Dissertationes Forestales 201

Use of remotely sensed auxiliary data for improving sample-based forest inventories

Svetlana Saarela

Department of Forest Sciences
Faculty of Agriculture and Forestry
University of Helsinki

Academic dissertation

To be presented, with the permission of the Faculty of Agriculture and Forestry of the University of Helsinki, for public criticism in Lecture Hall (Luentosali) 5 of B-building (Latokartanonkaari 7) on September 25th, 2015, at 1400 hrs.

Title: Use of remotely sensed auxiliary data for improving sample-based forest inventories

Author: Svetlana Saarela

Dissertationes Forestales 201

DOI: http://dx.doi.org/10.14214/df.201

Thesis supervisors:
Professor Bo Dahlin
Department of Forest Sciences, University of Helsinki, Finland

Anton Grafström, Senior Lecturer, Ph.D.
Department of Forest Resource Management, Swedish University of Agricultural Sciences, Sweden

Pre-examiners:
Professor Piermaria Corona
Director of the Forestry Research Centre (CRA-SEL), Full Professor at the University of Tuscia, Italy

Ronald E. McRoberts, Ph.D.
Forest Inventory and Analysis, USDA, Northern Research Station, U.S.A.

Opponent:
Professor Timothy G. Gregoire
School of Forestry and Environmental Studies, Yale University, U.S.A.

Cover photo:
The hybrid inference for population mean prediction based on ordinary least squares regression with homo- and heteroskedastic residuals; the background photo was taken in the Red Canyon (Utah, U.S.A.) by Svetlana Saarela.

ISSN 1795-7389 (online)
ISBN 978-951-651-491-18(pdf)
ISSN 2323-9220 (print)
ISBN 978-951-651-492-8 (paperback)

Printers:
Unigrafia, Helsinki 2015

Publisher:
Finnish Society of Forest Sciences
Natural Resources Institute Finland
Faculty of Agriculture and Forestry at the University of Helsinki
School of forest Sciences of the University of Eastern Finland

Editorial Office:
The Finnish Society of Forest Sciences
P.O. Box 18, FI-01301 Vantaa, Finland
http://www.metla.fi/dissertationes

Saarela, S. 2015. Use of remotely sensed auxiliary data for improving sample-based forest inventories. Dissertationes Forestales 201, 36 p.
Available at http://dx.doi.org/10.14214/df.201

ABSTRACT

Over the past decades it has been shown that remotely sensed auxiliary data have a potential to increase the precision of key estimators in sample-based forest surveys. This thesis was motivated by the increasing availability of remotely sensed data, and the objectives were to investigate how this type of auxiliary data can be used for improving both the design and the estimators in sample-based surveys. Two different modes of inference were studied: model-based inference and design-based inference. Empirical data for the studies were acquired from a boreal forest area in the Kuortane region of western Finland. The data comprised a combination of auxiliary information derived from airborne LiDAR and Landsat data, and field sample plot data collected using a modification of the 10th Finnish National Forest Inventory. The studied forest attribute was growing stock volume.

In Paper I, remotely sensed data were applied at the design stage, using a newly developed design which spreads the sample efficiently in the space of auxiliary data. The analysis was carried out through Monte Carlo sampling simulation using a simulated population developed by way of a copula technique utilizing empirical data from Kuortane. The results of the study showed that the new design resulted in a higher precision when compared to a traditional design where the samples were spread only in the space of geographical data.

In Paper II, remotely sensed auxiliary data were applied in connection with model-assisted estimation. The auxiliary data were used mainly in the estimation stage, but also in the design stage through probability-proportional-to-size sampling utilizing Landsat data. The results showed that LiDAR auxiliary data considerably improved the precision compared to estimation based only on field samples. Additionally, in spite of their low correlation with growing stock volume, adding Landsat data as auxiliary data further improved the precision of the estimators.

In Paper III, the focus was set on model-based inference and the influence of the use of different models on the precision of estimators. For this study, a second simulated population was developed utilizing the empirical data, including only non-zero growing stock volume observations. The results revealed that the choice of model form in model-based inference had minor to moderate effects on the precision of the estimators. Furthermore, as expected, it was found that model-based prediction and model-assisted estimation performed almost equally well.

In Paper IV, the precision of model-based prediction and model-assisted estimation was compared in a case where field and remotely sensed data were geographically mismatched. The same simulated population as used in Paper III was employed in this study. The results showed that the precision in most cases decreased considerably, and more so when LiDAR auxiliary data were applied, compared to when Landsat auxiliary data were used. As for the choice of inferential framework, it was revealed that model-based inference in this case had some advantages compared to design-based inference through model-assisted estimators.

The results of this thesis are important for the development of forest inventories to meet the requirements which stem from an increasing number of international commitments and agreements related to forests.

Keywords: design-based, Landsat, LiDAR, model-based, multivariate probability distribution, sampling.

ACKNOWLEDGMENT

If someone would have told me ten years ago that I would defend a doctoral dissertation at the University of Helsinki – I would not have believed it. In 2005, as an exchange student from the Saint-Petersburg North-West Technical University at the Joensuu University (the current University of Eastern Finland), I looked at the gaining of a doctoral degree at the University of Helsinki as an absolutely impossible mission! But, as my supervisor Dr. Anton Grafström says, everything is possible. My deep gratitude goes to him for his patience and ability to explain advanced mathematical statistical issues in an easy way. Without his support and supervision this dissertation would never have been written. I also thank my supervisor Prof. Bo Dahlin for his supervision during the final stages.

This dissertation is a result of hard work which has been supported by many researchers from Finland, Norway and Sweden. I wish to thank Prof. Annika Kangas, Dr. Sakari Tuominen and M.Sc. Andras Balazs from the Natural Resources Institute Finland (LUKE), Prof. Juha Hyyppä from the Finnish Geospatial Research Institute (FGI), Prof. Markus Holopainen from the University of Helsinki, and Dr. Liviu Theodor Ene from the Norwegian University of Life Sciences (NMBU) for believing in me at the beginning of this journey and supporting my doctoral studies. At the Swedish University of Agricultural Sciences (SLU) I am very grateful to the Remote Sensing division team – Dr. Eva Lindberg, M.Sc. Karin Nordkvist, Dr. Jonas Bohlin, Dr. Kenneth Olofsson and Dr. Mattias Nyström. Especially, my thanks go to the Forest Resource Analysis team – Dr. Anton Grafström – my main supervisor, Prof. Göran Ståhl who has inspired me as a teacher, and Dr. Sebastian Schnell. The several months I spent in Umeå were the most intensive learning months of my life and the turning point in my doctoral studies. I very much enjoyed being at SLU, where the environment was very stimulating in which to work and enabled me to achieve more.

I wish to express my appreciation and gratitude to my doctoral dissertation's pre-examiners Prof. Piermaria Corona (University of Tuscia, Italy) and Dr. Ronald E. McRoberts (Northen Research Station, U.S.A.) for their thorough evaluations, which have helped to improve this work considerably.

I thank my closest friend Helena Saarela for being my family in Finland. Every time I visited you I could just relax and enjoy the comfort and love around me ~ Kiitän läheisintä ystävääni, Helena Saarelaa, siitä, että hän on ollut tukeni ja turvani Suomessa. Luonasi olen aina saanut rentoutua ja nauttia mukavasta ja rakastavasta ilmapiiristä.

But my greatest love and deepest gratitude goes to my mother – my greatest fan and always a believer in me! You do not always approve of what I do, but you always support me, patiently listening to my tears of both sadness and happiness. Thank you mother for raising me into what I am today, this dissertation is your achievement as well! I also thank my brother for his love and everlasting support.

Но мою самую большую любовь и глубокую благодарность я выражаю моей маме – моему самому большому фанату, кто верит в меня! Ты не всегда одобряешь то, что я делаю, но всегда поддерживаешь меня, терпеливо выслушивая мои слезы печали и радости. Спасибо, мама, что вырастила меня такой, какая я есть сегодня, эта диссертация и твоя заслуга! Так же я благодарю моего брата за его любовь и нескончаемую поддержку.

Helsinki, September 2015
Svetlana Saarela

LIST OF ORIGINAL ARTICLES

This thesis consists of the following research articles, which are referred to in the text by Roman numerals. Articles I – III are reproduced with the permission of publishers, while study IV is the author's version of the submitted manuscript.

I. Grafström A., Saarela S., Ene L. T. (2014). Efficient sampling strategies for forest inventories by spreading the sample in auxiliary space. *Canadian Journal of Forest Research* 44(10): 1156-1164. http://dx.doi.org/10.1139/cjfr-2014-0202

II. Saarela S., Grafström A., Ståhl G., Kangas A., Holopainen M., Tuominen S., Nordkvist K., Hyyppä J. (2015). Model-assisted estimation of growing stock volume using different combinations of LiDAR and Landsat data as auxiliary information. *Remote Sensing of Environment* 158: 431-440. http://dx.doi.org/10.1016/j.rse.2014.11.020

III. Saarela S., Schnell S., Grafström A., Tuominen S., Nordkvist K., Hyyppä J., Kangas A., Ståhl G. (2015). Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume in Kuortane, Finland. *Canadian Journal of Forest Research* 45: 1524-1534. http://dx.doi.org/10.1139/cjfr-2015-0077

IV. Saarela S., Schnell S., Tuominen S., Balazs A., Hyyppä J., Grafström A., Ståhl G. (2015). Effects of positional errors in model-assisted and model-based estimation of forest resources using a combination of field plots and remotely sensed data. *Remote Sensing of Environment*. (Revised manuscript submitted.)

The contributions of Svetlana Saarela to the papers included in this thesis were as follows:

I. Planned the study and prepared the simulated population with co-authors. Wrote parts of the paper related to the data description. Was partly responsible for the responses to Reviewers.

II. Planned the study and processed data with co-authors. Developed the R-code for the simulator together with Grafström. Was responsible for the calculation and interpretation of the results. Carried out the literature review and wrote the major part of the manuscript. Was responsible for responses to Reviewers.

III. Planned the study with co-authors. Created the simulated population. Developed the R-code for the simulator together with Schnell. Carried out the literature review and wrote the major part of the manuscript. Was responsible for the responses to Reviewers.

IV. Planned the study with co-authors. Created the simulated population with the buffer zone. Developed the R-code for the simulator. Carried out the literature review and wrote the major part of the manuscript. Was responsible for the responses to Reviewers.

TABLE OF CONTENTS

SYMBOLS AND ABBREVIATIONS

NFI	National Forest Inventory
REDD+	Reducing Emissions from Deforestation and Forest Degradation
LiDAR	Light Detection And Ranging
RS	Remotely Sensed
InSAR	Interferometric Synthetic Aperture Radar
3D	3 Dimensional
Landsat 7 ETM+	Enhanced Thematic Mapper Plus, sensor on-board Landsat 7
DBH	Diameter at Breast Height
SD	Standard Deviation
DSM	Digital Surface Model
DEM	Digital Elevation Model
OPALS	Orientation and Processing of Airborne Laser scanning data
DVM	Digital Vegetation Model
CRR	Canopy Relief Ratio
HT	Horvitz-Thompson estimator
NN	Nonparametric
SI	Simple random sampling without replacement
πps	Probability-proportional-to-size sampling
OLS	Ordinary Least Square
NLS	Nonlinear Least Square
HC	Heteroskedasticity-Consistent
NHC	Nonlinear Heteroskedasticity-Consistent
LINEAR	Linear regression model
LOG-LOG	Log-transformed multiplicative regression model
SQRT	Square root transformed regression model
BIAS	Bias
RBIAS	Relative Bias
RSE	Relative Standard Error
RRMSE	Relative Root Mean Square Error

1. INTRODUCTION

1.1. General background

Forest resources are required for an increasing number of purposes globally, including wood- and fiber-based raw materials, the maintenance of biodiversity, and the mitigation of climate change (Mery et al. 2005). As a consequence, the demands for information derived from forests are steadily increasing (Tomppo 2006; Cienciala et al. 2008; UNECE and FAO 2011). National forest inventories (NFIs) have been established for a long time in many countries (e.g., Tomppo et al. 2010). Normally, they are based on statistical samples consisting of field plots (McRoberts et al. 2009, 2010; Woodall et al. 2009; Ståhl et al. 2012) as a means for ensuring trustworthy information, i.e. information derived from estimators that are unbiased and which have high precision.

Field-based forest inventories have many advantages. However, they become expensive when a large sample size is required to reach the needed levels of precision. Furthermore, sparse road networks or other conditions in a country may prevent easy access to the plots. Also, NFI information from field plots alone often leads to imprecise estimates for small regions within a country, due to rather small plot sample sizes and highly variable populations of interest. This has stimulated the development of solutions where field plots and remotely sensed (RS) data are combined in order to provide the required information (Holmström et al. 2001; Maltamo et al. 2007; Næsset et al. 2004).

Lately, the REDD+ mechanism (reducing emissions from deforestation and forest degradation; Angelsen and Brockhaus (2009)), which has been developed under the United Nations' Framework Convention on Climate Change, has led to an even stronger focus on forest information and NFIs, and on how remote sensing within NFIs is utilized, especially in countries with poor infrastructure conditions. Several approaches based on remote sensing have been developed and demonstrated (e.g., Næsset et al. 2006; Gobakken et al. 2012; Nelson et al. 2008, 2009). However, inventories that make use of auxiliary information from remote sensing are not only relevant for developing countries and REDD+ (e.g., Asner 2009; Saatchi et al. 2011), but also for remote areas in developed countries such as Siberia and Alaska (e.g., Andersen et al. 2009, 2012; Nelson et al. 2009; Ene et al. 2012a). Furthermore, in countries with well-established field-based NFIs, sample-based combinations of field and RS data may offer new possibilities to make inventories more cost-efficient.

