Lappeenranta
University of Technology

Virpi Junttila

AUTOMATED, ADAPTIVE METHODS FOR FOREST INVENTORY

Thesis for the degree of Doctor of Science (Technology) to be presented with due permission for public examination and criticism in Auditorium 1383 at Lappeenranta University of Technology, Lappeenranta, Finland on the 4th of February, 2011, at 12 pm.

Acta Universitatis
Lappeenrantaensis 424

Supervisor Docent, PhD Tuomo Kauranne
Faculty of Technology
Department of Mathematics and Physics
Lappeenranta University of Technology
Finland

Reviewers PhD Andrew O. Finley
Department of Forestry and Geography
Michigan State University
USA

Docent, D.Sc. Lauri Mehtätalo
School of Forest sciences
University of Eastern Finland
Finland

Opponent PhD Andrew O. Finley
Department of Forestry and Geography
Michigan State University
USA

ISBN 978-952-265-047-4
ISBN 978-952-265-048-1 (PDF)
ISSN 1456-4491

Lappeenrannan teknillinen yliopisto
Digipaino 2011

Abstract

Virpi Junttila
AUTOMATED, ADAPTIVE METHODS FOR FOREST INVENTORY
Lappeenranta, 2011
65 p.
Acta Universitatis Lappeenrantaensis 424
Diss. Lappeenranta University of Technology
ISBN 978-952-265-047-4, ISBN 978-952-265-048-1 (PDF), ISSN 1456-4491

Forest inventories are used to estimate forest characteristics and the condition of forest for many different applications: operational tree logging for forest industry, forest health state estimation, carbon balance estimation, land-cover and land use analysis in order to avoid forest degradation etc. Recent inventory methods are strongly based on remote sensing data combined with field sample measurements, which are used to define estimates covering the whole area of interest. Remote sensing data from satellites, aerial photographs or aerial laser scannings are used, depending on the scale of inventory.

To be applicable in operational use, forest inventory methods need to be easily adjusted to local conditions of the study area at hand. All the data handling and parameter tuning should be objective and automated as much as possible. The methods also need to be robust when applied to different forest types.

Since there generally are no extensive direct physical models connecting the remote sensing data from different sources to the forest parameters that are estimated, mathematical estimation models are of "black-box" type, connecting the independent auxiliary data to dependent response data with linear or nonlinear arbitrary models. To avoid redundant complexity and over-fitting of the model, which is based on up to hundreds of possibly collinear variables extracted from the auxiliary data, variable selection is needed.

To connect the auxiliary data to the inventory parameters that are estimated, field work must be performed. In larger study areas with dense forests, field work is expensive, and should therefore be minimized. To get cost-efficient inventories, field work could partly be replaced with information from formerly measured sites, databases.

The work in this thesis is devoted to the development of automated, adaptive computation methods for aerial forest inventory. The mathematical model parameter definition steps are automated, and the cost-efficiency is improved by setting up a procedure that utilizes databases in the estimation of new area characteristics.

Keywords: forest inventory, sparse Bayesian regression, sample plot database, remote sensing, histogram calibration, heuristic plot selection

UDC 519.23 : 528.7/.8 : 630*5

Preface

This work was carried out in the Laboratory of Mathematics and Physics in Lappeenranta University of Technology, Finland, between 2006 and 2010. The Finnish Graduate School of Inverse Problems is acknowledged with great gratitude for the financial support of this work.

There are numerous people, who have helped me in different ways during the course of this long process, and to whom I am greatly indebted. First of all, I would like to thank all my colleagues in the Laboratory of Mathematics and Physics for creating such a nice and caring atmosphere: it has been a pleasure to work with you. The interesting and fun conversations during the coffee (tea) breaks have been a refreshing joy of the working days.

Especially I want to thank my supervisor Tuomo Kauranne, not only for sharing his knowledge of the research area and its challenges, and good conversations about the science and everything else, but also for his constant positivity and encouragement: you have a great gift to turn other people's moments of lack of belief to feeling of optimism and belief of new possibilities. I want to thank you also for creating warm and caring working atmosphere where different feelings are allowed.

This work has been carried out between two areas of research - forestry and applied mathematics. I would like express my gratitude to Matti Maltamo and his research team at University of East Finland, in particular Petteri Packalén, for insightful discussions, and to Jussi Peuhkurinen for helping me out with many issues related to forest data. I want to thank Jussi, and other people at Arbonaut, especially Vesa Leppänen, Martin Gunia, Hanna Parviainen and Jaakko Ketola, for collecting and preparing the forest data used in this thesis. I am also grateful to the reviewers, Lauri Mehtätalo and Andrew Finley, for their valuable comments on this thesis.

Happy life at home is a good balance to the intensive research at work. So, most of all, I want to thank my friends and family for their care and understanding. All my friends from along the way, especially the oldest ones, Krista, Tanja and Paula: thank you for being there and reminding me where the real life is. My brother Tommi has been a great help for me in many ways, giving support as a friend but also in technical problems: I want to thank you for that, and especially for your patience with my numerous questions concerning Latex and computers. My parents Terttu and Antero have always encouraged me to challenge myself in the field of mathematics, even since I was a child: I want to thank you for leading me to the first steps of mathematical thinking. In the last nine years, you have also concretely helped me to combine my work and family life by being there in need. Thank you for all your care, patience and understanding. I also want to thank my son Oskari for being the lovely boy he is, and for all the interesting discussions with him: without you my life would be empty.

Lappeenranta, January 2011

Virpi Junttila

LIST OF THE ORIGINAL ARTICLES AND THE AUTHOR'S CONTRIBUTION

This thesis consists of an introductory part and three original refereed articles in scientific journals. The articles and the author's contributions in them are summarized below.

I **Junttila, V., M. Maltamo and T. Kauranne**, Sparse Bayesian Estimation of Forest Stand Characteristic from Airborne Laser Scanning, *Forest Science*, 54(5), 543-552, 2008.

II **Junttila, V., T. Kauranne and V. Leppänen**, Estimation of Forest Stand Parameters from LiDAR Using Calibrated Plot Databases, *Forest Science*, 56(3), 257-270, 2010.

III **Junttila, V. and T. Kauranne**, Evaluating the robustness of plot databases in species-specific LiDAR-based forest inventory, resubmitted to *Forest Science* 2010

V. Junttila is the principal author of all these publications. She has planned and written the mathematical algorithms and calculated all the results given in the publications. She also wrote most of the text and was the corresponding author in the publication process of each paper.

ABA	Area Based Approach
AIC	Akaike's Information Criterion
ALS	Airborne Laser Scanning
ARD	Automatic Relevance Detection
BIC	Bayesian Information Criterion
CBD	Conservation of Biological Diversity
CCA	Canonical Correlation Analysis
CHM	Canopy Height Model
CIR	Colour InfraRed
dgM	Diameter of basal area (G) Median tree
DIC	Deviance Information Criterion
DRD	Discrete-Return Device
DSM	Digital Surface Model
DTM	Digital Terrain Model
G	average breast height basal area per hectare
GIS	Geographical Information System
GLS	Generalized Least-Squares
GPS	Global Positioning System
hgM	Height of basal area (G) Median tree
INPE	Instituto Nacional de Pesquisas Espaciais
IR	InfraRed
ITC	Individual Tree Crown approach
LiDAR	Light Detection And Ranging
k-MSN	k-Most Similar Neighbours
k-NN	k-Nearest Neighbours
LOO	Leave-One-Out
LOSO	Leave-One-Stand-Out
N	Number of stems per hectare
NFI	National Forest Inventory

NIR	Near InfraRed
OLS	Ordinary Least Squares
PLS	Partial Least Squares regression
SBR	Sparse Bayesian Regression
SUR	Seemingly Unrelated Regression
UV	UltraViolet
REDD	Reducing Emissions from Deforestation and forest Degradation
RGB	Red, Green, Blue
RMSE	Root Mean Square Error
RVM	Relevance Vector Machine
V	Volume
WRD	Waveform Recording Device

PART I: OVERVIEW OF THE THESIS

CHAPTER I

Introduction

Forest resources are of great importance in Finland. During the last centuries, people have used forests for different aspects of living - from household use of timber such as firewood, slash-and-burn farming, and building with wood, to industrial use such as burning wood to make tar and using timber in sawmills. From the end of the 19th century, the role of large scale forest industry such as sawmills and pulp and paper mills, grew and became crucial for the Finnish economy. To supply enough raw material for the industry, the use of forest resources spread deeper to wilderness forests. A concern for the sufficiency of forest resources emerged, and the need to estimate forest resources of the country was established. In the 20th century, Finnish forest management became strongly controlled by the government and the main goal became to secure the supply of timber for industry.

In a global view, forests have different values depending on the countries and forest types in them. In addition to the industrial and economical use of timber, the importance of forests as a carbon sink has increased in value. Problems related to climate change have come to public knowledge and awareness of the role of forests has become greater. Carbon sinks will most probably have a significant role in the future as international treaties for reducing greenhouse gas concentration in the atmosphere are devised, and as a consequence, forests represent a financial asset. Also the international treaty on conservation of biological diversity (CBD) from Rio de Janeiro, 1992, demands sustainable use of forests. These days, different certifications of the sustainable management of forests are used to ensure that biodiversity is taken into account in forest management.

Verification of the current state or the direction of development of forests from the industrial or the ecological point of view, generate a strong need to measure and estimate forests and their characteristics. Since the end of the 19th century, different forest inventory methods have been developed to respond to concerns expressed locally, nationally and internationally for improved forest management and protection of forests.

Forest inventories can be based on purely statistical estimates of forest characteristics, estimated from field work measurements in sample plots, e.g. in national forest inventories (NFI), or as in many recent inventories, remote sensing data from different sources are widely used concurrent with the field work to estimate large and small area inventory parameters. For references about different approaches introduced here, see the following chapters. The most common remote sensing data sources used in these multi-source forest inventories are satellite images, digital aerial photographs and aerial laser scanning of the forest area. Selection of data source depends on the purpose and size of the inventory. Remote sensing data serves as auxiliary data, which covers the whole area of

interest, not only the field sample plots. This data can be used to estimate forest characteristics (forest stand parameters) of the whole target area with higher local accuracy and more cost-efficiently than pure statistical field sample plot based estimates.

Remote sensing data generally gives no direct estimates of forest characteristics, only variables that correlate more or less with them. Thus some suitable mathematical modeling approach, depending on the data sources and forest characteristics at hand, is needed. Mathematical models are built using the field measurements connected to the remote sensing data of the same area, giving a model that can be used to extrapolate the remote sensing data information to target areas without field measurements. To cover the variability of forest characteristics at total and species specific level in a given study area, a large number of field measurements at carefully selected plot areas is needed. Resulting estimates contain errors, depending on the suitability of the used method to the task at hand and on the correlation between the variables and the true values of the estimates. Different mathematical models can be used, from the individual tree estimation level to compartment based estimates. For forest management inventory purposes, compartment based approaches are often used since they produce estimates at the desired level and accuracy in an efficient manner.

Remote sensing data features of the same area are used as independent variables for the estimation of stand parameters. Different variables correlate with different stand parameters. The number of variables may be large, even hundreds, and the correlation within variables may be high. This may lead to serious problems in estimation accuracy outside field sample plot areas due to over-learning and possible multicollinearity of the variables of the model. For each mathematical stand parameter estimation approach in compartment based inventory, variable selection is a crucial task. It is performed e.g. by a cross-verification method or by step-wise regression with some stopping criteria. These methods are slow and laborious to perform. Each inventory area is modelled with different model parameters and variable sets, requiring a large amount of field measurements and model definition work. It is costly and time-consuming, and can be an obstacle for operative inventories.

Forest inventory modelling methods at management level, e.g. inventory for purposes of operational planning of a logging strategy, need to be easily adjusted to local forest characteristics and data sources. Large amount of time-consuming and expensive field work and any hand-work parameter tuning or variable selection in model preparation are undesirable. In this thesis, the main goal is to define automatic and adaptive methods to estimate forest stand parameters of a new, uninvented study area with low costs. A method which performs variable selection in regression automatically, Sparse Bayesian regression, is introduced to inventory tasks. The amount of required field measurement work is diminished by using formerly measured inventory areas, or databases, to define model parameters also for the new area. Database data is calibrated and preselected to fit the new area data quality and forest stand variability.

The thesis is organized as follows. Chapter 2 gives an overview of sampling methods and different remote sensing data sources used in forest inventory in Finland and also shortly discusses current challenges and approaches in forest inventory in a global view. Chapter 3 discusses the most commonly used mathematical estimation methods in remote sensing based forest inventory and problems concerning their performance accuracy. The objectives of the research work of the thesis are discussed in Chapter 4. The first part of the thesis - a new method for variable selection in forest inventory, Sparse Bayesian regression, is introduced and verified in Chapter 5 which also summarizes the main results of publication (**I**). The second part of the thesis - the use of existing, formerly measured data of other inventory areas in the estimation of a new site using aerial laser scanning data and digital aerial images as auxiliary data, is introduced in Chapter 6. The chapter

summarizes the test results of publication (**I**) concerning cases with a sparse set of field sample plots and unifies the procedures described in publications (**II**) and (**III**). Database assisted estimation results are given in Chapter 6.5. Pros and cons of the given method and future tasks for research are discussed in Chapter 7.

CHAPTER II

Background on forest inventory methods

In forest inventories, estimates of forest characteristics of the inventory area are based on the knowledge of field sample plots located in the area. The measurements of the forest characteristics, forest stand parameters, in the field sample plots are used as the "ground truth data" of the area. These days, the data of field measurements is generally augmented with other data - remote sensing measurements from different sources, which are achieved over the whole area of interest. In estimation of forest inventory parameters, data of field sample plots is extrapolated over the whole inventory area using suitable methods.

2.1 Field sample plot measurements

In Finnish forest inventories, forest inventory parameters are generally measured on field sample plots. Plot locations are determined with a sampling strategy that depends on the aims of the inventory, the shape of the inventory area, and possibly forest characteristics (each forest type of the area should be included in the samples). The number of plots required depends on the aspired accuracy of the estimates, and variability of the forest characteristics in the inventory area.

Field sample plot measurements serve as the "ground truth" for the estimates derived for larger areas with different methods. Errors made in the precision of measurements in the field sample plots accumulate to the estimates of other, unmeasured target plots, see e.g. Haara and Korhonen (2004) for a discussion of measurement errors in Finnish forests. Field measurement accuracy has always been an important issue in different inventory procedures, see e.g. Tomppo and Heikkinen (1999); Tomppo (2006) for the history of field sample measurement techniques used in Finland.

Forest stand inventory parameters of field sample plots in boreal forests are mainly measured using relascope sampling (Bitterlich, 1948). In relascope sampling measurements, the trees are viewed from the centre point of the plot, and included in it if the breast height diameter fills the horizontal angle of the relascope. Thus the inclusion probability is proportional to the to the basal area of the tree, i.e. the cross-section at breast height. The basal area of the plot can then be calculated using the number of trees included, multiplied with a basal area factor depending on the angle of the relascope. Different basal area factors can be used in targets with different stem density, see e.g. Tomppo (2006).

More accurate measurement information of the forest stand characteristics is acquired by more

detailed field measurements. Single-tree measurements of field sample plots are needed for reliable and precise estimates of the inventory area. In field data acquirement, only the species specific diameter and stem number of trees can be measured accurately. A hypsometer can be used to measure the height of trees, using the principles of triangles in geometry. Volume measurements are estimates derived from the other measurements. Height and volume models for different species have been given e.g. by Veltheim (1987); Laasasenaho (1982).

2.2 Remote sensing data in forest inventory

During the last decades, remote sensing data has changed the inventory strategies greatly - first in the form of digitized aerial photographs and satellite images as such data became available for forest inventory. Later on, experiments with airborne laser scanning in forest inventory were also performed. Many different sources of remote sensing data have been tested and used by now. An overall description of the various methods is given e.g. in Kangas and Maltamo (2006).

Remote sensing data of different types has been used as auxiliary data covering the whole inventory area. Plotwise estimates are derived by merging remote sensing data and field measurements using a suitable mathematical model. Utilizing auxiliary data that covers the whole study area, the plots in it are divided into two categories: those sample plots that cover the study area as a systematical grid and contain the auxiliary data, and the part of the sample plots which are also measured in the field and serve as the ground truth and the reference plots. The auxiliary data can then be utilized to predict the characteristics (i.e. forest stand variables or parameters) of the unmeasured sample plots, target plots. Using remote sensing data or field work measurement information, plots can be divided to larger entities, stands or clusters, containing plots of similar forest types. To cover the variation of the forest stand variables of the area, a sufficient number of plots needs to be measured. The methods using auxiliary data are found to be feasible for inventories on management planning level and large area inventories, see e.g. Holmgren (2004); LeMay and Temesgen (2005); Næsset (2004c); Katila and Tomppo (2001) for estimates of forest stand parameters made with different auxiliary data, approaches, and area sizes.

