CALIBRATION OF LiDAR HISTOGRAM DATABASES USING A NON-LINEAR MATHEMATICAL MODEL

Linus Nyarusanda

M.Sc. (Mathematical Modelling) Dissertation

University of Dar es Salaam

September 2011

CALIBRATION OF LiDAR HISTOGRAM DATABASES USING A NON-LINEAR MATHEMATICAL MODEL

By

Linus Nyarusanda

A Dissertation Submitted in (Partial) Fulfilment of the Requirements for the Degree of Master of Science (Mathematical Modelling) of the University of Dar es Salaam

University of Dar es Salaam

September, 2011

CERTIFICATION

The undersigned certify that they have read and hereby recommend for acceptance by the University of Dar es Salaam the dissertation entitled: ***Calibration of LiDAR Histogram Databases using a Non-linear Mathematical Model***, in partial fulfillment of the requirements for the degree of Master of Science (Mathematical modelling) of the University of Dar es Salaam.

.....

Prof. T. Kauranne

(Supervisor)

Date: June 27, 2011

.....

Dr. Mahera, C. W.

(Supervisor)

Date: 23-09-2011

DECLARATION

AND

COPYRIGHT

I, **Linus Nyarusanda**, declare that this dissertation is my own original work and that it has not been presented and will not be presented to any other university for a similar or any other degree award.

Signature _____

ACKNOWLEDGEMENT

The work reported in this thesis was carried out in collaboration with the Department of Mathematics University of Dar es slaam and Laboratory of Technomathematics and Technophysics (Technomathematics) at Lappeenranta University of Technology (LUT) from November 2010 to June 2011.

Many people have contributed in different ways to accomplish this work. First i would like to thank Prof. Tuomo Kauranne for his valuable instructions and also for providing me with the forest data used in this thesis. I would like to express my sincere gratitude to Dr. Virpi Juntilla for her support, readiness and encouragement without whom this work could not be possible.

I'm particularly indebted to Dr. Wilson C. Mahera for his technical support and continued guidance till the accomplishment of this work.

I gratefully thank the NORAD's Programme for Masters studies (NOMA) for sponsoring my masters programme at the University of Dar es salaam and LUT for admitting me as an exchange student through center for International Mobility - CIMO.

I would like to acknowledge my fellow students and friends both at the University of Dar es salaam and LUT. Isambi sailoni, Maombi Mkenyeleye, Goefrey Wingi and Saila Milau who gave me a warm welcome at LUT and my fellow exchange student Sihoja Hamis whose experience in pilau party gave me good moments and a memorable stay at LUT. Deogratias Emmanuel, Nshaija Muganyizi and Hugo assisted me in different ways and all mathematical modelling group for their lovely interaction.

I appreciate a happy and lovely moments from my father Joseph kisoma and my

mother Penina kisoma and all my brothers at home for encouraging and giving me all the necessary support. I will never forget their love and care.

DEDICATION

To my Parents

Mr.Joseph kisoma and Penina Kisoma

ABSTRACT

The airborne laser scanning technology (also known as Light Detection and Ranging-LiDAR) in Finland has been proved to be the efficient remote sensing technique used for forest data collection. The data obtained from this technology is used for prediction of forest stand characteristics for different forest inventories. Using airborne Laser scanning for the new inventory is very expensive. However as different forests have been laser scanned there is a need of utilizing such previous databases in new inventory areas. Since the new site and the database sites have been laser scanned with different scanning instruments with different flying altitudes, there is a need of calibrating the databases to match with the new site LiDAR histograms. Therefore, the aim of this study is to the calibrate the LiDAR histogram databases by a non-linear mathematical model using the mean values of the LiDAR histograms. The accuracy of the calibrated LiDAR data is verified by predicting the forest stand characteristics using sparse Bayesian regression. The results obtained show that it is possible to calibrated histogram databases and the calibrated LiDAR histograms can be used in new inventory areas for the forest stand parameters estimation. It is observed that when the small number of calibration set from the new area (approximately 50 plots) is combined with the calibrated plots from the database, the estimation accuracy is almost equal to that when using the whole new area plots. This process is cost effective since instead of scanning the whole new inventory area, only randomly selected plots (approximately 50) can be scanned and be complemented with calibrated database plots for prediction of the forest stand characteristics. In some forests the geographical location does not support the aeroplanes to scan the forests, hence using calibrated databases can help to predict the forest characteristics of such areas.

Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

dgM- Diameter of basal area (G) Median tree

G - average breast height average area per hectare

GPS - Global Positioning System

hgM - Height of basal area (G) Median tree

LiDAR - Light Detection And Ranging

k-MSN - k-Most Similar Neighbours

k-NN - k-Nearest Neighbours

LOO - Leave Out One

N - Number of stems per hectare

OLS- Ordinary Least Squares

PLS- Partial Least Squares

SUR -Seemingly Unrelated Regression

RMSE - Root Mean Square Error

V - Volume

CHAPTER ONE

INTRODUCTION

1.1 General Introduction

Airborne laser scanning (ALS) also known as Light Detection and Ranging (LiDAR) has become the most accurate remote sensing technology for forest inventories. LiDAR is based on a set of laser pulses transmitted from aeroplane flying above the target area. The technology returns a three dimensional cloud of pulses reflected back from reflective objects beneath the flight path. The information is recorded and pre-processed with respect to the measurement conditions, and produce geographical coordinates and the height of the hitting point augmented with the intensity of the returning pulse echo. The percentage of the measurements recorded are affected by not only the flight altitude, but also the scanning instrument.