The problems involved in reaching good inventory solutions include the variable and sometimes limited information in RS data, the need to combine remote sensing with field information in order to obtain reliable results, the lack of adequate field samples, the need to apply advanced statistical methods, and the challenge to make the solutions straightforward enough so that they can be easily employed in practice. Different inferential frameworks are available for the combination of field and RS data. A well-known approach is to use RS data only for stratification and post-stratification (e.g., McRoberts et al. 2002; Nilsson et al. 2003; Saarela et al. 2012). More advanced, and currently rather intensively studied methods include design-based model-assisted and model-based estimations approaches (e.g., Opsomer et al. 2007; Baffetta et al. 2009; Gregoire et al. 2011; Ståhl et al. 2011; Breidenbach and Astrup 2002; Næsset et al. 2013a).

1.2. Two general philosophies of inference – model-based and design-based

NFIs and other large-scale forest surveys are normally based on design-based inference, i.e. the populations of trees and other elements of interest within a country are considered as fixed, and thus there exist fixed but unknown population totals and means that can be estimated from sample data. Estimates of population parameters are random variables due to random selection of population elements into the sample (e.g., Särndal et al. 1992; Gregoire 1998). Design-based inference at least dates back to Neyman (1934). This paper shaped the domination of an inferential framework wherein inference is independent from any assumptions about population structure and distribution (Gregoire 1998). However, at the time Neyman published his paper, design-based inference had already been applied in forest surveys for more than a decade in the Nordic countries. Some key assumptions underlying design-based inference are (i) the values that are linked to the population elements are fixed (ii) the population parameters about which we wish to infer information are also fixed, (iii) estimates of the parameters are random because a

random sample is selected according to some design such as simple random sampling, and (iv) the probability of obtaining different samples can be deduced and used for the inference.

However, design-based sampling is not the only inferential mode that can be applied in survey sampling. Since Matérn (1960) presented his influential paper on model-based inference within forest surveys, there has been a dispute around whether or not classical design-based inference can be replaced by model-based inference (Cassel et al. 1977; Särndal 1978; Gregoire 1998; McRoberts 2010a). The assumption underlying model-based inference is that there is a model which generates random values of the population elements. This model is often known as a superpopulation model from which the actual population is a realisation (Cassel et al. 1977; Särndal 1978; Gregoire 1998; McRoberts 2010a). Since the individual values of population elements are random variables, the population total and mean are also random variables. Estimates (sometimes termed as predicted values in the case of model-based inference) are random variables, even if the sample is selected by following non-random principles. A detailed description of model-based inference can be found in Cassel et al. (1977). Several studies discuss the potential advantages of model-based inference in survey sampling (e.g., Cassel et al. 1977; Särndal 1978; Gregoire 1998; McRoberts 2010a). In forest surveys, model-based estimation has advantages in small-area estimation (e.g., Prasad and Rao 1990; Lappi 2001; Breidenbach and Astrup 2012) in which case it is typically called synthetic estimation. Synthetic estimators are based on models developed outside the target area, which is straightforward in cases such as when NFI data are applied for developing models that are applied within single stands (Breidenbach and Astrup 2012). The model-based estimation approach can also be useful for surveys of remote areas, where remotely sensed data can be combined with a small sample of purposively selected field plots (McRoberts 2006; Ståhl et al. 2011; Corona et al. 2014a; McRoberts et al. 2014). Some key assumptions underlying model-based inference are (i) the values linked to population elements are random variables, (ii) since the individual values are random variables, so is the population total or mean that we wish to predict, (iii) a model for the relationship between the target variable and some auxiliary variable(s) exist, (iv) auxiliary data are available for all population elements, and (v) after having selected a sample – that need not be random – for estimating the model parameters, we apply the estimated model for predicting the target population quantity.

1.3. Hybrid inference

Auxiliary data may not always be available prior to a forest survey and it may be very expensive to collect for all units in a population, in order to fulfil the standard assumption for model-based inference. In such cases, a sample of auxiliary data can be acquired upon which the population total of the auxiliary variable is estimated based on design-based inference. A model is applied for the relationship between the study variable and the sampled auxiliary variables, and thus model-based inference can be applied once the auxiliary variable totals (or means) have been estimated through design-based inference. This approach was termed as a hybrid inference by Corona et al. (2014b). In a previous study by Mandallaz (2014) it was termed as pseudo-synthetic estimation, in the context of small-area estimation. Previous studies by Ståhl et al. (2011) and Ståhl et al. (2014) proposed the same approach, but did not suggest any specific nomenclature other than model-based inference. The basic approach in all these studies is that the expected value of an estimator is evaluated over both the model and the design. Likewise, the variance of the estimator is obtained through a conditioning approach, and typically includes one component due to the sampling error and one component due to the model error.

1.4. Remotely sensed data for forest surveys

Nowadays several kinds of RS data are available from almost all parts of the world. These include spectral satellite data with low, medium and high resolution (e.g., Hill et al. 1999; Hansen et al. 2008; Tomppo et al. 2008), radar satellite data such as InSAR (e.g., Næsset et al. 2011), Light Detection And Ranging (LiDAR) data from airborne profilers and scanners (e.g., Nelson et al. 1988, 1997; Næsset 1997; Hyyppä and Inkinen 1999), and traditional air photos which are becoming increasingly important due to novel uses of 3D point-cloud techniques (Leberl et al. 2010; Bohlin et al. 2012; Breidenbach and Astrup 2012). In principle, the usefulness of the different image sources

depends on what correlations can be obtained between the target forest variables and the different metrics that can be derived from the images, the availability of images, the acquisition costs, and the possibility to link the image metrics to some appropriate source of field data. Statistical linkages of remote sensing metrics with ground-truth field data typically are conducted using various types of parametric or non-parametric regression techniques (Tomppo et al. 2011), as well as different types of classification schemes such as logistic regression and discriminant analysis. Opsomer et al. (2007) and Baffetta et al. (2009) were first to introduce model-assisted estimation in the realm of coupling RS digital imagery data with field forest inventory data.

LiDAR data are known to provide auxiliary data that are highly correlated with growing stock volume, biomass and aboveground carbon in forests (Nelson et al. 1988; Næsset 1997; Hyyppä and Inkinen 1999; Hyyppä et al. 2008; Maltamo 2009; Næsset 2009, 2011; Gobakken 2012; Næsset et al. 2013b). In many applications, LiDAR data have been acquired wall-to-wall over the target forest areas, and stand-level estimates have been derived either based on the area method (Næsset, 2002) or based on the identification of individual trees (Hyyppä et al. 2001). For applications over large areas, such as countries, the acquisition of LiDAR data is prohibitively expensive; however, the data acquisition can be carried out as part of a sampling scheme to improve the precision of estimation. For example, Nelson et al. (2008) used a profiling LiDAR to estimate the forest resources of Delaware, and Andersen et al. (2009) used data from an airborne laser scanner to estimate forest resources within a region of Alaska.

Over the last decade several studies have been conducted where remotely sensed and field data have been combined in order to enhance the precision of large-scale field based inventories, or to make forest surveys feasible in remote areas where field sampling is very costly. Important studies of this kind include Nelson et al. (2009), McRoberts (2010b, 2011), Andersen (2009), Gregoire et al. (2011), Ståhl et al. (2011) and Næsset et al. (2013a).

In this thesis, two different sources of RS auxiliary information were analysed within design-based and model-based survey sampling frameworks. The two data sources were airborne LiDAR and Landsat 7 Enhanced Thematic Mapper Plus (ETM+) data.

2. OBJECTIVES

The overall objective of the studies was to evaluate how RS auxiliary data could be used to improve the precision of estimators in large-area forest inventories carried out through probability sampling of field plots. Two different inferential frameworks were evaluated and compared for growing stock volume estimation: model-based inference and design-based inference, mostly through model-assisted estimators.

The specific objectives of the different papers included in the thesis were to

I. introduce spatially balanced sampling (spreading in auxiliary space) in order to evaluate if and how much this design and the use of auxiliary RS data would improve the precision of estimation;
II. evaluate what improvements in precision of model-assisted estimators of growing stock volume could be obtained using different combinations of RS data sources as auxiliary information;
III. analyze the impacts of sample size and model form on the precision of model-based prediction with different kinds of RS auxiliary data;
IV. compare the performance of model-assisted estimation and model-based prediction in cases where RS and field data are geographically mismatched.

3. MATERIAL AND METHODS

In this chapter the study area and data are described, as well as the methods used in the different studies. In Paper II real data were used in the analyses, whereas in Papers I, III and IV simulated populations were used. The simulated populations were created using a multivariate probability distribution technique.

3.1. Kuortane study area

All four studies involved data acquired from the Kuortane study area (300 km^2). This area is located in western Finland, in the Southern Osthrobotnia region. It is covered mainly by middle aged boreal forest in the Suomenselkä watershed area. It is dominated by Scots pine (*Pinus sylvestris L.*) which covers over 80% of forest area, whereas Norway spruce (*Picea abies*) and deciduous trees, mainly birches, usually occur as admixtures. The landscape is dominated by pine forests growing on mineral soil, peatlands drained for forestry, open peatlands (mires), and agricultural fields at lower elevations. Terrain depressions are covered by lakes, the largest being Kuortanejärvi. In 2006, the area was chosen for a pilot research project studying LiDAR applications for forest inventories.

The entire area was tessellated into 16 m x 16 m grid cells. Only grid cells which were located in the land use class forest were used. Other cells were masked out using digital map data provided by the Natural Resources Institute Finland (LUKE) (Tomppo et al. 2008). Overall, the forested parts of Kuortane comprise about 818000 grid cells, summing up to 210 km^2 of forest. Figure 1 presents an overview of the Kuortane study area.

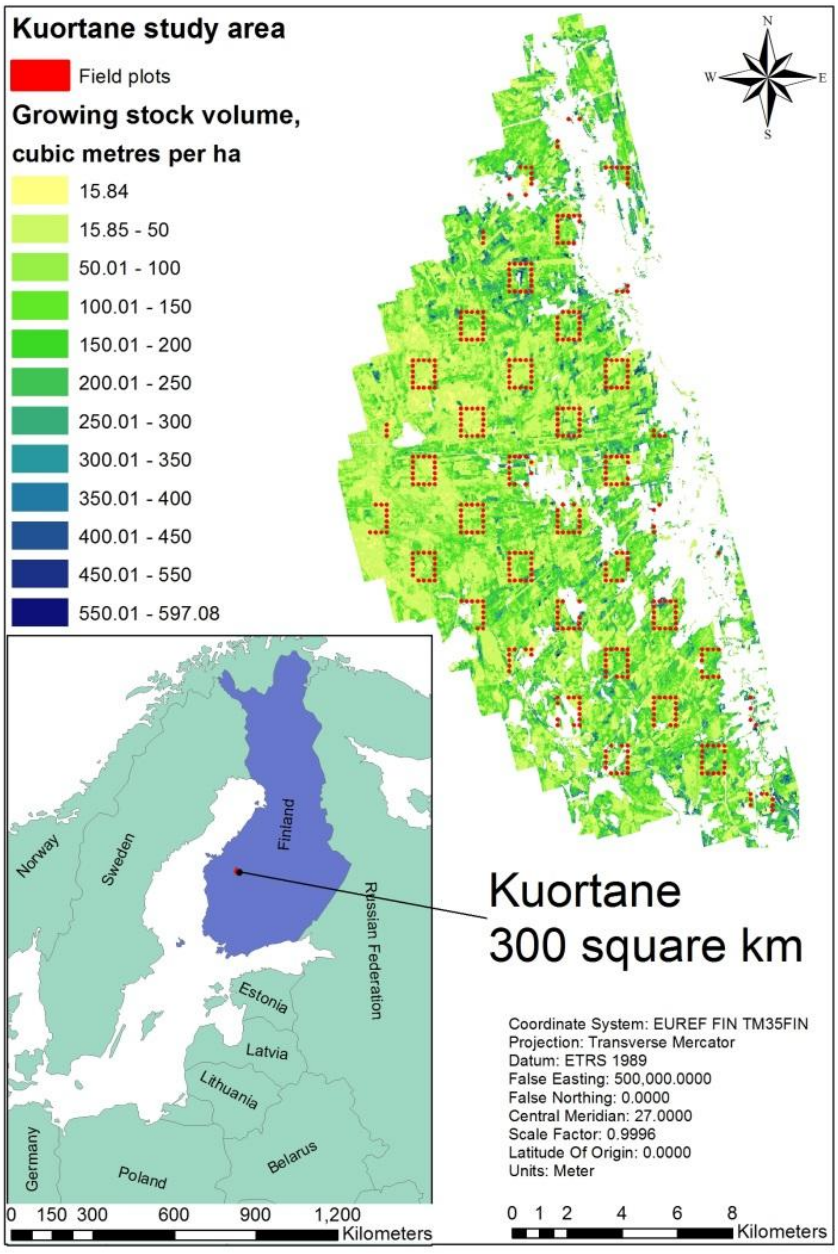

Figure 1: The location of the Kuortane study area (lower left), and details of the clusters of field plots and growing stock volume values of the simulated population developed for Papers III and IV (upper right).

Table 1: Overview of field data in the Kuortane study area (Paper II).

Variable	Scots pine		Norway spruce		Deciduous		All	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
DBH (cm)	11.4	9.1	4.4	7.3	4.2	6.2	-	-
Height (m)	9.0	7.2	3.8	6.2	4.4	6.2	-	-
Age (years)	32	32	13	24	15	23	-	-
Volume (m^3/ha)	60.6	70.3	13.7	45.0	8.0	23.2	82.3	91.5

3.1.1. Field data

Field data were collected using a modified version of the Finnish NFI measuring system. The Finnish NFI is based on a systematic cluster sampling design. L-shaped clusters of sample plots are located 7 km apart from each other in Central Finland, which includes the Kuortane region. But as the study area was rather small, the NFI sampling design was intensified for the purposes of this project. This was done by increasing the number of plots in a cluster to 18 to double the intensity of plots. The plots in a cluster were located along a rectangular tract, 300 m apart (see Figure 1). Overall, 39 clusters were laid out in the study area, and the total number of plots in the land use category forest was 441. Each plot was measured both as a truncated angle count sample plot and as a fixed area plot with a 9 m radius. In the studies included in this thesis, only the fixed area plots were used. The size of the grid cells was chosen to correspond to the size of the field plots (~255 m^2). An overview of the field data is provided in Table 1.