2.2.1 Geographical information systems

In order to estimate new areas outside the field measurement areas, forest inventory remote sensing variables must be linked to their measurement spatial location by geographical coordinates or some other method. Geographical information systems (GIS) are used. GIS in forest inventory means merging of inventory data and database technology. This information can be stored digitally in vector or raster form. Vector form defines areas by vectors limiting them, raster form by rows and columns of pixels, as small area units. In forest inventory, raster form is a natural approach since most of the remote sensing data is also in raster form. Remote sensing data is operated in a positioning system, and all measurements are handled together with given map information.

2.2.2 Satellite images

Optical satellite images, such as Landsat TM and Spot, cover large areas with cheap costs, and are thus favourable to large area forest inventory purposes, such as NFI's. With large covering, they are also more likely to yield essential cloud-free images, since there are likely to be multiple images

of the same area. Utilizing digital base maps, the images can be spatially located to geographical coordinates and areas not containing forests are disregarded. Also image analysis can be utilized to delineate e.g. waters and peat production areas before inventory analyses. Satellite image spatial resolution range varies depending on the equipment and the channel. For Spot and Landsat it is between some meters to approximately 30 meters. Spatial resolution of more expensive satellite data, very high resolution satellite imagery, varies from less than a meter to some meters, depending on the band mode (IKONOS, QuickBird). Satellite image data may consist of several spectral bands, or channels, with each channel representing an image with a different wavelength, varying from ultra violet light (UV) and visible light (RGB) to infrared (IR). See e.g. Holopainen and Kalliovirta (2006) for more information about different satellite imagery.

2.2.3 Aerial photographs

Aerial photographs are taken from aeroplanes above the study area, and multiple photographs are combined to cover the whole area. Aerial photographs cover large areas with relatively cheap costs, and can be used for small or large area inventories. For different purposes, e.g. visible light channels measuring red, green and blue (RBG) colour wavelengths or near infrared (NIR) or colour infrared (CIR) channels can be used. CIR is a combination of RGB (or RG) and NIR channels. Pixel size and flight altitude define the resolution and usability of the data. These days, digital aerial images have commonly replaced analogous images, since they give more stable radiometry and resolution, and no scanning of photographs is needed.

Photographing must be timed so that there are no clouds in the sky. Different conditions of lightning and shades depending on the weather and time of the day and year affect the colour range and shade directions in each photograph. Thus suitable correction methods are needed to standardize the photographs of a given area to fit the same conditions. After standardization, inventory data can be produced using human interpretation or in case of digital photographs, using automatized mathematical methods (see e.g. Tuominen and Pekkarinen (2005)).

2.2.4 Aerial laser scanning data

One of the most recent remote sensing data source applied to forest inventory is aerial (or airborne) laser scanning (ALS), which is often referred to as Light Detection And Ranging (LiDAR). LiDAR measurements are mainly used for small area inventories, e.g. in forest management planning.

LiDAR is based on a set of laser pulses transmitted from aeroplane flying above the target area, see Figure (2.1). Measurements are affected by the flight altitude and the angle of the lens of the instrument. Information of the pulses bouncing back from the obstacles is recorded and preprocessed with respect to the measurement conditions, and produce the geographical coordinates and the height of the hitting point augmented with the intensity of the returning pulse echo, see e.g. Wehr and Lohr (1999); Mallet and Bretar (2009) for general information of laser scanning, and Hyyppä et al. (2004) for a summary of its use in forest inventory.

LiDAR systems can be divided into two types, full waveform recording devices (WRD) and discrete-return devices (DRD). In WRD, the complete waveform of each back-scattered pulse can be recorded and then digitized and interpreted in a user controlled manner. The digitized, discrete information can be divided according to the travelling history of the pulses: the first echo pulse that bounces

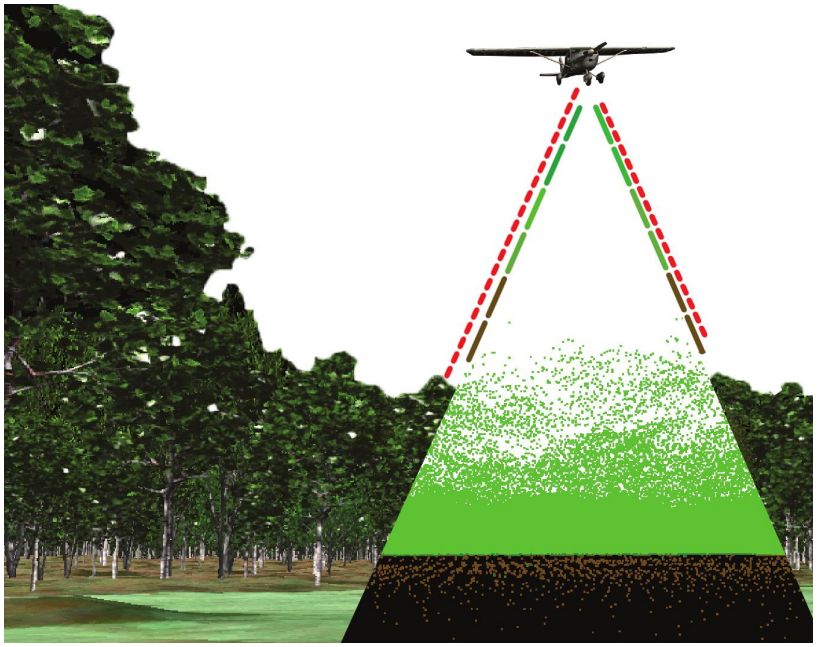

Figure 2.1: Laser scanning from aeroplane.

back usually from the crown of the tree, or from the ground; the echo pulses hitting obstacles between the crown and the ground; and the last echo pulse (ground hit). In DRD, generally only the first and the last, and in special cases the only, pulses are recorded.

Canopy height model (CHM) of the LiDAR measurements is defined as the difference between a digital surface model (DSM) and a digital terrain model (DTM). In practice, it can be calculated by means of first and last pulse echos. The LiDAR-histogram of a given plot (e.g. a round plot with given central coordinates) consists of pulses which bounce back from obstacles within the plot area. In LiDAR, the density of transmitted pulses generally varies from less than 0.5 to more than 10 hits per square meter. Data with dense LiDAR measurements can be utilized to obtain detailed estimates, e.g. estimates of individual trees, while data with lower resolution is generally sufficient for statistical estimates, e.g. total volume of trees within a given area.

LiDAR has definite benefits compared to other remote sensing data with regard to the confidence and objectivity of the data. Unlike for satellite and aerial photographs, the measurements of LiDAR can be performed even in cloudy weather if the flight altitude is below the cloud altitude or even at night, since LiDAR is an active sensor that provides its own energy. The measurements are handled automatically by physical or statistical methods, no human interpretation is included at any point. The histogram of measurements can be attached to spatial ground coordinates of the terrain with high accuracy.

2.2.5 Selection of remote sensing data source

In general, different sources of remote sensing data are utilized for different purposes. In national forest inventories of boreal forests, the estimated areas are large, at least communal level size, and relatively cheap and easily acquired data is needed. Satellite images and aerial photographs are used, giving tolerable estimates of forest inventory stand parameters. For operational use in forest management in Finland, more precise data for smaller size areas is needed. LiDAR combined with aerial photographs has shown to give promising results with tolerable costs. Recently, prices of LiDAR inventory have reduced as the method has seen wider use. Overall, if the inventory area is compact and unscattered, unit price of the inventory becomes cheaper than for a scattered inventory area.

2.3 Remote sensing in global forestry

Remote sensing methods have extended inventory methods to new approaches. In global scale, industrial forest use and management has a less significant role than in the northern countries. Today, remote-sensing data based inventory methods are used not only to management and nationwide inventories, but also to biodiversity monitoring (see. e.g. Goetz et al. (2007) for bird species richness predicted by LiDAR), carbon and biomass estimation (see e.g. Tomppo (2000) for carbon balance estimates using satellite images, Patenaude et al. (2004) for quantifying forest above ground carbon content using LiDAR and Næsset (2004b) for above- and below-ground biomass estimates using LiDAR) and to forest health estimation (see e.g. Solberg et al. (2004)).

An important application of remote sensing based forest inventories is its use in observation of changes in land-cover of tropical forests, e.g. in Brazilian Amazon (INPE, 2005; Asner et al., 2006, 2009) and in French Guiana (Häme et al., 2004; Rauste et al., 2007). High resolution satellite images can be utilized monitoring deforestation in terms of the UN-REDD Programme (United Nations Collaborative initiative on Reducing Emissions from Deforestation and forest Degradation (REDD) in developing countries). Organizations such as Instituto Nacional de Pesquisas Espaciais (INPE) in Brazil provide satellite maps of deforestation over a sequence of years.

CHAPTER III

Mathematical approaches to aerial forest inventory

In order to use multi-source data for forest stand parameter estimation, a suitable mathematical model is needed. There are different approaches to combine the auxiliary data with the field measurements of sample plots. Approaches to extract variables can be divided into two categories: area based approaches (ABA) and the individual tree crown approaches (ITC) as stated e.g. in Breidenbach et al. (2010) or correspondingly, statistical and image-processing based retrieval methods as stated in Hyyppä et al. (2004); Holopainen and Kalliovirta (2006). Individual tree crown approaches are straight-forward approaches to analyze the canopy surface and height estimates of remote sensing data, based e.g. on detection of individual tree location and estimation of crown size from remote sensing images. Area based approaches are based on compartments, plots, or stands consisting of homogeneous plots, i.e. a collection of plots of similar forest type located next to each other. The tree-level information is gathered to area sized entities of histograms or statistical values, and auxiliary remote sensing data is processed as compartment area units. To be usable in the model, the resolution of the remote sensing data must be comparable to unit sizes of the parameters that are estimated. For instance, individual tree level estimation is performed using auxiliary data with individual tree level resolution, i.e. high resolution remote sensing data. Lower resolution remote sensing is generally sufficient for plot-level forest stand parameter estimation, where the auxiliary data is gathered to plot-level units. Another, yet purely theoretical approach, is to recover the relationship between canopy height and forest stand parameters based on assumptions about the single tree crown, the distribution of tree height, and the spatial distribution of tree locations, i.e. discover a physical model connecting the laser scanner data to the forest attributes, see Mehtätalo and Nyblom (2009). Such models could be used to estimate the stand density and distribution of tree heights using observations of canopy height.

3.1 Error estimation

The performance of the mathematical model used is verified by the error of its estimates. Error of the model depends both on the model structure and the auxiliary variables used in it. Analytical error estimation of multi-source inventory results is difficult as it might contain errors from sampling strategies, location of the plots, remote sensing and field work measurement data and the mathematical estimates. See e.g. Kangas and Kangas (1999) for the effect of different error sources on the forest management planning solutions. Aerial data registration error is studied e.g. in Suvanto

et al. (2010), who simulated effect of error in GPS positioning of ALS on forest inventory results, and in McRoberts et al. (2002), who studied the effect of image registration and plot location errors of satellite imagery data on estimates of forest area.

In forest inventory, analytical error estimation is generally replaced with bias and root mean square error, RMSE. For a set of N estimated values $\hat{y}_i,\ i = 1, \ldots, N$, the bias and RMSE are estimated by verifying the estimated values against ground truth values y_i,

$$\mathrm{BIAS} = \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)}{N} \quad \text{and} \quad \mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{N}}. \tag{3.1}$$

Error estimates are often given in relative format, where precision of the estimates is compared to the average ground truth value of the data,

$$\overline{y} = \frac{\sum_{i=1}^{N} y_i}{N}. \tag{3.2}$$

Estimation precision of different areas can be better verified by these relative error estimates, BIAS% and RMSE%:

$$\mathrm{BIAS}\% = \frac{\mathrm{BIAS}}{\overline{y}} \times 100\% \quad \text{and} \quad \mathrm{RMSE}\% = \frac{\mathrm{RMSE}}{\overline{y}} \times 100\%. \tag{3.3}$$

For error estimation purposes, the existing measurement data is divided into two groups: the teaching set and the verification set, which do not overlap. The teaching set is used to estimate model parameters. Error is estimated comparing the ground truth data of the verification set to the estimates derived with a given model using verification set auxiliary data. If the error would be estimated from the teaching set of the model, the results would be unrealistic and over-optimistic. As there is only a limited number of measurements in forest inventories, dividing the set into two groups so that error estimation is reliable, is difficult.

The most realistic approach to error estimation is the leave-one-out method (LOO). In LOO, each measurement of the material is used in error estimation. One measurement at a time is left out from the teaching set to serve as the verification set. The model is prepared separately for each case, using the teaching set of all the measurements except the one left out. Estimates derived for each verification measurement are then used to error estimation. This method is a mathematically sound approach to error estimation and gives a realistic and reliable picture of the true error. For a large amount of data, it is, however, computationally demanding to calculate, especially if mathematical modeling requires any manual work at any stage.

3.2 Estimation of individual trees

A natural approach to analyze forests from remote sensing images is to locate and estimate tree characteristics from their crowns that can be detected from above, i.e. individual tree crown approach (ITC). Individual tree crowns can be estimated from different types of high resolution remote sensing data, see e.g. an early work of Gougeon (1995) for use of one band of one image from airborne multi-detector electro-optical imaging sensor, Brandtberg (1999); Korpela (2003) for use of high resolution aerial images, and Holmgren and Persson (2004); Peuhkurinen et al. (2007) for use of

ALS. The tree crowns can be depicted e.g. from stereo-pairs of large-scale digital photographs or high-pulse-rate laser-scanner images. Aerial photographs are in 2-D form or in 3-D form when stereoscopic photograph coverage is used, LiDAR in 3-D form as the digital terrain and crown models can both be retrieved by laser scanning. Individual trees can be located and their height and crown area estimated using segmentation algorithms. Other stand attributes can be estimated using that information combined with remote sensing data from different sources, see e.g. Hyyppä et al. (2001).

Individual tree-level approaches give relatively good estimates for certain inventory parameters. High resolution LiDAR and CIR- or NIR-images (Holmgren, 2004; Persson et al., 2004; Flewelling, 2006; Koch et al., 2006) and CIR-images (Korpela, 2004) have been utilized for classification of tree species. Several forest stand parameters, such as height and volume, are required for forest management purposes. Individual-tree level stand parameter estimates are relatively accurate, stand level RMSE% of total volume varies from 38% (aerial photographs, Anttila and Lehikoinen (2002)) to 10.5% (high-pulse-rate LiDAR, Hyyppä et al. (2001)). However, the bias of estimates tends to be large, giving systematic estimation error of the stand parameters, approximately 20%-40% depending on the study. Negative bias is explained by the fact that the small trees cannot be depicted from remote sensing data, since they are covered by the tall trees. Also the possibility that automatic segmentation cannot be conducted correctly with sparse data can cause error: either some large individual trees are split into many small ones (negative bias), or vice versa (positive bias). Both segmentation errors cause gross errors and thereby induce bias. In Breidenbach et al. (2010) an approach called "semi-ITC", that overcomes these problems by imputing ground truth data within crown segments from the nearest neighboring segment is proposed. Their analysis using mixed ITC and ABA approach shows to give good, unbiased results, and can thus be used as a showcase for how to use crown segments resulting from ITC algorithms in a forest inventory context.

3.3 Estimation of compartment-based forest stand parameters

Generally most reliable results have been derived from statistical approaches of area, or compartment-based forest stand parameter estimations (area based approach, ABA). Forests are analyzed as compartment-level (i.e. plot or stand level) parameters, which correlate with variables drawn from remote sensing data. Remote sensing auxiliary data covers the whole area of interest, while field work measurements are restricted to a given set of plots. In forest inventory, the dataset size (the number of plots measured) is generally several hundreds, say 400-600. Estimation methods are based on direct models of stand parameters as a function of remote sensing data variables. The most commonly used mathematical models in area based forest inventory are k-neighbours methods and linear regression, which will be discussed in the sections 3.3.2 and 3.3.3.

A suitable mathematical approach is required to derive reliable estimates for forest stand variables from the auxiliary data available. Since there generally exists no physical model between the multi-source auxiliary data and forest stand parameters, a "black-box" model is needed. That is, the data is modelled as a function of independent data (input vector) and forest stand parameter data (response data, target vector), and the parameters of the model are defined using the known dataset.

