ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Comparison of Small Area Estimators in Forest Inventory using Airborne Laserscanning Data

Vergleich von Kleingebietsschätzern in der Waldinventur unter Benutzung von flugzeugerhobenen Laserscanner Daten

Master Thesis

Georg-August-University Göttingen
Faculty of Forest Sciences and Forest Ecology
Department Ecoinformatics, Biometrics and Forest Growth

Andreas Hill

Supervisor:
Prof. Dr. Joachim Saborowski
Faculty of Forest Sciences University of Göttingen

Co-Supervisor:
PD Dr. Daniel Mandallaz
Department of Environmental System Science ETH Zürich

Göttingen, 11.07.2013

Acknowledgement

First of all, I want to express my deep gratitude to my supervisors, Prof. Dr. Joachim Saborowski and PD Dr. Daniel Mandallaz.

Joachim Saborowski provided the supervision of this thesis on the part of the Faculty of Forest Sciences and Forest Ecology of Göttingen. His confidence in me and his willingness of supervising an external master thesis only made this thesis possible. Whenever I had any concerns, he always took the time to give me helpful advice. I appreciated this very much.

Special gratitude is owed to Daniel Mandallaz who was my next-door supervisor at the ETH Zurich for sparking my interest on forest inventory methods in the framework of his lecture and for his excellent support throughout the whole thesis. It was a great pleasure to work with him and to benefit from his great experience in the field of statistics and forest inventory.

At this point, I would like to thank the whole team of the Chair of Land Use Engineering at the ETH Zurich for their support and their interest in my work. My special thanks go to Prof. Dr. Hans Rudolf Heinimann who gave me the chance to complete my master thesis at his chair and provided me with a perfect working environment. I also want to thank Dr. Jochen Breschan for his major support in organizing this thesis as well as for his supervision of the programming part. I want to express further gratitude to Patricia Moll who helped me getting started with the programming language Python, Daniel Trüssel who gave me helpful advice in advanced GIS applications as well as to Maja Messerli who always helped me in any organization concerns.

Special thanks also go to the “Amt für Wald und Naturgefahren Graubünden” for providing the terrestrial inventory data as well as any further information.

Last but not least, I would like to thank my parents for all their confidence and support in every moment of my studies. Without their help, I would never have had the chance to enjoy such interesting studies. For this I will always be incredibly grateful. I also want to thank my girlfriend Christina for all her love and support.

Abstract

The objective of this thesis was to compare new and extended versions of two-phase estimators proposed by Mandallaz (2013a, 2013b) with already existing estimators for design-based global and small area inventories. The newly proposed estimators especially aim at taking full advantage of auxiliary information which is *exhaustively* available over an inventory area in order to yield a further improvement in estimation precision of existing regression estimators. In this context, also a new approach for deriving a better estimation of the design-based variance (Mandallaz 2013a) was applied. The estimators were analyzed as part of a forest inventory study in eastern Switzerland with the objective to estimate the standing timber volume within the entire inventory area as well as within four sub-units (small areas). With thanks to the recent advancements in the use of high resolution remote sensing data in forest inventory, a canopy height model (CHM) constructed from high resolution airborne laser scanning data (LiDAR) provided the auxiliary information over the entire inventory area. The auxiliary variables derived from the CHM showed overall satisfactory correlations to the observed terrestrial standing timber volume. In particular, the mean canopy height and a timber volume estimation on plot level based on single tree detection revealed to be the auxiliary variables with the highest coefficient of determination ($R^2$of 0.5). The goodness of fit for the entire inventory area was improved to R^2 of 0.65 and 0.66 by using further descriptive statistics of the CHM as explanatory variables. The coefficients of determination in the small areas were, with one exception, at least equally good. The explanatory power of the variables was most likely due to the high quality of the LiDAR data, but additionally boosted by sophisticated processing concepts such as an adjustment for boundary effects.

The comparison of the design-based estimations for the standing timber volume revealed that the estimated standard error was substantially reduced by up to 56% when changing from the one-phase approach (sample mean) to the two-phase approaches (regression estimators), particularly in the case of the small area estimations. The new regression estimator proposed by Mandallaz (2013b), making use of exhaustive auxiliary information, yielded an additional improvement over the classical regression estimator. Here, the mean canopy height clearly appeared to be the variable of choice: whereas the exhaustive computation of other variables needed high computation effort, it was possible to compute the mean canopy height of the inventory area extremely fast in one step within a geographical information system (GIS) while providing the most important reduction of variance achieved with the new estimator. The new approach of deriving a better design-based variance estimate did in this case not reveal large differences to the classical approach. While this was most likely due to the availability of sufficiently large sample sizes in the small areas, the calculation of the more precise estimated variance is recommended if sample sizes in small areas are small.

Zusammenfassung

Ziel dieser Arbeit war es, die von Mandallaz (2013a, 2013b) neu entwickelten und erweiterten Versionen zweiphasiger, design-basierter Stichprobenschätzer mit bereits bestehenden, in der Waldinventur eingesetzten Schätzern zu vergleichen. Die neu entwickelten bzw. erweiterten Schätzer wurden insbesondere mit dem Ziel entwickelt, die Verfügbarkeit von Hilfsinformation, welche *flächendeckend* über ein gesamtes Inventurgebiet verfügbar ist, optimal zu nutzen um die geschätzte Varianz der Stichprobenschätzung gegenüber bestehenden Verfahren weiter zu reduzieren. In diesem Rahmen wurde auch ein neuer Ansatz zur verbesserten Berechnung der geschätzten design-basierten Varianz angewendet. Zum Vergleich wurde für ein Untersuchungsgebiet in der Ostschweiz (Kanton Graubünden) unter Benutzung dieser Schätzer eine Holzvorratsschätzung für das gesamte Inventurgebiet sowie für 4 Teilgebiete (Kleingebiete) vorgenommen. Die Hilfsinformation wurde aus einem Kronenhöhenmodell (CHM) abgeleitet, welches auf der Basis von hochaufgelösten Laserscanner Daten berechnet wurde und das Inventurgebiet flächendeckend abdeckt. Vor allem die mittlere Kronenhöhe sowie eine Schätzung des Holzvorrates auf der Basis von Einzelbaumdetektionen im CHM erwiesen sich als die Variablen mit dem höchsten Bestimmtheitsmaß (R^2 von 0.5). Als Voraussetzung für die bemerkenswert gute Vorhersagekraft fast aller Hilfsvariablen wurde die hohe Qualität der Laserscanner Daten verantwortlich gemacht, allerdings spielte auch die aufwendige Berechnung der Hilfsvariablen (einschließlich der Korrektur für Randeffekte) eine entscheidende Rolle. Durch die Benutzung zusätzlicher Variablen konnte die Modellgüte für das gesamte Inventurgebiet auf Bestimmtheitsmaße von 0.65 und 0.66 gesteigert werden, wobei, mit einer Ausnahme, diese in den vier Kleingebiete mindestens gleichgut waren. Die größte Reduktion des Standardfehlers wurde durch den Wechsel vom einphasigen Schätzverfahren (nur auf Basis der terrestrischen Stichproben) zu den zweiphasigen Schätzverfahren (Regressionsschätzer) erreicht (besonders deutlich im Falle der Kleingebietsschätzungen mit Reduzierung des Standardfehlers um bis zu 56%). Der von Mandallaz (2013b) neu entwickelte Regressionsschätzer führte nochmals zu einer Steigerung der Schätzgenauigkeit gegenüber dem klassischen Regressionsschätzer. Die mittlere Kronenhöhe stellte sich im Rahmen der flächendeckenden Berechnung der Hilfsvariablen als von besonderem Interesse heraus: im Vergleich zu anderen Hilfsvariablen konnte die mittlere Kronenhöhe für das gesamte Inventurgebiet und die Kleingebiete unter Benutzung eines Geoinformationssystems (GIS) extrem schnell ermittelt werden, wobei mit dieser Hilfsvariable auch die deutlichste Varianzreduktion bei Anwendung des neuen Regressionsschätzers erzielt wurde. Die genauere, design-basierte Varianzschätzung zeigte in dieser Studie nur geringe Unterschiede zu der herkömmlichen Varianzschätzung, sehr wahrscheinlich aufgrund der ausreichend hohen Anzahl an Stichprobenpunkten in den Kleingebieten. Allerdings ist die Verwendung der genaueren Varianzberechnung bei geringer Anzahl an verfügbaren terrestrischen Stichprobenpunkten zu empfehlen.

List of Abbreviations

AIC	Akaike Information Criterion
CHM	Canopy Height Model
C_p	Mallow's C_p
DBH	Diameter at Breast Height
DGPS	Differential Global Positioning System
DSM	Digital Surface Model
DTM	Digital Terrain Model
d7	Diameter at a Tree Height of 7 Meters
$\mathbb{E}$	Expected Value
F	Definition of Forest
FM	Forest Mask
GIS	Geo Information System
GPS	Global Positioning System
GRE / greg	Generalized Regression Estimator
IMU	Inertial Measurement Unit
K	Circle
LiDAR	Light Detection and Ranging (Laser Scanning)
lm	Large Model
ln	Logarithmus Naturalis
LSP	LiDAR Sample Plot
MAD	Median Absolute Deviation
MSE	Mean Square Error
N	Population Size
n	Sample Size
n_1	Number of First Phase Sample Points
n_2	Number of Second Phase Sample Points (Terrestrial Observations)
NFI	Swiss National Forest Inventory
p	p-value / predictor

P	Population
Pol-InSAR	Polarimetric Inferometric Synthetic Aperture Radar
PPE	Prediction Proportional to Error
PPP	Prediction Proportional to Prediction
PPS	Prediction Proportional to Size
PRF	Pulse Repetition Frequency
QQ-Plot	Quantile-Quantile Plot
r	Radius
$\mathbb{R}$	Set of real numbers
RE / reg	Regression Estimator
R^2	Coefficient of Determination
R^2_{adj}	Adjusted Coefficient of Determination
rm	Reduced Model
RSS	Residual Square Sum
s	sample
s_1	First Phase Sample
s_2	Second Phase Sample (Terrestrial Sample)
SA	Small Area
Sd / σ	Standard deviation
TLM3D	Topographic Landscape Model of Switzerland
TSS	Total Square Sum
V / σ^2	Variance
$\hat{V}$ / $\hat{\sigma}^2$	Estimated Variance
VIF	Variance Inflation Factor
X	Design Matrix
Y	Response Variable of Regression Model
$Y(x)$	Local Density at sample point *x*
$\hat{Y}(x)$	Predicted Local Density at sample point *x*
α	Significance Level (Type I Error)
κ	Condition Number
λ	Area / Eigenvalue
π	Inclusion Probability

Table of Contents

List of Figures

List of Tables

1 Introduction

The execution of forest inventories dates primarily back to at least the end of the Middle Ages (McRoberts et al. 2010), but has especially gained importance with the idea of sustainable forest management being introduced in literature for the first time by the German Hans Carl von Carlowitz (1713). For long time the needs on the forest sector mainly comprised the production of timber as raw material for industry and mining and it was for that reason that sustainable forest planning primarily required information about available timber resources as well as the timber increments. This information was, and still is, acquired and provided by forest inventories. Since the 1950s, however, the demands of forestry products and thus the information requirements have been steadily increasing (Köhl et al. 2006), including more diverse functions of forests which are often referred to as ecosystem services: beside the productive functions of forests, these amongst others comprise the recreational role of forests (socioeconomic functions), the conservation of biodiversity, forest health and vitality (McRoberts et al. 2010) and also their regulating and protective functions (Berger and Rey 2004; Frehner et al. 2005). Today's forest management also needs to comply with various reporting requirements from national as well as international conventions, such as those of the Forest Stewardship Council, the Food and Agricultural Organization or the Kyoto Protocol (McRoberts and Tomppo 2007).

Particularly within the framework of climate change (IPCC 2007; Raupach et al. 2007), forest ecosystems have lately been under increased scientific observation: being considered as strong indicators for potential responses of ecosystems to ongoing climate change impacts, forests play an enormous role as carbon sinks and sources. Various research projects have been dealing with the challenging task how to maintain forest productivity under vague climate change impacts (Geßler et al. 2007; Kramer et al. 2010). Although this research has yet been mainly based on model predictions, it is evident that improved information especially of potential growth rates and the state of health is needed to accomplish further progress in these matters. Consequently, today's forest inventories have to derive a huge amount of diverse information in order to satisfy the needs of society and scientists. In Europe for example, national forest inventories typically collect information on 100-400 variables (McRoberts et al. 2010).

A broad range of concepts and methods for collecting the desired forest information have been proposed (Cochran 1977; Gregoire and Valentine 2008; Köhl et al. 2006; Mandallaz 2008; Schreuder et al. 1993). The most straightforward approach is to derive the information by a full census of all trees in the field. However, this procedure is time-consuming and cost-intensive and in most cases even practically unfeasible due to the large spatial extension of forest areas (Gregoire and Valentine 2008). For this reason it has become increasingly attractive, not to say inevitable, to derive the information of interest by using only a subset of terrestrial observations and measurements by means of statistical sampling theory. Within the sampling framework, one no longer determines the true value of

the variable of interest, but aims at providing precise estimates. In this context one of the prime objectives in forest inventory is to keep the uncertainty of these estimations (i.e. the variance) as small as possible.

Among the variety of statistical methods, the most important concepts rely on two basic principles: one is based on exclusively collecting terrestrial samples to make estimations, so-called *one-phase sampling*. The derived information through terrestrial sampling is assumed to be very precise and increasing the precision of the estimates could primarily be realized by increasing the number of terrestrial samples. However, since field surveys are very time-consuming and expensive, the number of terrestrial samples is usually limited. Developments in the recent years showed an increasing need for alternative, less costly inventory methods to maintain estimation precision even under more and more limited resources for terrestrial samples. Progress in the development of sampling methods is therefore especially important for large scale forest inventories, where an increase of efficiency (i.e. either achieving the highest precision with predefined cost or achieving a predefined precision with minimal costs) can lead to considerable savings (von Lüpke 2013).

A method which has become particularly attractive with respect to this challenge is so called *double* – or *two-phase sampling* (Cochran 1977; Gregoire and Valentine 2008; Köhl et al. 2006; Mandallaz 2008; Schreuder et al. 1993) in which terrestrial samples are partially replaced by, and combined with, additional auxiliary data. The general principle of two-phase sampling is to enlarge the sample size in order to achieve a higher estimation precision as under exclusively using the terrestrial samples. This is achieved by using predictions of the terrestrial observations at additional sample locations where the terrestrial information has not been derived. These predictions are based on the relatively cheap auxiliary information which are combined with highly precise but expensive terrestrial samples in the framework of regression estimators (Rao 1988; Särndal et al. 2003). In this context one must however still consider the potential reduction in variance by the use of auxiliary information relative to the reduction that would be achieved by investing in a larger terrestrial sample (Gregoire and Valentine 2008). A rule of thumb for supporting this decision-making process is provided by Mandallaz (2008). It has also been shown that the efficiency of two-phase sampling can still be increased by extending this procedure to stratification (Saborowski et al. 2010, von Lüpke 2013).

In most two-phase inventories, the auxiliary information is provided by remote sensing data. For a long time, the information derived from these observations were exclusively based on visual interpretation of aerial photographs (Hildebrandt 1996) and often of qualitative manner. Whereas this method is still used for certain issues, e.g. forest-non forest decision and classification of development stages within the Swiss National Forest Inventory (Brassel and Lischke 2001), more recent studies have shown that qualitative and quantitative forest parameters can also be estimated from remote sensing data based on semi-automatic algorithms. These advancements have made the use of remote sensing data even more attractive for their application within forest inventories, as the automatic derivation of information can be time- as well as cost-saving and, owing to enormous

increases in computing capacities, provides the possibility to assess information at an excessively large number of sample points (McRoberts et al. 2010). This is further supported by the fact that remote sensing data can be derived from a broad range of instruments and data sources and that they provide access to information over large areas. Especially in cases of regional inventories or inventories at the enterprise level, remote sensing data may provide even exhaustive coverage of entire inventory areas. Beside passive remote sensing techniques (such as multispectral airborne and satellite imagery), which have already proven its suitability for various forestry applications (Schlerf et al. 2010; Stoffels et al. 2011; Vohland et al. 2007), an active remote sensing technique which attracts much attention in the field of forestry is Airborne Laserscanning (LiDAR). Especially the high spatial resolution of LiDAR data and its capacity to penetrate into vegetation canopies provides a variety of forest information to be acquired on stand level or even on the level of individual trees (van Leeuwen and Nieuwenhuis 2010). While LiDAR data can even be used for tree species classification (Brandtberg 2007; Holmgren and Persson 2004; Ørka et al. 2009), its specific strength lies in its ability to provided information about three-dimensional vegetation structures which has successfully been used to recover tree heights (Magnussen et al. 1999) as well as to estimate timber volume at plot level by regression models (Holmgren 2004) or k-nearest neighbour techniques (Nothdurft et al. 2009). Because of these possibilities, LiDAR data have already been used in various forest inventories (Hyyppä et al. 2008; Næsset 2002, 2007). However, it has to be noted that despite of the huge advancements of remote sensing data in forest inventories, terrestrial samples cannot be replaced completely (McRoberts et al. 2010).

Beside the need of forest inventories at the national or international scale, also information about forest attributes at the regional scale or even at enterprise level is often desired in the framework of forest planning. A problem which often arises here is that for such small geographic areas (also referred as small domains) the number of available terrestrial samples is too few for making estimates with acceptable precision, i.e. the standard errors of the estimates get considerably large (Köhl et al. 2006). The reason for the small sample sizes is often due to the fact that the overall number of samples has been originally designed to achieve a specific accuracy within much larger spatial entities than that of the small area. Statistical methods which cope with the problem of still deriving acceptable estimates under these conditions are known as small area estimations, and a wide range of estimators have been proposed and applied in this context (Rao 2003). Whereas the use of small area estimations has its origin mainly in socio-economic and demographic studies such as the local estimation of population (Ghosh and Rao 1994), they have also been applied since the 1980s in the field of agriculture (Battese et al. 1988) and forestry (Green et al. 1987). Also in the context of small area estimations for environmental surveys, the use of auxiliary information (being related to the terrestrial variable of interest) has been a very promising approach to derive precise estimations (Köhl et al. 2006). Whereas the auxiliary information has often been provided by satellite data (McRoberts and Tomppo 2007), also canopy height models derived from aerial imagery have successfully been used as auxiliary information within small area estimation of forest attributes in the Norwegian national forest inventory (Breidenbach and Astrup 2012). Latest forest research in Swit-

zerland has additionally demonstrated that the use of LiDAR-derived data products as auxiliary information within two-phase small area inventories can lead to very promising results (Steinmann et al. 2012).

As further reductions of the uncertainty of predictions are desired by authorities, the improvement of statistical estimation methods is one of the main topics of current forest inventory research. Optimally, the respective methods lead to an increase of efficiency while maintaining existing sampling schemes (von Lüpke 2013). With respect to these requirements, the work presented in this thesis had the objective to compare new and extended versions of two-phase estimators to already existing estimators for design-based global and small area inventories. These new estimators were proposed by Mandallaz (2013a, 2013b) and especially aim at taking advantage of the case when auxiliary information is exhaustively available within the entire inventory area. In contrast to existing estimators, where even in this case the auxiliary information is only derived at discrete sample points, the newly developed estimators use the provided information exhaustively to achieve a further reduction of the standard error of the estimations. In this context, also a new approach for deriving a better estimation of the design-based variance (Mandallaz 2013a, 2013b) was applied.

With respect to the ongoing huge advancements in the use of high resolution remote sensing data as auxiliary information, the estimators in this study were applied to an inventory area which was completely covered by available high resolution airborne LiDAR data. The objectives of the study were:

- to investigate whether the newly proposed estimators in this case yield a reduction of the prediction-uncertainty compared to already existing estimators. The new estimators may provide a huge advancement as exhaustive auxiliary information becomes increasingly available, but has still not been used to its full potential.

- to develop an optimal pre-processing scheme for the inventory data. In addition to state-of-the-art pre-processing of the laser scanning data, this included the implementation of a two-phase sampling design in accordance to available terrestrial sample points from a regional forest inventory and a consistent forest definition.

- to derive auxiliary variables from the laser scanning data which provide highest possible explanatory power for the terrestrial observations in order to use them as predictors in the applied regression estimators.

2 Study Site

The investigated forest inventory methods in this study have been tested in a study site located in the canton of Grisons (Eastern Switzerland). The site extends in north-south direction between Klosters and Davos, covering a total area of 2887.39 hectare (Fig. 2-1). In order to make estimations also for subterritories of the study site, the study area was arbitrarily sub-divided into four approximately equally sized small areas with corresponding sizes of 846.5, 762.6, 593.8 and 684.5 hectare in north-south direction.

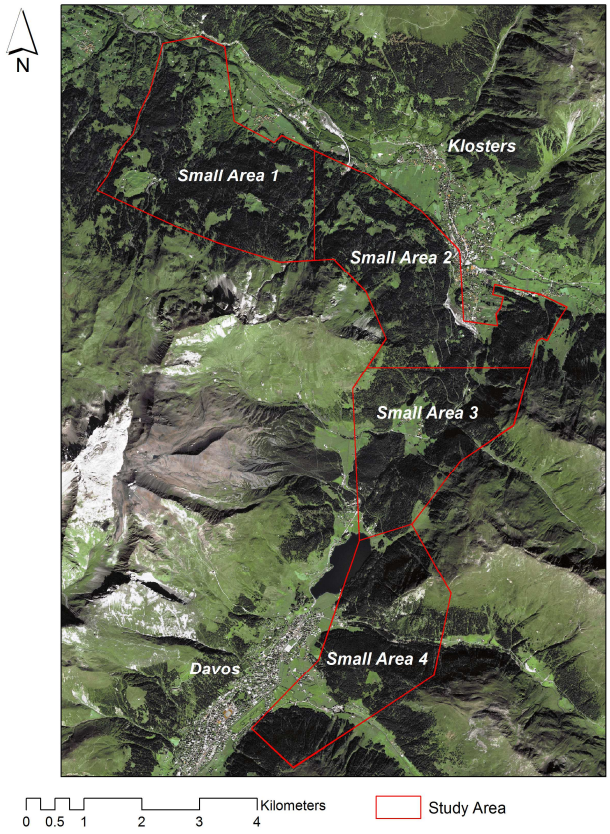


Fig. 2-1: Study area between Klosters and Davos. Also the division into four small areas is indicated. Background: Panchromatic Spot5 satellite image

According to the geographical segmentation of the Alps by Ott et al. (1997), the study area belongs to the region of the northern “Zwischenalpen” (Fig. 2-2). This part of the Alps has oceanic to continental climate conditions and is characterized by deeply incised valleys of northerly orientation. In normal years, the annual precipitation at an altitude of 1000 meters does not exceed 1200 mm. However, in comparison to the northern fringe of the Alps the growth potential of European beech is already considerably reduced: its stand-forming characteristic is limited to lower ranges at the valley entrances. At higher alti-

tudes, the occurrence of beech is rather sporadic. On the contrary, the proportion of the high montane vegetation zone, which is characterized by fir-spruce forests, considerably increases and tends to completely replace the upper montane zone (fir-beech forests), in particular at northerly exposed slopes (Fig. 2-2). On southern slopes, the colline vegetation height zone potentially rises up to altitudes of 800 meters, with oak and scots pine as dominating tree species. The alpine tree line in this part of the Alps is located at an altitude of approximately 2000 meters. At sites with increasing continental climate conditions, the upper tree line can already be found within the upper sub-alpine vegetation zone (characterized by larch-mountain pine forests).

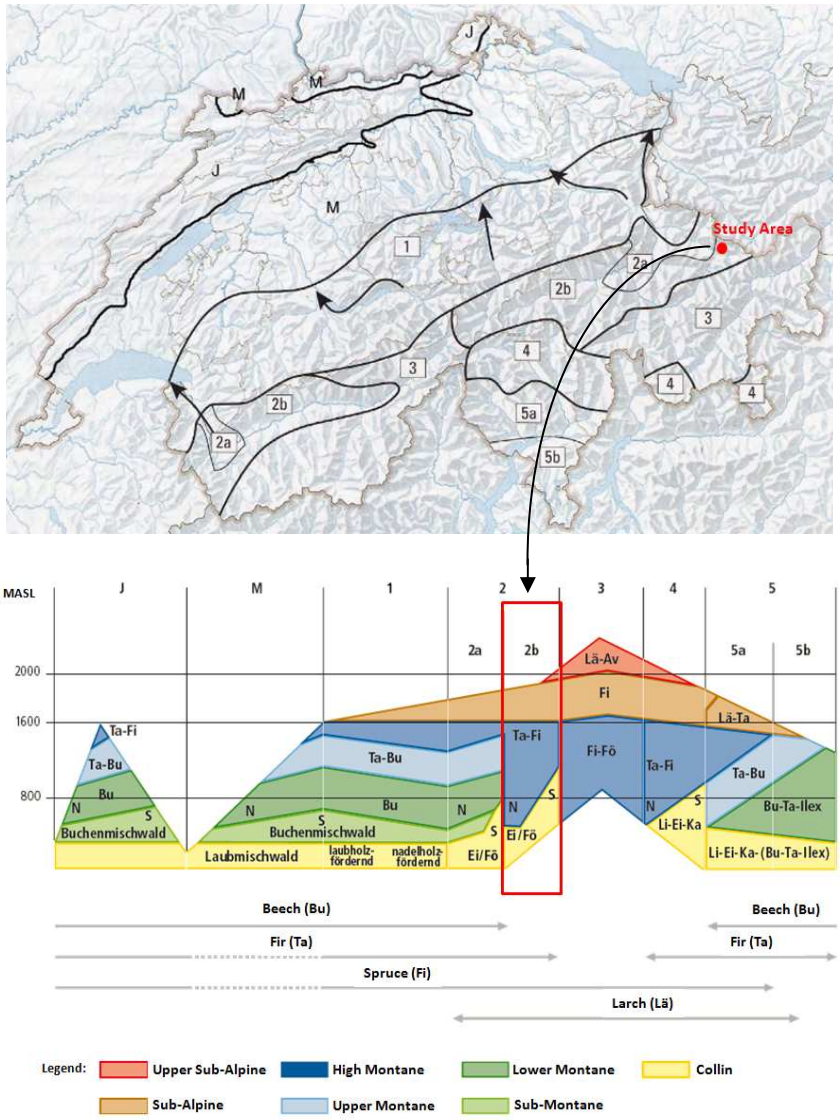


Fig. 2-2: *Above*: Geographical segmentation of the Alps taken from Ott et al. (1997): 1 (northern fringe of the Alps), 2a and 2b (northern "Zwischenalpen" with and without beech), 3 (continental High Alps), 4 (southern "Zwischenalpen") and 5a and 5b (southern edge of the Alps with and without Spruce). The location of the study area is indicated by a red point. *Below*: Vegetation Height Zones for the respective domains of the Alps (1-5) taken from Frehner et al. (2005). The study site is located at region 2b. Figure were modified by the author.

The study site is located at an altitude between 900 and 2200 meters above sea level, thus ranging between the high montane and the upper sub-alpine vegetation zone. Consequently, the forests within the study area were assumed to essentially consist of coniferous tree species, especially Norway spruce (*Picea abies*). The relief in the study site is primarily characterized by steep slopes (for illustration see Appendix A1 and A2).

The study site was chosen for the following reasons:

- The study site ideally represents typical forest areas at the regional and enterprise level within the domain of the Alps. According to the third Swiss National Forest Inventory (Brändli 2010), approximately 81 % of the complete forest area of the Alps (which makes up 34 % of the total swiss forest area) is located on higher altitudes (> 900 MASL). All respective height zones are covered by the study site.
- Terrestrial data from the latest cantonal forest inventory (2006/2007) were available.
- High-resolution Airborne Laserscanning data were available which have been acquired closely to the time period of the terrestrial inventory and cover the complete study site.
- Due to the local characteristics, the forest structure in the study site can be assumed to exhibit a high amount of variability in horizontal and vertical direction. Compared to homogeneous forests, a high variability in forest structure should better be suited to demonstrate the advantage of using two-phase- compared to one-phase estimation procedures.
- Mountainous landscape characteristics such as prevailing in the study site (steep slopes, considerably rough terrain, dense vegetation) can potentially hamper the acquisition of high-quality LiDAR data for forestry applications. The study area was hence perfectly suited to also investigate the applicability of LiDAR data in the framework of forest inventory in alpine and mountainous regions.

3 Data Management and Software

All available and produced geodata during this study were stored and managed in a consistent Geodatabase under the commercial software *ArcGIS* (vers. 10.1). All geodata were kept in the Swiss coordinate system CH1903 (geodetic datum). Simple geoprocessing operations as well as the creation of maps for visualization purposes were done in *ArcMap* (vers. 10.1). The Geodatabase amongst others comprised

- the laserscanning data for the study site (Raster Format)
- the inventory designs (Point and Polygon Features)
- the regional and national inventory data (Attribute Table)

Processing of the laserscanning data, comprising the pre-processing part and the computation of auxiliary variables, was carried out in the commercial software *MATLAB* (vers. R2011b) and the open-source software *Python* (vers. 2.7.2).

All statistical analyses, modelling as well as further data processing were performed in the open-source statistical programming language *R* (vers. 2.15.2). The statistical calculation of the estimators was performed within the procedure *iml* of the commercial statistical software *SAS* (vers. 9.2).

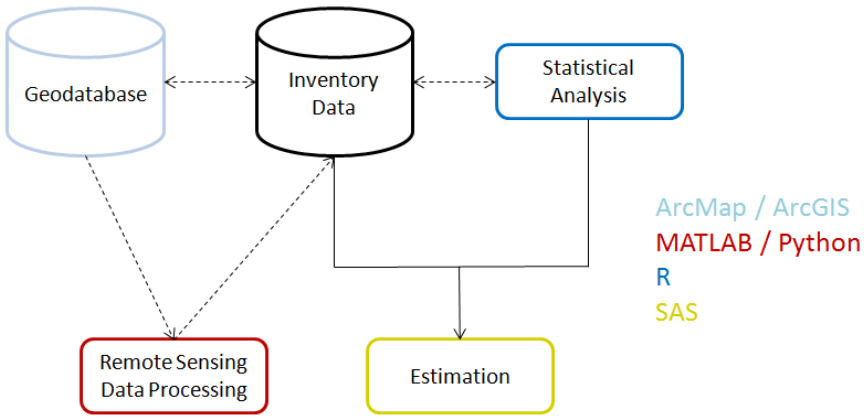


Fig. 3-1: Illustration of the data-handling for the present thesis.

4 Material and Methods

4.1 Terrestrial Inventory Data

The inventory in this study considerably used regional as well as national inventory data from terrestrial surveys. The following section will introduce the most important parameters derived by the surveys and provide detailed background information about the inventory methods.

4.1.1 National Forest Inventory in Switzerland

In this study, data from the third Swiss National Forest Inventory (NFI3), which took place from 2004 - 2007, were available for the domains Pre-Alps and Alps. The terrestrial field surveys of the NFI3 were performed at sample points of a systematic grid with a mesh size of 1.4 kilometers extending all over Switzerland (Fig. 4-2). At each sample plot, trees were included into the sample according to two concentric circles with a plot area of 200 and 500 m^2 around the sample centre. Within the inner circle (radius of 7.98 meters), all trees with a diameter at breast height (DBH) bigger than 12 cm were selected, where in the second circle (radius of 12.62 m), all trees with a DBH bigger than 36 cm were included into the sample. In addition, the relative position of each sample tree to the sample center as well as its species was recorded (Fig. 4-1).

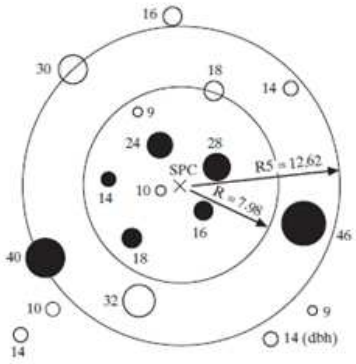


Fig. 4-1: The sample plots of the NFI3 consist of two concentric circles with radii of 12.62 and 7.98 meters. The DBH thresholds for the inner and outer circle are 12 and 36 cm. The blank circles indicate trees not included into the sample, where sampled trees are indicated by filled circles. The relative position to the sample center and the species is recorded for each sample tree. Source: Keller (2011)

If the sample plots had been located in sloped terrain, a slope correction (chapter 4.3) was performed according to the mean slope in a defined plot neighborhood (Keller 2011). One of the most important parameters recorded in the sample is the DBH, which e.g. allows to

estimate the timber volume of a tree by using so-called tariff-functions (Kaufmann 2000). This part of the terrestrial inventory is referred to the *first stage*, whose sampling scheme aims to approximate the so-called PPP-Principle (Prediction Proportional to Proportion) in the context of sampling theory, where larger trees possess a higher probability to be included into the sample than smaller trees. In the so-called *second stage* of the terrestrial inventory, also the height and diameter at 7 meters (*d7*) is obtained from a subset of selected trees of the first stage. This procedure aims at getting a more precise approximation of a trees timber volume by using the height and *d7* as additional information beside the DBH. Since for trees which are selected in the first as well as in the second stage, one gets two estimates for the timber volume (one of which is more precise), one can use the resulting residuals of the two estimates to correct the overall timber volume estimation of the sample plot (see also *One-phase-two-stage sampling*, Mandallaz 2008). The trees in the second stage are chosen according to the principle of *poisson sampling*, where the number of selected individuals is random and a selected tree can only occur once in the sample. The inclusion probability here is proportional to the DBH of a tree, thus favoring trees with larger DBH to be included into the sample. This aims at realizing the so-called PPE-Principle (Prediction Proportional to Error), since empirical observations revealed trees with larger DBH to cause more severe errors in estimating the timber volume than trees with smaller DBH. In the second NFI (1993-1995), on average two individuals were selected in the second stage for each sample plot, which overall makes approximately 12 % of all selected trees of the first stage (Brassel and Lischke 2001). It should be mentioned that the NFI is actually a two-phase two-stage inventory. Further information about the complete procedure can amongst others be found in Brändli (2010) and Brassel and Lischke (2001).

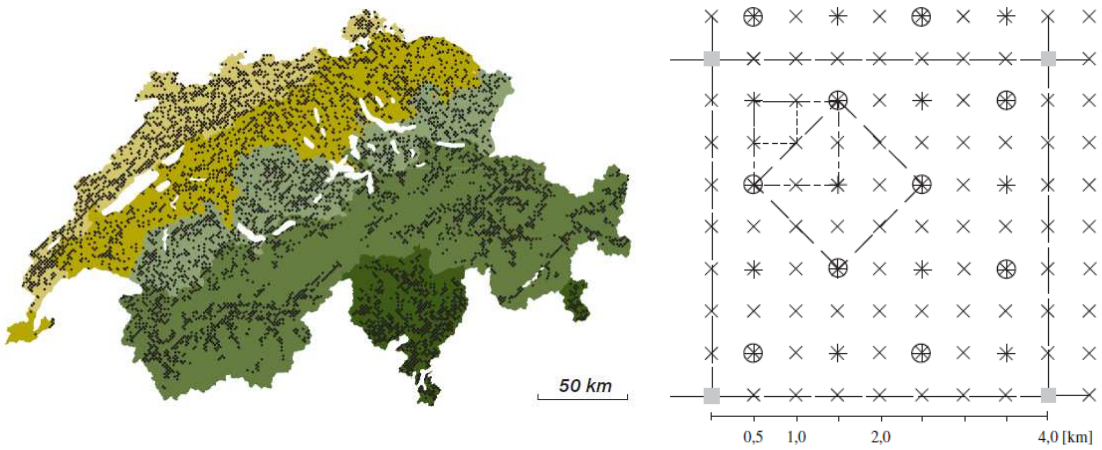


Fig. 4-2: *Left*: Distribution of the approximately 6500 terrestrial sample plots of the national inventory over Switzerland within the forest definition of the NFI. Colors indicating the five domains Jura (yellow), the Plateau regions (brown), Pre-Alps (bright green), Alps (green) and Southern-Alps (dark green). On the *right*: Systematic grid of the national terrestrial inventory (1.4 km, blank circles). Source: Brändli and Denzler (2011)

4.1.2 Regional Forest Inventory in the Canton of Grisons

Terrestrial data from a regional forest inventory at the canton of Grisons were provided by the "Amt für Wald und Naturgefahren Graubünden". The regional forest inventory in the canton of Grisons has been carried out during summer 2006/2007. The terrestrial sample plots of the regional inventory are arranged in a systematic grid with a meshsize of 500 meters, where the regional sampling scheme constitutes a sub-grid of the national terrestrial inventory grid of the NFI3 (Teufen 2004). Each sample plot is identified by an individual *identification number* (*Plot ID*). The survey methods at each sample plot as well as the evaluation of the inventory data were identical with those of the NFI (Brassel and Lischke 2001) and can be found in Teufen (2004). The important difference is that the regional inventory is a simplified *one-stage* approach, where the tree height as well as the *d7* was not recorded. As a consequence, the estimates of the terrestrial timber volume on plot level (local density), which were also contained in the dataset, were calculated only based on the DBH as predictor variable (Kaufmann 2000).

4.2 Airborne Laserscanning Data

In the present study, the auxiliary information within the two-phase sampling procedures was provided by Airborne Laserscanning data which were exhaustively available over the complete study site. This chapter will give an overview of the fundamentals of Airborne Laserscanning and a detailed description about the Laserscanning data of the study site as well as required pre-processing steps.

4.2.1 Fundamentals of Airborne Laserscanning

Laser Scanning, also-called LiDAR (**Li**ght **D**etection **A**nd **R**anging), belongs to the *active* remote sensing methods. LiDAR systems which are installed on an aircraft emit laser pulses in the near-infrared (with several wavelength ranges, preferably in the region of 1040 to 1060 nm) along a line perpendicular to the line of flight with a specific pulse rate often referred as the *pulse repetition frequency* (PRF). In most cases the PRF ranges between 5000 and 50000 laser pulses per second (Heritage and Large 2009). While an emitted laser pulse now interacts with an object on the ground, a part of its energy is reflected back in the direction of the LiDAR sensor. The concept of LiDAR sensing is based on measuring the time t between the emission of a pulse and receiving its backscattered energy (called *return* or *echo*). Since a laser pulse travels at the speed of light c (appr. $3*10^8$ ms^{-1}) one can calculate the distance R (termed *Range*) between the interacting object and the LiDAR sensor based on equation 4-1 (Baltsavias 1999).

$$R = \frac{1}{2}tc \qquad 4\text{-}1$$

A prerequisite for calculating the georeferenced positions of the illuminated ground objects is to know the precise location of the LiDAR sensor as well as its orientation in terms of roll, pitch and yaw at any time of the data acquisition (Jensen 2007). This is ensured by the usage of high precision GPS technology (DGPS) and an inertial measurement unit (IMU) onboard the aircraft. The exact position can then be determined by applying the derived rotation matrices. An additional important property of the laser pulse is that its *footprint*, i.e. the area illuminated by the pulse when it reaches an object on the ground, can vary according to the flight height and the *scan angle* θ because the pulse is widening in all directions with increasing travel time. This property can be quantified based on the so-called *beam divergence* angle γ. The effective footprint (called *instantaneous laser footprint*) thus mainly depends on the flight height, the scan angle and the beam divergence and can be calculated by equation 4-2 (Baltsavias 1999). The described procedure and relations are again illustrated in Fig. 4-3 (left).

$$Fp_{inst} = \frac{h}{2\cos^2(\theta_{inst})}\gamma \qquad 4\text{-}2$$

According to these system properties, scanning the area below the LiDAR sensor results in the collection of large amount of data points which define the distance between sensor and ground objects: these are often termed *masspoints*. The point density of the laser pulses on the ground, which is one of the most important quality features of LiDAR data, thereby depends on the flight height, the flight speed, the scan angle, the PRF and the instantaneous angular scanning angle (Jensen 2007).

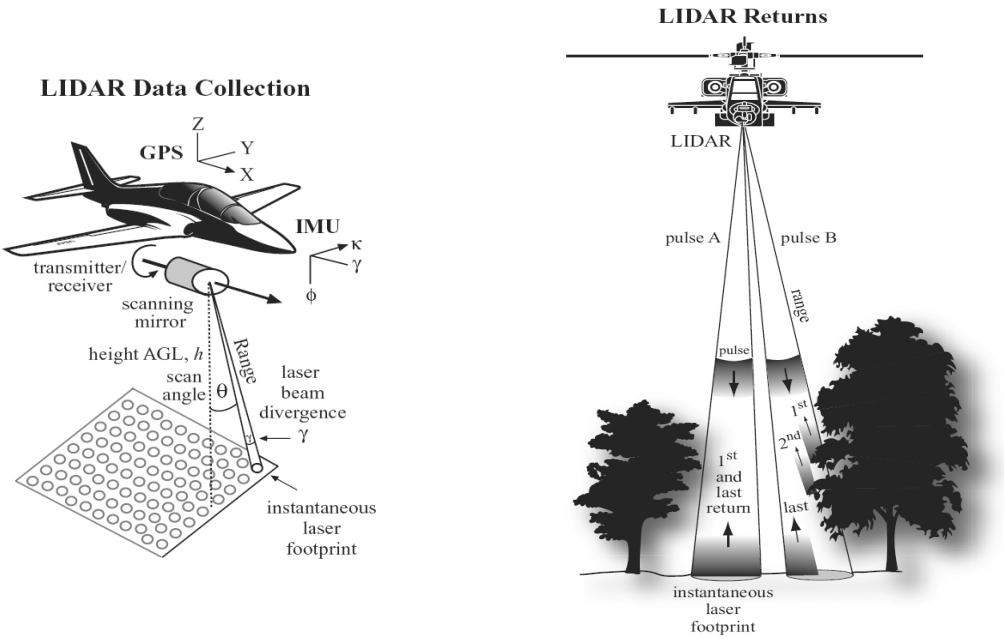


Fig. 4-3: *Left*: Illustration of the airborne LiDAR data collection. *Right*: Illustration of multiple echo returns from one emitted laser pulse. Source: Jensen (2007)

A remarkable property of most LiDAR systems is that a single pulse can create more than one return. This is the case if multiple objects are vertically distributed within the instantaneous laser footprint (shown in Fig. 4-3, right). In principle, the first and last echoes are of primary interest as they are assumed to stem from an object above the surface (e.g. a tree) and the ground itself respectively. The recording of multiple echoes results in a so-called *point cloud* which can be classified into *ground points* and *object points* (Fig. 4-4).

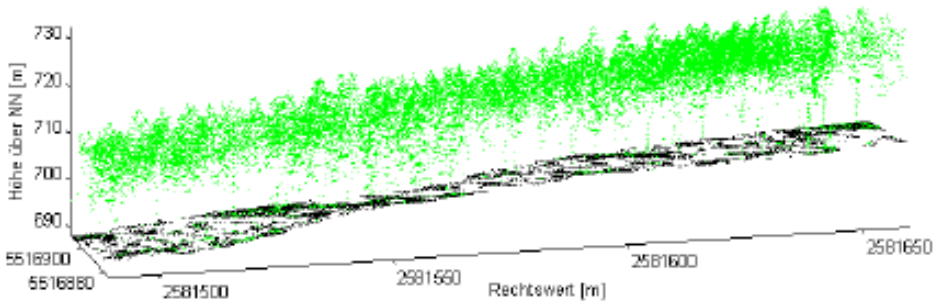


Fig. 4-4: LiDAR point cloud of a forest classified into ground points (dark green) and object points (bright green). The tree shapes are already recognizable. Source: Buddenbaum (2011)

There have been various studies in which point cloud data (especially derived by *Full Waveform* LiDAR systems) have been used in the field of forestry research, e.g. for geometric reconstruction of forest stands (Morsdorf et al. 2004). Alternatively, the point cloud data can also be converted into a raster (grid) of elevation values. This can be seen as a generalization procedure, since a raster cell then represents the statistical distribution of underlying cloud points. Preferably, this procedure aims at the creation of two raster datasets: the first one is the so-called *digital surface model* (DSM) which is derived by interpolating all first-pulse and only-pulse data and thus describes the elevation characteristics of the surface (including vegetation and man-made structures). The second raster is the so-called *digital terrain model* (DTM) which is constructed from the last- and only-pulse echoes: it can be expected to only represent the terrain surface. For generating DTMs of high accuracy, sophisticated iterative fitting algorithms have been developed (Kraus and Pfeifer 2001). However, one may still face the problem that even after classification and removal of the first echo data the last echoes do not always represent the ground. This can especially occur if the LiDAR data are acquired over dense vegetation like forests. For mid-European forests, the pulse penetration rate normally varies between 30 and 65%, which is sufficient for generating a high quality digital terrain model of the forest ground (Vosselmann and Maas 2010). The possibility of acquiring ground echoes can still be increased by the usage of larger footprints, but at the cost of spatial resolution. The ambiguous differentiation between vegetation and ground can also be amplified by large scan angles (large incidents angles) caused by steep terrain (Fig. 4-5).

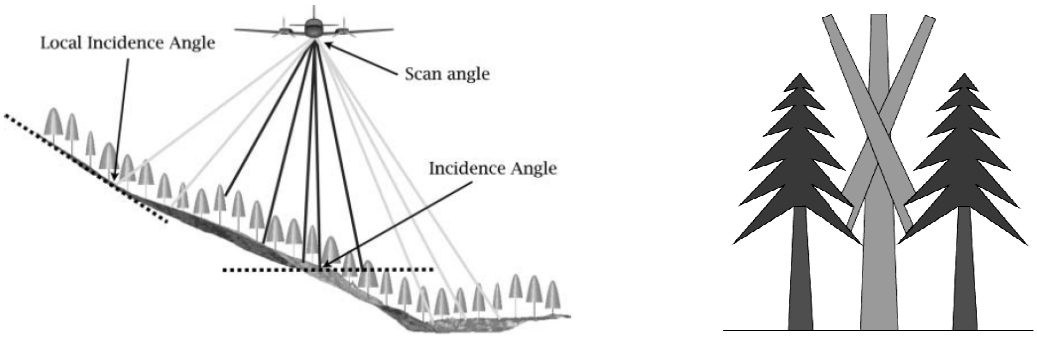


Fig. 4-5: *Left*: Illustration of scan angle and incident angle over forest in sloped terrain. *Right*: No availability of ground echoes caused by a combination of large scan angle and dense vegetation. Source: Morsdorf et al. (2008) and Morsdorf (2011)

LiDAR data are increasingly used in the field of forestry. A good overview can be found in van Leeuwen and Nieuwenhuis (2010). While it has been shown that LiDAR data (in particular the intensity of the return echoes) can even be used for tree species classification (Brandtberg 2007; Liang et al. 2007), the main application in the framework of forest inventory remains the derivation of height related structural parameters like the mean canopy height and other statistical parameters (Holmgren 2004; Næsset and Bjerknes 2001). To derive these parameters one often uses the so-called *Canopy Height Model* (CHM), often also referred to as *Vegetation Height Model* (VHM). The CHM is straightforwardly computed by subtracting the DTM from the DSM. As a result of this operation, the height information of all objects in the CHM equals their true object height because the underlying terrain elevation has been removed. Still, the so derived forest data may underlie the potential error sources described before. Morsdorf et al. (2008) found indications that the error caused by a large scan angle might be negligible for tree height estimation if the scan angle does not exceed ± 7°. In turn, they found out that an increased flight altitude of the LiDAR sensor can lead to increased underestimations of tree heights, mainly caused by a decreased point density or by less energy being reflected by the tree top. Despite their high spatial resolution due to high point densities, one should therefore always consider LiDAR-derived data still as an approximate representation of true conditions.

4.2.2 Laserscanning Data of the Study Site

LiDAR data for the study area were provided by the Remote Sensing Laboratories (RSL) of the University of Zurich. The data were acquired with a Trimble Harrier68 laser scanning system in the period between the 11th and the 15th of September 2010 at leaf-on status. The average flight height was approximately 700 m above ground. The sampling density was given as > 4 points m^{-2}. Unfortunately, no information concerning the maximum scan angle was available. The data were provided in raster format as a *digital terrain model* (DTM) and *digital surface model* (DSM) with a cell size of 0.5 m. The DTM had been generated by preferring the last echo data while the DSM had been calculated preferring the first echo data within the target raster. Outliers in the data had been eliminated by a using

threshold procedure. Gaps in the DTM had not been interpolated. After an evaluation with independent ground truth data, the provider gave the positional accuracy as < ± 50 cm and the height accuracy was given as < ± 0.15 cm. The data have already been used in forest research in the context of pre-harvest assessment and proved to be of suitable quality (Heinimann and Breschan 2012).

4.2.3 Pre-Processing Laserscanning Data

DTM and DSM constitute the data source for calculating the Canopy Height Model (CHM) for the whole study area. However, the DTM showed a considerably amount of missing height values over the whole study area. The reason for this was very likely the absence of ground returns due to a combination of dense vegetation (forest) and steep terrain. Fig. 4-6 shows that this phenomenon especially occurred within the forested areas.