3.1.2. LiDAR data

The LiDAR data were collected on 28th July 2006 with an Optech 3100 laser scanning system operated at an altitude of 2000 m above ground level, using a half-angle of 15^o and a side overlap of about 20%. This resulted in a swath width of 1070 m. The divergence of the laser beam (1064 nm) was 0.3 mrad, which produced a footprint of 60 cm at ground level. Altogether 21 laser strips were measured, of which 2 were used for calibration purposes. The Optech 3100 laser scanning system produces four types of echoes (only, first, last, intermediate) which were re-classified into first and last pulse data. First returns were used for the digital surface model (DSM) creation and last returns for the digital elevation model (DEM), using the OPALS (Orientation and Processing of Airborne Laser scanning data) software (Kraus and Pfeifer 2001). The DEM was used to extract the point cloud of returns corresponding to the vegetation – the digital vegetation model (DVM). An upper threshold of 35 m height was used for the DVM (Lindberg et al. 2012), and no lower threshold was applied.

In this research, an area-based approach (Næsset 2002) was used. Twenty-six LiDAR metrics were extracted from the DVM for each grid cell and field plot using FUSION software (McGaughey 2012). The metrics were maximum height (h_{max}), minimum height, mean height (h_{mean}), standard deviation, variance, coefficient of variation, skewness, kurtosis (a measure of whether the data are peaked or flat relative to a normal distribution), P01, P05, P10, P20, P25, P30, P40, P50, P60, P70, P75, P80, P90, P95, P99 (heights at different percentiles of the DVM), canopy relief ratio (CRR), and percentage of first returns above 2 m (p_{veg}) as a crown cover estimate. For details see Table A2 in Appendix A of Paper II.

3.1.3. Landsat 7 ETM + data

The Landsat 7 ETM+ data were acquired in June 2006 (path 190 and row 16). The orthorectified (L1T) imagery data were downloaded from the U.S. Geological Survey server (accessed in 2011). Landsat 7 ETM+ has 8 bands of different spatial resolutions. For Paper I and II, bands 1 to 5 and 7 corresponding to blue, green, red, near infra-red (NIR), and two shortwave infra-red (SWIR) bands with a spatial resolution of 30 m were used. For Paper III and IV, only bands 2, 3 and 5 were used. The image was geo-referenced to the ETRS35-FIN metric coordinate system,

and the pixel size was re-sampled to 16 m × 16 m using the nearest neighbour re-sampling method in ArcGIS 10 (ESRI 2011). Spectral values were extracted for each 16 m × 16 m grid cell and for each circular field plot.

3.1.4. Simulated populations

To create simulated study populations resembling the Kuortane study area, the copula technique (Nelsen 2006) was applied. A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. It is a popular tool in actuarial sciences and recently it has been used in forestry applications for multivariate modelling of tree diameters, heights and volumes (Wang et al. 2008, 2010), stochastic modelling of regeneration (Miina and Heinonen 2008), simulation of forest stand structures (Kershaw et al. 2010), estimating shrub cover in riparian forests (Eskelson et al. 2011), improving inference based on nearest neighbour imputation (Ene et al. 2012b), and for generating ground-truth populations in simulation studies related to large-area LiDAR-based biomass surveys (Ene et al. 2012a, 2013). The approach applied in this thesis was based on the methods developed by Ene et al. (2012a).

The empirical dataset used for constructing the simulated population in Paper I contains plot-level information regarding the field-based volume estimates, a selection of LiDAR-derived variables (h_{mean}, CRR, the 20%, 50%, and 95% height percentiles from the DVM (called h_{20}, h_{50} and h_{95}), and the spectral information of the Landsat bands 1-5 and 7 (called B10-B50 and B70). For Papers III and IV the set consists of h_{max}, the height of the 80^{th} percentile of the DVM distribution (h_{80}), CRR, and p_{veg}, and Landsat spectral values of green (B20), red (B30) and shortwave infra-red (B50) bands. In Papers III and IV, only plots with non-zero values of growing stock volume were used for the creation of the simulated copula population. Furthermore, in Paper IV a buffer zone was created to ensure that each unit of the population had eight neighbours in order to handle geographical mismatches. The grid cells in the buffer were randomly sampled from the entire population of grid cells in the study area.

The vine copula estimation was performed using the "VineCopula" package (Schepsmeier et al. 2013) in R (R Development Core Team 2013). A copula population of 3000000 observations was created and then a sample of about 818000 observations was extracted using nearest neighbour imputation across the entire Kuortane study area. This sample has become the simulated population resembling the Kuortane study area conditions (Figure 1).

3.2. Statistical approaches

In this section the main statistical approaches used in the thesis are described. One important method applied is spatially balanced sampling, (Grafström and Lundström 2013) in which auxiliary data are used to ensure that a probability sample is well spread in the space of the auxiliary variables, whereby precise estimators are obtained due to the low variability between different feasible samples. Another method that is frequently applied in this thesis is model-assisted estimation, in which auxiliary data are applied in the estimation stage rather than the sampling stage. Model-assisted estimation relies on probability sampling, and thus the mode of inference is design-based (Gregoire 1998). Design-based inference typically assumes a finite population, with fixed quantities of interest (such as the volume of a tree) linked to each element, from which random samples are selected. Estimators of population parameters are random variables due to the probability-based inclusion of population elements into the sample (Särndal et al. 1992).

Contrary to design-based inference, model-based inference assumes the quantities of interest linked to population elements to be random variables (Gregoire 1998). Thus, target variables of interest such as population totals and means are also random variables. Model-based inference is thus founded on different assumptions than design-based sampling. In this thesis, model-based inference was applied in two of the studies; in one case it is combined with probability sampling of the auxiliary variables and thus the approach can be considered as a hybrid between model-based and design-based inference (Corona et al. 2014).

In the following sections, balanced sampling, model-assisted estimation and model-based prediction are described in more detail. In addition, it describes what regression models were applied in the model-assisted

estimation and model-based prediction, as well as how sampling simulation was applied to facilitate comparisons between different sampling strategies.

3.2.1. Balanced sampling

In this study the estimated characteristic was the population total of growing stock volume. Two estimators were used: a design-based unbiased Horvitz-Thompson (HT) estimator (Horvitz and Thompson 1952)

$$\hat{Y}_{HT} = \sum_{k \in S} \frac{y_k}{\pi_k} \tag{1}$$

and a nonparametric NN estimator

$$\hat{Y}_{NN} = \sum_{k \in S} N_k y_k \tag{2}$$

where N_k is a number of population units that are closer to the sample unit k than to any other sample unit. As usual, y_k is the variable of interest for the k^{th} sampled unit and π_k is the probability of inclusion of this unit. S is the set of elements in the sample. The approximate variance estimator for HT and NN estimators under a spatially balanced design where the target variable y is well approximated by a smooth function of the variables in which the sample is well spread:

$$\hat{V}_{LPM}\left(\hat{Y}_{HT/NN}\right) = \frac{1}{2}\sum_{k \in S}\left(\frac{y_k}{\pi_k} - \frac{y_{j_k}}{\pi_{j_k}}\right)^2 \tag{3}$$

where *LPM* denotes the local pivotal method, introduced by Grafström et al. (2012), $j_k \in S$ is the nearest neighbour to k in the space in which the design is spatially balanced. Spatially balanced samples in x provides an approximate balance on smooth functions $f(x)$, which means that $\sum_{k \in S} \frac{f(x_k)}{\pi_k} \cong \sum_{k \in U} f(x_k)$, where U is the set of elements from entire population. Thus if y_k is close to (x_k), then $\sum_{k \in S} \frac{y_k}{\pi_k} \cong \sum_{k \in U} y_k = Y$ for every sample(Grafström and Lundström 2013).

Three different auxiliary spaces for sample selection and NN estimation were utilized. The spaces are

1. Geographical coordinates;
2. Landsat spectral values;
3. LiDAR metrics.

The following sampling designs were compared:

- SRS, which is mostly a baseline against which other equal probability designs can be compared. For SRS, NN estimation is performed with the three auxiliary spaces.
- LPM, which is used with equal probabilities and the three different auxiliary spaces. LPM with auxiliary space 1 corresponds to equal probability samples that are well spread geographically.
- RSY_3^*, systematic πps sampling with initial randomization of the order of the units. This design is a baseline for unequal probabilities. Because the unequal probabilities are connected to auxiliary space 3, this design only uses that space for NN estimation (using the other spaces for NN estimation could potentially lead to a massive bias of the NN estimation).
- LPM_3^*, which is used with unequal probabilities and auxiliary space 3.

3.2.2. Model-assisted estimation

Model-assisted (MA) estimation was applied in Papers II, III and IV. It assumes a probability sample to be available from the target population. Further, auxiliary data are either available from all population units or from a sample of units. The general structure of a model-assisted estimator in case auxiliary data available from all population elements is:

$$\hat{Y}_{MA} = \sum_{k \in U} \hat{y}_k + \sum_{k \in S} \frac{y_k - \hat{y}_k}{\pi_k} \tag{4}$$

Thus, the estimator is composed of a sum of model-predictions ($\hat{y}$)for all population elements, plus a correction term which can be interpreted as an estimator of the population total of the deviations between the true element values and the corresponding predictions obtained using the model. This estimator can be shown to be unbiased in the case where an external model is used, or approximately unbiased if a model developed from the sample is applied (Särndal et al. 1992).

In Paper II, model-assisted estimation was applied. The population of grid cells is denoted by *U*. The total $Y = \sum_{k \in U} y_k$ of the growing stock volume was estimated, where y_k is the true value of growing stock volume for unit *k*. The first phase sample was a sample of *n* out of *N* strips, denoted as S_a. Thus, S_a contains all grid cells in the *n* selected strips. The second phase sample *S* of field plots corresponded to sampled grid cells within selected strips. Thus, the second phase sample is seen as a subset of the strip sample ($S \subset S_a$).

Five cases were evaluated:

A. Estimation based on the field plots belonging to the sample *S*. To enable a comparison with the strip sampling designs, field plots from within sampled strips were utilised. The population total $Y = \sum_{k \in U} y_k$ of the growing stock volume for a finite population *U* of grid cells was estimated (Särndal et al. 1992, Eq. 9.3.5, p. 348) by:

$$\hat{Y}_A = \sum_{k \in S} \frac{y_k}{\pi_k^*} \tag{5}$$

where *A* denotes "Case A", π_k^* is the probability of inclusion obtained by a conditioning approach; the corresponding variance estimator (Särndal et al. 1992, Eq. 9.3.7, p. 348) is:

$$\hat{V}(\hat{Y}_A) = \sum\sum_{k,l \in S} \frac{\Delta_{akl}}{\pi_{kl}^*} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum\sum_{k,l \in S} \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}} \frac{y_k}{\pi_k^*} \frac{y_l}{\pi_l^*} \tag{6}$$

where $\Delta_{akl} = \pi_{akl} - \pi_{ak}\pi_{al}$, $\Delta_{kl|S_a} = \pi_{kl|S_a} - \pi_{k|S_a}\pi_{l|S_a}$and $\pi_{kl}^* = \pi_{akl}\pi_{kl|S_a}$ (see Särndal et al. 1992, Eq. 9.3.7, p. 348). The Δ:s are known as the covariances of the inclusion indicators. Details can be found in Table 2.

B. Two-phase model-assisted estimation with data from LiDAR strips as the first phase sample S_a, and field plot data as the second phase sample *S*. Only field plots within sampled LiDAR strips were utilised in the estimation. For this case, the population total and its variance estimator are estimated by (Särndal et al. 1992, Eq. 9.6.13 and Eq. 9.6.16, p. 358), where *B* denotes "Case B":

$$\hat{Y}_B = \sum_{k \in S_a} \frac{\hat{y}_k}{\pi_{ak}} + \sum_{k \in S} \frac{y_k - \hat{y}_k}{\pi_k^*} \tag{7}$$

$$\hat{V}(\hat{Y}_B) = \sum\sum_{k,l\in S} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k}{\pi_{ak}} \frac{y_l}{\pi_{al}} + \sum\sum_{k,l\in S} \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}} \frac{y_k - \hat{y}_k}{\pi^*_k} \frac{y_l - \hat{y}_l}{\pi^*_l}$$

(8)

C. Model-assisted estimation with full cover Landsat data U and field plots (sample S). In order to make straightforward comparisons with the other alternatives, the field plots were selected only from within strips corresponding to the LiDAR samples. For this case, the population total estimator and the corresponding variance estimator (Särndal et al. 1992, Eq. 9.6.12, p. 358) are; where *C* denotes "Case C"

$$\hat{Y}_C = \sum_{k\in U} \hat{y}_{1k} + \sum_{k\in S} \frac{y_k - \hat{y}_{1k}}{\pi^*_k}$$

(9)

$$\hat{V}(\hat{Y}_C) = \sum\sum_{k,l\in S} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k - \hat{y}_{1k}}{\pi_{ak}} \frac{y_l - \hat{y}_{1l}}{\pi_{al}} + \sum\sum_{k,l\in S} \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}} \frac{y_k - \hat{y}_{1k}}{\pi^*_k} \frac{y_l - \hat{y}_{1l}}{\pi^*_l}$$

(10)

D. Two-phase model-assisted estimation, with full cover U of Landsat data, using a LiDAR strip sample S_a as the first phase of sampling, and field plot data as the second phase sample S. Population total estimator and the corresponding variance estimator (Särndal et al. 1992, Eq. 9.6.8 and Eq. 9.6.10, p. 357) are; where *D* denotes "Case D"

$$\hat{Y}_D = \sum_{k\in U} \hat{y}_{1k} + \sum_{k\in S_a} \frac{\hat{y}_k - \hat{y}_{1k}}{\pi_{ak}} + \sum_{k\in S} \frac{y_k - \hat{y}_k}{\pi^*_k}$$

(11)

$$\hat{V}(\hat{Y}_D) = \sum\sum_{k,l\in S} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k - \hat{y}_{1k}}{\pi_{ak}} \frac{y_l - \hat{y}_{1l}}{\pi_{al}} + \sum\sum_{k,l\in S} \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}} \frac{y_k - \hat{y}_k}{\pi^*_k} \frac{y_l - \hat{y}_l}{\pi^*_l}$$

(12)

E. Model-assisted (MA) estimation with full cover U of Landsat and LiDAR data, and field plots (sample S). As for the other cases, only the field plots within sample strips were utilized in the estimation. For this case, the population total and its variance estimators are estimated by (Särndal et al. 1992, Eq. 9.6.12, p. 358); where *E* denotes "Case E"

$$\hat{Y}_E = \sum_{k\in U} \hat{y}_k + \sum_{k\in S} \frac{y_k - \hat{y}_k}{\pi^*_k}$$

(13)

$$\hat{V}(\hat{Y}_E) = \sum\sum_{k,l\in S} \frac{\Delta_{akl}}{\pi^*_{kl}} \frac{y_k - \hat{y}_k}{\pi_{ak}} \frac{y_l - \hat{y}_l}{\pi_{al}} + \sum\sum_{k,l\in S} \frac{\Delta_{kl|S_a}}{\pi_{kl|S_a}} \frac{y_k - \hat{y}_k}{\pi^*_k} \frac{y_l - \hat{y}_l}{\pi^*_l}$$

(14)

In addition to the five cases, the performance of simple random sampling without replacement of strips (SI,SI) was evaluated with a case where strips were selected proportionally to their size (πps,SI). Details of the inclusion probabilities can be found in Table 2.

