ALSyear*). $Area_{ALSyear}$: Area covered by ALS acquisition given in km^2. n: sample size of validation data. () give the respective values on the plot level.

ALSyear	$Area_{ALSyear}$	R^2	RMSE	RMSE%	n
2012	2807	0.65	98.52	29.62	156
		(0.61)	(135.84)	(44.87)	(408)
2011	4361	0.60	96.89	29.66	354
		(0.57)	(146.21)	(48.29)	(883)
2010	4182	0.64	76.38	27.57	420
		(0.51)	(120.90)	(39.93)	(1171)
2009	2100	0.53	92.22	33.31	218
		(0.42)	(133.42)	(44.07)	(559)
2008	2968	0.61	87.10	32.20	247
		(0.48)	(130.38)	(43.06)	(701)
2008_1	2116	0.43	117.99	33.64	157
		(0.33)	(175.43)	(57.94)	(394)
2007	3498	0.56	82.43	26.57	135
		(0.46)	(136.47)	(45.08)	(418)
2003	602	0.34	85.92	27.31	145
		(0.27)	(154.48)	(51.02)	(529)
2002	775	0.52	87.25	27.22	97
		(0.44)	(141.55)	(46.75)	(314)

Concerning the existence of outliers or leverage points in the training set for the model, it should be noted that it is more problematic for PSMALL, PSYNTH and EXTPSYNTH to simply remove them as one might be inclined to do in a model-dependent context. Strictly speaking,

outlier removal in the design-based context essentially means that those plots, and implicitly any potentially similar plots that were not realized in the selected sample, have been removed from the sampling frame and are no longer considered part of the forest area of interest. While this may be valid for some obvious typos or measurement errors, it is generally not advisable to manipulate the sampling frame after observing data collected from it, especially when the observation in question lies within the small area of interest. However, for sake of completeness, we conducted an analysis of influential observations (Fahrmeir et al., 2013, pp. 160–167) on the plot level for the full regression model. We calculated the leverage values and found that 10% of all observations exceeding a predefined critical threshold, i.e. twice the average of the hat matrix diagonal entries. Further investigation revealed that several leverage points showed unusually large *meanheight* values compared to their respective timber volume densities. They tended to occur in ALS acquisition years with longer time gaps to the terrestrial survey date and were thus more likely caused by harvesting activities in the sample plot area. Although these areas likely affected by harvest should clearly not be removed from the sampling frame, it does provided more justification for the inclusion of the *ALSyear* variable to mitigate these effects.

4.6 Results

4.6.1 General estimation results

An application of the SRS, PSMALL and EXTPSYNTH estimator was not feasible for 17 of all 405 FR-units due to an insufficient terrestrial sample size of $n_{2,G} < 2$. We further restricted the calculation of the PSMALL and EXTPSYNTH estimator to small area units with a minimum terrestrial sample size of $n_{2,G} \geq 4$ to avoid unstable estimates. This affected 65 additional FR units and limited unbiased two-phase estimations to 321 (79%) of the 405 FR units. It should be noted that also the PSYNTH estimator could not be applied for 2 FR-units since $n_{1,G} < 2$. Due to substantially larger sample sizes, all estimators could however be applied to all 45 FA units. The average value and the range of the mean timber volume estimates over the evaluated FA and FR units turned out to be very similar between all estimators (Table 4.6). An additional pairwise comparison of the 95% confidence intervals revealed that the four estimators did in fact not produce statistically different point estimates for all FA and FR units. This confirmed that the differences between the estimators are solely found in the precision which they provide for the point estimates.

Table 4.6: Descriptive summary of point estimates and estimation errors on the two forest district levels. N_u: number of evaluated small area units.

District level	Estimator		Point estimates			error$_{[\%]}$		
			mean	min	max	mean	min	max
FA	SRS	(N_u=45)	300.16	215.91	392.84	6.69	3.87	13.21
	PSMALL	(N_u=45)	307.29	209.26	417.10	5.16	3.46	14.33
	EXTPSYNTH	(N_u=45)	307.27	209.01	415.02	4.78	3.25	13.88
	PSYNTH	(N_u=45)	306.90	223.51	409.92	2.34	1.54	3.95
FR	SRS	(N_u=388)	301.83	99.89	612.13	18.32	0.34	104.97
	PSMALL	(N_u=321)	308.15	159.64	568.67	12.24	3.48	44.94
	EXTPSYNTH	(N_u=321)	308.38	154.07	544.34	11.34	3.60	40.91
	PSYNTH	(N_u=403)	307.82	166.01	444.29	4.65	2.56	62.51

4.6.2 Estimation error

On both small area levels, the design-unbiased estimators PSMALL and EXTPSYNTH led to a substantial reduction in the estimation error compared to the SRS estimator (Fig. 4.3). On the FA level, the SRS estimator yielded an estimation error of 6.7% on average compared to 5.2% and 4.8% under EXTPSYNTH and PSMALL respectively (Table 4.6). The cumulative error distribution (Fig. 4.3, left) reveals that under the SRS estimator, errors less than 5% were achieved for 17% of the FA units (8 of 45). This proportion could be increased to 62% (28 FA units) and 73% (33 FA units) by application of the PSMALL and EXTPSYNTH estimator. 95% of all estimates exhibited errors less than 9.5% under the SRS estimator and less than 6.6% when using PSMALL or EXTPSYNTH. Estimation errors higher than 10% only appeared twice for each of the three estimators.

Although the estimation errors were substantially larger overall on the FR level compared to the FA level due to smaller sample sizes, the error reduction from SRS by PSMALL and EXTPSYNTH were even more pronounced (Fig. 4.3, right). The average error under the SRS estimator was 18.3%, while it was 11.3% and 12.2% under PSMALL and EXTPSYNTH (Table 4.6). Errors smaller than 10% were achieved for 15% of the FR units by the SRS estimator, and for 46% by the PSMALL and PSYNTH estimator. 95% of the 321 FR units where PSMALL and EXTPSYNTH could be applied exhibited errors less than 20%. In comparion, the SRS estimates resulted in errors less than 36.6% for 95% of the 388 FR units.

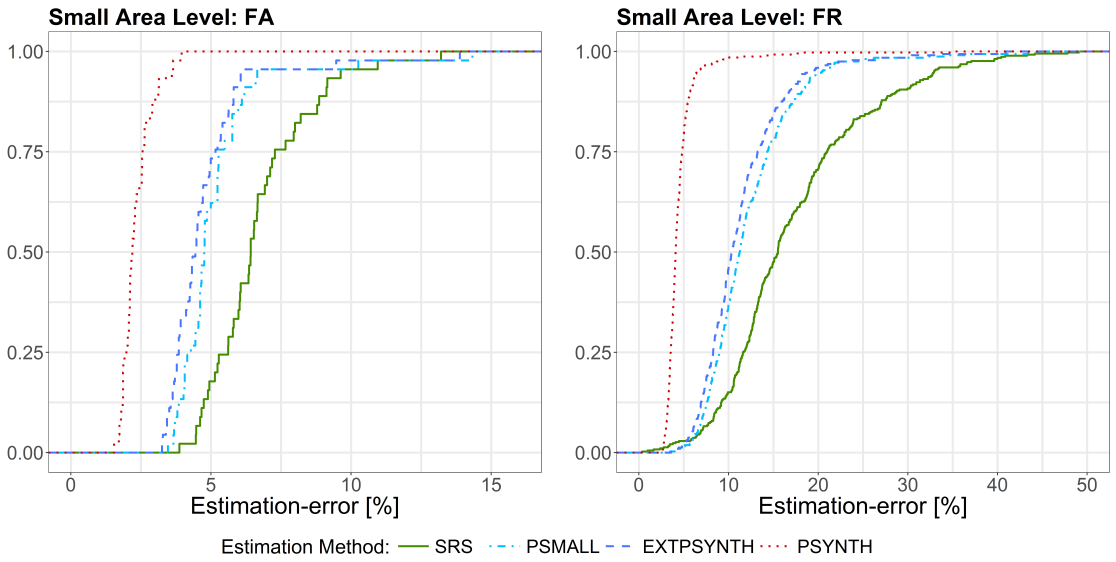


Figure 4.3: Cumulative distribution of estimation errors under SRS, PSMALL, EXTPSYNTH and the PSYNTH estimator. *Left*: Results for the 45 FA units. *Right*: Results for the 388 (SRS), 321 (PSMALL, EXTPSYNTH) and 403 (PSYNTH) FR units.

On both small area levels, the PSYNTH estimator resulted in much smaller estimation errors compared to PSMALL and EXTPSYNTH. This was as expected, since the PSYNTH variance estimate does not take the residual variation in each small area unit into account (Section 4.4.2). Compared to the asymptotically design-unbiased estimators PSMALL and EXTPSYNTH, the estimation errors produced by PSYNTH thus seem to be too optimistic. One should also recall that the estimates of the PSYNTH estimator are potentially design-biased.

4.6.3 Comparison of PSMALL and EXTPSYNTH

Figure 4.3 reveals that the error distribution of PSMALL and EXTPSYNTH are very similar, with PSMALL showing marginally higher estimation errors. In order to investigate the differences between PSMALL and EXTPSYNTH, we compared the g-weight variances of both estimators for all 321 FR units (Fig. 4.4, left). As obvious, PSMALL yielded slightly larger variances for the vast majority of the estimates. As addressed in Section 4.4.2, one possible explanation for differences was the effect of one or more clusters not entirely being included in a small area unit, as this would constitute an assumption violation of the EXTPSYNTH estimator. This violation was actually observed in 155 of the 321 FR units (48%). We compared the variances of PSMALL and EXTPSYNTH for all small areas that did not have the violations using a Wilcoxon Signed-Rank Test (Wilcoxon et al., 1970) on a 5% significance level. This test was also performed pairwise for groups $n_{2,G} \leq 6$, $n_{2,G} > 6$ and $n_{2,G} > 10$. The distribution of variances from EXTPSYNTH was found to be highly significantly lower than that of PSMALL except for the group of $n_{2,G} > 10$. The latter was expected since the variances of both estimators are asymptotically equivalent under large terrestrial sample sizes $n_{2,G}$ within the small area (Mandallaz et al., 2016, pp.17–18). This was also confirmed by a visual comparison of the absolute differences in the variances (Fig. 4.4, right) which decreased with increasing terrestrial sample size. Performing the same comparison for small areas with violations also revealed the EXTPSYNTH variances to be significantly smaller than the respective PSMALL variances until sample sizes $n_{2,G} > 10$. Based on these investigations, it was not possible to determine whether the differences for sample sizes smaller than 10 were caused by the violations or just reflect the general tendency of EXTPSYNTH to produce smaller variances than PSMALL under small sample sizes. However, a visual inspection provided some evidence that the violations created a statistically significant influence on the EXTPSYNTH variance (Fig. 4.4, left, red diamonds) that makes it appear to be slightly over-optimistic. For sample sizes of $n_{2,G} < 6$, a weakly significant difference between the EXTPSYNTH variances of those small areas with violations and the EXTPSYNTH variances without violation was also indicated by an unpaired Wilcoxon Rank-Sum Test. However, the differences were still marginal and a comparison of the confidence intervals of PSMALL and EXTPSYNTH revealed that the variance differences did not lead to statistically significant point estimates.

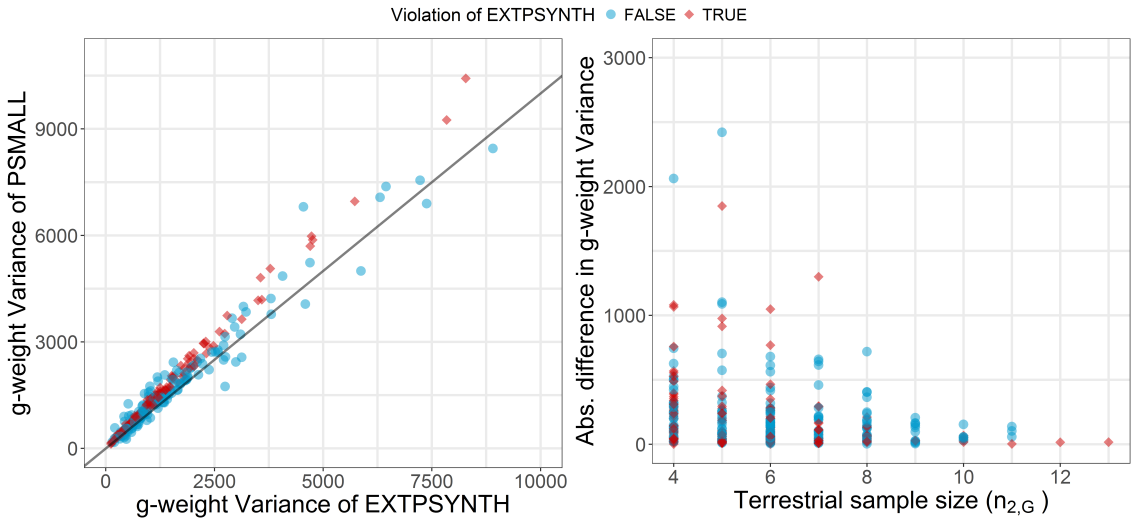


Figure 4.4: *Left*: Comparison of the g-weight variance between the PSMALL and the EXTPSYNTH estimator for the 321 FR units. *Right*: Difference in g-weight variance between the PSMALL and the EXTPSYNTH estimator in dependence of the terrestrial data ($n_{2,G}$) in the FR unit.

4.6.4 Variance reduction compared to SRS

The variance reduction relative to SRS for PSMALL and EXTPSYNTH are described in Figure 4.5 and Table 4.7. A direct comparison of the variances within the small area units revealed that the application of the design-unbiased estimators (PSMALL, EXTPSYNTH) led to a variance reduction compared to SRS in all FA units. In 75% of the FA units, the EXTPSYNTH estimator was able to reduce the variance by up to 54.1%. The reduction in variance can also be expressed in the relative efficiency values, which were 2.02 on average and ranged between 1.18 and 4.13 on the FA level. On FR level, the reduction in variance even reached values of 90% and relative efficiencies of 30 (Table 4.7 and Fig. 4.5). The PSMALL estimator again yielded slightly lower variance reductions and relative efficiencies due to the generally smaller variances of the EXTPSYNTH estimator (Section 4.6.3).