Forest inventory in Finland has traditionally relied on the partially visual assessment of species-specific stand characteristics by compartments in the field (Kangas *et al.*, 2004), however many alternative remote sensing techniques have been suggested, the most promising one is the LiDAR - based forest inventory method (Holmgren, 2004; Maltamo *et al.*, 20006). The area-based method has been the main approach to using ALS for large forest inventories because of its cost efficiency and the accuracy of the forest data obtained (Suvanto, A. and Maltamo, M. 2010). This has made it one of the principal operational methods used to collect information on forest structure and resources.

Mathematical models are needed to estimate forest characteristics (forest stand parameters) from the LiDAR data. Different kinds of regression models like Ordi-

nary Least Squares (OLS), seemingly Unrelated (SUR), k-Most Similar Neighbours $k-MSN$ and sparse Bayesian regression have been used to predict forest stand characteristics. Several studies have indicated that essential forest characteristics such as mean height, basal area, mean diameter, number of stems and stand volume can be accurately predicted using LiDAR data. Currently, new decision support systems are being developed, data collected from LiDAR provides comprehensive information about the state and dynamics of forests to assist decision making for strategic and management planning like accurate allocation and harvesting and thinning operations, and to optimize among a wide range of silvicultural treatments. LiDAR based inventories in operational projects have been carried out in Norway, Sweden and Finland (Næsset 2007, Maltamo *et al.,* 2007).

LiDAR histograms consists of data containing discrete returns of first and last pulse heights and intensities. The technology usually returns a three-dimensional cloud of point measurements from reflective objects scanned by the laser beneath the flight path. The percentage of the the pulses recorded as reflected at a particular point varies according to a number of factors including the physical weather, scanning parameters, flying altitude, footprint size, ground and forest characteristics. The LiDAR histogram distribution is affected by the quality of the LiDAR data obtained, therefore different sites may have different data qualities. Also different forests have different characteristics like tree species composition, development classes and annual heat sum variation. The variation of these characteristics in one site plot information requires coincidence of forest characteristics between the new site characteristics and the database plots, hence calibration involves database forests which have similar characteristics to the new site. The size of the first and last pulse columns may vary from one plot to another. In order to get a clear representativeness of these variables, this variation is controlled by taking the mean values of the LiDAR histograms, hence during calibrating, the mean values of the LiDAR histograms are used to calibrate the database histograms. In this study, a non linear mathematical model is used to

calibrate the LiDAR histogram databases in which the variables derived from the LiDAR measurements for each sample plot are the mean values of the first pulse intensity and height and the last pulse height and intensity, i.e., $v = \{H_f, H_l, I_f, I_l\}$. These variables are the non-ground hits of height above 2m from the ground. Here H refers to the height of the hit, and I to the intensity of the hit,and f to the first pulse and l to the last pulse .

1.2 Statement of the problem

Airborne laser scanning technology is currently the most efficient method of data collection used in different forest inventories. In order to improve the estimation of the LiDAR-based stand variable, regression methods have been employed, in which sparse Bayesian regression method was found to be better than other traditional methods. However using LiDAR data and sparse Bayesian methods need a large collection of sample plots (usually hundreds). Scanning such sample plots is very expensive. In order to reduce the operational costs there is a need of using few representative sample plots from the new area (calibration plots) and other plots from the databases that have been used in earlier missions. Since the new site plots and the databases are different and they have been collected in different conditions with different scanning instruments, they can lead to unrepresentative results. Hence the databases need to be calibrated to match with the new site characteristics before using them together for estimation of forest stand parameters. Therefore, the aim of this study is to calibrate the LiDAR histogram databases using a non-linear mathematical model.

1.3 Objectives of the study

1.3.1 General objective of the study

The general objective of this study is to calibrate LiDAR histogram databases for estimation of forest stand parameters.

1.3.2 Specific objectives

The specific objective include:

- Developing a non-linear mathematical model for calibration of the LiDAR histograms of the four databases.
- To verify the results obtained by using Sparse Bayesian Regression method.

1.4 Significance of the study

This study is very useful, because information on the forest statistics and other information produced are widely used to advise the ministry of natural resources about the usefulness of the technology and hence contribute to

- Forest policy making at both national and international levels
- Regional and national forest management planning
- Planning of forest industry investments
- Assessing sustainability of forestry and in forest certification

Studying this problem will develop expertise and contribute significantly to researchers for further research purposes, this is because the technology is currently

one of the most applicable remote sensing techniques for forestry inventory.

Now days many international statistics (e.g. FAO and Eurostat), processes (e.g. Ministerial Conference on the Protection of Forests in Europe, MCPFE) and agreements (e.g. Kyoto protocol) require information about development of forest resources. This study can provide useful information to serve the purpose.