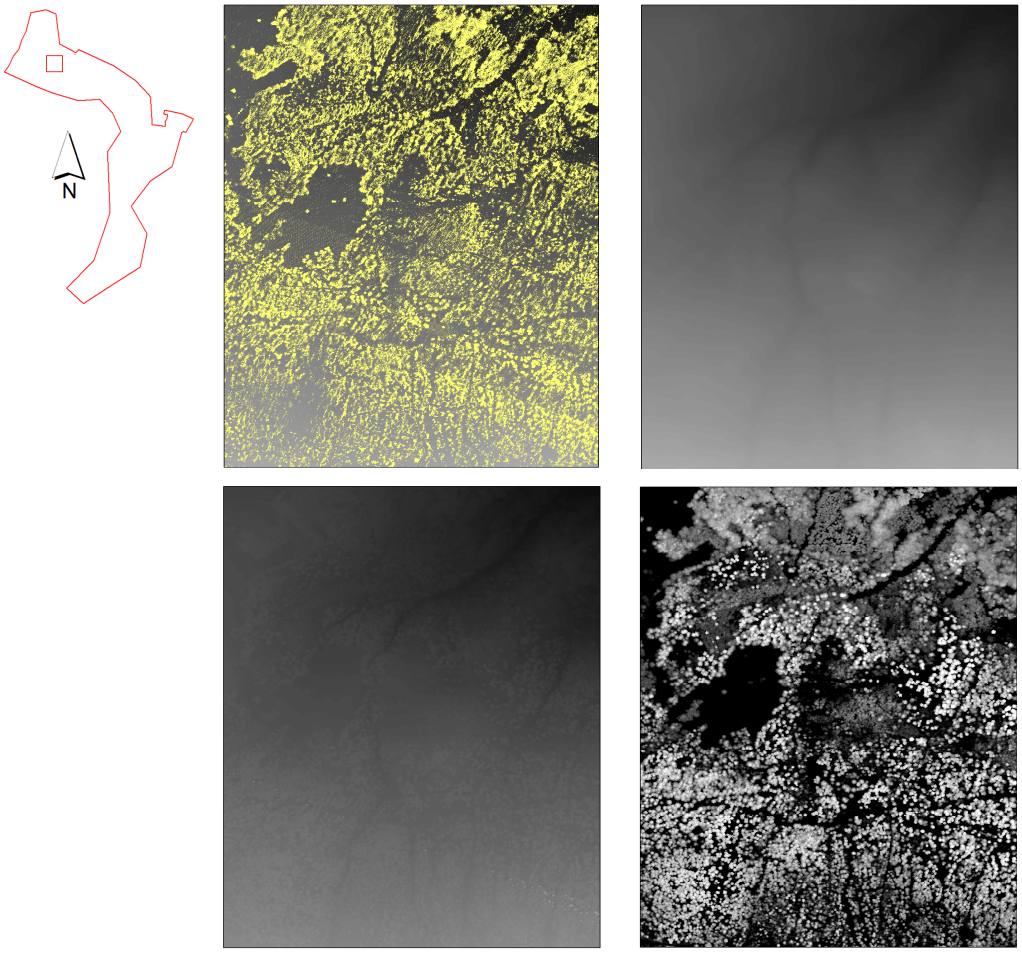


Fig. 4-6: *Upper left*: The raw DTM shows a high amount of not available ground information (yellow) especially within forest areas (compare with lower right). *Upper right*: After interpolation the DTM provides exhaustive terrain height information over the whole study area. *Lower left*: The raw DSM contains the terrain heights as well as the objects on the surface. The forests are already recognizable. *Lower right*: The Canopy Height Model (CHM) is calculated as the difference between the DSM and the DTM. It contains only the height of trees and buildings in meters above ground.

Consequently, these data gaps had first to be filled by representative height values before the calculation of the CHM could be performed. This was done by applying an inverse distance weighting interpolation algorithm (Davis 2002; Isaaks and Srivastava 1989), which assigns a weighted average of available height values from pixels within a defined neighborhood to raster cells with missing height information. The weight for an input value thereby decreases with increasing distance from the target pixel. After the interpolation process, the Canopy Height Model (CHM) was calculated as described in the previous section, i.e. the difference between the interpolated DTM and the DSM. The CHM finally contained the height of trees and buildings over the whole study area in meters above ground and thus represents the auxiliary information data for this study.

Additionally, a so-called "hillshade"-visualization of both, CHM and DTM (Fig. 4-7) was calculated using the commercial software ArcGIS (vers. 10.1). This visualization concept is particularly suited to reveal even very small height differences. In case of the CHM hillshade (Fig. 4-7, center), it is possible to visually identify fences, single trees and different developing stages of the forest stands. Also the DTM (Fig. 4-7, right) gives a very detailed representation of the terrain. It even allows the identification of ridges, channels and gullies within forest areas. This suggests that in spite of the relatively dense vegetation cover, a sufficient number of laser pulses had been able to reach the ground for providing a detailed terrain model. A visualization of the derived DTM and CHM for the complete study area can be found in Appendix A.

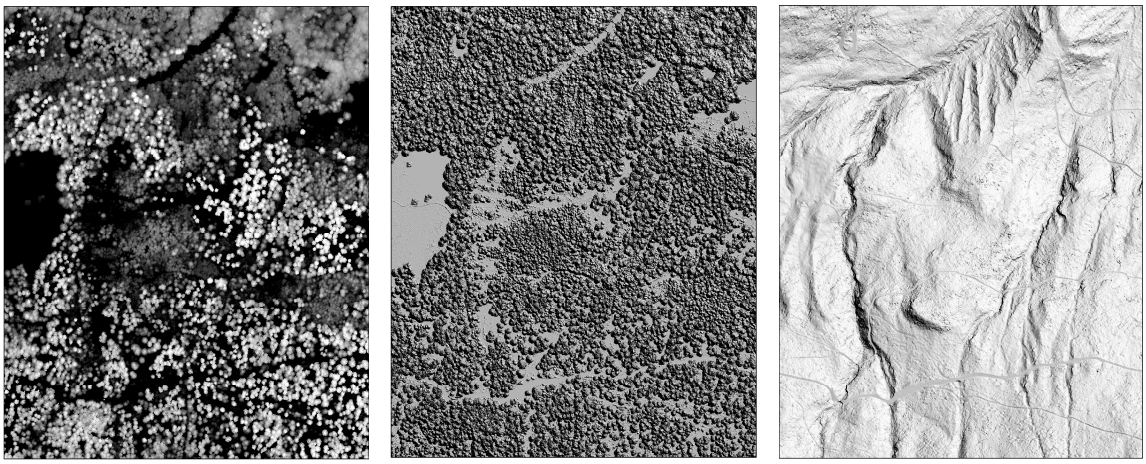


Fig. 4-7: *Left*: Part of the CHM (bright shaded regions indicate large tree heights). *Center*: Corresponding hillshade visualization of the CHM. *Right*: Corresponding underlying DTM (hillshade visualization) on which the CHM is based on.

4.3 Sampling Concepts of One- and Two-Phase Sampling

The following section will present all sampling and estimations methods which were applied to the study area and finally compared to each other. Before the methods of two-phase sampling will be introduced, the principle of one-phase-sampling in forest inventories will shortly be presented. For practical reasons, the terminology will be the same as used by Mandallaz (2008, 2012, 2013a, 2013b).

In forest inventory, one phase and two-phase sampling is used for estimating the quantity of specific forest parameters such as timber volume or basal area for a defined forest area *F*. In one-phase-sampling, one collects terrestrial information of interest at a set s_2 comprising n_2 sample points *x* within the forest area *F*. These sample points are randomly, uniformly and independently distributed within the forest area *F*. At each sample point *x*, trees from the defined and thus known population *P* (whose number of individuals *N* is unknown) are selected within a circle $K_r(x)$ with constant radius *r* and its center in *x*. Consequently, one can define a random indicator variable $I_i(x)$ for the *i*th tree of *P* at sample point *x* as

$$I_i(x) = \begin{cases} 1 \; if \; i \in K_r(x) \\ 0 \; if \; i \notin K_r(x) \end{cases} \qquad 4\text{-}3$$

According to the so-called *duality principle*, setting up a circle $K_r(x)$ with constant radius *r* around each sample point *x* is *equivalent* to imaging *N* circles $K_i(r) = K_r(u_i)$ with constant radius *r* being created around the trees of *P* at the center of the tree u_i. It follows that the *i*th tree lies within the circle $K_r(x)$ only if the sample point *x* lies within the circle $K_r(u_i)$. This duality principle can be written as

$$I_i(x) = 1 \Leftrightarrow x \in K_i(r) \qquad 4\text{-}4$$

One can thus calculate the *inclusion probability* π of the *i*th tree being included in the sample at point *x* by using the area λ of the circle $K_i(r)$ and the forest area $\lambda(F)$:

$$\pi_i = \mathrm{P}(I_i(x) = 1) = \mathrm{E}_x\big(I_i(x)\big) = \frac{\lambda(K_i(r) \cap \mathrm{F})}{\lambda(\mathrm{F})} \qquad 4\text{-}5$$

If the circle area of $K_i(r)$ (or $K_r(x)$ respectively) is not completely covered by the forest area *F*, one has to perform a so-called *boundary adjustment*, which means that the circle area has to be shrinked to the part lying in *F* (indicated by the numerator in equation 4-5). It has to be noted that neglecting boundary adjustments leads to underestimate the local density $Y(x)$ (equation 4-6).

It should further be mentioned that in most of the cases the radius r of $K_i(r)$ is *not* constant, but a function of an auxiliary variable related to the i^{th} tree. This leads to a generalized formulation of equation 4-5 where $K_i(r)$ has to be replaced by $K_i(r_i)$ being a circle centered on the i^{th} tree with radius r_i. For example, the radius can be a function of the tree DBH, where trees with a bigger diameter get a higher probability to be included into the sample. In the extreme case, an individual radius is thereby assigned to each tree. This procedure aims at approximating the so-called *Probability Proportional to Size* (PPS) sampling, which can be regarded as the optimal sampling scheme in one-phase sampling (Mandallaz 2008). In practice, PPS for the basal area is only implemented by the well known *angle count sampling* (Bitterlich 1984), where the inclusion probability is proportional to the square of the DBH. Strictly speaking, this is even only true if one assumes visibility of all trees of the population *P* (Ritter et al. 2013). In practice, PPS is often approximated by using a certain amount of *concentric circles*. Here, trees are included into the sample if they lie within one of the circles and if their diameter matches a certain threshold of the corresponding circle. It is in any case essential to perform a so-called *slope correction* if sample point *x*, i.e. the sample plot, is located on a slope. The projection of a circle created on a slope onto the horizontal plane will result in an ellipse whose area will be larger than the correct nominal surface area of the circle with radius r_i. One must therefore correct the radius in the slope by the *slope correction factor* $r_{i,slope} = \frac{r_i}{\cos(\beta)}$ to avoid miscalculation of the inclusion probability (see equation 4-5). In practice, the slope correction is performed at plot level and not on tree level.

For a given sampling scheme, the local density $Y(x)$ at each sample point *x* for an arbitrary response variable *Y* is defined according to the *Horwitz-Thompson Estimator* (equation 4-6) as

$$Y(x) = \frac{1}{\lambda(F)} \sum_{i=1}^{N} \frac{I_i(x) Y_i}{\pi_i} = \frac{1}{\lambda(F)} \sum_{i \in s_2(x)}^{N} \frac{Y_i}{\pi_i} \qquad 4\text{-}6$$

where $s_2(x) = \{i | I_i(x) = 1\}$. Having derived the local density $Y(x)$ at all s_2 sample points, one can calculate the *sample mean* as the arithmetic mean of all $Y(x)$ in s_2:

$$\hat{Y} = \frac{1}{n_2} \sum_{x \in s_2} Y(x) \qquad 4\text{-}7$$

The unbiased variance estimate of $\hat{Y}$ can be obtained by

$$\hat{V}(\hat{Y}) = \frac{1}{n_2} \frac{1}{n_2 - 1} \sum_{x \in s_2} \left(Y(x) - \hat{Y}\right)^2 \qquad 4\text{-}8$$

The estimated *Total* for the domain *F* can then be derived by just multiplying the sample mean by the forest area:

$$\hat{Y}_{Total} = \lambda(F)\,\hat{Y} \qquad 4\text{-}9$$

The local density $Y(x)$ is a function which is sampled to estimate the spatial mean $\bar{Y}_F$ within a given forest area. Because $\mathrm{E}_x\big(I_i(x)\big) = \pi_i$ (equation 4-5), one gets

$$\mathbb{E}_x\big(Y(x)\big) = \frac{1}{\lambda(F)}\sum_{i=1} Y_i = \bar{Y}_F \qquad 4\text{-}10$$

The above equation 4-10 (term in the middle) describes the estimation of a sum over a *finite* population *P* of *N* trees. However, this formulation reveals certain drawbacks: for instance, it is obviously impossible to achieve a census, i.e. to partition the forest area into a finite population of sample units with perfect tessellation, by using constant or concentric circles. This is the reason why one often uses the *infinite population approach*, which transforms the formulation of equation 4-10 into one of estimating the *integral* of a function over the domain *F* (see equation 4-11). This approach allows for viewing $\mathrm{Y}(x), x \in F$ as an infinite population (Mandallaz 2008) and is also known as the *Monte Carlo approach*. It yields the same properties as shown in equation 4-10, but is better suited for description of a forest inventory:

$$\mathbb{E}_x\big(Y(x)\big) = \frac{1}{\lambda(F)}\sum_{i=1} Y_i = \frac{1}{\lambda(F)}\int_F Y(x)\,dx = \bar{Y}_F \qquad 4\text{-}11$$

$Y(x)$ and therefore $\hat{Y}$ are unbiased estimates of $\bar{Y}_F$, which means that the expected value of the estimates is equal to the true value. The property of making unbiased estimates is here based on the random mechanism which creates the sample points in s_2: since the sample points *x* are randomly, uniformly and independently distributed within *F*, the indicator variable $I_i(x)$ and conclusively the local density $Y(x)$ at point *x* is a random variable, whereas the forest, in particular the tree population *P*, the number *N* of the tree individuals in *P* as well as the response variable Y_i is fixed. This conception is known as *design-based*. Alternatively, one can also consider the *model-dependent* approach (sometimes also-called *model-based*), where the forest is viewed as a realization of a complex stochastic process, and hence the number of trees as well as the response variable Y_i is also a random variable. An important feature of the *design-based* approach is that the inclusion probabilities are known for all trees included in the sample (equation 4-5).

In practice, instead of randomly distributed points one often uses a systematic grid of sample points that have random start and orientation (known as systematic random sampling) for practical reasons. Doing so, in the strict sense one cannot calculate the *design-based* variance anymore, since a systematic grid violates the condition that the pairwise inclusion probabilities π_{ij} for tree *i* and *j* being selected in the same sample is never zero (Mandallaz 2008). Still one treats these systematic points as randomly chosen and proceeds with the estimators based on simple random sampling. It has been shown that this is acceptable for point estimates, but in practice one generally overestimates the variance (Mandallaz 2008). However, theoretically also underestimation is possible. The reason for the over-or underestimation is that a systematic grid can be seen as one single large cluster where the response variables of adjacent sample points are correlated with each other. As this correlation has a multiplicative impact on the variance (intra-cluster correlation coefficient) one cannot strictly speaking derive the *design-based* variance using systematic grids. However, with overestimating the variance one is "on the safe side" and in practice, the corresponding bias seems to be negligible (Mandallaz 2008).

The method of two-phase sampling is regarded as an extension of the one-phase approach. It is based on the fact that deriving terrestrial information of high precision is time- as well as cost-expensive, and therefore the number of terrestrial sample plots is limited. The principle idea of two-phase sampling is that the terrestrial information, whose response variables are assumed to be very precise, can be combined with additional, so-called *auxiliary information* which is less precise as the terrestrial information, but available at broader range and reduced costs. By thus increasing the effective sample size, the two-phase method aims at decreasing the estimated variance of the point estimate, leading to more precise estimations. In practice, the auxiliary information is often derived by remote sensing data of qualitative nature (e.g. stand maps based on visual interpretation) or quantitative nature (e.g. timber volume estimated by algorithms).

The name "two-phase sampling" refers to the collection of inventory data within two phases: In the *first phase*, a *large* sample s_1 of n_1 points $x_i \in s_1$ are randomly, uniformly and independently distributed within the forest area *F*. At each of those points, the auxiliary information is derived. The *second phase* comprises a *small* sample $s_2 \subset s_1$ of n_2 points which are drawn according to equal probability sampling without replacement. At each sample point $x \in s_2$ the local density $Y(x)$ (equation 4-6) is derived by terrestrial field surveys. It is essential to see that at those points $x \in s_2$ one has the derived terrestrial information *and* the auxiliary information, since s_2 is a subsample of s_1. This property now allows for quantifying the relationship between the auxiliary variable(s) and the corresponding terrestrial response variable using all information collected at the s_2 sample points. This is often done by performing a simple or multiple linear regression analysis between $Y(x)$ as the response variable and the auxiliary variable(s) as predictor variable(s) by means of classical statistics. The resulting regression coefficient(s) of the regression model is (are) then used for *predicting* the local density $\hat{Y}(x)$ at the sample points $x \in s_1$ (thereby also at sample points where the terrestrial information is not available). Estimators using these regression-based predictions are hence called *regression estimators*.

With respect to the regression models used for predictions, a crucial distinction has to be made between the *external model* approach and the *internal model* approach: Within the external model approach, the model used for predictions is not adjusted to the local inventory data. For example, this is the case if the regression model, in particular the regression coefficient, is obtained from another inventory not located in the current inventory area. In this case, the regression coefficients have given fixed values. On the contrary, within internal models the model is fitted with the data provided by the current inventory and can therefore be regarded as adjusted to the local conditions. According to Mandallaz (2013a) one can neglect the error variance in the regression coefficients in very large samples, since in this case internal models can be treated as external ones.

The general principle of two-phase sampling presented so far and also applied in the present study belongs to the so-called *model-assisted* approach which is a special case of the *design-based* approach. Even under the use of regression models within two-phase sampling, the *validity* of the derived estimates is ensured by the *design-based* inventory scheme, i.e. by the randomization principle of the sample point, and ***not*** by the validity of the regression model. Thus the regression models do ***not*** have to satisfy any model assumptions as in classical statistics (e.g. constant error variance, uncorrelated errors etc.). It is of great importance to note that the regression models are used to *improve* the efficiency of the estimator (in terms of reducing the estimated variance of the point estimates), but they do ***not*** have to be *correct*. The regression models should however always be applicable from a *logical* point of view. This approach is in contrast to the *model-dependent* approach, where the model predictions are assumed to be correct (Mandallaz 2008).

4.4 Global Regression Estimators

In the following, all *global* regression estimators for the entire forest area *F* which were applied and compared in this study will be introduced and described. In practice, also in two-phase sampling most often a systematic grid is used instead of randomly distributed sample points. In this case, the sample s_2 is a subgrid of the large sample s_1. Again, the following design-based regression estimators derived under simple random sampling conditions are still applied using systematic grids.

4.4.1 Regression Estimator

First, the classical regression estimator will be introduced which has been used in various studies and inventories. The *linear regression model* using the terrestrial response variable and the auxiliary information given at the sample points $x \in s_2$ using ordinary least square regression (equation 4-12) is considered to be:

$$Y(x) = Z^t(x)\beta + R(x) \qquad 4\text{-}12$$

with $Z(x) \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^p$. For this purpose the auxiliary variables are coded into the vector $Z(x)$ including all auxiliary variables derived at sample point *x*. Consequently, $Z^t(x)$ is the transposed vector of auxiliary variables. Since $Z(x)$ has an intercept term, it is insured that the *mean residual* is *zero*, i.e. $\int_F R(x)\,dx = 0$ (Mandallaz 2013a). This is an important characteristic to which further attention will be paid when small area estimation is discussed (chapter 4.5).

One further defines

$$A_{s_2} = \frac{1}{n_2} \sum_{x \in s_2} Z(x)\, Z^t(x)$$

and

$$U_{s_2} = \frac{1}{n_2} \sum_{x \in s_2} Y(x)\, Z(x)$$

to formulate the sample-based normal equation as

$$A_{s_2} \hat{\beta}_{s_2} = U_{s_2}$$

One gets the asymptotical design-unbiased estimate $\hat{\beta}_{s_2}$ for the true regression parameter vector β by solving the normal equation for $\hat{\beta}_{s_2}$ (equation 4-13), which is thus the theoretical least square estimate found by minimizing the residual sum of squares of the regression model.

$$\hat{\beta}_{s_2} = A_{s_2}^{-1} U_{s_2} \qquad 4\text{-}13$$

The estimated regression coefficients contained in vector $\hat{\beta}_{s_2}$ can now be used to predict the local density $\hat{Y}(x)$ at any sample point *x* by

$$\hat{Y}(x) = Z^t(x) \hat{\beta}_{s_2} \qquad 4\text{-}14$$

The empirical model residuals $\hat{R}(x)$ can be calculated as the difference between the terrestrial local density $Y(x)$ and the predicted local density $\hat{Y}(x)$:

$$\hat{R}(x) = Y(x) - \hat{Y}(x) \qquad 4\text{-}15$$

The classical ***regression estimator*** for the entire domain *F* is then defined as

$$\hat{Y}_{F,reg} = \frac{1}{n_1}\sum_{x\in s_1}\hat{Y}(x) + \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x) =: \hat{\bar{Z}}_1^t\hat{\beta}_{s_2} \qquad 4\text{-}16$$

with $\hat{\bar{Z}}_1 = \frac{1}{n_1}\sum_{x\in s_1} Z(x)$. The *point estimate* $\hat{Y}_{F,reg}$ can be interpreted as simply the mean of all predictions over the s_1 sample *corrected* by the mean model residual over the s_2 sample. Since $\hat{\beta}_{s_2}$ is estimated by ordinary least square regression using current inventory data, it is insured that the zero mean residual assumption holds true for $\frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x)$. Consequently, the residual term in equation 4-16 vanishes and the point estimate is purely based on the model predictions.

The ***estimated external variance*** of $\hat{Y}_{F,reg}$ can now be calculated under the *external* model assumption, where the variance of the regression coefficients of $\hat{\beta}_{s_2}$ is neglected:

$$\hat{V}\left(\hat{Y}_{ext,F,reg}\right) = \frac{1}{n_1}\frac{1}{n_2-1}\sum_{x\in s_2}(Y(x)-\bar{Y}_2)^2 + \left(1-\frac{n_2}{n_1}\right)\frac{1}{n_2}\frac{1}{n_2-1}\sum_{x\in s_2}\left(\hat{R}(x)-\hat{\bar{R}}_2\right)^2 \qquad 4\text{-}17$$

with $\bar{Y}_2 = \frac{1}{n_2}\sum_{x\in s_2} Y(x)$ being the *sample mean* and $\hat{\bar{R}}_2 = \frac{1}{n_2}\sum_{x\in s_2}\hat{R}(x) = 0$. Using Taylor linearization technique, Mandallaz (2013a, 2013b) proposed a better estimation of the *design-based variance* which also explicitly takes the variance in the regression coefficients at least asymptotically into account. The ***estimated design-based variance*** of $\hat{Y}_{F,reg}$ is thus calculated by

$$\hat{V}\left(\hat{Y}_{F,reg}\right) := \hat{\bar{Z}}_1^t\hat{\Sigma}_{\hat{\beta}_{s_2}}\hat{\bar{Z}}_1 + \hat{\beta}^t{}_{s_2}\hat{\Sigma}_{\hat{\bar{Z}}_1}\hat{\beta}_{s_2} \qquad 4\text{-}18$$

and the estimated design-based covariance matrix of $\hat{\beta}_{s_2}$ is given by Mandallaz (2008) as:

$$\hat{\Sigma}_{\hat{\beta}_{s_2}} := A_{s_2}^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2(x)Z(x)Z^t(x)\right)A_{s_2}^{-1} \qquad 4\text{-}19$$

Interestingly, it can be shown that this is the same as the *robust* covariance matrix in the *model-dependent* ordinary least square theory described by Huber (1967) and Gregoire and Dyer (1989).

The estimated covariance matrix of $\hat{\hat{Z}}_1$ is given by

$$\hat{\Sigma}_{\hat{\hat{Z}}_1} = \frac{1}{n_1(n_1-1)} \sum_{x \in s_1} \left(Z(x) - \hat{\hat{Z}}_1\right)\left(Z(x) - \hat{\hat{Z}}_1\right)^t \qquad 4\text{-}20$$

One can express the ***estimated design-based variance*** $\hat{V}\left(\hat{Y}_{F,reg}\right)$ (equation 4-18) also using the so-called *g-weights* technique (Mandallaz 2008).

$$\hat{V}\left(\hat{Y}_{F,reg}\right) = \frac{1}{n_2^2} \sum_{x \in s_2} g_1^2(x)\hat{R}^2(x) + \frac{1}{n_1(n_1-1)} \sum_{x \in s_1} \left(\hat{Y}(x) - \bar{\hat{Y}}_1\right)^2 \qquad 4\text{-}21$$

where $\bar{\hat{Y}}_1 = \frac{1}{n_1}\sum_{x \in s_1} \hat{Y}(x)$ is the mean of all predictions $\hat{Y}(x)$ in the large sample s_1. The second term of the equation is the variance of the predictions over the domain *F*. The first term now uses the g-weights for weighting the square of the empirical residuals in s_2. The g-weights can be calculated for each sample point *x* and are defined as

$$g_1(x) = \hat{\hat{Z}}_1^t A_{s_2}^{-1} Z(x) \qquad 4\text{-}22$$

Where equation 4-21 is perfectly equivalent to equation 4-17, the g-weights provide better statistical properties especially in the case of post-stratification (Mandallaz 2008).

4.4.2 Generalized Regression Estimator

The auxiliary information is in practice often derived by remote sensing data. The generalized regression estimator proposed by Mandallaz (2013b) considers the fact that the remote sensing data are often not only available at the sample points of the large sample s_1, but are *exhaustively* available over the whole forest area *F*. The proposed generalized regression estimator takes advantage of this fact as these auxiliary variables which can be gained or computed without great expense (hence called *cost-effective*) are exhaustively derived. The auxiliary vector $Z^t(x)$ is therefore partitioned into two components, each including an intercept term:

$$Z^t(x) = \left(Z^{(1)t}(x), Z^{(2)t}(x)\right)$$

where the first one $Z^{(1)t}(x)$ contains the auxiliary variables derived exhaustively in *F*. One considers the so-called ***large model*** which uses all auxiliary variables (exhaustive and non-exhaustive ones) to derive the vector β containing the regression coefficients to explain the terrestrial local density $Y(x)$:

$$Y(x) = Z^t(x)\beta + R(x) = Z^{(1)t}(x)\beta^{(1)} + Z^{(2)t}(x)\beta^{(2)} + R(x) \quad \text{4-23}$$

One further defines the ***reduced model*** in which only the exhaustive variables are used to calculate the vector α containing the regression coefficients to explain the local density:

$$Y(x) = Z^{(1)t}(x)\alpha + R_1(x) \quad \text{4-24}$$

Again, the true regression coefficients of both models remain unknown and can only be estimated by $\hat{\beta}_2$ and $\hat{\alpha}_2$ using the sample points of the small sample s_2 (where the terrestrial local density is available). They are the ordinary least square estimates minimizing the residual sum of squares of their models and can be written as

$$\begin{aligned}\hat{\beta}_2 &= \left(\frac{1}{n_2}\sum_{x\in s_2} Z(x)\,Z^t(x)\right)^{-1} \frac{1}{n_2}\sum_{x\in s_2} Y(x)Z(x) := A_2^{-1}\frac{1}{n_2}\sum_{x\in s_2} Y(x)\,Z(x)\\ \hat{\alpha}_2 &= \left(\frac{1}{n_2}\sum_{x\in s_2} Z^{(1)}(x)\,Z^{(1)t}(x)\right)^{-1} \frac{1}{n_2}\sum_{x\in s_2} Y(x)Z^{(1)}(x)\\ &:= \left(A_2^{(1)}\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2} Y(x)\,Z^{(1)}(x)\end{aligned} \quad \text{4-25}$$

The large model yields the predictions $\hat{Y}(x) = Z^t(x)\hat{\beta}_2$ and the reduced model the predictions $\hat{Y}_1(x) = Z^{(1)t}(x)\hat{\alpha}_2$. The ***generalized regression estimate*** is then defined as

$$\hat{Y}_{F,greg} = \frac{1}{\lambda(F)}\int_F \hat{Y}_1(x)\,dx + \frac{1}{n_1}\sum_{x\in s_1}\left(\hat{Y}(x) - \hat{Y}_1(x)\right) + \frac{1}{n_2}\sum_{x\in s_2}\left(Y(x) - \hat{Y}(x)\right) \quad \text{4-26}$$

The interpretation of the point estimate $\hat{Y}_{F,greg}$ is as follows: The first term represents the spatial average of the predicted local densities which are exhaustively derived in F using the reduced model. Since we are still in the *model assisted* framework, the predictions and hence the average of $\hat{Y}_1(x)$ are not assumed to be correct, and so are the predictions $\hat{Y}(x)$ of the large model. However, the predictions $\hat{Y}_1(x)$ of the reduced model are assumed to be *less* accurate as the predictions $\hat{Y}(x)$ of the large model. This is because the large model uses the full range of auxiliary variables including the more *cost-intensive* ones (in terms of computing effort) which are assumed to significantly improve the fit of the regression model. On the other hand, the predicted local density $\hat{Y}(x)$ of the large model is again

considered to be *less* accurate than the terrestrial derived local density $Y(x)$. The second term is the mean difference between the predictions $\hat{Y}_1(x)$ of the large and $\hat{Y}(x)$ of the reduced model which are observable only in $x \in s_1$. One can further correct the predictions $\hat{Y}(x)$ of the large model by the terrestrial local density $Y(x)$. This is done in the third term of equation 4-26 which calculates the mean difference between $Y(x)$ and $\hat{Y}(x)$ and corresponds exactly to the mean residual of the large model. It is now obvious that the proposed generalized regression estimator (equation 4-26) is an extension to the regression estimator described in chapter 4.4.1.

Having described the principle of the generalized regression estimator, $\hat{Y}_{F,greg}$ can be rewritten after defining the following *mean values*:

$$\bar{Z}^{(1)} = \frac{1}{\lambda(F)} \int_F Z^{(1)}(x)\, dx, \qquad \hat{\bar{Z}}_1^{(1)} = \frac{1}{n_1} \sum_{x \in s_1} Z^{(1)}(x), \qquad \hat{\bar{Z}}_k = \frac{1}{n_k} \sum_{x \in s_1} Z(x)\,, k = 1{,}2 \tag{4-27}$$

where $\bar{Z}^{(1)}$ comprises the true means of all exhaustive auxiliary variables, $\hat{\bar{Z}}_1^{(1)}$ comprises the estimated means of all exhaustive variables at $x \in s_1$, $\hat{\bar{Z}}_1$ comprises the estimated means of all non-exhaustive variables at $x \in s_1$ and $\hat{\bar{Z}}_2$ comprises the estimated means of all non-exhaustive variables at $x \in s_2$. Now $\hat{Y}_{F,greg}$ can be rewritten as:

$$\begin{aligned} \hat{Y}_{F,greg} &= \left(\bar{Z}^{(1)} - \hat{\bar{Z}}_1^{(1)}\right)^t \hat{\alpha}_2 + \left(\hat{\bar{Z}}_1 - \hat{\bar{Z}}_2\right)^t \hat{\beta}_2 + \frac{1}{n_2} \sum_{x \in s_2} Y(x) \\ &= \left(\bar{Z}^{(1)} - \hat{\bar{Z}}_1^{(1)}\right)^t \hat{\alpha}_2 + \hat{\bar{Z}}_1^t \hat{\beta}_2 + \frac{1}{n_2} \sum_{x \in s_2} Y(x) - \hat{Y}(x) \\ &= \left(\bar{Z}^{(1)} - \hat{\bar{Z}}_1^{(1)}\right)^t \hat{\alpha}_2 + \hat{\bar{Z}}_1^t \hat{\beta}_2 \end{aligned} \tag{4-28}$$

$\frac{1}{n_2} \sum_{x \in s_2} \left(Y(x) - \hat{Y}(x)\right) = 0$ follows from the fact that the sum of the residuals are zero. Compared to equation 4-26, the generalized regression estimator $\hat{Y}_{F,greg}$ now becomes much more intuitive: The first term can be interpreted as a *correction* or *residual term*, where the *estimated* means of all exhaustive variables are corrected by their *true* mean values. The corresponding difference is then converted into a residual of the predicted local density of the reduced model by multiplication with the regression coefficients of $\hat{\alpha}_2$. This residual is then added to the second term of equation 4-28 (last line) which describes the predicted mean local density calculated by the large model. It now becomes obvious that the exhaustive variables in the large model are corrected using their true mean values.

The ***estimated variance*** of $\hat{Y}_{F,greg}$ can now be calculated under the *external* model assumption for both models (i.e. the variance of the regression coefficients in $\hat{\beta}_2$ and $\hat{\alpha}_2$ are neglected):

$$\hat{V}\left(\hat{Y}_{ext,F,greg}\right) = \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}\hat{R}_1^2\,(x) + \frac{1}{n_2}\left(1-\frac{n_2}{n_1}\right)\frac{1}{n_2^2}\sum_{x\in s_2}\hat{R}^2\,(x) \qquad 4\text{-}29$$

$\hat{R}_1(x) = Y(x) - \hat{Y}_1(x) = Y(x) - Z^{1(t)}(x)\hat{\alpha}_2$ are the empirical residuals between the terrestrial local density and the predicted local density from the *reduced* model. $\hat{R}(x) = Y(x) - \hat{Y}(x) = Y(x) - Z^t(x)\hat{\beta}_2$ are the empirical residuals between the terrestrial local density and the predicted local density from the *large* model. Both residual types are only observable at points $x \in s_2$ and suffice the zero mean residual assumption over *F*. According to Mandallaz (2013b) one can again derive a better estimation of the variance by using the g-weights, which are defined as

$$\begin{aligned} g_2(x) &= \hat{\bar{Z}}_1^t A_{s_2}^{-1} Z(x) \\ g_1^{(1)}(x) &= \bar{Z}^{1(t)}(A_{s_1}^{(1)})^{-1} Z^{(1)}(x) \end{aligned} \qquad 4\text{-}30$$

In equation 4-31, the estimated ***design-based variance*** of $\hat{V}\left(\hat{Y}_{F,greg}\right)$ is only given in its g-weights form. $\hat{V}\left(\hat{Y}_{F,greg}\right)$ can also be calculated using the covariance-matrix of $\hat{\beta}_2$ and $\hat{\alpha}_2$ (see Mandallaz (2013b).

$$\hat{V}\left(\hat{Y}_{F,greg}\right) = \frac{1}{n_1}\frac{1}{n_2}\sum_{x\in s_2}(g_1^{(1)}(x))^2\,\hat{R}_1^2\,(x) + \left(1-\frac{n_2}{n_1}\right)\frac{1}{n_2^2}\sum_{x\in s_2} g_2^2(x)\,\hat{R}^2\,(x) \qquad 4\text{-}31$$

4.4.3 Calculation of Confidence Intervals

For the global estimations, the two-sided $1-\frac{\alpha}{2}$ confidence interval must be based on the Student *t*-distribution with $n_2 - p$ degrees of freedom, where p is the number of parameters used in the estimator (equation 4-32). For the one-phase approach, the degrees of freedom are defined as $n_2 - 1$. Simulation results in Mandallaz (2013a) revealed that using $n_2 - 2p$ can lead to an even better achievement of the required 95% level for α=5%. However, for large samples such as within national inventories, one can rely on the central limit theorem and obtain the $1-\frac{\alpha}{2}$ confidence interval according to the normal distribution ($z_{1-\frac{\alpha}{2}}$). It should be noted that for large sample sizes, the t- or z-values and the corre-

sponding confidence interval will be small since n_2 will not considerably be reduced by subtracting the number of parameters.

$$CI_{1-\alpha}(\hat{Y}_F) = \left[\hat{Y}_F - t_{n_{2-p},1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_F)}\,, \hat{Y}_F + t_{n_{2-p},1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_F)}\right] \quad 4\text{-}32$$

4.5 Small Area Estimators

In this section, it is demonstrated how the global regression estimators described in chapter 4.4 can be used in the framework of *small area estimations*. The principle idea of small area estimation in forest inventory is to provide estimations of acceptable accuracy for forest areas, where the number of observable terrestrial plots is considerably "small". This is often the case in forest inventories at the enterprise level, where terrestrial data from regional inventories are used (Mandallaz 2008). The technique of small area estimation has gradually gained importance due to the high demand of cost-reduction of terrestrial field surveys. All subsequent estimators were applied and compared within this study.

4.5.1 Small Area Regression Estimator

One considers the small area $G \subset F$ (i.e. G lies in the forest area F). The external ***small area regression estimator*** can be formulated as

$$\hat{Y}_{ext,G,reg} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}}\hat{Y}(x) + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}(x) = \hat{\bar{Z}}^{t}_{1,G}\hat{\beta}_{s_2} + \frac{1}{n_{2,G}}\sum_{x\in s_{2,G}}\hat{R}(x) \quad 4\text{-}33$$

where

$$\hat{\bar{Z}}_{1,G} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}} Z(x) \quad 4\text{-}34$$

The corresponding estimated ***external variance*** is

$$\begin{aligned}\hat{V}\left(\hat{Y}_{ext,G,reg}\right) = {} & \frac{1}{n_{1,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_{2,G}}\left(Y(x) - \bar{Y}_{2,G}\right)^2 \\ & + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right)\frac{1}{n_{2,G}}\frac{1}{n_{2,G}-1}\sum_{x\in s_2,G}\left(\hat{R}(x) - \hat{\bar{R}}_{2,G}\right)^2\end{aligned} \quad 4\text{-}35$$

Obviously, $\hat{Y}_{ext,G,reg}$ is very similar to the regression estimator $\hat{Y}_{F,reg}$ introduced in chapter 4.4.1 (equation 4-16). The difference is that for calculating the point estimate and the estimated variance, one only uses the auxiliary variables of the sample points $x \in s_{1,G}$ and the terrestrial information only at sample points $x \in s_{2,G}$ (i.e. points of the large sample s_1 and the small sample s_2 which are located in the small area G). It is essential to see that one however still uses the regression coefficients of $\hat{\beta}_{s_2}$ which have been derived using all data at sample points $x \in s_2$ in the entire forest area F. Since the regression coefficients of $\hat{\beta}_{s_2}$ are hence not adjusted to the local inventory conditions of the small area (*external* model approach), the mean of the residuals $\frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} \hat{R}(x)$ over the small area $G \subset F$ are no longer zero in general. A regression estimator which does not account for this problem is the so-called *synthetic* or *pseudo-synthetic estimator* (Mandallaz 2008) . Although this estimator will in general have a small variance, its estimates are potentially biased.

Mandallaz (2012, 2013a) proposed a new technique to have the zero mean residual assumptions fulfilled in the whole inventory area F as well as in any arbitrary small area G within F by extending the auxiliary information vector $Z(x)$ with an indicator variable $I_G(x)$. This indicator variable indicates the small area of which the corresponding auxiliary information in $Z(x)$ belongs to. The extended auxiliary vector is given by $\mathcal{Z}^t(x) = \left(Z^t(x), I_G(x)\right)$ where

$$I_G(x) = \begin{cases} 1 \; if \; x \in G \\ 0 \; if \; x \notin G \end{cases} \qquad 4\text{-}36$$

Considering the case of more than one small area within F, $Z(x)$ can be extended by an arbitrary number of small area indicators. In the following section we will just consider the case of one small area, i.e. one indicator variable. The method can however be applied to an arbitrary number of small area indicators. Using the extended auxiliary vector $\mathcal{Z}(x)$, the regression coefficients of $\hat{\theta}_{s_2}$ can be estimated by

$$\hat{\theta}_{s_2} = \mathcal{A}_{s_2}^{-1} \left(\sum_{x \in s_2} Y(x) \mathcal{Z}(x) \right) \qquad 4\text{-}37$$

where $\mathcal{A}_{s_2} = \frac{1}{n_2} \sum_{x \in s_2} \mathcal{Z}(x) \, \mathcal{Z}^t(x)$. With $\mathcal{Z}(\mathrm{x})$ containing an intercept and the indicator variable $I_G(x)$, it is hence insured that the mean of the residuals in the whole inventory area F as well as in any small area $G \subset F$ is zero. The mathematical proof can be found in Mandallaz (2012, 2013a). It essential to note that the introduction of indicator variables is primarily a mathematical technique which insures the zero mean residual assumption also within the small areas and allows for deriving a simpler and better variance estimate.

The ***small area regression estimator*** can now be rewritten as

$$\hat{Y}_{G,reg} = \frac{1}{n_{1,G}} \sum_{x \in S_{1,G}} \mathcal{Z}(x)\hat{\theta}_{s_2} = \hat{\hat{\mathcal{Z}}}_{1,G}^{t}\hat{\theta}_{s_2} \qquad 4\text{-}38$$

Again, the g-weights can be used to give the estimated *design-based* variance of $\hat{Y}_{G,reg}$ which also takes the uncertainty in the regression coefficients in $\hat{\theta}_{s_2}$ into account. The g-weights are in this case defined as

$$\tilde{g}_{G,1}(x) = \hat{\hat{\mathcal{Z}}}_{1,G}^{t}\mathcal{A}_{s_2}^{-1}\mathcal{Z}(x) \qquad 4\text{-}39$$

The covariance matrix of $\hat{\hat{\mathcal{Z}}}_{1,G}$ is

$$\hat{\Sigma}_{\hat{\hat{\mathcal{Z}}}_{1,G}} = \frac{1}{n_{1,G}(n_{1,G}-1)} \sum_{x \in S_{1,G}} \left(\mathcal{Z}(x) - \hat{\hat{\mathcal{Z}}}_{1,G}\right)\left(\mathcal{Z}(x) - \hat{\hat{\mathcal{Z}}}_{1,G}\right)^{t} \qquad 4\text{-}40$$

With the empirical residuals $\hat{\mathcal{R}}(x) = Y(x) - \mathcal{Z}^{t}(x)\hat{\theta}_{s_2}$, the estimated ***design-based variance*** can be calculated by

$$\hat{V}(\hat{Y}_{G,reg}) = \frac{1}{n_2^2} \sum_{x \in S_2} \tilde{g}_{G,1}^{2}(x)\hat{\mathcal{R}}^{2}(x) + \hat{\theta}^{t}{}_{s_2}\hat{\Sigma}_{\hat{\hat{\mathcal{Z}}}_{1,G}}\hat{\theta}_{s_2} \qquad 4\text{-}41$$

4.5.2 Generalized Small Area Regression Estimator

Using the generalized regression estimator (chapter 4.4.2) in the framework of small area estimation, the simplest solution is again to restrict the samples in F to the small area $G \subset F$. The ***generalized small area estimator*** is then given by

$$\hat{Y}_{G,greg} = \left(\bar{Z}_G^{(1)} - \hat{\hat{Z}}_{1,G}^{(1)}\right)^{t}\hat{\alpha}_2 + \hat{\hat{Z}}_{1,G}^{t}\hat{\beta}_2 + \frac{1}{n_{2,G}} \sum_{x \in S_{2,G}} \hat{R}(x) \qquad 4\text{-}42$$

with the following *mean values*:

$$\bar{Z}_G^{(1)} = \frac{1}{\lambda(G)} \int_G Z^{(1)}(x)\,dx, \quad \hat{\hat{Z}}_{1,G}^{(1)} = \frac{1}{n_{1,G}} \sum_{x \in S_{1,G}} Z^{(1)}(x), \qquad \hat{\hat{Z}}_{1,G} = \frac{1}{n_{1,G}} \sum_{x \in S_1} Z(x) \qquad 4\text{-}43$$

where $Z^t(x = \left(\hat{\hat{Z}}_{1,G}^{(1)t}, \hat{\hat{Z}}_{1,G}^{(2)t}\right)$ comprises the exhaustive as well as the non-exhaustive auxiliary variables. Neglecting the uncertainty in the regression coefficients, the ***external variance*** estimate is

$$\hat{V}\left(\hat{Y}_{ext,G,greg}\right) = \frac{1}{n_{1,G}(n_{2,G}-1)} \sum_{x \in s_{2,G}} \left(\hat{R}_1(x) - \hat{\hat{R}}_{1,G}\right)^2 + \left(1 - \frac{n_{2,G}}{n_{1,G}}\right) \frac{1}{n_{2,G}(n_{2,G}-1)} \sum_{x \in s_2} \left(\hat{R}(x) - \hat{\hat{R}}_G\right)^2 \quad 4\text{-}44$$

where $\hat{\hat{R}}_{1,G} = \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} R_1(x)$ and $\hat{\hat{R}}_G = \frac{1}{n_{2,G}} \sum_{x \in s_{2,G}} R(x)$. Again, $\hat{Y}_{G,greg}$ is very similar to the generalized regression estimator $\hat{Y}_{F,greg}$ introduced in chapter 4.4.2 (equation 4-28). Also in this case, the regression coefficients in $\hat{\alpha}_2$ and $\hat{\beta}_2$ have been calculated using all data in *F* and are thus not adjusted to the conditions within the small area *G*. Consequently, the zero mean residual assumption does not longer hold true and the residual term (third term in equation 4-42) does not longer vanish in general.

The zero mean residual property over *F* and *G* can be realized by extending the model with the indicator variable $I_G(x)$ for the small area (Mandallaz 2013b). It thereby suffices to include $I_G(x)$ in the first, exhaustive component $Z^{(1)}(x)$ of $Z(x)$. The extended auxiliary vectors hence are $\mathcal{Z}^t(x) = \left(\mathcal{Z}^{(1)t}(x), \mathcal{Z}^{(2)t}(x)\right)$, where $\mathcal{Z}^{(1)t}(x) = \left(Z^{(1)t}(x), I_G^t(x)\right)$ and $\mathcal{Z}^{(2)t}(x) = Z^{(2)t}(x)$. Like in case of the small area regression estimator (chapter 4.5.1), the reason for introduction of the indicator variable is to insure the zero mean residual property within the small areas which allows for the calculation of a better variance estimate.

The ***large model*** can then be rewritten as

$$Y(x) = \mathcal{Z}^t(x)\theta + \mathcal{R}(x) = \mathcal{Z}^{(1)t}(x)\theta^{(1)} + \mathcal{Z}^{(2)t}(x)\theta^{(2)} + \mathcal{R}(x) \quad 4\text{-}45$$

and the ***reduced model*** as

$$Y(x) = \mathcal{Z}^{(1)t}(x)\gamma + \mathcal{R}_1(x) \quad 4\text{-}46$$

One now estimates the regression coefficients $\hat{\theta}_2$ for the large model and $\hat{\gamma}_2$ for the reduced model by equation 4-47. All regression coefficients in $\hat{\theta}_2$ and $\hat{\gamma}_2$ are the theoretical ordinary least square estimates of θ and γ and insure the zero mean residual properties in *F* and *G*.

$$\hat{\theta}_2 = \left(\frac{1}{n_2}\sum_{x\in s_2} \mathcal{Z}(x)\,\mathcal{Z}^t(x)\right)^{-1} \frac{1}{n_2}\sum_{x\in s_2} Y(x)\mathcal{Z}(x) := \mathcal{A}_2^{-1}\frac{1}{n_2}\sum_{x\in s_2} Y(x)\,\mathcal{Z}(x)$$

$$\hat{\gamma}_2 = \left(\frac{1}{n_2}\sum_{x\in s_2} \mathcal{Z}^{(1)}(x)\,\mathcal{Z}^{(1)t}(x)\right)^{-1} \frac{1}{n_2}\sum_{x\in s_2} Y(x)\mathcal{Z}^{(1)}(x) \qquad 4\text{-}47$$

$$:= \left(\mathcal{A}_2^{(1)}\right)^{-1}\frac{1}{n_2}\sum_{x\in s_2} Y(x)\,\mathcal{Z}^{(1)}(x)$$

The *new* ***generalized small area estimator*** can thus be calculated by

$$\hat{\bar{Y}}_{G,greg} = \left(\bar{\mathcal{Z}}_G^{(1)} - \hat{\bar{\mathcal{Z}}}_{1,G}^{(1)}\right)^t \hat{\gamma}_2 + \hat{\bar{\mathcal{Z}}}_{1,G}^t \hat{\theta}_2 \qquad 4\text{-}48$$

with the *mean values*:

$$\bar{\mathcal{Z}}_G^{(1)} = \frac{1}{\lambda(G)}\int_G \mathcal{Z}^{(1)}(x)\,dx, \qquad \hat{\bar{\mathcal{Z}}}_{1,G}^{(1)} = \frac{1}{n_{1,G}}\sum_{x\in s_{1,G}} \mathcal{Z}^{(1)}(x), \qquad \hat{\bar{\mathcal{Z}}}_{1,G} = \frac{1}{n_k}\sum_{x\in s_{1,G}} \mathcal{Z}(x) \qquad 4\text{-}49$$

Defining the estimated covariance matrices of $\hat{\theta}_2$ and $\hat{\gamma}_2$ as

$$\hat{\Sigma}_{\hat{\theta}_2} = \mathcal{A}_2^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2} \hat{\mathcal{R}}^2(x)\mathcal{Z}(x)\mathcal{Z}^t(x)\right)\mathcal{A}_2^{-1}$$

$$\hat{\Sigma}_{\hat{\gamma}_2} = (\mathcal{A}_1^{(1)})^{-1}\left(\frac{1}{n_2^2}\sum_{x\in s_2} \hat{\mathcal{R}}_1^2(x)\mathcal{Z}(x)\mathcal{Z}^t(x)\right)(\mathcal{A}_1^{(1)})^{-1} \qquad 4\text{-}50$$

with $\mathcal{R}(x) = Y(x) - \mathcal{Z}^t(x)\hat{\theta}_2$ being the residuals of the large model and $\mathcal{R}_1(x) = Y(x) - \mathcal{Z}^{(1)t}(x)\hat{\gamma}_2$. Mandallaz (2013b) gives the respective ***estimated design-based variance*** by

$$\hat{V}\left(\hat{\bar{Y}}_{G,greg}\right) = \frac{n_2}{n_1}\bar{\mathcal{Z}}_G^{(1)t}\hat{\Sigma}_{\hat{\gamma}_2}\bar{\mathcal{Z}}_G^{(1)} + \left(1 - \frac{n_2}{n_1}\right)\hat{\bar{\mathcal{Z}}}_{1,G}^t\hat{\Sigma}_{\hat{\theta}_2}\hat{\bar{\mathcal{Z}}}_{1,G} \qquad 4\text{-}51$$

4.5.3 Calculation of Confidence Intervals

For local estimations, such as in the framework of small area estimation, the two-sided $1-\frac{\alpha}{2}$ confidence interval has to be calculated based on the t-distribution with $n_{2,G}-1$ degrees of freedom (Mandallaz 2013a). The formula is given in equation 4-52. Whereas in

the case of global estimations, the degrees of freedom are calculated due to the (usually sufficiently large) amount of terrestrial observations in the entire inventory area, they are here only calculated based on the respective number of observations within the small area *G*. If the available number of terrestrial observations in *G* is small, this can lead to large t-values and hence to large confidence intervals. One possibility to still derive satisfactory confidence intervals is to use the synthetic estimator (Mandallaz 2008), since its variance will in general be small, but with the risk of its estimates being potentially biased. An alternative method is to embed the small area *G* in a larger inventory area *H* and to "borrow strength" from the higher amount of available observations in *H*. This approach can be applied with the proposed small area regression estimators, while insuring asymptotic consistency of its derived estimations by introducing the indicator variable for the small area *G*.

$$CI_{1-\alpha}\left(\hat{Y}_G\right) = \left[\hat{Y}_G - t_{n_{2,G-1},1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_G)}\,, \hat{Y}_G + t_{n_{2,G-1},1-\frac{\alpha}{2}}\sqrt{\hat{V}(\hat{Y}_G)}\right] \qquad 4\text{-}52$$

4.6 Implementation of Inventory Design

This section will demonstrate how the two-phase inventory grid described in chapter 4.3, 4.4 and 4.5 were implemented in the study area. This processing step comprised the forest definition by generating a forest mask, identifying the terrestrial inventory points within this forest definition and the implementation of the first phase sample points.