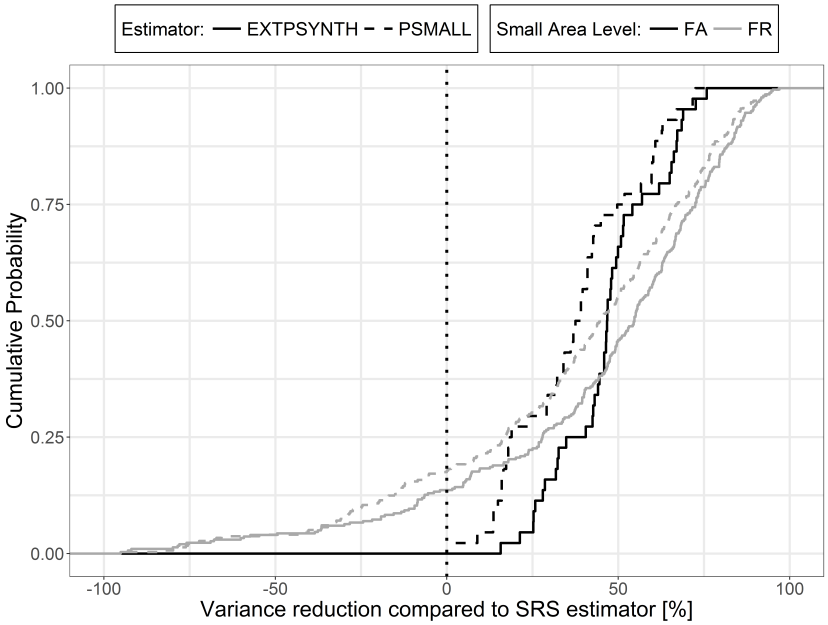


Figure 4.5: Cumulative distribution of variance reduction by the PSMALL and EXTPSYNTH compared to the SRS estimator for the 45 FA and 321 FR units.

Table 4.7: Descriptive summary of variance reduction compared to SRS and relative efficiencies on the two forest district levels. N_u: number of evaluated small area units.

District level	Estimator		Variance reduction [%]			relative efficiency		
			mean	min	max	mean	min	max
FA	PSMALL	(N_u=45)	33.51	2.6	72.5	1.74	1.03	3.64
	EXTPSYNTH	(N_u=45)	43.30	15.7	75.8	2.03	1.18	4.13
FR	PSMALL	(N_u=321)	12.48	-1203.9	96.8	2.54	0.08	31.61
	EXTPSYNTH	(N_u=321)	24.75	-892.7	97.0	2.95	0.10	33.70

Cases also occurred on the FR level where one or both two-phase estimators produced larger variance values than under the SRS estimator. This happened in 19% of the FR units under the EXTPSYNTH, and in 24% of the FR units under the PSMALL estimator. One possible reason for this was supposed to be a large residual variance due to a poor performance of the regression model

within the small area unit. In order to investigate this hypothesis, we analyzed the three variance terms of the PSMALL estimator (eq. 4.14b), i.e. the variance introduced by the uncertainty of the regression coefficients (term 1), the variance caused by estimating the auxiliary means (term 2), and the variance of the model residuals (term 3). In general, the residual term is expected to make the largest contribution to the overall variance since it's sample size is based on $n_{2,G}$ whereas the auxiliary term and the coefficient term are based on larger sample sizes, i.e. $n_{1,G}$ and n_2 respectively. Figure 4.6 illustrates the share of the overall variance by the residual term of the PSMALL estimator scaled by the overall percentage reduction or increase of the variance compared to SRS for various small area sample sizes $n_{2,G}$. The residual term generally constitutes the dominating part of the PSMALL variance (around 84% on average). Although high residual term dominance does not necessary indicate that the PSMALL variance will be disproportionately large, as apparent from Figure 4.6 (*right*), the vast majority of the small areas where the PSMALL variance was larger than the SRS variance had residual terms contributing over 75% to the overall PSMALL variance. Furthermore, the magnitudes of the worst cases tended to occur in lower sample sizes. For example, of the FR units that saw variance increases where $n_{2,G} = 4$, the average increase was 272%, compared to 62% for FR units with $n_{2,G} > 4$ (Fig. 4.6, *left*). In comparison, the magnitude of the variance decreases were far more homogeneous than for the variance increases regardless of terrestrial sample size. Since $n_{2,G}$ is the same for PSMALL and SRS, this implies that the sum of square residuals for the model are likely larger than the sum of square local densities for the clusters in $s_{2,G}$ indicating the presence of outliers with large residuals in the problematic small areas. This situation is likely to arise when there was forest loss after the ALS scanning but before the terrestrial survey year.

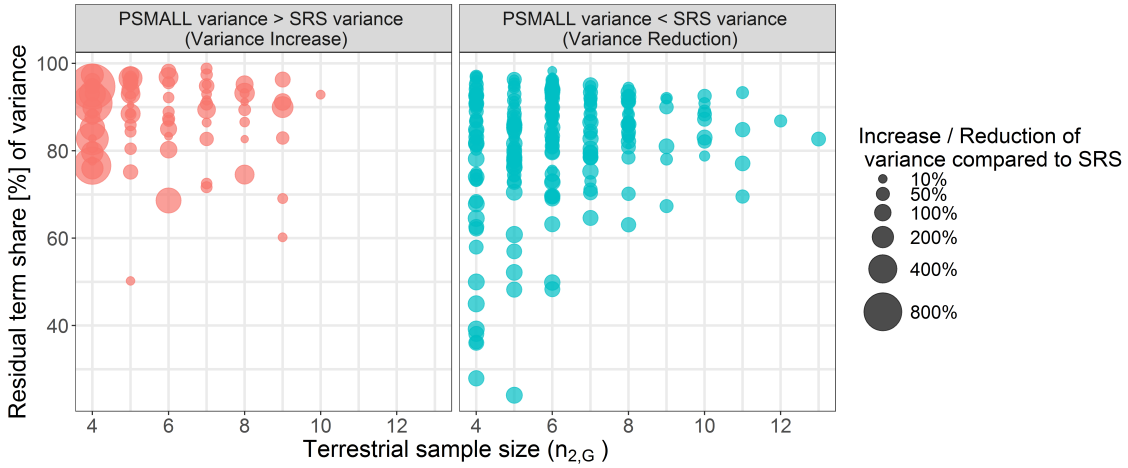


Figure 4.6: Share of the overall variance by the residual term of the PSMALL estimator for various small area sample sizes. Points are scaled by the overall percentage reduction/increase of the variance compared to SRS.

4.7 Discussion

4.7.1 Performance of estimators

The aim of this study was to investigate the performance of model-assisted design-based estimators for small area estimation of mean standing timber volume on two spatial forest management levels in Germany. It was of particular interest to gather information about the estimation error levels that can be attained using German NFI data that is characterized by low sampling intensities in the area of interests. To address these research questions, we applied the SRS, the PSMALL and

the EXTPSYNTH estimators for cluster sampling to two forest management levels consisting of 45 and 405 small area units respectively in the German federal state of Rhineland-Palatinate.

Our study showed that on both small area levels, the PSMALL and the EXTPSYNTH estimators generally led to a substantial reduction in estimation error compared to the standard one-phase SRS estimator. On the upper management level (FA districts), PSMALL and EXTPSYNTH produced estimation errors smaller than 5% for 73% of the small areas compared to only 17% under the one-phase SRS estimator. The same level of precision could not be achieved on the lower management level (FR districts) primarily due to substantially smaller terrestrial sample sizes. However, in 95% of the FR units, the estimation errors could be limited to 20% compared to 40% under SRS. A pairwise comparison of the confidence intervals revealed that the estimators did not produce significantly different point estimates. The much smaller estimation errors of the PSYNTH estimator reflected the fact that it does not try to correct for potential bias in the point estimate which can lead to overly optimistic estimation errors and confidence intervals. One should thus prefer the unbiased estimates of PSMALL or EXTPSYNTH.

For several FR units, it was observed that the PSMALL and the EXTPSYNTH estimator can occasionally produce larger variances than the SRS estimator. It is important to note that this is in perfect agreement with the theory of both two-phase estimators and can theoretically appear if the residual variance in the small area, which generally constitutes the dominating part of the two-phase variance, turns out to be much higher than the variance of the terrestrial data in the small area. The empirical findings of our study suggest that such cases can particularly occur if moderate or poor model fits within a small area are combined with small terrestrial sample sizes (≤ 5) in the small area. A closer look on these small areas thus might reveal the reason for the poor prediction performance and help to improve the model fit. Nonetheless, it should be kept in mind that small terrestrial sample sizes can also cause the SRS estimator to not reflect the actual variation of the local density within a small area. In this case, the two-phase variance estimate might be larger but more realistic. Whereas a visual analysis of aerial images, remote sensing data or stand maps might give some further evidence for or against this hypothesis, a definite proof is practically infeasible.

We were also able to empirically confirm that the EXTPSYNTH estimator generally produces slightly smaller variances and estimation errors than the PSMALL estimator. This is most probably caused by marginally smaller model residuals due to the intercept adjustment to the terrestrial data in the small area unit, which is primarily a means to ensure the zero mean residual property of the EXTPSYNTH estimator. However, our analysis indicated that the difference between the two estimators is negligible for sample sizes ≥ 10 due to their asymptotic equivalency. We further investigated a potential impact on the EXTPSYNTH variance caused by the assumption violation that one or more clusters are not entirely included in the small area unit and found a slight but statistically significant tendency to be over-optimistic for sample sizes smaller than 6. More empirical evidence must be gathered before generalizing this as a rule of thumb for the application of the EXTPSYNTH under cluster sampling. It thus seems recommendable to prefer the EXTPSYNTH to the PSMALL estimator if its assumptions are not violated since it yields slightly smaller variances under mathematically soundness. Even if the differences between both estimators were marginal and did not lead to significantly different point estimates, PSMALL can serve as a safe alternative if the EXTPSYNTH assumption is violated. Aside from this, calculating both PSMALL and EXTPSYNTH and subsequently compare their results is always recommended to reveal suspicious deviations.

4.7.2 Auxiliary data

The auxiliary data used in our study were derived from two remote sensing sources, i.e. an ALS canopy height model and a tree species classification map. Likewise in many similar studies, the

ALS mean canopy height proved to be the explanatory variable with highest predictive power. However, the large time-gaps of up to 10 years between the ALS acquisition and the terrestrial survey date caused the substantial introduction of artificial noise in the data. Whereas a post-stratification to the ALS acquisition years was an effective means to counteract the implied residual inflation, several leverage points were unambiguously caused by the temporal asynchronicity. Undetectable forest loss during the gap between the ALS acquisition and the NFI was also likely a cause for high residual variance in some small areas compared to the terrestrial data variance, which subsequently led to higher variances than the SRS estimator. As opposed to the ALS data, the availability of a country-wide tree species classification map has yet been unique among all German federal states. Whereas the study of Hill et al. (2018) already showed that the tree species information was able to improve the model fit, it has yet not been used to its full potential. One reason for this was the impossibility of modeling individual tree species within each ALS acquisition year, which would add further explanatory power. Another reason was the lack of available satellite data for classification in some parts of the country, which led to missing values in the inventory data and restricted 19 FR units to a simpler regression model. Promising steps with respect to more up-to-date canopy height information have already been made, as the topographic survey institution of RLP will from this year on provide a country-wide canopy height model derived from aerial imagery acquisitions. These campaigns will in the future be conducted in a two-year period and allow to derive canopy height information matching the dates of terrestrial forest inventories. A study of Kirchhoefer et al. (2017) recently indicated that similar model performance for German NFI data can be achieved using such imagery-based canopy height models. Due to the improved coverage and repetition rate of the Sentinel-2 satellite (ESA, 2017), the tree species classification map will in the future be updated each year. We consider these alternative auxiliary data sources to also solve the problem of missing explanatory variables at inventory plots. One could also make use of the exhaustive information within the two-phase estimators by using the true the auxiliary means (Mandallaz, 2013a; Mandallaz et al., 2013), which could further decrease estimation errors. Previous studies of Mandallaz et al. (2013) however showed that given a reasonable large sample size of the first phase, the differences in the estimation error are usually small. With respect to the substantial improvements in the temporal synchronicity between auxiliary and terrestrial inventory data, we consider the demonstrated double-sampling approach also to be very efficient for change estimation (Massey & Mandallaz, 2015b).

4.8 Conclusion

The study led to two major conclusions: (1) the EXTPSYNTH and PSMALL estimator generally achieved substantially smaller estimation errors on the two investigated forest district levels compared to the SRS estimator. The demonstrated double-sampling procedure thus constitutes a major contribution to an increase in value of the existing German NFI data on the federal state level. However, it is not possible to conclude from our study results alone whether the realized error levels are already sufficient enough in order to support forest planning decisions. Thus, further investigations are necessary in close cooperation with the forest authorities. A first study will concentrate on testing the EXTPSYNTH and PSMALL confidence intervals as a validation source for the stand-wise inventories. (2) Despite the quality restrictions in the ALS data and the tree species map, the two data sources were found to be well suited to model the mean timber volume on plot and cluster levels. With respect to frequently updated aerial canopy height models and tree species maps, it will thus be of interest to investigate the model and estimation performances that can be expected for future applications. In this framework, the incorporation of additional auxiliary data and the extension to change estimation seem the reasonable next steps to be explored towards an operational implementation of the demonstrated double-sampling procedure.

Acknowledgements

We want to express our gratitude to Prof. H. Heinimann (Chair of Land Use Engineering, ETH Zurich) for supporting this study. We also want to explicitly thank Johannes Stoffels and Henning Buddenbaum from the Environmental Sensing and Geoinformatic Group of University of Trier for providing the ALS data and tree species classification map, and Kai Husmann and Christoph Fischer from the Northwest German Forest Research Institution Göttingen for their advice in processing the terrestrial inventory data. Special gratitude is also owed to Thomas Riedel from the Thünen Institute for providing the densified NFI sample grid, and Alexander Massey for proofreading.

Chapter 5

Accuracy assessment of timber volume maps using forest inventory data and LiDAR canopy height models

Andreas Hill[1], Jochen Breschan[1], Daniel Mandallaz[1]

[1]ETH Zürich
Department of Environmental Systems Science, Universitätstrasse 16, 8092 Zürich, Switzerland

Published in:
Forests 5 (2014): 2253-2275
(DOI: 10.3390/f5092253)

– Jochen Breschan supported the formulation of the optimization model.

– Daniel Mandallaz supported the statistical data analysis.