3.3.1 Variables

In compartment based estimation, data is handled so that it is in uniform format. Instead of tree-level information, or remote sensing data pixels to classify, field work measurements and remote

sensing data variables are given in plot-level entities. Plot-level information of field measurements is given as histograms (single-tree data) or as statistical values of the trees inside a plot area, or both of them. This data serves as the dependent data, i.e. forest stand parameters, in the estimation procedure. Single-tree data consists e.g. of height, stem number and volume histograms of the trees in the plot. In industrial forest use, statistics of the single-tree data, such as the median tree height and diameter, the number of stems per hectare and the mean volume per hectare, are often used as forest stand parameters. All the measurements can be handled at species specific level or as total values containing all the species.

Remote sensing data consist of measurements located in the plot area: e.g. aerial photograph pixels and LiDAR measurements within the plot area boundaries. Independent variables for each plot are derived from these measurements. The number of independent variables and their transformations (powers, logarithms, etc.) drawn from the data may be large, several dozen or even hundreds.

Digital aerial photographs may be utilized for inventory purposes as aerial picture variables or classified pixels and areas. Variables derived directly from aerial photographs are e.g. mean and standard deviation of digital numbers in a given window for different colours, and variables derived by visual interpretation of photographs include estimates such as land use class, dominant tree species, proportion of deciduous tree species, site type class, mean height of trees and relative density of forest growing stock (see e.g. Poso et al. (1999); Packalén and Maltamo (2007)). Satellite image data is generally utilized in the form of intensity values on some number of channels.

LiDAR measurements are gathered in histograms of measurements which are located to a given plot area according to the spatial coordinates. There are generally four types of measurements: first and last pulse height and intensity measurements. Variables for modeling are drawn from the histograms according to different statistical approaches, e.g. mean and standard deviation of measurements, percentile part of the cumulative sum of ordered measurements and percentile part of measurements under given level (Næsset, 2002, 2004c; Hyyppä et al., 2004; Packalén and Maltamo, 2007).

3.3.2 k-nearest neighbour and k-most similar neighbour model estimation

A simple approach for plot level estimates would be to classify the plots into homogeneous strata, i.e. plots containing approximately equal values of different forest stand parameters, and to estimate the forest stand variables of interest of each plot in the stratum as averages of the measured field plots of that stratum. This approach, however, ignores the variation of plot characteristics, and the estimates are coarse. Estimation methods with a similar idea are the k-nearest neighbour (k-NN) method and its derivative, the k-most similar neighbour (k-MSN) method, see e.g. Kilkki and Päivinen (1987); Tomppo (1991, 1993); Moeur and Stage (1995); Korhonen and Kangas (1997) for early attempts to use these methods in forest inventory. These methods are based on searching plots similar to the one that is being estimated. Forest stand parameter estimates for the new plots, target set plots, are averages of the chosen neighbour forest stand parameters or histograms from the reference set. For instance, typically 100-400 characters concerning e.g. site, volume and increment of growing stock, are estimated in each plot of Finnish national inventories. Such a large number of inventory forest stand variables is hard to estimate separately, and thus k-NN and k-MSN methods are found to be applicable.

In the k-NN and k-MSN methods, k nearest neighbours are selected for each target set plot from the set of N reference plots available. The distance d_{ij} between plots i and j is defined in given metrics and feature space (Maltamo and Kangas, 1998; Poso et al., 1999). The feature space consists of

variable vectors $\mathbf{x}_i$ from different data sources, e.g. earlier inventory stand records or the features of the remote sensing data such as satellite image spectral channels, aerial photograph interpretations or aerial laser scanning measurements, or of their combination.

In the k-NN method, the distance between plots is given as a weighted linear difference model:

$$d_{ij} = \sum_{m=1}^{M} c_m |\mathbf{x}_{im} - \mathbf{x}_{jm}|, \tag{3.4}$$

where $\mathbf{x}_{im}$ is the variable m of plot i, c_m the weight of the variable and M the number of variables. Tokola et al. (1996) and Holmström (2002) define the distance of neighbours by forest stand variables using their regression estimates derived from auxiliary data features. Tomppo and Halme (2004) and Tomppo et al. (2009) use genetic algorithm to estimate weights for different variables in the distance equation. Restriction of geographical distance both in horizontal and vertical directions between the neighbouring plots has been shown to be advantageous, reducing the bias in estimates (Katila and Tomppo, 2001). Taskinen and Heikkinen (2004) use satellite image channels and geographical coordinates to estimate tree volume data and main site class data with nonparametric Bayesian partition model. The model they use can be considered as a Bayesian counterpart of k-NN method. An advantage of the model is that it provides model-based assessment of pixel level prediction error. k-NN can be used to estimate different forest stand parameters, e.g. total volume of the trees in the plot, combined with species composition classes (Mcroberts, 2009), and categorical forest variables such as site fertility and tree species dominance of a site (Tomppo et al., 2009).

Distance definition in the k-MSN method is based on a regression type analysis of the auxiliary data, canonical correlation analysis, CCA (Moeur and Stage, 1995). In CCA, correlation between two linear models is maximized. The linear models of the $N \times M$ feature variable matrix $\mathbf{X}$ drawn from the auxiliary data and the linear model of a $N \times P$ matrix of P forest stand parameters $\mathbf{Y}$ are used:

$$\mathbf{u}_r = \mathbf{X}\mathbf{w}_{xr}, \quad \mathbf{v}_r = \mathbf{Y}\mathbf{w}_{yr}. \tag{3.5}$$

Here $\mathbf{w}_{xr}$ is the rth column of linear auxiliary variable weight matrix, $\mathbf{w}_{yr}$ the rth column of stand parameter weight matrix. The maximization of the correlation of these linear models is performed using eigenvector-analysis. The $R = \min(M, P)$ largest eigenvalues r with corresponding eigenvectors are used to estimate the distance d_{ij} between different plots i and j:

$$d_{ij}^2 = (\mathbf{X}_i - \mathbf{X}_j)\, \mathbf{\Gamma}\mathbf{\Lambda}\mathbf{\Gamma}^{\boldsymbol{T}} (\mathbf{X}_i - \mathbf{X}_j)^T, \tag{3.6}$$

where $\mathbf{X}_i$ is the $1 \times M$ feature variable vector of plot i, $\mathbf{\Gamma}$ is the $M \times R$ matrix of canonical coefficients (eigenvectors) and $\mathbf{\Lambda}$ is the $R \times R$ diagonal matrix of canonical correlations (eigenvalues). With CCA, the whole forest stand parameter space is projected to a space of dependent variables (remote sensing auxiliary variables). Distance function can thus be estimated as a function of auxiliary data. k-MSN has been widely used in modern forest inventory, see e.g. Muinonen et al. (2001); Maltamo et al. (2006); Packalén and Maltamo (2006, 2007); Peuhkurinen et al. (2008). For instance, Packalén and Maltamo (2007) use LiDAR and digital aerial photograph variables to estimate total and species specific volumes (pine, spruce and deciduous trees). The estimates are derived using three variables of species specific volumes and 42 variables of remote sensing data with their logarithms, square roots, powers and inversion in CCA.

In both methods, the number of used neighbours, k, varies typically between 3 and 20, and it is defined by a cross-verification procedure using the reference set. Estimate for the new plot forest stand parameters is the average of the corresponding stand parameters of the k nearest neighbours. In most approaches, weighted average where the weight is defined by the distance values of the neighbours, have been used:

$$\mathbf{y}_i = \sum_{j=1}^{k} \left(\frac{d_{il_j}^{-s}}{\sum_{j=1}^{k} d_{il_j}^{-s}} \right) \mathbf{y}_{l_j}, \tag{3.7}$$

where $l_j,\ j = 1, \ldots, k$, is the set of k nearest or most similar neighbours defined by distance d_{ij} and s is a die-off parameter. The weight is largest for the plot with the smallest distance, and vice versa. The total sum of the k weights equals to one.

k-MSN method is a nonparametric-method. However, there are parameters that must be tuned: the number of neighbours k and the die-off parameter s. In the literature, optimal values are searched manually with partly heuristics cross-validation approaches, or by heavy algorithms which go through different combinations and choose the best result by testing, see e.g. Packalén and Maltamo (2007). Estimates for different forest stand variables with some RMSE and bias are used for comparison. To choose the best solution, the user must define a multi-criteria cost function. In the literature, the parameters are searched separately for each inventory study, e.g. three in LeMay and Temesgen (2005), five in Packalén and Maltamo (2007), etc.

LeMay and Temesgen (2005) verified results derived with different distance estimates: Euclidean distance, weighted distance of k-NN, i.e. equation (3.4), and distance of k-MSN, i.e. equation (3.6), together with different forest stand parameter estimates: Only one neighbour, average of the three neighbours or weighted average of the three neighbours, equation (3.7). In their study, k-MSN showed to perform best, and no large gain was noted in using the average of three neighbours rather than a single neighbour. Using the k-NN method with different approaches utilizing satellite and aerial image data, the RMSE% of the estimates of total volume in plot-level is at best approximately 30-70% (Poso et al., 1999; Holmström, 2002). For some studies, bias has been a problem. Estimates derived with the k-MSN method utilizing LiDAR and aerial photographs are rather accurate, plot-level total volume RMSE% being approximately 20% and bias close to zero (Packalén and Maltamo, 2007).

For the k-neighbours methods, the size of dataset must be large. The methods are interpolation methods, where each plot that is estimated must be an inner plot in terms of forest stand parameter distribution. Estimates of out-lying plots are prone to bias. If the scale of forest stand parameter variation of a site is large, a dense set of field sample plots is needed to guarantee the existence of close neighbours, see e.g. LeMay and Temesgen (2005) for tests with different reference dataset sizes.

3.3.3 Regression models

A common approach to solve black-box models is linear regression. Linear regression is a popular method thanks to the simplicity of the equation and its solution, and to its capability to give an analogous estimate also to the error of the prediction. Regression models are based on the linear equation

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}, \tag{3.8}$$

where $\mathbf{y}$ is the $N \times 1$ vector of dependent variables, $\mathbf{X}$ the $N \times M$ matrix of independent variables, containing the constant term 1, $\mathbf{w}$ the $M \times 1$ weight, or regression parameter vector and $\boldsymbol{\varepsilon}$ the $N \times 1$ vector of errors, which is assumed to be component-wise normally distributed with zero mean and variance σ^2. Ordinary least squares estimates (OLS) give an estimate for the weight vector

$$\hat{\mathbf{w}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}. \tag{3.9}$$

Regression models consisting of independent variables from different data sources have been widely used in forest multi-source inventories, see e.g. Lappi (1993); Næsset (1997); Means et al. (2000); Næsset and Bjerknes (2001); Holmgren and Jonsson (2004); Næsset (2004c); Suvanto et al. (2005). Linear and square root or logaritmic transformations of equations are used to predict the forest stand parameters of plots or stands. Especially in approaches using LiDAR-data, regression has shown to be a compatible method when compared to other approaches, such as k-MSN.

A drawback of the regression method is that different forest stand parameters are estimated separately, and the information of the correlation between different parameters and residuals of estimates is missed. Multivariate regression method can be used to estimate multiple forest stand parameters at once together with a multinormal estimate of their residual covariance matrix. However, it does not use the residual covariance in the model. Some regression methods take the residual correlation into account, e.g. seemingly unrelated regression (SUR). It gives realistic predictions for multivariate cases, but possible problems in SUR are the fact that the residual covariance is assumed multinormal, which may be a false assumption in real world problems, and the possibility of local optima. See e.g. Mardia et al. (1980) for the basic assumptions for these multivariate regression approaches. However, the estimates of any forest stand parameter derived with any method possible are only as good as the data is, that is, if there is no correlation between independent and dependent data, accurate estimation of the dependent data is impossible. For this reason, different approaches often result in approximately equally accurate estimates.

Estimation of values, which are not normally distributed or are close to zero but strictly positive, is somewhat cumbersome using regression. In forest inventory, such problems arise especially in estimation of species specific forest stand parameters. Linear estimates are not allowed to be negative and the feature of total volume being the sum of species specific volumes must be consistently adhered to. In the k-MSN method this feature is automatic, in regression the estimates need to be post-processed. To avoid negative estimate values, forest stand parameter transformations based on logarithms can be used.

The strength of the regression method lies in the feature that it is an extrapolation method, where the estimates are accurate as long as the linearity remains, independent of the location in the forest stand parameter distribution space. To estimate new plot forest stand parameters, the linear model must be established correctly. If the multivariate forest stand parameter distribution of reference plots is sparse, the distance between nearest neighbours in the k-neighbours methods may be large in terms of forest stand parameters and the estimate derived by weighted average of the k neighbours is prone to be biased. Also at the edges of the forest stand parameter space the estimates may be biased since the estimate is an average of the neighbours only from the inner points of the space. These problems can partly be circumvented by using a large number of measurements representing well the total variation of forest stand parameters, see e.g. LeMay and Temesgen (2005). The regression approach does not suffer from this feature, and even a small number of field sample plots, correctly representing the full feature space, is sufficient to establish accurate models if the correlation between independent and dependent variables is large.

3.3.4 Variable selection

The aim of all the mathematical approaches in forest compartment-level inventory is to estimate forest stand parameters using the given set of independent variables. Independent data variables correlate in different scales with the dependent parameters. If the relationship is strong, the variable is likely to explain the parameter well, an vice versa. For problems with small dataset size compared to the number of variables, a phenomenon called over-learning, or over-fitting, may occur. That is, variables explain the error or noise of the model instead of the underlying relationship. Over-fitting is likely to occur when a model is excessively complex, e.g. having too many variables compared to the amount of data. In such models, variables with weak correlation are not only unnecessary in the model, but harmful, since the model tends to use those and give them too large a weight to explain the noise. As a consequence, predictive performance of the mathematical model is poor, since the given weights of the variables are misleading. Also internal correlation between the variables is likely to occur since independent variables are to a large degree derived from the same data, only with different approaches (multicollinearity). Such data may lead to poor estimates, since different variables tend to explain not only the response, but also each other, resulting in exaggerated fluctuations to predictions. Also the input vector of multicollinear independent data is likely to be singular, which causes problems to many mathematical linear approaches such as OLS and CCA.

A common feature in all approaches to solve compartment-based estimates in forest inventory is the need to evaluate the feasibility of variables from different sources in terms of prediction of forest stand parameters. The variable selection in k-NN, k-MSN or regression methods is usually performed manually for each site, or by automatized algorithms which search through a large number of different variable subset combinations. Criteria for the selection, and the number of approved variables need to be established. Common approaches utilized in forest inventory are e.g. step-wise regression used e.g. in Næsset (2002), model definition with cross-validation which can be assumed to be used in many studies where the model is defined beforehand and used set of variables are just given, e.g. LeMay and Temesgen (2005), genetic algorithm for k-NN used in Tomppo and Halme (2004) and cross-validation based predictor selection algorithm for k-MSN used in Packalén and Maltamo (2007). In Næsset (2002) a criterion to avoid serious collinearity of the variables was added to the step-wise regression algorithm. Another approach to avoid over-learning is e.g. the leaps and bounds algorithm (Furnival and Wilson, 1974). To circumvent problems of collinearity, methods such as James-Stein multiple regression (Efron and Morris, 1975), ridge regression (Hoerl and Kennard, 1970) or shrinking (Copas, 1983) can be used. To define the number of variables e.g. Akaike's information criteria, AIC (Hall et al., 2005) can be utilized. For Bayesian approaches, e.g. the Bayesian information criterion (BIC) can be used to regularization, or a combination of AIC and BIC, deviance information criterion (DIC) can be used in Markov Chain Monte Carlo simulations (Spiegelhalter et al., 2002). DIC is a criterion which favors a good fit of the model, but also small number of parameters. It has been used e.g. in a multivariate spatial process discussed by Finley et al. (2008). Other Bayesian variable selection methods have been discussed and verified e.g. in O'Hara and Sillanpää (2009) and in references therein.

Selection of suitable algorithms depends on the modelling task and mathematical model used. Variable selection performance is generally estimated by cross-validation, either dividing the material to model the teaching set and the verification set, or utilizing the leave-one-out procedure. Overall, variable selection is strongly related to a number of methods, e.g. regularization, early stopping, Bayesian priors on parameters and model comparison, and can be seen as a regularization technique for ill-posed estimation problems.

As the forest circumstances and characteristics vary greatly, it is highly unlikely that the model parameters designed to one inventory area would be appropriate to another area. Suvanto et al. (2005) discussed the demand of inclusive estimation models, which would cover the whole area of Finland. Regression models with defined parameters predicting the forest stand parameters of distinct spatial areas of certain parts of Finland were found feasible. However, the differences in forest types are large, and it is not likely that the mathematical models of one area would consistently give sufficient estimates to other, different areas. Also the differences in remote sensing methods and equipment are likely to produce inaccuracies to such nationwide inclusive models.