Table 2: First and second order probabilities of inclusion and covariances of inclusion indicators for the two design approaches, with simple random sampling in the first phase of selecting strips and simple random sampling of plots beneath the strips (SI,SI), and probability-proportional-to-size sampling in the first phase and simple random sampling in the second (πps,SI) (Paper II).

Term	SI,SI	πps,SI	Description
π_{ak}	$\frac{n}{N}$	$\frac{n\sum_{j\in U(k)}\hat{y}_{1j}}{\sum_{i\in U}\hat{y}_{1i}}$	The probability that unit *k* included in S_a.
$\pi_{(k\mid S_a)}$	$\frac{\sum_{S_a} m_i}{\sum_{S_a} M_i}$	$\frac{\sum_{S_a} m_i}{\sum_{S_a} M_i}$	The conditional probability that unit *k* included in S, given that $k \in S_a$.
π_{akl}	$\begin{cases}\pi_{ak}, & if\ in\ same\ strip,\\ \frac{n}{N}\frac{n-1}{N-1}, & otherwise.\end{cases}$	$\begin{cases}\pi_{ak}, & if\ in\ same\ strip,\\ Appendix\ B\ in\ II, & otherwise.\end{cases}$	The second order inclusion probability of units *k* and *l* included in S_a.
$\pi_{kl\mid S_a}$	$\begin{cases}\pi_{k\mid S_a}, & if\ in\ same\ strip,\\ \frac{\sum_{S_a} m_i}{\sum_{S_a} M_i}\frac{\sum_{S_a} m_i - 1}{\sum_{S_a} M_i - 1}, & otherwise.\end{cases}$	$\begin{cases}\pi_{k\mid S_a}, & if\ in\ same\ strip,\\ \frac{\sum_{S_a} m_i}{\sum_{S_a} M_i}\frac{\sum_{S_a} m_i - 1}{\sum_{S_a} M_i - 1}, & otherwise.\end{cases}$	The conditional second order inclusion probability of units *k* and *l* included in S, given that $k, l \in S_a$.
π_k^*	$\pi_{ak}\pi_{k\mid S_a}$	$\pi_{ak}\pi_{k\mid S_a}$	The probability that unit *k* included in S.
π_{kl}^*	$\pi_{akl}\pi_{kl\mid S_a}$	$\pi_{akl}\pi_{kl\mid S_a}$	The second order probability of units *k* and *l* included in S.
$\Delta_{kl\mid S_a}$	$\pi_{kl\mid S_a} - \pi_{k\mid S_a}\pi_{l\mid S_a}$	$\pi_{kl\mid S_a} - \pi_{k\mid S_a}\pi_{l\mid S_a}$	The covariance of inclusion indicator conditional to S_a.
Δ_{akl}	$\pi_{akl} - \pi_{ak}\pi_{al}$	$\pi_{akl} - \pi_{ak}\pi_{al}$	The covariance of inclusion indicator.

In Paper III, a model-assisted estimator was applied as a baseline method for comparison with the model-based prediction. In this case a ratio estimator was applied (Särndal et al. 1992, p. 327):

$$\hat{\mu}_{MA} = \frac{\sum_{i=1}^{n}\hat{G}_{MA_i}}{\sum_{i=1}^{n} M_i} \tag{15}$$

where

$$\hat{G}_{MA_i} = \sum_{t=1}^{M_i}\hat{y}_t + \sum_{t=1}^{m_i}\frac{y_t - \hat{y}_t}{\pi_{t\mid i}} \tag{16}$$

where $\pi_{t\mid i}$ is the conditional probability that grid cell *t* is included in the second phase sample given that the i^{th} strip is included, m_i is the second phase sample size within i^{th} strip, and M_i is the total number of grid cells in strip *i*.

In Paper IV, model-assisted estimation (Eq. 17) following simple random sampling of population elements and complete cover of auxiliary data was applied to evaluate the effects in case data with positional errors was applied (Särndal et al. 1992, Eq. 6.3.4 and Eq. 6.3.6, p. 222-223):

$$\hat{\mu}_{MA} = \frac{1}{M}\left[\sum_{k\in U}\hat{y}_k + \frac{M}{m}\sum_{k\in S}(y_k - \hat{y}_k)\right] \tag{17}$$

$$\hat{V}(\hat{\mu}_{MA}) = \left(\frac{M-m}{M}\right)\frac{\sum_{k\in S}(y_k - \hat{y}_k)^2}{m(m-1)} \tag{18}$$

where *M* is a total number of grid cells in population *U*, *m* is the selected number of grid cells in sample *S*, and *MA* denotes "model-assisted". There are no selected strips in the sampling design, hence no sample S_a.

3.2.3. Model-based prediction

Model-based prediction was applied in Papers III and IV. In Paper III, a hybrid between model-based and design-based inference was applied, since the auxiliary data were assumed to be collected through probability sampling. The estimator for the population mean value μ was:

$$\hat{\mu}_{MB} = \frac{\sum_{i=1}^{n} \hat{G}_{MB_i}}{\sum_{i=1}^{n} M_i} \tag{19}$$

where *MB* denotes "model-based", strip totals $\hat{G}_{MB_i} = \sum_{t=1}^{M_i} g(x_t, \hat{\beta}) = \sum_{t=1}^{M_i} \hat{y}_t$, and $g(x_t, \hat{\beta})$ is a function to predict *y* for t^{th} grid cell. This ratio estimator was applied since strips varied considerably in size. The variance of $\hat{\mu}_{MB}$ can be estimated as (Ståhl et al. 2011, Eq. 15, p. 101):

$$\hat{V}(\hat{\mu}_{MB}) \approx \frac{1}{\bar{M}^2}\left[\frac{(N-n)}{N}\frac{\sum_{i=1}^{n}\left(\hat{G}_{MB_i} - \hat{\mu}_{MB}M_i\right)^2}{n(n-1)} + \sum_{k=1}^{p+1}\sum_{l=1}^{p+1}\widehat{Cov}(\hat{\beta}_k, \hat{\beta}_l)\,\hat{\bar{G}}_k'\hat{\bar{G}}_l'\right] \tag{20}$$

where *n* is a number of selected strips out of *N* as the first phase sample, $\bar{M}$ is the first phase sample mean of the M_i, (p+1) is the number of model parameters (including the constant), $\widehat{Cov}(\hat{\beta}_k, \hat{\beta}_l)$ is the estimated covariance between the parameter estimates $\hat{\beta}_k$ and $\hat{\beta}_l$, $\hat{\bar{G}}_k' = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{M_i} g_k'(x_{it}, \hat{\beta})$, and $g_k'(x_{it}, \hat{\beta})$ is the derivative of $g(x_{it}, \hat{\beta})$ with respect to the k^{th} model parameter. As can be seen, the covariances of the model parameter estimates play an important role in the model-based variance estimator.

Four different estimators of the covariance matrix were investigated in Paper III, the ordinary and nonlinear least squares (OLS and NLS) and the heteroskedasticity-consistent for linear and nonlinear regression (HC and NHC). The least squares (OLS and NLS) covariance matrix estimator is:

$$\widehat{Cov}(\hat{\beta}) = \hat{\sigma}^2(\boldsymbol{X}(\beta)^T\boldsymbol{X}(\beta))^{-1} \tag{21}$$

where $\hat{\sigma}^2 = \frac{\hat{e}^T\hat{e}}{m-(p+1)}$ is the residual variance ($\hat{e}^T\hat{e}$ is a sum of squared estimated residuals $\hat{e}$), and $\boldsymbol{X}(\beta)$ is a matrix of dimension *m* times (p+1) of partial derivatives with respect to the model parameter. It can be seen that if a model is of a linear form, then the matrix of partial derivatives $\boldsymbol{X}(\beta)$ becomes a matrix of predictors $\boldsymbol{X}$ (including a column of 1 for intercept).

The HC and NHC covariance matrices are estimated as (White 1980):

$$\widehat{Cov}(\hat{\beta}) = (\boldsymbol{X}(\beta)^T\boldsymbol{X}(\beta))^{-1}\left[\sum_{t=1}^{m}\hat{e}_t^2 X(\beta)_t X(\beta)_t^T\right](\boldsymbol{X}(\beta)^T\boldsymbol{X}(\beta))^{-1} \tag{22}$$

In Paper IV, the target parameter was the population mean μ, of growing stock volume. The estimators used for the model-based inference were:

$$\hat{\mu}_{MB} = \frac{1}{M}\sum_{i\in U}\hat{y}_i \tag{23}$$

$$\hat{V}(\hat{\mu}_{MB}) = \frac{1}{M^2}\sum_{k=1}^{p+1}\sum_{l=1}^{p+1}\widehat{Cov}(\hat{\beta}_k,\hat{\beta}_l)\,\hat{g}'_k\hat{g}'_l \tag{24}$$

where M is the total number of grid cells in population U and $\hat{g}'_k$ is estimated as

$$\hat{g}'_k = \sum_{i\in U}\frac{\partial g(X_i,\hat{\boldsymbol{\beta}})}{\partial\hat{\beta}_k} \tag{25}$$

Five cases of positional errors were evaluated in Paper IV:

A. Perfect positions (denoted perfect), where reference plot data and RS auxiliary data always coincided perfectly, i.e. were taken from the same grid cells.
B. Fair positions (denoted fair), in which case for 50 % of the positions – for randomly selected plots – coincided perfectly. For the remaining 50 % of plots, RS data were retrieved from a neighbouring grid cell, selected in a random direction (including diagonal grid cells).
C. Poor positions (denoted poor), in which case for 100 % of the positions, RS data were retrieved from a neighbouring grid-cell, selected in a random direction (including diagonal grid cells).
D. Fair positions with perfect models (denoted semi fair); models from Case A were in this case applied to RS data from grid cells of fair positioning quality.
E. Poor positions with perfect models (denoted semi poor); models developed in Case A were in this case applied to RS data from grid cells of poor positioning quality.

3.2.4. Regression models

Regression modelling is an important basis for both model-based prediction and model-assisted estimation. In this section the main model types used in the studies are described. The following regression model forms were used in the studies:

1. The linear regression model, denoted “LINEAR”

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{26}$$

2. The multiplicative regression model

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} \ldots e^{\epsilon} \tag{27}$$

which can be linearized by taking the natural logarithm of both sides (denoted “LOG-LOG”)

$$\ln y = \beta_o + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \cdots + \beta_p \ln x_p + \epsilon \tag{28}$$

3. The model of the following type

$$y = \left(\beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon\right)^2 \tag{29}$$

which is linearized through a square root transformation of the vector of responses (denoted "SQRT")

$$\sqrt{y} = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \tag{30}$$

4. Nonparametric regression denoted "NN".

In Papers II and IV the "SQRT" model form was applied, in Paper III the models "LINEAR", "LOG-LOG" and "SQRT" were investigated, and in Paper I the "NN" model form was applied.

3.2.5. Sampling simulation

In studies I, III and IV, Monte Carlo sampling simulation with 10000 repetitions was performed, but in study II only 1000 iterations were applied. Based on the outcome of the simulations, the empirical variance was estimated as:

$$V(\hat{\theta})_{emp} = \frac{1}{R-1}\sum_{i=1}^{R}\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2 \tag{31}$$

Where R is the number of repetitions, $\hat{\theta}$ is an arbitrary estimator, $\hat{\theta}_i$ is the estimated value after iteration i, $\bar{\hat{\theta}}$ is the mean value of estimated values over all R repetitions. This variance was taken as the "true" variance of an estimator and was also compared with the average of the variance estimators.