4.6.1 Forest Area Definition

Since for a forest inventory only inventory plots *within* the forest should be taken into account, one of the first issues is to decide whether an inventory point falls into the forest or non-forest area. Furthermore, one needs to derive the exact area extent of the forest for which the inventory will be performed in order to estimate the *inclusion probabilities* as well as the *Total* (chapter 4.3). For these purposes one needs a consistent *definition* of forest which satisfies not only the major requirement of defining the total forest area but also allows an unambiguous decision on whether the centre of an inventory plot is located inside the forest or not (Lanz 2005). It has been shown that this can be a demanding task, especially if a plot centre is located close to the forest edge (Kleinn et al. 2011). Various studies in the field of forest change assessment (Kattenborn 2013) or estimation of net carbon emissions by land cover change in the tropics (Achard et al. 2004) use the forest definition of the Food and Agriculture Organization of the United Nations FAO (2000). The FAO definition is mainly based on criteria regarding a minimum forest area, a minimum canopy cover and a potentially reachable tree height. The NFI layout also provides a detailed forest definition which is based on the just mentioned main criteria, but the rules for a forest-/non-forest-decision are far more detailed (Stierlin et al. 1999). The decision

process in the NFI is carried out by visual interpretation of aerial photographs and is in cases of ambiguity also validated by field surveys of the terrestrial inventory.

Preferably, the forest mask for this study should have represented the forest definition of the NFI. Unfortunately, no forest layer according to this definition was available. Since the implementation of a forest mask according to NFI rules was beyond the scope of this study, the TLM3D data (Topographic Landscape Model) provided by the Swiss Federal Institute of Topography swisstopo (2013) were used instead. The TLM3D is under development since spring 2008 and provides a large-scale topographical landscape model of Switzerland, including natural as well as artificial landscape features in vector format. The land cover map, which is one of nine data products derived by the TLM3D model, comprises the geometries and features of nine land cover classes with high positional accuracy (Fig. 4-8), among those three classes concerning forest-types (open forest, closed forest, shrub forest). The positional accuracy for well-defined landscape features (e.g. rocks, glaciers, water bodies) is given with 0.2 – 1.5 m, the accuracy for forest areas is slightly worse (1 – 3 m).

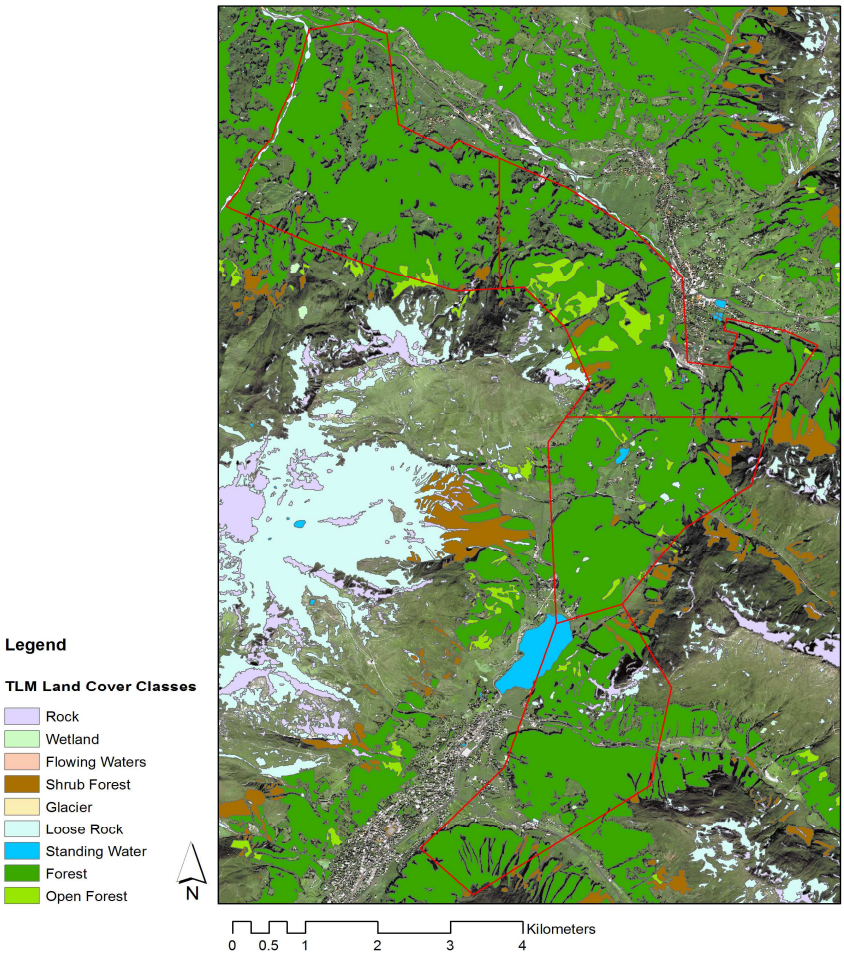


Fig. 4-8: TLM land cover classes covering the study area (delineated by red line). An RGB Spot5 Satellite Image was used as background for visualization purposes.

A forest mask (in raster and vector format) for the study area was then generated by selecting only those TLM3D classes concerning forest (open forest, closed forest, shrub forest) within the study area. Raster cell size (0.5 m) and the orientation of the forest raster mask were compliant to the CHM. Unfortunately, no documentation about the rules of forest-/non-forest-decision of the TLM3D data was available. For that reason, the suitability of the mask was assessed by a visual comparison of the derived forest mask to the CHM and a Spot5 true color satellite image (spatial resolution of 5 meters) at random locations within the study site. An illustration of the complete forest mask as well as the validation procedure is given in chapter 5.1.

Although other forest definitions are possible, it should be underlined that the most important point is to perform the inventory in accordance to a consistent forest definition. Smaller differences between definitions are less relevant.

4.6.2 Terrestrial Data Processing

The terrestrial data of the regional inventory (chapter 4.1.2) were imported into the geodatabase using the center coordinates of the sample plots. The provided dataset comprised information of sample plots lying within forest area according to NFI definition. Originally, 76 terrestrial sample plots were thus located within the study area. However, a visual inspection revealed that the NFI forest definition did not match in all cases the forest definition of the current inventory, i.e. there were terrestrial plots not lying within the forest mask (chapter 4.6.1). For that reason the terrestrial plots were again masked by the forest mask. Doing so, the number of terrestrial plots in the study area was reduced from 76 to 67 plots, of which 19 were located in small area 1, 17 located in small area 2, 15 located in small area 3 and 16 located in small area 4 (Fig. 4-9). An investigation revealed that there were no theoretical sample plot locations within the forest mask where terrestrial information had not been derived. The 67 plots of the second phase provided the terrestrial information in the estimation framework.

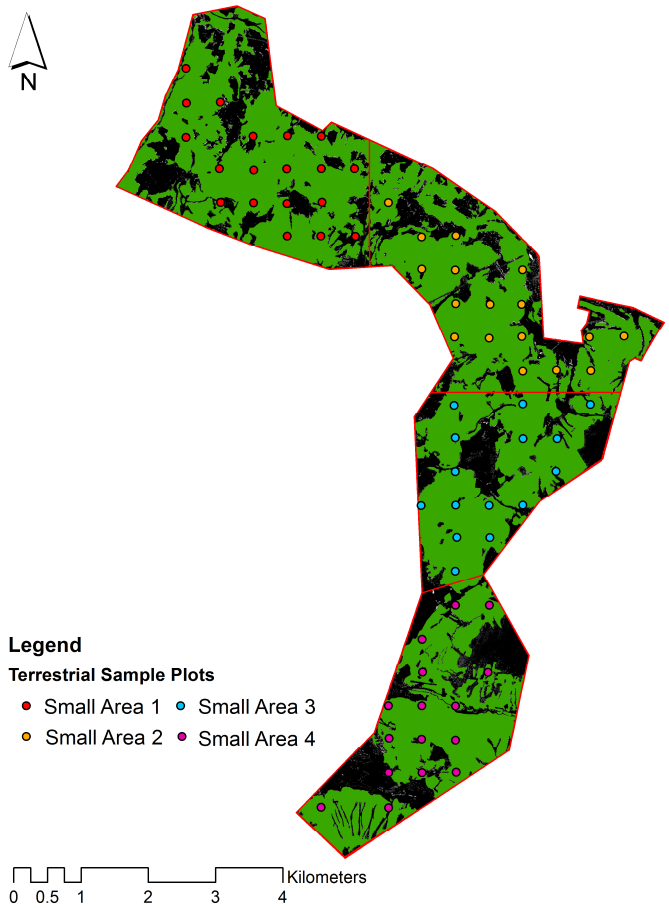


Fig. 4-9: Location of all 67 terrestrial inventory plots within the inventory area, providing the second phase (terrestrial) information.

4.6.3 First Phase Inventory Grid

The first phase sample points were created as a regular grid with a meshsize of 250 meters, which is half the meshsize of the terrestrial grid. The first phase grid was calculated within the open source software *R* (R Development Core Team 2012) and imported into a GIS using the *R*-package *shapefiles* (Stabler 2013). The grid was generated by catching a NFI sample point in the upper left neighborhood of the study area and defining it as the starting point of the first phase grid, whose extent covers the entire study area (Fig. 4-10, left). Choosing the meshsize of the first phase grid as being divisible by the second phase meshsize without remainder, one ensures that the second phase grid is actually a subgrid of the first phase grid (see *Two-Phase Sampling*, chapter 4.3). To derive the final two-phase inventory design for the inventory area, also the first phase sample points were masked by the forest mask (Fig. 4-10, right). A larger illustration can be found in Appendix A. The whole inventory area finally comprised 306 first phase sample points.

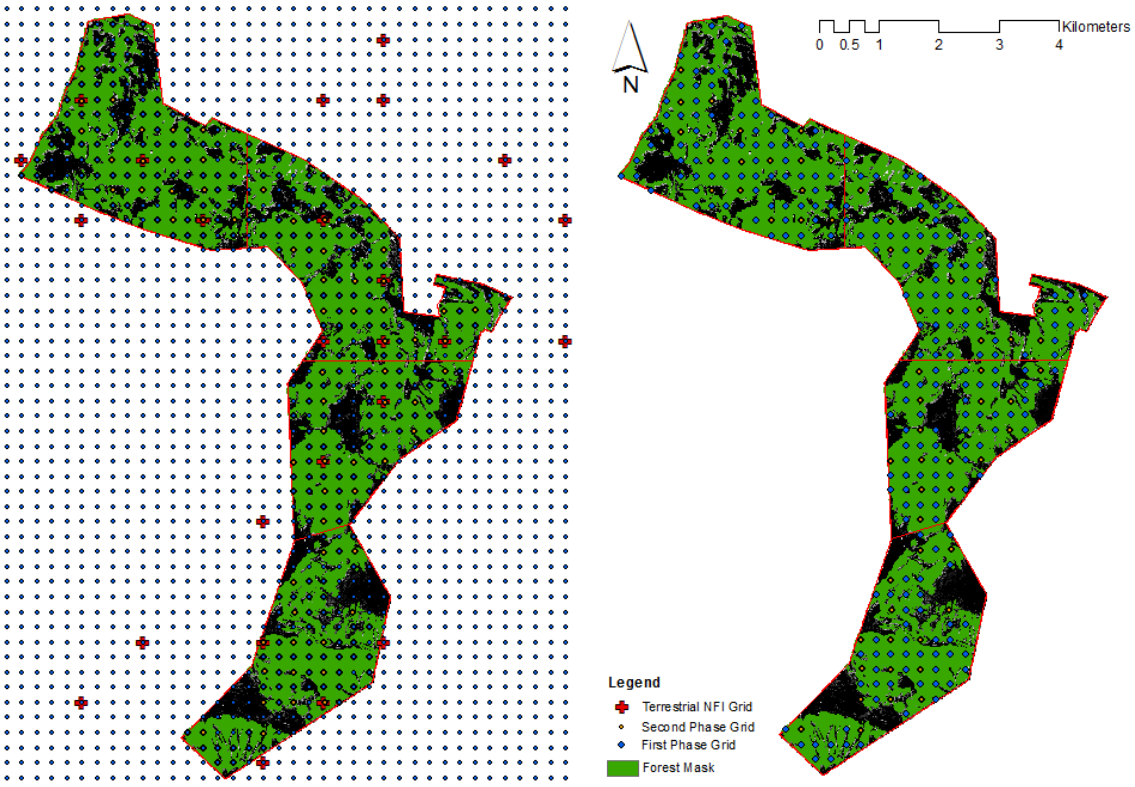


Fig. 4-10: *Left*: Study Area overlaid by first phase grid. It can be observed that the regional terrestrial sample points are a subgrid of the NFI sample points as well as a subgrid of the first phase grid. *Right*: Final two-phase sampling design for the study area (also the first phase sample points have been masked by the forest mask).

Tab. 1 shows the ratios between the number of second phase and first phase sample points. The ratio in the small areas (between 1:4 and 1:5) did not differ much from the ratio of the entire inventory area (1:4.6).

Tab. 1: Ratios between number of second phase and first phase sample points in the inventory area

Entire Area	*Small Area 1*	*Small Area 2*	*Small Area 3*	*Small Area 4*
67:306	19:94	17:81	15:66	16:65
(1:4.57)	(1:4.95)	(1:4.76)	(1:4.4)	(1:4.06)

4.7 Computation of Auxiliary Variables

This chapter introduces the auxiliary variables derived from the CHM and gives a detailed description of how they are computed. Before going into further detail, it is essential to recapitulate some crucial terminology concerning the *type* of auxiliary variables: In the framework of the generalized regression approach (chapter 4.4.2 and 4.5.2) it becomes necessary to partition the auxiliary variables into so-called *exhaustive* variables and *non-exhaustive* variables (which are assumed to need considerably more computation time),

because only the exhaustive ones are going to be calculated exhaustively over the whole inventory area. However, in the classic two-phase sampling approach there is no differentiation between the exhaustive and non-exhaustive variables since the variables are here only calculated at the first phase sample points (see chapter 4.4.1 and 4.5.1). This issue is again illustrated in Fig. 4-11. Since the focus of this study was mainly on implementing the alternative approach using exhaustive information, the following two sections will describe the calculation of the exhaustive and non-exhaustive variables separately. In the last section it is demonstrated how the exhaustive variables were calculated exhaustively over the entire inventory area.

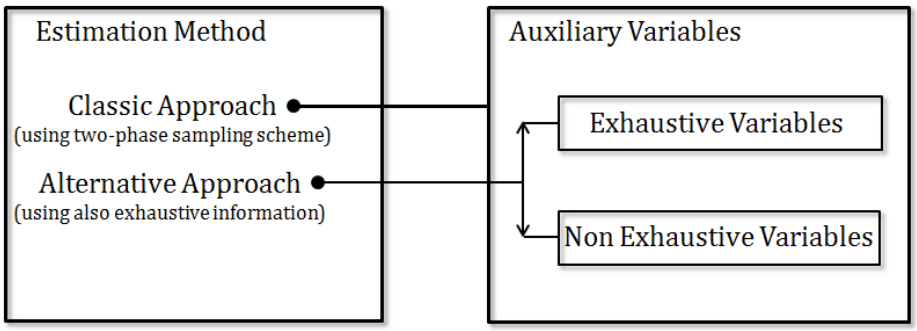


Fig. 4-11: Illustration of the terminology for the auxiliary variables: In the framework of the generalized estimator which uses exhaustively derived information, the auxiliary variables have to be partitioned into the *non-exhaustive*ly and *exhaustive*ly derived variables. This differentiation is not necessary in the framework of the classical two-phase sampling estimation approach.

4.7.1 Non-exhaustive Auxiliary Variables

Two non-exhaustive auxiliary variables were calculated at a given sample point. The extraction of the variables is based on a circle with a radius of 12.62 meters around the sample point, which is the same as the outer field plot radius. This circle is further referred to as the *LiDAR sample plot* (*LSP*) of the non-exhaustive variables (representing the plot of the terrestrial inventory). The extraction of the variables is also adjusted for boundary effects at the forest edge. The variables are the LiDAR derived *stem number* per hectare and the LiDAR-derived *Volume Density* per hectare. In the following, the processing steps of these variables will be described in detail. As mentioned in chapter 4.4.2, the effort of deriving these variables is assumed to be computer intensive compared to the exhaustive variables as they rely on more sophisticated algorithms.

Stem number

The computation of the *stem number* per hectare in general comprises seven processing steps which are illustrated by Fig. 4-12. The algorithm purely consists of raster operations. The procedure is mainly based on the identification of tree locations by the detection of local height maxima on the canopy height model CHM. A raster cell in the CHM is defined

as a local maximum if its height value is larger than the height of its eight grid neighbors. This procedure has amongst others been proposed by Solberg et al. (2006) and was also used by Heinimann and Breschan (2012). The position of a raster cell which has been identified as a local maximum is then interpreted as the location of a tree. A respective study of Persson et al. (2002) in a boreal forest in southern Sweden using LiDAR- and available ground truth data revealed a detection of 71% of all trees and 90% of the trees with a DBH >20 cm. The forest stands comprised 795 spruce and pine trees, thus showing a very similar tree species composition as the current study site. Also the point-density of Laserscanning data (5 points/m^2) was equal to the one used in this study. However, this technique is usually not able to detect understory trees or small trees standing close to large trees (Vosselmann and Maas 2010).

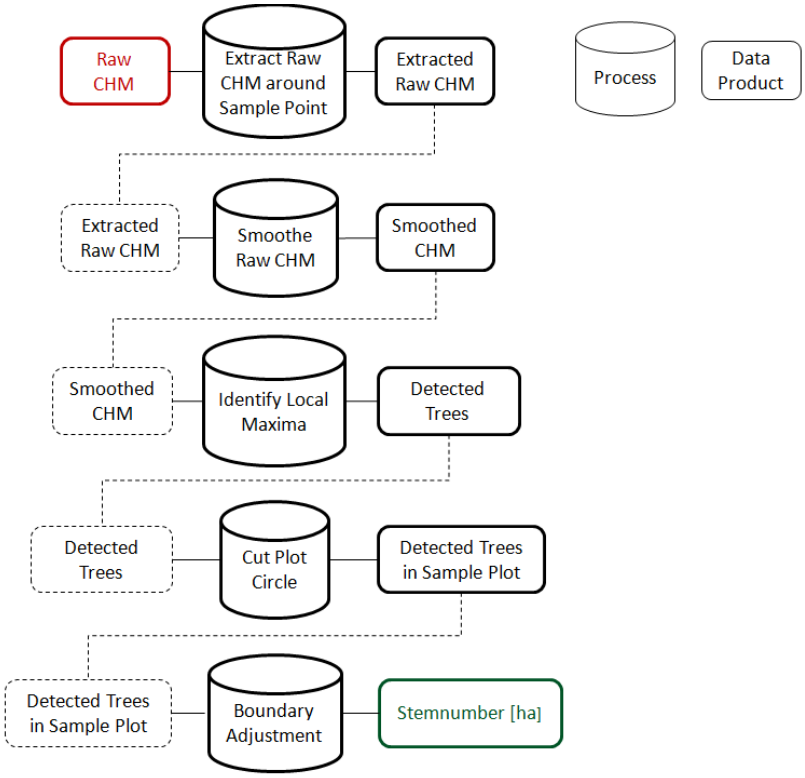


Fig. 4-12: Simplified illustration of the algorithm to compute the LiDAR-estimated *Stem number* per hectare based on single tree detection within a sample plot.

Two processing steps have to be performed before the detection of the local maxima can be applied: Firstly, the CHM is extracted in a defined surrounding around the first phase sample point (i.e. a square of certain edge length). By this operation, the CHM raster cell which spatially covers the sample point becomes a generalized representation of the originally sample point location. The next step comprises the *smoothing* of the extracted CHM. Various studies (Hyyppä et al. 2001; Solberg et al. 2006) revealed that smoothing the CHM is essential to insure an accurate tree detection. This becomes necessary as the *raw* CHM (i.e. the difference between DSM and DTM without any further processing) shows a

high roughness (Fig. 4-13, left) which may lead to an overdetection of trees due to a high amount of local maxima. Finding an appropriate degree of smoothing can be a challenging task as a mild smoothing might not sufficiently decrease the roughness in order to solve the problem of overdetection, whereas a tough smoothing can result in missing a considerable amount of truly existing trees (Solberg et al. 2006). In this study, a 5x5 Gaussian Kernel filter was applied to smooth the CHM before detecting the local maxima. A visual inspection of the CHM after smoothing (Fig. 4-13, right) shows that this leads to a decreased roughness and thereby to more defined height maxima located at the tree top positions.

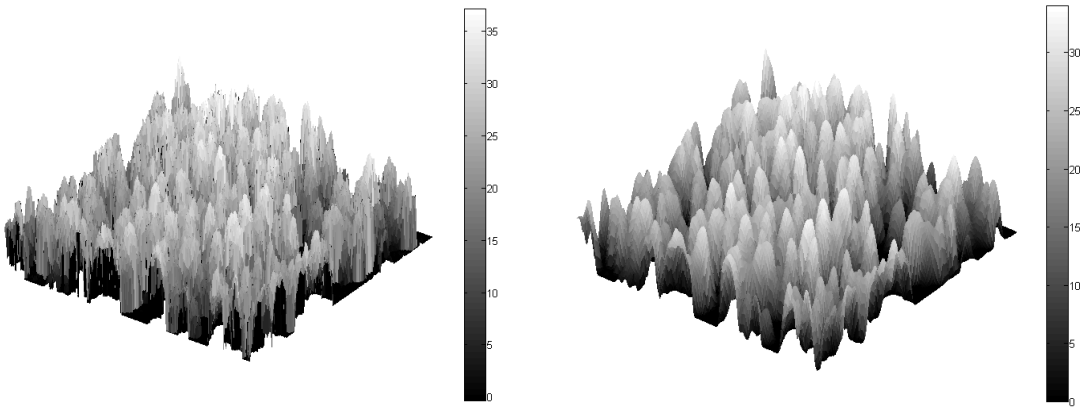


Fig. 4-13: *Left*: The raw CHM shows a high roughness which could lead to an overdetection of the trees due to a high amount of local maxima. *Right*: After application of a Gaussian Kernel filtering, the CHM shows a decreased roughness and well defined local maxima localized at the locations of the tree tops. The color bar in each figure displays the object height in meters.

Following this operation, the local maxima detection is performed on the smoothed CHM. An example of the detected local maxima for an extracted and smoothed CHM is illustrated in Fig. 4-14 (left). Afterwards, the detected trees have to be limited to only those lying within the LiDAR sample plot. This was realized by applying a binary circlemask in raster format with its center pixel representing the plot center. Fig. 4-14 (upper right) demonstrates that the original circle can only be approximated by the raster mask. The approximation accuracy between the circle and its raster representation thereby highly depends on the spatial raster resolution. As the circlemask has to be applicable to the CHM, its raster resolution was compliant to the CHM (i.e. 0.5 meters). This obviously led to an appropriate approximation of the original circle. The circlemask was then used to select only the detected tree positions exclusively within the LiDAR sample plot (Fig. 4-14, lower right).

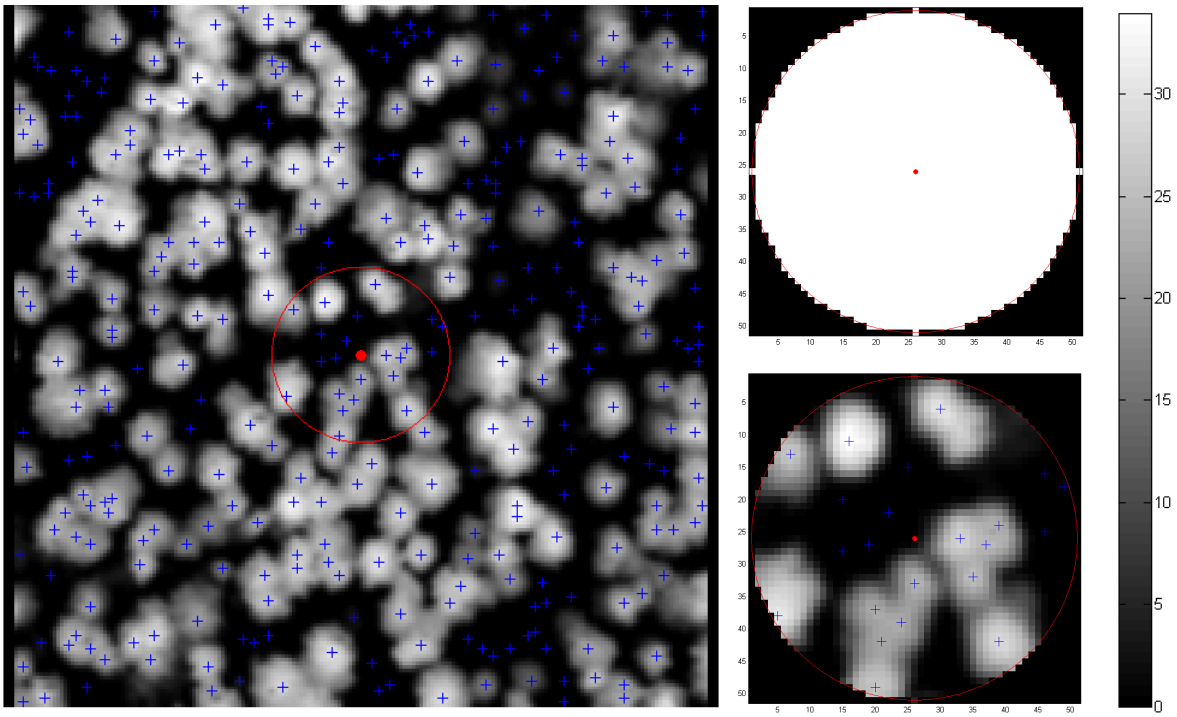


Fig. 4-14: Example for tree detection within the LiDAR sample plot. *Left*: Trees are identified (blue crosses) by detection of local height maxima in a specified surrounding of a sample plot (indicated by red circle). *Upper right*: A binary plot mask (black=non-forest, white=forest) representing the LiDAR sample plot is used to restrict the identified trees to only those lying within the sample plot (*lower right*). The color bar shows the object height in meters.

The last operation in order to derive the *stem number* per hectare comprises the boundary adjustment. For each extracted CHM the corresponding forest mask was therefore also extracted and then restricted to the LiDAR sample plot using the circlemask. The so far detected trees within the sample plot are then finally reduced to those also located within the forest mask (Fig. 4-15). The *stem number* per hectare can then finally be calculated by

$$Stem\ number_i[per\ hectare] = \frac{ndtrees}{\lambda(LSP \cap FM)} \cdot 10000 \qquad 4\text{-}53$$

with $ndtrees$ being the number of detected trees within the *i*th LiDAR Sample Plot (LSP) and the forest mask (FM); $\lambda(LSP \cap FM)$ describes the intersection of the LiDAR Sample Plot and the forest mask.

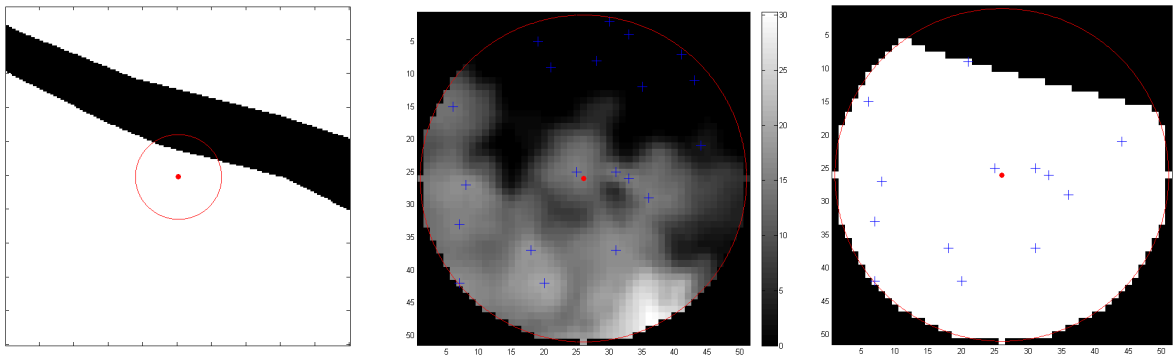


Fig. 4-15: Illustration of Boundary Adjustment. *Right*: The binary forest mask is extracted for the corresponding CHM (white=forest area, black=non-forest area) and then restricted to the sample plot. The so far detected trees (blue crosses) in the sample plot (*center*) are then reduced to those located within the forest mask (*right*). The extrapolation to hectare is then based on the forest area within the sample plot (intersection).

Volume Density

Also the calculation of the *Volume Density* per hectare is based on the detection of tree positions within the LiDAR sample plot. Fig. 4-16 shows a simplified illustration of the algorithm. First, a detection of all tree positions lying within the LiDAR sample plot as well as within the forest definition is done with the same procedure used in the *Stem number* algorithm. The next step comprises the extraction of the CHM height value at each detected tree position. The extracted height is consequently interpreted as the tree height of the respective identified tree. A regression model is then used to predict the timber volume of each detected tree by its tree height. By summing up these predictions, one gets an estimation of the total timber volume within the boundary adjusted LiDAR sample plot.

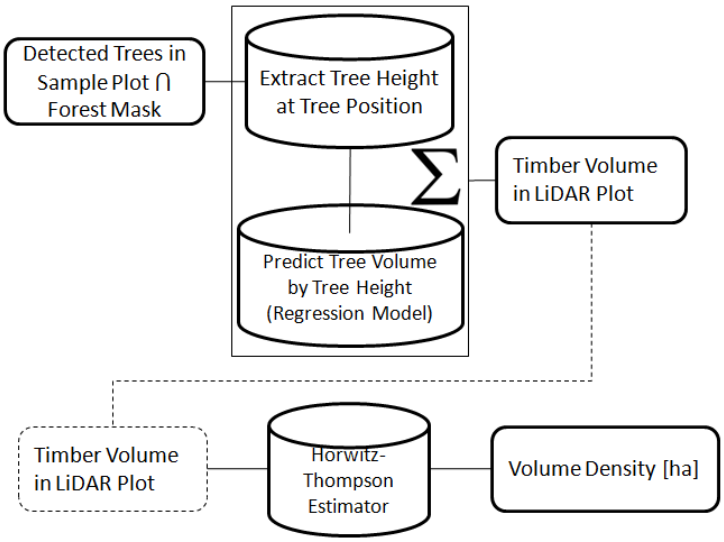


Fig. 4-16: Simplified illustration of the algorithm to compute the LiDAR-estimated *Volume Density* [ha] on sample plot level.

The *Volume Density* per hectare for each sample plot is then calculated as

$$Volume\ Density\ [per\ hectare] = \frac{\sum_{i=1}^{ndtrees} Vol_{LiDAR,i}}{\lambda(LSP \cap FM)} \cdot 10000 \qquad 4\text{-}54$$

with $ndtrees$ being the number of detected trees within a LiDAR sample plot *LSP* (corrected for boundary effects), $Vol_{LiDAR,i}$ being the estimated timber volume for the *i*th detected tree and $\lambda(LSP \cap FM)$ being the intersection of the LiDAR sample plot and the forest mask. Equation 4-54 exactly corresponds to the Horwitz-Thompson estimator introduced in chapter 4.3. The LiDAR-estimated *Volume Density* per hectare can therefore be seen as the estimated *local density* for the timber volume on plot level.

The regression model for predicting the timber volume of an individual tree based on its tree height was calculated using the second stage data (comprising DBH *and* tree height information) from the inventory domains *Pre-Alps* and *Alps* of the NFI. The dataset initially comprised 8420 observations, including an estimation of the total overbark stem volume for each sample tree. These stem volume estimations had been calculated by NFI tree-specific tariff-functions using the diameter at breast height DBH as predictor variable (Brassel and Lischke 2001; Kaufmann 2000). Unfortunately, the dataset also included a certain amount of shrub species. Since it was of interest to build the regression model purely on tree species, shrub species were excluded by using a simple threshold regarding the percentage share of the species in the second stage data. The remaining data were then used to calculate the regression model with the given tree stem volume as response variable and the corresponding tree height as predictor variable. A strong correlation between the two parameters has amongst others been reported by Husmann (2013) who found a respective correlation coefficient of 0.86 after linearization of tree data derived by random branch sampling at study sites in Germany. The correlation between DBH and tree height as well as DBH and tree volume (Zianis et al. 2005) has often been described by the theory of *allometry* (Niklas 2004; Pretzsch 2009), which focuses on the size relations in organisms and refers to the *Metabolic Scaling Theory* (Niklas 1994). Under the assumption that these allometric relations also hold true for the current dataset, the relation between tree volume and tree height can be expressed by an allometric function (also referred as a *power law*) of the form

$$Tree\ Timber\ Volume = \beta \cdot Tree\ Height^{\alpha} \qquad 4\text{-}55$$

This approach has also been used by Mette et al. (2004) in the context of estimating forest biomass by Pol-InSAR derived forest height information. The allometric relation in the form of equation 4-55 can be calculated as a non-linear least-square model. However, there are several reasons why a log-transformation of the response and predictor variable is often appropriate: Firstly, the relationship between the two variables can be expressed as a linear relationship which simplifies the regression model (Draper and Smith 1966;

Warton et al. 2006). Secondly, a log-transformation can reduce the severity and number of outliers in a dataset as well as stabilize the error variance. It can also decrease the severity of an increasing measurement error with increasing dimension of the measured objects (here the tree height and the DBH), which is often visible as a spreading of the data cloud in the original scale. The transformed model then reads as

$$\ln(Tree\ Timber\ Volume) = \ln(\beta) + \alpha\ \cdot \ln\ (Tree\ Height) \qquad 4\text{-}56$$

To calculate the predicted tree volume in the original scale, the regression coefficients derived under the linear model (equation 4-56) have to be backtransformed by the formula

$$Tree\ Timber\ Volume_{predicted} = e^{\ln(\beta)+\alpha\cdot\ln\ (Tree\ Height)} \cdot e^{\frac{\sigma^2}{2}} \qquad 4\text{-}57$$

where $e^{\frac{\sigma^2}{2}}$ is a factor to correct for the bias resulting from the logarithmic transformation using the standard deviation of the model residuals (Beauchamp and Olson 1973). The quality of the regression model was validated by the coefficient of determination (R^2), a diagnostic of the model residuals (constant error variance and uncorrelated errors) as well as a visual inspection of the tree volume predictions against their true values. The model results are presented in chapter 5.2.

4.7.2 Exhaustive Auxiliary Variables

The calculation of the exhaustive variables at a given sample point was based on a *square area* centered on the respective sample point. The square extent was chosen according to tangentially circumscribing a field- and LiDAR sample plot (Fig. 4-17, left). The square area consequently comprised 637 m^2. By analyzing the distribution of the (not smoothed) CHM raster values within a square, nine descriptive auxiliary variables were considered by means of simple statistical parameters: they were the *mean*, the *median*, the *standard deviation*, the *coefficient of variation*, the *MAD* (Median Absolute Deviation), the *maximum value* and the *25%-*, *75%-* and *90%-quantiles*. The reasons for the choice of these parameters were as follows: the *mean*, representing the average canopy height for a sample point, has been used in various studies (Hyyppä et al. 2008) and proved to be well correlated to field-derived forest attributes such as basal area and standing timber volume. However, the mean canopy height is a rather generalized parameter and does not provide any further information about the heterogeneity or homogeneity of the observed forest (i.e. theoretically, the same mean canopy height can result from a perfect homogeneous as well as from a very diverse forest). Information about the spatial variation in the canopy structure is assumed to be represented by the *standard deviation*. Additionally, the *quantiles* provide information about a potential skewness in the distribution of canopy height val-

ues, while the *maximum* height value is assumed to represent the highest tree within the sample plot (which should contribute most to the timber volume on plot level). All these variables have already been used as predictors in regression models for estimating the timber volume of forest stands (e.g. Holmgren 2004; Lefsky et al. 1999; Magnussen et al. 1999; Næsset 2002). Considering outliers in the height data due to potential errors in the acquisition and processing of the LiDAR data (chapter 4.2), for some of the hence proposed variables also their corresponding *robust* measures were derived (i.e. the *median* for the *mean*, the *coefficient of variation* for the *standard deviation* and also the *MAD* as a robust measure of the variability). The calculations of the exhaustive variables were also adjusted for boundary effects by only considering those raster values within a square which are covered by the forest mask (Fig. 4-17, center and right).

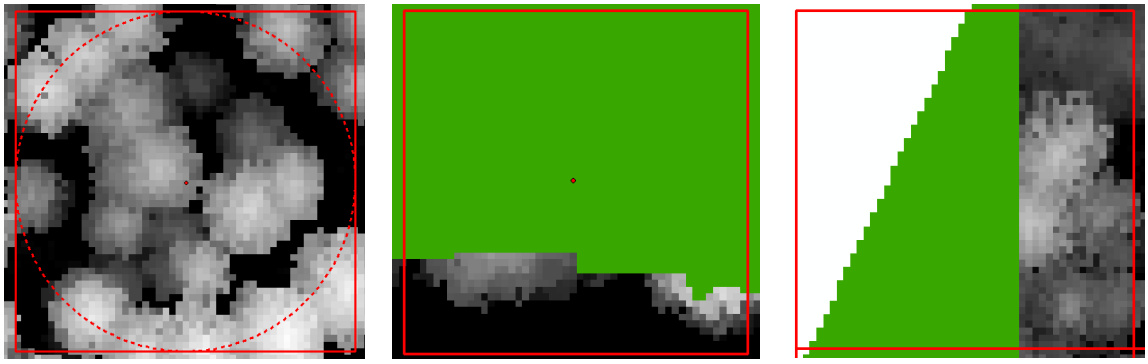


Fig. 4-17: *Left*: The calculations of the exhaustive variables were based on squares (red) which tangentially circumscribed the LiDAR sample plot of the non-exhaustive variables (12.62 meter radius, indicated by dotted circle). *Center* and *right*: The forest mask (green=forest area) was used to account for boundary effects (right shows a case were a square exceeds the limits of the study area).

To recapitulate, the application of the classic regression estimator (chapter 4.4.1 and 4.5.1) implies the calculation of the empirical means of all exhaustive variables over all observations at the first phase sample points. The empirical means are the *estimations* of their true means, which are in this case unknown. The empirical, i.e. *estimated* means are summarized in the auxiliary vector $\hat{\bar{Z}}_1$ and then multiplied with their corresponding regression coefficients in order to derive the point estimate for the inventory area (equation 4-16). The performance of the boundary adjustment now requires accounting for the fact that the calculation of the variables in the squares may be based on different numbers of raster values. To insure an unbiased calculation of the estimated means, the estimated means have to be calculated as the *weighted means* (equation 4-58). The weight $W_{i,square}$ for square $i \in s_1$ is the proportion of the number of raster cells of the forest mask within the square and the total number of raster cells of a square. These weights are hence proportional to the intersection area $\lambda(Square(x) \cap FM)$.

The vector $\hat{\bar{Z}}_1$ containing the estimated weighted means of the exhaustive variables (derived at the first phase sample points s_1) is calculated by

$$\hat{\bar{Z}}_1 = \frac{\sum_{i=1}^{n_1} W_{i,square} \cdot Z_i}{\sum_{i=1}^{n_1} W_{i,square}} \qquad 4\text{-}58$$

with Z_i being the vector containing all exhaustive variables derived in the *i*th square, $W_{i,square}$ being the weight for the *i*th square and the important property that the sum of all *relative* weights p_i is equal to one:

$$p_i = \frac{W_{i,square}}{\sum_{i=1}^{n_1} W_{i,square}} \qquad \Rightarrow \qquad \sum_{i=1}^{n_1} p_i = 1 \qquad 4\text{-}59$$

In case of the generalized regression estimator (chapter 4.4.2 and 4.5.2), the empirical means of the exhaustive variables are corrected by their corresponding *true* means which are contained in the vector $\bar{Z}^{(1)}$ (equation 4-28). To recapitulate, this is *theoretically* done by deriving the exhaustive variables exhaustively over the entire forest area. *In practice*, it is obvious that only the *mean canopy height* could exhaustively be calculated in one step (i.e. without the use of squares by calculating the mean for all raster cells of the whole CHM covered by the forest definition). The true means of the remaining exhaustive variables however have to be calculated based on the same sample unit as their corresponding estimates (i.e. in our case the square). For this reason, the whole study site was covered by squares of equally size as the ones created around the first phase sample points. It is essential that by doing so, the squares at the first phase points are actually a subset of the exhaustive grid of squares (Fig. 4-18, upper left and right). The exhaustive variables are then finally calculated within each square whose center is located in the forest area of the study site according to the same procedure as before (i.e. including adjustment for boundary effects). The forest-/ non-forest decision was again based on the forest mask (Fig. 4-18, below). In this study, 31622 squares had their center located within the forest definition and were thus used to calculate the respective true means of the exhaustive auxiliary variables. To ensure asymptotic consistency of the empirical means for $n_1 \to \infty$, the true means of the exhaustive variables were also calculated as the weighted means over all squares of the exhaustive grid. The individual weights were calculated according to the same procedure as before (equation 4-58 and 4-59), only replacing *n* (sample of the exhaustive grid) by *N* (all squares of the exhaustive grid). Using the hence proposed technique for deriving the true means, it might occur more suitable to refer to the collection of the necessary information as being *partially exhaustive*, as mathematically speaking, the vector function to derive the exhaustive variables $Z(x)$ is assumed to be stepwise constant over the squares lying in the forest area of the study site ($Square(x) \cap F$). Finally, one has to note that the demonstrated procedure requires a perfect tessellation of the forest area, which cannot be realized by using sample units like circles. A comprehensive illustration of the exhaustive grid over the complete study site is given in Appendix A.

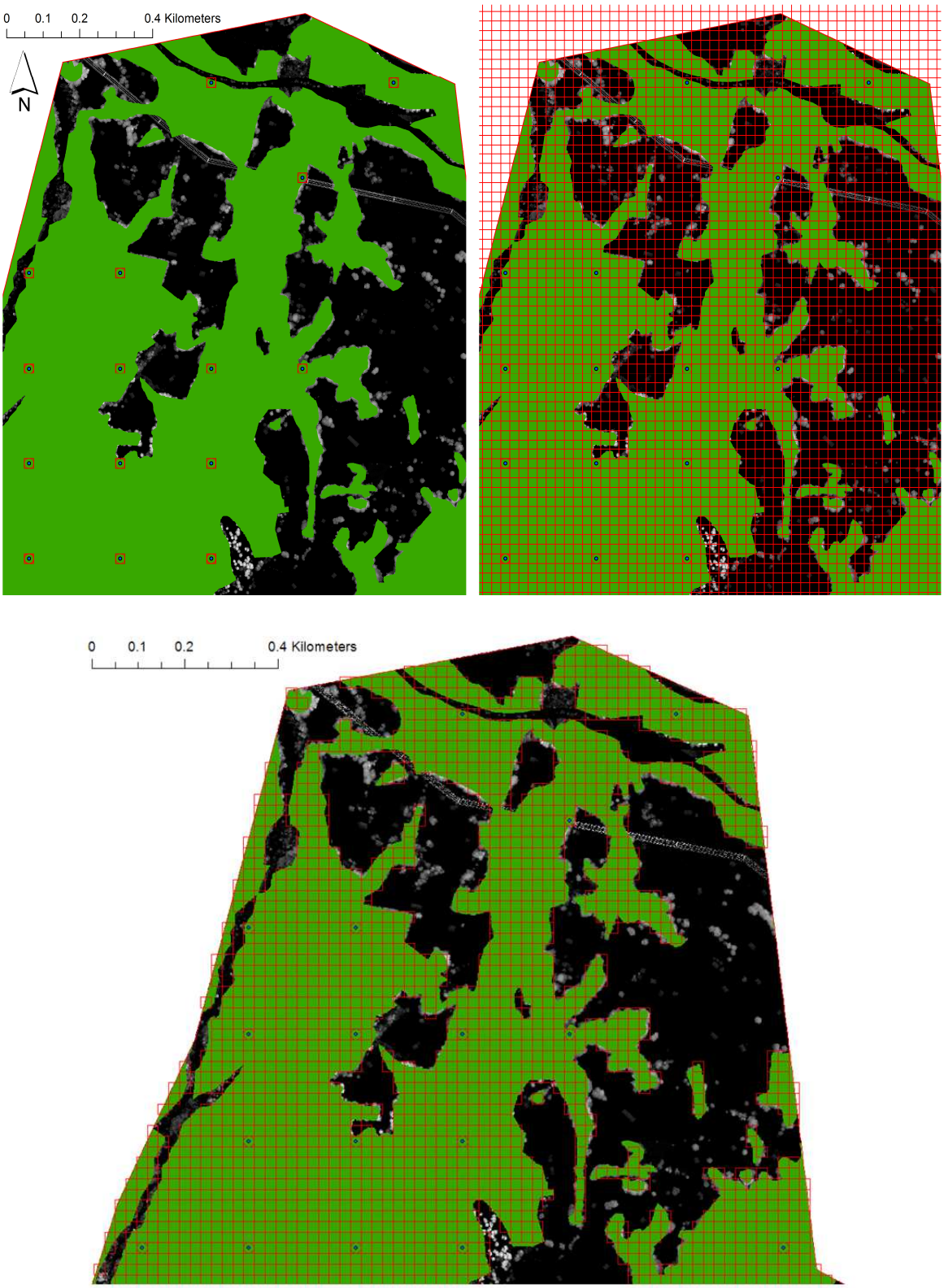


Fig. 4-18: Illustration of the *exhaustive grid* (figures show a part of the northern region of the study site): *Upper left* and *right:* The complete study area is overlaid by a grid of squares with the squares at the first phase sample points (blue) being a subset. *Below*: The exhaustive variables are then calculated within all squares whose center is located within the forest (decision based on the forest mask (green)). The CHM is shown in the background.

To summarize the described auxiliary variables, Tab. 2 provides again an overview about the kind of auxiliary variables considered in this study, their corresponding *type* (exhaustive or non-exhaustive), the *sampling unit* in which there are extracted (circle or square) and the *sample locations* at which they are computed according to the respective *estimation method*.

Tab. 2: Overview on all auxiliary variables considered in the inventory. They can be divided into their *Type* (exhaustive or non-exhaustive), the geometry *in* which they are computed (*Sampling Unit*) and the *Location* in the inventory area *at* which they are computed dependent on the Estimation *Method*. RE stands for the classical Regression Estimator and GRE for the Generalized Regression Estimator which also uses exhaustive information additionally (+). (SA) indicates the potential extension to small area estimation.

Variable	*Type*	*Sampling Unit*	*Location derived*	*Method*
Mean	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Median	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Sd	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
MAD	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Max	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Q25	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Q75	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Q90	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Varcoef	exhaustive	Square	1st phase points / + whole inventory area	(SA)RE / (SA)GRE
Stem number	non-exhaustive	Circle	1st phase points / 1st phase points	(SA)RE / (SA)GRE
Volume Density	non-exhaustive	Circle	1st phase points / 1st phase points	(SA)RE / (SA)GRE

4.8 Selection of Predictor Variables

With respect to the two-phase estimators described in chapter 4.4 and 4.5, auxiliary variables are used as *predictor variables* in regression models to estimate the local density (in this case the standing timber volume) at given sample points. It is essential to recapitulate, that in the *design-based* framework (also applied in this study) the regression models are used to improve the accuracy of the estimates, but they do not have to be correct. Within the *design-based* approach, the asymptotic consistency (i.e. the estimates converge towards the true values) of the ***point estimates*** and its ***estimated variances*** is ***not dependent*** on any model or residual assumptions regarding the regression model(s). The validity of the estimates purely relies on the randomization principle of the sample points (chapter 4.3). This also implies that even using all considered auxiliary variables as predictors in the regression models would be methodological sound. However, in accordance to good statistical practice (Occam's razor), it was tried to keep the regression models as simple and small as possible, meaning that only a subset of all predictors was going to be used in the final regression models. Again, this is a rather philosophical approach and not required for making *design-based* estimations. The entire process of predictor selection carried out belongs to the framework of *classical*, *model-dependent* statistics and has to be clearly differentiated from the subordinate *design-based* approach.

As a guideline for the variable selection, an explorative data analysis of the derived auxiliary variables as well as the application of classical statistical selection algorithms based on simple and multiple regressions of the auxiliary variables on the observed terrestrial timber volume (chapter 4.1.2) as response variable were performed. Therefore, all considered auxiliary variables (Tab. 2) were first calculated at the terrestrial s_2 sample point positions, since only at those locations a comparison to the terrestrial response variable is possible. It should however be emphasized that the choice of variables was considered not to be a purely statistical issue: prior knowledge from other inventories as well as the interpretability of the models was also attached great importance to. Based on the variable selection process, only a chosen subset of exhaustive and non-exhaustive auxiliary variables were computed at the first phase sample points (and the exhaustive variables also exhaustively in the case of the generalized regression estimator). The functionality of the variable selection procedure in the general auxiliary computation process is again illustrated in Fig. 4-19.