CHAPTER IV

Objectives of the thesis

The main goal of the thesis is to introduce cost-efficient, automated estimation procedures to forest inventory that could be easily adapted to inventorying on a new site. The results of the thesis are divided into two approaches that can be applied successively. The first approach is to introduce a new automatic and adaptive approach to variable selection in forest inventory regression methods. The second approach is the utilization of formerly measured areas, databases, in forest inventory with the aim of reducing the field sample measurement work and costs. The goal is to produce precise and unbiased estimates while keeping expensive field measurements in the new site to a minimum.

All the estimates in the publications included in the thesis are based on sparse Bayesian regression (SBR). SBR is a form of the relevance vector machine (RVM) approach which has been introduced for kernel-based linear equation estimation by Tipping (2001). Publication **I** introduces this new approach to forest inventory and computes test results derived with forest inventory data utilizing LiDAR-measurements as auxiliary data. The results are compared to results of other linear regression methods. SBR automates the estimation procedure by selecting linear model variables from a set of candidate variables using a Bayesian prior distribution for variable weights.

Publications **II** and **III** introduce an algorithm which utilizes formerly measured databases for new site estimation. New site LiDAR is used to select a small amount of calibration plots (50-70) that represent forest stand parameter distributions of the site. Field measurements and LiDAR histograms of calibration plots are used to calibrate the database LiDAR-histograms. Database plots fitting the calibration set distributions are selected to form SBR estimates for the new site. The method is first introduced in publication **II** to estimate total forest stand variables. Three databases are utilized and five forest stand parameters are estimated using a calibration set from the new site and selected plots from the calibrated databases. Estimation results are verified to optimal estimates derived with a high number of field plots (400-600) in the new site and to estimates derived with only the calibration plots.

Publication **III** expands the method to using a larger number of databases and to estimation of species specific forest stand variables. In addition to LiDAR measurements, digitized aerial photographs with subjective interpretation are used as auxiliary data. Species specific forest stand information is taken into account in all steps of the database utilization algorithm. In the case of a large number of databases, the number of selected plots from databases may be much larger than the number of measured calibration plots from the new site. The distribution of the forest stand charac-

teristics of selected database plots may also differ from the calibration set distribution. Publication **III** introduces methods to avoid the bias caused by this distortion.

The performance of the expanded, automated method of forest stand parameter estimation procedure utilizing database information is verified in publication **III**. Seven spatially different sites and twenty forest stand parameters (total and species specific forest stand parameters) are used in a cross-verification procedure where one site at a time serves as the new site, and the others as databases. For each site as the new site, 50 repetitions of the procedure with randomly selected calibration sets are calculated. Results of the procedure are verified to the optimal results and to the results derived with only the selected calibration plots.

CHAPTER V

Bayesian regression approach for variable selection

In plot-based remote sensing forest inventory, there are multiple layers of data. The data consist of the field measurements supplemented with remote sensing data, which generally is assembled to field measurement area size entities and transformed to plot level scalar variables. For example, utilizing aerial laser scanning with discrete-return devices, auxiliary data is given as four histograms (first and last echo height and intensity of scanning measurements) covering the areas of interest. Response data, forest stand variables, are given at plot-level entities. To integrate the response and LiDAR data, histogram information is gathered to plot-size units by some statistical models.

The task of estimating forest stand parameters from given auxiliary data with no physical model attached to the phenomenon, is most often solved with linear regression. Since the level of knowledge is limited to the existing data, it is favourable to keep the complexity of the chosen mathematical model as simple as possible. In regression, this equals to minimizing the number of used variables. However, deleting relevant variables, the estimate accuracy diminishes. Sparse Bayesian regression is a method to search optimal combination of variables that are required to accurate estimates.

5.1 Sparse Bayesian regression in forest inventory

Sparse Bayesian regression (SBR) is based on probabilistic regularization approaches, see e.g. MacKay (1992, 1999), and for kernel based approaches, see e.g. Tipping (2001, 2004). Linear regression is stated in probability function form, enabling Bayesian approach to variable selection. Parameters of the linear equation with normally distributed errors are defined in fully probabilistic framework, where prior information of the parameter behaviour moderates the regression model. In hierarchical Bayesian terminology, prior distributions with hyperparameters are given over the parameters. The hyperparameters are also estimated in the process.

The likelihood function form of linear regression equation (3.8) is

$$p(\mathbf{y}|\mathbf{w},\sigma^2) = \prod_{i=1}^{N} \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-||y_i - \mathbf{X}_i\mathbf{w}||^2/2\sigma^2\right) \tag{5.1}$$

$$= \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)^T \beta \left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)/2\right) \tag{5.2}$$

which is a normally distributed (Gaussian) function over the response vector $\mathbf{y}$, $N(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2)$, with mean value $\mathbf{X}\mathbf{w}$ and variance σ^2. Here $\beta = \sigma^{-2}$ and weights $w_m,\ m = 1, \ldots, M$, are the parameters of the distribution.

A complexity penalty term is added to the model to constrain the weight parameters. In the hierarchical Bayesian framework, an explicit prior probability distribution is defined over the parameters. A zero mean normal distribution over each weight w_m,

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} N(w_m|0, \alpha_m^{-1}) = N(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \tag{5.3}$$

is defined to shrink the weights close to zero. Importantly, an individual hyperparameter, α_m, is associated independently with every weight m, defining the strength of the prior. Here $\mathbf{A}$ is an $M \times M$ diagonal matrix of the $M \times 1$ hyperparameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^T$. If a hyperparameter is large, $\alpha_m \to \infty$, the variance of the prior distribution $p(w_m|0, \alpha_m^{-1})$ goes to zero, forcing the weight to a peak with no variance around zero and thus rendering the variable $\mathbf{X}_m$ to be insignificant in the model. For small hyperparameter values, $\alpha_m \to 0$, the variance is large and the weight prior distribution almost flat, allowing nonzero weight for variable m. The hyperparameters are responsible for the sparsity of the model, "deleting" unnecessary variables out by zero weights.

To complete the hierarchical Bayesian formulation, hyperpriors are defined over the variance parameter of the hierarchical prior, α_m, and over variance parameter of the likelihood, β. Suitable priors for such scale parameters are Gamma distributions. More discussion about prior distributions for variance parameters can be found e.g. in Gelman et al. (2004); Gelman (2006). If the parameters of Gamma distribution are given small values, the prior converts to non-informative. With zero parameters, the hyperprior becomes uniform over a logarithmic scale and the effect of hyperpriors is eliminated in the model. Thus the hyperpriors are left out from the mathematical derivation of the sparse regression method. A convenient consequence of the use of such priors is also that the predictions are independent of the linear scaling of the measurement target vectors values, $\mathbf{y}$, and variables $\mathbf{X}$, that is, scale-invariance.

Bayesian inference is carried out by computing the posterior distribution over all unknowns. The posterior probability distribution of weight parameters conditioned on the teaching set data is given by Bayes' rule

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = \frac{p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)}, \tag{5.4}$$

where $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = \int p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w}$ is called evidence. Being a product of normal distributions, the posterior can also be stated as normal distribution,

$$p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = p(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{5.5}$$

where posterior mean and covariance are analytically solved from the exponents,

$$\boldsymbol{\Sigma} = \left(\beta \mathbf{X}^T\mathbf{X} + \mathbf{A}\right)^{-1} \tag{5.6}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}\beta\mathbf{X}^T\mathbf{y}. \tag{5.7}$$

With zero hyperparameters, $\alpha_m = 0,\ \forall m$, posterior mean and covariance equal to OLS solution giving unbiased parameters. A constant hyperparameter, $\alpha_m = \alpha,\ \forall m$, modifies them to ridge

regression (Hoerl and Kennard, 1970; Goldstein and Smith, 1974), or to Tikhonov regularization with Tikhonov matrix $\alpha \boldsymbol{I}$. The goal of ridge regression is to circumvent the problem of collinearity of independent variables $\mathbf{X}$. Even though the estimates are somewhat biased, the variance of ridge regression parameters has shown to be smaller than that of OLS. Also the predictions of ridge regression have been shown to be more accurate with nearly collinear independent variables. Since the estimation accuracy is a sum of bias and variance of errors, such an approach will give desirable results, being an illustration of the bias-variance trade-off issue. A similar approach is also at the heart of the SBR.

However, in SBR, each regulator parameter α_m is defined individually, modifying also the significance of each parameter m in the model. Parameters and hyperparameters are estimated with a type-II maximum likelihood method (MacKay, 1992; Tipping, 2001), where the first step is to integrate over the analytically solved parameters $\mathbf{w}$, and then maximize evidence $p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$ over the hyperparameters. The evidence can be stated analytically by the normal distribution,

$$p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = N(\mathbf{y}|\mathbf{0}, \mathbf{C}), \tag{5.8}$$

where $\mathbf{C} = \sigma^2\mathbf{I} + \mathbf{X}^T\mathbf{A}\mathbf{X}$. Differentiation of the log-likelihood of this distribution with respect to $\log(\alpha_m)$ and $\log(\beta)$ yields the maximum likelihood point estimates to be solved. The solution is not obtained in a closed form, and thus the parameters are based on an iterative procedure, where the updates for ML-estimates are

$$\alpha_m^{\text{new}} = \frac{\gamma_m}{\mu_m^2} \tag{5.9}$$

and

$$(\sigma^2)^{\text{new}} = \frac{||\mathbf{y} - \mathbf{X}\mathbf{w}||^2}{N - \sum_m(\gamma_m)}, \tag{5.10}$$

where $\gamma_m = 1 - \alpha_m\Sigma_{mm}$, Σ_{mm} denoting the m:th diagonal element of $\boldsymbol{\Sigma}$.

Giving some suitable small initial values for the scale parameters α_m and β (allowing large deviation for the weight parameters and model precision), the iterative process is the following: first solve analytically $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, then re-estimate α_m and $\beta = \sigma^{-2}$. The solution procedure is fast, the solution converges to stable set of elements α_m within some seconds and a few hundred iterations are sufficient even for models with several hundred datapoints and tens of variables.

Within the iterations, as hyperparameter values become large, the covariance of the weight posterior (5.6) becomes sparse. Large diagonal elements in $\mathbf{A}$ cause the corresponding rows and columns in the covariance matrix to approach zero. Thus the corresponding weight posterior mean goes to zero and the variable is "deleted".

Overall, the variable selection procedure is automated, both in terms of variable subset contents and size determination. The precision of the model is maximized by the likelihood (5.2) and the weight divergence from zero is minimized by maximizing the prior distribution (5.3). In other words, model complexity is minimized while demanding sufficiently well explained data. The level of model accuracy depends on the dataset size N, which emphasizes the likelihood importance compared to the effect of prior distributions of each weight parameter. The ratio of importance affects the number of variables needed, circumventing the problems concerning over-learning without any cumbersome cross-verification procedures. The idea is similar to that of automatic relevance detection (ARD),

which has been used in neural networks (Mackay, 1994; Neal, 1996), and also tested in forest inventory to classify forest scenes (Vehtari et al., 1998). Also the deviance information criterion (DIC) is based on similar approach: the model fit is favored while the number of used parameters is penalized. This approach has been used in a Bayesian multivariate spatial process model for prediction of forest attributes by Finley et al. (2008).

The weight parameters are estimated as normal distributions in SBR. To predict new values with given auxiliary variable $\mathbf{x}_*$, a marginal likelihood is given:

$$p(y_*|\mathbf{y}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}) = \int p(y_*|\mathbf{w}, \sigma^2_{\mathrm{MP}}) p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}) d\mathbf{w} = N(y_*|\hat{y}, \hat{\sigma}^2), \tag{5.11}$$

where the normality of the distributions convert the integral to a normal distribution with mean and variance

$$\hat{y} = \mathbf{x}_* \boldsymbol{\mu}, \tag{5.12}$$

$$\hat{\sigma}^2 = \sigma^2_{MP} + \mathbf{x}_* \boldsymbol{\Sigma} \mathbf{x}_*^T. \tag{5.13}$$

The mean and variance define the estimated value and its precision using the given model. The variance depends both on the model variance using the teaching set, and on the weight parameter covariance structure.

5.2 Results of SBR verification

SBR is implemented for forest inventory estimations in publication **I**. SBR is compared to other regression methods, OLS and SUR, which are performed using cross-verification to define optimal forest stand parameter conversion and selection of independent variables. A site consisting of 472 plots with field measurements divided into 67 homogeneous stands was used to verify the models. Five forest stand parameters, diameter of basal area median tree ($\mathbf{y_1} = \mathrm{dgM}$), height of basal area median tree ($\mathbf{y_2} = \mathrm{hgM}$), number of stems per hectare ($\mathbf{y_3} = \mathrm{N}$), breast height basal area per hectare ($\mathbf{y_4} = \mathrm{G}$) and total volume per hectare ($\mathbf{y_5} = \mathrm{V}$) were estimated. Independent variables consisted of a constant term complemented with 27 candidate variables which were drawn from the LiDAR histograms of first and last pulse heights and intensities. Leave-one-stand-out method (LOSO) was used to verify the performance of different methods. In LOSO repetitions, each homogeneous stand and plots in it are left out at a time, and the rest of the plots serve as the model teaching set. The plots in the stand that is left out are the verification set. Total error is calculated for the full set of plots where model estimates are derived by the LOSO repetitions.

SBR is based on distributions: model weights are mean values of the weight distributions, which implicitly states that also the forest stand parameter estimates are distributions. Forest stand parameters extracted from the data include errors from different sources, and should thus be considered as random numbers, not as fixed values. In publication **I**, estimated weight distributions were shown to be relatively robust with respect to the different teaching sets of LOSO-algorithm. Using SBR, 6-16 variables were selected to the model depending on the forest stand parameter in each repetition of LOSO. Some deviation in the variable set size and composition occurred also depending on the teaching set data of each LOSO-repetition. The new method is shown to be fast, for each test the solution converged within some seconds. The accuracy of the estimates is competitive with that of OLS and SUR.

CHAPTER VI

Databases

Teaching set size together with its quality affect the estimation accuracy. Teaching set data is required to represent well the heterogeneity of the area characteristics in terms of response parameters that are measured, otherwise the estimates are prone to be biased. In forest inventory, generally several hundreds of plots with field measurements are required for each inventory site to guarantee accurate estimates. This is costly, and thus other approaches based on sparser teaching sets containing a fraction of the standard set of sample plot measurements have been tested.

In publication **I**, the total forest stand parameter estimates dgM, hgM, N, G and V were verified using a sparse set of teaching stands, i.e. using a smaller number of field sample plots. To maintain the representativeness of all forest stand qualities of the area, the teaching set was randomly selected so that all the development classes are sufficiently represented in it and that the forest stand parameter variability is large enough. Using SBR, the estimates remained tolerably good even with models defined with 9 teaching set stands (approximately 63 plots). Maltamo et al. (2009) continued to verify different sampling strategies for field training plot selection. They verified the estimate accuracies using different number of training plots (21 - 181) with different sampling strategies: random sampling, random sampling within pre-stratification according to forest type, selection of plots according to geographical location and selection of plots based on properties of the LiDAR data. LiDAR data based criteria were based on the 90% height points (see Næsset (2004c)) and the proportion of ground echoes versus canopy echoes using a threshold value of 2m. Maltamo et al. (2009) showed that LiDAR based selection provides generally the most accurate results, especially for volume estimates and that plot sample size of approximately 50 is enough to give reasonably accurate results. However, they did not discuss the robustness, i.e. deviation of RMSE and bias of the estimation errors depending on the random selections. The variable selection and other parameter tuning of the k-MSN method used for estimation were based on cross-validation with the entire training data (181 plots), so the possible variation in variable selection and number of neighbours when training data size is small, is omitted. Thus the estimation accuracy can be expected to be over-optimistic for small training data sizes.

Magnussen et al. (2010) have tested model estimates, treated as functions of model parameters that arise from different laser pulse densities, in a linear regression model with the assumption of random predictors. The authors conclude that total forest stand characteristics in a boreal forest are relatively insensitive to small variations in pulse density, but in practice the variation in LiDAR extracted variables due to model parameter invoked randomness should be addressed.

In publications **II** and **III** a similar sparse teaching set approach to the new site is sustained. One aim of the studies is to derive estimates which are robust against the selection of the sparse teaching set. In these publications, the selection criterion to the new site field sample plots is modified to include the information obtained from LiDAR measurements. Variation of 85% first pulse height points of LiDAR is required to cover well its variability in the new site plots. This criterion alone is used in publication **III** which contains sites without information of development class. In publication **II**, the criterion is combined with the criterion of development class variability.