The bias of the estimator was estimated as:

$$BIAS = \frac{1}{R}\sum_{i=1}^{R}\left(\hat{\theta}_i - \theta\right) \tag{32}$$

where θ is the true parameter value, which was known only in Papers I, III and IV, where simulated populations were used; the relative bias is:

$$RBIAS = 100\%\frac{BIAS}{\theta} \tag{33}$$

The relative bias of the variance estimator was estimated as:

$$RBIAS = 100\%\frac{\hat{V}(\hat{\theta}) - V(\hat{\theta})_{emp}}{V(\hat{\theta})_{emp}} \tag{34}$$

In Paper III the difference between model-based and model-assisted empirical variances was estimated as:

$$Relative\ Difference = 100\% \frac{V(\hat{\theta}_{MB})_{emp} - V(\hat{\theta}_{MA})_{emp}}{V(\hat{\theta}_{MA})_{emp}} \quad (35)$$

The relative standard error was estimated as:

$$RSE = 100\% \frac{\sqrt{\hat{V}(\hat{\theta})}}{\theta} \quad (36)$$

In Paper I the empirical relative root mean square error is given by:

$$RRMSE = \frac{\sqrt{\frac{1}{R}\sum_{i=1}^{R}(\hat{\theta}_i - \theta)^2}}{\theta} \quad (37)$$

4. **RESULTS**

In Paper I, the results show that spreading the sample geographically didn't improve the precision of estimation. Interestingly, it can be seen that for the HT estimator, the local pivotal method (LPM_3) used in the auxiliary space of LiDAR metrics performed much better than simple random sampling without replacement (SRS), e.g. LPM_3 with a sample size of 100 performed better than SRS with a sample size of 480. It would appear that great savings on improved precision by selecting a field sample with a more sophisticated design can be achieved. From Figure 2, it can be seen that the design effects for the HT estimator are rather stable and the same order of the designs for all sample sizes is achieved. All of the designs seem to show the general pattern that if the sample size is increased by a factor of 4, then the RRMSE decreases by about a factor of 2. There is a similar pattern for the NN estimator, but the order of designs change. The NN estimator becomes more efficient with unequal probabilities compared with equal probabilities when the sample size increases, and this effect comes at a smaller sample size for LPM_3^* versus LPM_3, than for RSY_3^* versus SRS_3.

In Paper II, the standard errors (the square root of the estimated variances) of the growing stock volumes were found to decrease with increasing LiDAR sample size (Figure 3). Furthermore, the introduction of additional auxiliary data clearly improved the precision of the estimators, i.e. adding Landsat wall-to-wall data improved the precision of the LiDAR strip-based model-assisted estimators (compare case B to case D in Figure 3). The designs with probability-proportional-to-size sampling in the first phase always resulted in a lower relative standard error than their counterparts, based on simple random sampling. As expected, case E – the combination of all available auxiliary data – was the most precise strategy. Studying the SI,SI design approach, the results show how the precision of estimates increases from Case A to Case E, i.e. when additional auxiliary information is added. However, the standard errors are fairly large. For a given case and sample size, the design approach πps,SI considerably increased the precision when compared to SI,SI. However, the trend with increased precision from Case A to Case E is not as clear for πps,SI, and LiDAR sample data performed better than wall-to-wall Landsat data in this case. Very high precision was attained for the πps,SI two-phase sampling strategy in moderate and large LiDAR strip sample sizes.

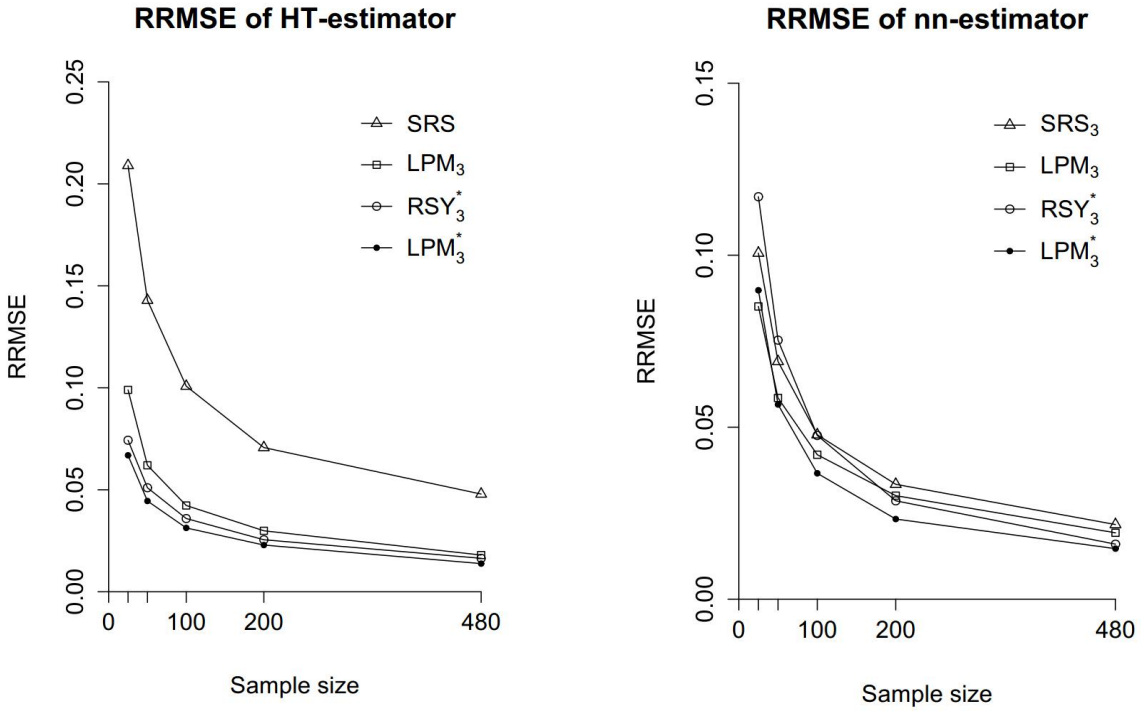

Figure 2: RMSE of HT and NN estimators for sample sizes 25, 50, 100, 200, and 480 under the different sampling designs (Paper I).

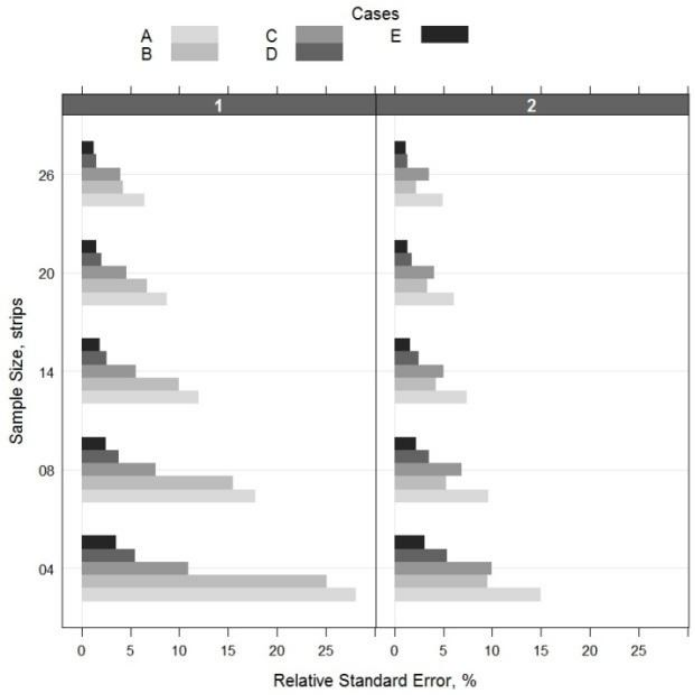

Figure 3: Relative standard errors for the different sampling strategies in Paper II. A – design-based estimation based on field plots only; B – two-phase model-assisted estimation with data from LiDAR strips as the first phase and field plot data as the second phase; C – model-assisted estimation with wall-to-wall data and field plots; D – two-phase model-assisted estimation with wall-to-wall Landsat data, a LiDAR strip sample as the first phase of sampling, and field plot data as the second phase; E – model-assisted estimation with wall-to-wall Landsat and LiDAR data and field plots (1- SI,SI, and 2- πps,SI).

Table 3: Estimated growing stock volume and model-based variance for Landsat models when using wall-to-wall data, i.e. *n=N.* The subscripts at $\hat{V}(\hat{\mu}_{MB})$ indicate the type of covariance matrix estimator applied with Eq. (20); OLS – ordinary least square, NLS – nonlinear least square, HC – heteroskedasticity-consistent, and NHC – nonlinear heteroskedasticity-consistent (Paper III).

Plots	$\hat{\mu}_{MB}$, m^3ha^{-1}	$\hat{V}(\hat{\mu}_{MB})_{OLS}$		$\hat{V}(\hat{\mu}_{MB})_{NLS}$		$\hat{V}(\hat{\mu}_{MB})_{HC}$		$\hat{V}(\hat{\mu}_{MB})_{NHC}$		$V(\hat{\mu}_{MB})_{emp}$	RSE, %
		est.	diff.	est.	diff.	est.	diff.	est.	diff.		
						"LINEAR"					
15	96.00	544.54	37.94	-	-	376.29	206.19	-	-	582.48	23.90
25	98.20	280.36	11.41	-	-	229.35	62.43	-	-	291.77	16.92
75	100.44	86.50	0.77	-	-	81.88	5.39	-	-	87.26	9.25
250	100.79	25.15	0.67	-	-	24.77	0.29	-	-	24.48	4.90
1000	100.96	6.25	0.13	-	-	6.23	0.10	-	-	6.13	2.45
						"LOG-LOG"					
15	131.18	1.21×10^{08}	1.17×10^{08}	1.28×10^{08}	1.24×10^{08}	5.92×10^{07}	5.49×10^{07}	8.76×10^{07}	8.33×10^{07}	4.27×10^{06}	2.05×10^{05}
25	101.86	433.87	150.55	292.58	9.26	335.00	51.69	299.76	16.44	283.32	16.67
75	101.23	108.11	33.44	72.20	2.47	89.99	15.32	71.04	3.63	74.67	8.56
250	101.02	30.55	9.71	20.60	0.24	26.67	5.83	20.65	0.19	20.84	4.52
1000	101.00	7.54	2.33	5.09	0.11	6.68	1.47	5.15	0.06	5.21	2.26
						"SQRT"					
15	102.27	422.51	68.22	503.32	12.60	334.97	155.76	427.68	63.04	490.73	21.94
25	100.77	202.83	48.52	250.35	1.01	182.12	69.24	226.54	24.82	251.36	15.70
75	100.99	61.78	18.32	79.54	0.55	59.12	20.98	74.64	5.46	80.10	8.86
250	100.93	17.95	5.00	23.31	0.36	17.48	5.47	22.06	0.88	22.95	4.74
1000	100.98	4.46	1.34	5.81	0.00	4.36	1.44	5.50	0.30	5.80	2.39

In Paper III, the performances of the OLS, NLS, NC and NHC covariance matrix estimators in the model-based variance estimator Eq. (20) with wall-to-wall first phase data were compared. For each estimated variance, an absolute difference from the corresponding empirical variance was estimated and is presented in Table 3 and Table 4 (denoted as "diff"). Similar magnitudes of variance were obtained, but the NLS estimator resulted in the smallest variance. LiDAR models always led to more precise results. Comparing the performances of different models, the square-root transformed models resulted in the smallest variances. Regarding the comparison of the model-based and model-assisted estimators' performance [Eq. (19) and Eq. (15)], their empirical variances estimated by Eq. (31) were compared. Figure 4 shows the relative difference estimated by Eq. (35) as a function of strip sample size for "SQRT" models.

In Paper IV, substantial differences in the effect of positional errors can be observed between the use of LiDAR and Landsat auxiliary data. With LiDAR data, the difference between using perfectly and poorly geo-located data is very large, whilst this is not the case when Landsat data are applied (Figure 5). Comparing the two inferential approaches, model-based estimation often resulted in higher precision than model-assisted estimation. Figure 6 shows the relative bias obtained in the different cases. In terms of comparison of estimated variances, it can be seen from Figure 7 that the variance estimators in the case of the model-based prediction were almost always substantially less biased than the variance estimators for model-assisted estimation.

Table 4: Estimated growing stock volume and model-based variance for LiDAR models when using wall-to-wall data, i.e. *n=N.* The subscripts at $\hat{V}(\hat{\mu}_{MB})$ indicate the type of covariance matrix estimator applied with Eq. (20); OLS – ordinary least square, NLS – nonlinear least square, HC – heteroskedasticity-consistent, and NHC – nonlinear heteroskedasticity-consistent (Paper III).