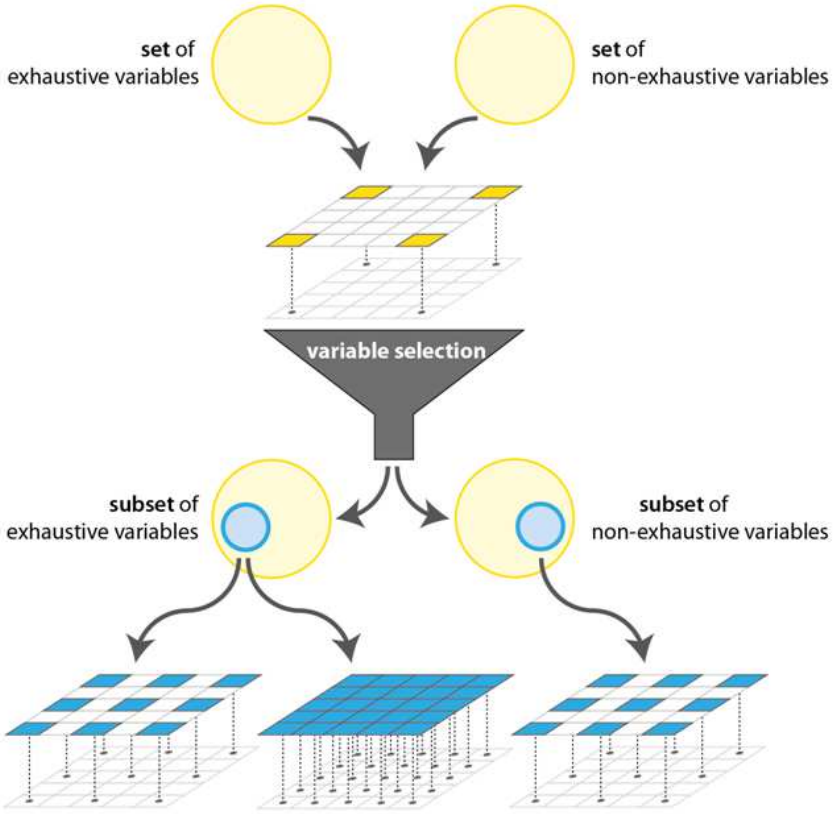


Fig. 4-19: The variable selection procedure requires the computation of *all* considered auxiliary variables at the s_2-sample points (yellow), since only at those points a comparison to the terrestrial response variable is possible. After selecting a subset of auxiliary variables, only these variables are subsequently calculated at the first phase sample points (blue) and, in the case of the generalized regression estimator, also exhaustively.

4.8.1 Principle of Parsimony

The variable selection procedure was motivated by the idea to satisfy the so-called *Principle of Parsimony* (also referred to as the Occam's razor). For statistical modeling, this principle means that - having the choice between several models which sufficiently explain a given response variable - the *simplest model* should always be preferred. This especially implies that a model should have *as few parameters as possible* (Crawley 2013). Consequently the *smallest* model that *fits* the data has to be preferred and thus any redundant information in terms of predictor variables or factor levels should be avoided (Faraway 2002). Ignoring this principle bears the risk of *overfitting*: if the most meaningful predictors have already been included in a model, further adding of more and more independent predictors will only slightly decrease the residual mean square error (MSE), but not significantly improve the fit of the data. Draper and Smith (1966) suggest as rule of thumb that a model should at least comprise five to ten times as much observations as predictor variables. Regarding the potential maximal number of 11 predictors in the large model explaining 67 terrestrial observations, it was thus attempted to limit the number of predictors to those with the highest predictive power with respect to the terrestrial response variable.

4.8.2 Simple Linear Regression Models

To get an overview of how well each auxiliary variable individually performs as predictor variable in order to estimate the observed terrestrial timber volume, a simple linear regression of each auxiliary variable against the terrestrial response variable was calculated. An inspection of the fitted regression line as well as of the residuals (constant error variance) was carried out to insure that a linear regression of first polynomial (i.e. fitting a straight line) was appropriate to describe the potential relationship between predictor and response, and no other kind of relationship was missed. The analysis results were considered to be used as additional information within the variable selection process, which was mainly based on a *multiple* regression analysis.

4.8.3 Multiple Linear Regression Models

With respect to the regression estimators applied within this study (chapter 4.4 and 4.5), the following multiple regression models were formulated:

- In the framework of the *generalized regression estimator* (chapter 4.4.2 and 4.5.2), the variable selection of the auxiliary variables has to be done *separately* for the *reduced model* and the *large model*. The reduced model (*rm*) potentially uses only all exhaustive variables (chapter 4.7.2) as predictor variables. The maximal possible regression model thus is:

$$Y = \beta_{0,rm} + \beta_{1,rm}Mean + \beta_{2,rm}Median + \beta_{3,rm}Sd + \beta_{4,rm}MAD + \beta_{5,rm}MAX + \beta_{6,rm}Q25 + \beta_{7,rm}Q75 + \beta_{8,rm}Q90$$

  The maximal possible large model (*lm*) comprises all exhaustive and non-exhaustive variables (chapter 4.7.1) as potential predictors and reads as:

$$Y = \beta_{0,lm} + \beta_{1,lm}Mean + \beta_{2,lm}Median + \beta_{3,lm}Sd + \beta_{4,lm}MAD + \beta_{5,lm}MAX + \beta_{6,lm}Q25 + \beta_{7,lm}Q75 + \beta_{8,lm}Q90 + \beta_{9,lm}Stem\ Number + \beta_{10,lm}Volume\ Density$$

- The variable selection for the "classical" *regression estimator* (chapter 4.4.1 and 4.5.1) is identical to the large model of the generalized regression estimator, since there is no distinction between exhaustive and non-exhaustive variables.

The variable selection was carried out for the *global* regression model and the identified best-subset of auxiliary variables was then also used as the set of predictors in the respective small area estimation models (chapter 4.4.2 and 4.5.2). The multiple regression models for the variable selection (large and reduced model) were thus calculated using the information of the respective auxiliary predictor variables and the terrestrial response variable at all s_2-sample points of the entire inventory area. It should be mentioned that theoretically it would also be possible to perform an individual variable selection for each

small area, but this would have contradict the idea of the proposed small area estimators (chapter 4.5), which was to take advantage from the larger number of observations in the global model ("borrow strength"), but to use the extension by the indicator variables as a mathematical technique in order to achieve the zero mean residual assumption also within the small areas, where the number of available terrestrial observation can be considerably limited. The multiple regression analyses thus comprised 67 observations (number of terrestrial plots in the complete inventory area).

4.8.4 Goodness-of-Fit and Testing-Based Selection Methods

As a guideline for the variable selection, selection methods relying on *goodness of fit criteria* as well as methods relying on *hypothesis testing* were performed, but the goodness of fit criteria were preferred for the following reasons:

The *testing* based methods rely on finding out, whether a predictor x has a significant influence on the response variable by testing the null hypothesis H_0: $\beta_x=0$ (no significant influence of x) against H_A: $\beta_x\neq0$ (significant influence of x). This is done by a *t-test* (Stahel 1999), whose acceptance and rejection regions for a given significance level can be determined by a Student distribution with n-p degrees of freedom (with p including the intercept). The decision about a significant relation between a certain predictor variable and the response variable (also referred to as *inference*) can then be based on the corresponding *t-value* of the test statistic or directly on the equivalent *p-value* (Draper and Smith 1966). In the framework of multiple regression analysis, each individual t-test quantifies the effect of the predictor on the response after having subtracted the linear effect of all other predictor variables in the model on the response. For all models being considered, the model residuals were investigated for constant error variance and uncorrelated errors. These assumptions satisfy the Gauss Markov Theorem which suggests that the ordinary least square estimate of the regression coefficients are the best linear unbiased estimate (BLUE). To rely on the inference on the predictors, the theoretical model residuals have also to be independently and identical normally distributed. These assumptions were validated by a visual inspection of the Tukey-Anscombe- (Anscombe and Tukey 1963) and the Scale-Location Plots as well as the QQ-Plots (Fox 2008) and the application of the Shapiro-Wilk-test (Royston 1982) for testing the model residuals for normality (Gaussian distribution). The Shapiro-Wilk test was performed with a significance level of $\alpha=5\%$.

However, the testing based variable selection can be seriously hampered by existing correlation among the predictor variables, i.e. some predictors are linear combinations of others. This phenomenon is also referred to as *collinearity*. Collinearity among predictors may lead to imprecise estimates of the regression coefficients (e.g. misleading signs of the coefficients) and cause the t-tests to fail in revealing the significance and thus the importance of predictor variables (Faraway 2002). Moreover, a high amount of collinearity may indicate the presence of redundant predictors in the model, which should in any case be avoided according to the principle of parsimony. Especially the exhaustive predictor

variables of the present study (chapter 4.7.2) were expected to show a considerable amount of collinearity, since statistical parameters such as mean and median or 90%-quantile and maximum can be located very close to each other (mean and median can even be identical). The amount of collinearity between all auxiliary predictor variables was investigated by inspecting the *correlation matrix* of the predictors and the computation of the *condition number* κ as well as the *variance inflation factor* (VIF): while the correlation matrix of the predictors (comprising the Pearson correlation coefficients) can only reveal *linear* and *pairwise* correlations, the condition number κ can also be used to assess the presence of collinearity. The latter was calculated for each predictor p_i using the respective design matrix *X*, where κ is defined as the square root of the largest eigenvalue λ_i of $X^T X$ divided by the eigenvalue λ_{p_i} of $X_{-i}^T X_{-i}$ (equation 4-60).

$$\kappa_i = \sqrt{\frac{\lambda_{p_1}}{\lambda_{p_i}}} \qquad 4\text{-}60$$

A condition number larger than 30 was assumed to indicate collinearity for the respective predictor (Weisberg 1985).

Deriving the VIF requires to perform a multiple regression of each predictor as response variable with all other predictors as predictor variables. The VIF for each predictor p_i is then calculated using the coefficient of determination R_i^2 of the respective model according to equation 4-61. The *R*-package *HH* (Heiberger 2013) already provided the automatic calculation of VIFs for the given design matrix *X*. The square root of the VIF approximately tells how many times larger the standard error of the predictor's regression coefficient is than it would be without collinearity (Faraway 2002).

$$VIF_i = \frac{1}{1 - R_i^2} \qquad 4\text{-}61$$

The *goodness of fit criteria* used in the variable selection procedure were the *Akaike Information Criterion* (AIC), the *Mallow's* C_p Statistic and the *adjusted* R^2 (R_{adj}^2). Considering that the more predictors will be used in a model, the better the model fit will become (overfitting), these methods aim at quantifying how well a model statistically fits while also accounting for the number of predictors used to achieve the fit. These methods are not based on hypothesis-testing and do not underlie the uncertainty of the t-tests due to collinearity effects. The AIC is defined as:

$$\begin{aligned} AIC &= -2\max(\log likelihood) + 2p \\ &= const + n\log(RSS/n) + 2p \end{aligned} \qquad 4\text{-}62$$

The AIC uses the so-called *deviance* ($n\log(RSS/n)$) as a measure for the goodness of fit, with a small deviance indicating a good fit of the model. Larger models will fit better and have a smaller deviance, but at the cost of using more parameters. Consequently, the best model choice would be a balance between fit and model size. Therefore, the AIC penalizes the achieved deviance by adding two times the number of predictors used in the model. When comparing different models based on the AIC criterion it is suggested to select that one which minimizes the AIC. The AIC selection method was applied in the framework of *forward*, *backward* and *stepwise regression* (Draper and Smith 1966), using the AIC as decision criterion which is obtained by adding or removing a certain predictor to or from a model. The potential number of models tested against each other can become considerably large due to the number of predictors (p predictors would generate 2^p models to be calculated). To keep the effort as small as possible, the AIC based variable selection was performed using the *stepAIC* procedure of the *R*-package *MASS* which evaluates only a subset of potential models within backward, forward and stepwise regression (Venables and Ripley 2002).

In comparison, the Mallow's C_p uses the ratio of the residual square sum (RSS) from a model with a subset of *p* predictors and the estimated error variance from the model with *all* predictors to express the quality of the model fit. This term is penalized by adding two times the number of predictors *p* in the model minus the number of observations *n*:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n \qquad \text{4-63}$$

For the model using all predictors, it holds that $C_p = p$ exactly. The Mallow's C_p thus has the nice property that, if a model with p predictors fits well, then $\mathbb{E}(RSS_p) = (n-p)\sigma^2$ and $\mathbb{E}(C_p) \approx p$ (Faraway 2002). Conclusively, a model with good fit will have C_p very close to p, whereas a model with bad fit will show C_p much larger than p. The C_p-values were plotted against *p* for all calculated models using the *faraway*-package in *R* (Faraway 2011) under the assumption that the best models will be located close to the $C_p = p$ line (identity line). Accounting for parsimony, models with small p as well as C_p around or less than p should be preferred. The calculation of the C_p-values for all possible models was done using the *leaps*-package in *R* (Lumley and Miller 2009).

Finally, the adjusted R^2 (R^2_{adj}) is an extension of the R^2 which also accounts for the number of predictors within a model (Srivastava et al. 1995). The R^2 is calculated by $1 - RSS/TSS$ (with *RSS* being the residual square sum and *TSS* being the square sum of the response variable) and can only increase if more and more predictor variables are added to the model. In comparison, the R^2_{adj} is calculated by using the number of observations *n* and the number of predictors *p* in the model (equation 4-64) and can only be increased by adding meaningful predictors.

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2) \quad \text{4-64}$$

The R^2_{adj} was calculated for all possible predictor variable combinations using again the *leaps*-package of *R*. Afterwards, those models having produced the 50 best R^2_{adj}-values were investigated for the respective predictor combinations.

4.8.5 Model Validation

The performance of the resulting global regression models was analysed by plotting the model predictions against the terrestrial response variable for the entire inventory area. It was also checked how well the respective global regression models perform within the individual small areas. This was done by performing the same procedure for the fitted values and the response variables within each small area individually.

In case that the final large model consisted of a subset of the predictors that had been chosen for the final reduced model, it could be checked whether the large model contributes a significant improvement to the fit of the reduced model. This can be done by applying a partial F-test, defining the null hypothesis H_0 that the additional predictors in the large model do not have an effect on the response. The respecting test statistic has an F-distribution with $p_{lm} - p_{rm}$ (with p being the number of predictors in the models) and $n - (p_{lm} + 1)$ degrees of freedom. The test statistic reads as:

$$F = \frac{n-(p_{lm}+1)}{p_{lm}-p_{rm}} \frac{RSS_{H_0} - RSS_{H_A}}{RSS_{H_A}} \sim F_{p_{lm}-p_{rm},n-(p_{lm}+1)} \quad \text{4-65}$$

As already mentioned, the selected predictors in the global models were also used when applying them for small area estimations. A special case of the small area estimators in this study was the introduction of an indicator variable for a respective small area in the reduced and large model as a mathematical technique to achieve the zero mean residual assumption within the small areas (chapter 4.5). Among the various possible versions of how to use the indicator variables, the following were considered in the present study:

- All indicator variables for the small areas are introduced at once (four indicator variables in the model, whereas the intercept is implicated by the sums of the four indicator variables).
- Only one indicator variable per small area is introduced at a time, i.e. using an indicator variable for the kth small area.

For all these model variations, the coefficient of determination R^2 as a measure of the respective model performance was calculated for each small area as well as for the entire inventory area by comparing the respective model predictions to the terrestrial response. It was further investigated if:

- the terrestrial timber volume significantly differs between the small areas.
- the introduction of indicator variables in the global regression model were statistical significant in the *model-dependent* sense
- significant interactions between the indicator variables and the (most powerful) predictor variables were indicated in the *model-dependent* sense

It should be emphasized that the motivation for testing potential differences in the intercepts and slopes of the small areas was purely to improve the understanding of the respecting regression model performance within the small areas. Referring to chapter 4.4, the introduction of the indicator variables within the *design-based* framework is a mathematical technique which insures the zero mean residual property within the small areas to provide a simpler and better variance estimate and to insure unbiased small area estimates. It has however nothing to do with variable selection. That means that indicator variables should be introduced even if they are not significant in the *model dependent* sense. Furthermore, the significance of an interaction term does not imply to include this interaction in the regression model, as this would again contradict the motivation of the proposed small area estimators, i.e. *not* to fit individual models for each small area, but to "borrow strength" from the global model while insuring the zero mean residual in the small areas by introduction of indicator variables: in other words, the global model provides a large number of observations on which the main relationship between auxiliary variable(s) and terrestrial response can be derived. It is clear that the *design-based* variance can be further reduced by also considering different slopes, i.e. fitting separate models for each small area, but this is neither always possible nor adequate due to considerable few terrestrial observations in the small areas. With respect to this, one borrows the explanatory strength from the global model, while insuring unbiased estimates for the small area by introducing the indicator variable. Additionally, especially under the assumption that the selection of the small areas within the global inventory area is in most cases (as also in the case of the present study) completely arbitrary (e.g. not chosen according to vegetation height zones), the slopes are assumed not to differ considerably.

The tests for different intercepts and slopes in the *model-dependent* sense were carried out by means of one-way analysis of variance (ANOVA) and analysis of covariance (ANCOVA), using the type III square sums. All decisions for potential significance were done on a significance level of α=5%. With respect to the first investigation point, the applicability of a parametric ANOVA was insured by testing if the terrestrial response variable was normally distributed within each small area (group variable) and showed equal variance in each small area. This was done by the application of the Shapiro-Wilk test for normality and Levene's test for homogeneity of variance across groups (Fox and Weisberg 2010). If these assumptions were not satisfied, the non-parametric Kruskal-Wallis-Test (Wolfe and Hollander 1973) was considered instead. The ANCOVAs were carried out under the following assumptions: equal error variance within the groups, normality of the model residuals, uncorrelated errors and zero mean of the model residuals.

4.9 Computation Time of Auxiliary Variables

The separation of the auxiliary variables into exhaustive and non-exhaustive variables was based on the assumption that the non-exhaustive variables require considerably more computation time than the exhaustive variables. In order to investigate whether these assumptions were justified in the present case, the computation time was tracked for each of the auxiliary variables when being calculated at the location $x \in s_2$ of the terrestrial sample points including the performance of the boundary adjustment (this calculation was done within the commercial software *MATLAB,* vers. R2011b). Thereby, it was possible to carry out an analysis of the computation time based on 67 observations for each auxiliary variable. A first visual inspection for potential differences in the computation time between the variables was done by means of boxplots for all variables. Additionally, a test for equal means in the computation time for all variables was considered by the application of an ANOVA under the required preconditions of equal variances of the computation times among the variables as well as normal distribution of the computation times for all variables. In case the ANOVA revealed significant differences in the computation time among at least one of the auxiliary variables on a significance level of α=5%, Tukey's Honest Significant Difference (HSD)-method (Yandell 1997) was considered to reveal differences between certain variables. If the ANOVA assumptions were not met, the application of the non-parametric Kruskal-Wallis test and the pairwise Wilcox-Test with Bonferroni correction for the p-values was considered instead.

Furthermore, the auxiliary variables were coded into exhaustive and non-exhaustive variables according to the classification made in section 4.7.1 and 4.7.2. This allowed for a visual inspection of the computation time distributions of the two groups via boxplots and could give a first impression if the classification was justified. In addition, an unpaired t-test was considered to test the two groups for equal mean computation times. The precondition to apply the t-test was normal distribution of the computing times within the two groups. If this assumption was not satisfied, a Wilcoxon rank sum test (Wolfe and Hollander 1973) was considered instead.

Additionally, it was of particularly interest to compare the computing times for deriving the finally used auxiliary information at all first-phase sample locations (derived by using ArcGIS and *MATLAB*) to the computing time needed to derive the finally selected exhaustive variables over the entire forest area of the study site (realized within the programming language *Python*, vers. 2.7.2, and the *Python*-package *arcpy*). With respect to the alternative reduced model with only the *mean canopy height* as predictor variable, it was of further interest to investigate whether the true mean of the *mean canopy height* could be computed significantly faster within one step (i.e. without the use of squares). This was alternatively done within the ArcGIS software (vers. 10.1) by calculating the mean for all raster cells of the entire CHM covered by the forest definition.

All processing was done using a Dell Latitude E6420 laptop with an Intel® Core™ i5-2520M processor (2.5 GHz) and 4 GB RAM internal memory (64 Bit operating system, Windows 7 Professional).

4.10 Global and Small Area Estimations

After the predictor variables for the regression models had been selected, the estimators introduced in chapter 4.4 and 4.5 were finally applied to the current inventory area. It should be mentioned again that in contrast to the entire variable selection procedure (chapter 4.8) which was carried out by the application of methods belonging to classical, *model-dependent* statistics, the estimations were derived by estimators in the *design-based* approach (chapter 4.3, 4.4 and 4.5). Several variations of the proposed estimators are possible in terms of using the indicator variables. A brief description of the estimators that were actually calculated is therefore provided by Tab. 3 and Tab. 4. The generalized regression estimator (*greg*) was also calculated using the *alternative* reduced model (only with *Mean* as predictor variable). The estimators were subsequently compared by their point estimates (estimate of the standing timber volume per hectare) and their standard errors (square root of the estimated variance). The calculation of the estimations was provided by the co-supervisor of this thesis, who had already implemented the respecting estimators under the statistical software *SAS* (vers. 9.2).

Tab. 3: Estimators used to estimate the standing timber volume for the entire inventory area *F*

Estimator	*Description*
$\hat{Y}_F$	One-Phase Estimator (here, the point estimate is referred to as the *sample mean*)
$\hat{Y}_{F,reg}$	Regression Estimator considering the error in the regression coefficients
$\hat{Y}_{ext,F,reg}$	Regression Estimator *without* considering the error in the regression coefficients (external model approach). The zero mean residual assumption over *F* leads to the same point estimate as $\hat{Y}_{F,reg}$, but a different variance
$\hat{Y}^{(k)}{}_{F,reg}$	Extended Regression Estimator using only the indicator variable for the kth small area *and* considering the error in the regression coefficients
$\hat{Y}^{(c)}{}_{F,reg}$	Extended Regression Estimator using indicator variables for all small areas *and* considering the error in the regression coefficients
$\hat{Y}_{F,greg}$	Generalized Regression Estimator considering the error in the regression coefficients
$\hat{Y}_{ext,F,greg}$	Generalized Regression Estimator *without* considering the error in the regression coefficients (external model approach). The zero mean residual assumption over *F* leads to the same point estimate as $\hat{Y}_{F,greg}$, but a different variance
$\hat{Y}^{(k)}{}_{F,greg}$	Extended Generalized Regression Estimator using only the indicator variable for the kth small area *and* considering the error in the regression coefficients
$\hat{Y}^{(c)}{}_{F,greg}$	Extended Generalized Regression Estimator using indicator variables for all small areas *and* considering the error in the regression coefficients

Tab. 4: Small area estimators used to estimate the standing timber volume for each small area *G*

Estimator	*Description*
$\hat{Y}_G$	One-Phase Estimator
$\hat{Y}_{ext,G,reg}$	Regression Estimator *without* considering the error in the regression coefficients and without using indicator variables (external model approach).
$\hat{Y}^{(k)}{}_{G,reg}$	Extended Regression Estimator for small area *k* using only the indicator variable for the kth small area *and* considering the error in the regression coefficients
$\hat{Y}^{(c)}{}_{G,reg}$	Extended Regression Estimator with all indicator variables for the small areas included, but without intercept. Also considering the error in the regression coefficients
$\hat{Y}_{ext,G,greg}$	Generalized Regression Estimator *without* considering the error in the regression coefficients and without using indicator variables (external model approach)
$\hat{\bar{Y}}^{(k)}{}_{G,greg}$	Extended Generalized Regression Estimator for small area *k* using only the indicator variable for the kth small area *and* considering the error in the regression coefficients
$\hat{\bar{Y}}^{(c)}{}_{G,greg}$	Extended Generalized Regression Estimator with all indicator variables for the small areas included, but without using intercept. Also considering the error in the regression coefficients

5 Results

5.1 Forest Area Definition

Fig. 5-1 (left) shows the forest mask for the complete study site based on the selection of the TLM3D forest classes. The quality of the mask was validated via a detailed visual inspection of the superposition of the forested areas (appearing in the CHM and a Spot5 true color satellite image) by the forest mask at 50 randomly chosen locations in the study site. The main quality criteria were the following: firstly, the mask should appropriately describe the forest edges, and secondly, the mask should adequately present gaps inside the forest cover. Fig. 5-1 (right) shows an example for such a visual inspection. Based on this validation, the forest mask provided a suitable forest definition for the study objectives.

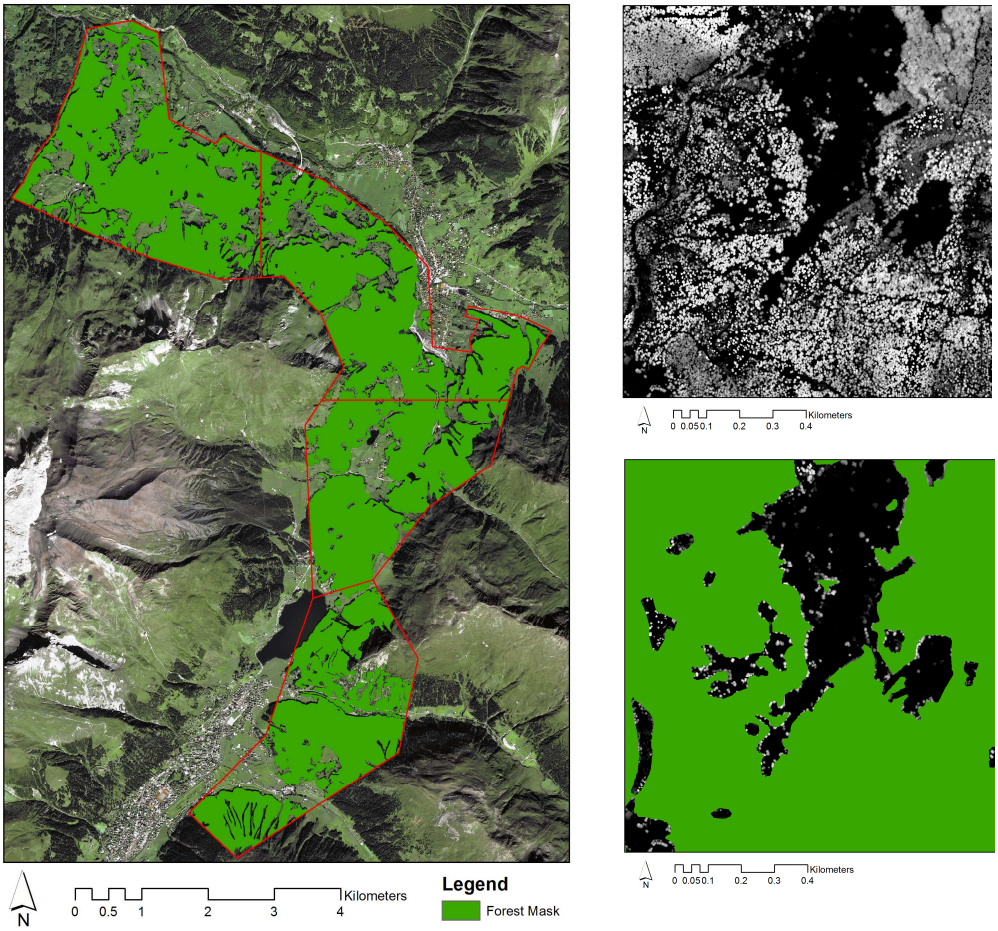


Fig. 5-1: *Left:* Forest mask of the inventory area based on the TLM land cover map. The forest mask is used as the forest definition in the current inventory. *Right*: The forest mask appropriately covers the stocked regions in the CHM.

Based on this mask, the forest area extent was calculated for the complete study area as well as for the small areas (Tab. 5).

Tab. 5: Forest area of the entire study area and the small areas according to the forest mask (forest definition)

	Area [ha]	*Forest Area [ha]*	*Forest Area [%]*
Entire Study Area	2887.39	1974.49	68.4
Small Area 1	846.48	586.60	69.3
Small Area 2	762.58	531.88	69.7
Small Area 3	593.85	425.55	71.7
Small Area 4	684.48	430.46	62.9

5.2 Computation of Auxiliary Variables

This chapter will present the results of the regression analysis which was performed in order to find a suitable model for estimating a trees timber volume by its height. This model was then implemented in the algorithm to derive the non-exhaustive variable *Volume Density*.

In order to provide an estimate for the stem timber volume of a LiDAR detected tree, a regression model had to be formulated which is capable of modeling the stem volume of a tree as a function of its tree height (chapter 4.7.1). This regression model has been established using the available NFI tree data of the domains Pre-Alps and Alps. For optimizing this model with respect to trees, it was attempted to exclude shrub species by selecting all species whose percentage share in the dataset was larger than 1%. As it turned out, this 1%-threshold was not only suited for excluding the shrub species, but also to reduced the dataset to the main coniferous and three of the five main broadleaf tree species according to NFI (Brändli 2010). The reduced dataset, on which the regression analysis was then based yet comprised 8121 observations in total, with a proportion of 59.1% spruce, 12.9% silver fur, 8.5% European larch, 2.6% Scots pine, 2% Swiss pine, 10.4% European beech, 2.5% great maple and 2% European ash (Fig. 5-2).

Fig. 5-2**:** Percentage share of tree species [%] in the second stage NFI-inventory of the domains Pre-Alps and Alps. A 1% threshold was used to exclude the shrub-species and reduce the dataset to the main tree species according to NFI.

A log-transformation was applied to both the tree volume and the tree height data. A linear regression (first polynomial) of the transformed data revealed an R^2 of 0.65. A visualization of the regression lines in the log-scale and after back transformation in the original scale (using the bias correction for the regression coefficients) is presented in Fig. **5-3** and reveals the expected allometric relationship. The relative high amount of scattering was attributed to the wide range of sites included in the dataset. It suggests that for a given tree height the DBH and therefore the trees timber volume can considerably vary due to natural variability, its growing space and competitive situation in the stand (Pretzsch 2009). This phenomenon is well-known in the field of forest yield and growth and goes back to the self-thinning rule described by Yoda and Reinecke (Pretzsch and Biber 2005).

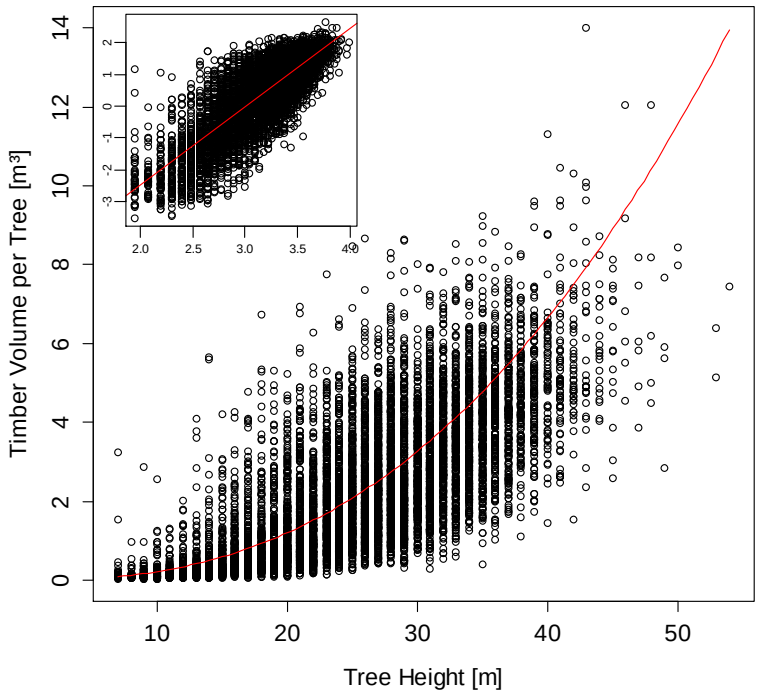


Fig. 5-3: Tree timber volume vs. tree height in original (large window) and double-logarithmic scale (small window). The red lines visualize the regression function in the respective scale.

The regression analysis revealed the following relationship:

$$Tree\ Timber\ Volume_{predicted} = e^{-7.44308+2.47508 \cdot \ln(Tree\ Height)} \cdot e^{\frac{\sigma^2}{2}} \qquad 5\text{-}1$$

with $\frac{\sigma^2}{2} = 0.2063553$.

The residual diagnostics revealed no severe violation of the model assumptions. The Tukey-Anscombe plot (Fig. 5-4, left) shows that the model residuals sufficiently scatter around zero (also indicated by red smoothing line). The Scale-Location plot (Fig. 5-4, right) can be used to investigate the variance of the model residuals. In contrast to the Tukey-Anscombe plot, the Scale-Location plot visualizes the square root of the absolute residuals against the predicted values. This ensures that an inconstancy in the residual variance cannot be visually compensated by residuals of unequal signs. For the current model residuals, the Scale-Location plot shows a constant error variance for the first half of the fitted data, whereas a slight non-constant error variance is recognizable for the second half of the fitted data. This problem could not be overcome by alternative models and regression methods which are less sensitive to a non-homogeneous error variance (standard major axis- (SMA) and non-linear regression methods). Consequently, this drawback had to be accepted, leading to an increasing prediction uncertainty for larger tree heights in the original scale.

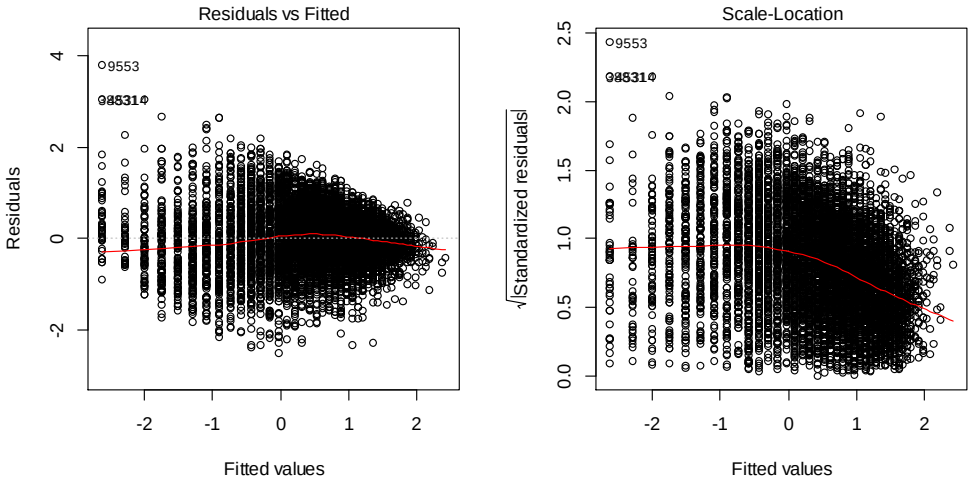


Fig. 5-4: *Left*: The Tukey-Anscombe plot hints at a satisfaction of the zero expectation assumption of the model residuals. *Right:* The Scale-Location plot shows a slightly non-constant error variance for the second half of the data.

Fig. 5-5 finally shows the tree volume of the NFI versus their model predictions in the log- and original scale. It can be seen that the resulting point clouds scatter symmetrical along the identity-line. This indicates that the model does not systematically over-or underestimates the true tree volume. The model was finally decided to be applicable in order to estimate the timber volume of a LiDAR detected tree by its LiDAR tree height and the model equation (equation 5-1) was implemented in the *Volume Density* algorithm.

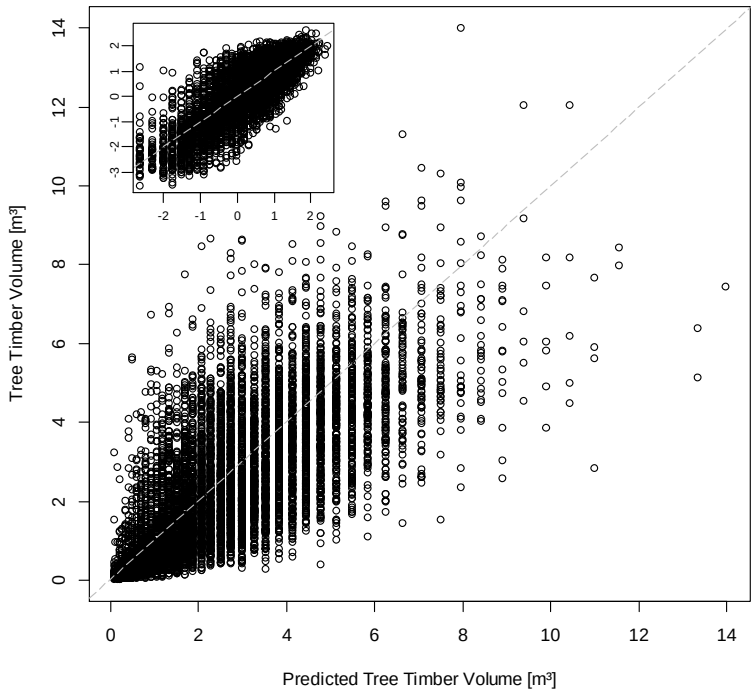


Fig. 5-5: Predicted tree timber volume vs. response tree timber volume in original (large window) and double logarithmic scale (small window). The identity line is indicated by the dotted grey line.

5.3 Selection of Predictor Variables

The selection of the auxiliary variables finally used as predictors in the regression model(s) was carried out by comprehensive, *model-dependent* explorative data analysis. However, the results of these analyses were considered to only be a guideline to find a final set of predictors. Great importance was also attached to the interpretability of the results.

5.3.1 Simple Linear Regression Models

First, a simple linear regression was computed for each auxiliary variable with the terrestrial timber volume as response variable (Fig. 5-6). The analysis revealed that the exhaustive variable *Mean* and *Median* as well as the non-exhaustive variable *Volume Density* each achieve the highest coefficients of determination among all variables (R^2 of 0.5). Also *Q90* and *Q75* performed quite well, with R^2's of 0.47 and 0.45, whereas *Q25* showed considerably less correlation to the response. The coefficients of determination of the remaining exhaustive variables ranged between 0.15 and 0.29. Surprisingly, the LiDAR estimated *stem number* did show only weak correlation to the response ($R^2 = 0.08$). However the negative slope of the regression line was significant on a 5% level and seemed to indicate the well known rule of self-thinning. In accordance to this rule, increasing timber volume within a stand can only be realized by decreasing the stem number (Pretzsch and Biber 2005). So far, the analysis suggested that all of the considered auxiliary variables are related to the observed terrestrial timber volume. Still, the identification of a best subset of predictor variables for the reduced and large model was also based on the results of the multiple regression analysis.

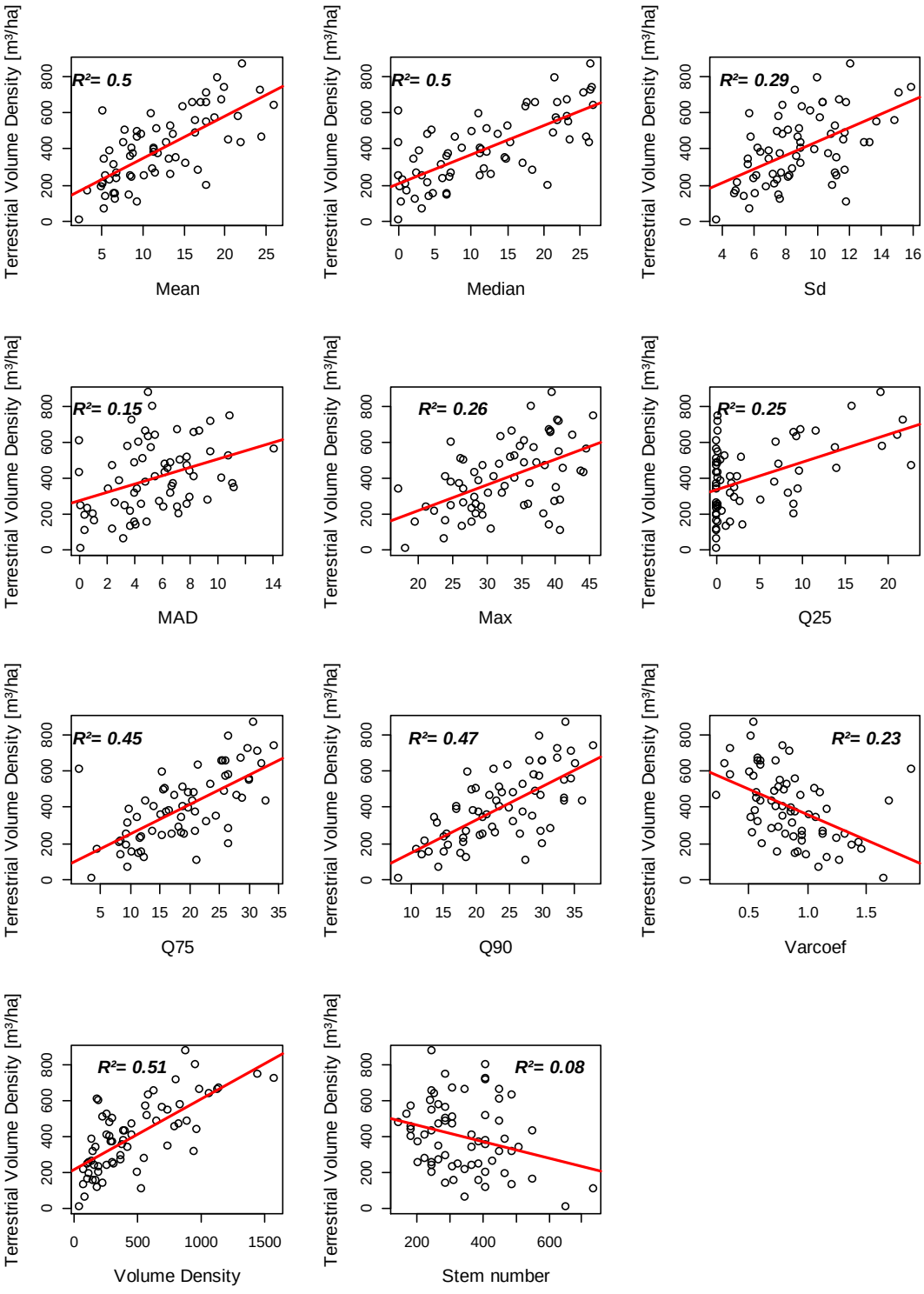


Fig. 5-6: Simple linear regression of all auxiliary variables against the observed terrestrial response variable (terrestrial volume density [m^3/ha]). The exhaustive variables *Mean* and *Median* as well as the non-exhaustive variable *Volume Density* achieve an R^2 of 0.5.

5.3.2 Multiple Regression Analysis

Before the variable selection methods (chapter 4.8.4) based on the multiple regression analyses were performed for the reduced and large model, the complete set of predictors was investigated for the presence of collinearity. A visual inspection of a graphical representation where the predictor variables derived at the terrestrial sample points were plotted against each other (Fig. 5-7) revealed a high amount of *linear* correlations between several predictors. The severity of correlation was quantified by calculating the matrix of Pearson correlation coefficients for all predictors (Fig. 5-8). With respect to Fig. 5-8 it becomes obvious that all predictors show a certain correlation to at least one of the remaining predictors. Especially the variables *Mean*, *Median*, *Q75*, *Q90* and *Volume Density* revealed several correlation coefficients greater than 0.8. However, Fig. 5-6 suggests that most of the predictors are also correlated to the response variable, i.e. the terrestrial volume density. Since the Pearson correlation coefficient only detects linear relationships, it was visually insured that no other kind of relationship between predictors was missed.

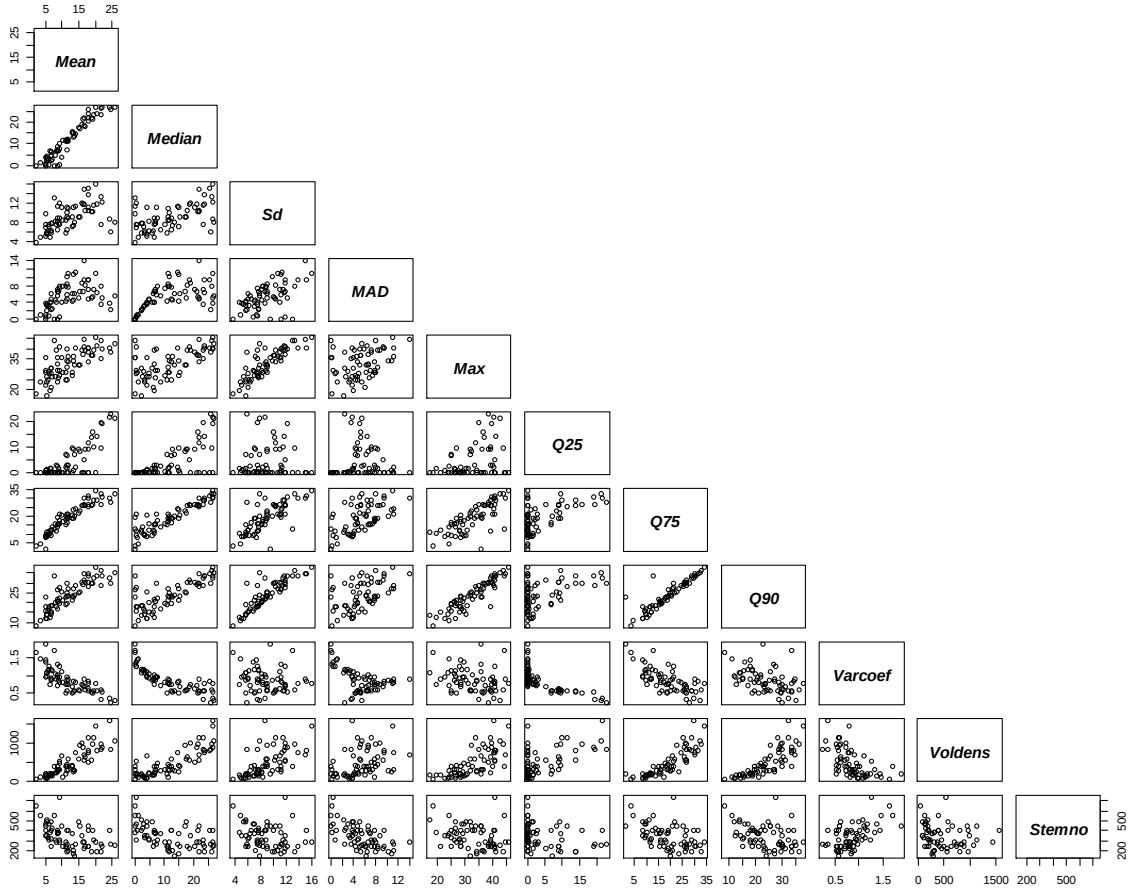


Fig. 5-7: Plotting all predictor variables against each other using the data of the terrestrial sample points (n=67).

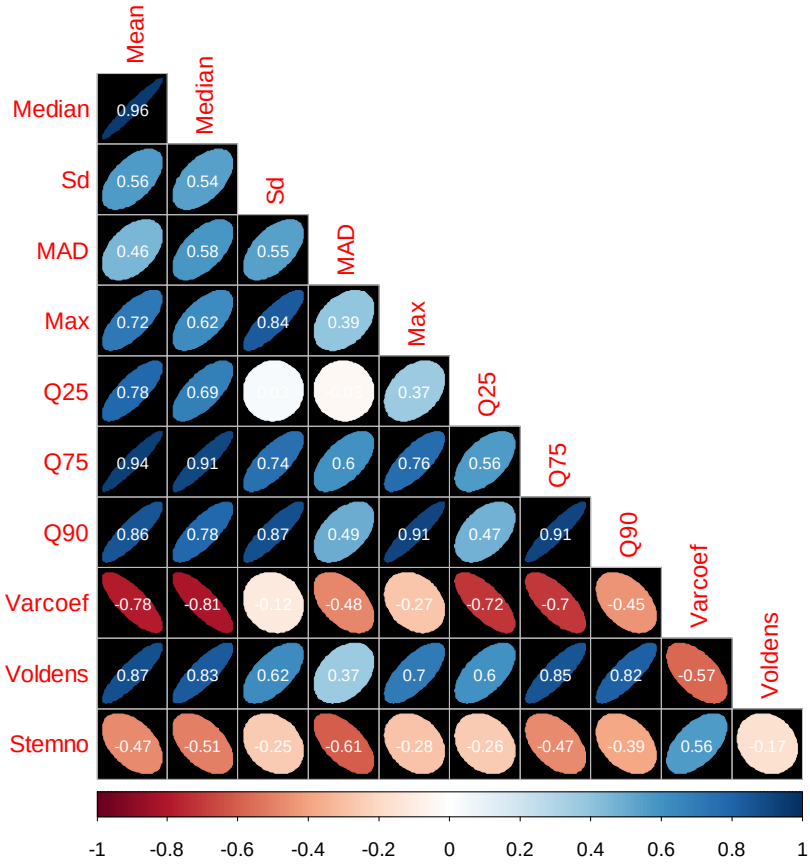


Fig. 5-8: Visualization of the correlation matrix of all predictors. Increasing elliptical shapes indicate increasing linear correlation, the orientation of the ellipses indicate a negative (to the left) or positive (to the right) relationship. The Pearson correlation coefficients are given by the numbers and colors of the shapes.