CHAPTER TWO

LITERATURE REVIEW

Different studies have been released concerning the accuracy of the forest data produced by this technology, and different kinds of regression models like Ordinary Least Squares(OLS), seemingly Unrelated(SUR), $k-MSN$ etc have been applied for modeling forest stand characteristics from Airborne Laser Scanning. The results have indicated that the area-based method could produce at least as accurate results as traditional field measurements.

Maltamo *et al.*, (2009) studied on Predicting tree attributes and quality characteristics of Scots pine using airborne laser scanning data. In their study two modeling methods non-parametric $k-MSN$ and SUR were compared and the variables were estimated simultaneously. The results indicated that the $k-MSN$ method can provide more accurate tree-level estimates than SUR models. The $k-MSN$ estimates were in fact highly accurate in general, the $RMSE$ being less than 10% except in the case of tree volume and height of the lowest dead branch.

Naesset *et al.*, (2005) compared the ordinary least squares(OLS), seemingly unrelated (SUR), and the partial least-squares methods in the modeling of stand variables. Only small divergences were observed between these estimation techniques.

Different modeling techniques have been employed to improve the LiDAR data based stand variable estimation. Junttila, *et al.*, (2009) employed sparse Bayesian regression (as introduced by Tipping 2001) approach to predict the stand characteristics from LiDAR data. The accuracy of the method was found to be as good as or even slightly better than that of conventional regression model based methods. Maltamo *et al.*, (2006a) and Packerlen and Maltamo (2006b) applied non-parametric k-most similar neighbor (MSN) estimation to predict stand characteristics. The

accuracy was found better than that of the corresponding OLS models, and it was observed that it is possible to estimate several stand variables simultaneously. Also, species specific variables were compatible with total stand characteristics. Sulvanto and Maltamo (2010), studied on Using mixed estimation for combining airborne laser scanning data in two different forest areas in which mixed estimation and OLS were compared in predicting six forest stand parameters in the new area. These parameters include basal area median tree diameter and height, mean tree height, stem number, basal area and volume. The LiDAR height histograms and number of chosen sample plots from the new area (10-212 plots) were combined with the 472 database plots. The results of the study indicated that in LiDAR- based forest inventory, mixed estimation with a combination of datasets from the existing database and the new area improved the accuracy of derived plot-level characteristics compared with OLS-based regression models. The study had neither calibration of LiDAR histograms nor plots selection based on the new site forest stand parameter distribution.

Junttila, *et al.*, (2010) studied on estimation of forest stand parameters from airborne laser scanning using calibrated plot databases. In their study three databases from different sites were calibrated and combined with the new site information to form a teaching set for estimating five forest stand parameters, these were mean height, basal area, mean diameter, number of stems and stand volume. This methodology reduced the number of sample plots (calibration set) needed from the new site substantially (between 3 and 9 stands including 20 to 70 plots) which were supported by the plots from the three databases. The LiDAR histograms of each site was calibrated using a linear mathematical model and sparse Bayesian regression model was used to predict the stand parameters and they observed that it is possible to effectively use the sample plots from earlier studies plot databases to predict forest stand characteristics in the new area. This method was found to correct the errors due to the use of different scanner, scanning parameters and scanning conditions, furthermore, the

bias due to the use of alien plots was significantly reduced. This method was found to save the plots collection costs, when airborne laser scanning and Bayesian regression are used to estimate the total forest stand parameters. The calibration of the LiDAR histograms performed was based on the percentile points.

The aim of this study is to calibrate the databases of the LiDAR histograms using a non-linear mathematical model based on the mean values of the LiDAR histograms. The assumption is that the mean values of the LiDAR histograms represent the salient features of the forest better than the percentiles, hence we expect that using mean values will give better calibration results than using the percentiles.

CHAPTER THREE

Model Development

3.1 The study area and the field data

The airborne laser scanning data source in this study come from Arbonaut Company in Finland which five different sites Matalansalo, Juuka, Loppi, Pello and karttula were measured. The sample plots from these areas were measured with different sampling strategies, some sites were equipped with a regular sample plot grid, some others were equipped with regular sample plot clusters while others were randomly selected. These differences were ignored in the estimation process as they are likely to vary in operational use as well. The sample plots were circular with a radius of $9m$. The centre of each plot was determined by the hand-held GPS, and the exact position of the plot centre was later calculated during measurement and differentially corrected off-line. This method generally ensured a positioning accuracy of less than one meter, which has been deemed adequate to align LiDAR data and field plots so that their areas overlap to a degree exceeding ninety percent. LiDAR data was clipped to plot extent before extracting LiDAR parameters from it.