Plots	$\hat{\mu}_{MB}$, m^3ha^{-1}	$\hat{V}(\hat{\mu}_{MB})_{OLS}$		$\hat{V}(\hat{\mu}_{MB})_{NLS}$		$\hat{V}(\hat{\mu}_{MB})_{HC}$		$\hat{V}(\hat{\mu}_{MB})_{NHC}$		$V(\hat{\mu}_{MB})_{emp}$	RSE, %
		est.	diff.	est.	diff.	est.	diff.	est.	diff.		
						"LINEAR"					
15	101.37	78.52	28.18	-	-	42.21	64.50	-	-	106.70	10.23
25	101.15	37.89	9.44	-	-	28.72	18.60	-	-	47.33	6.81
75	101.14	11.52	1.07	-	-	10.84	1.75	-	-	12.59	3.51
250	100.99	3.38	0.08	-	-	3.34	0.12	-	-	3.46	1.84
1000	101.01	0.85	0.00	-	-	0.85	0.01	-	-	0.85	0.91
						"LOG-LOG"					
15	100.62	6.65×10^{10}	6.65×10^{10}	3.49×10^{05}	3.49×10^{05}	9.30×10^{10}	9.30×10^{10}	7.98×10^{03}	7.88×10^{03}	101.05	9.96
25	100.51	4.18×10^{04}	4.17×10^{04}	62.75	13.74	3.57×10^{04}	3.56×10^{04}	54.91	5.89	49.02	6.93
75	100.89	14.70	0.28	14.27	0.70	14.81	0.16	14.29	0.68	14.97	3.83
250	100.92	3.19	1.24	4.06	0.37	3.26	1.16	4.05	0.38	4.42	2.08
1000	100.99	0.74	0.36	1.01	0.10	0.77	0.34	1.00	0.11	1.11	1.04
						"SQRT"					
15	100.90	48.18	32.26	72.42	8.01	39.51	40.93	49.91	30.53	80.44	8.88
25	100.75	24.19	17.14	37.17	4.16	27.63	13.70	31.78	9.55	41.33	6.37
75	100.95	7.29	5.16	11.94	0.51	10.75	1.69	11.03	1.42	12.45	3.49
250	100.94	2.12	1.53	3.53	0.12	3.38	0.27	3.29	0.37	3.65	1.89
1000	100.99	0.53	0.38	0.89	0.03	0.86	0.05	0.82	0.09	0.91	0.95

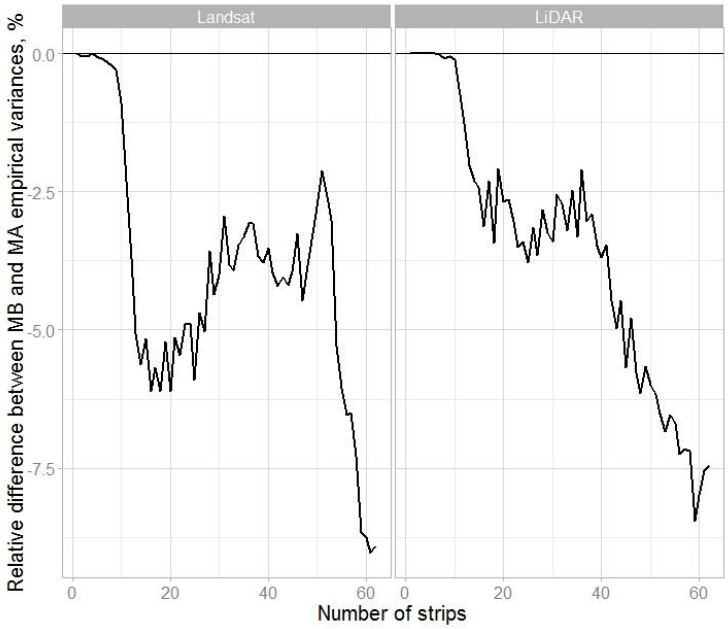

Figure 4: Relative difference between empirical model-based and model-assisted variances as a function of strip sample size for 'SQRT' models (Paper III).

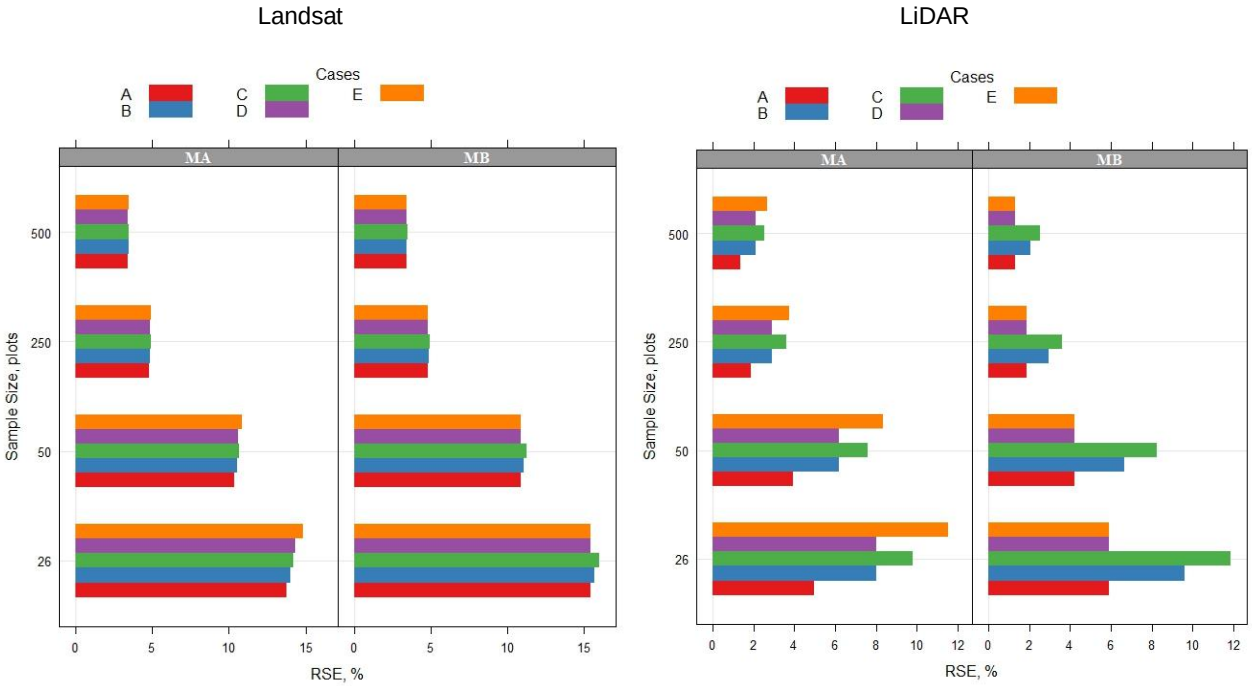

Figure 5: The precision (relative standard error) of the estimators in model-based and model-assisted estimation for different cases of positional errors, sources of auxiliary data and sample size. A – perfect positions; B – fair positions; C – poor positions; D – fair positions with perfect models (semi fair); E – poor positions with perfect models (semi poor) (Paper IV).

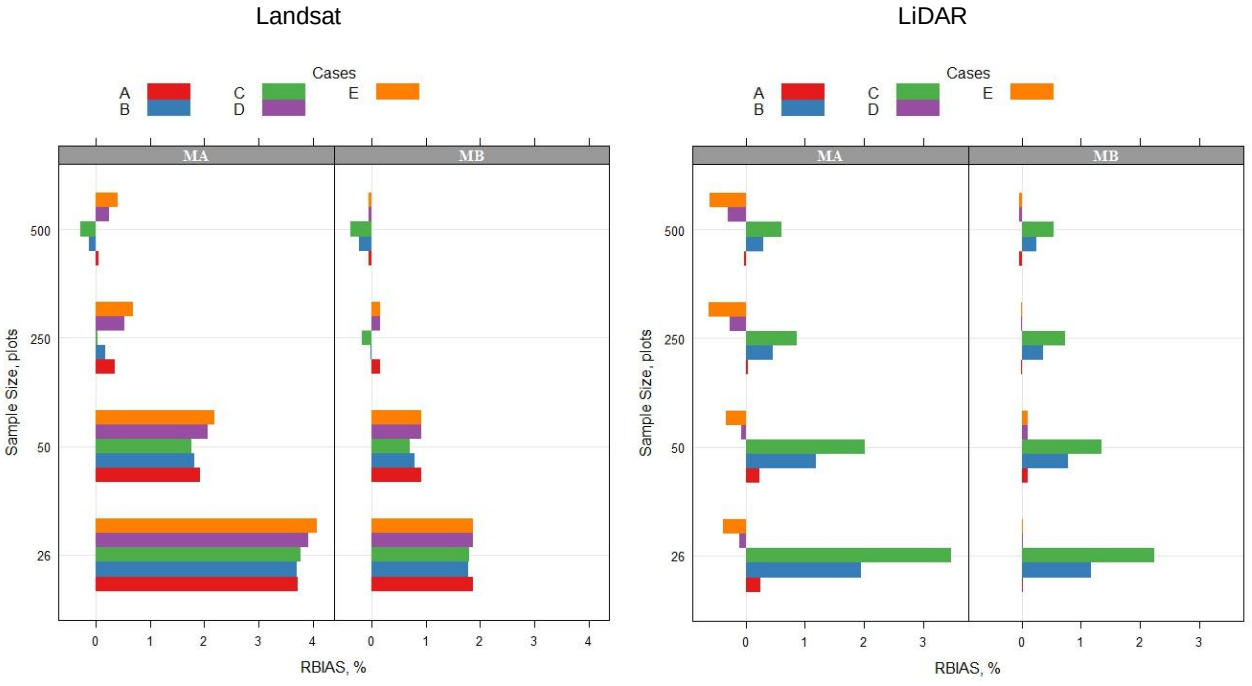

Figure 6: The relative bias of estimators in model-assisted and model-based estimation for different cases of positional errors, sources of auxiliary data and sample sizes; A – perfect positions; B – fair positions; C – poor positions; D – fair positions with perfect models (semi fair); E – poor positions with perfect models (semi poor) (Paper IV).

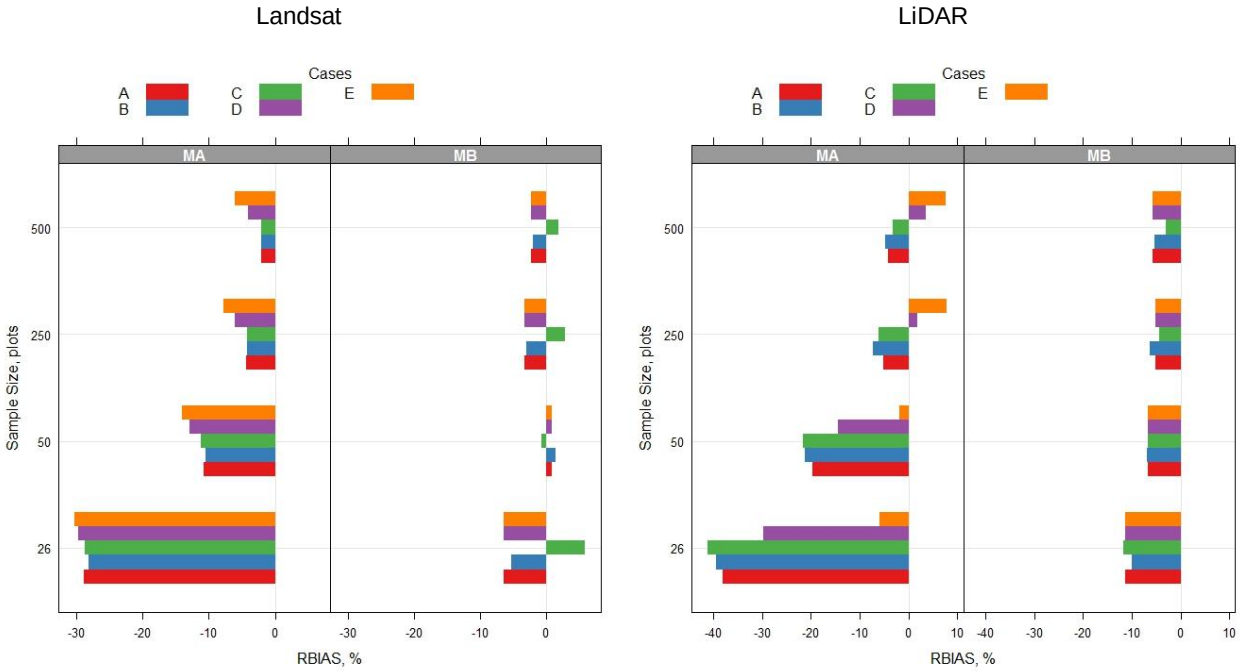

Figure 7: The relative bias of estimated variances; A – perfect positions; B – fair positions; C – poor positions; D – fair positions with perfect models (semi fair); E – poor positions with perfect models (semi poor) (Paper IV).

5. DISCUSSION

This thesis investigates the usefulness of RS auxiliary data in a series of design-based sampling studies. In Paper I, it was shown that the use of RS data improved the precision of estimators in a newly developed balanced sampling design (Grafström et al. 2012; Grafström and Lundström 2013). Through this design, the samples will always be well spread in the space of auxiliary data. Also, when the auxiliary data – as in the case of Paper I – are well correlated with the target variable, different samples will yield similar estimates and thus the estimator has high precision. Similar results have been reported by Grafström and Ringvall (2013) for a design resembling that studied in Paper I.

In Paper II it was shown that Landsat and LiDAR auxiliary data improved the precision of model-assisted estimation compared to using field data alone. Use of LiDAR data lead to the largest increase in precision, however, use of Landsat data in addition to LiDAR data lead to further improvements of the precision. Similar results using LiDAR auxiliary data have been reported by Ene et al. (2012a) from a sampling simulation study based on data obtained from Hedmark in Norway, and also by Gobakken et al. (2012). A special feature of Paper II was the use of probability-proportional-to-size sampling of LiDAR strips as a means to address the problem of variable flight-line length, which inflates the variance of estimators in cases where simple random sampling is applied. As expected, higher precision was obtained (Paper II) when strips were selected by a probability-proportional-to-size procedure, using Landsat data for a preliminary estimate of growing stock volume in each strip. An alternative way to handle the issue of variable strip length could have been to apply a ratio estimator (Särndal et al. 1992).

In Paper III, as expected, using LiDAR data in the model-based prediction of growing stock volume lead to more precise results than using Landsat data. The reason for this is that LiDAR data are strongly related to variables such as growing stock volume and biomass (Næsset 2002). Although no major differences were found between the models evaluated, the “SQRT” model (i.e. a model that is linearized through a square root transformation of the vector of responses) led to slightly more precise results than the other models evaluated in the study. This is in line with several other studies (Gregoire et al. 2008, Næsset 2011) where this model has been suggested for biomass prediction based on RS data. The “SQRT” model fitted our data well, and thus the variances of the model parameter estimates were lower when compared to using other models. Furthermore, Paper III investigated the choice of

covariance matrix estimator, since the covariances of the model parameter estimates affect the model-based variance estimators. One important issue in connection with this is whether or not the variance of the residual terms is homogeneous (e.g., Davidson and MacKinnon 1993). Often it is not, and thus alternative estimators should be applied. In general the best results were obtained when the NLS covariance matrix estimator was applied. However, for moderate to large sample sizes no major differences between the different estimators were observed, and thus the results of Paper III suggest that the choice of covariance matrix estimator is not a crucial part of model-based inference of growing stock volume based on RS data. Still, as pointed out by White (1980), in the case of non-homogeneous data and large sample sizes it might be advisable not to use the standard outputs from regression software, but rather to apply covariance matrix estimators that are adapted to this case.