As the correlation matrix can only reveal pairwise collinearity, also the condition number κ and the variance inflation factor *VIF* (chapter 4.8.4) were calculated for all predictors. The results are provided in Tab. 6. It can be seen that there is a wide range in the eigenvalues and except for *Median* all condition numbers exceed the critical threshold of 30, implying that collinearity problems might be caused by more than just one linear combination of predictors. The *VIF*–values do not seem as critical as the condition numbers. The most worrying predictors here are *Mean*, *Median*, *Sd*, *Q75* and *Q90*. For example, the *VIF*'s square root for *Mean* suggests that the standard error of its regression coefficient can be 24.75 times larger than it would be without collinearity to other predictors. Conclusively, all applied methods confirmed the existence of substantial collinearity among the predictors and thereby emphasized the necessity to limit the number of predictors in the regression models down to the most meaningful ones, avoiding the usage of probably redundant predictors. It also confirmed the assumption that the variable selection methods relying on goodness of fit criteria should in this case be preferred to the ones using testing-based procedures, since the partially huge variance inflation can lead to misleading individual parameter t-tests.

Tab. 6: Condition number κ and square root of the variance inflation factor (VIF) for all predictors

	Mean	*Med*	*Sd*	*MAD*	*Max*	*Q25*	*Q75*	*Q90*	*Varcoef*	*Stem number*	*Volume Density*
κ	1.0	2.93	46.22	110.91	164.45	298.8	429.8	535.78	1036.47	3242.9	5463.7
$\sqrt{VIF}$	24.75	11.19	8.53	2.59	3.26	4.40	9.33	10.92	3.38	2.82	1.78

The results of the AIC based variable selection (*forward, backward and stepwise regression*) for the reduced model are presented in Tab. 7. Using the backward elimination, the lowest AIC was found by the predictor combination *Mean*, *Sd*, *Max* and *Q75* (AIC=646.10), achieving an R^2 of 0.65. All these predictors were also significant on a 5% significance level (individual parameter t-test). The same predictor combination was found by the stepwise regression procedure. Using the forward selection, the best identified subset of predictors were *Median*, *Q90*, *Max* and *Q75* (AIC=649.76). Also these predictors were significant on a 5% level and the corresponding R^2 was only slightly lower (0.63).

Tab. 7: Results of the backward, forward and stepwise variable selection for the reduced model according to AIC criterion. Potential predictor variables were Mean, Median, Sd, MAD, Max, Q25, Q75 and Q90. * indicates significance of a predictor based on individual parameter t-test (5% significance level).

Reduced model	*Backward elimination*	*Forward selection*	*Stepwise regression*
Proposed Selection	Mean*, Sd*, Max*, q75*	Median*, Q90*, Max*, Q75*	Mean*, Sd*, Max*, Q75*
AIC	646.10	649.76	646.10
R^2	0.6453	0.6253	0.6453

Tab. 8 provides the results of the AIC based variable selection for the large model. Also in this case, the backward elimination- and the stepwise regression procedure revealed identical results: the lowest AIC value was produced by the predictor combination *Mean*, *Median*, *Sd*, *MAD*, *Max*, *Q75* and *Volume Density* (AIC=644.75, R^2=0.68). However, *Median* and *Max* were not significant on a 5% significance level. The forward selection procedure proposed using the predictors *Median*, *Q90*, *Max*, *Q75* and *Volume Density* as the best subset (AIC=646.16, R^2=0.66) with all variables being significant.

Tab. 8: Results of the backward, forward and stepwise variable selection for the large model according to AIC criterion. Potential predictor variables were Mean, Median, Sd, MAD, Max, Q25, Q75, Q90, Volume Density and Stemnumber. * indicates significance of a predictor based on individual parameter t-test (5% significance level).

Large model	*Backward elimination*	*Forward selection*	*Stepwise regression*
Proposed Selection	Mean*, Median, Sd*, MAD, Max*, Q75*, Volume Density*	Median*, Q90*, Max*, Q75*, Volume Density*	Mean*, Median, Sd*, MAD, Max*, Q75*, Volume Density*
AIC	644.75	646.16	644.75
R^2	0.6821	0.6554	0.6821

For all possible predictor combinations within the reduced and large model, the regression models and the corresponding *Mallow's* C_p–values were calculated. Fig. 5-9 shows the respective *Mallow's* C_p-values plotted against the number of predictors used. Within the plots, the models are denoted by the indices of the predictors (1=*Mean*, 2=*Median*, 3=*Sd*, 4=*MAD*, 5=*Max*, 6=*Q25*, 7=*Q75*, 8=*Q90*, 9=*Varcoef*, 10=*Volume Density* and 11=*Stemnumber*). For the reduced model (Fig. 5-9, above), the competition for the best model (less predictors and closest to the $C_p = p$ line) is between the model *1357* with *Mean*, *Sd*, *Max* and *Q75* as predictor variables and model *2578* with *Median*, *Q90*, *Max* and *Q75* as predictor variables. It is reminded that these were also the best subset models identified by the backward and stepwise variable selection according to AIC criterion.

According to the *Mallow's* C_p-plot, the best subset of predictors for the large model (Fig. 5-9, below) are *Mean*, *Sd*, *Max*, *Q75* and *Volume Density* (*135710*) or *Median*, *Q90*, *Max*, *Q75* and *Volume Density* (*257810*). It should be recognized that these predictor combinations are simply the best two predictor sets found for the reduced model being extended by the non-exhaustive variable *Volume Density*.

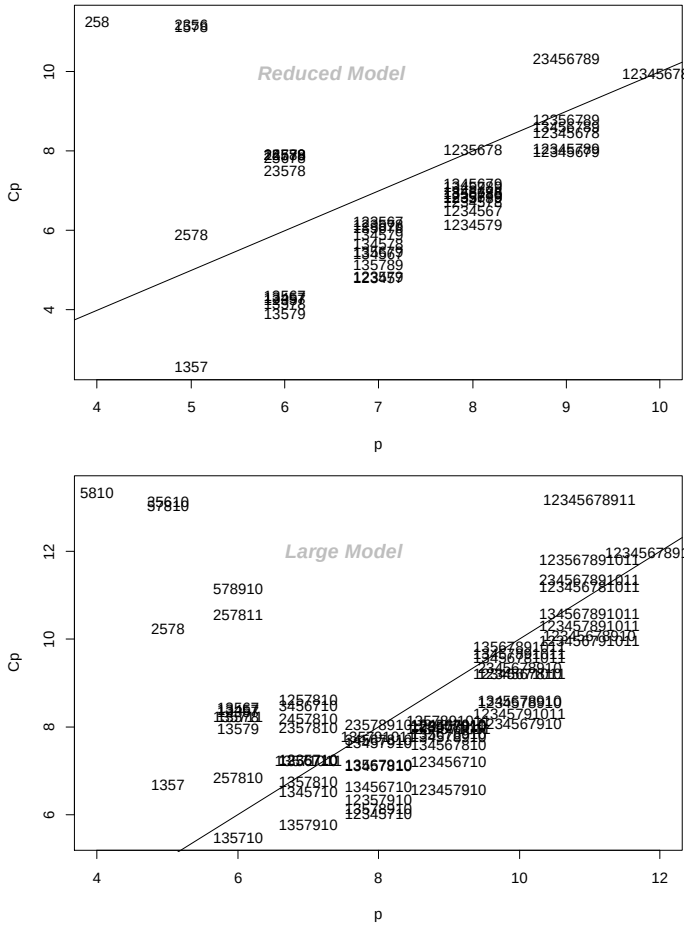


Fig. 5-9: For each predictor combinations of the reduced model (*above*) and the large model (*below*), the Mallow's C_p is plotted against the number of predictors used in the respecting model. The models are denoted by the indices of the predictors (1=*Mean*, 2=*Median*, 3=*Sd*, 4=*MAD*, 5=*Max*, 6=*Q25*, 7=*Q75*, 8=*Q90*, 9=*Varcoef*, 10=*Volume Density* and 11=*Stemnumber*). Smaller models with less than three predictors (above) or less than four predictors (below) are not shown as their C_p-values are large.

One could here consider as well the variable combination *13571* (*Mean*, *Sd*, *Max*, *Q75*) as the potential best model, since it is also located close to the identity line while using only 5 predictors. However, this would not have been in the sense of testing the generalized regression estimator (chapter 4.4.2), as the large model per definition has to include at least one non-exhaustive variable.

Finally, the R^2_{adj} was calculated for all possible parameter combinations of the reduced and large model. The highest R^2_{adj} for the reduced model was found by the predictor combination *Mean*, *Sd*, *Max* and *Q75* (R^2_{adj}=0.62). The next model using only four predictors was located on rank 33 and comprised the predictors *Median*, *Q90*, *Max* and *Q75* (R^2_{adj}=0.60). In case of the large model, the smallest model containing at least one non-exhaustive variable was located on rank 28, using the predictors *Mean*, *Sd*, *Max*, *Q75* and *Volume Density* (R^2_{adj}=0.64). A similar model revealed an R^2_{adj} of 0.63 and used *Median*, *Q90*, *Max* and *Q75* and *Volume Density* as predictor variables.

5.3.3 Model Validation and Final Predictor Selection

Summarizing, all goodness of fit criteria (AIC, *Mallow's* C_p, R^2_{adj}) indicated for the reduced model, that the usage of *Mean*, *Sd*, *Max* and *Q75* **or** *Median*, *Q90*, *Max* and *Q75* as set of predictors performs almost equally good while optimizing the balance between model fit and model size. Both models did not differ much in terms of the presence of collinearity: in the first model, *Mean* and *Q75* are highly correlated (Pearson Correlation Coefficient of 0.94), but leaving out one of the both predictors considerably decreased the fit (R^2 and R^2_{adj}). The same is true for the second model: there is still strong collinearity between *Median* and *Q75*, *Max* and *Q90* as well as *Q75* and *Q90*, but also in this case the fit considerably decreased by leaving out one of these variables. These explorative data analysis showed that several models are possible with essentially the same fit. The final selection can thus not be left to statistical algorithms alone but has to be done by the "manually".

For the reduced model, *Mean*, *Sd*, *Max* and *Q75* were finally selected as set of predictors. The final choice of predictors for the large model was *Mean*, *Sd*, *Max*, *Q75* and *Volume Density*. This predictor constellation was proposed by the *Mallow's* C_p, the analysis of R^2_{adj} and also indicated by the forward and stepwise regression using the AIC criterion (t-tests for *Median* and *MAD* were not significant). Choosing this model had the nice property, that the exhaustive variables in the reduced and large model were the same, and thus the estimated mean of all exhaustive variables could be corrected by their true mean in the framework of the generalized regression estimator (chapter 4.4.2 and 4.7.2). With respect to the interpretability of the models, the only auxiliary variables which can honestly be clearly interpreted are the *Mean* (since it is very intuitive that the mean canopy height must be related to the timber volume) and *Volume Density* (as derived based on a well known allometric function). For this reason, an *alternative* reduced model was considered with only *Mean* as predictor variable. This decision was also motivated by the inspection of the individual linear regressions for each predictor, showing that *Mean* alone already

yields an R^2 of 0.5, compared to R^2= 0.65 when also including *Sd*, *Max* and *Q75*. Tab. 9 provides again an overview of the final set of predictors for the reduced, the alternative reduced and the large model.

Tab. 9: Final set of predictors of the reduced model, the *alternative* reduced model (*alt*) and large model with their respecting regression coefficients (the predictors of the large model were also the predictors used for the "classical" regression estimator (*RE*))

Model	Final set of predictors	AIC	R^2
Reduced model	*Mean*, *Sd*, *Max*, *Q75*	646.10	0.65
Alternative Reduced model	*Mean*	662.75	0.5
Large model / RE-model	*Mean*, *Sd*, *Max*, *Q75*, *Volume Density*	644.66	0.66

The consideration of an alternative reduced model with only *Mean* as predictor was also underlined by an F-test which was used to investigate whether the additional predictors in the large model yield a substantial improvement to the two considered reduced models. In the case of the reduced model, there was just one additional predictor in the large model, i.e. the non-exhaustive variable *Volume Density*. On the contrary, the extension from the alternative reduced model to the large model comprised four variables. The F-tests were calculated using the *type III* residual square sums. A p-value of 0.07807 for the respective test statistic indicated that, strictly speaking, the extension from the reduced to the large model by the variable *Volume Density* did not yield a significant improvement on a significance level of α=5% (Tab. 10). In case of the alternative reduced model, the enlargement to the large model yield a considerable significant improvement of the fit (Tab. 11).

Tab. 10: Partial F-Test for the reduced and large model (all predictors of the reduced model being a subset of the large model). On a 5% significance level, the extension of the reduced model did not yield a significant improvement of the fit.

Model	*RSS (Type III)*	*Diff. RSS (Type III)*	*F*	*p(>F)*
Reduced Model	889750			
Large Model	845245	44504	3.2118	0.07807

Tab. 11: Partial F-Test for the alternative reduced and large model (all predictors of the reduced model being a subset of the large model). On a 5% significance level, the extension of the reduced model did yield a significant improvement of the fit.

Model	*RSS (Type III)*	*Diff. RSS (Type III)*	*F*	*p(>F)*
Alternative Reduced Model	1247796			
Large Model	845245	402550	7.2629	$7.51*10^{-5}$

With respect to the final selection of predictor variables for the reduced and large model, it was now possible to formulate the respective model equations to predict the local density $\hat{Y}(x)$, i.e. the terrestrial standing timber volume on plot level [m^3/ha], at any given sample point. In the *model-dependent* inference framework, all predictors in these regression models had a significant influence on the terrestrial response variable. The validity of these tests rest upon certain residual assumptions (constant error variance, uncorrelated errors, normality of residuals) which were decided to be fulfilled (the respective Tukey-Anscombe-, Scale-Location and QQ-Plots can be found in Appendix B). Thus the testing-based procedures as well as any further t- and F-tests for these models were justifiable in the *model-dependent* sense.

The reduced model yields

$$\hat{Y}(x) = 233.19 + 63.03 \cdot Mean(x) + 77.38 \cdot Sd(x) - 19.96 \cdot Max(x) - 33.56 \cdot Q75(x) \qquad (R^2=0.65)$$

and the *alternative* reduced model with only *Mean* is

$$\hat{Y}(x) = 116.16 + 23.46 \cdot Mean(x) \qquad (R^2=0.5)$$

The large model is

$$\hat{Y}(x) = 276.88 + 54.53 \cdot Mean(x) + 71.77 \cdot Sd(x) - 19.49 \cdot Max(x) - 32.57 \cdot Q75(x) + 0.16 \cdot Volume\ Density(x) \qquad (R^2=0.66)$$

As an additional guideline for variable selection, it was observed how the considered global reduced- and large models perform within the entire inventory area by plotting the predictions $\hat{Y}(x)$ of the reduced- and large model against their corresponding terrestrial observations $Y(x)$. Fig. 5-10 shows the respective inspection for the finally selected reduced and large model. In both cases, the resulting point cloud scatters almost equally around the identity line (which illustrates the position of perfect predictions). This suggested that both models did perform well for the entire inventory area, not leading to a systematic over- or underestimation in the *model-dependent* sense. The same was true for the alternative reduced model. However, plotting the model predictions $\hat{Y}(x)$ against the observed local density $Y(x)$ individually for each small area revealed that the fit in small area 3 was far below average (Fig. 5-11 and Fig. 5-12): while the coefficient of determination R^2 for small area 1, 2 and 4 ranged between 0.6 and 0.8, the R^2 for small area 3 was only 0.36 for the large and 0.46 for the reduced model. A simple linear regression of each predictor of the large model on $Y(x)$ for each small area revealed that in small area 3 the

slopes of all predictors were considerably smaller than in the other small areas, suggesting that here the predictors do provide less additional information as in the other small areas.

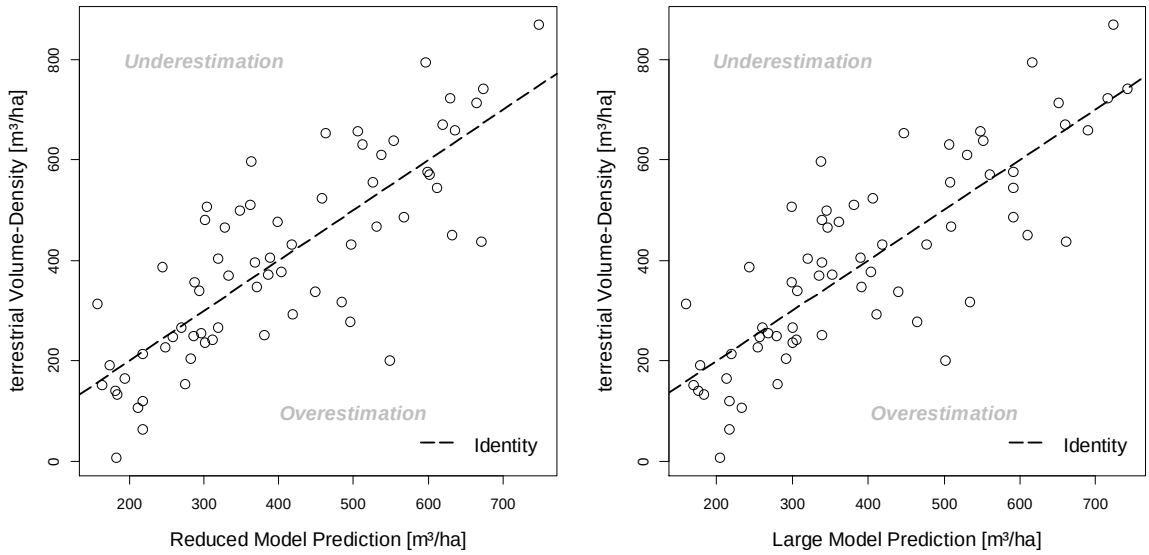


Fig. 5-10: Terrestrial Volume Density [m³/ha] vs. predicted Volume Density by the reduced (*left*) and large model (*right*) over entire inventory area. Both models did not show any systematic over- or underestimation.

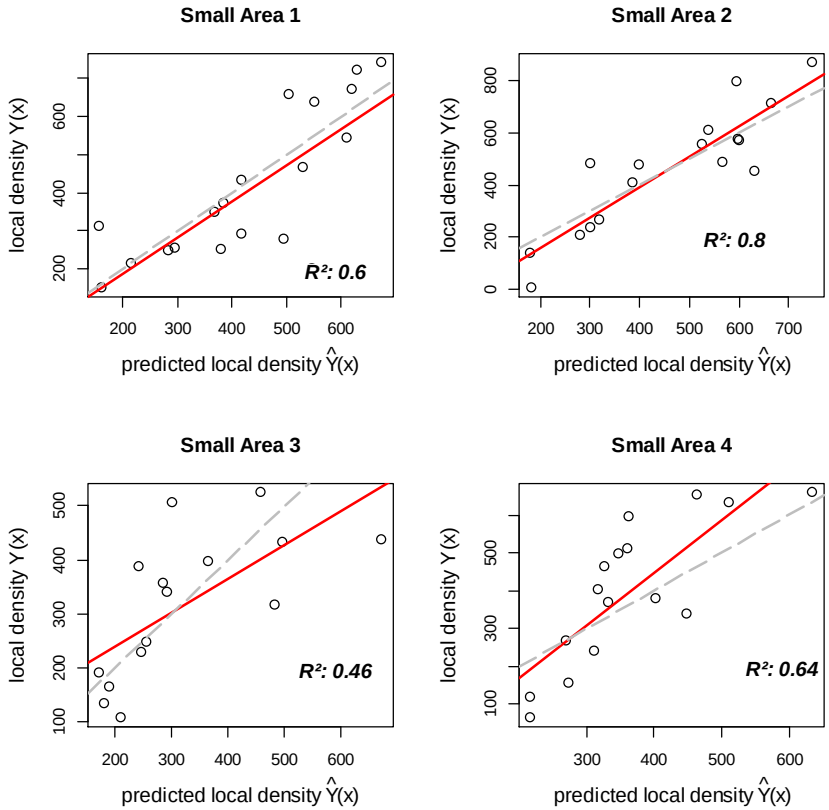


Fig. 5-11: Performance of the global reduced model within the small areas. Red line indicates a linear regression that was used to calculate the R^2 within the small areas. The dotted grey line illustrates the identity line.

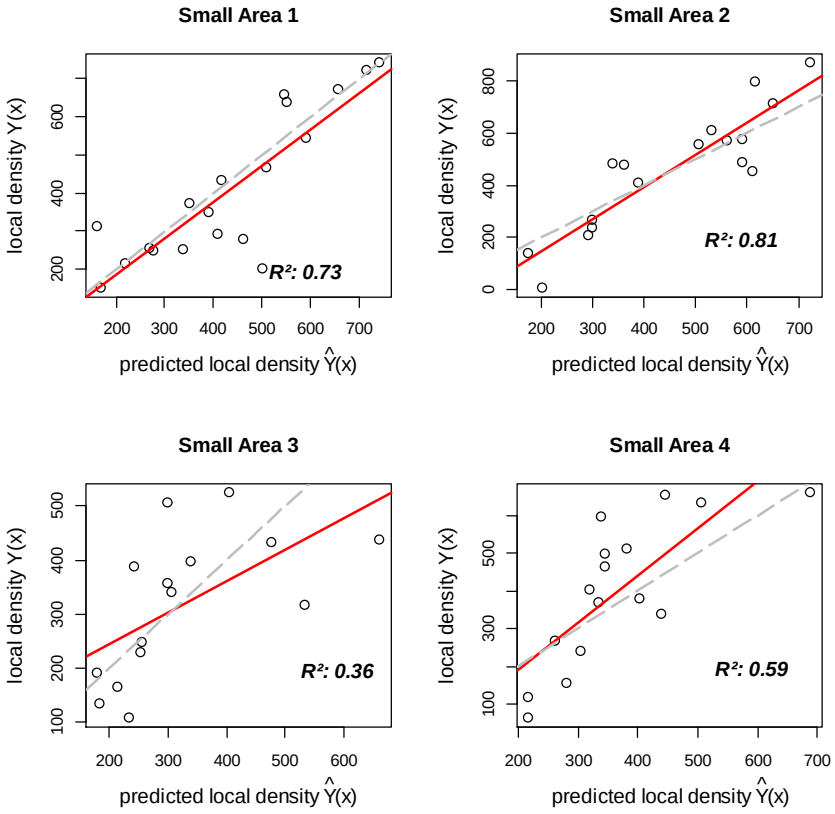


Fig. 5-12: Performance of the global large model within the small areas. Red line indicates a linear regression that was used to calculate the R^2 within the small areas. The dotted grey line illustrates the identity line.

The same analysis was carried out for the reduced and large models after their extensions for small area estimations, i.e. introducing one indicator variable for a small area at a time as well as introducing all indicator variables all at once. To assess the respective model performance for the entire inventory area, the model predictions were plotted against their response while not being distinguished by the small areas. Fig. 5-13 shows the respective graphics for the case of including all indicator variables at once, the illustrations for the remaining models are given in Appendix B. It turned out that introducing one or all indictor variable(s) did not considerably change the model fit, with the R^2's for the entire inventory area varying between 0.65 and 0.68, and thereby almost identical to the R^2's of the global reduced and large model (0.65 and 0.66). Plotting the model predictions against the terrestrial response variable within each small area revealed the same property for all models as already seen when analyzing the global models: whereas small area 1, small area 2 and small area 4 did in all cases achieve R^2's between 0.6 and 0.8, small area 3 revealed R^2's far below average (around 0.3 - 0.4). Fig. 5-14 and Fig. 5-15 illustrate this in the case of the reduced and large model including all indicator variables for the small areas at once (an illustration for the remaining models can also be found in Appendix B).

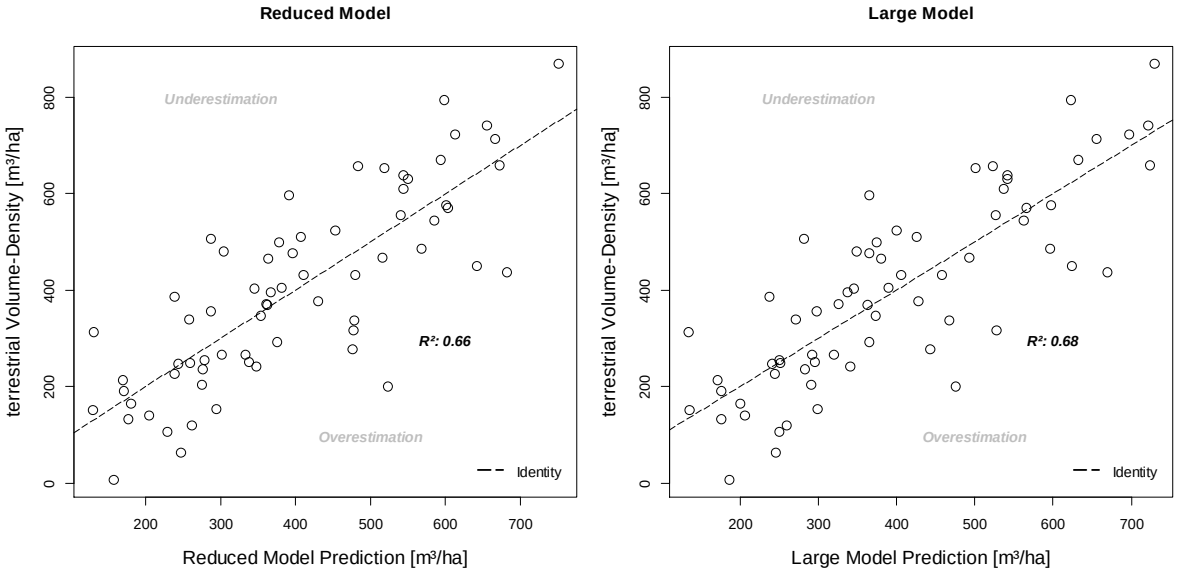


Fig. 5-13: Fit for the entire inventory area using the reduced (*left*) and large (*right*) model with all indicator variables included in the model.

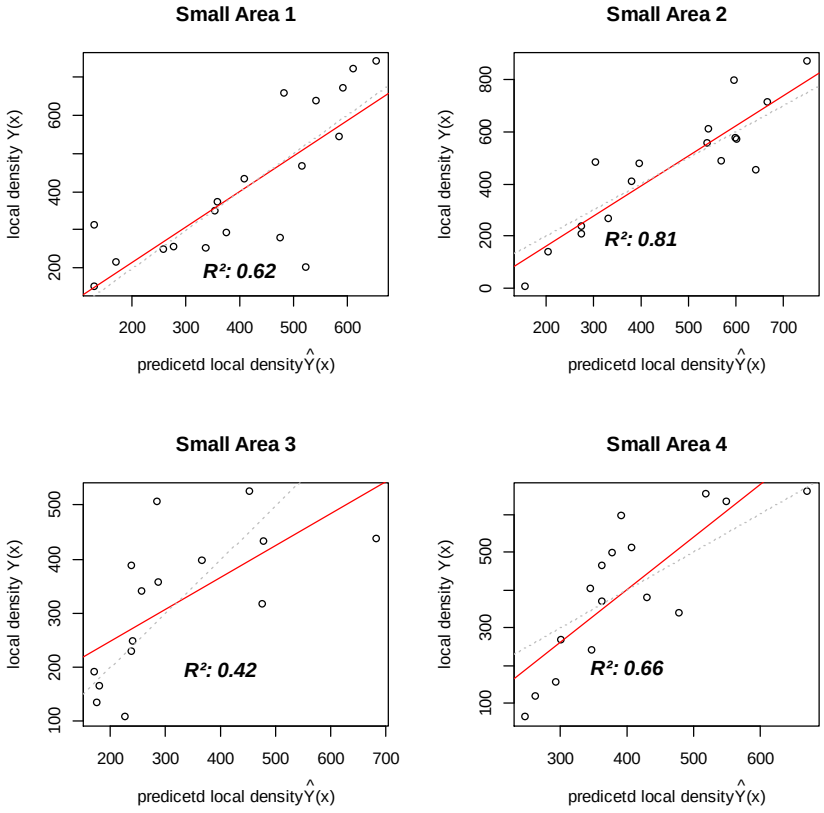


Fig. 5-14: Performance of the reduced model with all indicator variables included in the model. Red line indicates a linear regression that was used to calculate the R^2 within the small areas

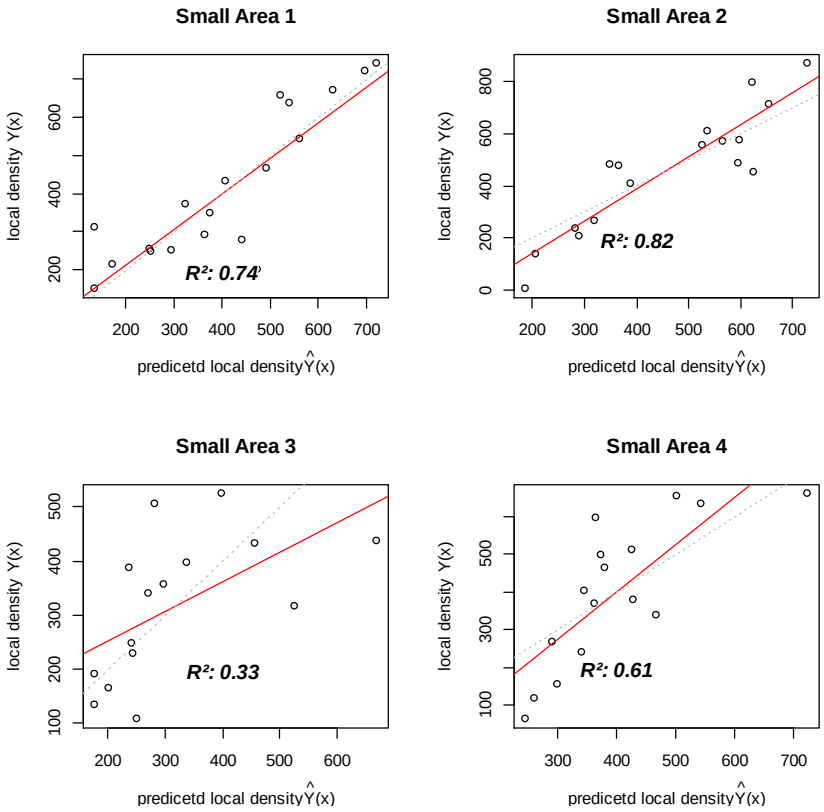


Fig. 5-15: Performance of the large model with all indicator variables included in the model. Red line indicates a linear regression that was used to calculate the $\boldsymbol{R^2}$ within the small areas.

It was further investigated if the indicator variables per small area in the reduced and large model were significant in the *model-dependent* sense when introducing them at a time as well as introducing them all at once. This was carried out by means of ANCOVA analyses. The satisfaction of the model assumptions for the ANCOVA analysis were insured by the application of the Shapiro-Wilk test for normality, Levene's test for homogeneity of variance across groups (Fox and Weisberg 2010) and a visual inspection of the Tukey-Anscombe- as well as the Scale-Location Plots for insuring constant error variance and uncorrelated errors (Appendix B). For the finally selected models, the indicator variables were in all cases not significant on a level of α=5%, indicating that the intercepts did not differ significantly between the small areas and using the same intercept for all small areas would be appropriate in the *model-dependent* sense. Nevertheless, within the *design-based* approach applied in this study the introduction of the indicator variables had nothing to do with variable selection, but was purely a mathematical method to achieve the zero mean residuals also within the small areas and to allow for an easier and better calculation of the estimated variance (chapter 4.5). That means that the indicator variables have to be introduced even if they are in this case not significant in the *model dependent* sense.

The fact that the indicator variables were not significant was rather taken as a hint that the characteristics of the small areas did not considerably differ. This consideration was also underlined by testing the small areas for differences in the timber volume. This was done using the observed terrestrial standing timber volume on plot level within each small area

(illustrated by boxplots in Fig. 5-16). Satisfying the assumption of equal variances and normal distribution in all small areas on a significance level of α=5%, an ANOVA was performed to test for significant differences of the observed terrestrial timber volume between the small areas. The p-value of the underlying F-test was 0.223, suggesting that the standing timber volume did not differ significantly between the small areas on a significance level of α=5%.

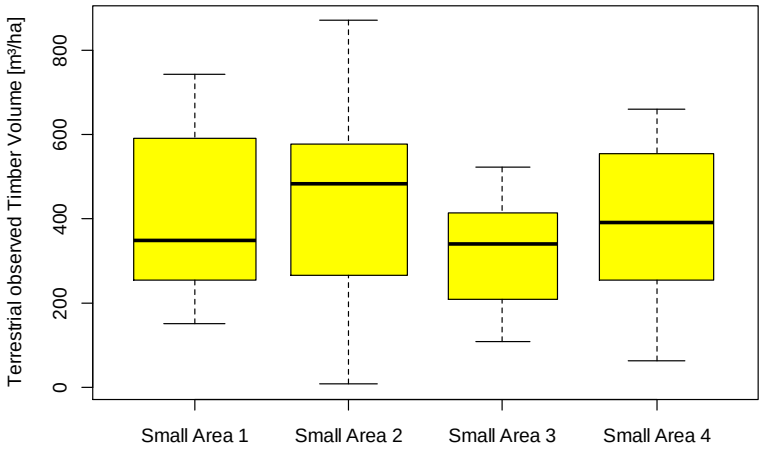


Fig. 5-16: Boxplots of the observed terrestrial standing timber volume [m^3/ha] on plot level for the small areas.

It was further investigated whether the slopes of the predictors used in the regression models showed considerable differences within the small areas. This investigation was carried out by means of ANCOVA analyses using the respective predictors as covariates and *all* indicator variables (included at once) as group variable. For each of the following considered models the ANCOVA assumptions (chapter 4.8.5) were satisfied. With respect to all tests within these analyses, a p-value below the critical value of α=5% was considered to be significant. The software outputs of the statistical analyses can be found in Appendix B. The results were as follows:

- Initially, the slopes of the most powerful predictors *Mean* and *Volume Density* were tested separately by two ANCOVA models: With *Mean* as covariate, using different intercepts for each small area was on the edge of being significant (p=0.0516). Also the interaction term for *Mean* was not significant. However, testing for the significance of *Mean* as predictor within each small area here revealed that the slope in small area 3 was only borderline significant (p=0.0497), i.e. *Mean* seems to have decreased explanatory power compared to the remaining part of the inventory area.

- With *Volume Density* as covariate, using the indicator variables as well as the interactions term was significant. However, the slope of *Volume Density* was not sig-

nificantly different from zero in small area 3, suggesting that *Volume Density* is not able to explain the terrestrial response within this small area.

- An ANCOVA model was also considered including both, *Mean* and *Volume Density*, as covariates at the same time. Here, the introduction of the indicator variables was significant. Introducing the interaction terms (i.e. interaction between *Mean* and the indicator variables and *Volume Density* and the indicator variables) was significant. However, the individual parameter tests (testing the slopes of the covariates to be unequal zero) revealed results that were more difficult to interpret and partially also in contrast to the previous described results: *Mean* and *Volume Density* were in this case significant in small area 3, whereas in the previous investigations small area 3 was the only one with both predictors being *not* significant. In contrast to this, *Mean* was in this case not significant for small area 1 and small area 2, and also *Volume Density* was not significant in small area 2 and small area 4. Furthermore, the signs of the slopes for *Mean* and *Voldensity* showed no clear pattern in the small areas, e.g. for *Volume Density* the slopes in small area 1 and 2 were clearly positive, whereas they were clearly negative in small area 3 and 4. A similar behaviour was observed for *Mean.* A potential explanation for this unstable behaviour could be the strong collinearity among the predictors *Mean* and *Volume Density*. Additionally, the model already comprised 12 parameters compared to 67 observations and was thus at the edge of violating the rule of parsimony in terms of overfitting (i.e. at least 5 to ten times as much observations as parameters used in the model).

- Finally, the *maximal* model was tested for different slopes within the small areas, i.e. beside the introduction of the indicator variables, the large model was extended by an interaction term for each predictor with the indicator variables (*Mean*, *Max*, *Sd*, *Q75* and *Volume Density*). By extending the large model in this way, the R^2 increased from 0.66 to 0.83. However, besides the fact that the model is thereby considerably overfitted (24 parameter compared to 67 observations), the interpretability of the results was even worse than for the previous ANCOVA model: in this case, neither the introduction of the indicator variables, nor the introduction of the interaction terms was significant. With respect to the slopes of each predictor in the small areas, there was no systematic pattern at all: the signs of all predictors varied considerably depending on the respective small area. Furthermore, *Volume Density* was only significant as predictor in small area 1 whereas *Mean*, *Max*, *Sd* and *Q75* were only significant in small area 2. With respect to the investigations before, which revealed a good prediction performance of these predictors for the entire inventory area as well as for the small areas, this result could not be well interpreted anymore and clearly contradicts several observations made before.

5.4 Computation Time of Auxiliary Variables

Before the main analysis of the computation times was carried out, the time measurements of the first computed sample location were removed from the dataset since considered as clear outliers: for unknown reasons, here the computation time for all variables was approximately four times larger as for all other computations. The study was hence based on 66 observations. An inspection of the boxplots for all variables (Fig. 5-17) already revealed substantial differences in the computation time of the statistical parameters (Fig. 5-17, left) compared to the variables *Stem number* and *Volume Density* (Fig. 5-17, right): the statistical parameters required less than a millisecond whereas the *Stem number* and *Volume Density* both ranged between 4 and 6 milliseconds per plot (all times including the performance of the boundary adjustment). The computation times also seemed to differ between the statistical parameters: here, the lowest computation times were achieved by the variables *Mean*, *Max* and *Varcoef*. On the contrary, the computation times of *Stem number* and *Volume Density* seemed to be almost identical. Since the distribution of several variables were significantly different from a normal distribution on α=5%, these visual impressions were ensured by the application of a Kruskal-Wallis test, suggesting significant differences in the computation time between at least one and the remaining variables. The pairwise Wilcoxon test subsequently indicated that only the variables *Q75* and *Q90* as well as *Stem number* and *Volume Density* were not significantly different (α=5%) in their computation time.

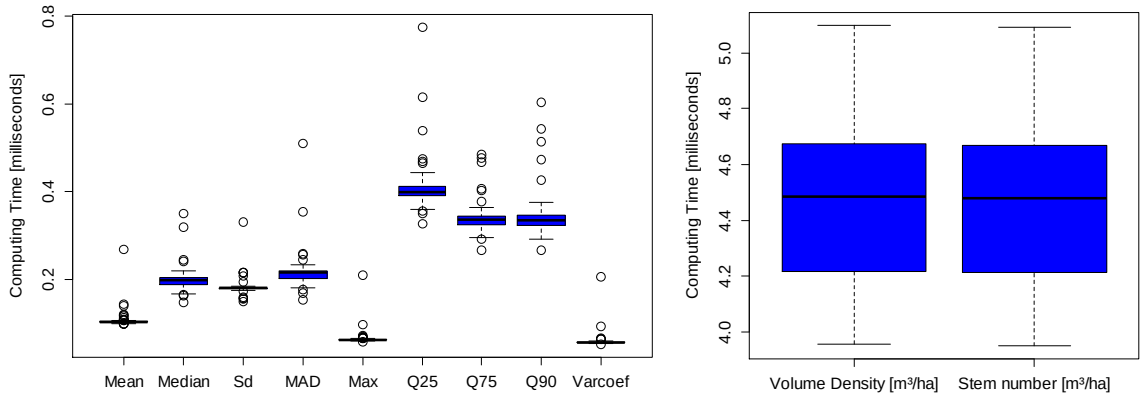


Fig. 5-17: Boxplots of the computation times [milliseconds] for all auxiliary variables (including boundary adjustment) based on their computation at the terrestrial sample point locations. For visualization purposes (scaling), two boxplots are used.

More interestingly, the auxiliary variables were also grouped in exhaustive and non-exhaustive variables according to the assumptions made in this study (i.e. *Mean*, *Median*, *Sd*, *MAD*, *Max*, *Q25*, *Q75*, *Q90*, *Varcoef* as exhaustive- and *Stem number* and *Volume Density* as non-exhaustive variables). An inspection of the resulting boxplots (Fig. 5-18) indicated that by this classification, the computation time of all variables was well separated in two groups. Furthermore, the two groups of variables revealed considerable differences in

their computation time: obviously, the non-exhaustive variables require considerable more computation time than the exhaustive variables. The significance of this difference ($p=2.2*10^{-6}$) was - not surprisingly - underlined by the application of an unpaired Wilcoxon rank sum test (the computation time of the exhaustive variables was significantly different from a normal distribution).

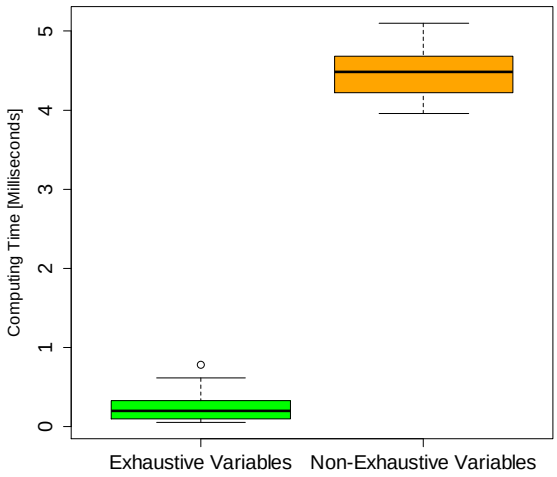


Fig. 5-18: Boxplots of the computation times [milliseconds] after classifying the auxiliary variables into exhaustive and non-exhaustive variables according to the assumption of the current study (exhaustive=*Mean*, *Median*, *Sd*, *MAD*, *Max*, *Q25*, *Q75*, *Q90*, *Varcoef*; non-exhaustive variables=*Stem number* and *Volume Density*)

When analysing the total computing time for these auxiliary variables that were finally used in the regression estimators, it has to be considered that the computation process is generally divided into two parts, i.e. the access and provision of the CHM data for the respective sample plot and the subsequent computation of the auxiliary variables (exhaustive and non-exhaustive ones) for this plot according to the algorithms described in chapter 4.7. For the calculation of all 306 first phase sample point locations it was possible to differ the computing time according to these two processing parts. As it turned out, most of the total computing time was needed for the access and provision of the CHM information: approximately 76 minutes were needed in total for the calculation of all 306 first phase sample plots, but only 3 out of 76 minutes were actually due to the calculation of the variables (i.e. the statistical parameters as well as the *Volume Density* and the *Stem number*).

In order to derive the true means of all exhaustive variables in case of the general regression estimator, the exhaustive variables had to be calculated within 31622 squares covering the entire forest of the inventory area. The respective calculation took approximately 2 days. In contrast to this considerably high computation effort, calculating the mean value of the exhaustive variable *Mean* (mean canopy height) within one step in a GIS needed a computation time of only 1 minute and 45 seconds (only 30 seconds respectively without calculating the true mean of *Mean* for each small area individually).

5.5 Global and Small Area Estimations

This section presents the *design-based* estimations for the standing timber volume [m^3/ha] within the entire inventory area (global estimations) and within the four small areas (small area estimations). The estimations were based on the variations of the estimators defined in chapter 4.10. The estimated and true weighted means of all exhaustive variables as well as the estimated mean of the non-exhaustive variable *Volume Density* are given in Tab. 12. Additionally, also the true mean of the exhaustive variable *Mean* (mean canopy height) which was alternatively calculated within a GIS in one step without using the squares (chapter 4.7.2) is given ($\bar{Z}_{1Mean,direct}$). It should be noted that the applied generalized estimators in this study were however calculated using the true mean of *Mean* derived by using the square-technique ($\bar{Z}_{1Mean}$). By comparing the two derived true means of *Mean* it becomes obvious that they are only marginally different. This was taken as a confirmation that the calculation of the true means of all remaining exhaustive variables as *weighted means* was justified as well as necessary when using the square-technique.

Tab. 12: Estimated and true means of auxiliary variables of *F* (entire inventory area) and the Small Areas ($\mathbf{SA_k}$). As additional information, also the true mean for *Mean* (mean canopy height) which was calculated in one step within a GIS is given (these values were however not used for the calculation of the estimators).

	F	SA_1	SA_2	SA_3	SA_4
$\hat{\bar{Z}}_{Mean}$	11.56	13.29	13.02	9.19	9.60
$\bar{Z}_{1Mean}$	11.39	12.85	12.21	9.33	10.45
$\bar{Z}_{1Mean,direct}$	11.37	12.84	12.18	9.31	10.40
$\hat{\bar{Z}}_{Sd}$	9.05	9.86	9.78	7.75	8.27
$\bar{Z}_{Sd}$	8.84	9.31	9.47	7.90	8.36
$\hat{\bar{Z}}_{Max}$	32.77	35.66	35.31	28.23	29.95
$\bar{Z}_{Max}$	32.68	34.92	35.36	28.81	30.22
$\hat{\bar{Z}}_{Q75}$	18.62	20.93	20.45	15.34	16.26
$\bar{Z}_{Q75}$	18.03	19.77	19.16	15.40	16.91
$\hat{\bar{Z}}_{Volume\ Density}$	476.55	592.12	554.23	325.47	366.03

5.5.1 Global Estimation Results

Tab. 13 provides the estimations for the entire inventory area: for all applied estimators, the corresponding point estimates (estimated standing timber volume [m^3/ha]) and standard errors of the point estimates (square root of the estimated variance in [m^3/ha] and percentage of the point estimates) are given. The estimations of the generalized regression estimator (*greg*) were obtained using the reduced model (predictor variables *Mean*, *Sd*, *Max* and *Q75*) and the large model (predictor variables *Mean*, *Sd*, *Max* and *Q75* and *Volume Density*). Tab. 14 presents the estimation results for the generalized regres-

sion estimator obtained by using the alternative reduced model with only *Mean* as predictor variable. For both approaches, the estimations for the regression estimator (*reg*) were identical since the regression estimator uses the large model, which in both cases was the same.

Tab. 13: Estimation results for the standing timber volume [m^3/ha] in the entire inventory area *F* (n_1=306, n_2=67).

Estimator	*Point Estimate*	*Standard Error*
$\hat{Y}_F$	399.43	23.82 (5.96%)
$\hat{Y}_{F,reg}$	387.68	16.16 (4.17%)
$\hat{Y}_{ext,F,reg}$	387.68	16.41 (4.25%)
$\hat{Y}^{(1)}{}_{F,reg}$	386.92	16.17 (4.18%)
$\hat{Y}^{(2)}{}_{F,reg}$	387.92	16.16 (4.17%)
$\hat{Y}^{(3)}{}_{F,reg}$	387.94	16.05 (4.14%)
$\hat{Y}^{(4)}{}_{F,reg}$	386.65	16.03 (4.15%)
$\hat{Y}^{(c)}{}_{F,reg}$	386.47	15.96 (4.13%)
$\hat{Y}_{F,greg}$	381.56	13.42 (3.67%)
$\hat{Y}_{ext,F,greg}$	381.56	13.80 (3.76%)
$\hat{Y}^{(1)}{}_{F,greg}$	381.40	13.39 (3.63%)
$\hat{Y}^{(2)}{}_{F,greg}$	381.80	13.42 (3.65%)
$\hat{Y}^{(3)}{}_{F,greg}$	381.79	13.33 (3.64%)
$\hat{Y}^{(4)}{}_{F,greg}$	381.16	13.27 (3.63%)
$\hat{Y}^{(c)}{}_{F,reg}$	381.32	13.18 (3.58%)

Tab. 14: Estimation results for the standing timber volume [m^3/ha] in the entire inventory area *F* (n_1=306, n_2=67) using only *Mean* as predictor variable in the reduced model.

Estimator	*Point Estimate*	*Standard Error*
$\hat{Y}_{F,greg}$	383.59	14.08 (3.52%)
$\hat{Y}_{ext,F,greg}$	383.59	14.42 (3.62%)
$\hat{Y}^{(1)}{}_{F,greg}$	383.30	13.92 (3.51%)
$\hat{Y}^{(2)}{}_{F,greg}$	384.11	14.02 (3.51%)
$\hat{Y}^{(3)}{}_{F,greg}$	383.92	13.96 (3.49%)
$\hat{Y}^{(4)}{}_{F,greg}$	382.73	13.90 (3.48%)
$\hat{Y}^{(c)}{}_{F,reg}$	383.08	13.73 (3.46%)

To achieve a better comparability, Fig. 5-19 illustrates the point estimates and standard errors of all applied global estimators in a graphical representation. The point estimates (Fig. 5-19, above) seemed to decrease when switching from the one-phase estimator to the two phase estimators, and again from the regression- to the generalized regression estimator. However, the differences between all point estimates were rather small (less than 20 m^3/ha) and calculating the 95%-confidence intervals of the point estimates (chapter 4.4.3) revealed no significant differences in the point estimates (all intervals are overlapping).