In this study there were 20 forest stand parameters, based on field measurements. The forest stand parameters consist of the total parameters median tree diameter (Dgm), median tree height (Hgm), stem number(N), basal area(G) and volume(V) supplemented with corresponding species-specific parameters:
Dgm1, Dgm2, Dgm3, Hgm1, Hgm2, Hgm3, N1, N2, N3, G1, G2, G3, and V1, V2, V3. Here the indices 1-3 refer to the species :1 for scots pine (pinus silvestris), 2 for Norway spruce (Picea Abies) and 3 for hardwoods treated as a group, but mostly comprising birch (betula pendula or Betula pubescent). The LiDAR scanning of the different areas was conducted from 2004 to 2008. Three different types of

scanners were used, they include the Optech ALTM 3600, the Leica ALS-50, and the Leica ALS-60. Flying height varied between 700m and 2000m and the scanner pulse frequency varied between 58900Hz and 125100Hz. The characteristics of different sites are shown in the table 3.1.

	Matalansalo	Juuka	Loppi	Pello	karttula
Annual heat sum (degree days)	1150	1000	1250	850	1100
Mean volume m^3/ha	203.4	145.5	203.2	102.8	205.9
mean timber height m	17.0	14.7	17.4	11.7	19.7
Mean basal area (m^2/ha)	24.7	20.3	22.9	17.8	24.0
Mean number of stems	1506.9	1284.6	1109.0	1325.2	1150.7
Max.timber height m	30.6	25.2	33.7	20.9	35.5
Scots pine vol- %	53.2	67.2	45.3	38.6	29.1
Norway spruce vol-%	34.5	21.7	41.5	34.1	45.0
Hardwoods vol-%	12.3	11.0	13.2	27.3	25.9
No of measured plots	472	511	441	553	538

Table 3.1: Mean characteristics of the test sites

3.1.1 LiDAR histogram

LiDAR histograms consist of data containing discrete returns of first and last pulse heights and intensities. The technology usually returns a three-dimensional cloud of point measurements from reflective objects scanned by the laser beneath the flight path. The percentage of the the pulses recorded as reflected at a particular point vary according to a number of factors including the physical weather, scanning parameters, flying altitude, footprint size, ground and forest characteristics. The Li-

DAR histogram distribution is affected by the quality of the LiDAR data obtained, therefore different sites may have different data qualities. Also different forests have different characteristics like tree species composition, development classes and annual heat sum variation. The size of the first and last pulse columns may vary from one plot to another. In order to get a clear representativeness of these variables, this variation is controlled by taking the mean values of the LiDAR histograms, hence during calibrating, the mean values of the LiDAR histograms are used to calibrate the database histograms. In this study a non-linear mathematical model is used to calibrate the LiDAR histograms of the variables in which the variables derived from the LiDAR measurements for each sample plot are the mean values of the first pulse intensity and height and the last pulse height and intensity, i.e., $v = \{H_f, H_l, I_f, I_l\}$. Here H refers to the height of the hit, and I to the intensity of the hit, and f to the first pulse and l to the last pulse. These variables are the non-ground hits of height above 2m from the ground.

3.1.2 Selection of calibration set from the new area

Before calibration of the LiDAR histograms, a new area N is laser scanned and the histograms $D_{n,i}$ of the LiDAR data measurements of its plots $i = 1, .., n$ are served as

$$D_{ni} = \{H_{f,ni}, H_{l,ni}, I_{f,ni}, I_{l,ni}\}. \tag{3.1}$$

Then a calibration set from the new site is selected. The fieldwork is performed in the new site at a selected set of stands. The set must contain plots that well represent the variability of the stand parameters in the whole new area. The aim is to find a representative set of stands (plots) from the new area by random selection that has sufficient properties to justify the new area characteristics, this helps us to gain information about the LiDAR measurement properties and the forest stand parameters distributions of the whole new site. The calibration set histograms $D_{c,i}$

of each plot i in the new area LiDAR information is defined as

$$D_{c,i} = \{H_{f,ci}, H_{l,ci}, I_{f,ci}, I_{l,ci}\} \qquad \forall c \in n. \tag{3.2}$$

The considered LiDAR measurements are first and last pulse height and intensities and let the new site be Loppi. Assume only one LiDAR measurement, first pulse height H_f. The figure (3.1) shows the distribution of the mean values of the first pulse height H_f from the new site.

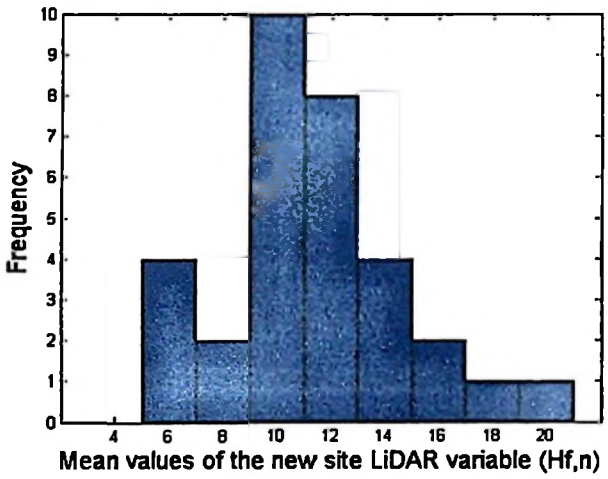

Figure 3.1: The new site first pulse height LiDAR Variable (Loppi)

Figure (3.1) shows the first pulse height H_f histograms in the new site, this information provides the calibration criteria for the database first pulse height.