Abstract

Maps of standing timber volume provide valuable decision support for forest managers and have therefore been the subject of recent studies. For map production, field observations are commonly combined with area-wide remote sensing data in order to formulate prediction models, which are then applied over the entire inventory area. The accuracy of such maps has frequently been described by parameters such as the root mean square error of the prediction model. The aim of this study was to additionally address the accuracy of timber volume classes, which are used to better represent the map predictions. However, the use of constant class intervals neglects the possibility that the precision of the underlying prediction model may not be constant across the entire volume range, resulting in pronounced gradients between class accuracies. This study proposes an optimization technique that automatically identifies a classification scheme which accounts for the properties of the underlying model and the implied properties of the remote sensing support information. We demonstrate the approach in a mountainous study site in Eastern Switzerland covering a forest area of 2000 hectares using a multiple linear regression model approach. A LiDAR-based canopy height model (CHM) provided the auxiliary information; timber volume observations from the latest forest inventory were used for model calibration and map validation. The coefficient of determination ($R^2 = 0.64$) and the cross-validated root mean square error ($RMSE_{cv} = 123.79$ m^3/ha) were only slightly smaller than those of studies in less steep and heterogeneous landscapes. For a large set of pre-defined number of classes, the optimization model successfully identified those classification schemes that achieved the highest possible accuracies for each class.

5.1 Introduction

5.1.1 Context and problem

Among the multitude of information that forest inventories are expected to provide (McRoberts et al., 2010), knowledge about available standing timber volume on the national, regional, as well as enterprise level is still of high interest. Since on these spatial levels, a full census is too cost-intensive and, in most cases, even practically unfeasible, a broad range of methods in the framework of sampling theory has been developed and applied to estimate this quantity (**?**Mandallaz, 2008; Schreuder et al., 1993). The strength of forest inventory methods relying on design-based procedures is that (at least asymptotically) unbiased point and variance estimates can be obtained, and this without assuming the applied prediction models to be correct in the classical statistical (model-dependent) sense. An important advancement in increasing this accuracy without, at the same time, increasing the number of costly terrestrial samples has been achieved by combining terrestrial samples with auxiliary information provided by remote sensing data, so-called two-phase or double-sampling procedures (**?**Mandallaz, 2008; Cochran, 2007; Köhl et al., 2006). In this context, especially airborne laser scanning (ALS) data have proven to provide a high degree of information for timber volume estimation (Holmgren, 2004; Næsset, 2002, 2007). It has recently been shown that the efficiency of two-phase sampling can be further increased by extending this procedure to stratification (Saborowski et al., 2010; von Lüpke, 2013) or by using part of the auxiliary information exhaustively when remote sensing data are covering the entire inventory area (Mandallaz et al., 2013). The two-phase procedure is thus not restricted to large forest areas, but has also been applied in the context of small area estimation (Breidenbach & Astrup, 2012). Given that the number of terrestrial samples in the small area is sufficiently large (i.e., one is not restricted to the application of synthetic estimations), even for small areas the accuracy specifications are ensured to be unbiased (Mandallaz, 2013a). While these forest inventory methods have the advantage of supplying reliable accuracy specifications for their estimates, they do not provide information about the spatial distribution of the estimated quantity. However, the availability of spatially explicit

stand information is of prime importance for efficiently locating forest management operations.

Accordingly, mapping the spatial distribution of standing timber volume has been the subject of various recent studies. The statistical models that have been used for mapping can be divided into parametric models, particularly linear regression models (Tonolli et al., 2011; Van Aardt et al., 2008), and non-parametric models (Franco-Lopez et al., 2001; Latifi et al., 2010; Nothdurft et al., 2009). Among the non-parametric models, k-NN imputation has become increasingly popular due to its simplicity and easy implementation (Magnussen & Tomppo, 2014). k-NN approaches have been investigated and applied in the model-dependent framework of forest inventory with promising results (McRoberts et al., 2007) and have also been used for the mapping of various forest attributes (Beaudoin et al., 2014; Chirici et al., 2012; Tomppo, 2006). Haara & Kangas (2012) compared the k-NN method to linear regression in a simulation study and found the two methods to perform similarly well. Especially in the case where the relationship between observations and the auxiliary variable followed a linear trend, the regression model performed better than the k-NN approach. Fehrmann, et al. (Fehrmann et al., 2008) came to a similar result when comparing linear and linear mixed effect models to an instance-based k-NN approach for single-tree biomass estimation. Also in their case, the performance of the k-NN approach and the linear mixed model only differed marginally, and both methods were slightly superior to simple linear regression. On the other hand, they also confirmed that the application of k-NN methods can be an effective and promising method if no a priori knowledge about the relationship between target and auxiliary variable(s) exists, particularly if the relationship is considered to be complex due to random and interaction effects. However, they also raised the question of whether a k-NN approach should be used in situations where the functional relationship among variables is approximately known. Additionally, the performance of k-NN estimation and its potential superiority to already existing methods has also been investigated in the design-based framework of forest inventory (Baffetta et al., 2010, 2009). While in several cases, the proposed k-NN estimator of the population mean achieved smaller errors than the Horwitz-Thompson estimator (Mandallaz, 2008), the result also turned out to be dependent on the underlying model of the investigated population.

Irrespective of the model choice, a core issue of these mapping approaches is to characterize the accuracy of the resulting maps. If map predictions are made on a continuous prediction scale, the map precision is commonly characterized by quality parameters of the applied prediction model, such as the cross-validated root-mean-squared error (RMSE) and the coefficient of determination (R^2). However, the derived map predictions are often visualized using constant class intervals in order to provide users, such as forest managers, with a better visual interpretation of the map. In this case, it could be misleading to still use the previously mentioned RMSE and R^2 in order to provide information about the accuracy of resulting timber volume classes. This is because these parameters only describe the overall model performance on a continuous prediction scale, but do not quantify the accuracy of individual timber volume ranges (classes). A more appropriate validation strategy would then be to adopt the concept of confusion matrices, which provides a differentiated accuracy assessment (user's and producer's accuracy) for each particular volume class, as well as the complete mapping system. Franco-Lopez et al. (2001), for example, used these metrics. However, their classification scheme of constant class intervals exhibited a strong gradient of degrading class accuracies towards higher volume classes (most likely due to saturation effects in the remote sensing data). Such a severe gradient in class accuracies, however, reveals the following problems: (1) it implies that the chosen classification scheme with constant class intervals is not accounting for the fact that the performance of the underlying model may not be constant across the entire volume range; and (2) it severely hampers the usability of the maps in forest practice due to the high uncertainty within higher timber volume classes.

The motivation of this study was to improve the usability of volume maps for forest management operations by avoiding classification schemes of this kind. We hypothesized that this can be achieved by optimizing the class intervals with respect to the accuracy potential of the underlying

prediction model. This implies using smaller class intervals in those volume ranges where the model ensures precise prediction performance and enlarging these intervals in ranges where the model performs worse. If the class boundaries are allocated according to this concept, it becomes possible to design classification schemes that provide highest possible accuracies for each class, while avoiding a severe gradient between class accuracies. This concept was investigated by implementing an optimization algorithm which can be applied to any type of prediction model that provides estimates on a continuous scale.Implicitly, the method also provides an additional option for evaluating the precision of prediction models.

We demonstrate the method in a case study in the canton of Grisons using a LiDAR-derived canopy height model and regional forest inventory data. The workflow included: (1) the production of a map of estimated standing timber volume for the entire study area; (2) the calculation of reliable accuracy metrics for this map; and (3) the application of the proposed optimization algorithm in order to identify the classification schemes which provide the highest possible accuracies. In this particular case, we decided to use a multiple linear regression model, because most auxiliary variables exhibit a pronounced linear relationship to the terrestrial inventory (Hill, 2013). Additionally, the number of available terrestrial observations in our study was small (n = 67) compared to similar studies, whereas it has been indicated that a good performance of k-NN requires larger datasets (Fehrmann et al., 2008; Magnussen et al., 2010b).

5.1.2 Background on Heuristic Search Methods

Heuristic Search Methods (HSMs) are often used when coping with combinatorial optimization problems (Pirlot, 1996) or, in general, problems whose structure cannot be satisfactorily represented and performed by means of classical optimization techniques, such as Linear Programming (Rayward-Smith et al., 1996). Basically, HSMs aim at improving an objective function by subsequent inspection and adoption of neighboring solutions. Inspection rules are often inspired by nature (Pirlot, 1996) and have the property of occasionally accepting inferior solutions for further inspection to avoid getting trapped within a local minimum or maximum. Simulated Annealing (SA) (Kirkpatrick et al., 1983) is such an HSM, borrowing its accepting rule for inferior solutions from metallurgy. It is based on the assumption that a configuration of atoms in a metal can move to configurations of higher energy (i.e., inferior solution) with a certain probability at a given temperature. Given that probability is a function of temperature and energy difference to the inferior solution, SA adopts a cooling scheme that aims at annealing the metal to the point of minimum configuration energy (i.e., objective function). As opposed to classical optimization methods, optimality of HSM-derived solutions cannot be proven. However, one can assume to find a solution close to the true but unknown optimum if the heuristic is appropriately parameterized.

5.2 Materials and Methods

5.2.1 Materials

Study area

The methods proposed in this article were applied to a study site located in the canton of Grisons, Eastern Switzerland (Fig. 1). The site extends in the north-south direction between Klosters and Davos and covers a total area of 2887.39 hectares. According to a forest mask (in raster format) of the study site derived by the use of the Swiss TLM3D (Swiss Topographic Landscape Model) data with the approval of the Swiss Federal Institute of Topography (for details, see Hill (2013)), the forest area of the study site comprises 1974.49 hectares (68.4%). The study site is located at an altitude between 900 and 2200 meters above sea level, and its relief is mainly characterized

by rough terrain and steep slopes. Classifying the forest area of the study site according to the scheme given by (Ott et al., 1997) revealed 49.7% of the forest area belonging to the high montane vegetation zone, 49.5% to the sub-alpine zone and 0.8% to the upper sub-alpine vegetation zone. Consequently, the forests within the study area were assumed primarily to consist of coniferous tree species, especially Norway spruce (Picea Abies).

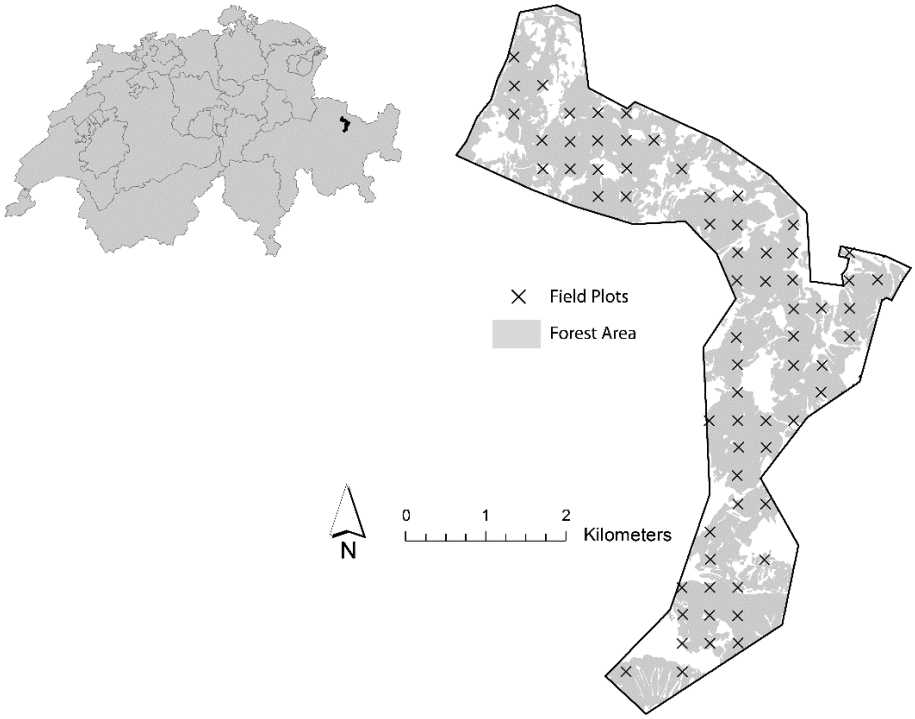


Figure 5.1: Study site, including the distribution of the 67 field plots (regional forest inventory) that are part of the forest area derived from the TLM3D (Topographic Landscape Model) data (with approval of swisstopo JA100120/JD100042).

Timber volume densities from field inventory

The methods applied in this study are based on timber volume densities obtained from terrestrial field surveys. For the study site, the timber volume densities were provided from field surveys of the latest regional forest inventory at the canton of Grisons and provided by the canton's forest service. Any uncertainty of the timber volume densities associated with their acquisition was ignored (e.g., potential measurement errors). The forest area of the study site comprised 67 terrestrial plots, which had been surveyed in the year 2007 (Fig. 5.1). Part of the survey was also the re-measurement of the plot centers with GPS technique. Unfortunately, no reliable information about the positional accuracy could be provided. The field surveys were conducted using circular sample plots with their center (i.e., the sample point) belonging to a regional permanent systematic inventory grid with a mesh size of 500 m. This regional sampling scheme thus constitutes a sub-grid of the nation-wide terrestrial inventory grid of the National Swiss Forest Inventory (NFI) with a mesh size of 1.4 km. Each sample plot consists of two concentric circles with a plot area of 200 and 500 m^2 around the sample center. Within the inner circle (radius of 7.98 m), all trees with a diameter at breast height (DBH) larger than 12 cm were selected, whereas in the second circle (radius of 12.62 m), all trees with a DBH larger than 36 cm were included in the sample. Boundary and slope adjustments were performed on plot level. The explicit survey methods and the evaluation of the regional inventory surveys were identical to those of the NFI and can be found

in detail in Brassel & Lischke (2001) and Keller (2011). To obtain the standing timber volume on plot level, the overbark timber volume of each sample tree was estimated by measuring its DBH and using it as the main predictor variable in the tariff models provided by the NFI. These tariff models are based on the general function proposed by Hoffmann (1982) and have been extended by further explanatory variables, such as the production region and additional tree and plot attributes (Brassel & Lischke, 2001). The standing timber volume for each plot was then estimated according to the Horwitz-Thompson estimator, which provides an unbiased estimation of the actual timber volume on plot level (Mandallaz, 2008). The volume distribution of the 67 terrestrial observed field plots is illustrated in Figure 5.2, and a brief statistical summary is given in Table 5.1.

Table 5.1: Summary statistics of the timber volume observed at the 67 terrestrial sample plots (given in m^2/ha).