The idea of reducing the size of teaching set gathered from the new inventory site is taken further to consider the use of formerly measured sites, databases, in publications **II** and **III**. New site forest stand parameters are estimated measuring only 50-100 plots in it. These plots are called calibration plots (native plots). Additional information is achieved from other, formerly measured sites (alien plots). In these studies, auxiliary remote sensing data of the old and new sites, namely LiDAR, is used to estimate new site parameters. In publication **III** also two variables drawn from digital aerial images are used. These variables are defined in a similar manner by subjective classification for each site used.

LiDAR measurement quality of the new site must be comparable to the LiDAR measurements of databases. LiDAR measurement instrument and flying altitudes of the measuring flight may vary between different sites, leading to different scales of measurements. The histograms of different sites need to be calibrated before using them together. Another issue that requires attention in database utilization, is the forest characteristics of databases. Different sites contain different qualities of forests (e.g. tree species composition, development classes and annual heat sums vary), and variability even within one site may be large. Use of other site plot information requires coincidence of forest characteristics between the new site plot characteristics and database plots. For instance, estimation of a young forest with model designated for old forests may lead to biased values. Thus, only plots similar to the new site characteristics should be attached to the new site model definition.

Use of databases has been studied quite little in earlier forest inventory publications. Næsset et al. (2005) used LiDAR height measurements and inventories from two different sites located relatively close to each other in geographical coordinates. These sites were used to predict six biophysical stand properties using different regression models: OLS and SUR with variable selection and partial least squares regression (PLS). The effect of two separate data sources was included to regression models with dummy variables. Three different stratum specific prediction models were used to predict the forest stand parameters using either only the new inventory area native information or using also the alien information from the other site, the database. LiDAR height histograms and different site sample plot sets were used without any manipulation, and since the two datasets fitted well in these contexts, the results were promising. Only in the case of the mean height model, the use of the database was unprofitable in terms of dummy variable significance. The combined models using the new site and the database gave estimates with better or equal accuracy than the new site model alone. With the given teaching set of 233 plots (133 in site A, 100 in site B) and verification set of 115 plots (61 in site A, 54 in site B), none of the three regression techniques was superior to the others. However, only one test composition with fixed teaching and verification sets was performed, and the general effect of the method with e.g. different sizes of new site teaching sets compared to the database teaching sets were not tested.

Suvanto and Maltamo (2010) tested different regression model based algorithms (standard OLS and mixed estimation) to predict six new target area forest stand parameters with use of one database. They used LiDAR height histograms and different numbers of randomly chosen target area (new

site) sample plots (10-212 plots) combined with 472 database plots. No calibration of LiDAR histograms and no plot selection based to new site forest stand parameter distribution were used. They verified the results to cases where only the new site sample plots were used in the teaching set of the model, either with variable selection based on teaching set plot information, or with variable selection performed earlier with the database information. They tested each sample plot size with 100 repetitions of different, randomly selected sample plot sets, and gave the results as averages of the repetitions. When the plot number was at least 50-120, the best results were obtained using the model with only randomly selected new site teaching set plots and variable selection based on these plots. This result confirms the results given earlier e.g. in publication **I** and Maltamo et al. (2009). For sample plot set sizes less than that, the use of a database improved the RMSE%. However, BIAS% was a problem with every database based model, depending on the weight the database was given in the model. This problem most probably originates from the differences in the forest stand parameter distributions in the two separate areas. Also the differences in the LiDAR scanning equipment may have some influence to the results.

6.1 LiDAR histogram calibration

LiDAR histograms of first and last pulse heights and intensities vary depending not only on forest and ground characteristics, but also on the measurement instrument, footprint size and flying conditions. The flight altitude has been shown to have some effect on dominant tree species height estimates and on LiDAR height histogram scale, but overall the estimates have been shown to be quite robust against variability in these attributes (Næsset, 2004a; Yu et al., 2004; Næsset, 2009).

The users of the estimation process often only know the numerical values of measurements and their coordinates, and the effect of the measurement equipment settings and flying altitude on laser pulse measurements is unknown. A common mathematical method to translate sample plot histograms of different scanning targets to a uniform "metric" is needed to facilitate the use of common estimation models.

6.1.1 Most similar pairs

Initially, the only information of the new site is the remote sensing auxiliary data that covers the whole site area. Using this auxiliary information or some other sampling criteria, a number of N_c sample plots are measured (calibration plots, c) and are assumed to be known. With a dense set of field measurements, different forest stand parameters can been estimated with LiDAR histogram data with different accuracies. For instance, median tree height, hgM, of boreal forests is known to be quite precisely estimated, with an RMSE% approximately 10% at plot-level, while the number of stems per hectare, N, is estimated only with RMSE% of approximately 20-40%. If the forest stand parameters are estimated with high precision, plots where these forest stand parameters are similar contain also similar LiDAR histograms. Calibration of LiDAR-histograms is thus based on the most similar pairs, for which LiDAR-histograms are assumed to be equivalent.

Selection of most similar pairs for each calibration plot is performed using information on forest stand parameter values and accuracy of LiDAR-histogram based SBR-models. Most similar plot pairs, d_{jc}, for plots in the calibration set c from each database d_j are defined by minimizing the

weighted square distances of forest stand parameter values,

$$d_{jc,i} = \arg\min_{l} \sum_{k} \frac{1}{\sigma^2_{k,d_j}} \left(\mathbf{y}_{d_j,l,k} - \mathbf{y}_{c,i,k}\right)^2 \quad \forall i \in c, \tag{6.1}$$

where σ^2_{k,d_j} is the residual variance of the LiDAR based forest stand parameter k SBR estimates in plots of database d_j, $\mathbf{y}_{d_j,l,k}$ is the forest stand parameter k value of plot l in the database d_j and $\mathbf{y}_{c,i,k}$ is the forest stand parameter k value of plot i in the calibration set of the new site.

Obviously, depending on the relation of forest stand parameter distributions, there can be plots in the calibration set for which no pairs similar enough can be found. If the forest characteristics of different areas are too distinct, the most similar neighbour of a calibration plot is far in terms of weighted forest stand parameter distance. Such neighbours are not accepted to LiDAR measurement calibration. The definition of accepted plots is based on normal distributions, where the weighted square distance from the distribution mean, in this case the calibration plot i, follows the χ^2 distribution, see e.g.Mardia et al. (1980). With a given tolerance, $\hat{N}_{c_j}$ calibration set plots $\hat{c}_j$ with $\hat{N}_d$ corresponding neighbours $\hat{d}_{jc}$ are used to calibrate the LiDAR histograms of database j.

6.1.2 Database histogram calibration

LiDAR histograms consist of data containing discrete returns of first and last pulse height and intensity measurements. The distributions are affected by not only the forest and ground qualities, but also by scanning instrument quality and flight altitude. The physics of the latter are somewhat complex, but since the end user possesses only the histogram data, histogram rectification is performed by mathematical means. Calibration of the histograms is also of a "black-box"-model type, where exact information of the physics lying in the background is not available. In such cases, the complexity of the model is minimized with the aim to avoid over-learning. Thus, a simple linear model is used.

The linearity is motivated also by the physical basis of LiDAR histogram formation as the latency of monochromatic light when it is reflected back to the air from a surface beneath the aircraft. The intensity of the recorded pulse varies greatly due to the surface conditions of the reflection point, scanning height and calibration of the instrument. Scanning height can be assumed to scale roughly linearly to the intensity, calibration affects the amplitude. By the linear mapping between different scans, it is implicitly assumed that all LiDAR histogram measurements are reduced or increased by roughly the same percentage with pulse intensity. This assumption definitely fails at some point, when the intensity of returning pulses fails to reach the threshold value by which they are registered. This affects particularly last pulse data, but for first pulse data the assumption appears plausible as a first approximation. For pulses which do no bounce back the intensity is zero, before and after linear histogram calibration.

Height is the most important variable used in choosing plots for calibration. First pulse height can be assumed to correlate linearly with the height above the Digital Terrain Model (DTM) of the forest canopy that reflected it. For last pulses, correlation with height of canopy may well be non-linear, but it will still statistically be a monotonically increasing function of canopy height, the forest otherwise staying the same.

The percentage of pulses that are recorded as reflected at a particular height will vary according to flying altitude, LiDAR power and even weather. But such effects can be assumed to be smooth

functions of scanning conditions and are further ameliorated by the adoption of histograms - i.e. percentiles, as already used in Næsset (1997) and Means et al. (2000) - instead of absolute heights. These histograms of first and last pulse heights and the intensities of both locally correlate linearly, or at least in a monotonically increasing fashion, with small changes in scanning parameters and conditions.

Percentile points $\mathbf{X}_{\mathrm{var},c}$ of given calibration set c histograms $\mathbf{D}_{c,i}$ of plot i,

$$\mathbf{D}_{c,i} = \{\mathbf{H}_{\mathrm{f},ci}, \mathbf{H}_{\mathrm{l},ci}, \mathbf{I}_{\mathrm{f},ci}, \mathbf{I}_{\mathrm{l},ci}\}, \tag{6.2}$$

are defined for each measurement variable, $\mathrm{var} = \{\mathrm{H}_f, \mathrm{H}_l, \mathrm{I}_f, \mathrm{I}_l\}$, separately. Here H refers to height of the hit, I to the intensity of the hit, and f to first pulse and l to last pulse. The size of first and last pulse columns may vary. Percentile points are defined as points, where the cumulative sum of ordered measurements reaches a given percentile $p\%$ of the total sum of measurements. Here $p\% = 20\%, 40\%, \ldots, 100\%$ is used. Measurements classified as ground points, i.e. height under 2m, are neglected. These ground hits are not supposed to have direct correlation with tree qualities and quantities, and thus they are not acceptable for the histogram calibration which utilizes pairs selected by tree characteristics. If the ground hit measurements correlate with ground quality, there can be some correlation also with the trees, but it is assumed insignificant here. Percentile points are linear equations of histograms, and thus admissible for linear histogram calibration. Percentile points of the database, $\mathbf{X}_{\mathrm{var},d_j}$, are defined similarly.

To get information on the correlation between LiDAR variable at hand, $\mathbf{X}_{\mathrm{var}}$, and forest stand parameters, $\mathbf{Y}$, a multidimensional regression estimation is carried out for the selected plots in the calibration set. It can be defined by the normal likelihood:

$$N(\mathbf{Y}_{\hat{c}_j} | \mathbf{X}_{\mathrm{var},\hat{c}_j} \mathbf{W}_{\mathrm{var},\hat{c}_j}, \mathbf{\Sigma}_{\mathrm{var},\hat{c}_j}), \tag{6.3}$$

where the covariance matrix is diagonal, i.e. the forest stand parameter models are independent of each other. The model covariance defines the weight of different forest stand parameters in histogram calibration, corresponding to the LiDAR-variable model ability to explain the parameter.

The statistical variables of the database are corrected by the calibration coefficient a_{var,d_j} such that the LiDAR variables of selected plots of the calibration set and their calibrated neighbours from the database are assumed to be sampled from identical distributions

$$\mathbf{X}_{\mathrm{var},\hat{c}_j} \sim a_{\mathrm{var},d_j} \mathbf{X}_{\mathrm{var},\hat{d}_{jc}}. \tag{6.4}$$

Assuming that this identity is true, the forest stand parameter estimates $\hat{\mathbf{Y}}_{\hat{c}_j} = \mathbf{X}_{\mathrm{var},\hat{c}_j} \mathbf{W}_{\mathrm{var},\hat{c}_j}$ equal to $\hat{\mathbf{Y}}_{\hat{d}_{jc}} = a_{\mathrm{var},d_j} \mathbf{X}_{\mathrm{var},\hat{d}_{cj}} \mathbf{W}_{\mathrm{var},\hat{c}_j}$. The calibration model takes into account the fact that forest stand parameter values of pairs are not necessarily exactly equal, and different scales of inequality appear within different pairs and different forest stand parameters. The errors between forest stand parameters of pairs, $\mathbf{e}_i = \mathbf{Y}_{\hat{c}_j i} - \mathbf{Y}_{\hat{d}_{jc} i},\ i \in \hat{c}$, are assumed to be multinormally distributed with estimation mean $\hat{\mathbf{e}}_i = \hat{\mathbf{Y}}_{\hat{c}_j} - \hat{\mathbf{Y}}_{\hat{d}_{jc}}$,

$$\mathbf{e} \sim \prod_{i=1}^{\hat{N}_{cj}} N\left(\mathbf{e}_i | \hat{\mathbf{e}}_i, \mathbf{\Sigma}_{\mathrm{var},\hat{c}_j}\right). \tag{6.5}$$

Distribution mean depends on the difference of the forest stand parameter estimates of plots, and the covariance is the diagonal matrix of error variances of these estimates. Thus different forest stand

parameters in each plot are weighted, with a higher weight given to forest stand parameters that are well estimated with the LiDAR variables at hand, and a lower weight given to stand parameters that cannot be estimated well.

The calibration coefficient a_{var,d_j} is estimated by maximizing the likelihood (6.5). For each LiDAR measurement variable, an estimate of the calibration coefficient is defined with a similar procedure. Thus the new, calibrated database LiDAR-histograms

$$\hat{\mathbf{D}}_{d_j} = \{a_{\mathrm{H_f},d_j}\mathbf{H}_{\mathrm{f},d_j}, a_{\mathrm{H_l},d_j}\mathbf{H}_{\mathrm{l},d_j}, a_{\mathrm{I_f},d_j}\mathbf{I}_{\mathrm{f},d_j}, a_{\mathrm{I_l},d_j}\mathbf{I}_{\mathrm{l},d_j}\}, \tag{6.6}$$

and the statistical variables derived from them, $\hat{\mathbf{X}}_{d_j}$, are now assumed to be from the same distribution as the LiDAR-measurement of the new site.

6.2 Plot selection

The distribution and range of forest stand parameters in the database are rarely equal to the ones on the new site. Using all the database information, including information from plots of distinctly different character compared to the new site characteristics, causes error and bias to estimates of forest parameters on the new site. Thus it is necessary to select only the plots which can be expected to represent the new site characteristics to the teaching set.

Data from each site is independent of the data from other sites. The reference data from the new site is selected according to the auxiliary information of the site, in this case LiDAR histograms. Database plots are taken into the model teaching set as plots simulating the data that is not measured in the new site, and thus they must fit in the characteristics of the new site. If they would fit the distribution of the new site forest characteristics perfectly, estimation results would approach to those achieved using a dense set of field sample plots from the new site. Optimal approach in database plot selection would be to use the auxiliary data of the new site target plots to define the acceptability of the database plots in the model. However, there is no prior knowledge of the correlation between auxiliary variables and different forest stand parameters of the new site, i.e. whether a given variable is needed in the estimation model at all, and the distributions of different auxiliary variables may be very complicated. Thus, the use of auxiliary target plot variables as selection criteria, i.e. in verification of the new site qualities versus database qualities, is very complicated and yet unresolved task. However, since the database measurements are independent of the new site measurements, and the new site calibration set forest stand parameter distribution is known, the database plots fitting in the distribution may be taken as replicas of the new site plots. Such an approach does, however, miss the auxiliary information of the target plots of the new site, and thus ignores the possibility of undercoverage of the calibration set in terms of the forest stand parameter values of the whole new site set.

Considering the five forest stand parameters used in the publications, $\mathrm{Dgm}, \mathrm{Hgm}, \mathrm{N}, \mathrm{G}$ and V, a probability distribution of the new site can be established using the calibration plots. If the calibration set is not representing the new site characteristics well, it is likely that the probability distribution will also be unrepresentative. Thus, the error in calibration sample plot set selection strategy accumulates to plot selection of the databases.

An approach where the calibration set is used to span a p-dimensional multinormal distribution is used in the plot selection. The distributions of each forest stand parameter $\mathbf{y}_k$ are transformed to normal or close to normal, $\mathbf{y}_k^t$, and the distribution mean and covariance are calculated according

to them. The transformed calibration set is stated as c^t, transformed databases as d_j^t. In publication **II** five total forest stand parameters ($p = 5$) are used. The method is expanded to species specific parameters in publication **III**, where a similar approach is utilized to the five nonzero values of each species specific parameter. A square root transformation is used for some parameters, others remain untransformed. Same transformation is used for the same parameter total and species specific values. The transformation is based on visual validation of the common distribution behaviour of different parameters. The transformed distributions and their covariances are not exactly multinormal, and different sites may have different qualities of distribution. Thus it is still left open, if a better approach to do the plot selection exists. However, the results achieved with this approach are acceptable (the chosen plots are within the range of the calibration set distribution) and robust against changes in distributions.