3.1.3 The databases

The database consists of the LiDAR histograms used in old sites in earlier missions. These forests are assumed to have similar characteristics with the new site. The

mean values of the database histograms $D_{d_j,i}$ for each plot i is given as

$$D_{d_j,i} = \{H_{f,d_j i}, H_{l,d_j i}, I_{f,d_j i}, I_{l,d_j i}\} \tag{3.3}$$

for each measurement variable. The figure (3.2) shows the distribution of the the mean values of one database site, the first pulse height H_{If,d_j}

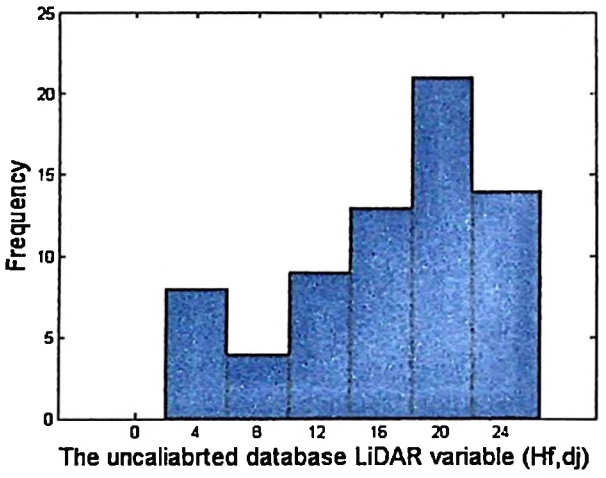

Figure 3.2: The Database LiDAR Variables

Therefore, figure (3.2) represents database histograms of one database site, the first pulse height H_{Hf,d_j}

3.2 Selection of plot pairs

3.2.1 Introduction

After getting the calibration set from the new site where the forest stand parameters are known, plot pairs for which the LiDAR histograms of the new site and the databases are found. The plot pairs will be used for the calibration of the LiDAR

histogram databases. Calibration of the LiDAR histogram is based on the most similar pairs for which the LiDAR histograms of the new site and database are assumed to be equivalent. The aim is, having a calibration set from the new area where the forest stand parameters are known, this set will be used to form the plot pair with the database. The plot pair will be used for the calibration of the LiDAR histogram of database. Therefore, once the forest stand parameters are known from the calibration set of the new area, Y_c they are combined with those from the database sets Y_{d_j}. A set of forest stand parameter pair is found from each database. If for some calibration plots no similar pair is found, that plot is not used for stand parameter estimation.

The process of selection of most similar pairs of each calibration plot is done using information on forest stand parameter values and accuracy of LiDAR histogram based on sparse Bayesian Regression models.

3.2.2 pairs selection

The process of pairs selection used in this study is based on the model developed by Junttila, *et al.*, 2010 who used the linear mathematical model to calibrate the LiDAR histogram databases. The calibration pairs are selected from each database, j for each plot in the calibration plots, where a most similar neighbour plot $d_{j_c} \in d_j$ in terms of measured forest stand parameter values is chosen. When estimations with LiDAR data are evaluated, forest stand parameter that are well estimated with LiDAR should be equal between pairs. For each forest stand parameter k, the regression estimates of each database is defined as

$$Y_{d_j,k} = X_{d_j} w_{d_j,k} + \varepsilon_{d_j,k} \tag{3.4}$$

where X_{d_j} is a full normalized database LiDAR variable, the linear weight vector $w_{d_j,k}$, and the noise term $\varepsilon_{d_j,k}$ which is assumed to be normal $N(0, \sigma^2_{k,d_j})$. The

weight of the stand parameter k in pair selection is the inverse of the regression estimate error variance, σ_{k,d_j}^{-2}, and most similar neighbor is defined by minimizing

$$d_{jc,i} = \underset{d_j}{\text{argmin}} \sum_k \frac{1}{\sigma_{k,d_j}^2}(Y_{d_j,k} - Y_{cik})^2 \tag{3.5}$$

for each plot i in the calibration set. Here Y_{cik} is the forest stand parameter k in plot i in the calibration set, argmin is the argument of the minimum for which the value $d_{jc,i}$ attains its minimum value. This means that for the LiDAR variables to correlate well with the the stand parameter k (σ_k^2 small) the selection of the pairs is done in such a way that the stand parameter k is almost equal in both members of the pair. Contrary to this notion, if the estimation is not expected to correlate well with the stand parameter k (σ_k^2 large), the effect of the stand parameter k is not large.