Range	**Mean**	**Median**	**SD**	n
7.3 - 869.57	399.4	386.9	194.94	67

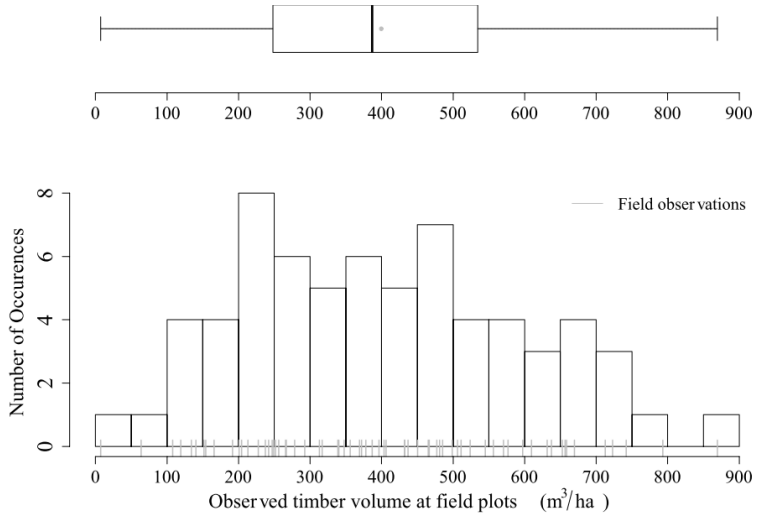


Figure 5.2: Terrestrial observed timber volume on plot level for all 67 sample plots of the study area (histogram bandwidth = 50 m^3/ha; origin = 0 m^3/ha).

LiDAR Data

A LiDAR dataset covering the entire study area was acquired with a Riegl LMS Q 560 laser scanning system in the period September 11-15, 2010. The LiDAR acquisition was conducted as a part of a campaign to gather data for the Swiss National Park. A digital terrain model (DTM) and a digital surface model (DSM) with a spatial resolution of 0.5 m were computed by the provider Toposys by application of their company-internal processing software TopPit. Gaps in the DTM due to the absence of last echoes had not yet been interpolated. The average flight height was 700 m above ground, and the average echo density was 27 points m^2. The provider specified the positional accuracy as $< \pm$ 0.50 m and the height accuracy as $< \pm$ 0.15 m. Further specifications of the LiDAR acquisition are summarized in Table 5.2.

Table 5.2: Summary statistics of the timber volume observed at the 67 terrestrial sample plots (given in m^2/ha).

Beam deflection	Rotating mirror
Pulse Repepition Frequency (kHz)	70
Average Flying Altitude (m above ground)	700
Max. scan angle (°)	± 15
Wavelength (nm)	1550
Beam divergence (mrad)	≤ 0.5
Average echo density (m−2)	27.4

5.2.2 Methods

The conceptual model in Figure 5.3 captures the general workflow of creating the timber volume map for the study area and the subsequent accuracy assessment. It consists of the following steps.

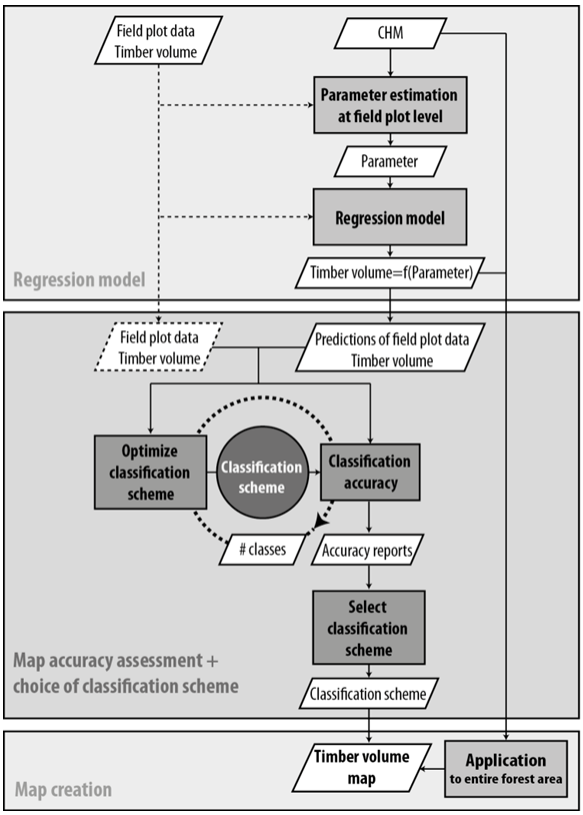


Figure 5.3: Conceptual model illustrating the workflow of producing the timber volume map and assessing its accuracy.

Step I: A regression model for predicting the standing timber volume in m^3/ha is formulated using the terrestrial observed timber volume of the field plots as the response variable and parameters extracted from the canopy height model (CHM) at the respective plot locations as predictor variables. The model then allows for predicting the timber volume at each point of the CHM that lies within the inventory area.

Step II: The model predictions (Step I) at the plot locations are then compared to the corresponding field data. These comparisons are used to assess accuracy metrics of the timber volume map. Regarding the classification accuracies, which can be estimated under class representations of the timber volume map (i.e., timber volume classes), this part also comprises the application of an optimization model aiming to find an optimal classification scheme, i.e., the choice of class boundaries and class widths for a given number of classes.

Step III: The regression model (Step I) is applied to the entire study area, and the result is then represented using a classification scheme that best satisfies the required accuracies for map users (Step II).

Step I: Computation of canopy height model

A canopy height model (CHM) with a spatial resolution of 0.5 m, completely covering the study site, was calculated by subtracting the LiDAR-acquired digital terrain model (DTM) from the digital surface model (DSM). As the DSM represents the elevation characteristics of the surface including vegetation and man-made structures, whereas the DTM describes only the elevation of the terrain, this operation equals removing the underlying terrain information from all features in the DSM. The height information of all objects in the CHM hence describes their estimated object height. Before calculating the CHM, an interpolation step for the DTM was crucial, since it exhibited a considerable amount of missing height information-most likely due to the absence of last pulse LiDAR returns over densely-covered forest areas. Missing height values in the DTM were predicted by applying an inverse distance weighting (IDW) interpolation algorithm [**38**]. Due to a locally varying number of adjacent missing raster values, the IDW algorithm was iteratively applied five times using varying neighborhood distances within which available height values were considered for interpolation. The use of small neighborhood distances aimed at providing a high amount of local precision where sufficient height values are available in the direct neighborhood of a missing value, whereas the use of large neighborhood distances was necessary for the interpolation of large gaps of missing values. Starting with a threshold distance of 2 neighbors, i.e., 1 m, and iteratively increasing the distance up to 20 neighbors, i.e., 10 m, all missing height values of the DTM were replaced by height predictions.

Step I: Regression model

To predict the standing timber volume (TV) at location x, a multiple regression model was formulated using the observed timber volume of the field plots as the response variable and certain CHM metrics as predictor variables. The CHM metrics were extracted at all 67 field plots within squares of constant size, which were centered at the respective plot centers. To ensure high spatial consistency between the circular field plots and the CHM metrics, the square extent was chosen in order to tangentially circumscribe a field plot. The side length of a square was 25 m, resulting in an area of 625 m2 compared to the field plot area of 500 m2. In the following, these squares are also referred to as the support of the estimates. By analyzing the distribution of the raster values of the CHM within each square, we calculated the following metrics of the LiDAR CHM at each plot location: the MEAN, the standard deviation (SD), the maximum value (MAX), as well as the 25%, 75% and 90% quantiles (Q25, Q75, Q90). The reason for the choice of these parameters was that they have often been used as predictors for estimating timber volume of forest stands (Holmgren, 2004; Næsset, 2002; Lefsky et al., 1999; Magnussen et al., 1999). The calculations of these variables were also adjusted for boundary effects by only considering those raster values within a square that are covered by the forest mask (Section 5.2.2). Table 5.3 provides the main statistics for the observed CHM metrics at the 67 terrestrial sample plots. The ordinary least square regression model containing the maximum number of predictor variables conclusively reads

as:

$$TV(x) = \beta_0 + \beta_1 MEAN(x) + \beta_2 SD(x) + \beta_3 MAX(x) + \beta_4 Q25(x) + \beta_4 Q75(x) + \beta_4 Q90(x) + \varepsilon(x) \quad (5.1)$$

A variable selection procedure was applied in order to restrict the regression model to the most meaningful variables and to avoid an overfitting effect of the model (Draper & Smith, 2014) due to the small number of observations. As most of the predictor variables were considered to be correlated to each other (collinearity), criterion-based selection procedures, by means of AIC (Akaike information criterion) (Akaike, 1992) and adjusted R-square criteria (Srivastava et al., 1995), as well as Mallow's Cp statistic (Mallows, 2000), were preferred to the testing-based selection procedures relying on p-values.

Table 5.3: Summary statistics of the CHM metrics calculated at the 67 terrestrial sample plots (in meters).

Metrics	Range	Mean	Median	SD
MEAN	2.26-26.03	12.07	11.31	5.89
SD	3.71-15.97	8.93	8.64	2.73
MAX	17.03-45.35	32.63	32.74	7.05
Q25	0-22.88	4.25	0.67	6.35
Q75	1.43-34.21	18.92	18.81	7.96
Q90	8.11-37.78	23.77	23.48	7.25

Step II: Assessment of map accuracy

Prediction performance

Assessing accuracy metrics of the timber volume map was realized by validating the timber volume predictions made by the regression model at the field plot locations using the corresponding observed timber volume of the field plots as reference data. Since the terrestrial sample was considered too small to be split into separate calibration and validation subsets, a leave-one-out cross-validation was performed to estimate the root-mean-squared error (RMSE) as a measure for the prediction performance of the timber volume map.

Classification accuracy

We applied the concept of representing the timber volume map by prior defined timber volume classes (i.e., assigning the prediction of each raster cell to a prior defined interval). This representation has commonly been used by various research studies producing maps of forest parameters, such as basal area or timber volume (Tonolli et al., 2011; Latifi et al., 2010; Clementel et al., 2012), in order to facilitate the interpretability, as well as the readability of a map. We propose to treat this map representation as a classification procedure and, consequently, adopted the concept of confusion matrices to provide a differentiated assessment for each particular class as well as the complete mapping system. Using the available field data as reference data, we estimated the following accuracy metrics for the resulting classified timber volume map (for details, see Congalton & Green (2008) and Richards & Richards (1999):

– The overall accuracy (OAA) is the proportion of correctly classified pixels of the entire map. The true overall accuracy of the map is unknown, since we only have references for the classified raster cells at a small subspace of the map. The OAA is therefore estimated by the ratio

of the total number of correctly classified pixels and the total number of reference/classified pixels. The 95% confidence interval for the OAA was calculated according to the binomial distribution.

- The producer's accuracy (PA) is a measure of the classification performance. It indicates the probability that if a ground observation belongs to a certain class, this class will be reflected in the map. The producer's accuracy can be estimated for each class by dividing the number of correctly classified pixels of a class by the total number of reference pixels in this class.
- The user's accuracy (UA) of each class is the most interesting information for a user of the map. It indicates the probability that if the map shows a certain class, this class will actually be validated by a terrestrial survey. The producer's accuracy is estimated by the number of correctly classified pixels in a class divided by the total number of classified data in this class.
- Cohen's kappa coefficient is a measure to assess to what degree the classification accuracy was realized by a chance agreement. The kappa coefficient ranges between -1 (accuracy was realized under pure chance agreement) and 1 (accuracy was reached by no chance agreement)

Class Selection Problem

One of the main benefits of classifying the timber volume map and assessing its classification accuracy is to provide information on the accuracies for individual timber volume ranges. However, classifying the model predictions into classes produces the problem of having to choose an appropriate classification scheme, i.e., choosing the class boundaries of the timber volume classes. A classic approach would be to use equally-sized classes with origin at zero and constant class width, but we consider three reasons not to do so: a constant class width (i) is likely to create classes for which no reference data are available, especially if the class width is chosen small; for such classes, PA and UA cannot be estimated, and the overall accuracy would give the user of the map an overoptimistic impression of the actual map precision; (ii) may separate a reference from its prediction (or vice versa) even if the two values were almost identical (i.e., their difference is very small and even negligible from a user's point of view); (iii) does not account for saturation effects in the remote sensing data, occasionally leading to a strong gradient of degrading class accuracies towards higher volume classes. To overcome these problems, we propose a locally-adaptive selection of class boundaries which satisfies the following rules (Class Selection Problem):

Rule I: Choose the class boundaries to ensure that each timber volume class at least contains a minimum number of reference data. This also aims at using a smaller class width where a sufficient number of references is available (thus providing locally higher detail), whereas the class width is increased for regions where references are rare.

Rule II: Avoid cases where a reference and its (closely located) corresponding prediction are separated by a class boundary. This implies not only taking into account the distribution of the reference data, but also considering the distribution of their corresponding predictions. A slight adaptation of the class width may thereby increase the classification accuracy.

Optimization model

The class selection problem (CSP) can be solved by repeatedly moving the boundaries and evaluating the resulting classification schemes until Rules I and II of the CSP are optimally satisfied. Since the number of combinations of alternative classification schemes can become big (e.g.,

2.3 million alternatives to distribute four boundaries along a range of 10-900 m3/ha specified by a lower and an upper boundary and discretized into 88 steps of 10 m3/ha), it risks becoming too computationally intensive to evaluate all alternatives. For this reason, we formulate the following multi-objective optimization model to automate the design of an (approximately) optimal classification scheme and solve it using Simulated Annealing (Section 5.1.2).

Its variables are specified as follows:

$$x_j : \text{class boundary value } j \ (m^3/ha) \tag{5.2}$$

$$\text{class } j = (x_j, x_{j+1}), \text{ j=1, ..., m+1} \tag{5.3}$$

$$n_{ref,j} : \text{number references / number plots in class } j \tag{5.4}$$

$$y_{ij} = \begin{cases} 1, \text{ if } V_{tp} \text{ and } V_{pred} \text{ at plot } i \text{ in same class } j \\ 0, \text{ if otherwise} \end{cases} \tag{5.5}$$

Where V_{tp} is the timber volume of terrestrial plot and V_{pred} the predicted timber volume. The decision variables x_j define the class boundary positions. A class j is then defined via the interval which is bound to x_j and x_{j+1} (Eq. 5.3). The number of pre-defined classes is m. The number of reference data for class j is given by $n_{ref,j}$ and the binary indicator variable y_{ij} (Eq. 5.5) captures whether V_{tp} and V_{pred} at plot i are assigned to the same class j (yes = 1, no = 0).

The optimization model can then be formulated as follows when n plots are subject to assignment:

$$MAX \sum_{j=1}^{m} \sum_{i=1}^{n} y_{ij} - w_1 \sum_{j=1}^{m} \left(x_{j+1} - x_j\right)^2 - w_2 \sum_{j=1}^{m} \left(\frac{n}{m} - n_{ref,j}\right)^2 \tag{5.6}$$

subject to:

$$x_{j+1} - x_j \geq t_w \tag{5.7a}$$

$$x_j \geq 0, y_{ij} \in \{0; 1\} \tag{5.7b}$$

The objective function (Equation 5.6) implements Rules I and II with three terms. The first term captures Rule II by maximizing for the number of cases where terrestrial observed volumes V_{tp} and corresponding estimated volumes V_{pred} are assigned to the same class j. In principle, the first term can be maximized by proposing a 'super class' covering almost the entire range. The second term is thus introduced to penalize the selection of a 'super class' by minimizing for local class widths, thereby also accounting for Rule I. This is realized by minimizing the square class width for each class j. Equation 5.7a limits the minimum class width to a threshold t_w (m^3/ha) and concurrently satisfies that x_{j+1} is chosen bigger than x_j. The third term captures Rule I and aims at equally distributing the reference data over all classes by minimizing the squared difference between the average number of plots per class and the actual number of reference data for each class j.