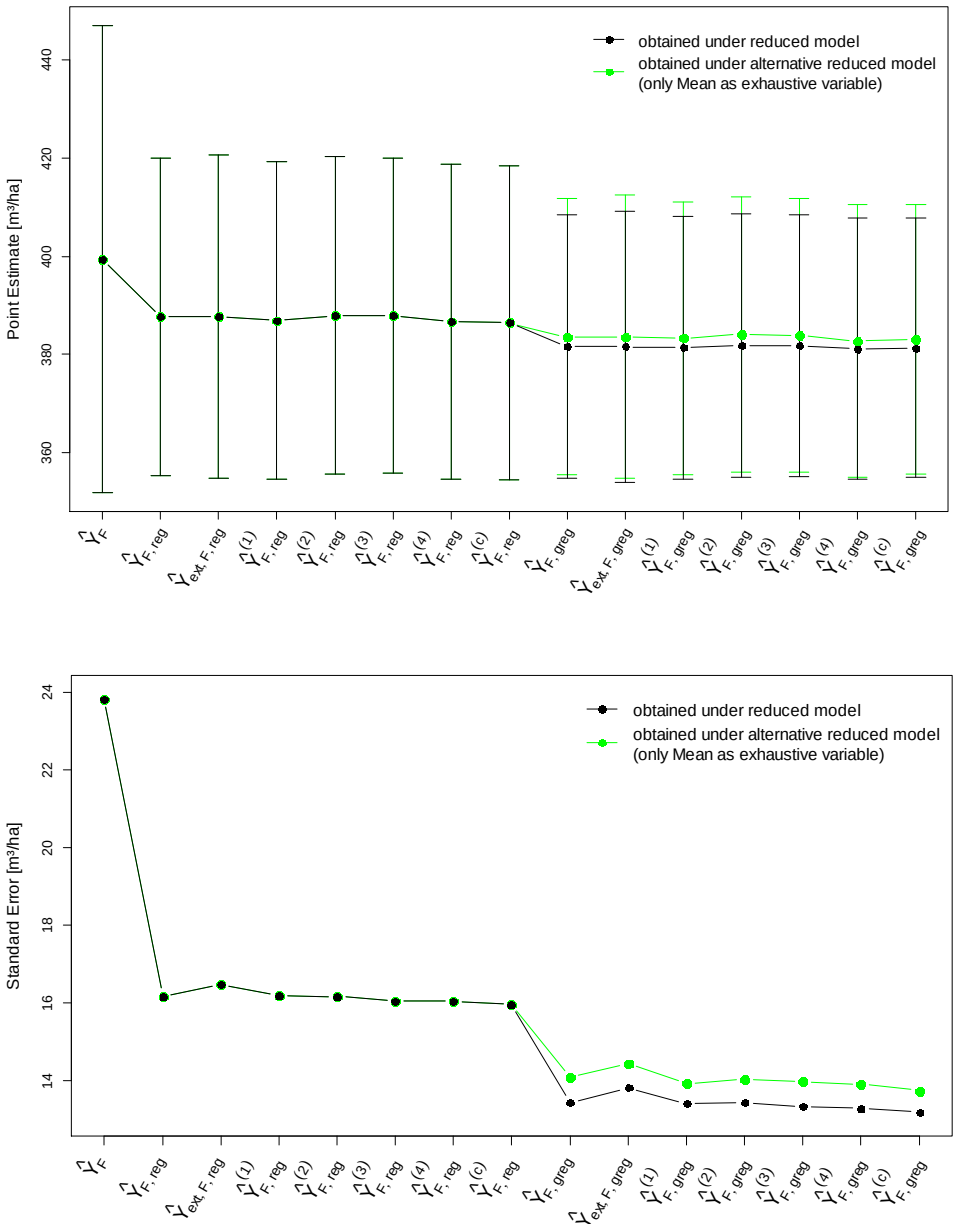


Fig. 5-19: *Above*: Visualization of the point estimates with 95%-confidence intervals. The point estimates do not differ significantly. *Below*: Standard errors for the global estimations. The main reduction is achieved by applying the two-phase methods compared to the one-phase approach. The generalized regression estimator yields an additional improvement in the estimation precision. (The connecting lines have been added for improving the visual perception).

It was of most interest to investigate whether one of the applied estimators show a considerable reduction in the standard error (Fig. 5-19, below): as already indicated by the visualization of the confidence intervals, obviously the main reduction was achieved by using the two-phase methods as an alternative to the one-phase estimation methods: the standard error was decreased by approximately 33% by applying the regression estimator and by approximately 44% in case of the generalized regression estimator. An investigation of the various versions of the regression estimator revealed that the corresponding generalized regression estimator (i.e. using exhaustive information) again yielded an additional reduction of the standard error compared to the regression estimator. This was true for both reduced models, though the standard error of the alternative reduced model was slightly larger. Additionally, the extension of the regression estimator by either one indicator variable or all indicator variables at a time showed only marginal differences. The same was true for the generalized regression estimator. In both cases, the standard error derived under the external model approach was only slightly increased compared to those derived under the more sophisticated calculation (i.e. using the g-weights technique).

5.5.2 Small Area Estimation Results

Tab. 15 presents the estimations for the small areas of the inventory site: for each applied small area estimator, the corresponding point estimate (standing timber volume [m^3/ha]) as well as the standard error (square root of the estimated variance) in [m^3/ha] and in percentage of the corresponding point estimate is given. The estimations of the generalized small area regression estimator were obtained using the predictors Mean, Sd, Max and Q75 in the reduced model. Additionally, Tab. 16 shows the estimation results for the generalized small area regression estimator which were obtained by using the alternative reduced model with only *Mean* as predictor variable. The estimations for the small area regression estimator were identical to those in Tab. 15 since the small area regression estimator uses the large model, which in both cases was the same.

Tab. 15: Estimation results for the standing timber volume [m^3/ha] within the small areas. The standard error is given in brackets.

Estimator	*Small Area 1*	*Small Area 2*	*Small Area 3*	*Small Area 4*
$n_{1,G}: n_{2,G}$	94:19	81:17	66:15	65:16
$\hat{Y}_G$	410.40	461.44	318.00	396.85
	(44.58 / 10.86%)	(56.35 / 12.21%)	(34.36 / 10.81%)	(47.86 / 12.06%)
$\hat{Y}_{ext,G,reg}$	400.93	428.09	314.36	370.92
	(28.85 / 7.27%)	(34.84 / 7.70%)	(31.99 / 10.79%)	(36.12 / 8.93%)
$\hat{Y}^{(k)}{}_{G,reg}$	398.91	428.71	314.38	374.31
	(28.99 / 7.25%)	(33.01 / 7.59%)	(33.93 / 10.83%)	(33.67/ 9.00%)
$\hat{Y}^{(c)}{}_{G,reg}$	398.86	430.54	324.93	375.21
	(28.91 / 7.20%)	(32.68 / 8.14%)	(35.18 / 10.18%)	(33.51 / 9.74%)
$\hat{Y}_{ext,G,greg}$	383.77	396.04	335.84	403.94
	(24.32 / 5.98%)	(26.29 / 7.10%)	(30.73 / 9.61%)	(31.18 / 7.83%)
$\hat{\hat{Y}}^{(k)}{}_{G,greg}$	381.66	396.68	335.73	407.31
	(22.42 / 5.96%)	(26.49 / 6.96%)	(30.02 / 9.65%)	(30.99 / 7.81%)
$\hat{\hat{Y}}^{(c)}{}_{G,greg}$	381.29	400.06	332.12	408.04
	(22.35 / 6.42%)	(25.88 / 7.08%)	(31.38 / 9.86%)	(30.77 / 8.03%)

Tab. 16: Estimation results for the standing timber volume [m^3/ha] within the small areas using only *Mean* as predictor variable in the reduced model. The standard error is given in brackets.

Estimator	*Small Area 1*	*Small Area 2*	*Small Area 3*	*Small Area 4*
$n_{1,G}: n_{2,G}$	94:19	81:17	66:15	65:16
$\hat{Y}_{ext,G,greg}$	390.58	409.11	317.50	390.90
	(25.09 / 5.87%)	(28.98 / 6.68%)	(31.32 / 8.94%)	(31.37 / 7.61%)
$\hat{\hat{Y}}^{(k)}{}_{G,greg}$	387.86	410.07	317.47	391.24
	(23.18 / 5.86%)	(29.13 / 6.47%)	(30.52 / 9.45%)	(30.62 / 7.54%)
$\hat{\hat{Y}}^{(c)}{}_{G,greg}$	388.04	410.70	328.21	396.10
	(23.11 / 6.34%)	(28.57 / 6.64%)	(31.67 / 9.15%)	(30.94 / 7.72%)

To provide again a better comparability, the estimation results of Tab. 15 and Tab. 16 were visualized in Fig. 5-20 and Fig. 5-21. The point estimates did not show a common pattern, i.e. a decrease in the point estimate from the one-phase- to the two-phase estimator, and from the regression- to the generalized regression estimator (as observed for the global point estimates). More interestingly, the calculation of the 95%-confidence intervals (chapter 4.5.3) revealed that also within the small areas, the point estimates did not significantly differ due to the applied estimator. Furthermore, the confidence intervals also indicated that the standing timber volume did not significantly differ between the small areas (overlapping intervals), which was also suggested by the ANOVA analyses of the terrestrial observations (chapter 5.3.3).

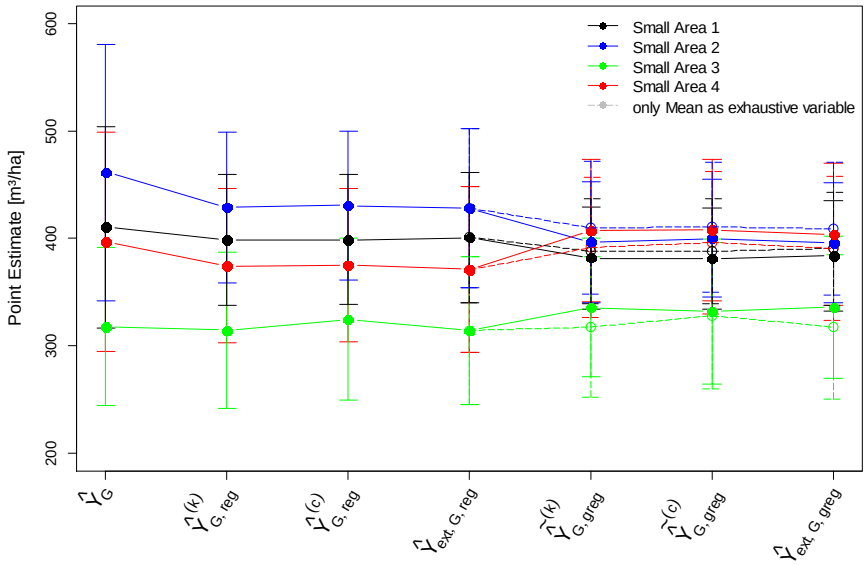


Fig. 5-20: Visualization of the point estimates for the small area estimations with 95%-confidence intervals. The connecting horizontal lines have been added for improving the visual perception.

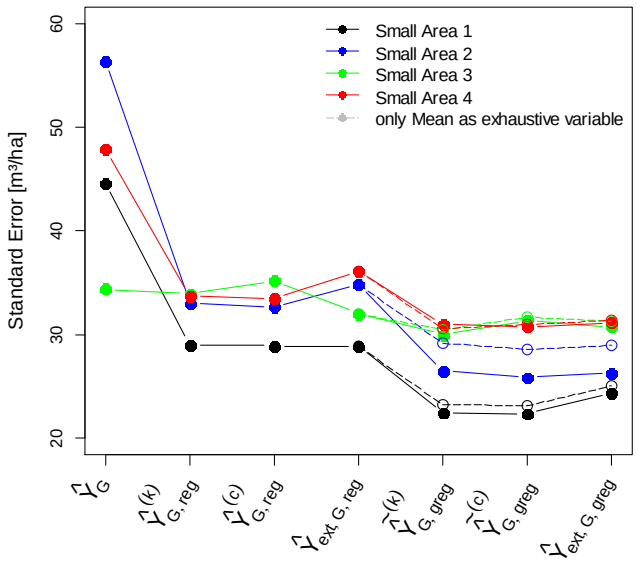


Fig. 5-21: Visualization of the standard errors for the small area estimations (the connecting lines have been added for improving the visual perception).

As already indicated by the confidence intervals, the standard error in general showed the same properties as in the case of the global estimators, i.e. the main variance reduction was achieved by changing from the one-phase estimator to the regression estimators. For most of the small areas the reduction of the estimated standard error was even more pronounced as for the global estimations: the application of the regression estimators yielded a reduction of the standard error between 44% (in small area 1) and 56% (small area 2). The estimation precision was further improved, but not dramatically, by using the generalized regression estimator. An exception was small area 3: here, the regression and generalized regression estimators did not perform much better than the sample mean.

Within all small areas, using the alternative reduced model (with *Mean* as the only predictor variable) in the framework of the generalized regression estimator led to slightly increased standard errors compared to the application of the reduced model (with the predictors *Mean*, *Sd*, *Max* and *Q75*). However, the standard errors achieved by the reduced model were still smaller than those achieved by the respective regression estimator.

With some evidence, the external model approach yielded larger errors as the more sophisticated approach (g-weights technique) at least for small area 2 and small area 4, and also for small area 1 in case of the generalized regression estimator. However, the differences were only small.

6 Discussion

The objective of this thesis was to investigate whether the new and extended version of design-based regression estimators proposed by Mandallaz (2013a, 2013b) can yield a further reduction in estimation uncertainty compared to already existing design-based estimators in the framework of operational forest inventory. The estimators comprised approaches for global as well as small area estimations. The estimators were tested in a forest inventory for a study site in eastern Switzerland (canton of Grisons) with the objective to estimate the standing timber volume of the entire inventory area as well as of four arbitrary chosen sub-units (small areas).

6.1 Pre-Processing of LiDAR Data

As auxiliary information to be used in the regression estimators, high resolution airborne LiDAR data were employed. The LiDAR data had been recorded under leaf-on condition and covered the entire inventory area. In order to derive auxiliary variables from the LiDAR data with highest possible explanatory power for the terrestrial target variable (standing timber volume on plot level), the LiDAR data had to be pre-processed in several steps: a fundamental step was the generation of a canopy height model (CHM) as the difference of the Digital Surface- and the Digital Terrain Model (DSM and DTM). Concerns about a potential loss of quality and precision due to the difficult acquisition conditions (steep slopes, rough terrain and dense vegetation) proved to be unfounded: in spite of locally sparse amounts of terrain echoes (last-pulses) - most likely due to the considered landscape features - the application of an inverse-distance weighting algorithm appeared capable of reconstructing a detailed and precise DTM. The generation of a hillshade representation of the DTM proofed to be very efficient for visualizing the quality of the DTM as well as the subsequently derived Canopy Height Model.

6.2 Computation of Auxiliary Variables

Beside these visual inspections, also the overall strong relationship between the auxiliary variables and the terrestrial observed target variable (standing timber volume on plot level) was an indirect confirmation of the high quality of the LiDAR derived canopy height model. However, it should be mentioned that the relatively high goodness of fit was also due to the short time lag between the terrestrial inventory (2006/2007) and the acquisition of the LiDAR data (2010). Additionally, the effort to optimize the agreement of shape and extension between the terrestrial and LiDAR plots together with a boundary adjustment at the forest edge was found to be a key factor in order to enable highest possible coefficients of determination.

Among the statistical parameters derived from the CHM at plot level, especially the *mean canopy height*, the *median* of the height values but also the *75%*- and *90% height quantiles* for each sample plot each revealed considerably strong correlation to the terrestrial response variable (R^2's of 0.47 – 0.5). The analysis of the computation time for all statistical parameters additionally showed that they can be computed very quickly. Considering the auxiliary variables which were derived by more sophisticated algorithms, *Volume Density* turned out to be the most powerful of them, achieving a coefficient of determination of 0.51. This was not too much surprising, as this information was derived at a considerable high level of detail, i.e. by detection of individual trees within a sample plot and subsequent prediction of the timber volume of each detected tree according to an allometric regression model with an R^2 of 0.65 at the tree level. This allometric regression model had been calculated based on a huge dataset including the main coniferous and broadleaf tree species of Switzerland (8121 LFI tree data of the domains *Alps* and *Pre-Alps*). An alternative regression model based only on the dominant tree species in the case study site (i.e. spruce and silver fir) could probably yield a further improvement.

In contrast to the good performance of *Volume Density*, the poor predictive power of the auxiliary variable *stem number* (R^2 of 0.08) was quite disappointing, since its influence on standing timber volume is well known from other studies as well as from fundamental rules in the field of yield and growth (self-thinning rule). An explanation for this unexpected performance is that detected local maxima in the CHM also occurred within canopy gaps. While the difference between DTM and DSM should be zero at locations were no object above ground exists, it is assumed that these minor peaks (in the range of several centimeters to one meter) are artifacts introduced by noise in the DTM and DSM. Although very small, the applied algorithm interprets them as individual trees. The peaks could however also be caused by stocks, deadwood, meadows or natural regeneration. Various studies have used a height threshold (e.g. 3 meters) in order to eliminate such low objects (Holmgren 2004). Unfortunately, the phenomenon was recognized in a late stage of this thesis and time was too short to improve the respective algorithms and repeat all respective steps (i.e. from the variable selection process up to recalculating the estimations). In case the proposed algorithms will be used in the framework of other inventories, it is of course highly recommended to implement such height-thresholds. Obviously, the explanatory power of the auxiliary variable *Volume Density* (which was also calculated based on the tree detection) has not been substantially affected by this effect: this is most likely due to the fact that false detections with low height values produce correspondingly small timber volumes (owing to the underlying allometric regression model). *Volume density* therefore proved to be more robust against the detection of these unwanted objects.

The choice of the auxiliary variables was based on prior knowledge from other studies as well as knowledge of existing allometric relationships. However, there may still be auxiliary variables which could be derived from the CHM and could further improve the goodness of fit. For example, even if the 75%- and 90%-quantiles each showed high correlation to the terrestrial response variable, the number of possible other quantiles is large and their might be quantiles which still provide more information.

6.3 Selection of Predictor Variables

A further step which attracted much attention in this study was the selection of a subset of auxiliary variables which were finally used in the design-based regression estimators. This selection process was motivated by satisfying the principle of parsimony according to which a model should always be as small as possible. The choice of the predictors which were finally used in the regression estimators was guided by testing-based procedures and goodness-of-fit-criteria (R^2_{adj},C_p, AIC) with respect to forward, backward and stepwise selection procedures. The various selection procedures lead to similar subsets of explanatory variables with higher R^2 (0.65 and 0.66), whereas the R^2 was around 0.5 with just one auxiliary variable (*mean canopy height* or *Volume Density*). This justified the use of several predictors.

However, due to substantial collinearity between almost all auxiliary variables several choices of predictor combinations would have been possible with essentially the same fit. Additionally, it became obvious that the collinearity effects can seriously hamper variable selection as the results of the selection algorithms become extremely unstable and maybe even not reliable. Consequently, the "best" model could possibly be missed by these algorithms, whereas even the description "best" can be misleading: while the applied selection algorithms identify the "best" model by searching for an appropriate balance between model size and model fit, the "best" model in the framework of regression estimators may be defined by additional or even different requirements which are not considered by classical selection algorithms. This was the case within the framework of the newly proposed general regression estimator, where the reduced model is defined as a subset of predictors of the respective large model: The workload for the reduced model, in particular the computation time, should per definition be much smaller than for the large model. Although a comparison of the computing times revealed that the exhaustive variables (used in the reduced model) can much faster be calculated than the non-exhaustive variables (which were only allowed in the large model), the calculation of the reduced model predictors over the entire inventory area still requested considerable large computation time (approximately 2 days). Applying an alternative reduced model with only *mean canopy height* as predictor variable however led to enormous savings in computation time for the exhaustive derivation (approximately 2 minutes) while this variable selection by hand did not lead to considerable different standard errors. A further drawback of only relying on selection algorithms is the fact that in large datasets (e.g. for national inventories) almost all predictor variables will be significant due to the increasing power of the tests, while the question which of the predictors are really relevant might still be left unanswered.

Summarizing, it was one of the fundamental conclusions of the study that the final selection of predictors should not be left to statistical algorithms alone, but is in the responsibility of the subject scientist, and that also the interpretability as well as traceability of regression models might be more important than relying solely on the results of statistical tests. In the case of the reduced model in this study, strictly speaking, the only comprehen-

sibly interpretable auxiliary variables were the *mean* and *median canopy height* and the non-exhaustive variables *Volume Density* and s*tem number*. It should however be mentioned again that the process of variable selection was performed due to good statistical practice, but is not fundamental for deriving the design-based estimations. The entire selection procedure including the model validations belong to classical, model-dependent statistics and have to be clearly separated from the design-based inventory methods, which do not require any assumptions on the regression models used in the estimators.

6.4 Comparison of Estimators

The comparison of the design-based estimators revealed that the standard error was substantially reduced from the one-phase approach (sample mean) to the two-phase approach (regression estimators), particularly in the case of the small area estimations. The accuracy was further improved within the two-phase approach by applying the newly proposed generalized regression estimator (however not as substantially). For the global as well as for the small area estimations, the introduction of either one indicator variable per small area or the introduction of the indicator variables all at once had little impact on the point estimates and the standard errors within the regression- and generalized regression estimator.

An interesting observation was that for the global and small area estimations the external approach (i.e. not accounting for the errors in the regression coefficients) was very close to the more sophisticated procedure used to derive the design-based estimated variances. This was a little bit disappointing, but on the other side rather reassuring. The strong similarity of the standard errors between the external approach and the more sophisticated approach was probably due to the fact that the sample sizes in the small areas (around 17 terrestrial observations) were sufficiently large. From a mathematical point of view the introduction of models extended by the indicator variables of small areas is very appealing and yields excellent asymptotic properties of the resulting point and variance estimates (Mandallaz 2013a). The external model approach, however, is easier to implement as it requires only predictions and residuals that can be obtained from available packages of statistical software such as *R*, which can be an advantage especially for complex regression models including interactions. One should in this case keep in mind that the derived errors of the regression coefficients given by such software packages are model-dependent and not design-based, although the differences were small in the present study. However, the calculation of the more precise estimated variance as described in this study is highly recommended if sample sizes in small areas are small.

Another interesting observation was that the goodness-of-fit in small area 3 was far below average for all calculated variations of the estimators ($R^2$of 0.3-0.4 as compared to R^2 of 0.7-0.8 in the other small areas). For this reason, in small area 3 the regression estimators did not perform much better than the sample mean. The ANCOVA analyses for testing the significance of potential interactions in the framework of the variable selection process indicated that the most important auxiliary variables, i.e. the *mean canopy height* and

Volume Density, had less predictive power in small area 3. A possible explanation is that the forest in small area 3 is more homogeneous. In other words, this suggested that there was an interaction between the small areas and the most important auxiliary variables. However the ANCOVA analyses were here also seriously hampered by the strong collinearity among the predictors and did not lead to interpretable results when using more than one predictor variable as covariate at a time: Considering the maximal model with all predictors of the large model including their interaction terms, the model fit was admittedly improved from R^2= 0.66 to R^2=0.83, but at the cost of losing any interpretability of the regression model (which underlines the recommended more pragmatic approach of keeping a model as simple and interpretable as possible). In addition, the approach of fitting an individual model for each small area (with individual slopes) would have clearly contradict the more pragmatic approach of the applied estimators, i.e. to "borrow strength" (Ghosh and Rao 1994) from a large sample size of the surrounding of a small area while insuring asymptotic properties by introducing an indicator variable of the small area (Mandallaz 2013a).

7 Conclusion

To summarize the main conclusions of this study, it should firstly be mentioned that the LiDAR-derived explanatory variables yield a substantial improvement in comparison to regression estimators using stand maps obtained by the visual interpretation of aerial photographs, most often with R^2 in the range of 0.3-0.4 only (Mandallaz 2008). Additionally, the use of LiDAR data in forest inventory seems thereby not to be limited to flat or hilly terrain, but can provide suitable information of high quality even in mountainous landscapes as long as appropriate sensor systems settings (footprint size, flight altitude, scan angle), suitable acquisition period (leaf-on condition) and appropriate pre-processing techniques are guaranteed.

The generalized regression estimator proposed by Mandallaz (2013b), which takes advantage of exhaustively available auxiliary information, yielded a further improvement over the classical regression estimator. With respect to the exhaustive calculation of predictor variables, the *mean canopy height* clearly appears to be the variable of choice: it could be shown that - compared to other variables which have to be calculated "partially" exhaustive by the use of sub-units (squares) – the *mean canopy height* can be computed extremely fast in one step via geographical information systems (GIS) without the prerequisite of any programming skills. This model choice (referred to as the "alternative" model in this study) was further confirmed since the *mean canopy height* was one of the most powerful predictors in the regression analyses and yielded alone most of the supplementary variance reduction in the framework of the generalized regression estimator. The *mean canopy height* is well interpretable and it can be conjectured that also in other inventories this variable will provide important information, particularly in the context of standing timber volume estimation. If the canopy height is available exhaustively it is highly recommended to use this variable in the generalized regression estimator to reduce the standard error, as shown in this study.

8 Outlook

The theory of two-phase inventories implies that at the second phase sample points the auxiliary variables are computed at exactly the positions of the terrestrial observations. A practical difficulty is that the nominal coordinates of the terrestrial sample points are possibly not equal to the actual, true location of acquisition. In the recent years, the application of high precision Differential GPS Systems (DGPS) for identifying the terrestrial sample locations in the field surveys have become increasingly popular. However, also this technology may produce position errors of up to 10 meters (Steinmann et al. 2012), essentially due to difficult topographical conditions and/or dense forest canopies. The influence of such position errors on predictive models using LiDAR data in forestry has already been in the focus of research (Mauro et al. 2010), but has not been consequently implemented in operational forest inventories. Considering that location errors can lead to outliers or even observations with high leverage in regression models, research is currently done at the chair of land use engineering of the ETH Zurich to cope with the problem of location uncertainty of terrestrial sample plots in the framework of two-phase forest inventories. Besides the identification of terrestrial sample plots whose potential location errors may have considerable influence on the regression coefficients in regression estimators, matching algorithms are being developed in the framework of optimization techniques which allow for the reconstruction of the original locations of terrestrial sample points. Although such a correction of terrestrial sample points is not fundamental in the framework of design-based estimation methods, a successful correction of the location may lead to further increased model fits and an associated reduction of the estimated variance.

In some countries, institutions responsible for forest inventories have already been provided with LiDAR data for free by the respective authorities. In those cases, the use of these data as auxiliary information has become very attractive since it enables a potential reduction of estimation uncertainty without increasing costs. Where this is however not the case, the acquisition of suitable LiDAR data often remains prohibitive. For this reason, research is currently being done considering alternative remote sensing data sources which also allow for constructing canopy height models. In Switzerland, the federal office of topography has decided not to continue the country-wide acquisition of LiDAR data in the coming years, but they will continue the repeatedly, exhaustive acquisition of aerial images over entire Switzerland. As the derivation of canopy height models based on photogrammetric technologies may provide an interesting, cost-effective alternative to the LiDAR-based product, research has lately been focused on comparing canopy height models derived by these two methods in the framework of forest inventory (Steinmann et al. 2011; Steinmann et al. 2012). Following this thesis, further research will be done within this topic by investigating how aerial imagery-derived CHMs acquired over steep mountainous terrain (e.g. the Alps) perform compared to respective LiDAR-derived CHMs. Since pre-investigations of the photogrammatically-based CHMs have already revealed a loss in

detail (spatial resolution) compared to the LiDAR-derived CHMs, it is of particularly interest to understand whether auxiliary variables based on single tree detection – providing considerable predictive power in the framework of timber volume estimation – can still be derived.

Although the use of the exhaustive information in the framework of the generalized regression estimator (Mandallaz 2013b) yielded a further reduction of the estimated variance, it became obvious that the computation time can become considerably large (especially when considering more or other variables than the *mean canopy height*). Whereas the required computation time can be acceptable for relatively small inventory areas (as within this study) it can however be prohibitive for large scale forest inventories. With respect to this consideration, an alternative three-phase estimator has just been developed at the chair of land use engineering to cope with that problem and will be tested within a follow-on case study: The idea of this three-phase estimator is to keep the main principle of the proposed design-based generalized regression estimator, but to derive some of the auxiliary variables not exhaustively anymore, but at a large number of sample points, the so-called null-phase. The first phase used in this study is thereby obtained by simple random sampling without replacement of the null-phase. The true means of the exhaustive variables are then replaced by empirical means in the null phase which are still more precise than the empirical means derived by the first phase. The sample size of the null-phase can be defined by setting upper thresholds on errors and costs (especially with respect to the computation time). First simulation results indicate that the increase in the standard error is small compared to the generalized regression estimator used in this study.

Literature

Achard, F., Eva, H.D., Mayaux, P., Stibig, H.-J., & Belward, A. (2004). Improved estimates of net carbon emissions from land cover change in the tropics for the 1990s. *Global Biogeochemical Cycles, 18 (2)*, GB2008

Anscombe, F.J., & Tukey, J.W. (1963). The Examination and Analysis of Residuals. *Technometrics, 5 (2)*, 141-160

Baltsavias, E.P. (1999). Airborne laser scanning: basic relations and formulas. *ISPRS Journal of Photogrammetry and Remote Sensing, 54 (2-3)*, 199-214

Battese, G.E., Harter, R.M., & Fuller, W.A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association, 83 (401)*, 28-36

Beauchamp, J.J., & Olson, J.S. (1973). Corrections for Bias in Regression Estimates After Logarithmic Transformation. *Ecology, 54 (6)*, 1403-1407

Berger, F., & Rey, F. (2004). Mountain Protection Forests against Natural Hazards and Risks: New French Developments by Integrating Forests in Risk Zoning. *Natural Hazards, 33 (3)*, 395-404

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

Brändli, U.-B. (2010). *Schweizerisches Landesforstinventar. Ergebnisse der dritten Erhebung 2004–2006*. Birmensdorf: Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft WSL. Bern: Bundesamt für Umwelt BAFU

Brändli, U.-B., & Denzler, L. (2011). Ergebnisse aus der dritten LFI 2004-06. Posterserie in 15 Teilen. Birmensdorf: Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft WSL

Brandtberg, T. (2007). Classifying individual tree species under leaf-off and leaf-on conditions using airborne lidar. *ISPRS Journal of Photogrammetry and Remote Sensing, 61 (5)*, 325-340

Brassel, P., & Lischke, H. (eds.) (2001). *Swiss National Forest Inventory: Methods and Models of the Second Assessment*. Birmensdorf: Swiss Federal Institute of Forest, Snow and Landscape Research WSL

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

Buddenbaum, H. (2011). Charakterisierung von Forstbeständen mit Hilfe von Laserscanning und Reflexionsmodellierung. Dissertation Fachbereich VI - Umweltfernerkundung und Geoinformatik, Universität Trier, 193 pages, http://ubt.opus.hbz-nrw.de/volltexte/2011/615/ (30.06.2013)

Cochran, W.G. (1977). *Sampling Techniques*. New York, Chichester, Brisban, Toronto, Singapore: John Wiley & Sons

Crawley, M.J. (2013). *The R Book*. Chichester: John Wiley & Sons

Davis, J.C. (2002). *Statistic and data analysis in geology*. New York, Chichester, Brisban, Toronto, Singapore: John Wiley & Sons

Draper, N.R., & Smith, H. (1966). *Applied Regression Analysis*. New York: Wiley

FAO (2000). FRA 2000-On Definitions of Forest and Forest Cover Change *Forest Resources Assessment, Working Paper 33, http://www.fao.org/docrep/006/ad665e/ad665e03.htm#P183_8898* (23.05.2013)

Faraway, J. (2011). faraway: Functions and datasets for books by Julian Faraway. R package version 1.0.5. http://CRAN.R-project.org/package=faraway (12.04.2013)

Faraway, J.J. (2002). *Practical Regression and ANOVA using R.* http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf (05.02.2013)

Fox, J. (2008). *Applied regression analysis and generalized linear models*. Thousand Oaks: SAGE Publications Inc.

Fox, J., & Weisberg, S. (2010). *An R companion to applied regression (2nd Ed.)*. Thousand Oaks: SAGE Publications, Incorporated

Frehner, M., Wasser, B., & Schwitter, R. (2005). *Nachhaltigkeit und Erfolgskontrolle im Schutzwald - Wegleitung für Pflegemaßnahmen in Wäldern mit Schutzfunktion*. Bern: Bundesamt für Umwelt, Wald und Landschaft (BUWAL)

Geßler, A., Keitel, C., Kreuzwieser, J., Matyssek, R., Seiler, W., & Rennenberg, H. (2007). Potential risks for European beech (*Fagus sylvatica* L.) in a changing climate. *Trees, 21 (1)*, 1- 11

Ghosh, M., & Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science, 9 (1)*, 55-76

Green, E.J., Thomas, C.E., & Strawderman, W.E. (1987). Notes: Stein-Rule Estimation of Timber Removals by County. *Forest Science, 33 (4)*, 1054-1061

Gregoire, T.G., & Dyer, M. (1989). Model fitting under patterned heterogeneity of variance. *Forest Science, 35 (1)*, 105-125

Gregoire, T.G., & Valentine, H.T. (2008). *Sampling strategies for natural resources and the environment*. Boca Raton: Chapman & Hall/CRC

Heiberger, R.M. (2013). HH: Statistical Analysis and Data Display: Heiberger and Holland. R package version 2.3-37. http://CRAN.R-project.org/package=HH (12.04.2013)

Heinimann, H.-R., & Breschan, J. (2012). Pre-Harvest Assessment based on LiDAR Data. *Croation Journal of Forest Engineering, 33 (2)*, 169-180

Heritage, G., & Large, A. (2009). *Laser Scanning for the Environmental Sciences*. Chichester: Wiley-Blackwell

Hildebrandt, G. (1996). *Fernerkundung und Luftbildvermessung*. Heidelberg: Wichmann

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

Holmgren, J., & Persson, Å. (2004). Identifying species of individual trees using airborne laser scanner. *Remote Sensing of Environment, 90 (4)*, 415-423

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

Husmann, K. (2013). *Forecasting the economic optimal intensity of beech crown usage*. Master Thesis Department of Forest Economics and Forest Utilization, Faculty of Forest Sciences and Forest Ecology Göttingen, 124 pages

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., & Maltamo, M. (2008). Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. *International Journal of Remote Sensing, 29 (5)*, 1339-1366

Hyyppä, J., Kelle, O., Lehikoinen, M., & Inkinen, M. (2001). A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. *Ieee Transactions on Geoscience and Remote Sensing, 39 (5)*, 969-975

Isaaks, E.H., Srivastava, R.M. (1989). *Applied geostatistics*. Oxford, New York: Oxford University Press

IPCC (2007). Climate Change 2007 - The Physical Science Basis. Contribution of Working Group 1 to the Fourth Assessment Report of the IPCC, http://www.ipcc.ch/publications_and_data/publications_ipcc_fourth_assessment_report_wg1_report_the_physical_science_basis.htm (20.03.2011)

Jensen, J.R. (2007). *Remote Sensing of the Environment - An Earth Resource Perspective (2nd Ed.)*, Upper Saddle River, New Jersey: Prentice Hall Series in Geographic Information Science

Kattenborn, T. (2013). Forest Change Assessment and Corresponding Driver Analysis in the Magdalena Departement, Colombia (1985-2010). *Tagungsband der 33. Wissenschaftlich-Technische Jahrestagung der DGPF 27.02 - 1.03.2013 in Freiburg*, 108-115

Kaufmann, E. (2000). *Tarife für Schaftholz in Rinde und Rundholz-Sortimente*. Birmensdorf: Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft WSL

Keller, M.R. (2011). *Swiss National Forest Inventory. Manual of the Field Survey 2004-2007*. Birmensdorf: Swiss Federal Institute of Forest, Snow and Landscape Research WSL

Kleinn, C., Kändler, G., & Schnell, S. (2011). Estimating forest edge length from forest inventory sample data. *Canadian Journal of Forest Research, 41 (1)*, 1-10

Köhl, M., Magnussen, S., & Marchetti, M. (2006). *Sampling methods, remote sensing and GIS multiresource forest inventory*. Berlin, Heidelberg: Springer

Kramer, K., Degen, B., Buschbom, J., Hickler, T., Thuiller, W., Skykes, M.T., & de Winter, W. (2010). Modelling exploration of the future of European beech (*Fagus sylvatica* L.) under climate change- Range, abundance, genetic diversity and adaptive responses. *Forerst Ecology and Management, 259 (11)*, 2213- 2222

Kraus, K., & Pfeifer, N. (2001). Advanced DTM generation from LIDAR data. *International Archives Of Photogrammetry Remote Sensing And Spatial Information Sciences, 34 (3/W4)*, 23-30

Lanz, A. (2005). Stichproben am Waldrand - Probleme und nene Lösungsansätze. In Wunn, U. (ed) *Deutscher Verband Forstlicher Forschungsanstalten, Sektion Biometrie und Informatik, 17. Tagung.* Freiburg: 177-190

Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A., & 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

Liang, X., Hyyppä, J., & Matikainen, L. (2007). Deciduous-Coniferous Tree classification using Difference betwenn First and Last Pulse Laser Signatures. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, XXXVI, Part 3, W52*, 253-257

Thomas Lumley using Fortran code by Alan Miller (2009). leaps: regression subset selection. R package version 2.9. http://CRAN.R-project.org/package=leaps (12.04.2013)

Magnussen, S., Eggermont, P., & LaRiccia, V.N. (1999). Recovering Tree Heights from Airborne Laser Scanner Data. *Forest Science, 45 (3)*, 407-422

Mandallaz, D. (2008). *Sampling Techniques for Forest Inventories*. Boca Raton: Chapman & Hall/CRC

Mandallaz, D. (2012). Design-based properties of small area estimators in forest inventory with two phase sampling. *Technical report, ETH Zurich, Department of Environmental System Science, http://e-collection.library.ethz.ch* (07.01.2013)

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 two-phase sampling and partially exhaustive information with application to small-area estimation. *Technical report, ETH Zurich, Department of Environmental System Science, http://e-collection.library.ethz.ch* (01.03.2013)

Mauro, F., Valbuena, R., Manzanera, J.A., & García-Abril, A. (2010). Influence of Global Navigation Satellite System errors in positioning inventory plots for tree-height distribution studies. *Canadian Journal of Forest Research, 41 (1)*, 11-23

McRoberts, R.E., & Tomppo, E.O. (2007). Remote sensing support for national forest inventories. *Remote Sensing of Environment, 110 (4)*, 412-419

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

Mette, T., Papathanassiou, K., Hajnsek, I., Pretzsch, H., & Biber, P. (2004). Applying a common allometric equation to convert forest height from Pol-InSAR data to forest biomass. *Geoscience and Remote Sensing Symposium, 2004. IGARSS '04. Proceedings. 2004 IEEE International*, 511-514

Morsdorf, F. (2011). Accuracies and systematic effects in LiDAR assessment. In *Lecture LiDAR & SAR.* Zürich: Remote Sensing Laboratories, University of Zurich

Morsdorf, F., Frey, O., Meier, E., Itten, K.I., & Allgöwer, B. (2008). Assessment of the influence of flying altitude and scan angle on biophysical vegetation products derived from airborne laser scanning. *International Journal of Remote Sensing, 29 (5)*, 1387-1406

Morsdorf, F., Meier, E., Kötz, B., Itten, K.I., Dobbertin, M., & Allgöwer, B. (2004). LIDAR-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. *Remote Sensing of Environment, 92 (3)*, 353-362

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

Næsset, E., & Bjerknes, K.-O. (2001). Estimating tree heights and number of stems in young forest stands using airborne laser scanner data. *Remote Sensing of Environment, 78 (3)*, 328-340

Niklas, K.J. (1994). *Plant allometry: the scaling of form and process*. Chicago: University of Chicago Press

Niklas, K.J. (2004). Plant allometry: is there a grand unifying theory? *Biological Reviews, 79 (4)*, 871-889

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

Ørka, H.O., Næsset, E., & Bollandsås, O.M. (2009). Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. *Remote Sensing of Environment, 113 (6)*, 1163-1174

Ott, E., Frehner, M., Frey. H.U., & Lüscher, P. (1997). *Gebirgsnadelwälder. Ein praxisorientierter Leitfaden für eine standortgerechte Waldbehandlung*. Bern, Stuttgart, Wien: Paul Haupt

Persson, A., Holmgren, J., & Söderman, U. (2002). Detecting and measuring individual trees using an airborne laser scanner. *Photogrammetric Engineering and Remote Sensing, 68 (9)*, 925-932

Pretzsch, H. (2009). *Forest Dynamics, Growth and Yield: From Measurement to Model*. Berlin, Heidelberg: Springer

Pretzsch, H., & Biber, P. (2005). A Re-Evaluation of Reineke's Rule and Stand Density Index. *Forest Science, 51*, 304-320

R Development Core Team (2012). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ (02.06.2013)

Rao, J.N.K. (2003). *Small area estimation*. Hoboken, New Yersey: Wiley Series in Survey Sampling

Rao, P.S.R.S. (1988). 18 Ratio and regression estimators. In P.R. Krishnaiah & C.R. Rao (Eds.), *Handbook of Statistics* (449-468), Elsevier

Raupach, M., Marland, G., Ciais, P., Le Quéré, C., Canadell, J.G., Klepper, G., & Field, C.B. (2007). Global and regional drivers of accelerating CO2 emissions, *Proceedings of the National Academy of Sciences of the United States of America PNAS, 104 (24),* 10288-10293

Ritter, T., Nothdurft, A., & Saborowski, J. (2013). Correcting the nondetection bias of angle count sampling. *Canadian Journal of Forest Research, 43 (4)*, 344-354

Royston, J.P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 31 (2)*, 115-124

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.H. (2003). *Model assisted survey sampling*. New York: Springer Series in Statistics

Schlerf, M., Atzberger, C., Hill, J., Buddenbaum, H., Werner, W., & Schüler, G. (2010). Retrieval of chlorophyll and nitrogen in Norway spruce (Picea abies L. Karst.) using imaging spectroscopy. *International Journal of Applied Earth Observation and Geoinformation, 12 (1)*, 17-26

Schreuder, H.T., Wood, G.B., & Gregoire, T.G. (1993). *Sampling methods for multiresource forest inventory*. New York, Brisbane, Toronto, Singapore: John Wiley & Sons

Solberg, S., Naesset, E., & Bollandsås, O.M. (2006). Single tree segmentation using airborne laser scanner data in a structurally heterogeneous spruce forest. *Photogrammetric Engineering & Remote Sensing, 72 (12)*, 1369-1378

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

Stabler, B. (2013). shapefiles: Read and Write ESRI Shapefiles. R package version 0.7, http://CRAN.R-project.org/package=shapefiles (25.01.2013)

Stahel, W. (1999). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler (2.überarbeitete Auflage)*. Braunschweig, Wiesbaden: Vieweg

Steinmann, K., Ginzler, C., & Lanz, A. (2011). Kombination von Landesforstinventar- und Fernerkundungsdaten für Kleingebietsschätzungen. *Schweizerische Zeitschrift für Forstwesen, 162 (9)*, 290-299

Steinmann, K., Mandallaz, D., Ginzler, C., & Lanz, A. (2012). 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

Stierlin, H.-R., Brändli, U.-B., Herold, A., & Zinggeler, J. (1999). *Schweizerisches Landesforstinventar. Anleitung zur Erhebung für die Feldaufnahme der Erhebung 1993-1995*. Birmensdorf: Eidgenössische Forschungsanstalt für Wald, Schnee und Landschaft WSL

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

Swiss Federal Institute of Topography swisstopo (2013). http://www.swisstopo.admin.ch/internet/swisstopo/de/home/products/landscape/swissTLM3D.html (06.07.2013)

Teufen, B. (2004). Waldinventur Graubünden - Davos. Ausgewählte Resultate / Interpretationen. Erhebung 2006/07. Chur: Amt für Wald und Naturgefahren Graubünden

van Leeuwen, M., & Nieuwenhuis, M. (2010). Retrieval of forest structural parameters using LiDAR remote sensing. *European Journal of Forest Research, 129 (4)*, 749-770

Venables, W.N., & Ripley, B.D. (2002). Modern Applied Statistics with S. Fourth Edition. New York: Springer

Vohland, M., Stoffels, J., Hau, C., & Schüler, G. (2007). Remote Sensing Techniques for Forest Parameter Assessment: Multispectral Classification and Linear Spectral Mixture Analysis. *Silva Fennica, 41(3)*, 441-456

von Carlowitz, H.C. (1713). *Sylvicultura Oeconomica - Haußwirtliche Nachricht und Naturmäßige Anweisung Zur Wilden Baum-Zucht*. Leipzig: Johann Friedrich Braun

von Lüpke, N. (2013). *Approaches for the optimisation of double sampling for stratification in repeated forest inventories*, Dissertation Department of Ecoinformatics, Biometrics and Forest Growth, Faculty of Forest Sciences and Forest Ecology Göttingen, 87 pages

Vosselmann, G., & Maas, H.-G. (2010). *Airborne and Terrestrial Laser Scanning*. Dunbeath, Scotland: Whittles Publishing

Warton, D.I., Wright, I.J., Falster, D.S., & Westoby, M. (2006). Bivariate line-fitting methods for allometry. *Biological Reviews, 81 (2)*, 259-291

Weisberg, S. (1985). *Applied linear regression (2nd. Ed)*. Hoboken, New Yersey: Wiley Series in Probability and Statistics

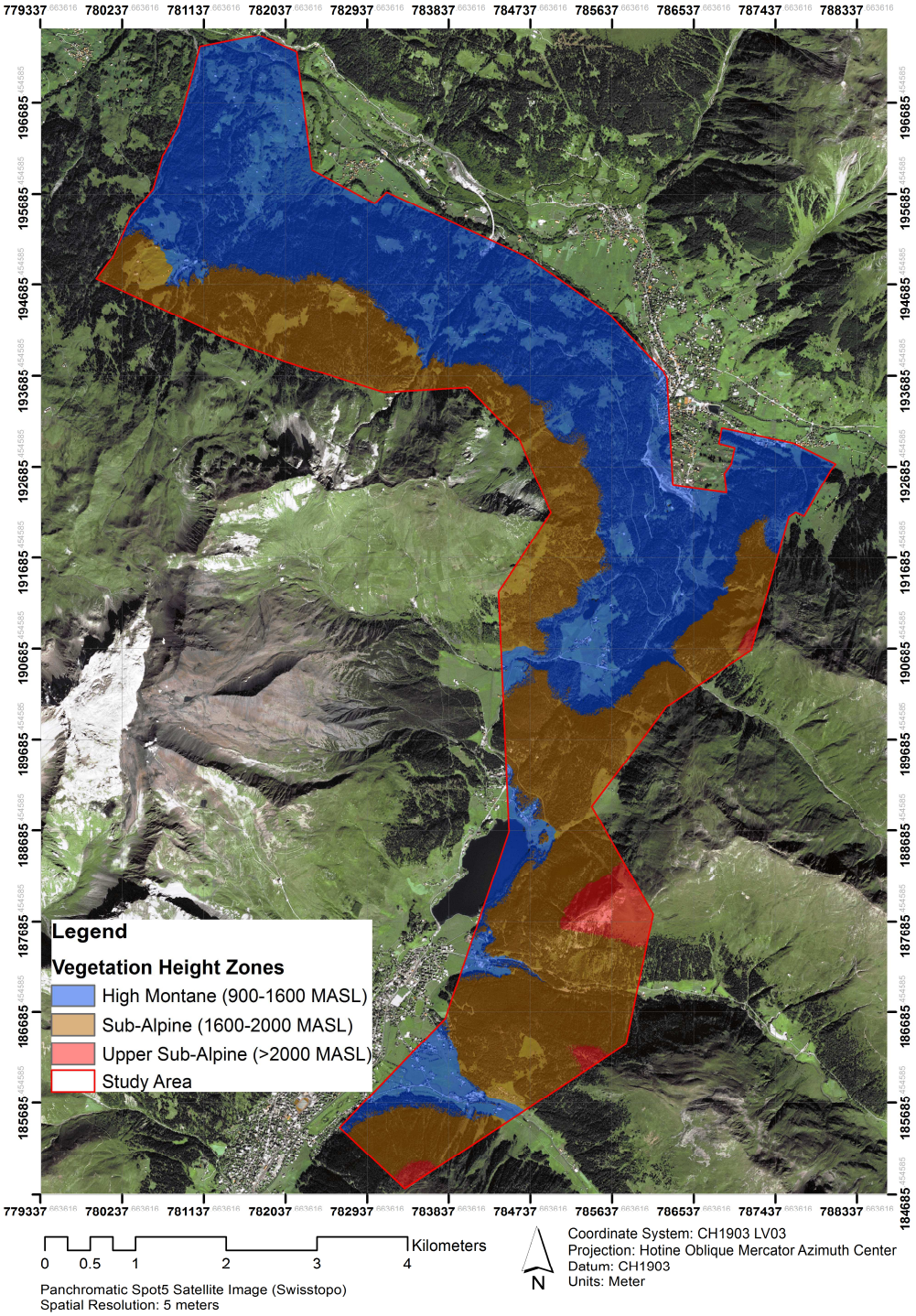
Wolfe, D.A., & Hollander, M. (1973). *Nonparametric statistical methods*. New York: John Wiley & Sons

Yandell, B.S. (1997). *Practical data analysis for designed experiments*. New York: Chapman & Hall/CRC

Zianis, D., Muukkonen, P., Mäkipää, R., & Mencuccini, M. (2005). Biomass and stem volume equations for tree species in Europe. *Silva Fennica Monographs No. 4*