The set of the five stand parameters considered in this study of each calibration plot, are treated as vector of the multinormal distribution, $\mu_{y,i}$ with expected variance matrix given in the LiDAR estimate of the database,
$\Sigma_y = diag((\sigma_{1,d_j}^2, ..., \sigma_{k,d_j}^2)^T)$, where *diag* represents the diagonal elements. The chosen database forest stand parameter values of each pair, $Y_{d_{jc},i}$, need to be close to the values of the of the calibration plot which is the center of the distribution. According to Mardia et al., (1980) the pair distance from the center is χ^2 distributed with K degrees of freedom

$$U_i = (Y_{d_{jc},i} - \mu_{y,i})^T \Sigma_y^{-1} (Y_{d_{jc},i} - \mu_{y,i}) \sim \chi^2(K) \tag{3.6}$$

whose range probability is within 10% from each other, that is

$$U_i \leq \chi_{0.1}^2(K), \tag{3.7}$$

and otherwise discarded. In this case $\chi_{0.1}^2(K)$ is the point where the cumulative distribution of $\chi^2(K)$ has the valuc 0.1. Thc pairs fulfilling this condition are labeled as $\hat{c}_j$ and $\hat{d}_{jc}$ and the number of these pairs is $\hat{N}_{cj}$. We need at least 10 pairs with the smallest possible distances that are chosen into the measurement calibration

procedure to get enough amount of data for calibration. If no sufficient pairs we should re consider whether the intended database is suitable for use in the current case for the new area.

3.3 Database LiDAR histogram calibration

3.3.1 Introduction

During histogram calibration only the ground points, i.e., height below $2m$ were neglected. These ground hits were assumed to have no direct correlation with tree qualities as well as quantities because histogram calibration requires pairs selection by tree characteristics.

The challenge encountered during selection of calibration plots from the new area is about handling variables drawn from the last pulse LiDAR measurements. Some plots may have either only a few or no measurements with last pulse heights above $2m$. In such cases, there are no enough non-ground data for height and intensity variables. These plots cannot be used for LiDAR calibration, since there are no pairs for last pulse height and intensity in calibration procedures.

3.3.2 Calibration process

The calibration process is based on the linear equation $X_{H_f,\hat{c}_j} = h_{H_f,d_j} X_{H_f,\hat{d}_{jc}}$ developed by Juntilla, *et al.*, 2010. Consider one LiDAR measurement first pulse height H_f. To get information on the correlation between the LiDAR variable X and forest stand parameter Y of the new site, a multidimensional regression estimate is carried out for the selected plots $\hat{c}_j$ in the calibration set which can be defined by normal

likelihood $N(Y_{\hat{c}_j}|X_{H_f,\hat{c}_j}, W_{H_f,\hat{c}_j}, \Sigma_{H_f,\hat{c}_j})$, hence

$$Y_{\hat{c}_j} = X_{H_f,\hat{c}_j} W_{H_f,\hat{c}_j} \tag{3.8}$$

and the model covariance

$$\Sigma_{H_f,\hat{c}_j} = diag\left\{ \frac{(Y_{\hat{c}_j} - X_{H_f,\hat{c}} W_{H_f,\hat{c}_j})^T (Y_{\hat{c}_j} - X_{H_f,\hat{c}} W_{H_f,\hat{c}_j})}{\hat{N}_{cj}} \right\} \tag{3.9}$$

is a diagonal matrix,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 LiDAR variable ability to explain the parameter, and the weight is given by

$$W_{H_f,\hat{c}_j} = (X^T_{H_f,\hat{c}_j} X_{H_f,\hat{c}_j})^{-1} X^T_{H_f,\hat{c}_j} Y_{\hat{c}_j} \tag{3.10}$$

The statistical variables of the database are corrected by the calibration coefficient h_{H_f,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, i.e., the calibrated database variables and LiDAR histogram measurements are assumed to be linearly correlated with the calibration coefficient $X_{H_f,\hat{c}_j} \sim h_{H_f,d_j} X_{H_f,\hat{d}_{jc}}$, then

$$X_{H_f,\hat{c}_j} = h_{H_f,d_j} X_{H_f,\hat{d}_{jc}}, \tag{3.11}$$

Taking the assumption that this identity is true, the forest stand parameter estimates $\hat{Y}_{\hat{c}_j} = X_{H_f,\hat{c}_j} W_{H_f,\hat{c}_j}$ equal to

$$\hat{Y}_{\hat{d}_{jc}} = h_{H_f,d_j} X_{H_f,\hat{d}_{c_j}} W_{H_f,\hat{c}_j} \tag{3.12}$$

The errors between the forest stand parameters of pairs $e_i = Y_{\hat{c}_j i} - Y_{\hat{d}_{jc} i}, i \in \hat{c}$ are assumed to be multinormally distributed with mean $\hat{e}_i = \hat{Y}_{\hat{c}_j} - \hat{Y}_{\hat{d}_{jc}}$, that is $e \sim \prod_{i=1}^{\hat{N}_{cj}} N(e_i|\hat{e}_i, \Sigma_{H_f,\hat{c}j})$ hence,

$$e \sim \prod_{i=1}^{\hat{N}_{cj}} N\left(Y_{\hat{c}_j i} - Y_{\hat{d}_j i} | (X_{H_f,\hat{c}_j i} - h_{H_f,d_j} X_{H_f,\hat{d}_{jc} i}) W_{H_f,\hat{c}_j}, \Sigma_{H_f,\hat{c}j} \right) \tag{3.13}$$

where the mean of the distribution depends on the difference of the forest stand parameter estimates of plots, and the covariance is the diagonal matrix of the error variances of the estimates. Thus different forest stand parameters in each plot were weighted with a higher weight given to forest stand parameters that are well estimated with the LiDAR variables at hand and a lower weight to stand parameters that cannot be estimated well.