Both the second and the third terms are implemented as penalty terms for the first term that aims at maximizing the number of correctly classified plots. The corresponding weights w_1 and w_2 can be used to control the emphasis of both penalty terms and provides the user with the

possibility to prefer one of the three objectives within the optimization process.

In order to find an appropriate choice for the weight factors, several weights have to be evaluated to achieve satisfactory classification schemes. In our study, we decided to give the weights equal emphasis ($w_1 = w_2 = 2$). Class boundary selection was restricted to a range discretized into 10 m^3/ha units to reduce the problem size and simultaneously create useable class boundaries in the final classification scheme. A satisfying alternative was then identified by picking out the best overall solution from 100 runs of Simulated Annealing, where each run included the computation of 1000 alternatives.

Step III: Computation of the timber volume map

After model selection, the regression model (Section 5.2.2) was applied at any location x of the CHM, i.e., over the entire study area. We chose the design of the timber volume map in accordance with the setup of the regression model: the spatial resolution of the map was defined as the size of the support used for ground calibration, i.e., 25x25 m. We further orientated the map in such a way that the supports at the field plots actually became an almost exact subset of all raster cells of the timber volume map. Within each of the raster cells, the CHM metrics used in the regression model were calculated by the same procedure as described in Section 5.2.2 (i.e., by the same support and technique) and then used to predict the timber volume. The estimated value was then assigned to the entire raster cell, resulting in a timber volume map with a spatial resolution of 25x25 m. The entire procedure is again illustrated in Figure 5.4.

The design of the map as proposed here has two main advantages that allow for relying on the provided classification accuracy metrics: (i) each estimate of a raster cell is based on exactly the same support that was used to calibrate the prediction model; and (ii) as the reference data are an (almost) exact subset of the map raster cells, this allows for a valid accuracy assessment of the classified timber volume map (Congalton & Green, 2008). Another advantage of the map design is that it is in perfect agreement with the inventory design of the generalized two-phase regression estimator proposed by Mandallaz et al. (2013), where part of the auxiliary information is derived exhaustively over the entire inventory area. It thereby provides a link between the derived timber volume map and sample designs of classical forest inventory.

Once the timber volume map was calculated for the entire study area, each raster cell of the map, still carrying estimates on a continuous scale, was classified into timber volume classes according to a classification scheme (Step II, Section 5.2.2).

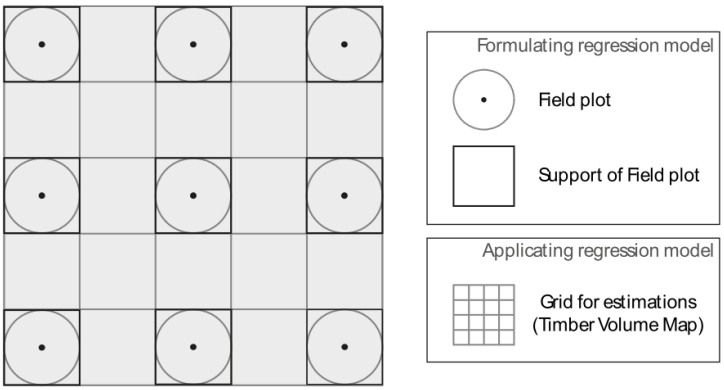


Figure 5.4: Schematic design of the timber volume map.

5.3 Results

5.3.1 Regression model

Performing individual simple linear regressions revealed that all predictor variables derived from the CHM were correlated to the field-obtained timber volume on plot level. Here, the highest coefficient of determination (R^2) was achieved by the variable mean ($R^2 = 0.5$). A forward, backward, as well as bidirectional selection procedure according to minimize the AIC all revealed mean, standard deviation, maximum value and the 75% quantile as the model of choice (adjusted $R^2 = 0.62$, AIC = 646.4). These predictor variables also revealed a significant influence on the field-obtained timber volume (individual parameter t-test, 5% significance level). The model was also suggested by Mallow's Cp and the adjusted R2-criterion selection procedure. Summary statistics of the model are presented in Table 5.4. The estimated timber volume $\hat{TV}(x)$ at location x was consequently calculated according to the following regression model formula:

$$TV(x) = 228.83 + 62.76 * MEAN(x) + 76.65 * SD(x) + 19.59 * MAX(x) + 33.44 * Q75(x) \quad (5.8)$$

Figure 5.5 shows the predicted timber volume on plot level plotted against the observed timber volume of the field plots. The leave-one-out cross-validated $RMSE_{cv}$ of the regression model was 123.79 m^3/ha and, thus, only slightly larger than the RMSE without cross-validation (115.5 m^3/ha).

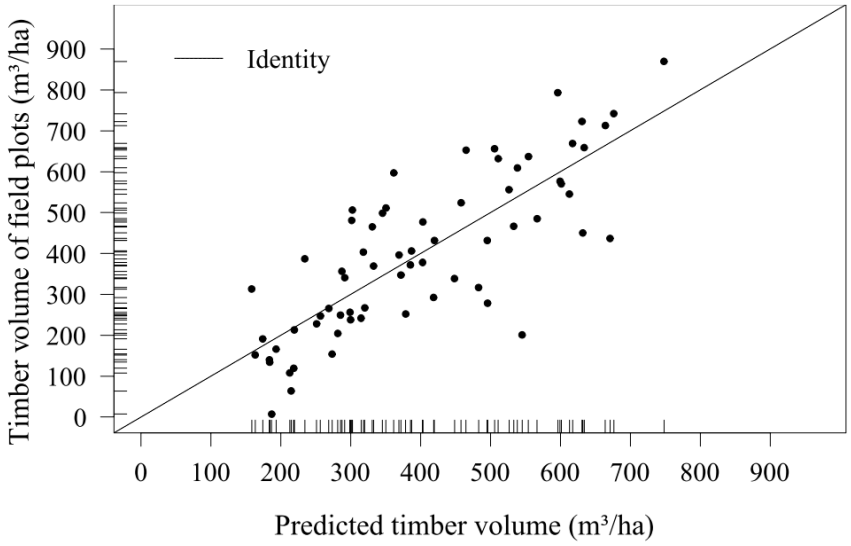


Figure 5.5: Model-predicted timber volume on plot level against the observed timber volume of the 67 field plots; the distribution of the predictions and observations are also indicated on the respective axis.

Table 5.4: Summary statistics of the regression model.

Predictors	AIC	R^2	Adjusted R^2	Mallow's C_p	$RMSE_{cv}$	Model Range
MEAN*, SD*, MAX*, Q75*	646.6	0.64	0.62	2.56	123.79	0 - 900 m^3/ha

5.3.2 Assessment of map classification accuracy

The classification accuracies described in Section 5.2.2 were estimated for nine possible constant class widths (100, 125, 150, 175, 200, 250, 275 and 300 m^3/ha), which were assumed to be of interest for representing the timber volume map. We applied the optimization model (Section 2.2.3) for each of those constant class widths in order to find a better classification scheme using the same number of corresponding classes, but locally adaptive class widths. The accuracies were then also estimated for the best-found classification schemes. Figure 5.6 provides a graphical summary of the results, giving the overall accuracies with their 95% confidence intervals, as well as the kappa coefficients. In all cases but one, the optimized locally-adapted class widths led to a higher overall classification accuracy compared to the corresponding constant class width. However, the 95% confidence intervals revealed that the overall accuracies using the constant class widths were, in all cases, not significantly different from their corresponding optimized alternatives (i.e., the confidence intervals are overlapping). In other words, even if the differences between the two accuracies seemed in several cases to be quite distinct, the true (but unknown) overall accuracies of the corresponding maps could actually be identical. However, with respect to the overlapping areas of the confidence intervals, the probability of acquiring a better overall accuracy by using the locally-adapted class widths was highest for smaller constant class widths (i.e., for larger numbers of classes). Interestingly, at the same time, a constant class width of 225 m^3/ha (four classes) was also the best-found solution for locally-adapted class widths. In all cases but one, the kappa coefficients were higher under the application of the optimized class widths, even more distinct for larger numbers of classes (i.e., smaller constant class widths).

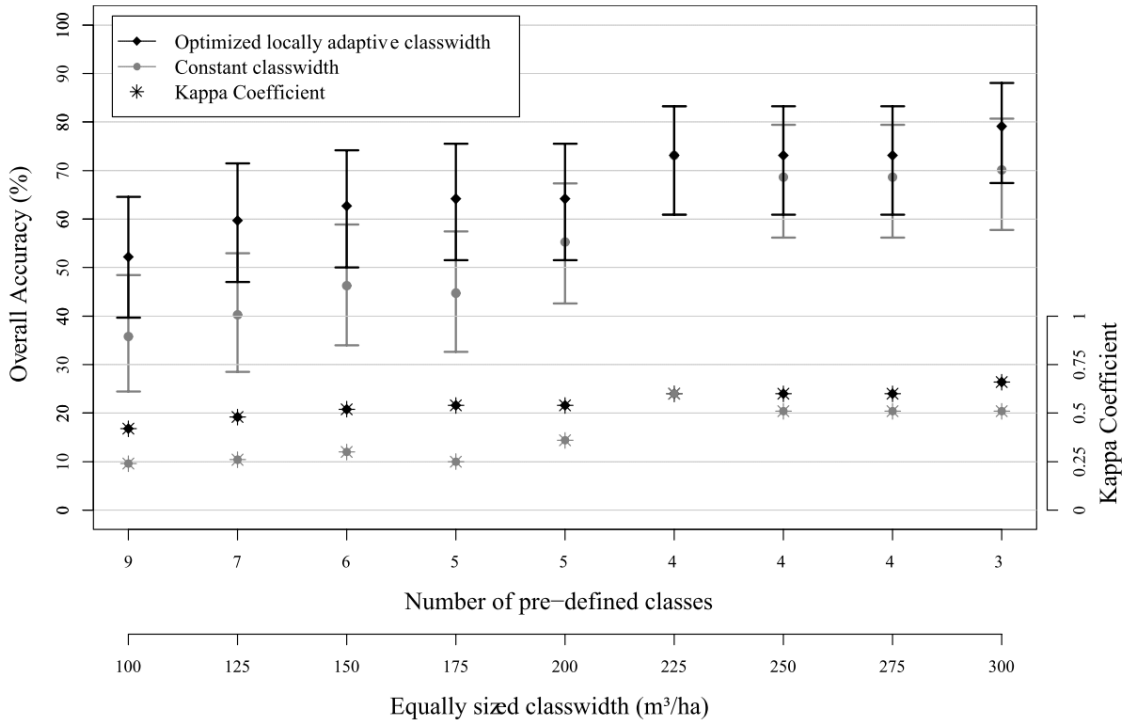


Figure 5.6: Overall accuracies with 95% confidence intervals and corresponding kappa coefficients for a pre-defined number of classes, calculated for constant and locally-adapted class widths.

We further compared the properties of the confusion matrices obtained by the use of constant and corresponding locally-adapted class widths, especially with respect to the estimated producer's

and user's accuracies. The producer's, as well as the user's accuracies under the optimized classification schemes were consistently higher than those of the corresponding constant class width approach. A phenomenon that, in several cases, occurred for the constant class width approach was that some of the classes did not include at least one prediction. This phenomenon especially concerned the classes comprising the upper scale region of the field data where the number of reference data was small. Consequently, the producer's accuracy for these classes was zero, whereas the user's accuracy is not even defined for this case (Section 5.2.2). The reason for this undesired property of the confusion matrix is likely to be that the fit of the regression model for the upper timber volume range was rather poor due to a limited number of field data or/and a saturation effect in the CHM (i.e., beyond a certain canopy height, different timber volumes cannot longer be discriminated by the regression model). The constant class width approach did, however, not account for this effect, whereas the optimization model proposed larger class widths for the upper-range classes, including a larger number of field data and predictions. An illustration of the findings described can be found in Table 5.5 and Table 5.6, using a constant class width of 200 m^3/ha (five locally-adapted classes, respectively). The described locally-adaptive classification scheme provided a satisfactory trade-off between the number of classes and classification accuracies and was therefore exemplary used for the visualization of the timber volume map (Fig. 5.7).

Table 5.5: Constant class width of 200 m^3/ha (five timber volume classes); OAA (± 95% confidence interval) (%): 55.22 (42.58, 67.4); kappa: 0.36; square sum of class width: 200'000.

Classes	Class Width	Producer's Accuracy	User's Accuracy	No. of References
(0,200)	200	60	85.71	10
(200, 400)	200	72	60	25
(400, 600)	200	40	40	20
(600, 800)	200	45.45	50	11
(800, 1000)	200	0	-	1

Table 5.6: Optimized class width for five timber volume classes; OAA (± 95% confidence interval) (%): 64.18 (51.53, 75.53); kappa: 0.54; square sum of class width: 176'600. Class widths smaller than 200 m^3/ha are indicated by *.

Classes	Class Width	Producer's Accuracy	User's Accuracy	No. of References
(0,220)	220	76.92	90.91	13
(220, 330)	110*	61.54	50	13
(330, 450)	120*	57.14	53.33	14
(450, 660)	210	66.67	66.67	21
(660, 900)	240	50	75	6

Additionally, we investigated how often the locally-adaptive class widths were capable of satisfying Rules I and II of the optimization model (Section 5.2.2) more successfully than their constant class width counterparts. This was done by comparing the respective sum of squared class widths (second term of Equation 5.6), as well as the respective squared difference between the average number of plots per class and the actual number of reference data for each class (third term of Equation 5.6). It turned out that, particularly for the larger constant class widths (175, 200, 225, 250, 275 and 300 m^3/ha), the locally-adapted approach worked very successfully for both

objectives: in six out of nine cases, each objective was better solved by the locally-adaptive approach, and in five out of nine cases, the optimization approach even succeeded in both objectives simultaneously.