In studying different sample sizes, expected results in terms of decreased standard errors over increasing sample sizes were obtained. However, an interesting feature of model-based prediction is that the magnitude of the model error variance component will remain approximately the same, regardless of the sample size. Thus, at a certain point the variance of estimators can only be negligibly decreased by increasing the sample size. Interestingly however, a slight tendency of decreased model error variance with increasing (first phase) sample size was observed. This was most likely due to the second phase sample being more evenly spread across the population, and thus the model parameters were more precisely estimated.

Model-based prediction for large-area surveys has been addressed in many recent studies. For example, McRoberts (2006 and 2013) applied model-based inference for estimating forest area based on combinations of Landsat data and field plots. Ståhl et al. (2011) applied model-based estimation in a study area in Norway, and the basic set of estimators from that study was applied in Paper III of this thesis. Magnussen (2015) presents some case examples and argues that model-based inference has several advantages over traditional design-based inference. Among others, Breidenbach and Astrup (2012) and Mandallaz (2014) have demonstrated the potential of model-based inference in small-area estimation. However, Paper III appears to be the first article in the context of applications in forest surveys that has addressed details linked to the choice of models and to model parameter estimation on the precision of estimators.

In Paper IV, the effects of positioning mismatches between field and RS data were studied. When such mismatches are present, the models developed for model-based prediction and model-assisted estimation will not be as good as when they are developed from perfectly matched data. Furthermore, the mismatches are likely to inflate the variance of model-assisted estimators since the differences between the actual state and the model estimation will normally increase, and a study by McRoberts (2010b) reports that positioning mismatches can often be substantial.

The results showed that the precision of estimation in both model-based prediction and model-assisted estimation decreased when the magnitude of the positional errors increased. In cases where the models were both developed from and applied to the same set of data, with positional errors, model-based prediction and model-assisted estimation were seen to be about equally prone to positional errors. The reason is that the correction term of the model-assisted estimator will always be zero (for a certain set of models; see Särndal et al. 1992), and thus the model-based and the model-assisted estimators will always lead to similar results. However, the model-based variance estimators in this case were less biased than the model-assisted variance estimators. In cases where the models were developed from an external set of data (without positional errors) but applied to a set of data with positional errors, the model-based prediction was superior to model-assisted estimation. The reason for this is that RS positional errors in this case will not affect the predicted model-based totals and means, but they will affect the correction terms employed in the model-assisted estimators.

Methodologically, it may be argued that a model-based prediction and model-assisted estimation cannot be compared through sampling simulation using a fixed study population, and this was seen to be the case in Papers III and IV. The reason is that model-based inference assumes the values of individual population elements to be random variables, and thus the population total and mean are also random variables. In principle, such comparisons should be based on the generation of multiple random populations over which the average performance of model-based prediction and model-assisted estimation is assessed. This approach was not used in the presented studies. However, since a large population was studied, the relative differences between the different possible outcomes of random population totals and means would be very small and thus the difference in results between the approach (using a single population for the simulations) and an alternative approach (using multiple random populations)

should be small. Instead, the main elements contributing to the variability within a given sampling strategy is the random selection of sampling units and the model parameter estimation errors. Also, Breidenbach et al. (2014) showed that the influence of random components linked to each element is typically negligible when surveys are conducted over large regions. Thus, as an approximation and simplification, a fixed population was assumed, and the fact that population totals and means are random variables in model-based inference was not accounted for. In this thesis, comparisons between different sampling strategies were made mainly through the precision of estimators. Costs were not explicitly considered. However, in designing forest inventories in general, costs are an important factor to consider. One approach to include costs is to search for the sampling strategy that maximises precision given the inventory budget (Cochran 1977). Another is to search for the sampling strategy that minimises the inventory cost, plus the expected loss due to non-optimal decisions (Ståhl et al. 1994). However, in the different sampling strategies assessed in this thesis, all of the alternatives involving a certain source of auxiliary data at a certain sample size would be of similar cost.

Considering the use of models in the estimation phase rather than in the design phase, at least three main approaches can be identified. These are:

1. Use of models in the context of design-based inference through model-assisted estimation.
2. Use of models in the context of model-based inference through model-based prediction.
3. Used of models in the context of hybrid inference.

The third approach might be considered a special case of model-based inference that can be applied in cases where auxiliary data are not available wall-to-wall, but must be acquired in the first phase of sampling. In this thesis, all three approaches were used and evaluated. An advantage of the hybrid approach is that it does not require auxiliary data to be available for all population elements. However, a disadvantage of all of the model-based approaches is the reliance on a model which may be incorrect.

Through the use of models in connection with model-assisted estimation, we stand by design-based inference, which is often considered to be an objective approach since the estimators will be (almost) design-unbiased, and thus on average the estimated value will (almost) coincide with the true value. Thus we do not take the risk that a poor model would make the estimator biased.

6. CONCLUSIONS

A general conclusion from this thesis is that RS auxiliary data can be used for improving the precision of estimators using field data from probability sample surveys for growing stock volume estimation. The use of LiDAR auxiliary data led to improvements, but it was shown that a combination of LiDAR and Landsat auxiliary data was superior to using LiDAR data alone.

Several comparisons of design-based inference using model-assisted estimators and model-based inference for large-area forest surveys were conducted. Yet, no general recommendation about what mode of inference is best suited for large-area forest surveys can be made. Model-assisted estimation can be safely applied even if the model relationship between the target and auxiliary variables is poor. In a worst case scenario where there is no relationship, the precision of the model-assisted estimation will be about the same as estimations that do not involve auxiliary data but employ only field data. Under favourable conditions the model-assisted estimators perform very well. Such conditions are characterised by a high correlation between the target variable and the auxiliary variable(s), the availability of a probability field sample, and no positional mismatches between the data sources. In the case where there are positional mismatches, the variance estimators may be severely biased, and may thus mislead the users of the survey results.

While model-assisted estimation has many advantages, so has model-based prediction. Especially, this is the case when there is a high correlation between the target and the auxiliary variable(s) and when access to the forest for field sampling is expensive. Regarding the effects of positional errors, model-based prediction appears to perform slightly better than model-assisted estimation. However, model-based prediction relies on the availability of a good model, and if no such model can be constructed, then model-based prediction is a poor alternative. Although this aspect is not specifically studied in this thesis, this would be the case for a large number of parameters that are traditionally assessed in national forest inventories, such as tree species composition, site

quality, soil condition and forest floor vegetation. Thus, the usefulness of model-based inference based on non-design-based field samples should be restricted to variables that are known to be highly correlated with RS data, such as growing stock volume and biomass.

However, in the case that a probability-based field sample is available and used for constructing the models, then model-based prediction and model-assisted estimation will in many cases lead to similar results in terms of bias and the precision of estimators, if the summed estimated regression model residuals divided by corresponding inclusion probabilities is equal zero, e.g. $\sum_{i=1}^{n} \frac{\hat{e}_i}{\pi_i} = 0$ (Särndal et al. 1992, p. 231-232).

REFERENCES

Andersen H. E. (2009). Using airborne light detection and ranging (LIDAR) to characterize forest stand condition on the Kenai Peninsula of Alaska. Western Journal of Applied Forestry, 24(2), 95-102.

Andersen H. E., Barrett T., Winterberger K., Strunk J., Temesgen H. (2009). Estimating forest biomass on the western lowlands of the Kenai Peninsula of Alaska using airborne lidar and field plot data in a model-assisted sampling design. In Proceedings of the IUFRO Division 4 Conference: "Extending Forest Inventory and Monitoring over Space and Time" (pp. 19-22).

Andersen H. E., Strunk J., Temesgen H., Atwood D., Winterberger K. (2012). Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: a case study in the boreal forests of interior Alaska. Canadian Journal of Remote Sensing, 37(6), 596-611.
http://dx.doi.org/10.5589/m12-003

Angelsen A., Brockhaus M. (Eds.). (2009). Realising REDD+: National strategy and policy options. CIFOR.

Asner G. P. (2009). Tropical forest carbon assessment: integrating satellite and airborne mapping approaches. Environmental Research Letters, 4(3), 034009.
http://dx.doi.org/10.1088/1748-9326/4/3/034009

Baffetta F., Fattorini L., Franceschi S., Corona P. (2009). Design-based approach to k-nearest neighbours technique for coupling field and remotely sensed data in forest surveys. Remote Sensing of Environment, 113(3), 463-475.
http://dx.doi.org/10.1016/j.rse.2008.06.014

Bohlin J., Wallerman J., Fransson J. E. (2012). Forest variable estimation using photogrammetric matching of digital aerial images in combination with a high-resolution DEM. Scandinavian Journal of Forest Research, 27(7), 692-699.
http://dx.doi.org/10.1080/02827581.2012.686625

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.
http://dx.doi.org/10.1007/s10342-012-0596-7

Breidenbach J., Antón-Fernández C., Petersson H., McRoberts R. E., Astrup R. (2014). Quantifying the model-related variability of biomass stock and change estimates in the Norwegian National Forest Inventory. Forest Science, 60(1), 25-33.
http://dx.doi.org/10.5849/forsci.12-137

Cassel C. M., Särndal C. E., Wretman J. H. (1977). Foundations of inference in survey sampling. New York: John Wiley and Sons.

Cienciala E., Tomppo E., Snorrason A., Broadmeadow M., Colin A., Dunger K., Exnerova Z., Ståhl G. (2008). Preparing emission reporting from forests: use of National Forest Inventories in European countries.

Cochran, W. G. Sampling techniques. (1977). New York: John Wiley and Sons.

Corona P., Chirici G., Franceschi S., Maffei D., Marcheselli M., Pisani C., Fattorini L. (2014a). Design-based treatment of missing data in forest inventories using canopy heights from aerial laser scanning. Canadian Journal of Forest Research, 44(8), 892-902.
http://dx.doi.org/10.1139/cjfr-2013-0521

Corona P., Fattorini L., Franceschi S., Scrinzi G., Torresan C. (2014b). Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based, and hybrid perspectives. Canadian Journal of Forest Research, 44(11), 1303-1311.
http://dx.doi.org/10.1139/cjfr-2014-0203

Davidson R., MacKinnon J. G. (1993). Estimation and inference in econometrics. OUP Catalogue.

Ene L. T., Næsset E., Gobakken T., Gregoire T. G., Ståhl G., Nelson R. (2012a). Assessing the accuracy of regional LiDAR-based biomass estimation using a simulation approach. Remote Sensing of Environment, 123, 579-592.
http://dx.doi.org/10.1016/j.rse.2012.04.017

Ene L.T., Næsset E., Gobakken T. (2012b). Model-based inference for k-nearest neighbours predictions using a canonical vine copula. Scandinavian Journal of Forest Research, 28, 266-281.
http://dx.doi.org/10.1080/02827581.2012.723743

Ene L., Næsset E., Gobakken T., Gregoire T. G., Ståhl G., Holm S. (2013). A simulation approach for accuracy assessment of two-phase post-stratified estimation in large-area LiDAR biomass surveys. Remote Sensing of Environment, 133, 210–224.
http://dx.doi.org/10.1016/j.rse.2013.02.002

Eskelson B.N.I., Madsen L., Hagar J.C., Temesgen H. (2011). Estimating riparian understory vegetation cover with beta regression and copula models. Forest Science, 57, 212-221.

ESRI (2011). ArcGIS Desktop: Release 10. Redlands, CA: Environmental System Research Institute.

Europe F. (2011). UNECE and FAO (2011) State of Europe's forests 2011. Status and trends in sustainable forest management in Europe.

Gobakken T., Næsset E., Nelson R., Bollandsås O. M., Gregoire T. G., Ståhl G., Holm S., Ørka H.O., Astrup R. (2012). Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. Remote Sensing of Environment, 123, 443-456.
http://dx.doi.org/10.1016/j.rse.2012.01.025

Gobakken T., Korhonen L., Næsset E. (2013). Laser-assisted selection of field plots for an area-based forest inventory. Silva Fennica, 47(5).
http://dx.doi.org/10.14214/sf.943

Gregoire T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. Canadian Journal of Forest Research, 28(10), 1429-1447.
http://dx.doi.org/10.1139/x98-166

Gregoire T. G., Lin Q. F., Boudreau J., Nelson R. (2008). Regression estimation following the square-root transformation of the response. Forest Science, 54(6), 597-606.

Gregoire T. G., Ståhl G., Næsset E., Gobakken T., Nelson R., Holm S. (2011). Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Research, 41(1), 83-95.
http://dx.doi.org/10.1139/X10-195

Grafström A., Lundström N.L.P., Schelin L. (2012). Spatially balanced sampling through the Pivotal method. Biometrics, 68(2), 514-520.
http://dx.doi.org/10.1111/j.1541-0420.2011.01699.x

Grafström A., Lundström N.L.P. (2013). Why Well Spread Probability Samples are Balanced. Open Journal of Statistics, 3(1), 36-41.
http://dx.doi.org/10.4236/ojs.2013.31005

Grafström A., Ringvall A. H. (2013). Improving forest field inventories by using remote sensing data in novel sampling designs. Canadian Journal of Forest Research, 43(11), 1015-1022.
http://dx.doi.org/10.1139/cjfr-2013-0123

Hansen M. C., Stehman S. V., Potapov P. V., Loveland T. R., Townshend J. R., DeFries R. S., Pittman K. W., Arunarwati B., Stolle F., Steininger M. K., Carroll M., DiMiceli C. (2008). Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. Proceedings of the National Academy of Sciences, 105(27), 9439-9444.
http://dx.doi.org/10.1073/pnas.0804042105

Hill J., Diemer C., Stöver O., Udelhoven T. (1999). A local correlation approach for the fusion of remote sensing data with different spatial resolutions in forestry applications. International Archives of Photogrammetry and Remote Sensing, 32(Part 7), 4-3.

Holmström H., Nilsson M., Ståhl G. (2001). Simultaneous estimations of forest parameters using aerial photograph interpreted data and the k nearest neighbour method. Scandinavian Journal of Forest Research, 16(1), 67-78.
http://dx.doi.org/10.1080/028275801300004424

Horvitz D.G., Thompson D.J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663-685.
http://dx.doi.org/10.1080/01621459.1952.10483446

Hyyppä J., Inkinen M. (1999). Detecting and estimating attributes for single trees using laser scanner. The Photogrammetric Journal of Finland, 16(2), 27-42.