The distance from the mean of the probability distribution can be stated as the Mahalanobis distance in the multinormal space (Mardia et al., 1980). The average Mahalanobis distance $\mathrm{m}_{\mathrm{y}_i,f}$ between $1 \times p$ random vector $\mathbf{y}_i$ and $1 \times p$ mean $\mathbf{y}_{f,m}$ according to the $p \times p$ covariance matrix $\mathbf{C}_f$ of the p-dimensional distribution of plot set f is stated as

$$\mathrm{m}(\mathrm{y}_i, f) = \mathrm{Mah}(\mathbf{y}_i, \mathbf{y}_{f,m}, \mathbf{C}_f) = \frac{1}{p}\mathrm{trace}\left((\mathbf{y}_i - \mathbf{y}_{f,m})\mathbf{C}_f^{-1}(\mathbf{y}_i - \mathbf{y}_{f,m})^T\right). \tag{6.7}$$

Mahalanobis distance is small if the random vector value is close to the mean of the distribution, otherwise it tends to become larger according to the ellipsoidal form of the multinormal distribution. In database selection, the plot set f defining the mean and covariance is the transformed calibration set c^t.

Plot selection is performed with heuristic selection criteria. Since the multinormal distribution is denser around the mean vector value, plots close to the mean in terms of covariance structure are selected with higher probability than those far from it. The heuristic selection criterion is based on random variables $r_i \in [0, 1]$: The database d_j plot i is accepted to the model if

$$e^{-\mathrm{m}(\mathbf{y}_{d_{ji}}^t, c^t)/2} > r_i. \tag{6.8}$$

The accepted plots form the new database $\tilde{d}_j$.

The criterion can be used as such for the total forest stand parameters only. However, considering the species specific parameters, the criterion is less simple. Total forest stand parameter distribution may be similar for different combinations of species, and using only the total parameter selection criterion may lead to false database distributions $\tilde{d}_j$. In the case of total forest stand parameters, the lack of species-wise selection does not disturb the estimation accuracy, since the total parameters cannot "see" the species specific combination, and as long as the criterion (6.8) is fulfilled, the calibration set and selected database set can be expected to be from similar forests.

For the species specific case, selection criteria for the database plots follow a similar rule. For each species s, a combination of nonzero species specific forest stand parameter elements of the calibration set, c_{sp_s}, and the database j, d_{j,sp_s}, are used to define the species specific mean Mahalanobis distance $m(\mathbf{y}^t_{d_{j,\mathrm{sp}_s},i}, c^t_{\mathrm{sp}_s})$. For the plots in database j, where there are no trees of species sp_s, the mean Mahalanobis distance is zero. For instance, the mean and covariance matrix of spruce are calculated from the transformed space of nonzero calibration set values of $\mathrm{Dgm}, \mathrm{Hgm}, \mathrm{N}, \mathrm{G}$ and V of spruce. The database plots with nonzero values of spruce are verified with the calibration set spruce distribution. Criterion (6.8) must be fulfilled for the total values and each species specific value

in the case. The limiting distance of database plot i is thus the one for which mean Mahalanobis distance has its maximum

$$\mathrm{m}(\mathbf{y}^t_{d_j,i}, c^t) = \max\left\{m(\mathbf{y}^t_{d_j,\mathrm{tot},i}, c^t_{\mathrm{tot}}), m(\mathbf{y}^t_{d_j,\mathrm{sp}_1,i}, c^t_{\mathrm{sp}_1}), \ldots, m(\mathbf{y}^t_{d_j,\mathrm{sp}_S,i}, c^t_{\mathrm{sp}_S})\right\}. \tag{6.9}$$

Here the subindex tot refers to the total values of the corresponding forest stand parameter, and S is the number of species classes. Thus, for each species, only the plots that fit in the multinormal space of the nonzero values of the calibration set parameters of both the species and total values, are included in the model.

After the plot selection, only the database plots which follow the same forest stand parameter distributions of total and species-wise forest stand parameters as in the calibration set, are accepted. For regression estimates, zero values of species specific forest stand parameters are challenging. To maintain the equality between the cumulative total values of parameters N, G and V and the sum of the corresponding species specific values, which are all estimated separately, calibration of the estimates is needed. A similar challenge is the presence of zero values, i.e. the plots where some species are absent. Such plots could be extracted from the estimation model, if the independent variable correlation would correlate with the phenomenon. However, with given LiDAR and digital aerial image data, such correlation does not occur. Thus, the zero values are kept in the estimation models. In regression analysis with homoscedastic variance, each data point, i.e. plot, has equal weight in the model. Thus, the ratio of zero values in plot selection has an effect, and if being too small or too large would result in biased estimates. For this reason, it is important that the database plots are reselected to achieve a selection with species specific zero ratios equal to that of the calibration set and thus the whole new site.

To define the forest characteristics in terms of contents of given species, classification of plots is performed in terms of species composition. In case of three different species (pine, spruce and hardwoods), each plot consists of either only one of these species or of a combination of two or all three species. If the presence of species (nonzero values of given species specific forest stand parameter values in plots) is labeled with one, and absence (zero values) with zero, and the order of species is given, the class of plot i is defined by:

$$\mathrm{cl}(i) = \{(1,0,0), (0,1,0), (0,0,1), (1,1,0), (1,0,1), (0,1,1), (1,1,1)\}, \tag{6.10}$$

where the classes $\mathrm{cl}(i) = \mathrm{cl}_l$ are labeled as $l = 1, \ldots, 7$, respectively. For instance, if the species order is given by (pine, spruce, hardwoods), and plot i contains nonzero values of forest stand parameters of spruce and hardwoods, it belongs to class 6, $\mathrm{cl}(i) = \mathrm{cl}_6$. Species ratios of plots of the database plot selection $\tilde{d}_j$ of size $\tilde{N}_{d_j}$ and plots of calibration set are defined by

$$\pi_{\tilde{d}_j,\mathrm{cl}_l} = \frac{\tilde{N}_{d_j,\mathrm{cl}_l}}{\tilde{N}_{d_j}} \quad \text{and} \quad \pi_{c,\mathrm{cl}_l} = \frac{N_{c,\mathrm{cl}_l}}{N_c}, \quad l = 1, \ldots, 7, \tag{6.11}$$

respectively. Here $\tilde{N}_{d_j,\mathrm{cl}_l}$ and N_{c,cl_l} are the number of selected database and calibration plots belonging to class l.

Reselection among the selected database plots is performed heuristically, with the aim to maintain similar ratios of different species compositions in the database as in the calibration set. Thus the selection criterion for each class is defined according to the under- or over-presentation of the classes in the database selection:

$$\pi_{\mathrm{cl}_l} = \frac{\pi_{c,\mathrm{cl}_l}}{\pi_{\tilde{d}_j,\mathrm{cl}_l}}, \quad l = 1, \ldots, 7. \tag{6.12}$$

If the ratio of a given class l in the calibration set is larger than that in the database plot selection, the class is under-presented in the database, and $\pi_l > 1$. If there are relatively more plots with class l in database plot selection than in the calibration set, class l is over-presented in the database and $\pi_l < 1$. To attain a criterion which keeps the order between the classes, criterion (6.12) is calibrated so that the plots belonging to a class with the largest ratio π_l are always selected and other classes are accepted with probabilities given in the same order as their appearance in database selection provides. Thus, the acceptance criterion of database $\tilde{d}_j$ plot i belonging to class l is given by the heuristic rule,

$$\frac{\pi_{\mathrm{cl}(i)}}{\max_l(\pi_l)} > r_i, \tag{6.13}$$

where $r_i \in [0, 1]$ is a random number. Plots accepted also in the reselection form the final database set $\tilde{d}_j$, which follows the same forest stand parameter distributions as the calibration set both in terms of nonzero total and species specific parameters, and in terms of species specific zero value ratios.

6.3 Model weighting

After the LiDAR histogram calibration and database selection procedures, each database j data is compatible with the new site data. Since the procedures are strongly based on the calibration characteristics, defining both the calibration coefficients of LiDAR histograms and database plot selection criteria, database plots are in fact representing "replicas" of the plots belonging to the new site calibration set. No plots outside the range of the calibration set are chosen, and thus the possible lack of new site information in the calibration set is copied also to database plots.

Another possible source of error arises from the fact that the database distribution density may differ from the calibration set density, producing biased estimates if used as such in the model teaching set. This phenomenon can be expected to happen whenever an alien site, i.e. a database, is used in the estimation of another site.

Overall, database utilization resembles the case, where new site sample plot selection is performed with a method that leads to under- or over-presentation of different forest characteristics. The selected plots of databases are considered as samples from the new site together with calibration set plots. Calibration set sample plots are the only true data from the new site, expected to be unbiased and representing well the forest stand parameter distribution density and range on the new site. Database plots are only as good as the correlation between the database and new site characteristics is. If the database does not cover the full calibration set parameter distribution, it is prone to be biased in terms of new site parameters. The lack of plots with extreme values of different forest stand parameters does not necessarily lead to biased estimates with a linear model, but the lack of different combinations of forest stand parameters in some area of the distribution may be a more severe a problem.

As the number of databases becomes larger, adding to the number of alien plots which cumulate the bias caused by database distribution density distortion, the effect of calibration set sample plots needs more weight in regression analysis. At any step of estimation, one should not rely more on database alien plots than on the native calibration set plots. Thus the native plots of the calibration set are used to define the mean of the estimated parameter distribution and weighted in regression in terms to rule the model definition over the alien plots.

The aim of forest stand parameter estimation using databases is to achieve a smaller RMSE on the new site verification set than using only the calibration set, keeping also the BIAS close to zero. Thus the task is a sample of bias-variance trade-off issue discussed already in Chapter 5.1. From the regression point of view, the calibration set is assumed to be the only true, unbiased dataset that contains the correct distribution density. Thus, the mean of the estimates should be equal to the calibration set mean. In practice, each input in SBR is normalized with calibration set characteristics:

$$\mathbf{y}_k \rightarrow \frac{\mathbf{y}_k - \overline{\mathbf{y}}_{kc}}{\sigma_{kc}},\ \forall k, \tag{6.14}$$

where $\mathbf{y}_k$ is the vector of forest stand parameter k, $\overline{\mathbf{y}}_{kc}$ is the calibration set mean and σ_{kc} the calibration set standard deviation of forest stand parameter k. Similar normalization is advisable to be performed also to independent candidate variables of the model in order to avoid any numerical problems due to different scales of response and independent variable values. The regression itself is then performed without a constant term, i.e. the final regression line goes through the calibration set mean with a slope defined by the calibration set combined with the databases.

Weights of the calibration set and database sample plots in regression need to be such, that the total weights of these two sources of data are equal. The total number of sample plots from J databases with plot selection and one calibration set is $N = N_c + N_{\tilde{d}_1} + \ldots + N_{\tilde{d}_J} = N_c + N_{\tilde{d}}$. From the viewpoint of the new site, all the accepted database plots are equally weighted, since the plot selection procedure is used to ensure that the selected plots fit in the span of new site characteristics. Thus the combination of different database plots is treated as one database containing alien plots.

In the weighted regression, the ordinary regression estimate for the weight (3.9) is replaced by the definition

$$\hat{\mathbf{w}} = \left(\mathbf{X}_s^T \mathbf{\Pi}_s^{-1} \mathbf{X}_s\right)^{-1} \left(\mathbf{X}_s^T \mathbf{\Pi}_s^{-1} \mathbf{y}_s\right), \tag{6.15}$$

which is known as the generalized least-squares (GLS) estimator. Here $\mathbf{X}_s = (\mathbf{X}_c^T, \mathbf{X}_{\tilde{d}}^T)^T$ and $\mathbf{y}_s = (\mathbf{y}_c^T, \mathbf{y}_{\tilde{d}}^T)^T$ are the sample variables and the response vector, respectively. The variance of the normally distributed error vector, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_N)$, is weighted to become heteroscedastic, i.e. each sample plot error variation is defined by the constant variance multiplied by a sample dependent weight π_i^{-1}. Thus the residual variance has spatial variation, and $\sigma^2 \mathbf{I}_N$ is replaced with $\sigma^2 \mathbf{\Pi}_s$, where $\mathbf{\Pi}_s$ is an $N \times N$ diagonal matrix with diagonal elements π_i corresponding to the samples $i = 1, \ldots, N$.

Inverse of the weight of each sample is defined by π_i, which may be different for different elements $i = 1, \ldots, N$ in the sample. In database model sampling, π_i is defined so that it varies according to the source of data such that the total weight of each source is equal in the regression model. For sample plot i, which belongs to the calibration set, it is

$$\pi_i = \pi_c = N_c/N,\ \forall i \in c, \tag{6.16}$$

and for sample plot i from the database $\tilde{d}$,

$$\pi_i = \pi_d = N_{\tilde{d}}/N,\ \forall i \in \hat{d}. \tag{6.17}$$

Thus, in case that $N_{\tilde{d}} > N_c$, π_i is smaller for the calibration sample plot, describing that the calibration set samples are under-represented in the total amount of sample plots.

For Bayesian formulation of regression, similar weighting can be defined by modifying the variance of the likelihood distribution, equation (5.2):

$$N(\mathbf{y}_s|\mathbf{X}_s\mathbf{w}, \sigma^2\mathbf{\Pi}_s) \sim \prod_{i=1}^{N} \left(\beta\pi_i^{-1}\right)^{N/2} \exp\left(-\beta\pi_i^{-1}||\mathbf{y}_{s,i} - \mathbf{X}_{s,i}\mathbf{w}||^2/2\right) \tag{6.18}$$

Variance allowed to the estimates follows thus the structure defined in matrix $\mathbf{\Pi}_s$, binding the regression weight to the allowed error size. As the regression weight of the sample becomes smaller, deviation from the true value of the response vector is allowed to be larger, and vice versa. With given π_i, the samples which belong to data sources with a smaller sample size are weighted more than those with a larger sample size in the model. With given regression weights of the calibration set and the database, representativeness and total weight of both is equal in the likelihood.

Since the variable selection is crucial in the regression models of such possibly small dataset size problems, a similar weighted regression approach needs to be adopted to the sparse Bayesian regression. Reconsidering the likelihood (6.18), the formulation can be transformed to

$$\left(\beta\mathbf{\Pi_s}^{-1}\right)^{N/2} \exp\left(-\left(\mathbf{y}_s - \mathbf{X}_s\mathbf{w}\right)^T \mathbf{\Pi}_s^{-1/2}\beta\mathbf{\Pi}_s^{-1/2}\left(\mathbf{y}_s - \mathbf{X}_s\mathbf{w}\right)/2\right). \tag{6.19}$$

Parameters and hyperparameters of SBR can be solved by replacing the measurements $(\mathbf{X}, \mathbf{Y})$ by π weighted values $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \mathbf{\Pi}_s^{-1/2}(\mathbf{X}_s, \mathbf{Y}_s)$. This approach is sufficient and valid through the whole type-II maximum likelihood method discussed earlier. In this method, due to the given prior distribution on the weights, the number of selected variables depends on the size of the data compared to the precision of the estimate, not on the scale of the data. Thus the weighting in this form affects only the mutual weighting of the different sources of data, as in ordinary weighted regression.

Using this formulation in the Bayes' rule (5.4), the analytical posterior mean and covariance for the weight are,

$$\mathbf{\Sigma}_w = \left(\beta\mathbf{X}_s^T\mathbf{\Pi}_s^{-1}\mathbf{X}_s + \mathbf{A}\right)^{-1} \tag{6.20}$$

$$\boldsymbol{\mu}_w = \mathbf{\Sigma}\beta\mathbf{X}_s^T\mathbf{\Pi}_s^{-1}\mathbf{y}_s. \tag{6.21}$$

With zero values of weight prior hyperparameters, $\alpha_m = 0,\ \forall m$, these equations have an equal formulation as the mixed estimator of Theil and Goldberger (1960); Theil (1963):

$$\mathbf{w}_{\mathrm{tc}} = \left(\mathbf{X}_c^T\mathbf{\Pi}_c^{-1}\mathbf{X}_c + \mathbf{X}_{\tilde{d}}^T\mathbf{\Pi}_{\tilde{d}}^{-1}\mathbf{X}_{\tilde{d}}\right)^{-1}\left(\mathbf{X}_c^T\mathbf{\Pi}_c^{-1}\mathbf{y}_c + \mathbf{X}_{\tilde{d}}^T\mathbf{\Pi}_{\tilde{d}}^{-1}\mathbf{y}_{\tilde{d}}\right) \tag{6.22}$$

$$= \left(\mathbf{X}_c^T\mathbf{X}_c + \lambda\mathbf{X}_{\tilde{d}}^T\mathbf{X}_{\tilde{d}}\right)^{-1}\left(\mathbf{X}_c^T\mathbf{y}_c + \lambda\mathbf{X}_{\tilde{d}}^T\mathbf{y}_{\tilde{d}}\right), \tag{6.23}$$

where $(\mathbf{X}_c, \mathbf{y}_c)$ and $(\mathbf{X}_{\tilde{d}}, \mathbf{y}_{\tilde{d}})$ are the measurements of the calibration set and those of selected plots of the database, correspondingly. Here $\lambda = N_c/N_{\tilde{d}}$ corresponds to the ratio between alien plot and native plot weights, formulated to be the weight for the database data. With given values, both the calibration set and the database affect the estimate with a total weight of N_c plots: N_c plots from the calibration set and $\lambda \times N_{\tilde{d}}$ plots from the database.