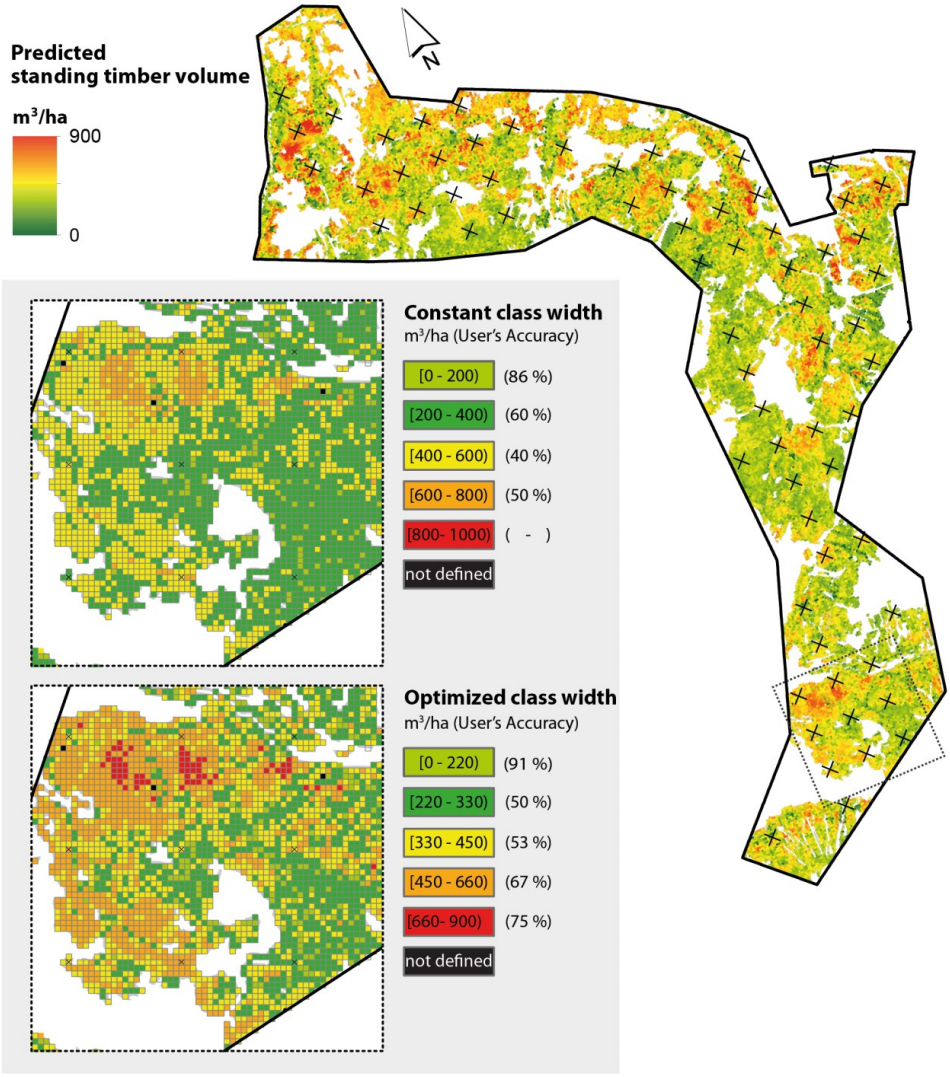


Figure 5.7: Volume map with model-predictions on a continuous scale for the entire study area covering 2000 hectares of forest with a spatial resolution of 25 meters; the subareas show the classified volume map using a constant class width of 200 m^3/ha (upper) and locally-optimized class widths for five classes (lower).

5.3.3 Calculation of the timber volume map

Figure 5.7 shows the timber volume map for the entire study area estimated by the application of the regression model (Equation 5.8). An undesired property of the map was the occasional appearance of negative predictions, which are most likely caused by the negative signs of two of the regression coefficients (Equation 5.8). Additionally, the upper range of the predictions of the entire map (1002 m^3/ha) turned out to exceed the upper valid model range (900 m^3/ha). The values of these raster cells were consequently changed into 'Not Available' (NA). The map was additionally classified according to a constant class width of 200 m^3/ha, as well as to the

corresponding optimized classification scheme presented in Section 5.3.2. As the classification schemes were also based on the valid model range, cells whose predictions did not cover any of the pre-defined classes were consequently marked as 'not defined' in the classified map. The number of these cases was, however, small compared to the total number of predicted cell values of the entire volume map (170 of 31,622 raster cells, i.e., 0.5%).

5.4 Discussion

5.4.1 Assessment of map accuracy

A core issue of this study was to use a class representation of continuous map predictions to estimate accuracy metrics of individual timber volume ranges (user's and producers accuracies). An evaluation for a large number of classification schemes revealed that the class accuracies can vary according to (i) which class width is used, (ii) how the class boundaries are chosen and (iii) a combination of both aspects. This variation in accuracy is not indicated by metrics such as the cross-validated $RMSE_{cv}$ of the prediction model, which can, however, be used to describe the overall prediction performance. When investigating the map for explicit timber volume ranges, purely relying on $RMSE_{cv}$ or R^2 metrics could, however, lead to over- or underestimation of the provided accuracy. Building the map according to classes and, consequently, estimating their individual accuracies is therefore considered a valuable step towards coping with the uncertainty of such maps.

5.4.2 Class Selection by optimization model

Regarding the choice of appropriate classes to represent the timber volume map, the proposed optimization model turned out to be of high value. While the overall accuracies between the optimized class width and constant class width approach did not differ significantly on a 5% significance level, the true benefit of the optimized class width approach were the properties of the confusion matrices thus obtained. The proposed classes revealed a more uniform distribution of references among the classes, ensuring that the estimated producer's and user's accuracies of each class were estimated by the highest possible number of references. The class width was thus successfully chosen smaller where a sufficient number of reference data were available, leading to a higher degree of detail and, with respect to an optimized adaptation of the class boundaries, without a loss of accuracy in those classes. In many of the evaluated classification schemes, the adaptation of the class boundaries led to higher producer's and user's accuracies compared to those obtained under constant class widths.

The optimization model has yet provided a method which allows for (i) finding an optimal classification scheme for a given number of classes and (ii) finding a number of classes, such that the class accuracies are acceptable for a user. If forest managers are interested in identifying stands with timber volumes greater than a selected minimum, perhaps for possible thinning treatments, one of the class boundaries can be fixed to that minimum volume.

While the maximization for the number of predictions and references included in the same class (Rule II, Section 5.2.2) gives a more realistic representation of the underlying regression model, one could argue that optimizing for this objective can lead to an 'overfitting' of the confusion matrix and, thus, to an increased generalization error of the accuracy metrics. This could not be investigated, since an appropriate method of bootstrapping and cross-validating for this kind of classification process has yet to be implemented. However, one could also slightly change the optimization model in order to optimize only for Rule I (i.e., to locally achieve smaller class widths and an even distribution of the reference data over the classes).

5.4.3 Regression model

In the present study, the mean of the canopy height turned out to be the best predictor, providing an R^2 of already 0.5. The fit was further improved to an adjusted R^2 of 0.62 by extending the model using the predictor variables standard deviation, maximum height and the 75% quantile. This, however, came at the price of losing interpretability of the model (negative signs of some regression coefficients), as well as the occurrence of negative predictions in the timber volume map. The latter reassuringly happened only for 141 of 31,622 pixels (0.4%). To avoid negative predictions, we also computed the regression model using the log-response of the terrestrial timber volume and then calculated the map by back-transformation of the log-predictions into the original scale (Beauchamp & Olson, 1973). The log-response model, however, tended to predict unrealistically large timber volume values (up to 2400 m^3/ha), exceeding the valid model range in 267 cases (0.16%) and, thus, not providing an improvement in prediction performance. The use of a log-response or non-linear regression model is also methodically sound in the design-based framework, with the restriction that only the variance under the external model assumption (Mandallaz, 2008) can be calculated. This is because the design-based variance using the g-weight technique (Mandallaz et al., 2013; Mandallaz, 2013a) is only available within linear models on the original scale. Within the model-dependent framework, other alternatives to ensure non-negative predictions can be the use of nonlinear logistic regression models (McRoberts et al., 2013) or k-NN approaches. One could also consider the alternative use of external models (e.g., taken from literature) based on data that provide, for example, a larger model range or better coverage of a required timber volume range. It would be statistically sound to estimate the classification accuracies by applying an external model and consequently validating its classified predictions by available reference data of the study site. Using such models in the design-based framework of inventory would, however, again require calculating the variance under the external model approach.

5.4.4 Availability and quality of reference data

In the present study, we used existing forest inventory data as reference and validation data, instead of acquiring these data in a special campaign. While this is in general both time and cost saving, the number of available reference data for calibrating the regression model and for validating the timber volume map was considerably limited, finally resulting in large confidence intervals of the accuracy metrics. However, in order to produce maps for operational forest management, probably over considerably large areas, one will realistically always depend on the use of existing inventory data due to limited financial resources. It should also be mentioned that the proposed methods assume the nominal coordinates of all field plots to be equal to the actual, true location of acquisition. This is in practice almost never the case, due to potential location errors, which can still be in the range of up to 10 meters, even under the use of GPS technique (Mauro et al., 2010; Steinmann et al., 2013). Although the remote sensing data can reveal positioning errors as well, they can be expected to be considerably smaller than the largest possible GPS location errors occasionally caused by dense vegetation and shielding of GPS reception on the ground. However, severe location errors should have appeared as outliers (or even leverage points) in the regression model. Since, in our case, neither the cross-validation nor an inspection of the regression model revealed any such outliers, the severity of location errors was assumed to be small. However, the R^2 of the regression model is expected to be higher if the exact positions of the plot centers are known (Fuller, 2009), and the same can be assumed for the classification accuracies under the restriction that the locally-adaptive class width must be recomputed if the model is changed.

5.5 Conclusion

The methods proposed in this study provide better knowledge about the actual accuracy of a timber volume map. The classification of predictions into classes and the consequent computation of classification accuracy metrics improved the knowledge about the map accuracy, especially the accuracies of timber volume intervals. Additionally, considering the distribution of the reference data, as well as their corresponding predictions turned out to be key factors in choosing an appropriate classification scheme: the application of optimized locally-adaptive class widths ensured good statistical properties of the confusion matrices and also led to higher class accuracies and kappa coefficients compared to the approach of using constant class widths. The proposed methods, including the optimization of classification schemes, are thus not restricted to maps based on linear regression models, but can be applied to a larger class of prediction methods (e.g., k-NN estimation). Finally, the proposed design of the timber volume map was considered to be crucial for the reliability of the estimated accuracy metrics. Another advancement of the entire map set up is that the continuous map predictions can be directly used in the framework of design-based inventory methods. For example, several raster cells of the volume map could be merged, e.g., to 0.5 or one hectare, and the resulting larger cell could be used for true small area (synthetic) estimation (Mandallaz, 2013a). We expect this approach to also solve the problem of negative predictions. The design-based confidence intervals thus obtained for these estimates could then serve as a measure for local accuracy. The accuracy assessment of timber volume classes could also improve an automatic delineation of possible harvesting units by additionally considering the probability of each raster cell belonging to a certain class.

Acknowledgements

We express our thanks to the reviewers for their comments and suggestions; to Professor Hans Rudolf Heinimann (Chair of Land Use Engineering, Swiss Federal Institute of Technology ETH Zurich) for his support; to the Forest Service of the canton of Grisons (Switzerland) for providing the forest inventory data; and to the Remote Sensing Laboratories (University of Zurich) for information on the LiDAR acquisition.

Chapter 6

Synthesis

6.1 Main findings

The investigated design-based small area regression estimators show great potential for the German NFI data to be additionally used for estimation of forest attributes on forest district levels. In the case study presented in this thesis, the variance for the estimation of timber volume on two district levels was reduced by 43% and 25% on average by the unbiased regression estimators compared to the one-phase estimator exclusively using the terrestrial inventory data. The results thus strongly support the findings of similar studies from other countries that the precision of estimates based on national forest inventory data on regional scales can be considerably improved if the terrestrial inventory data are combined with remote sensing data in the framework of small area estimation techniques. It was also empirically confirmed that the synthetic estimator who neglects the prediction performance of the regression model produces considerably smaller variances and confidence intervals compared to the estimates of the unbiased estimators. In addition to their potential design-bias, this strongly suggests that synthetic estimates are over-optimistic and should be treated with caution. The application of design-unbiased estimators provides the advantage of not having to rely on the validity of the prediction model and should thus be preferred to synthetic estimation whenever it is possible. Fitting the regression model used in the estimators internally seems obvious if the terrestrial inventory phase provides sufficient data for modeling. This is most likely the case if the terrestrial inventory phase is a national forest inventory covering large areas and diverse forest structures. In case the regression model is fitted internally, providing the g-weight variance is preferable to the external variance since it accounts for the dependency of the regression coefficients on the realized sample.

Whereas calculating the g-weight variance restricts the internal model of the estimators to OLS regression, it was demonstrated that OLS modelling possibilities provide sufficient flexibility to reflect real world dependencies between predictor and response variables and also to mitigate residual inflations caused by quality restrictions in the auxiliary data. If the quality of auxiliary data varies between known strata, including these strata as categorical predictors in the regression model can be an effective means to improve the model precision. Among the auxiliary variables for timber volume prediction, the main tree species of a sample plot revealed to be a powerful predictor when combined with vegetation height information that can be used to boost prediction performance. If categorical data such as tree species information is derived from classification of remote-sensing data and thus prone to classification errors, the calibration technique suggested in this thesis provides a simple and effective method to remove the bias in the regression coefficients caused by misclassifications and thereby increase the model precision. Using categorical variables including their interactions with other categorical or continuous predictors in the regression model thereby extends double-sampling for regression to post-stratification, which is a well-known and effective means to reduce the variance of estimates.

The investigated mapping scenario constitutes a special case of small area estimation where the target area, i.e. a map cell, corresponds to the extent of a sample plot and thus exclusively comprises one model prediction. In the case of prediction maps for continuous response variables, the suggested approach of calculating the user's accuracy for prediction value intervals according to accuracy assessment techniques constitutes an alternative way of providing a confidence interval for each map cell. A difference to classical regression prediction intervals is that the confidence level for the intervals can vary between map cells instead of being fixed. This seems more appropriate for practical usage as it provides the map user with precisions associated to self-defined map value intervals opposed to providing him with map value intervals based on a fixed confidence level. The proposed optimization method can additionally be used to automatically identify prediction intervals that provide highest possible confidence levels while ensuring statistically sound properties of the underlying error matrix.

6.2 Limitations and criticisms

The most obvious criticism on design-based estimators is that they are in practice applied to systematic inventory grids although they rely on the assumption of independently and uniformly distributed sample locations. Accounting for the systematic grid design would thus imply the application of model-dependent estimators. Among those, Mandallaz (1993) described geostatistic ordinary and double kriging estimators as the most accurate method that provide the advantage of modeling the spatial covariance function. Intensive studies reassuringly revealed that while the design-based estimators tend to overestimate the variance under systematic grids, the difference to the variance under geostatistic kriging is marginal for domains whose spatial extent considerably exceed the range of the spatial correlation. The minimum spatial extent of such domains will more quickly be exceeded in the case of double-sampling, as the spatial correlation range of the model residuals, which in turn contribute most to the double-sampling variance, is usually much smaller than the range of the local density used for one-phase estimations (i.e. ordinary kriging). The overestimation of the variance by design-based estimators under systematic sampling will thus particularly occur for small area estimation, but reassuringly result in conservative confidence intervals.