In order to get the calibration coefficient h_{H_f,d_j} the equation of the likelihood (3.13) is minimized subject to the condition that $h_{H_f,d_j} > 0$. The procedure is done for each LiDAR measurement variable $v = \{H_f, H_l, I_f, I_l\}$.

The weight due to different error variances on different forest stand parameters can be included in the model by variable transformations:

$$Y_{z,c_j} = Y_{\hat{c}_j} \Sigma^{-1/2}_{H_f,\hat{c}_j} \tag{3.14}$$

$$Y_{z,d_j} = Y_{\hat{d}_{jc}} \Sigma^{-1/2}_{H_f,\hat{c}_j} \tag{3.15}$$

$$W_z = W_{H_f,\hat{c}_j} \Sigma^{-1/2}_{H_f,\hat{c}_j} \tag{3.16}$$

$$\hat{Y}_{z,c_j} = X_{H_f,\hat{c}_j} W_z \tag{3.17}$$

$$\hat{Y}_{z,d_j} = X_{H_f,\hat{d}_{jc}} W_z \tag{3.18}$$

solving the multinormally distributed model (3.13) we get an estimate for the calibration coefficient $\hat{h}_{H_f,d_j}$

$$\hat{h}_{H_f,d_j} = trace(\hat{Y}^T_{z,d_j} \hat{Y}_{z,d_j})^{-1} trace(\hat{Y}^T_{z,d_j} dY_{H_f,\hat{d}_{jc}}) \tag{3.19}$$

where $dY_{H_f,\hat{d}_{jc}} = \hat{Y}_{z,c_j} - Y_{z,c_j} + Y_{z,d_j}$

The calibrated database LiDAR histograms distribution is assumed to be equal to that of the LiDAR measurement of the new site. The calibrated database is defined as

$$\hat{D}_{d_j,i} = \{h_{H_f,d_j} H_{f,d_j i}, h_{H_l,d_j} H_{l,d_j i}, h_{H_f,d_j} I_{f,d_j i}, h_{H_l,d_j} I_{l,d_j i}\}. \tag{3.20}$$

3.4 Database Plot selection

The challenge in plot pair selection between the new site and the databases is that the distribution of the forest stand parameters in the new site is not always equal to the databases. Hence Error and bias are likely to occur when using the database plots to estimation the new site characteristics. We need to select only the plots which are expected to represent new site characteristics. Data from each site is independent from the other site. The data from the new site is selected according to the available information from the new site (which is the new site LiDAR histograms). Database plots that are taken into the model must fit the characteristics of the new site. If this data fit the distribution of the new site characteristics perfectly, estimation results would approach to those achieved using a dense set of field sample plots from the new site. The best way 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. Nevertheless there is no prior knowledge of the correlation between the auxiliary variables and different forest stand parameters of the new site, i.e., whether the given variable is needed in the given estimation model at all, and the distributions of different auxiliary variables may be very complicated. Hence the use of auxiliary target plot variables as selection criteria, i.e., in verification of new site qualities versus database qualities, is very complicated. 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 can lead to unrepresentative result to the new site. Considering the five forest stand parameters Dgm, Hgm, N, G, and V, the probability of the distribution of the new site can be established using the calibration plot. If the calibration set is not representing the new site, it is likely that the probability distribution will also be unrepresentative, thus the error in calibration sample plot selection strategy

accumulates to plot selection of the databases. The calibration set size N_c may be smaller than the selected plots from each database $\hat{d}_j$, and when using all the j new databases available, its relative size becomes even smaller. Since the databases contain only alien plots, one should never rely more on them than the calibration set.

Having known the distribution of the forest stand parameters of the calibration plots of the new site and transforming each of them to normal or close to normal. The distribution $(\sqrt{Dgm}, Hgm, \sqrt{N}, G, \sqrt{V})$, one may heuristically calculate an estimation whether a given plot i in each database belongs to the same distribution using uniformly distributed random variables $r_i \in [0, 1]$: The database plot d_{ji} is accepted to the model if

$$exp\{-trace[(Y_{d_{j_i}} - \bar{Y}_c)\Sigma_Y^{-1}(Y_{d_{j_i}} - \bar{Y}_c)^T/2/K]\} > r_i \tag{3.21}$$

In this case, $\bar{Y}_c$ is the mean and $\Sigma_Y = cov(Y_c)$ is the covariance of the measured forest stand parameters of the calibration set from the new area. The plots accepted from the calibrated database are $\hat{d}_j$.