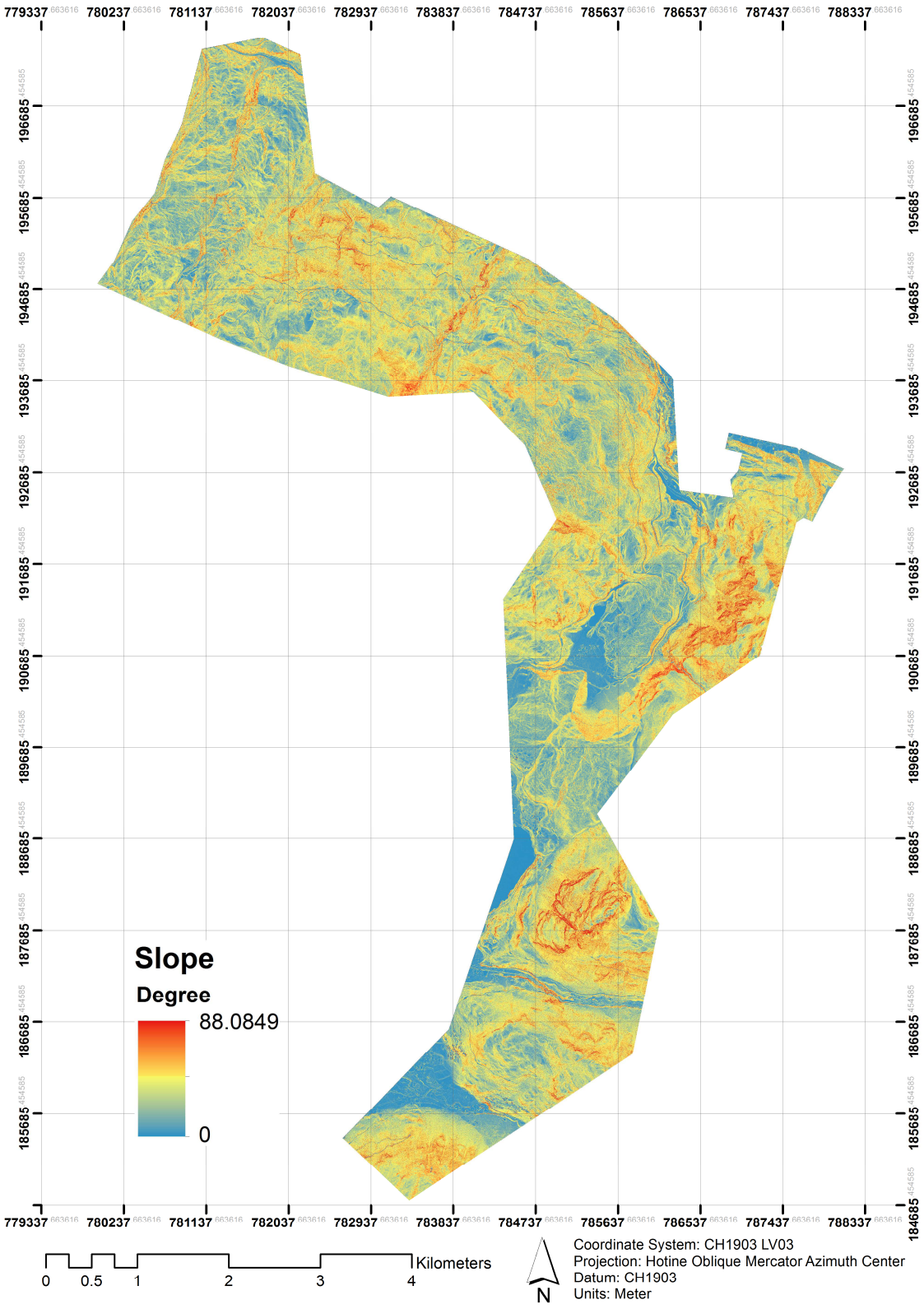
Appendix A Maps

A.1 Vegetation Height Zones (based on LiDAR Digital Terrain Model)



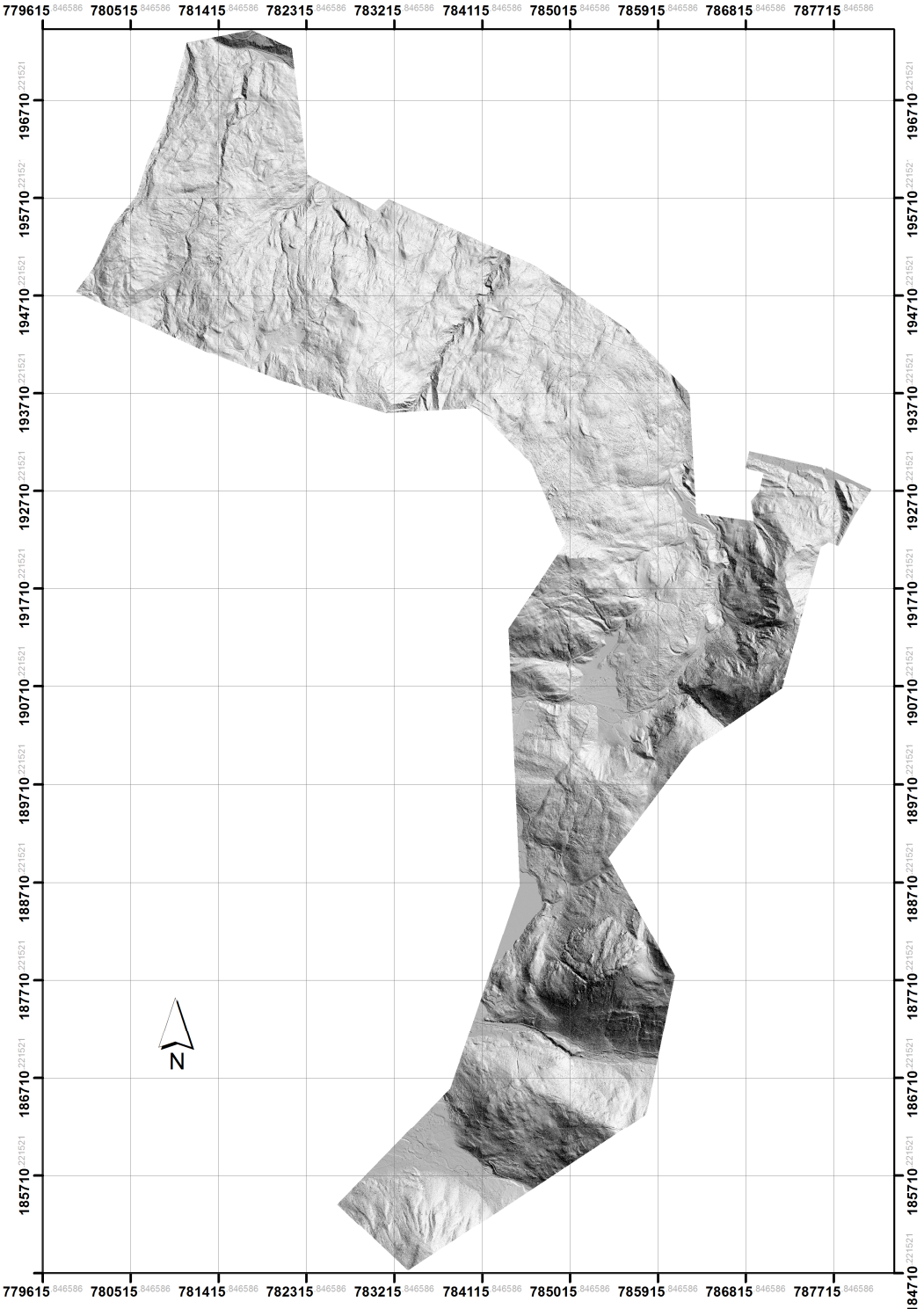
Map was created by standard geoprocessing tool “Reclassify” under ArcGIS 10.1

A.2 Slope Analyses (based on LiDAR Digital Terrain Model)



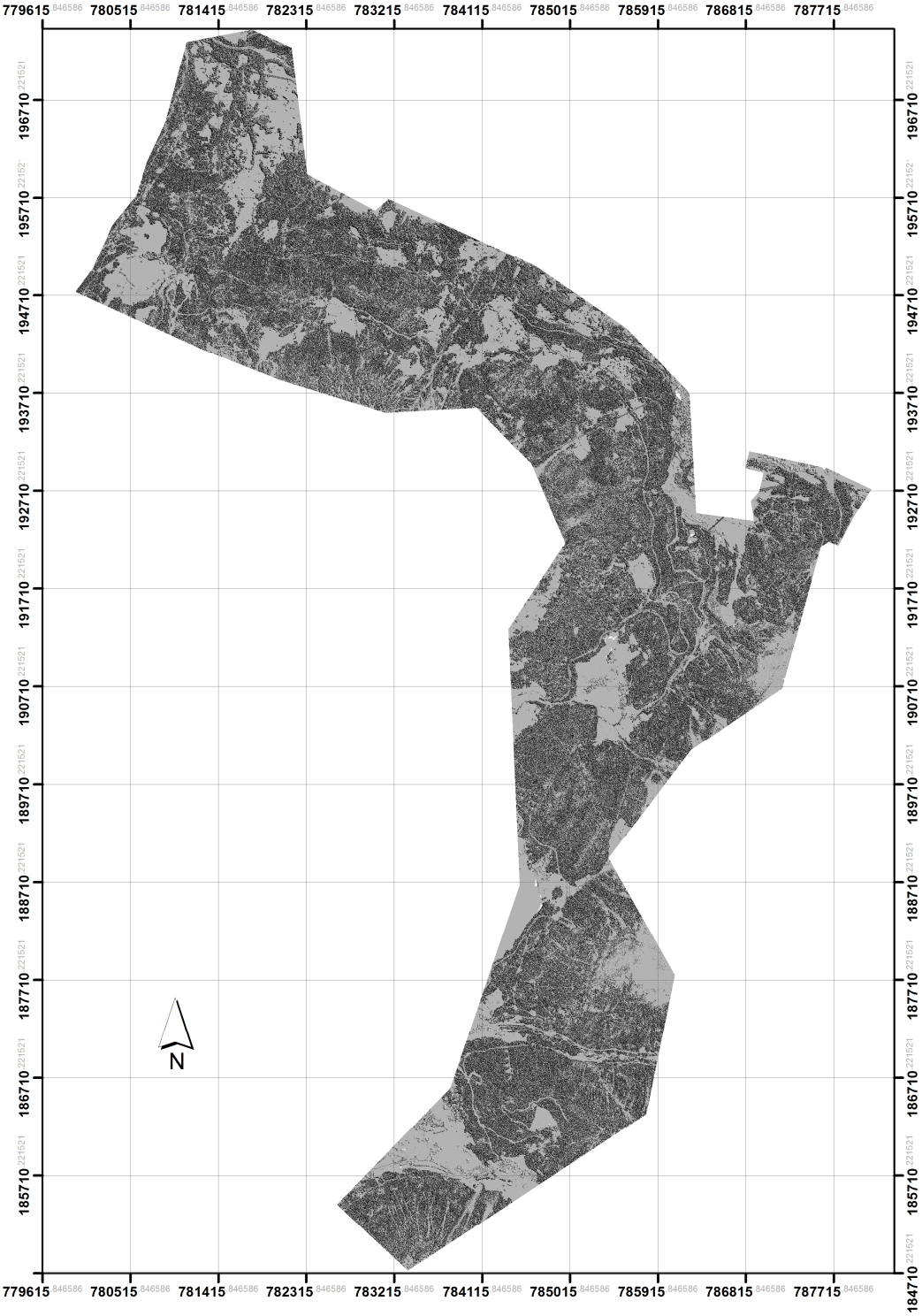
Slope Map was created by standard geoprocessing tool “Slope” under ArcGIS 10.1

A.3 LiDAR Digital Terrain Model (Hillshade)



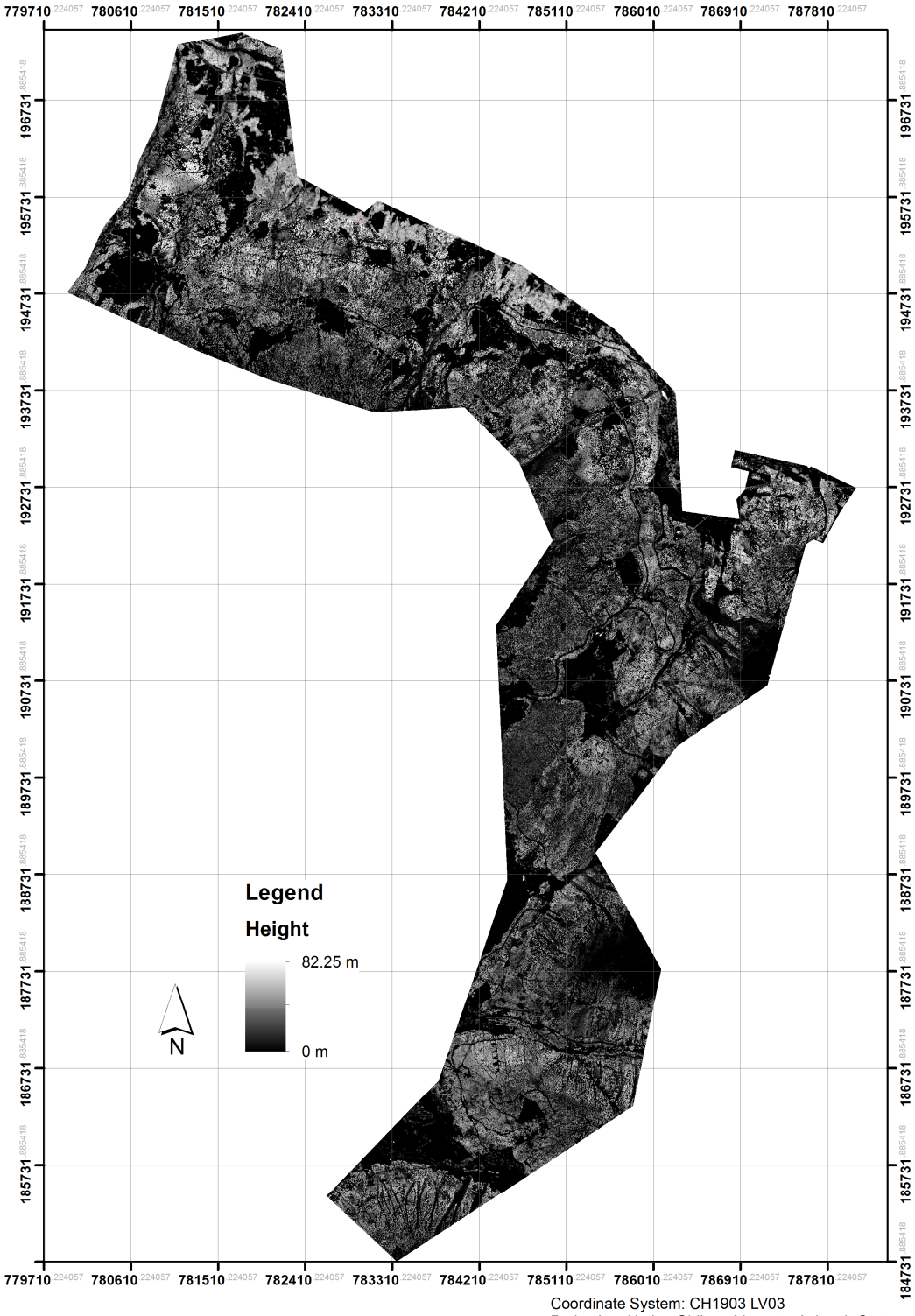
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
Units: Meter

A.4 LiDAR Canopy Height Model (Hillshade)



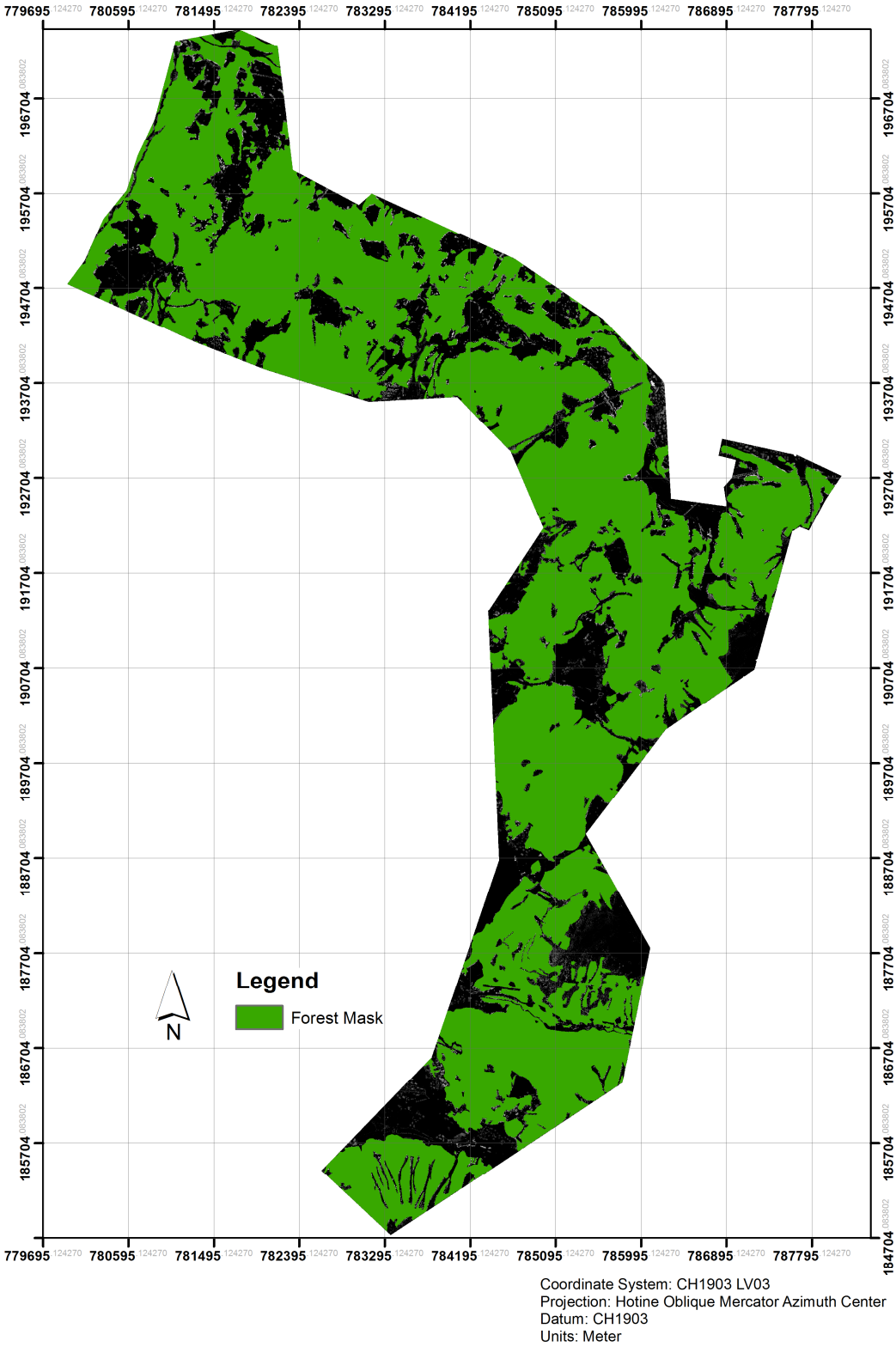
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
Units: Meter

A.5 LiDAR Canopy Height Model



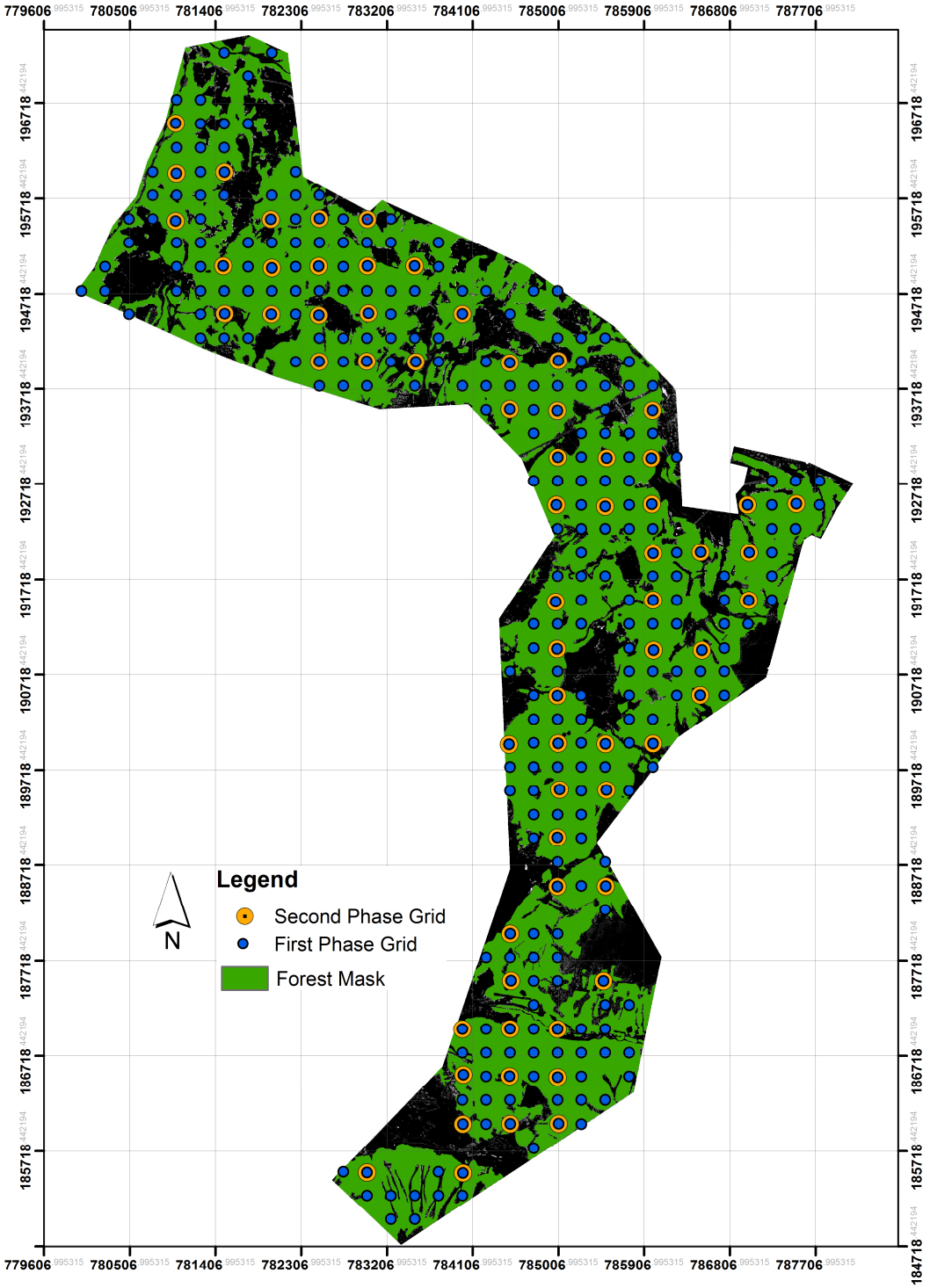
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
Units: Meter

A.6 Forest Mask



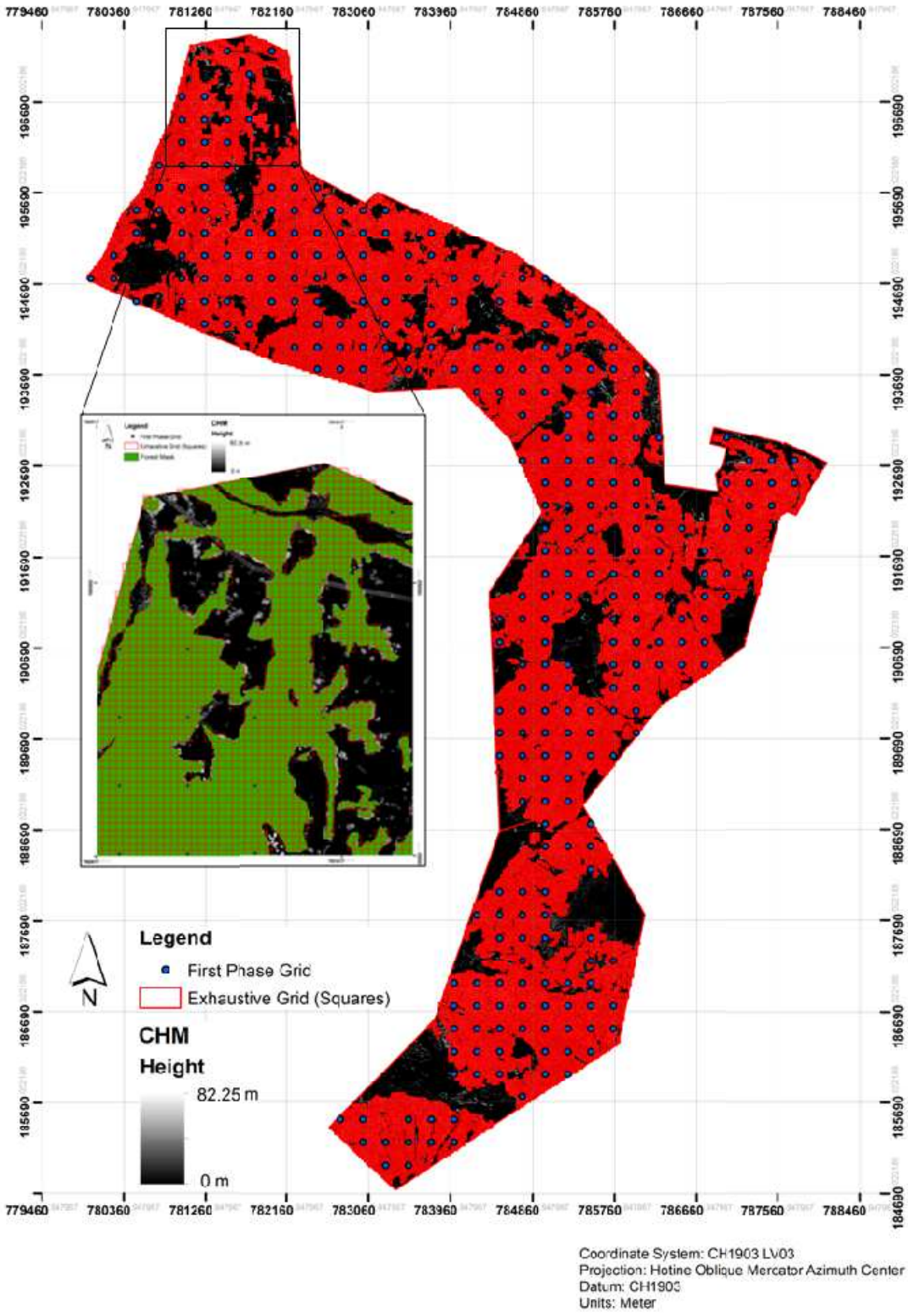
Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
Units: Meter

A.7 Two-Phase Inventory Grid



Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
Units: Meter

A.8 Exhaustive Inventory Grid

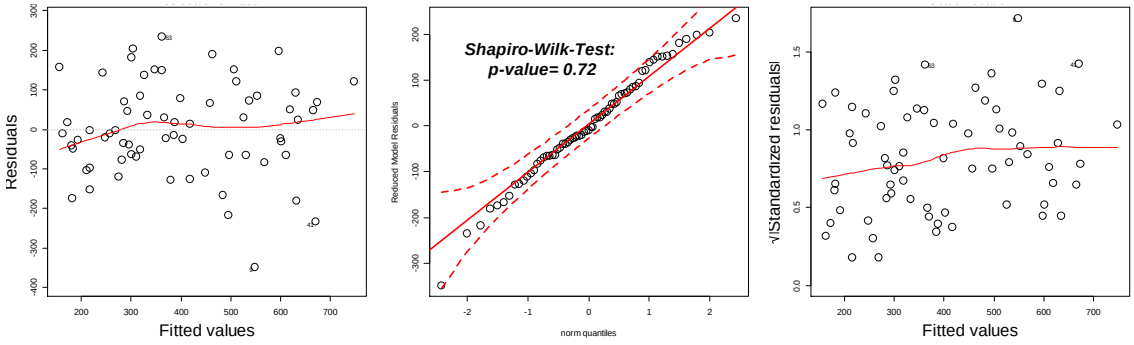


Coordinate System: CH1903 LV03
Projection: Hotine Oblique Mercator Azimuth Center
Datum: CH1903
Units: Meter

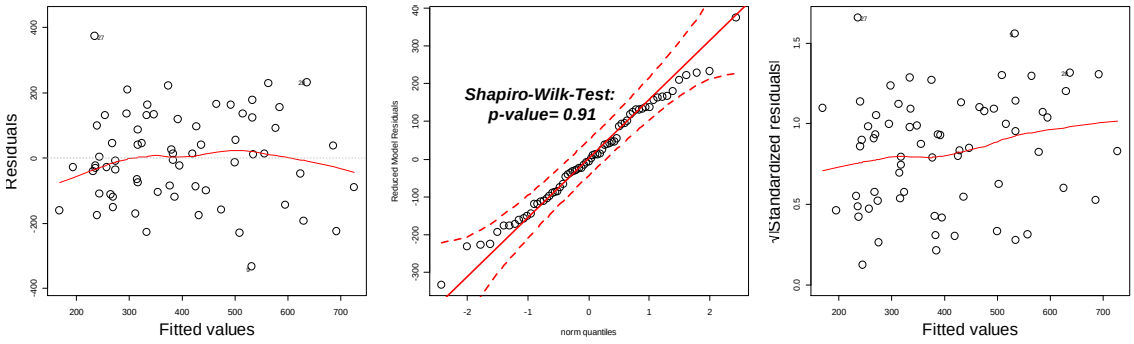
Appendix B Explorative Data Analysis

B1. Global Models (no indicator variables) - Model Diagnostic for entire inventory area

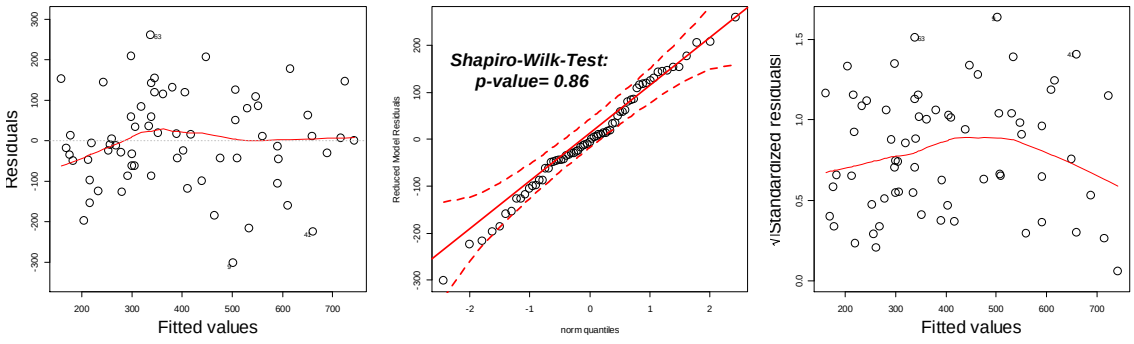
Global Reduced Model Diagnostics



Global Alternative Reduced Model Diagnostics



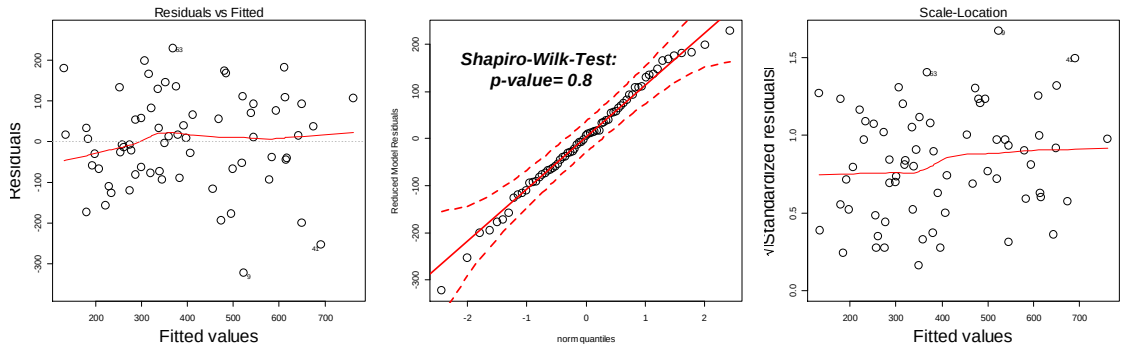
Global Large Model Diagnostics



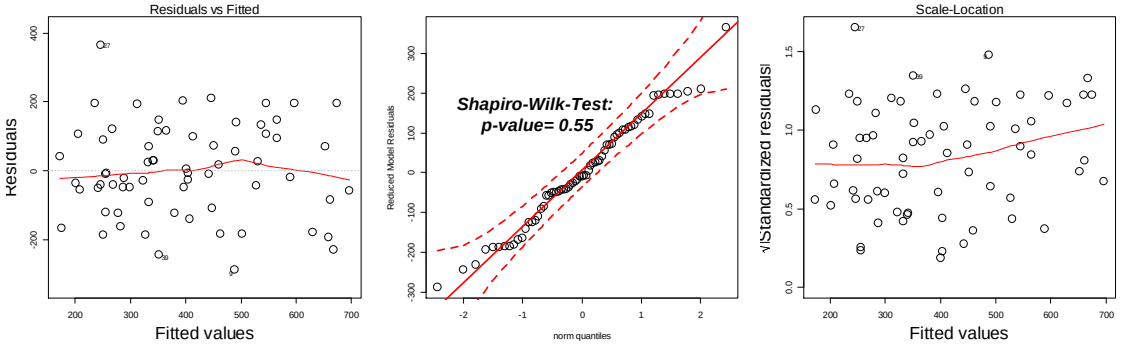
From *left* to *right*: Tukey-Anscombe Plot (fitted values vs. model residuals), QQ-Plot including 95% confidence intervals (dotted red lines) and result of the Shapiro-Wilk-test, and Scale-Location Plot (fitted values vs. square root of the standardized residuals).

B2. Indicator for 1st SA –Model Diagnostic for entire inventory area

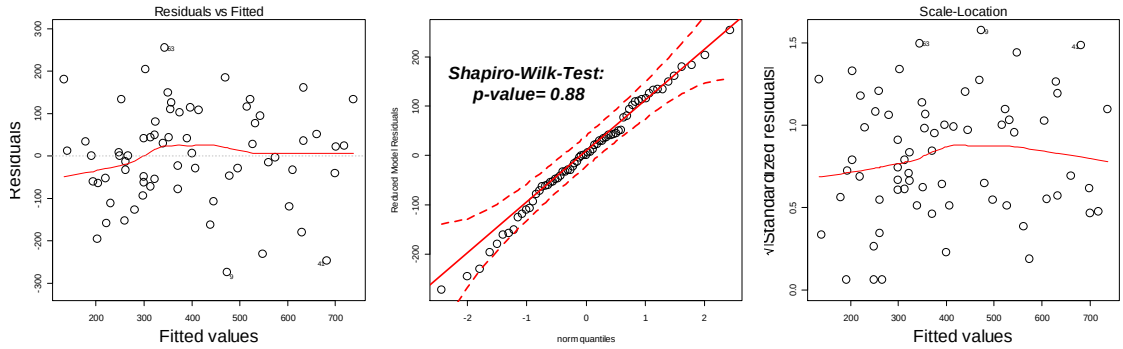
Global Reduced Model Diagnostics



Global Alternative Reduced Model Diagnostics



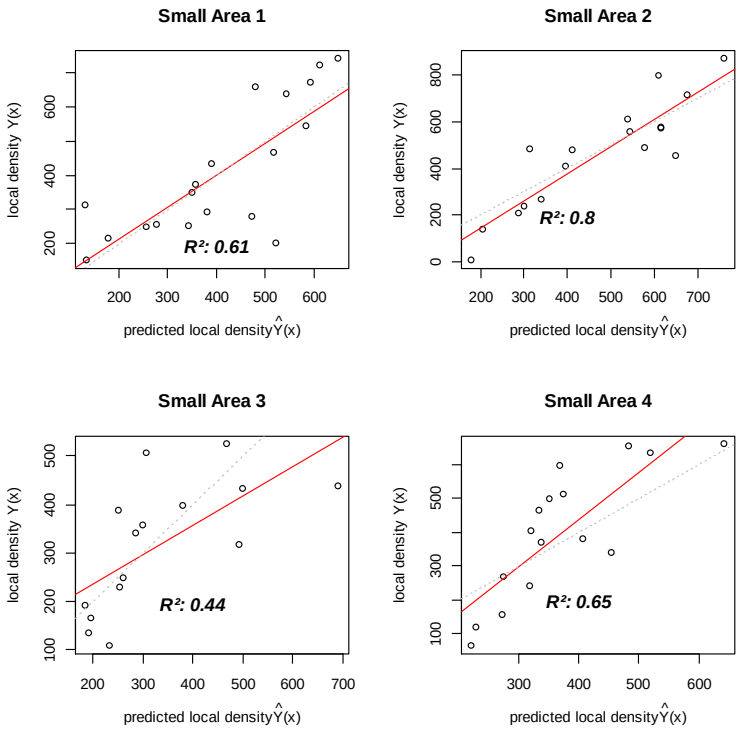
Global Large Model Diagnostics



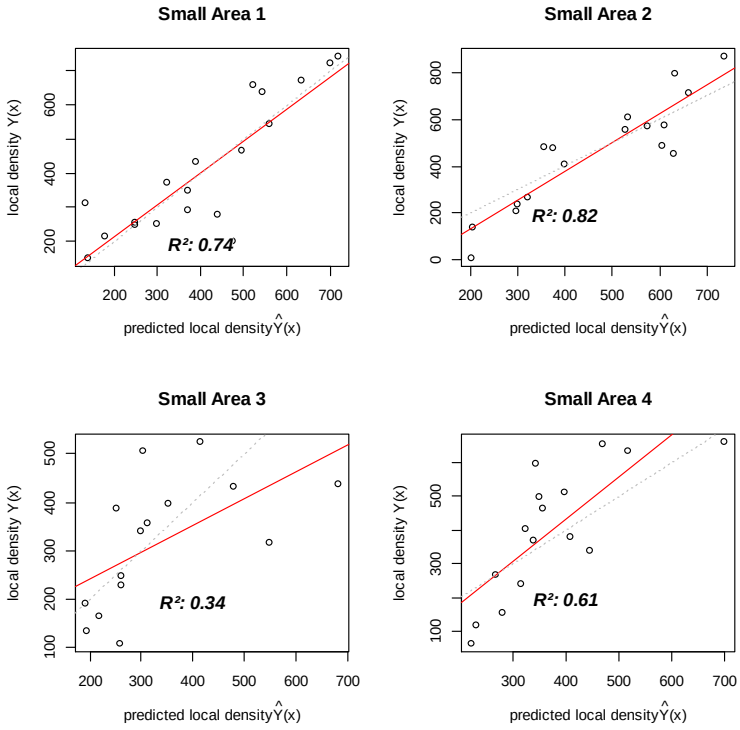
From *left* to *right*: Tukey-Anscombe Plot (fitted values vs. model residuals), QQ-Plot including 95% confidence intervals (dotted red lines) and result of the Shapiro-Wilk-test, and Scale-Location Plot (fitted values vs. square root of the standardized residuals).

Indicator for 1st SA - Model Performance in Small Areas

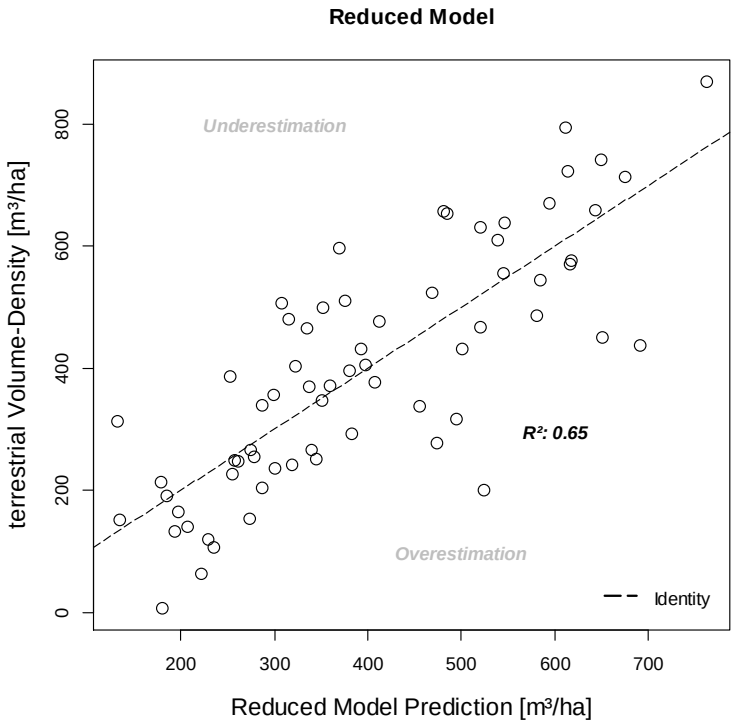
Reduced Model

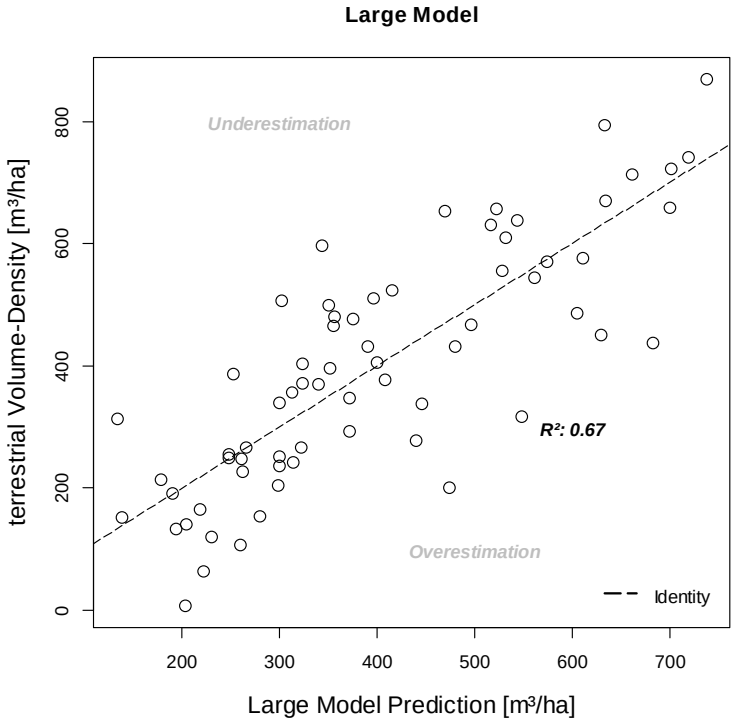


Large Model



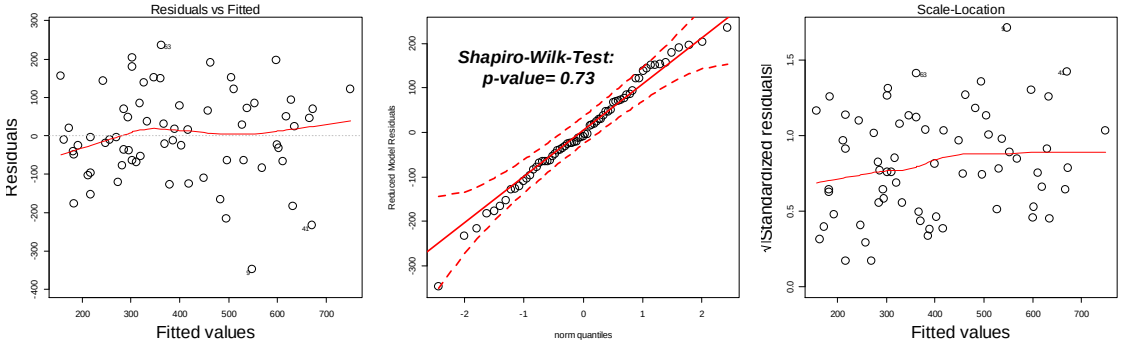
Indicator for 1st SA - Global Model Performance in F



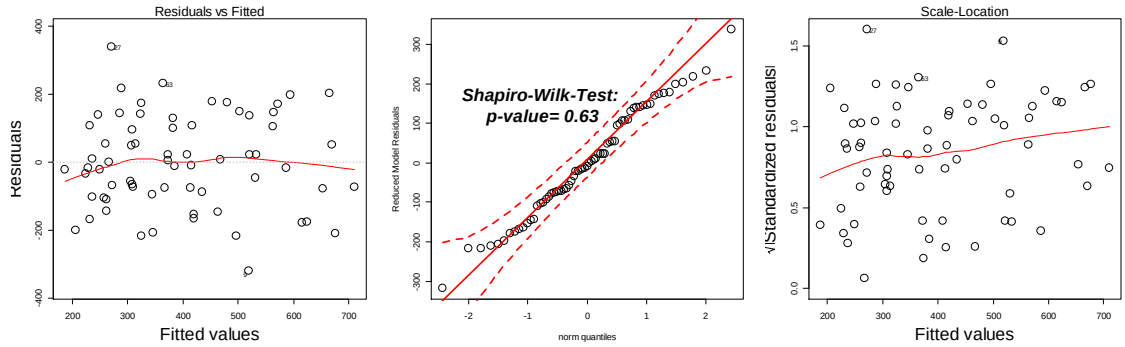


B3. Indicator for 2nd SA - Model Diagnostic for entire inventory area

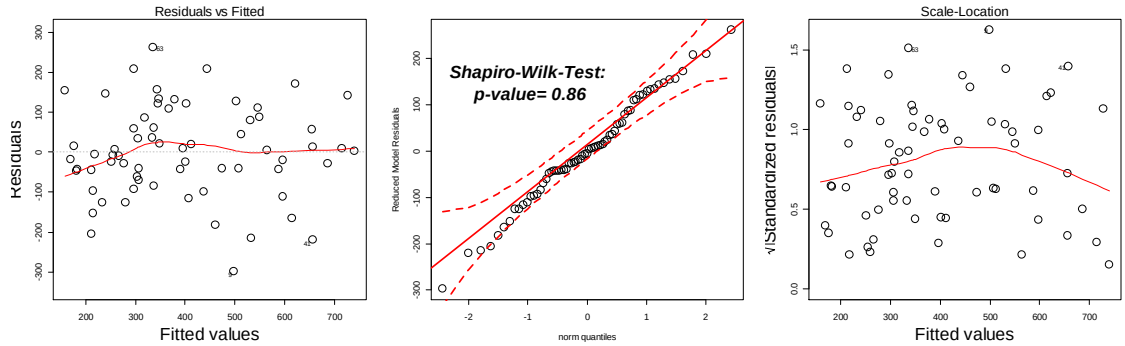
Global Reduced Model Diagnostics



Global Alternative Reduced Model Diagnostics



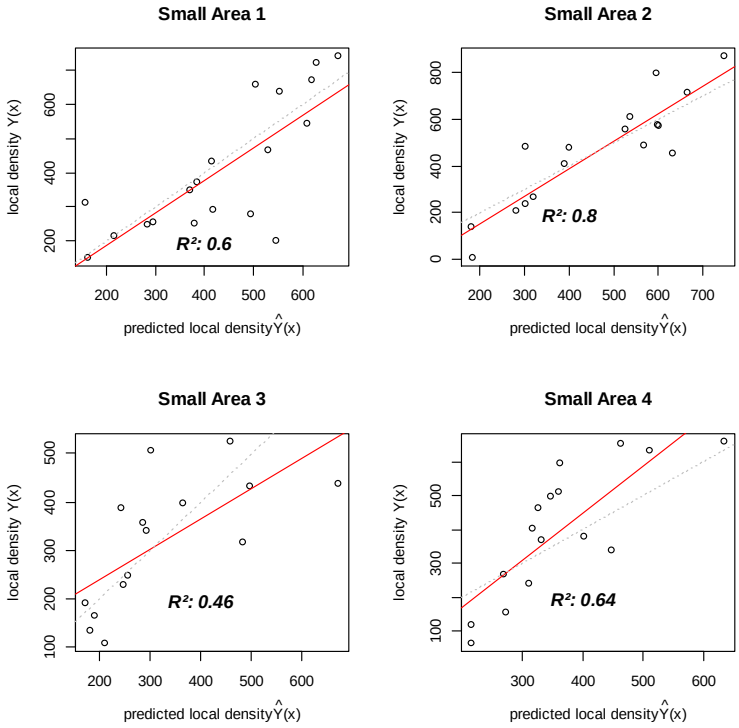
Global Large Model Diagnostics



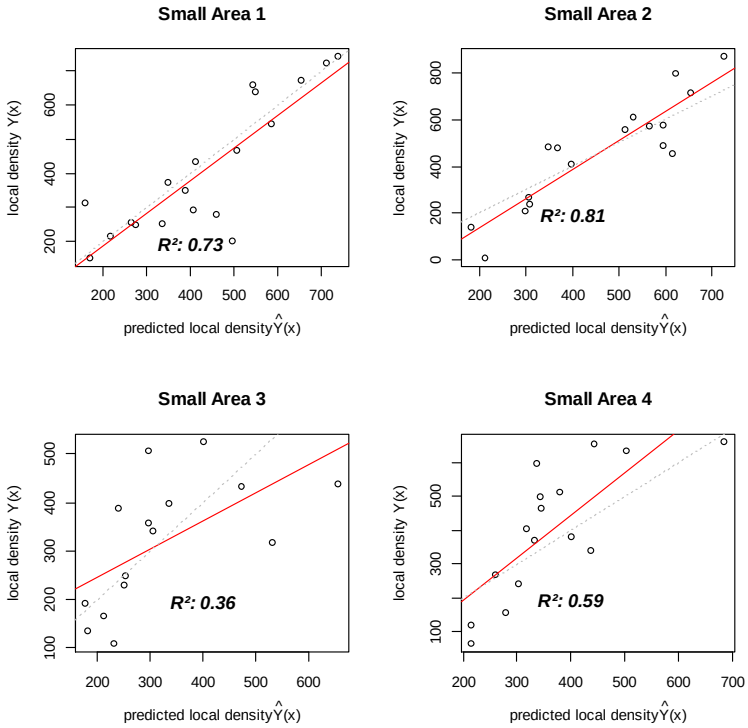
From *left* to *right*: Tukey-Anscombe Plot (fitted values vs. model residuals), QQ-Plot including 95% confidence intervals (dotted red lines) and result of the Shapiro-Wilk-test, and Scale-Location Plot (fitted values vs. square root of the standardized residuals).