A general criticism on cluster sampling as applied in the German NFI is an increase of the design-based variance compared to simple random sampling, which is however accepted to the benefit of reduced transport costs. However, the inflation in variance for change estimation can be expected to be less than for state estimation. For small area estimation, cluster sampling has also the beneficial effect of increasing the terrestrial sample size within the small area domain as it implies an increased probability for at least one plot of a cluster to be included in the small area. However, it has to be emphasized that only part of a cluster being included in a small area domain constitutes an assumption violation of the extended pseudo-synthetic estimator. Whereas this thesis provided empirical evidence that this results in slightly over-optimistic variances, it has also been indicated that the impact is negligible for small area sample size larger than 6. Recommending a minimum small area sample size of 6 for design-unbiased estimation has also been supported by simulation studies in Mandallaz et al. (2013) for simple random sampling which indicated that the nominal coverage rates are ensured for small area sample sizes equal and larger than 6. Re-evaluating the simulation example confirmed the same results for cluster sampling.

Another criticism concerns the practical implementation of double-sampling in the special case of small area estimation. A theoretical necessity that was neglected in this thesis is that boundary adjustment on the sample plot and support level should be performed also at the boundaries of a small area unit for both the field and auxiliary data respectively. For the field data, neglecting boundary adjustment can lead to including trees from outside the small area domain in the sampling frame which thus might cause an overestimation of the plot local density. On the part of the auxiliary data, including information from outside the small area domain can weaken the relationship between the derived predictor variable value and the boundary corrected plot local density and thus decrease the model precision (Mandallaz et al., 2013). On the contrary, performing boundary adjustment also at the small area level would imply a considerable increase in effort for data storage and handling, particularly for large scale applications such as presented in this thesis. It could be conjectured that the impacts on estimates are marginal because the occurrence of boundary plots within a domain is usually rare. Since the implications have yet not been investigated, this could be a topic for future work.

Regarding the modeling as being a substantial part of regression estimation procedures, time-lags between the date of the terrestrial survey and the acquisition of the auxiliary data can constitute a severe limitation. While for state estimation the implied loss in model precision can be partially accounted for by post-stratifying to acquisition dates, it has been shown that such time-lags can hamper the application of double sampling for the estimation of change (Massey,

2015). Another limiting factor in the modeling framework is angle-count sampling, as it implies the problem that the extraction of the auxiliary data cannot be matched to an exact spatial extent in which the terrestrial data was gathered. Whereas it was confirmed that this can affect model performance, it was revealed that the impacts are not as critical for auxiliary data of medium to low spatial resolution. It could however be conjectured that impacts on model precision are more pronounced when high resolution data are used to make predictions on sample plot level, for example timber volume prediction based on individual tree detection.

A further criticism on the presented accuracy assessment for prediction maps is that even the user's accuracies that serve as confidence levels for the prediction value intervals are themselves associated with uncertainties caused by the sampling process of the reference data. Whereas this has also rarely been addressed in similar studies, quantifying these uncertainties is feasible and could be a potential area of future work. In addition, the suggested optimization algorithm is limited by the fact that it constitutes a heuristic approach that rather finds a solution close to the optimum than finding the exact optimal solution. Another implication of the heuristic approach is that the solution for the same input data can slightly vary due to the randomization part of the algorithm and the various possibilities of parameterization. Concerning the latter, it is most important to ensure an appropriate cooling scheme of the simulated annealing.

6.3 Conclusion and implications for future work

The results of this thesis strongly confirmed that combining terrestrially gathered forest inventory data with auxiliary information constitutes a very effective means to increase the efficiency of inventories and to expand the range of terrestrial data application. Mapping provides predictions in very high spatial detail and can thus support decisions on very small scales, but on the other side imposes high requirements on the reliability of the predictions as they are purely model-dependent. The findings of this thesis provided evidence that the error associated to specific value intervals are often considerably larger than expected from overall precision accuracies. Conclusively, reliable and detailed information on the map accuracy should always be provided to the practical user. Design-based small area estimators can not meet the spatial detail of model-dependent approaches since they require a minimum number of terrestrial data per small area domain. However, they provide the huge advantage of being robust to model misspecification and are thus very attractive to be used operationally. The recent interest on these family of estimators is also supported by the acception of the R-package software that was developed and used in this thesis (4000 downloads since release in June 2016). Both approaches, model-dependent mapping and design-based small area estimation, suggest that one should in any case not discontinue the conduction of terrestrial inventories as they are crucial in order to quantify the reliability of estimates and predictions.

This thesis also supports the recommendation to increasingly use already widely available auxiliary data in the service of forest inventories. It should be emphasized that beside high resolution remote sensing data, these auxiliary data can also be provided by maps based on expert judgment (e.g. reflecting soil properties and bioclimatic growing conditions). For change estimation, it could also be beneficial to include information indicating tree species specific growth impacts of intermediate dry events that are likely to occur in increasing frequencies with ongoing climate change. With regards to more frequently updated and alternative remote sensing data, it can be conjectured that the problem of time-lags between terrestrial survey and the auxiliary information will not be an issue in the future. This will also improve the applicability of double-sampling and mapping techniques for the estimation of change. Two challenges of future applications are the reduction of positional inaccuracies of the terrestrial sample locations as well as the synchronization between auxiliary data and terrestrial data derived by angle-count sampling. Concerning the first, various matching approaches to reconstruct the exact plot locations have been suggested. However,

the effects of such procedures on design-based estimates should first be investigated via simulation studies before applying them in practice. With regards to the problem associated to angle-count sampling, the alternative use of 2-3 concentric circles has been proven to be very close to optimal PPS sampling for the Swiss NFI (Mandallaz, 2008) while providing an unambiguous extent in which the auxiliary data can be derived. This should be considered in case angle-count sampling turns out to be a major limiting factor for future NFIs that are increasingly using double-sampling.

Finally, geostatistical double-kriging constitutes an estimation technique which could lead to a further increase in estimation precision but has scarcely been investigated empirically. First evaluations of the spatial variograms for the German NFI data that were used in this thesis however revealed that the correlation range is already reached for the minimum distance within a cluster. Thus, the gain in efficiency can expected to be small.

Bibliography

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics* (pp. 610–624). Springer.

Akaike, H. (2011). *Akaike's Information Criterion*, (pp. 25–25). Springer Berlin Heidelberg: Berlin, Heidelberg.

Baffetta, F., Corona, P., & Fattorini, L. (2010). Design–based diagnostics for k-nn estimators of forest resources this article is one of a selection of papers from extending forest inventory and monitoring over space and time. *Canadian Journal of Forest Research*, 41(1), 59–72.

Baffetta, F., Fattorini, L., Franceschi, S., & Corona, P. (2009). Design–based approach to k–nearest neighbours technique for coupling field and remotely sensed data in forest surveys. *Remote Sensing of Environment*, 113(3), 463–475.

Beauchamp, J. J. & Olson, J. S. (1973). Corrections for bias in regression estimates after logarithmic transformation. *Ecology*, 54(6), 1403–1407.

Beaudoin, A., Bernier, P., Guindon, L., Villemaire, P., Guo, X., Stinson, G., Bergeron, T., Magnussen, S., & Hall, R. (2014). Mapping attributes of canada's forests at moderate resolution through k nn and modis imagery. *Canadian Journal of Forest Research*, 44(5), 521–532.

Bitterlich, W. (1984). *The relascope idea. Relative measurements in forestry.* Commonwealth Agricultural Bureaux.

Böckmann, T., Saborowski, J., Dahm, S., Nagel, J., & Spellmann, H. (1998). A new conception for forest inventory in lower saxony. *Forst und Holz (Germany)*.

Bohlin, J., Bohlin, I., Jonzén, J., & Nilsson, M. (2017). Mapping forest attributes using data from stereophotogrammetry of aerial images and field data from the national forest inventory. *SILVA FENNICA*, 51(2).

Bont, L. & Heinimann, H. R. (2012). Optimum geometric layout of a single cable road. *European journal of forest research*, 131(5), 1439–1448.

Bont, L. G., Heinimann, H. R., & Church, R. L. (2015). Concurrent optimization of harvesting and road network layouts under steep terrain. *Annals of Operations Research*, 232(1), 41–64.

Brassel, P. & Lischke, H. (2001). *Swiss National Forest Inventory: Methods and Models of the Second Assessment. WSL Swiss Federal Research Institute, CH-8903 Birmensdorf.* Technical report, ISBN 3-905620-99-5. URL http://www. lfi. ch/publikationen/publ/methods/methods. pdf.

Breidenbach, J. (2015). ***JoSAE****: Functions for some Unit-Level Small Area Estimators and their Variances. R* package version 0.2.3.

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.

Breidenbach, J., Kublin, E., McGaughey, R., Andersen, H.-E., & Reutebuch, S. E. (2008). Mixed–effects models for estimating stand volume by means of small footprint airborne laser scanner data. *Photogrammetric Journal of Finland*, 21(1), 4–15.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Brosofske, K. D., Froese, R. E., Falkowski, M. J., & Banskota, A. (2014). A review of methods for mapping and prediction of inventory attributes for operational forest management. *Forest Science*, 60(4), 733–756.

Bundesministerium für Ernährung, L. u. V. (2011). Aufnahmeanweisung für die dritte Bundeswaldinventur BWI3 (2011 - 2012).

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. CRC press.

Chirici, G., Corona, P., Marchetti, M., Mastronardi, A., Maselli, F., Bottai, L., & Travaglini, D. (2012). K–nn forest: a software for the non–parametric prediction and mapping of environmental variables by the k–nearest neighbors algorithm. *European Journal of Remote Sensing*, 45(1), 433–442.

Clementel, F., Colle, G., Farruggia, C., Floris, A., Scrinzi, G., & Torresan, C. (2012). Estimating forest timber volume by means of 'low–cost' lidar data. *Italian Journal of Remote Sensing/Rivista Italiana di Telerilevamento*, 44(1).

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

Congalton, R. G. & Green, K. (2008). *Assessing the accuracy of remotely sensed data: principles and practices*. CRC press.

Cullmann, A. D. (2016). ***maSAE**: Mandallaz' Model-Assisted Small Area Estimators*. *R* package version 0.1-5.

Deo, R. K., Froese, R. E., Falkowski, M. J., & Hudak, A. T. (2016). Optimizing variable radius plot size and lidar resolution to model standing volume in conifer forests. *Canadian Journal of Remote Sensing*, 42(5), 428–442.

Dowle, M. & Srinivasan, A. (2017). ***data.table**: Extension of 'data.frame'*. R package version 1.10.4-1.

Draper, N. R. & Smith, H. (2014). *Applied regression analysis*. John Wiley & Sons.

ESA (2017). Sentinel–2 earth observation mission. Accessed: 2017-03-29.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: models, methods and applications*. Springer Science & Business Media.

Fehrmann, L., Lehtonen, A., Kleinn, C., & Tomppo, E. (2008). Comparison of linear and mixed–effect regression models and a k–nearest neighbour approach for estimation of single–tree biomass. *Canadian Journal of Forest Research*, 38(1), 1–9.

Franco-Lopez, H., Ek, A. R., & Bauer, M. E. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k–nearest neighbors method. *Remote sensing of environment*, 77(3), 251–274.

Fuller, W. A. (2009). *Measurement error models*, volume 305. John Wiley & Sons.

Gauer, J. & Aldinger, E. (2005). Waldökologische Naturräume Deutschlands–Wuchsgebiete. *Mitteilungen des Vereins für Forstliche Standortskunde und Forstpflanzenzüchtung*, 43, 281–288.

Goerndt, M. E., Monleon, V. J., & Temesgen, H. (2011). A comparison of small–area estimation techniques to estimate selected stand attributes using lidar–derived auxiliary variables. *Canadian journal of forest research*, 41(6), 1189–1201.

Goodspeed, A. (1934). A modified plot method of timber cruising applicable in southern new England. *Journal of Forestry*, 32(1), 43–46.

Grafström, A., Schnell, S., Saarela, S., Hubbell, S., & Condit, R. (2017). The continuous population approach to forest inventories and use of information in the design. *Environmetrics*.

Gregoire, T. G. & Valentine, H. T. (2007). *Sampling strategies for natural resources and the environment*. CRC Press.

Grosenbaugh, L. R. (1958). *Point sampling and line sampling: probability theory, geometric implications, synthesis*. Southern Forest Experiment Station, Forest Service, US Department of Agriculture.

Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

Haara, A. & Kangas, A. (2012). Comparing k nearest neighbours methods and linear regression–is there reason to select one over the other? *Mathematical and Computational Forestry & Natural Resource Sciences*, 4(1), 50.

Hansen, M. H. & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14(4), 333–362.

Hartig, G. L. (1795). *Anweisung zur Taxation der Forste, oder zur Bestimmung des Holzertrags der Wälder: Ein Beytrag zur höheren Forstwissenschaft: Nebst einer illuminirten Forst-Charte und mehreren Tabellen*. Heyer.

Hasel, A. A. (1938). Sampling error in timber surveys. *Journal of Agricultural Research*, 57(10), 713–736.

Hill, A. (2013). Comparison of small area estimators in forest inventory using airborne laserscanning data. Master's thesis, ETH Zurich, University of Göttingen.

Hill, A. (2017). *forestinventory*. GitHub repository.

Hill, A., Breschan, J., & Mandallaz, D. (2014). Accuracy assessment of timber volume maps using forest inventory data and lidar canopy height models. *Forests*, 5(9), 2253–2275.

Hill, A., Buddenbaum, H., & Mandallaz, D. (2018). Combining canopy height and tree species map information for large scale timber volume estimations under strong heterogeneity of auxiliary data and variable sample plot sizes. submitted for publication.

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.

Hill, A. & Massey, A. (2017). *forestinventory: Design–Based Global and Small-Area Estimations for Multiphase Forest Inventories. R package version 0.3.1.*

Hoffmann, C. (1982). Die Berechnung von Tarifen für die Waldinventur. *Forstwissenschaftliches Centralblatt*, 101(1), 24–36.

Hollaus, M., Wagner, W., Maier, B., & Schadauer, K. (2007). Airborne laser scanning of forest stem volume in a mountainous environment. *Sensors*, 7(8), 1559–1577.

Holmgren, J. (2004). Prediction of tree height, basal area and stem volume in forest stands using airborne laser scanning. *Scandinavian Journal of Forest Research*, 19(6), 543–553.