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. Geoscience and Remote Sensing, IEEE Transactions on, 39(5), 969-975.
http://dx.doi.org/10.1109/36.921414

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.
http://dx.doi.org/10.1080/01431160701736489

Kershaw J. A., Jr. Richards E. W., McCarter J. B., Oborn S. (2010). Spatially correlated forest structures: A simulation approach using copulas. Computers and Electronics in Agriculture, 74, 120-128.
http://dx.doi.org/10.1016/j.compag.2010.07.005

Kraus K., Pfeifer N. (2001). Advanced DTM generation from LiDAR data. International Archives of Photogrammentry Remote Sensing And Spatial Information Sciences, 34, 23-30.

Lappi J. (2001). Forest inventory of small areas combining the calibration estimator and a spatial model. Canadian Journal of Forest Research, 31(9), 1551-1560.
http://dx.doi.org/10.1139/x01-078

Leberl F., Irschara A., Pock T., Meixner P., Gruber M., Scholz S., Wiechert A. (2010). Point Clouds. Photogrammetric Engineering & Remote Sensing, 76(10), 1123-1134.
http://dx.doi.org/10.14358/PERS.76.10.1123

Lindberg E., Olofsson K., Holmgren J., Olsson H. (2012). Estimation of 3D vegetation structure from waveform and discrete return airborne laser scanning data. Remote Sensing of Environment, 118, 151-161.
http://dx.doi.org/10.1016/j.rse.2011.11.015

Magnussen S. (2015). Arguments for a model-dependent inference? Forestry, cpv002.
http://dx.doi.org/10.1093/forestry/cpv002

Maltamo M., Packalén P., Peuhkurinen J., Suvanto A., Pesonen A., Hyyppä J. (2007, September). Experiences and possibilities of ALS based forest inventory in Finland. In Proceedings of ISPRS Workshop on Laser Scanning (pp. 270-279).

Maltamo M., Packalen P., Suvanto A., Korhonen K. T., Mehtätalo L., Hyvönen P. (2009). Combining ALS and NFI training data for forest management planning: a case study in Kuortane, Western Finland. European Journal of Forest Research, 128(3), 305-317.
http://dx.doi.org/10.1007/s10342-009-0266-6

Mandallaz D. (2014, May). Regression estimators in three-phase sampling. In ForestSAT2014 Open Conference System.

Matérn B. (1960). Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations. Meddelanden fran statens Skogsforskningsinstitut, 49(5).

McGaughey R. (2012). FUSION/LDV: Software for LIDAR Data Analysis and Visualization, Version 3.01. http://forsys.cfr.washington.edu/fusion/fusionlatest.html. Accessed: 24 August 2012.

McRoberts R. E. (2006). A model-based approach to estimating forest area. Remote Sensing of Environment, 103(1), 56-66.
http://dx.doi.org/10.1016/j.rse.2006.03.005

McRoberts R. E., Wendt D. G., Nelson M. D., Hansen M. H. (2002). Using a land cover classification based on satellite imagery to improve the precision of forest inventory area estimates. Remote Sensing of Environment, 81(1), 36-44.
http://dx.doi.org/10.1016/S0034-4257(01)00330-3

McRoberts R. E., Tomppo E., Schadauer K., Vidal C., Ståhl G., Chirici G., Lanz A., Cienciala E., Winter S., Smith W. B. (2009). Harmonizing national forest inventories. Journal of Forestry, 107(4), 179-187.

McRoberts R. E. (2010a). Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sensing of Environment, 114(5), 1017-1025.
http://dx.doi.org/10.1016/j.rse.2009.12.013

McRoberts R. E. (2010b). The effects of rectification and Global Positioning System errors on satellite image-based estimates of forest area. Remote Sensing of Environment, 114(8), 1710-1717.

http://dx.doi.org/10.1016/j.rse.2010.03.001

McRoberts R. E., Ståhl G., Vidal C., Lawrence M., Tomppo E., Schadauer K., Chirici G., Bastrup-Birk A. (2010). National forest inventories: prospects for harmonised international reporting. In National Forest Inventories (pp. 33-43). Springer Netherlands.
http://dx.doi.org/10.1007/978-90-481-3233-1_3

McRoberts R. E. (2011). Satellite image-based maps: Scientific inference or pretty pictures? Remote Sensing of Environment, 115(2), 715-724.
http://dx.doi.org/10.1016/j.rse.2010.10.013

McRoberts R. E., Næsset E., Gobakken T. (2013). Inference for lidar-assisted estimation of forest growing stock volume. Remote Sensing of Environment, 128, 268-275.
http://dx.doi.org/10.1016/j.rse.2012.10.007

McRoberts R. E., Næsset E., Gobakken T. (2014). Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information. Remote Sensing of Environment, 154, 226-233.
http://dx.doi.org/10.1016/j.rse.2014.08.028

Mery G., Alfaro R. I., Kanninen M., Lobovikov M. (2005). Forests in the global balance-changing paradigms. IUFRO.

Miina J., Heinonen J. (2008). Stochastic simulation of forest regeneration establishment using a multilevel multivariate model. Forest Science, 54, 206-219.

Nelsen R.B. (2006). An introduction to copulas (2nd ed). New York: Springer.

Nelson R., Krabill W., Tonelli J. (1988). Estimating forest biomass and volume using airborne laser data. Remote Sensing of Environment, 24(2), 247-267.
http://dx.doi.org/10.1016/0034-4257(88)90028-4

Nelson R., Oderwald R., Gregoire T. G. (1997). Separating the ground and airborne laser sampling phases to estimate tropical forest basal area, volume, and biomass. Remote Sensing of Environment, 60(3), 311-326.
http://dx.doi.org/10.1016/S0034-4257(96)00213-1

Nelson R., Gobakken T., Ståhl G., Gregoire T. G. (2008). Regional Forest Inventory using an Airborne Profiling LiDAR (< Special Issue> Silvilaser). Journal of Forest Planning, 13, 287-294.

Nelson R., Boudreau J., Gregoire T. G., Margolis H., Næsset E., Gobakken T., Ståhl G. (2009). Estimating Quebec provincial forest resources using ICESat/GLAS. Canadian Journal of Forest Research, 39(4), 862-881.
http://dx.doi.org/10.1139/X09-002

Neyman J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society, 558-625.
http://dx.doi.org/10.2307/2342192

Nilsson M., Folving S., Kennedy P., Puumalainen J., Chirici G., Corona P., Marchetti M., Olsson H., Ricotta C., Ringvall A., Stahl G., Tomppo E. (2003). Combining remote sensing and field data for deriving unbiased estimates of forest parameters over large regions. In Advances in forest inventory for sustainable forest management and biodiversity monitoring (pp. 19-32). Springer Netherlands.
http://dx.doi.org/10.1007/978-94-017-0649-0_2

Næsset E. (1997). Estimating timber volume of forest stands using airborne laser scanner data. Remote Sensing of Environment, 61(2), 246-253.
http://dx.doi.org/10.1016/S0034-4257(97)00041-2

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.
http://dx.doi.org/10.1016/S0034-4257(01)00290-5

Næsset E. (2009). Effects of different sensors, flying altitudes, and pulse repetition frequencies on forest canopy metrics and biophysical stand properties derived from small-footprint airborne laser data. Remote Sensing of Environment, 113(1), 148-159.
http://dx.doi.org/10.1016/j.rse.2008.09.001

Næsset E. (2011). Estimating above-ground biomass in young forests with airborne laser scanning. International Journal of Remote Sensing, 32(2), 473-501.

http://dx.doi.org/10.1080/01431160903474970

Næsset E., Gobakken T., Holmgren J., Hyyppä H., Hyyppä J., Maltamo M., Nilsson M., Olsson H., Persson Å., Söderman U. (2004). Laser scanning of forest resources: the Nordic experience. Scandinavian Journal of Forest Research, 19(6), 482-499.
http://dx.doi.org/10.1080/02827580410019553

Næsset E., Gobakken T., Nelson R. (2006, October). Sampling and mapping forest volume and biomass using airborne LIDARs. In Proceedings of the eight annual forest inventory and analysis symposium (pp. 297-301).

Næsset E., Gobakken T., Solberg S., Gregoire T. G., Nelson R., Ståhl G., Weydahl D. (2011). Model-assisted regional forest biomass estimation using LiDAR and InSAR as auxiliary data: A case study from a boreal forest area. Remote Sensing of Environment, 115(12), 3599-3614.
http://dx.doi.org/10.1016/j.rse.2011.08.021

Næsset E., Bollandsås O. M., Gobakken T., Gregoire T. G., Ståhl G. (2013a). Model-assisted estimation of change in forest biomass over an 11year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide "activity data". Remote Sensing of Environment, 128, 299-314.
http://dx.doi.org/10.1016/j.rse.2012.10.008

Næsset E., Gobakken T., Bollandsås O. M., Gregoire T. G., Nelson R., Ståhl G. (2013b). Comparison of precision of biomass estimates in regional field sample surveys and airborne LiDAR-assisted surveys in Hedmark County, Norway. Remote Sensing of Environment, 130, 108-120.
http://dx.doi.org/10.1016/j.rse.2012.11.010

Opsomer J. D., Breidt F. J., Moisen G. G., Kauermann G. (2007). Model-assisted estimation of forest resources with generalized additive models. Journal of the American Statistical Association, 102(478), 400-409.
http://dx.doi.org/10.1198/016214506000001491

Prasad N. G. N., Rao J. N. K. (1990). The estimation of the mean squared error of small-area estimators. Journal of the American Statistical Association, 85(409), 163-171.
http://dx.doi.org/10.1080/01621459.1990.10475320

R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 (URL: http://www.R-project.org/, last accessed March 17, 2014).

Saarela S., Kangas A., Tuominen S., Holopainen M., Hyyppä J., Vastaranta M., Kankare V. (2012, September). Comparing performances of ALS and Landsat 7 ETM+ satellite optical data in stratification-based sampling method for large-area forest inventory. Full Proceedings, SilviLaser 2012, Sept. 16-19 2012, Vancouver, Canada/Ed. Coops, NC & Wulder, MA.

Saatchi S. S., Harris N. L., Brown S., Lefsky M., Mitchard E. T., Salas W., Zutta B. R., Buermann W., Lewis L., Hagen S., Petrova S., White L., Silman M., Morel A. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. Proceedings of the National Academy of Sciences, 108(24), 9899-9904.
http://dx.doi.org/10.1073/pnas.1019576108

Schepsmeier U., Stoeber J., Brechmann E. C., Graeler B. (2013). VineCopula: Statistical inference of vine copulas. R package version 1.2. http://CRAN.R-project.org/package=VineCopula

Ståhl G., Carlsson D., Bondesson L. (1994). A method to determine optimal stand data acquisition policies. Forest Science, 40(4), 630-649.

Ståhl G., Holm S., Gregoire T. G., Gobakken T., Næsset E., Nelson R. (2011). Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Research, 41(1), 96-107.
http://dx.doi.org/10.1139/X10-161

Ståhl G., Cienciala E., Chirici G., Lanz A., Vidal C., Winter S., McRoberts E. E., Rondeux J., Schadauer K., Tomppo E. (2012). Bridging national and reference definitions for harmonizing forest statistics. Forest Science, 58(3), 214-223.
http://dx.doi.org/10.5849/forsci.10-067

Ståhl G., Heikkinen J., Petersson H., Repola J., Holm S. (2014). Sample-Based Estimation of Greenhouse Gas Emissions From Forests – A New Approach to Account for Both Sampling and Model Errors. Forest Science, 60(1), 3-13.

http://dx.doi.org/10.5849/forsci.13-005

Särndal C. E. (1978). Design-based and model-based inference in survey sampling. Scandinavian Journal of Statistics, 27-52.

Särndal C. E., Swensson, B., Wretman J. (1992). Model assisted survey sampling. Springer. http://dx.doi.org/10.1007/978-1-4612-4378-6

Tomppo E. (2006). The Finnish national forest inventory. In: Kangas A., Maltamo M. (eds) Forest Inventory. Methodology and Applications, Managing Forest Ecosystem. Springer Netherlands 10, 179-194. http://dx.doi.org/10.1007/1-4020-4381-3_11

Tomppo E., Haakana M., Katila M., Peräsaari J. (2008). Multi-source National Forest Inventory: Methods and Applications. Springer New York 18.

Tomppo E., Gschwantner M., Lawrence M., McRoberts R. E. (2010). National Forest Inventories. Pathways for Common Reporting. European Science Foundation. http://dx.doi.org/10.1007/978-90-481-3233-1

Tomppo E., Heikkinen J., Henttonen H. M. (2011). Designing and Conducting a Forest Inventory-case: 9th National Forest Inventory of Finland: Designing and Conducting a Forest Inventory-Case: 9th National Forest Inventory of Finland (Vol. 22). Springer. http://dx.doi.org/10.1007/978-94-007-1652-0

U.S. Geological Survey (2014). Landsat Missions. http://landsat.usgs.gov/index.php, Accessed: 28 March 2011.

Wang M., Rennolls K., Tang S. (2008). Bivariate distribution modelling of tree diameters and heights: Dependency modeling using copulas. Forest Science, 45, 284-293.

Wang M., Upadhyay A., Zhang L. (2010). Trivariate distribution modeling of tree diameter, height, and volume. Forest Science, 56, 290-300.

White H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.Econometrica: Journal of the Econometric Society, 817-838. http://dx.doi.org/10.2307/1912934

Woodall C. W., Rondeux J., Verkerk P. J., Ståhl G. (2009). Estimating dead wood during national forest inventories: a review of inventory methodologies and suggestions for harmonization. Environmental management, 44(4), 624-631. http://dx.doi.org/10.1007/s00267-009-9358-9