Theil and Goldberger (1960) stated that weight λ should be estimated using the model residual error ratio. However, such an approach does not take into account the reliability of different sites: the weight of databases should not be larger than the calibration set weight, which is possible when we are using only residual error information. A mixed estimation procedure is used in forest inventory

approaches e.g. when combining two different NFI datasets (Korhonen, 1993) or when combining data from two different forest areas (Suvanto and Maltamo, 2010). The given weight λ corresponds to the 50% proportion of calibration plots from the total population, which was given in Suvanto and Maltamo (2010).

Given regression weights are suitable whenever the number of calibration plots is small compared to the total number of plots selected, i.e. when $\pi_c^{-1} > \pi_{\tilde{d}}^{-1}$, otherwise the relative weight of database plots is incorrectly increased from the original. If less than N_c database plots are selected, then $\pi_{\tilde{d}}^{-1} > \pi_c^{-1}$. In such a case, however, databases most likely do not fit the calibration set characteristic's distributions, and databases have no additional information to contribute to the new site characteristics. Thus the weight of an individual database plot should never be more than the weight of any of the calibration set plots, and some other formulation of the regression weights needs to be found.

6.4 Process of validation

The procedures utilizing plot databases in estimation of a new site forest stand parameters were validated similarly, with some exceptions, in publications **II** and **III**. Only a small number of sample plot measurements were used in the new site which is estimated, serving as calibration plots. The rest of the new site data served as the verification set of the estimation procedure. In publication **II**, between 50 and 100 calibration plots were selected with given criteria, utilizing LiDAR data as widely as possible. In publication **III**, 50 calibration plots were used. In the calculations, LiDAR histograms were assumed to be available before any field measurements are performed, and utilized at many levels of the estimation procedure. Field measurements at given calibration plots were derived in a similar manner as in the databases.

The first crucial task in combining datasets from multiple sites was to calibrate the auxiliary data to compatible form. The linear correlation coefficient in LiDAR histograms derived with linear multinormal models utilizing most similar pairs in the forest stand parameter space showed to produce acceptable correlations to histograms. Distributions with even very large differences in the scale of intensity measurements were calibrated to scale with distributions correlating both in terms of shape and mean.

The second task was plot selection from databases to fit the forest characteristics on new site. For total forest stand parameter estimation only, the task is simple and a relatively large number of database plots are accepted in most test cases. For the species specific parameters, the task is more complicated, and depending on the match between the database and new site calibration set forest characteristics, possibly only a few plots are accepted.

Estimates of five different total forest stand parameters, Dgm, Hgm, N, G, and V, were evaluated with respect to the calibration set size in publication **II**. In publication **III**, the set of forest stand parameters recorded on each field plot was extended to altogether 20 parameters. The total parameters were supplemented with corresponding species-wise parameters Dgm_1, Dgm_2, Dgm_3, Hgm_1, Hgm_2, Hgm_3, N_1, N_2, N_3, G_1, G_2, G_3, V_1, V_2 and V_3. Here the indices 1-3 refer to the species, 1 for Scots pine (Pinus sylvestris), 2 for Norway spruce (Picea abies) and 3 for hardwoods treated as a group, but mostly comprising birch (Betula pendula and Betula pubescens). Thus, three different species are handled in database calibration and plot selection procedures, and estimated with separate SBR models. The accuracy of the estimate is verified against each site being as the new site

by repeating the method 50 times with different randomly selected calibration sets which fulfill the given criteria.

6.5 Results of database utilization procedure

Use of databases was tested in publications **II** and **III**. In publication **II**, LiDAR and plot sample measurement data were available from four different sites in Finland. The site qualities were quite homogeneous, representing typical forests of central Finland, except of one site which is located in north Finland. LiDAR data was collected with different instruments and flight altitudes, and with some variation in true mean point density on plots. In publication **III**, seven separate sites were available. In addition to earlier data, also digital aerial image data was available in all but one site. LiDAR data measurement instrument information was not available for some sites at all, but only the final preprocessed data. The preprocessed LiDAR histogram data in both publications contained first and last and only pulse data. The only pulse data was used as both first and last pulse data in the estimations.

Total forest stand parameters Hgm, Dgm, N, G and V are estimated relatively accurately with LiDAR variables. Especially Hgm correlates highly with first pulse height percentile points. The parameters with less correlation with LiDAR are more sensitive to teaching set data size. Especially the species specific parameters suffer easily when dataset size is not sufficiently large. With diminished dataset size, RMSE% and BIAS% increase depending on how well the teaching set data represents the parameter variation of the site. To maintain robust results and estimates with reliable expectation of accuracy, databases are used to complement new site plots.

The test results for total forest stand parameter estimates are encouraging. Different randomly selected combinations of calibration plots are used in distinct repetitions of the algorithm. Using only the calibration set information, estimation accuracy varies a lot. This variation can be substantially diminished with adding database information. In publication **II**, the mean of verification set RMSE% of different repetitions remained close to the optimal RMSE% achieved with a dense set of teaching set plots (400-600 plots) even with a calibration set size of approximately 50 plots. In publication **II**, model weighting was not used, since only three databases were used. However, the effect of bias due to unweighting can be seen cumulating as the number of databases increases from one separate database to a combination of all databases available. The average bias% of different repetitions is not remarkably large, but the trend of cumulation is clear.

The results of species specific estimates vary greatly depending on the new site at hand, see publication **III**. Total forest stand parameters are estimated with distinctly more stable RMSE%'s when using the databases than when using only the calibration set as the teaching set of the model. For the species specific parameters, the advantage of databases is not as clear. Generally, if the new site is well covered by the databases in terms of forest stand parameter distributions, the species specific estimates are robust against the calibration set contents. For almost all the 20 parameters in the seven different new sites, the average results of the database assisted estimates were consistently better or at least as good as the estimates based only on the calibrations set. Only four spruce specific parameters which were located on two sites, two parameters in both, caused slight deterioration in accuracy. Using the weighted SBR, the problem of bias was circumvented. The average of the BIAS% of repetitions was close to zero regardless of the number or size of the used databases.

CHAPTER VII

Discussion and future prospects

Operational forest inventory requires methods and algorithms which enable automated and adaptive use of given data with low costs. Site specific algorithm tuning should be automated so that it is fast and objective and thus forest stand parameter prediction methods easily adapt to site specific circumstances. These objectives are achieved for the standard forest stand parameter estimation cases with the algorithms presented in this thesis.

A regression algorithm with fast automatic variable selection was shown to give equally good results as standard regression methods. With sparse Bayesian regression, a local variable selection procedure can be attached even to problems with a small dataset, since the algorithm automatically notices the uncertainty of information due to small dataset size, and cuts down the number of used variables accordingly. The problems of over-fitting and variable collinearity are thus circumvented. In the future, a similar sparse approach with seemingly unrelated regression (SUR) could be tested in forest stand parameter estimation, even though SUR estimates often give results that improve only a little from OLS estimates. Similar methods could be also extended to cases, in which the spatial correlation between forest stand parameters affects the model, see e.g. Finley et al. (2008) for a Bayesian approach of such analysis.

Use of databases in new site forest stand parameter prediction proved to be an intricate task. Each step of the method had a significant effect for producing accurate predictions. The selection of new site calibration plots is crucial, since the calibration set determines the characteristics which are used for LiDAR histogram calibration, database plot selection and at least half of the weight in the final regression estimates. Thus it is important, that the calibration set covers the variability of the forest characteristics and represents well the true distribution of the site's forest stand parameters. Otherwise bias is prone to appear. This is even more important when species specific parameters are predicted, since the variability of the different combinations to be predicted is larger due to higher number of degrees of freedom.

LiDAR histogram calibration and plot selection produce robust estimation results, even though some of the assumptions made concerning linearity and normality of the material are not always true, and more general approaches should be searched for in future work. LiDAR calibration may also be performed by standardization of instrument operating system and measurement strategy planning, if this is possible. However, if the only data available are the numerical values of the measurement histograms, the linear calibration introduced in this thesis shows to give adequate results. Calibration of other remote sensing data, such as digital aerial photographs or satellite data, should be

performed in the image processing stage preceding feature extraction.

The plot selection procedure introduced in this thesis is based on heuristic selection of plots from an existing dataset of databases within the framework of characteristics given in the calibration set. The method is strongly based on multinormal distributions, which are a generalization of given forest stand parameter distributions. The generalization does not hold true in all cases. In different sites the forest stand parameter distribution shape may deviate from this assumption. However, with only a minor set of calibration sample plots (e.g. 50 plots), the true form of the mutually dependent distributions with even 20 forest stand parameters is impossible to predict. Thus, the general prior information about the shape of distributions is an adequate approach to plot selection. Tests with different forms of the distributions (different transformations of forest stand parameter values) show that this stage of the database utilization procedure is rather robust to changes, and final estimation accuracy remains at the same level independent of the specific transformation, as long as the transformation is somewhat normal in shape. However, other approaches to plot selection should be tested, especially ones that keep also distribution densities similar to the calibration set density.

Use of calibrated LiDAR histograms and other auxiliary data combined with field measurements of the selected plots of databases is still prone to bias due to the fact that the plots are alien to the new site. The weight of such alien plots is kept below a native plot weight with weighted regression, where the dataset sizes of both sources, the databases and the calibration set, are used to define the weights. The method can be incorporated to sparse Bayesian regression, which is an automatic method for variable selection. The results so far are promising, the bias of the estimates is kept close to zero while using the direct, unweighted regression the bias tends to become a significant problem as the database size increases. The ratio of the weights of different data sources used in this study is 50%. Also other ratios, and possibly ratios based on the database accuracy in calibration set forest stand parameter estimation could be tested.

In this study, the weighted sparse Bayesian regression utilize the full information of plots in terms of variable selection, i.e. the number of selected variables follows the total number of sample plots N_s, not the weighted sum of plots, $N_c \times 1 + N_{\tilde{d}} \times \lambda < N_s$. Thus, if the number of database plots used in the estimation is large, the number of variables used is allowed to be large, even though the calibration set size is small and would allow the use of only a small number of variables in terms of avoiding over-fitting. To preserve the uncertainty of variable selection compared to the trusted dataset size, a weighted log-likelihood function approach could be utilized e.g. in the spirit given in Shimodaira (2000).

Adequate use of databases is strongly ruled by the characteristics of the new site and the databases available. For accurate estimates, databases must correlate enough with the new site in terms of forest stand parameters that can be predicted with the existing auxiliary data. If the parameters do not correlate with the auxiliary data, estimation will fail regardless of the database plot selection. However, from the viewpoint of estimation, it is not important, if there are differences in forest stand parameter distribution values of the parameters which cannot be estimated with given auxiliary data. For instance, new site total forest stand parameters can be estimated correctly even with plots containing wrong species of trees. In future studies, also the knowledge of auxiliary data model accuracy could be included to the database plot selection procedure, which has shown to be a rather complicated task.

Database coverage over the new site forest stand distribution has proven to be crucial to get improved estimation accuracy. Over-coverage can be handled by the plot selection procedure, but

under-coverage may be a serious problem. For multiple forest stand parameter estimation in species specific analysis, it is likely that there are not enough plots similar to new site plots unless the number of plot databases is large and covers a wide range of different forest types. If certain types of plots are under-represented in the teaching set of the model, a forest stand parameter which correlates well with the auxiliary data may be estimated with a bias. Knowledge of the under-coverage can be gained by comparing the calibration set distribution with the database distribution. Weighting of plots with respect to under-representation should be performed, once under-coverage is discovered. However, with the 20 degrees of freedom and no definite information of distribution shape, this is a challenging task and must be considered in future work.

When the auxiliary data and forest stand parameter distributions of selected database plots fit the new site characteristics, the databases could be used also to another approach which has shown to give accurate results in forest stand parameter estimation, k-MSN. With only a sparse set of new site sample plots, k-MSN is not likely to produce accurate estimation results, since the accuracy is highly dependent on the local neighbours in the feature space. Thus, with database plots, the problem could be solved. However, serious planning must be performed considering the distribution density and distribution shape of selected database plots and database plot weighting. Also variable selection must be carefully implemented, since it affects the eigenvectors and -values in the linear equations of canonical correlation analysis (CCA) of the k-MSN method. Automated variable selection in CCA is another issue, which should be considered using a prior weighting scheme.

BIBLIOGRAPHY

Anttila, P., Lehikoinen, M., 2002. Kuvioittaisten Puustotunnusten Estimointi Ilmakuvilta Puoliautomaattisella Latvusten Segmentoinnilla. Metsätieteen aikakauskirja 3/2002, 381–389, (In Finnish.).

Asner, G. P., Broadbent, E. N., Oliveira, P. J. C., Keller, M., Knapp, D. E., Silva, J. N. M., 2006. Condition and Fate of Logged Forests in the Brazilian Amazon. Proceedings of the National Academy of Sciences USA (PNAS) 103 (34), 12947–12950.

Asner, G. P., Knapp, D. E., Balaji, A., Páez-Acosta, G., 2009. Automated Mapping of Tropical Deforestation and Forest Degradation: CLASlite. Journal of Applied Remote Sensing 3 (033543), 24p.

Bitterlich, W., 1948. Die Winkelzählprobe. Allgemeine Forst- und Holzwirtschaftliche Zeitung 59 (1), 4–5.

Brandtberg, T., 1999. Automatic Individual Tree Based Analysis of High Spatial Resolution Aerial Images on Naturally Regenerated Boreal Forests. Canadian Journal of Forest Research 29, 1464–1478.

Breidenbach, J., Næsset, E., Lien, V., Gobakken, T., Solberg, S., 2010. Prediction of Species Specific Forest Inventory Attributes Using a Nonparametric Semi-Individual Tree Crown Approach Based on Fused Airborne Laser Scanning and Multispectral Data. Remote Sensing of Environment 114 (4), 911–924.

Copas, J. B., 1983. Regression, Prediction and Shrinkage. Journal of the Royal Statistical Society. Series B (Methodological) 45 (3), 311–354.

Efron, B., Morris, C., 1975. Data Analysis Using Stein's Estimator and its Generalizations. Journal of the American Statistical Association 70 (350), 311–319.

Finley, A. O., Banerjee, S., Ek, A. R., McRoberts, R. E., 2008. Bayesian Multivariate Process Modeling for Prediction of Forest Attributes. Journal of Agricultural, Biological, and Environmental Statistics 13 (1), 1–24.

Flewelling, J. W., 2006. Forest Inventory Predictions From Individual Tree Crowns: Regression Modeling Within a Sample Framework. In: Proceedings of the Eighth Annual Forest Inventory and Analysis Symposium. pp. 203–210.

Furnival, G. M., Wilson, R. E. J., 1974. Regression by Leaps and Bounds. Technometrics 16 (4), 499–511.

Gelman, A., 2006. Prior Distributions for Variance Parameters in Hierarchical Models. Bayesian Analysis 1 (3), 515 – 533.

Gelman, A., B., C. J., S., S. H., Rubin, D. B., 2004. Bayesian Data Analysis, 2nd Edition. Chapman & Hall/CRC.

Goetz, S., Steinberg, D., Dubayah, R., Blair, B., 2007. Laser Remote Sensing of Canopy Habitat Heterogeneity as a Predictor of Bird Species Richness in an Eastern Temperate Forest, USA. Remote Sensing of Environment 108 (3), 254–263.

Goldstein, M., Smith, A. F. M., 1974. Ridge-Type Estimators for Regression Analysis. Journal of the Royal Statistical Society. Series B (Methodological) 36 (2), 284–291.

Gougeon, F. A., 1995. A Crown-Following Approach to the Automatic Delineation of Individual Tree Crowns in High Spatial Resolution Aerial Images. Canadian Journal of Remote Sensing 21 (3), 274–284.

Haara, A., Korhonen, K. T., 2004. Kuvioittaisen Arvioinnin Luotettavuus. Metsätieteen Aikakauskirja 4/2004, 489–508, in Finnish.