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260), 663–685.

Husmann, K., Rumpf, S., & Nagel, J. (2017). Biomass functions and nutrient contents of european beech, oak, sycamore maple and ash and their meaning for the biomass supply chain. *Journal of Cleaner Production*.

Jakubowski, M. K., Guo, Q., & Kelly, M. (2013). Tradeoffs between lidar pulse density and forest measurement accuracy. *Remote Sensing of Environment*, 130, 245–253.

Kangas, A. & Maltamo, M. (2006). *Forest inventory: methodology and applications*, volume 10. Springer Science & Business Media.

Keller, M. (2011). Swiss national forest inventory. manual of the field survey 2004–2007. *Swiss Federal Research Institute WSL, Birmensdorf, CH*.

Kirchhoefer, M., Schumacher, J., Adler, P., & Kändler, G. (2017). Considerations towards a novel approach for integrating angle–count sampling data in remote sensing based forest inventories. *Forests*, 8(7), 239.

Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P., et al. (1983). Optimization by simulated annealing. *science*, 220(4598), 671–680.

Koch, B. (2010). Status and future of laser scanning, synthetic aperture radar and hyperspectral remote sensing data for forest biomass assessment. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6), 581–590.

Köhl, M., Magnussen, S. S., & Marchetti, M. (2006). *Sampling methods, remote sensing and GIS multiresource forest inventory*. Springer Science & Business Media.

Kublin, E. (2003). Einheitliche beschreibung der schaftform–methoden und programme–bdatpro. *Forstwissenschaftliches Centralblatt*, 122(3), 183–200.

Kublin, E., Breidenbach, J., & Kändler, G. (2013). A flexible stem taper and volume prediction method based on mixed–effects b–spline regression. *European journal of forest research*, 132(5-6), 983–997.

Kuliešis, A., Tomter, S. M., Vidal, C., & Lanz, A. (2016). Estimates of stem wood increments in forest resources: comparison of different approaches in forest inventory: consequences for international reporting: case studyof european forests. *Annals of Forest Science*, 73(4), 857–869.

Lamprecht, S., Hill, A., Stoffels, J., & Udelhoven, T. (2017). A machine learning method for co–registration and individual tree matching of forest inventory and airborne laser scanning data. *Remote Sensing*, 9(5).

Langballe, N. & Fogh, I. (1938). Improved methods of procedure in cruising for operating purposes. *The Forestry Chronicle*, 14(2), 84–87.

Latifi, H., Nothdurft, A., & Koch, B. (2010). Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/lidar-derived predictors. *Forestry*, 83(4), 395–407.

Latifi, H., Nothdurft, A., Straub, C., & Koch, B. (2012). Modelling stratified forest attributes using optical/lidar features in a central european landscape. *International Journal of Digital Earth*, 5(2), 106–132.

Lefsky, M. A., Cohen, W., Acker, S., Parker, G. G., Spies, T., & Harding, D. (1999). Lidar remote sensing of the canopy structure and biophysical properties of douglas-fir western hemlock forests. *Remote sensing of environment*, 70(3), 339–361.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3), 18–22.

Lumley, T. (2016). ***survey**: Analysis of Complex Survey Samples*. R package version 3.32.

LWaldG (2000). *Landeswaldgesetz Rheinland-Pfalz (Forest Act Rhineland-Palatinate)*. Rhineland-Palatinate, Germany.

Maack, J., Lingenfelder, M., Weinacker, H., & Koch, B. (2016). Modelling the standing timber volume of baden-württemberg–a large-scale approach using a fusion of landsat, airborne lidar and national forest inventory data. *International Journal of Applied Earth Observation and Geoinformation*, 49, 107–116.

Magnussen, S., Eggermont, P., & LaRiccia, V. N. (1999). Recovering tree heights from airborne laser scanner data. *Forest science*, 45(3), 407–422.

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.

Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., & Kändler, G. (2017). Area-level analysis of forest inventory variables. *European Journal of Forest Research*, (pp. 1–17).

Magnussen, S., Næsset, E., & Gobakken, T. (2010a). Reliability of lidar derived predictors of forest inventory attributes: A case study with norway spruce. *Remote Sensing of Environment*, 114(4), 700–712.

Magnussen, S. & Tomppo, E. (2014). The k-nearest neighbor technique with local linear regression. *Scandinavian journal of forest research*, 29(2), 120–131.

Magnussen, S., Tomppo, E., & McRoberts, R. E. (2010b). A model-assisted k-nearest neighbour approach to remove extrapolation bias. *Scandinavian Journal of Forest Research*, 25(2), 174–184.

Mallows, C. L. (2000). Some comments on cp. *Technometrics*, 42(1), 87–94.

Mandallaz, D. (1993). *Geostatistical methods for double sampling schemes*. Habilitation, ETH Zurich.

Mandallaz, D. (2008). *Sampling techniques for forest inventories*. CRC Press.

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.

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.

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.

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.

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.

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.

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.

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.

Massey, A. & Mandallaz, D. (2015b). Design-based regression estimation of net change for forest inventories. *Canadian Journal of Forest Research*, 45(12), 1775–1784.

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.

Massey, A. F. (2015). *Multiphase estimation procedures for forest inventories under the design-based Monte Carlo approach.* PhD thesis, ETH Zurich.

Mathworks (2017). Matlab version 9.2.0.538062 (r2017a).

Mauro, F., Valbuena, R., Manzanera, J., & García-Abril, A. (2010). Influence of global navigation satellite system errors in positioning inventory plots for tree-height distribution studies this article is one of a selection of papers from extending forest inventory and monitoring over space and time. *Canadian journal of forest research*, 41(1), 11–23.

McRoberts, R. E. (2010). The effects of rectification and global positioning system errors on satellite image–based estimates of forest area. *Remote Sensing of Environment*, 114(8), 1710–1717.

McRoberts, R. E., Næsset, E., & Gobakken, T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. *Remote Sensing of Environment*, 128, 268–275.

McRoberts, R. E., Næsset, E., Gobakken, T., & Bollandsås, O. M. (2015). Indirect and direct estimation of forest biomass change using forest inventory and airborne laser scanning data. *Remote Sensing of Environment*, 164, 36–42.

McRoberts, R. E., Tomppo, E. O., Finley, A. O., & Heikkinen, J. (2007). Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. *Remote Sensing of Environment*, 111(4), 466–480.

McRoberts, R. E., Tomppo, E. O., & Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research*, 25(4), 368–381.

Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote sensing of environment*, 80(1), 88–99.

Næsset, E. (2007). Airborne laser scanning as a method in operational forest inventory: status of accuracy assessments accomplished in scandinavia. *Scandinavian Journal of Forest Research*, 22(5), 433–442.

Nink, S., Hill, J., Buddenbaum, H., Stoffels, J., Sachtleber, T., & Langshausen, J. (2015). Assessing the suitability of future multi-and hyperspectral satellite systems for mapping the spatial distribution of norway spruce timber volume. *Remote Sensing*, 7(9), 12009–12040.

Nothdurft, A., Saborowski, J., & Breidenbach, J. (2009). Spatial prediction of forest stand variables. *European Journal of Forest Research*, 128(3), 241–251.

NÃ¦sset, E. (1997). Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment*, 61(2), 246 – 253.

NÃ¦sset, E. (2014). Area-based inventory in norway – from innovation to an operational reality. In *Forest Applications of Airborne Laser Scanning – Concepts and Case Studies* chapter 11, (pp. 216–240). Springer.

Ott, E., Frehner, M., Frey, H.-U., & LÃ¼scher, P. (1997). *GebirgsnadelwÃ¤lder: Ein praxisorientierter Leitfaden fÃ¼r eine standortgerechte Waldbehandlung*. Haupt. .

Pinheiro, J. & Bates, D. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Pirlot, M. (1996). General local search methods. *European journal of operational research*, 92(3), 493–511.

Polley, H., Schmitz, F., Hennig, P., & Kroiher, F. (2010). Germany. In *National Forest Inventories - Pathways for Common Reporting* chapter 13, (pp. 223–243). Springer.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, J. N. (2015). *Small-Area Estimation*. Wiley Online Library.

Rayward-Smith, V. J., Osman, C., Reeves, C. R., & Smith, G. D. (1996). *Modern heuristic search methods*. John Wiley.

Richards, J. A. & Richards, J. (1999). *Remote sensing digital image analysis. An introduction*, volume 3. Springer.

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.

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

SAS Institute Inc. (2015). *SAS/STAT Software, Version 9.4.* Cary, NC.

Schmitz, F., Polley, H., Hennig, P., Dunger, K., & Schwitzgebel, F. (2008). Die zweite Bundeswaldinventur - BWI2: Inventur- und Auswertmethoden, Bundesministerium fur Ernahrung, Land-Wirtschaft und Verbraucherschutz (Hrsg).

Schreuder, H. T., Gregoire, T. G., & Wood, G. B. (1993). *Sampling methods for multiresource forest inventory.* John Wiley & Sons.

Srivastava, A. K., Srivastava, V. K., & Ullah, A. (1995). The coefficient of determination and its adjusted version in linear regression models. *Econometric reviews*, 14(2), 229–240.

Steinmann, K., Mandallaz, D., Ginzler, C., & Lanz, A. (2013). Small area estimations of proportion of forest and timber volume combining lidar data and stereo aerial images with terrestrial data. *Scandinavian journal of forest research*, 28(4), 373–385.

Stoffels, J., Hill, J., Sachtleber, T., Mader, S., Buddenbaum, H., Stern, O., Langshausen, J., Dietz, J., & Ontrup, G. (2015). Satellite-based derivation of high-resolution forest information layers for operational forest management. *Forests*, 6(6), 1982–2013.

Stoffels, J., Mader, S., Hill, J., Werner, W., & Ontrup, G. (2012). Satellite-based stand-wise forest cover type mapping using a spatially adaptive classification approach. *European journal of forest research*, 131(4), 1071–1089.

Straub, C., Dees, M., Weinacker, H., & Koch, B. (2009). Using airborne laser scanner data and cir orthophotos to estimate the stem volume of forest stands. *Photogrammetrie-Fernerkundung-Geoinformation*, 2009(3), 277–287.

Thünen-Institut (2014). Dritte Bundeswaldinventur 2012. Accessed: 2017-02-03.

Tomppo, E. (2006). The finnish multi-source national forest inventory–small area estimation and map production. *Forest inventory–methodology and applications. Springer, Dordrecht, NL*, (pp. 195–224).

Tonolli, S., Dalponte, M., Vescovo, L., Rodeghiero, M., Bruzzone, L., & Gianelle, D. (2011). Mapping and modeling forest tree volume using forest inventory and airborne laser scanning. *European Journal of Forest Research*, 130(4), 569–577.

Van Aardt, J., Wynne, R., & Scrivani, J. (2008). Lidar-based mapping of forest volume and biomass by taxonomic group using structurally homogenous segments. *Photogrammetric Engineering & Remote Sensing*, 74(8), 1033–1044.

von Carlowitz, H. C. (1713). *Sylvicultura oeconomica.* J.F. Braun, Leipzig.

von Lüpke, N. (2013). *Approaches for the Optimisation of Double Sampling for Stratification in Repeated Forest Inventories.* PhD 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.

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.

White, J. C., Coops, N. C., Wulder, M. A., Vastaranta, M., Hilker, T., & Tompalski, P. (2016). Remote sensing technologies for enhancing forest inventories: A review. *Canadian Journal of Remote Sensing*, 42(5), 619–641.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilcoxon, F., Katti, S., & Wilcox, R. A. (1970). Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test. *Selected tables in mathematical statistics*, 1, 171–259.

Zianis, D., Muukkonen, P., Mäkipää, R., Mencuccini, M., et al. (2005). Biomass and stem volume equations for tree species in Europe. *Silva Fennica*, Monographs 4.

Zöhrer, F. (1980). *Forstinventur: Leitfaden für Studium und Lehre*. Practice/Hamberg, Berlin: Parey,(Parey Studientexte: Nr. 26).

Curriculum Vitae

Personal Details

Name	Andreas Christian Hill
Date of birth	15.01.1986
Place of birth	Trier (Germany)

Education

02/2014–03/2018	Ph.D. Student ETH Zurich, Department of Environmental Systems Science, Chair of Forest Engineering
04/2015–05/2017	Diploma of Advanced Studies ETH ETH Zurich, Applied Statistics
03/2011–03/2012	ERASMUS study abroad at ETH Zurich, Switzerland
04/2010–08/2013	Master of Science (M.Sc.) University of Göttingen, Forest Sciences and Forest Ecology with study focus on Forest Ecosystem Analysis and Information Processing
10/2006–03/2010	Bachelor of Science (B.Sc.) University of Göttingen, Forest Sciences and Forest Ecology

Professional Experience

Since 10/2013	Researcher Department of Environmental System Science, Chair of Forest Engineering, ETH Zurich
10/2011–04/2012	Research Assistant Department of Environmental System Science, Chair of Forest Engineering, ETH Zurich
09/2007–03/2011	Research Assistant Department of Forest Inventory and Remote Sensing, Faculty of Forest Sciences, University of Göttingen

Publications in Conference Proceedings and Workshops

[1] **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 German National Forest Inventory. In *3rd International Workshop on Forest Inventory Statistics*. Freiburg, Baden-Württemberg Germany.

[2] **Hill, A.**, Stoffels, J., Langshausen, J. (2016): . In *CARISMA-workshop on large-scale mapping and estimation of forest resources*. Ås, Norway

[3] **Hill, A.**, Breschan, J. (2014): Automatic Design of Efficient Harvesting Units using Remote Sensing and Field Data. In *24th IUFRO World Congress*. Salt Lake City, Utah, USA

[4] Breschan, J., **Hill, A.** (2014): Validation of timber volume maps derived from remote sensing data. In *24th IUFRO World Congress*. Salt Lake City, Utah, USA

Publications in Scientific Journals

[1] Lamprecht, S., **Hill, A.**, Stoffels, J., Udelhoven, T.(2017): A Machine Learning Method for Co-Registration and Individual Tree Matching of Forest Inventory and Airborne Laser Scanning Data. *Remote Sensing*, 9 (5). doi: 10.3390/rs9050505

[2] **Hill, A.**, Breschan, J., Mandallaz, D. (2014): Accuracy Assessment of Timber Volume Maps using Forest Inventory Data and LiDAR Canopy Height Models. *Forests*, 5 (9). 2253-2275. doi: 10.3390/f5092253

[3] 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

Other Publications

[1] 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

[2] **Hill, A.**, Massey, A., Mandallaz D. (2016): forestinventory: Design-Based Global and Small-Area Estimations for Multiphase Forest Inventories. R package version 0.1.0 *CRAN Repository* url: https://CRAN.R-project.org/package=forestinventory