Indicator for 2nd SA - Model Performance in Small Areas

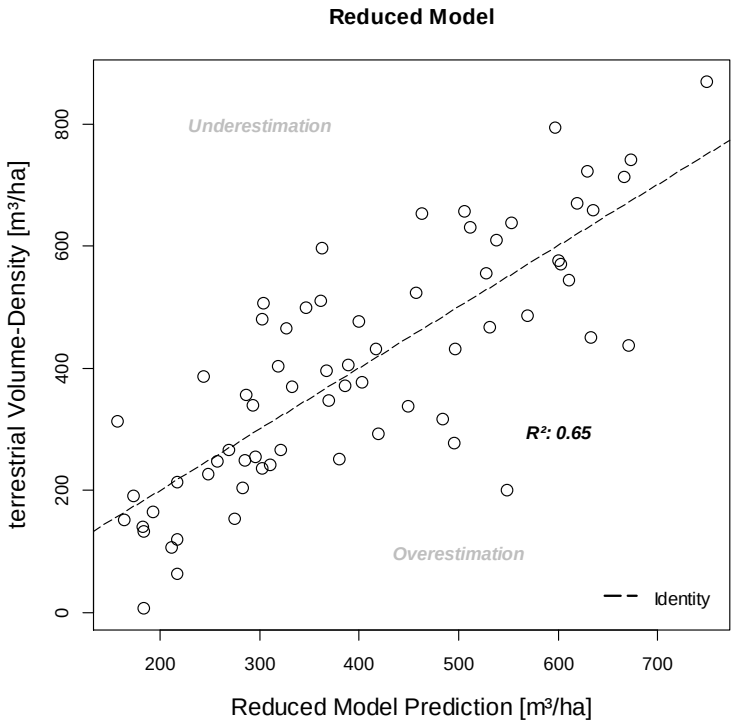
Reduced Model

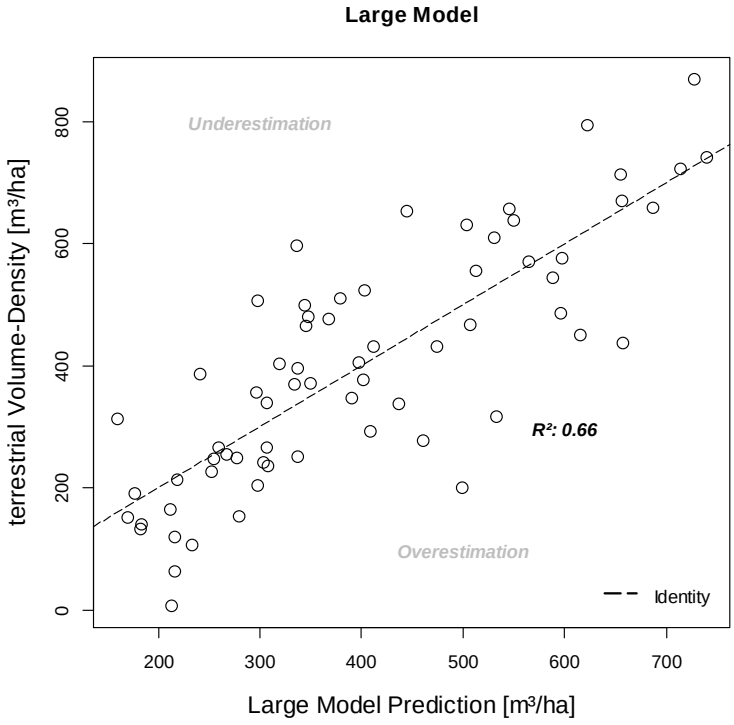


Large Model



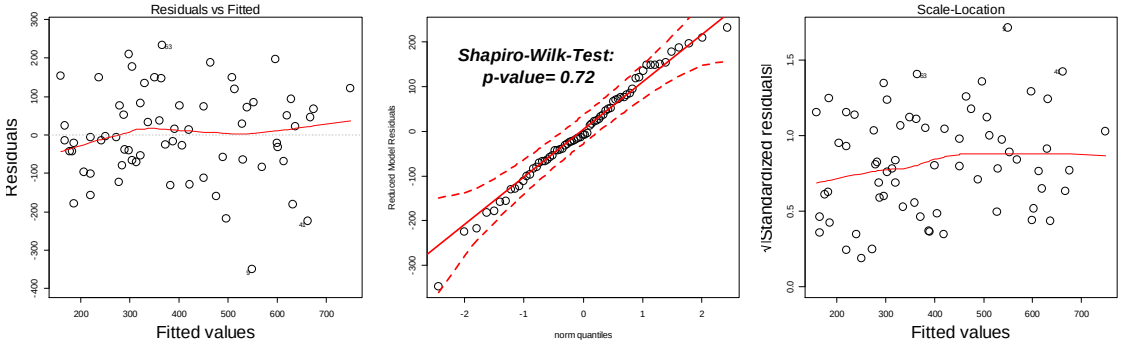
Indicator for 2nd SA - Global Model Performance in F



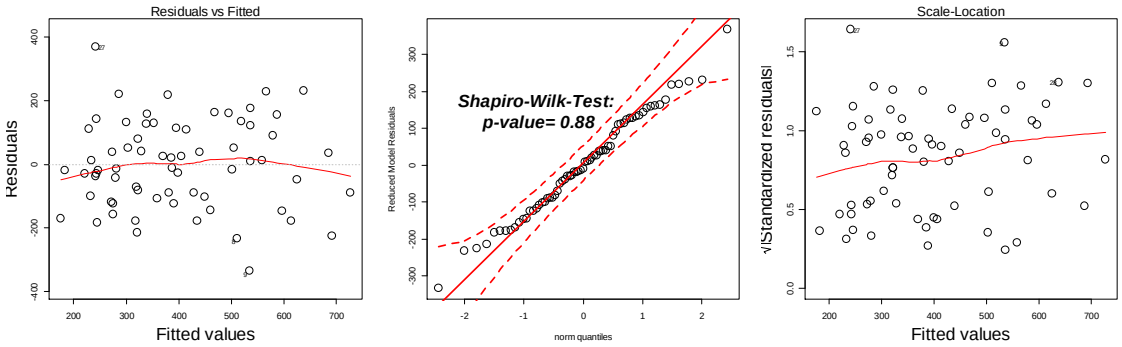


B4. Indicator for 3rd SA - Model Diagnostic for entire inventory area

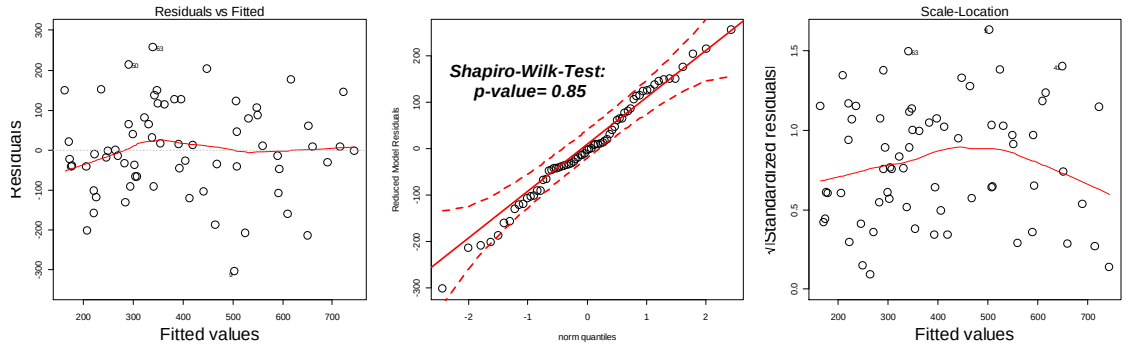
Global Reduced Model Diagnostics



Global Alternative Reduced Model Diagnostics



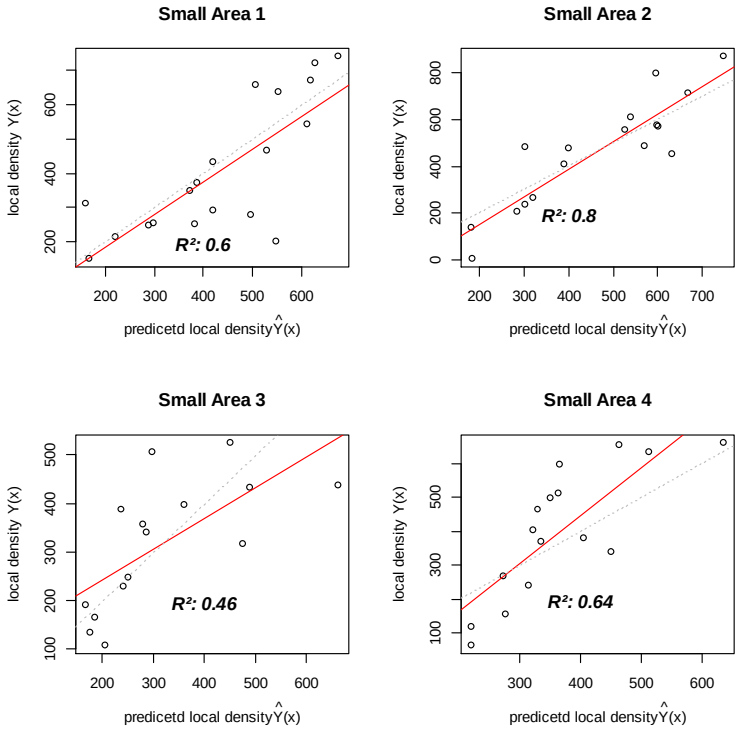
Global Large Model Diagnostics



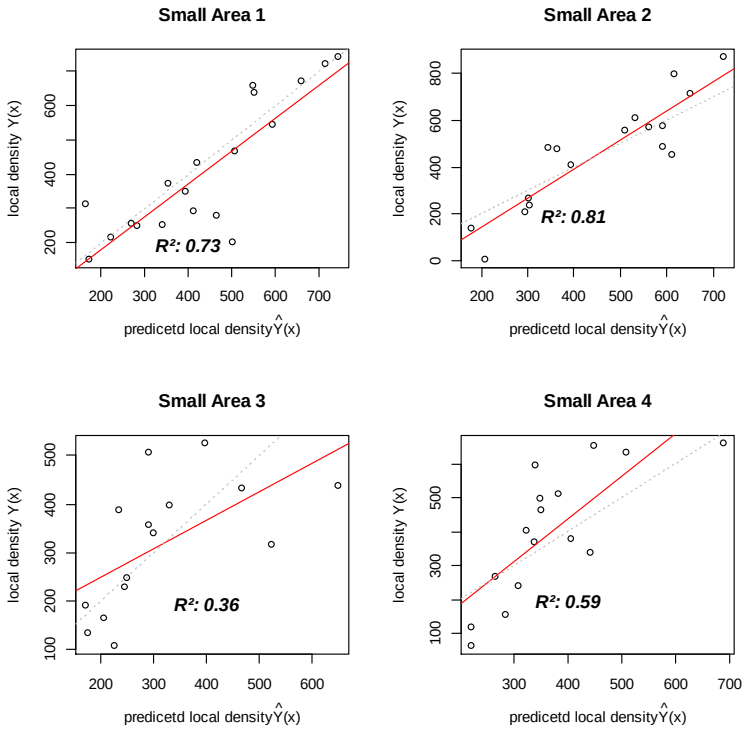
From *left* to *right*: Tukey-Anscombe Plot (fitted values vs. model residuals), QQ-Plot including 95% confidence intervals (dotted red lines) and result of the Shapiro-Wilk-test, and Scale-Location Plot (fitted values vs. square root of the standardized residuals).

Indicator for 3rd SA - Model Performance in Small Areas

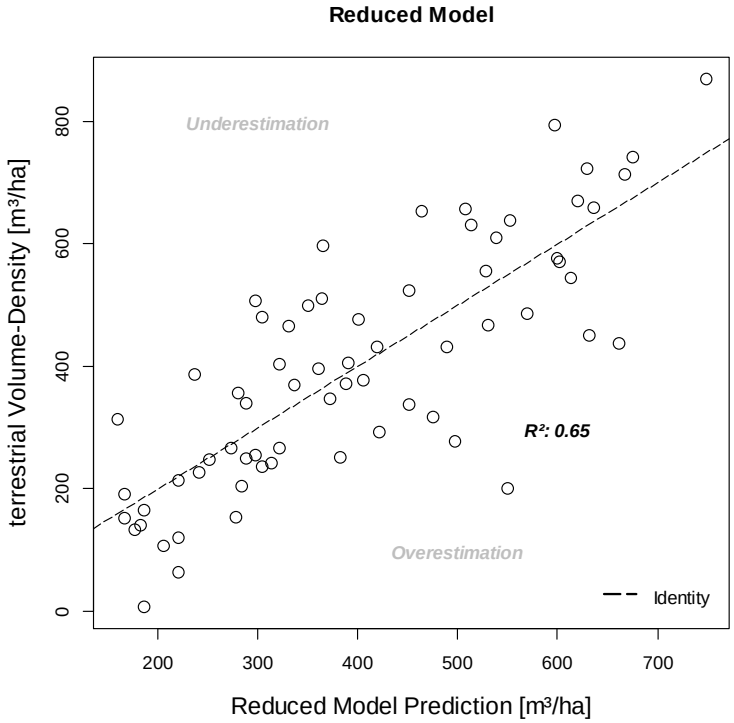
Reduced Model

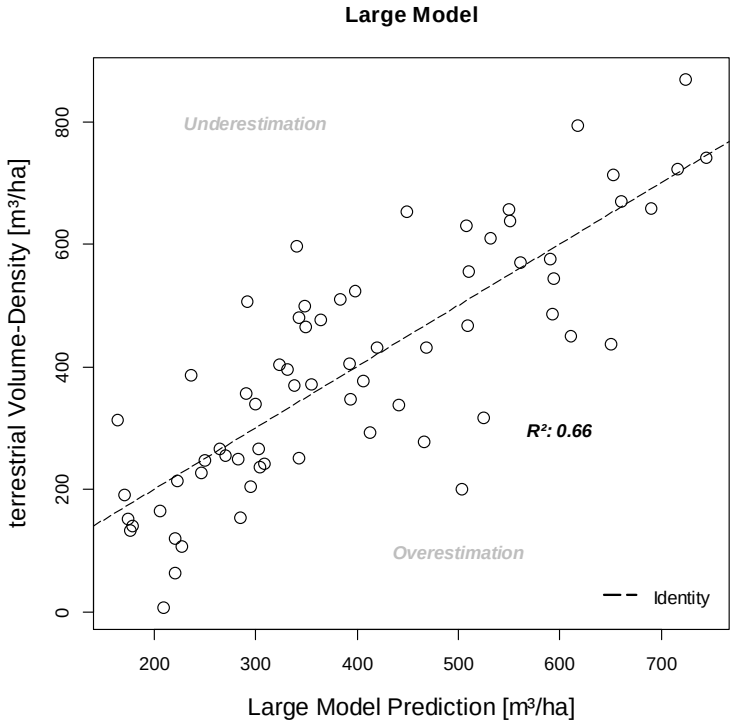


Large Model



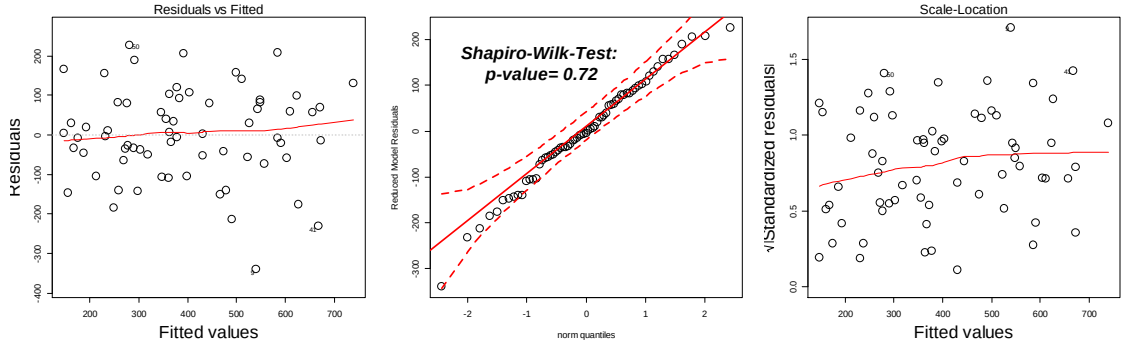
Indicator for 3rd SA - Global Model Performance in F



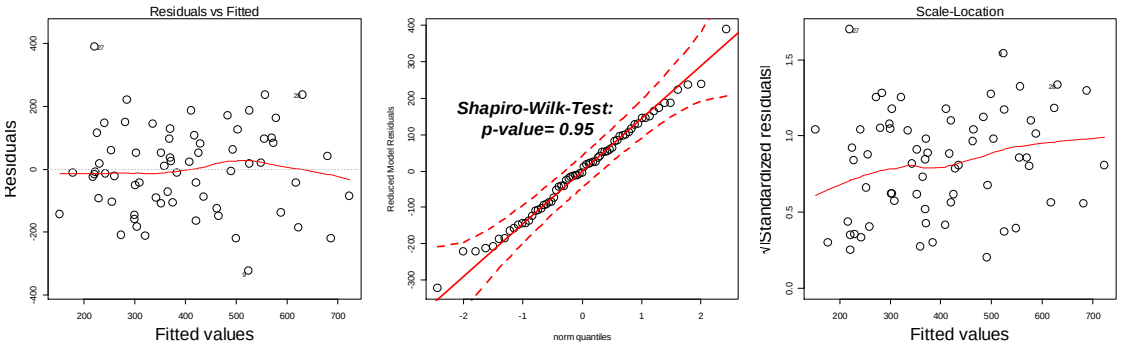


B5. Indicator for 4th SA - Model Diagnostic for entire inventory area

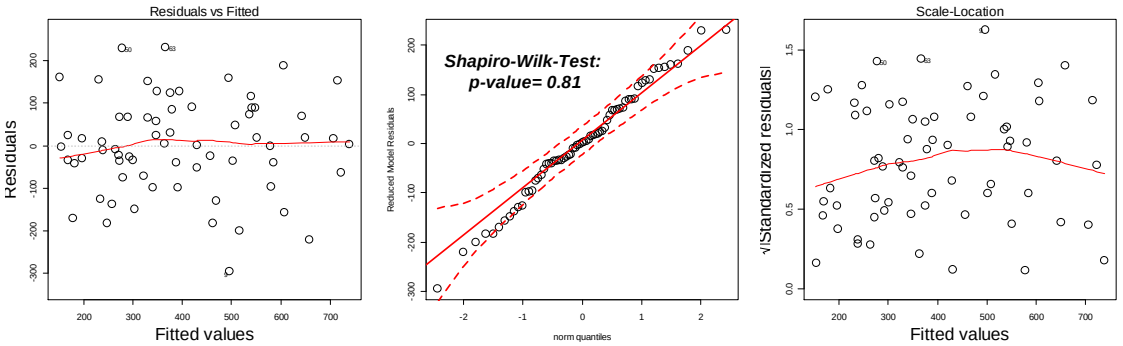
Global Reduced Model Diagnostics



Global Alternative Reduced Model Diagnostics



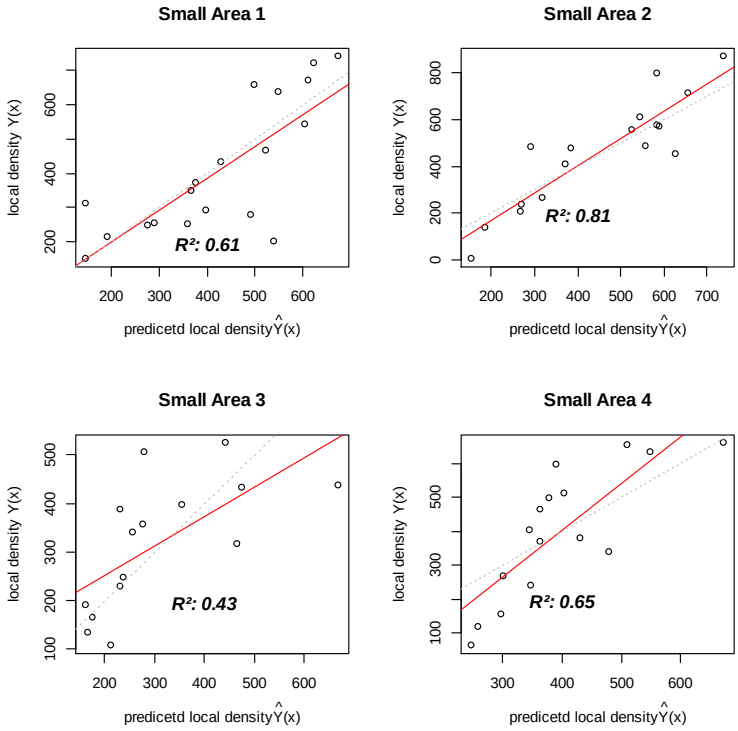
Global Large Model Diagnostics



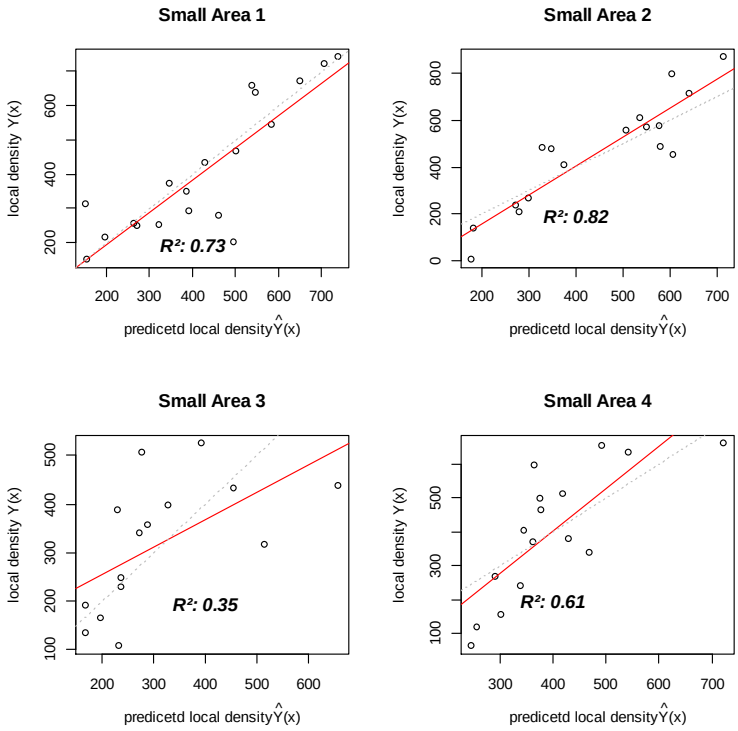
From *left* to *right*: Tukey-Anscombe Plot (fitted values vs. model residuals), QQ-Plot including 95% confidence intervals (dotted red lines) and result of the Shapiro-Wilk-test, and Scale-Location Plot (fitted values vs. square root of the standardized residuals).

Indicator for 4th SA - Model Performance in Small Areas

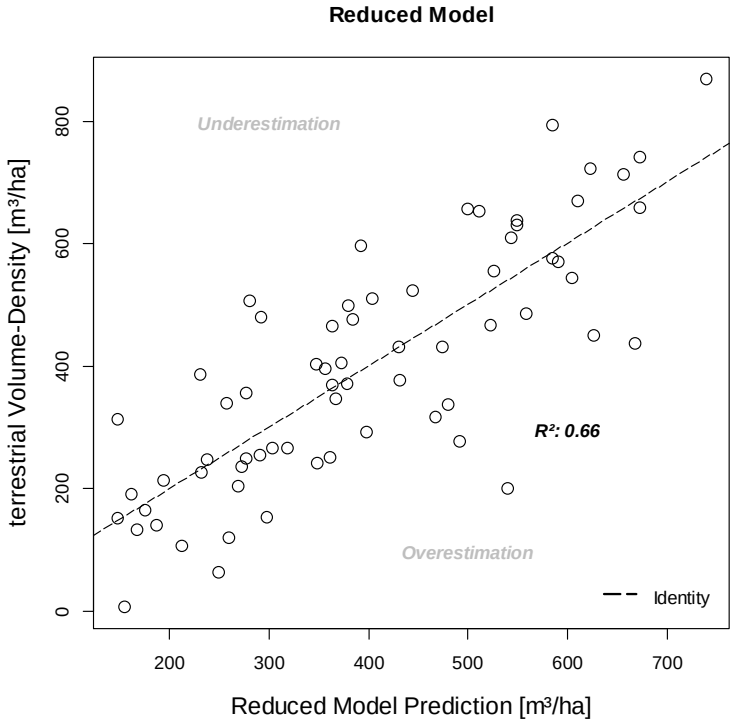
Reduced Model

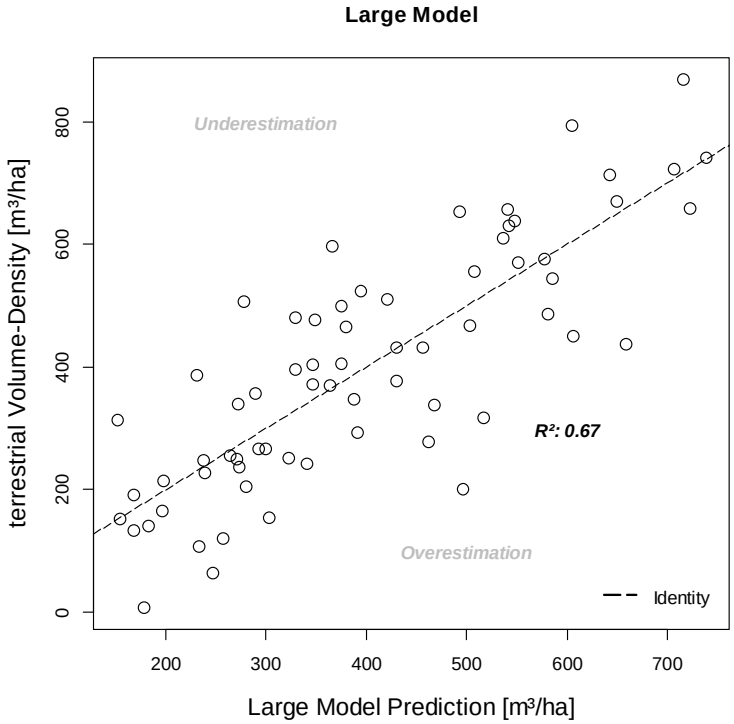


Large Model



Indicator for 4th SA - Global Model Performance in F





ANCOVA Results

B6. Testing different intercepts and slopes – *Mean*

Parameter	Estimate	Std Error	t-value	Pr(>|t|)
meanheight	21.464	4.826	4.447	3.91e-05
areaID1	96.141	76.916	1.250	0.2163
areaID2	92.109	76.510	1.204	0.2334
areaID3	189.059	72.859	2.595	0.0119
areaID4	-53.020	100.421	-0.528	0.5995
meanheight:areaID2	6.209	7.097	0.875	0.3852
meanheight:areaID3	-7.347	8.539	-0.860	0.3931
meanheight:areaID4	21.649	10.288	2.104	0.0396

Residual standard error: 132.4 on 59 degrees of freedom
Multiple R-squared: 0.587793
F-statistic: 86.77 on 8 and 59 DF, p-value: <2.2e-16

Anova Table (Type III tests)

	Sum Sq	*Df*	*F-value*	*Pr(>F)*
meanheight	346560	1	19.7770	3.908e-05
areaID	175650	4	2.5059	0.0516.
meanheight:areaID	125057	3	2.3789	0.0788.
Residuals	103388059	59		

Testing Nested Effects

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr(>|t|)*
areaID1	96.1408949	76.9158700	1.25	0.2163
areaID2	92.1092963	76.5099429	1.20	0.2334
areaID3	189.05866011	72.8591245	2.59	0.0119
areaID4	-53.019905	100.4213613	-0.53	0.5995
meanheight(areaID1)	21.4635530	4.8263783	4.45	<.0001
meanheight(areaID2)	27.6728734	5.2034728	5.32	<.0001
meanheight(areaID3)	14.1167693	7.0440788	2.00	0.0497
meanheight(areaID4)	43.1128014	9.0861747	4.74	<.0001

Anova Table (Type III tests)

	Df	*Type III Square*	*Mean Square*	*F-value*	*Pr (>F)*
areaID	4	175649.658	43912.415	2.51	0.0516
Meanheight(areaID)	4	1307068.142	326767.036	18.65	<.0001

ANCOVA Results

B7. Testing different intercepts and slopes –*Volume Density*

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr(>|t|)*
voldensity	0.39071	0.06636	5.888	1.98e-07
areaID1	168.48812	50.40844	3.342	0.00145
areaID2	154.06988	62.66235	2.459	0.01690
areaID3	282.93518	50.93748	5.555	7.00e-07
areaID4	193.45195	57.35570	3.373	0.00132
voldensity:areaID2	0.19137	0.12274	1.559	0.12432
voldensity:areaID3	-0.28325	0.13644	-2.076	0.04226
voldensity:areaID4	0.14582	0.14228	1.025	0.30962

Residual standard error: 127.3 on 59 degrees of freedom
Multiple R-squared: 0.618737
F-statistic: 94.41 on 8 and 59 DF, p-value: <2.2e-16

Anova Table (Type III tests)

	Sum Sq	*Df*	*F value*	*Pr(>F)*
voldensity	561949	1	34.6712	1.981e-07
areaID	963507	4	14.8616	1.862e-08
voldensity:areaID	167053	3	3.4356	0.02251
Residuals	956269	59		

Testing Nested Effects

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr (>|t|)*
areaID1	168.4881228	50.40844212	3.34	0.0014
areaID2	154.0698778	62.66235227	2.46	0.0169
areaID3	282.9351838	50.93748048	5.55	<.0001
areaID4	193.4519452	57.35569876	3.37	0.0013
voldensity (areaID1)	0.3907147	0.06635526	5.89	<.0001
voldensity (areaID2)	0.5820837	0.10325923	5.64	<.0001
voldensity (areaID3)	0.1074628	0.11921982	0.90	0.3711
voldensity (areaID4)	0.5365333	0.12586409	4.26	<.0001

Anova Table (Type III tests)

	Df	*Type III Square*	*Mean Square*	*F-value*	*Pr (>F)*
areaID	4	963506.955	240876.739	14.86	<.0001
Meanheight (areaID)	4	1384679.470	346169.867	21.36	<.0001

ANCOVA Results

B8. Testing different intercepts and slopes – *Mean, Volume Density*

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr(>|t|)*
voldensity	0.45230	0.10975	4.121	0.000128
areaID1	205.41509	71.41281	2.876	0.005713
areaID2	110.06313	66.77373	1.648	0.104993
areaID3	131.27518	68.62047	1.913	0.060952
areaID4	-67.43086	106.98991	-0.630	0.531138
meanheight	-5.12652	7.67720	-0.668	0.507079
voldensity:areaID2	-0.09555	0.23408	-0.408	0.684717
voldensity:areaID3	-0.90940	0.24459	-3.718	0.000472
voldensity:areaID4	-0.50752	0.26470	-1.917	0.060397
areaID2:meanheight	17.33896	12.62306	1.374	0.175141
areaID3:meanheight	41.90130	14.60211	2.870	0.005822
areaID4:meanheight	51.62643	18.40179	2.806	0.006932

Residual standard error: 114.1 on 55 degrees of freedom
Multiple R-squared: 0.714421
F-statistic: 79.87 on 12 and 55 DF, p-value: < 2.2e-16

Anova Table (Type III tests)

	Sum Sq	*Df*	*F-value*	*Pr(>F)*
voldensity	221196	1	16.9847	0.000128
areaID	195972	4	3.7620	0.008932
meanheight	5807	1	0.4459	0.507079
voldensity:areaID	204855	3	5.2433	0.002965
areaID:meanheight	172043	3	4.4035	0.007576
Residuals	716277	55		

Testing Nested Effects

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr (>|t|)*
areaID1	205.4150866	71.4128101	2.88	0.0057
areaID2	110.0631258	66.7737283	1.65	0.1050
areaID3	131.2751843	68.6204731	1.91	0.0610
areaID4	-67.4308631	106.9899123	-0.63	0.5311
meanheight (areaID1)	-5.1265240	7.6771985	-0.67	0.5071
meanheight (areaID2)	12.2124399	10.0200889	1.22	0.2281
meanheight (areaID3)	36.7747714	12.4210380	2.96	0.0045
meanheight (areaID4)	46.4999079	16.7238303	2.78	0.0074
voldensity (areaID1)	0.4523047	0.1097493	4.12	0.0001
voldensity (areaID2)	0.3567564	0.2067532	1.73	0.0900
voldensity (areaID3)	-0.4570926	0.2185885	-2.09	0.0412
voldensity (areaID4)	-0.0552172	0.2408805	-0.23	0.8195

ANCOVA Results

Anova Table (Type III tests)

	Df	*Type III Square*	*Mean Square*	*F-value*	*Pr (>F)*
areaID	4	195971.5916	48992.8979	3.76	0.0089
Meanheight (areaID)	4	239991.4486	59997.8622	4.61	0.0028
Voldensity (areaID)	4	317602.7763	79400.6941	6.10	0.0004

B9. Testing different intercepts and slopes – All Predictors

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr(>|t|)*
voldensity	0.4240	0.1105	3.837	0.000402
areaID1	214.7216	187.2353	1.147	0.257805
areaID2	118.7903	142.2488	0.835	0.408284
areaID3	343.0902	147.8761	2.320	0.025148
areaID4	-54.0863	184.7536	-0.293	0.771123
meanheight	20.8419	25.2659	0.825	0.413981
maxheight	-4.1660	10.1736	-0.409	0.684212
sd	27.5711	28.9438	0.953	0.346131
quantil75	-21.9654	19.5105	-1.126	0.266483
voldensity:areaID2	-0.2126	0.2345	-0.907	0.369713
voldensity:areaID3	-0.6730	0.2543	-2.646	0.011335
voldensity:areaID4	-0.4771	0.3530	-1.352	0.183568
areaID2:meanheight	39.5260	29.5383	1.338	0.187889
areaID3:meanheight	38.2716	45.4622	0.842	0.404539
areaID4:meanheight	-63.8643	78.4886	-0.814	0.420314
areaID2:maxheight	-14.6761	13.1271	-1.118	0.269776
areaID3:maxheight	-28.4807	20.6840	-1.377	0.175657
areaID4:maxheight	-0.7700	23.5092	-0.033	0.974025
areaID2:sd	57.8498	37.1247	1.558	0.126503
areaID3:sd	105.0388	115.8544	0.907	0.369648
areaID4:sd	-99.9019	124.8262	-0.800	0.427921
areaID2:quantil75	-15.4784	21.9344	-0.706	0.484202
areaID3:quantil75	-16.5350	51.8560	-0.319	0.751373
areaID4:quantil75	118.8240	72.1931	1.646	0.107071

Residual standard error: 100.6 on 43 degrees of freedom
Multiple R-squared: 0.826357
F-statistic: 52.5 on 24 and 43 DF, p-value: < 2.2e-16

ANCOVA Results

Anova Table (Type III tests)

	Sum Sq	*Df*	*F-value*	*Pr(>F)*
voldensity	149155	1	14.7263	0.0004023
areaID	75773	4	1.8703	0.1330799
meanheight	6892	1	0.6805	0.4139808
maxheight	1698	1	0.1677	0.6842123
sd	9191	1	0.9074	0.3461308
quantil75	12838	1	1.2675	0.2664828
voldensity:areaID	81699	3	2.6887	0.0581628.
areaID:meanheight	34387	3	1.1317	0.3469272
areaID:maxheight	25570	3	0.8415	0.4786685
areaID:sd	41014	3	1.3498	0.2708885
areaID:quantil75	40651	3	1.3379	0.2746018
Residuals	435525	43		

Testing Nested Effects

Parameter	*Estimate*	*Std Error*	*t-value*	*Pr (>|t|)*
areaID1	214.7215874	187.2353222	1.15	0.2578
areaID2	118.7903040	142.2487862	0.84	0.4083
areaID3	343.0901926	147.8761220	2.32	0.0251
areaID4	-54.0862764	184.7536152	-0.29	0.7711
meanheight (areaID1)	20.8418792	25.2659261	0.82	0.4140
meanheight (areaID2)	60.3678922	15.3018285	3.95	0.0003
meanheight (areaID3)	59.1134755	37.7947590	1.56	0.1251
meanheight (areaID4)	-43.0223995	74.3108309	-0.58	0.5656
voldensity (areaID1)	0.4240060	0.1104904	3.84	0.0004
voldensity (areaID2)	0.2114242	0.2068421	1.02	0.3124
voldensity (areaID3)	-0.2489797	0.2290968	-1.09	0.2832
voldensity (areaID4)	-0.0530785	0.3352359	-0.16	0.8749
max (areaID1)	-4.1659849	10.1735741	-0.41	0.6842
max (areaID2)	-18.8420648	8.2957682	-2.27	0.0282
max (areaID3)	-32.6466847	18.0090612	-1.81	0.0768
max (areaID4)	-4.9359397	21.1939141	-0.23	0.8169
sd (areaID1)	27.5711058	28.9438306	0.95	0.3461
sd (areaID2)	85.4209342	23.2486277	3.67	0.0007
sd (areaID3)	132.6098619	112.1806419	1.18	0.2437
sd (areaID4)	-72.3307993	121.4242400	-0.60	0.5545
Q75 (areaID1)	-21.9654348	19.5104713	-1.13	0.2665
Q75 (areaID2)	-37.4437930	10.0229693	-3.74	0.0005
Q75 (areaID3)	-38.5004830	48.0456815	-0.80	0.4273
Q75 (areaID4)	96.8585493	69.5067045	1.39	0.1706

ANCOVA Results

Anova Table (Type III tests)

	Df	*Type III Square*	*Mean Square*	*F-value*	*Pr (>F)*
areaID	4	75772.9688	18943.2422	1.87	0.1331
Meanheight (areaID)	4	192705.4560	48176.3640	4.76	0.0029
Voldensity (areaID)	4	171954.4082	42988.6021	4.24	0.0055
Max (areaID)	4	87782.3253	21945.5813	2.17	0.0889
Sd (areaID)	4	163672.4968	40918.1242	4.04	0.0072
Q75 (areaID)	4	180364.8603	45091.2151	4.45	0.0042

Appendix C Selection of Source Codes

C.1 Calculation of True Means of Exhaustive Variables

```
#-------------------------------------------------------------------------------#
# Name:       exhaustive_auxvars                                                #
# Language: Python                                                              #
#                                                                               #
# Author:     Andreas Hill, Patricia Moll                                       #
# Created:    27.03.2013                                                        #
#-------------------------------------------------------------------------------#

##############################
### Pakete laden             ##
##############################
import math
import sys
import arcpy as arc
import numpy as np
import numpy.ma as npma
import calculate_statistics as stat

##############################
### Vorprozessierung         ##
##############################

# hier liegen alle Geodaten
mainworkspace = r"C:\Users\ahill\Documents\Masterarbeit\Geodata\LiDAR_Davos"

# Datengrundlage:
CHM = mainworkspace + r"\LiDAR_Parts.gdb\CHM_GR"
maskvector = mainworkspace + r"\addGeodata.gdb\Waldmaske_Gesamtperimeter_LiDAR"
smallareavector = mainworkspace + "\Plots_LFI.gdb\SmallArea"

# exhaustivegrid zu Raster umwandeln:
from arcpy import env
env.workspace=mainworkspace
env.snapRaster= CHM
env.cellSize= CHM
env.overwriteOutput = True

# Convert to Raster
maske = mainworkspace + r"\Plots_LFI.gdb\waldmaske_raster"
smallarea = mainworkspace + r"\Plots_LFI.gdb\smallarea_raster"

##############################
### Hauptteil                ##
##############################
# Definitionen:
step = 25                           # Berechnungseinheit in [m]
halfstep = step/2.0
cellsize = 0.5
pixel = int(step/cellsize           # Anzahl Pixel pro Berechnungseinheit
Ankerpunkt = [778000, 200000]       #  LFI-Ankerpunkt, visuell bestimmen
desc=arc.Describe(maske)
extent_maske=desc.extent
```

```
xmin=extent_maske.XMin
xmax=extent_maske.XMax
ymin=extent_maske.YMin
ymax=extent_maske.YMax

xdiff = abs(xmin - Ankerpunkt[0])
ydiff = abs(ymax - Ankerpunkt[1])

# Koordinaten des Startpunktes (sind Mittelpunkte der Kacheln)
xAnker = int(Ankerpunkt[0] + math.floor(xdiff / step) * step)
yAnker = int(Ankerpunkt[1] - math.floor(ydiff / step) * step)

# Anzahl Berechnungseinheiten in x- und y-Richtung
xCount = int(math.ceil(abs(xmax-xAnker) / step))
yCount = int(math.ceil(abs(ymin-yAnker) / step))

# Koordinaten des Endpunktes
xEnd = xAnker + xCount * step
yEnd = yAnker - yCount * step

# Anzahl an Iterationen gesamt
iterations=xCount*yCount
no_iterations=0

# Initialisierung:
narea = 4
stats = { 'ratio': [0]*narea,
          'mean': [0]*narea,
          'median': [0]*narea,
          'std': [0]*narea,
          'mad': [0]*narea,
          'maxim': [0]*narea,
          'quants': [[0]*3, [0]*3, [0]*3, [0]*3],
           'varcoef': [0]*narea }

result ={  'mean': [0]*narea,
           'median': [0]*narea,
           'std': [0]*narea,
           'mad': [0]*narea,
           'maxim': [0]*narea,
           'quants': [[0]*3, [0]*3, [0]*3, [0]*3, [0]*3],
           'varcoef': [0]*narea }

# Statistische Auswertung pro Berechnungseinheit
for x in range(xAnker, xEnd, step): # xEnd
   for y in range(yAnker, yEnd, -step): # yEnd
      no_iterations = no_iterations+1
      print "iteration",no_iterations,"of",iterations,"iterations calculating"
      print "X-Coordinate is: ",x
      print "Y-Coordinate is: ",y
      llcorner = arc.Point(x-halfstep, y-halfstep)
      mask = arc.RasterToNumPyArray(maske, llcorner, pixel, pixel, 0)
      if np.max(mask)==0:
         continue
      center = mask[pixel/2, pixel/2]
      if center > 0:
         print "center within forest mask!"
         area = arc.RasterToNumPyArray(smallarea, llcorner, pixel, pixel, 0)
         chm = arc.RasterToNumPyArray(CHM, llcorner, pixel, pixel, np.nan)
         chm = npma.masked_invalid(chm) # NaNs werden ausmaskiert
         saID = area[pixel/2, pixel/2]
```

```
        print "area: ",saID
        if saID==0:
            print "%d, %d ---Error! No Small Area information available" %(x, y)
            sys.exit()
        stats = stat.Statistik(chm,mask,saID,stats,narea,pixel)
print "Statistics successfully calculated"

# Write out indices of ratio values > 0
inx = np.where(np.array(stats['ratio'])!=0)
for item in stats.items():
    # No statistics for summarized ratio
    if item[0] == 'ratio':
        continue
    if item[0]=='quants':
        for a in inx[0]:
            result[item[0]][a] = list(np.array(item[1][a])/stats['ratio'][a])
            resmat=np.matrix(item[1])
            for k in range(resmat.shape[0]-1):
                result[item[0]][narea][k]=np.sum(resmat[:,k])/sum(stats['ratio'])
        continue
    # Summarized statistics for every other case
    result[item[0]]=list(np.array(item[1])/np.array(stats['ratio']))
    result[item[0]].append(sum(item[1])/sum(stats['ratio']))

# writing output-file:
import os
filename = r'C:\Users\ahill\Documents\Masterarbeit\Local Inventory Data
GR\MeanExhaustiveAuxvars/exhaustive_varmeans.txt'
quantiles = ['q25', 'q75', 'q90']
with open(filename, 'w') as f:
    for item in result:
        if item == 'quants':
            mat=np.matrix(result['quants'])
            for j in range(mat.shape[1]):
                f.write(quantiles[j] + '\n')
                for k in range(mat.shape[0]):
                    f.write(str(mat[k,j])+'\n')
            continue
        f.write(item + '\n')
        for i in range(len(result[item])):
            f.write(str(result[item][i])+'\n')
print "Output-File has been written"
```

```
#-------------------------------------------------------------------------------#
# Name:       calculate_statistics                                              #
# Language: Python                                                              #
#                                                                               #
# Author:     Andreas Hill, Patricia Moll                                       #
# Created:    27.03.2013                                                        #
#-------------------------------------------------------------------------------#

def Statistik(chm,mask,saID,stats,narea,pixel):

    # Libraries laden:
    import numpy as np

    valid=np.where(mask>0)
    chm=chm[valid].compressed() # compressed:Nur unmaskierte Werte (== Wald)
    count = len(chm)
    ratio = count/pixel**2.0
    stats['ratio'][saID-1] = stats['ratio'][saID-1] + ratio


    # Berechnung Mittelwert
    mittel = np.mean(chm)
    stats['mean'][saID-1] = stats['mean'][saID-1] + (ratio*mittel)

    # Berechnung Median:
    median = np.median(chm)
    stats['median'][saID-1] = stats['median'][saID-1] + (ratio*median)

    # Berechnung Standardabweichung:
    std = np.std(chm)
    stats['std'][saID-1]= stats['std'][saID-1] + (ratio*std)

    # Berechnung median absolut deviation:
    mad = np.median(np.abs(chm - median))
    stats['mad'][saID-1]= stats['mad'][saID-1] + (ratio*mad)

    # Berechnung maximum:
    maxim = np.max(chm)
    stats['maxim'][saID-1] = stats['maxim'][saID-1] + (ratio*maxim)

    # Berechnung Quantile:
    quants =np.percentile(chm,[25,75,90])
    for i in range(3):
        stats['quants'][saID-1][i] = stats['quants'][saID-1][i] + (ratio*quants[i])

    # Berechnung varcoef:
    varcoef = std/mittel
    stats['varcoef'][saID-1] = stats['varcoef'][saID-1]+ (ratio*varcoef)

    return stats
```

C.2 Calculation of Auxiliary Variables of First Phase Sample Plots

```
%------------------------------------------------------------------------------- %
% Calculation of auxiliary variables for 1st-phase plots                        %
% Language: MATLAB                                                              %
% Calculates auxiliary variables for all 1st-phase plots                        %
%                                                                               %
% Author: Andreas Hill                                                          %
% Date: 24.02.2013                                                              %
% ------------------------------------------------------------------------------%

%% Choice of Parameter :

% Filtering-Parameter:
filtsize = 5;
stdv = 1;

% Cellsize of CHM:
CellSize = 0.5;

% Probekreis-Parameter:
outer_Radius   = 12.62;
PRadius_outer = floor(outer_Radius/CellSize);
[Circlemask,CircleArea] = createCirclemask(outer_Radius,CellSize);

% Boundary-Adjustment:
boundadj='yes';

% Load List of 1st-phase-plot-IDs (CHM- Cuts of 1st-Phase-Sample Points) and Small Area IDs
ID=load('C:\Users\ahill\Documents\Masterarbeit\LocalInventoryDataGR\1stphasegrid\List_1stphasepoints_sma
llarea.txt');
if size(ID,2)> 3
   ID = ID(:,3:4); % original 1st and 2nd column are coordinates, new 1st column: ID, new 2nd
                   % column: small area ID (if existing)
   PlotID = ID(:,1);
   smallareaID = ID(:,2);
else
   PlotID = ID(:,3); % 1st column: ID
   smallareaID = NaN*ones(size(ID,1));
end
n_ID = size(PlotID,1);

% Initialising:
plot_terr = 0;
plot_results = zeros(n_ID,15);

%% Loop over all 1st-Phase-Points:
for i=1:n_ID
    disp(['Calculating Plot ',int2str(PlotID(i)),' from ',int2str(n_ID),' Plots'])
   % Load and Import the respective CHM (large Cut)
     CHM_Filename=['C:\Users\ahill\Documents\Masterarbeit\Geodata\LiDAR_Davos\Probeflächen_CHM_GR\1
                   stphase_Asci_buf50\chm_', int2str(PlotID(i)),'.asc'];
   [CHM,projection]=ImportRaster(CHM_Filename);
   [CHM_filtered]= smoothCHM(CHM,filtsize,stdv);
    CHM_center = [floor(size(CHM_filtered,1)/2)+1, floor(size(CHM_filtered,2)/2)+1];

   % Load and Import the respective Forestmask (if available) (large Cut):
   if strcmp(boundadj,'yes')==1
```

```
        fmask_Filename=['C:\Users\ahill\Documents\Masterarbeit\Geodata\LiDAR_Davos\Probeflächen_CHM_G
                        R\1stphase_Asci_buf50\waldmaske_', int2str(PlotID(i)),'.asc'];
      [fmask,projection]=ImportRaster(fmask_Filename);
    else
      fmask='not available'; % Forestmask not available
    end

    % Calculate Plot-Centre and calculate Cutting-Parameters:
    CHM_Coordinate_x          = CHM_center(1,2);
    CHM_Coordinate_y          = CHM_center(1,1);
    cutoriginal.x_left        = CHM_Coordinate_x -(PRadius_outer);
    cutoriginal.x_right       = CHM_Coordinate_x +(PRadius_outer);
    cutoriginal.y_up          = CHM_Coordinate_y +(PRadius_outer);
    cutoriginal.y_down        = CHM_Coordinate_y -(PRadius_outer);

    % Calculate Auxiliary Variables for Sample Points:
    [CHM_Parameters] = calculate_CHMParameters(CHM,cutoriginal,fmask);
    [VolumeDensity,Volume,Stemnumber_ha]  =
    calculate_PlotVolumeDensity(CHM_filtered,cutoriginal,outer_Radius,CellSize,fmask);

     plot_results(i,1:15)=[PlotID(i) CHM_Parameters.meanheight CHM_Parameters.med  CHM_Parameters.sd
    CHM_Parameters.MAD CHM_Parameters.maxheight  CHM_Parameters.quants CHM_Parameters.varcoeff
    CHM_Parameters.weight VolumeDensity Stemnumber_ha plot_terr smallareaID(i)];

end

% Write Results for 1st-Phase-Plots to Textfile:
fid=fopen('C:/Users/ahill/Documents/Masterarbeit/LocalInventoryData
          GR/Auxvariables_1stphasePlots/FirstPhasePlot_results_nobadj.txt','wt');
fprintf(fid,'%s\n %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t %s\t
        %s\n',['BoundaryAdjustment:',boundadj],'ID','meanheight','median','sd','mad','maxheight','quantil25','q
        uantil75','quantil90',,'varcoef','weight','voldensity','stemnumber_ha','plot.terr','areaID');
for h=1:n_ID
   fprintf(fid,'%g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\t %g\n'
           ,plot_results(h,1:15));
end
```

Erklärung / Non-Plagiarism Statement

Hiermit versichere ich gemäß § 7 Abs. 5 der Master-Prüfungsordnung vom 23.09.2010, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Göttingen, 11.07.2013

(Andreas Hill)