Hall, S. A., Burke, I. C., Box, D. O., Kaufmann, M. R., Stoker, J. M., 2005. Estimating Stand Structure Using Discrete-return Lidar: an Example from Low Density, Fire Prone Ponderosa Pine Forests. Forest Ecology and Management 208, 189–209.

Häme, T., Rauste, Y., Sirro, L., Ahola, H., Lappi, J., Rudant, J., Mascret, A., Sept. 6 - 10 2004. Using ERS 1 and ASAR imagery for mapping forest in French Guiana. In: Proceedings of the 2004 Envisat and ERS Symposium. Salzburg, Austria, pp. 441–446, (ESA SP-572, April 2005).

Hoerl, A. E., Kennard, R. W., 1970. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 12 (1), 69–82.

Holmgren, J., 2004. Prediction of Tree Height, Basal Area and Stem Volume in Forest Stands Using Airborne Laser Scanning. Scandinavian Journal of Forest Research 19, 543–553.

Holmgren, J., Jonsson, T., October, 3-6 2004. Large Scale Airborne Laser Scanning of Forest Resources in Sweden. In: Thies, M., Koch, B., Spiecker, H., Weinacker, H. (Eds.), Laser Scanners for Forest and Landscape Assessment. Proceedings of the ISPRS Working Group VIII/2. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Freiburg, Germany, pp. 157–160.

Holmgren, J., Persson, Å., 2004. Identifying species of Individual trees using Airborne Laser Scanner. Remote Sensing of Environment 90, 415–423.

Holmström, H., 2002. Estimation of Single-tree Characteristics Using the kNN Method and Plot-wise Aerial Photograph Interpretations. Forest Ecology and Management 167, 303–314.

Holopainen, H., Kalliovirta, J., 2006. Modern Data Acquisition for Forest Inventories. In: Kangas, A., Maltamo, M. (Eds.), Forest Inventory – Methodology and Applications. Springer, pp. 343 – 362.

Hyyppä, J., Hyyppä, H., Litkey, P., Yu, X., Haggrén, H., Rönnholm, P., Pyysalo, U., Pitkänen, J., Maltamo, M., 2004. Algorithms and Methods of Airborne Lase Scanning for Forest Measurements. In: Laser Scanners for Forest and Landscape Assessment. Proceedings of the ISPRS Working Group VIII/2. Freiburg, Germany, pp. 82–89.

Hyyppä, J., Kelle, O., Lehikoinen, M., Inkinen, M., 2001. A Segmentation-Based Method to Retrieve Stem Volume Estimates from 3-D Tree Height Models Produced by Laser Scanners. IEEE Transactiions on Geoscience and Remote Sensing 39 (5), 969–975.

INPE, 2005. Instituto Nacional de Pesquisas Espaciais. www.obt.inpe.br.

Kangas, A., Maltamo, M. (Eds.), 2006. Forest Inventory – Methodology and Applications. Springer.

Kangas, A. S., Kangas, J., 1999. Optimization Bias in Forest Management Planning Solution Due to Errors in Forest Variables. Silva Fennica 33 (4), 303 – 315.

Katila, M., Tomppo, E., 2001. Selecting Estimation Parameters for the Finnish Multisource National Forest Inventory. Remote Sensing of Environment 76, 16–32.

Kilkki, P., Päivinen, R., 1987. Reference Sample Plots to Combine Field Measurements and Satellite Data in Forest Inventory. In: Remote Sensing-Aided Forest Inventory. Vol. 19 of Research Notes. University of Helsinki, Department of Forest Mensuration and Management, pp. 209–215.

Koch, B., Heyder, U., Weinacker, H., 2006. Detection of Individual Tree Crowns in Airborne Lidar Data. Photogrammetric Engineering and Remote Sensing 72 (4), 357–363.

Korhonen, K. T., 1993. Mixed Estimation in Calibration of Volume Functions of Scots Pine. Silva Fennica 27, 269–276.

Korhonen, K. T., Kangas, A., 1997. Application of Nearest-neighbour Regression for Generalizing Sample Tree Information. Scandinavian Journal of Forest Research 12, 97–101.

Korpela, I., 2003. Individual Tree Measurements by Means of Digital Aerial Photogrammetry. Ph.D. thesis, University of Helsinki, Department of Forest Resource Management.

Korpela, I., 2004. Individual Tree Measurements by Means of Digital Aerial Photogrammetry. Silva Fennica monographs 3, 93p.

Laasasenaho, J., 1982. Taper Curve and Volume Function for Pine, Spruce and Birch. Communicationes Instituti Forestalis Fenniae 108, 74 p.

Lappi, J., 1993. Metsäbiometrian Menetelmiä. Vol. 24 of Silva Carelica. University of Joensuu, (Study book, in Finnish).

LeMay, V., Temesgen, H., 2005. Comparison of Nearest Neighbor Methods for Estimating Basal Area and Stems per Hectare Using Aerial Auxiliary Variables. Forest Science 51 (2), 109 – 119.

MacKay, D. J. C., 1992. Bayesian Interpolation. Neural Computation 4 (3), 415–447.

Mackay, D. J. C., 1994. Bayesian Methods for Backpropagation networks. In: Domany, E., van Hemmen, J. L., Schulten, K. (Eds.), Models of Neural Networks III. Springer, Ch. 6, pp. 211–254.

MacKay, D. J. C., 1999. Comparison of Approximate Methods for Handling Hyperparameters. Neural Computation 11 (5), 1035–1068.

Magnussen, S., Næsset, E., Gobakken, T., 2010. Reliability of LiDAR Derived Predictors of Forest Inventory Attributes: A Case Study with Norway Spruce. Remote Sensing of Environment 114, 700–712.

Mallet, C., Bretar, F., 2009. Full-Waveform Topographic Lidar: State-of-the-Art. ISPRS Journal of Photogrammetry & Remote Sensing 64, 1 – 16.

Maltamo, M., Bollandsås, O. M., Næsset, E., Gobakken, T., Packalén, P., 2009. Different Sampling Strategies for Field Training Plots in ALS inventory. In: Proceeding of the SilviLaser 2009 Conference. p. 9p.

Maltamo, M., Kangas, A., 1998. Methods Based on k-Nearest Neighbor Regression in the Prediction of Basal Area Diameter Distribution. Canadian Journal of Forest Research 28 (8), 1107–1115.

Maltamo, M., Malinen, J., Packalén, P., Suvanto, A., Kangas, J., 2006. Non-parametric Estimation of Stem Volume Using Laser Scanning, Aerial Photography and Stand Register Data. Canadian Journal of Forest Research 36, 426–436.

Mardia, K. V., Kent, J. T., Bibby, J. M., 1980. Multivariate Analysis. Academic Press.

Mcroberts, R., 2009. A Two-Step Nearest Neighbors Algorithm Using Satellite Imagery for Predicting Forest Structure within Species Composition Classes. Remote Sensing of Environment 113 (3), 532–545.

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, 36–44.

Means, J. E., Acker, S. A., Fitt, B. J., Renslow, M., Emerson, L., Hendrix, C., 2000. Predicting Forest Stand Characteristics with Airborne Scanning Lidar. Photogrammetric Engineering and Remote Sensing 66, 1367–1371.

Mehtätalo, L., Nyblom, J., 2009. Estimating Forest Attributes Using Observations of Canopy Height: A Model-Based Approach. Forest Science 55 (5), 411 – 422.

Moeur, M., Stage, A. R., 1995. Most Similar Neighbour: an Improved Sampling Inference Procedure for Natural Resource Planning. Forest Science 41 (2), 337–359.

Muinonen, E., Maltamo, M., Hyppänen, H., Vainikainen, V., 2001. Forest Stand Characteristics Estimation Using a Most Similar Neighbor Approach and Image Spatial Structure Information. Remote Sensing of Environment 78, 223–228.

Næsset, E., 1997. Determination of Mean Tree Height of Forest Stands Using Airborne Laser Scanning Data. ISPR Journal of Photogrammetry and Remote Sensing 52 (2), 49–56.

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, 88–99.

Næsset, E., 2004a. Effects of Different Flying Altitudes on Biophysical Stand Properties Estimated from Canopy Height and Density Measured with a Small-Footprint Airborne Scanning Laser. Remote Sensing of Environment 91, 243–255.

Næsset, E., 2004b. Estimation of Above- amd Below-Ground Biomass in Boreal Forest Ecosystems. In: Laser Scanners for Forest and Landscape Assessment. Proceedings of the ISPRS working group VIII/2. Freiburg, Germany, pp. 145–148.

Næsset, E., 2004c. Practical Large-scale Forest Stand Inventory Using a Small Airborne Scanning Laser. Scandinavian Journal of Forest Research 19, 164–179.

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.

Næsset, E., Bjerknes, K. O., 2001. Estimating Tree Heights and Number of Stems in Young Forest Stands Using Airborne Laser Scanner Data. Remote Sensing of Environment 78, 328–340.

Næsset, E., Bollandsås, O. M., Gobakken, T., 2005. Comparing Regression Methods in Estimation of Biophysical Properties of Forest Stands from Two Different Inventories Using Laser Scanning Data. Remote Sensing of Environment 94, 541–553.

Neal, R. M., 1996. Bayesian Learning for Neaural Networks. Vol. 118 of Lecture notes in Statistics. Springer-Verlag.

O'Hara, R. B., Sillanpää, M. J., 2009. A Review of Bayesian Variable Selection Methods: What, How and Which. Bayesian Analysis 4 (1), 85–118.

Packalén, P., Maltamo, M., 2006. Predicting the Volume by Tree Species Using Airborne Laser Scanning and Aerial Photographs. Forest Science 52, 611–622.

Packalén, P., Maltamo, M., 2007. The k-MSN Method for the Prediction of Species-Specific Stand Attributes Using Airborne Laser Scanning and Aerial Photographs. Remote Sensing of Environment 109, 328–341.

Patenaude, G., Hill, R. A., Milne, R., Gaveau, D. L. A., Briggs, B. B. J., Dawson, T. P., 2004. Quantifying Forest Above Ground Carbon Content Using LiDAR Remote Sensing. Remote Sensing of Environment 93 (3), 368–380.

Persson, Å., Holmgren, J., Söderman, U., Olsson, H., 2004. Tree species classification of individual trees in Sweden by combining high resolution laser data with high resolution near-infrared images. In: Proceedings of the Natscan Conference. pp. 204–207.

Peuhkurinen, J., Maltamo, M., Malinen, J., Pitkänen, J., Packalén, P., 2007. Preharvest Measurement of Marked Stands Using Airborne Laser Scanning. Forest Science 53 (6), 653–661.

Peuhkurinen, J., Maltamo, M., Vesa, L., Packalén, P., 2008. Estimation of Forest Stand Characteristics Using Spectral Histograms Derived from an Ikonos Satellite Image. Photogrammetric Engineering & Remote Sensing 74 (11), 1335–1341.

Poso, S., Wang, G., Tuominen, S., 1999. Weighting Alternative Estimates when Using Multi-Source Auxiliary Data for Forest Inventory. Silva Fennica 33 (1), 41–50.

Rauste, Y., Häme, T., Ahola, H., Stach, N., Henry, J.-B., April, 23–27 2007. Detection of Forest Changes Over French Guiana Using ERS-1 and ASAR Imagery. In: Proceedings of Envisat Symposium 2007 (ESA SP-636). European Space Agency, Montreux, Switzerland, p. 6p.

Shimodaira, H., 2000. Improving Predictive Inference Under Covariate Shift by Weighting the Log-likelihood Function. Journal od Statistical Planning and Inference 90, 227–244.

Solberg, S., Næsset, E., Lange, H., Bollandsås, O. M., 2004. Remote Sensing of Forest Health. In: Laser Scanners for Forest and Landscape Assessment. Proceedings of the ISPRS Working Group VIII/2. Freiburg, Germany, pp. 161–166.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., van der Linde, A., 2002. Bayesian Measures of Model Complexity and Fit. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 64 (4), 583–639.

Suvanto, A., Maltamo, M., 2010. Using Mixed Estimation for Combining Airborne Laser Scanning Data in Two Different Forest Areas. Silva Fennica.

Suvanto, A., Maltamo, M., Packalén, P., Kangas, J., 2005. Kuviokohtaisten Puustotunnusten Ennustaminen Laserkeilauksella. Metsätieteen Aikakauskirja 4/2005, 413–428, (In Finnish).

Suvanto, A., Packalén, P., Maltamo, M., 2010. A Simulation of GPS Location Error to ALS Based Estimates in Two Separate Forest Areas in Finland, revised to Forestry.

Taskinen, I., Heikkinen, J., 2004. A Nonparametric Bayesian Method for Assessing Uncertainty in Thematic Maps of Forest Variables, revision submitted to Journal of Agricultural, Biological, and Environmental Statistics.

Theil, H., 1963. On the Use of Incomplete Prior Information in Regression Analysis. Journal of the American Statistical Association 58 (302), 401–414.

Theil, H., Goldberger, A. S., 1960. On Pure and Mixed Statistical Estimation in Economics. International Economic Review 2 (1), 65–78.

Tipping, M. E., 2001. Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning Research 1, 211–244.

Tipping, M. E., 2004. Bayesian Inference: An Introduction to Principles and Practice in Machine Learning. In: Bousquet, O., von Luxburg, U., Rätsch, G. (Eds.), Advanced Lectures on Machine Learning. Springer, pp. 41–62.

Tokola, T., Pitkänen, J., Partinen, S., Muinonen, E., 1996. Point Accuracy of a Non-parametric Method in Estimation of Forest Characteristics with Different Satellite Materials. Int. J. of Remote Sens. 17, 2333–2351.

Tomppo, E., 1991. Satellite Image-based National Forest Inventory in Finland. In: Int. Arch. Photogr. Remote Sensing. Vol. 28 of Proceedings of the Symposium on Global and Environmental Monitoring, Techniques and Impacts. Victoria, British Columbia, Canada, pp. 419–424.

Tomppo, E., 1993. Multi-source National Forest Inventory of Finland. In: Proceedings of Ilvessalo Symposium. Vol. 444 of Metsäntutkimuslaitoksen Tiedonantoja – The Finnish Forest Research Institute, Research Papers. pp. 52–60.

Tomppo, E., 2000. National Forest Inventory of Finland and Its Role Estimating the Carbon Balance of Forests. Biotechnol. Agron. Soc. Environ. 4 (4), 281–284.

Tomppo, E., 2006. The Finnish National Forest Inventory. In: Proceedings of the Eighth Annual Forest Inventory and Analysis Symposium. pp. 39–46.

Tomppo, E., Gagliano, C., De Natale, F., Katila, M., McRoberts, R., 2009. Predicting Categorical Forest Variables Using an Improved k-Nearest Neihgbour Estimator and Landsat Imagery. Remote Sensing of Environment 113, 500–517.

Tomppo, E., Halme, M., 2004. Using Coarse Scale Forest Variables as Ancillary Information and Weighting of Variables in k-NN Estimation: a Genetic Algorithm Approach. Remote Sensing of Environment 92, 1–20.

Tomppo, E., Heikkinen, J., 1999. National Forest Inventory of Finland – Past, Present and Future. In: Alho, J. (Ed.), Statistics, Registries, and Science – Experiences from Finland. Statistics Finland, Helsinki, pp. 89–108.

Tuominen, S., Pekkarinen, A., 2005. Performance of Different Spectral and Textural Aerial Photograph Features in Multi-Source Forest Inventory. Remote Sensing of Environment 94 (2), 256 – 268.

Vehtari, A., Heikkonen, J., Lampinen, J., Juujärvi, J., 1998. Using Bayesian Neural Networks to Classify Forest Scenes. In: Casasent, D. P. (Ed.), Intelligent Robots and Computer Vision XVII: Algorithms, Techniques, and Active Vision. Vol. 3522. SPIE, pp. 66–73.

Veltheim, T., 1987. Pituusmallit Männylle, Kuuselle ja Koivulle. Master's thesis, Helsingin Yliopisto, Metsänarvioimistieteen Pro Gradu – Tutkielma, 59 p. (In Finnish).

Wehr, A., Lohr, U., 1999. Airborne Laser Scanning – an Introduction and Overview. ISPRS Journal of Photogrammetry & Remote Sensing 54, 68 – 82.

Yu, X., Hyyppä, J., Hyyppä, H., Maltamo, M., 2004. Effects of Flight Altitude on Tree Height Estimation Using Airborne Laser Scanning. In: Laser Scanners for Forest and Landscape Assessment. Proceedings of the ISPRS working group VIII/2. Freiburg, Germany, pp. 96–101.

PART II: PUBLICATIONS