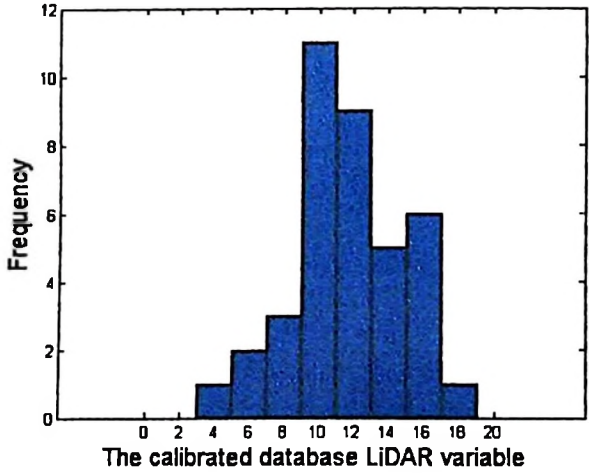

Figure 3.3: The calibrated Database LiDAR Variable

Figure (3.3) shows the LiDAR histograms of the calibrated database. From this it is observed that the calibrated database histograms is similar to the new site LiDAR histograms.

3.5 The calibration equation

Two non-linear equations are considered here and from which only one equation will be taken as the calibration equation. The calibration is done by using the mean values of the LiDAR histograms and the calibration equation is defined as follows: Let d_j be the databases information, the considered LiDAR measurements are first and last pulse height and intensities.
Consider one database LiDAR measurement first pulse height H_{Hf,d_j}. Let the calibration coefficient be $h_{H_{f,d_j}}$.

The first non-linear calibration equation is given by

$$\bar{X}_{H_{f,\hat{c}_j}} = m + a * \bar{H}_{f,d_j} + b * \bar{H}^q_{f,d_j} \tag{3.22}$$

where m is a constant term, $\bar{H}$ is the mean value of the LiDAR histogram $a, b \in \mathbb{R}$ $a, b > 0$ are calibration coefficients, and q is the calibration power, $0.5 \leq q \leq 1.2$, and $h_{H_{f,d_j}}$=[a b] and the corresponding figure is figure 3.4

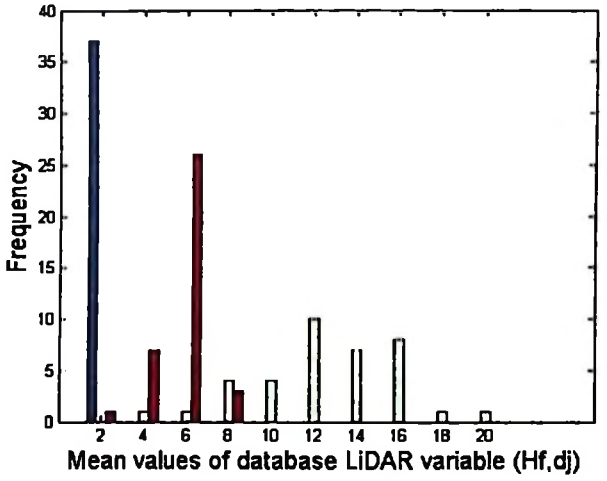

Figure 3.4: The mean value of the LiDAR histogram using the first equation

In the figure 3.4 , while the blue bar is the result of the constant term and the red bars are the result of the non linear term $\bar{H}^q_{f,d_j}$, the green bars are a result of the linear term $\bar{H}_{f,d_j}$.

The second non-linear calibration equation is given by

$$\bar{X}_{Hf,\hat{c}_j} = m + b * \bar{H}^q_{f,d_j} \tag{3.23}$$

where m is a constant term, $\bar{H}$ is the mean value of the LiDAR histogram, $b \in \mathbb{R}$ $b > 0$ is a calibration coefficient, and q is calibration power with $0.5 \leq q \leq 1.2$. and the calibrated mean values of the LiDAR histograms are given in the figure(3.5)

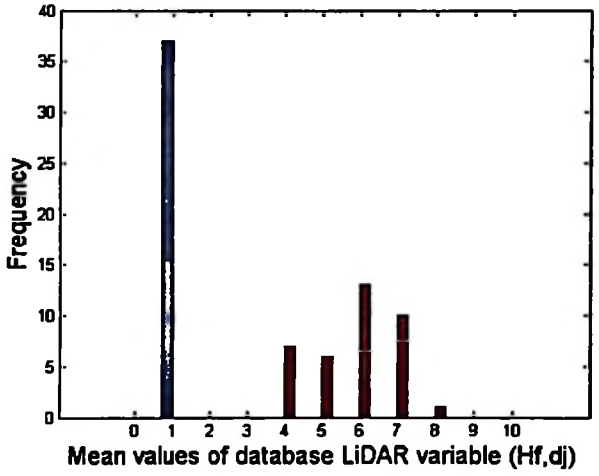

Figure 3.5: The mean value of the LiDAR histogram using the second equation

In figure (3.5), the blue bars come from the constant term while the red bars come from the power term $\bar{H}^q_{f,d_j}$.

The process of getting the calibration equation shows that the constant term seems to haves no significant effect in the calibration equation. This is because it shifts the position of mean values from one point to another, while the LiDAR variables remain unaffected. Considering the calibration equations 1 and 2, the mean values

of the first equation are better than that of the second equation because range of the mean values from the first equation is higher than the range of the mean values from the second equation (see table 4.1).

There fore the first equation is going to be used as the calibration equation where the constant term m will be neglected. The calibration equation is defined as follows;

$$\bar{X}_{H_f,\hat{c}_j} = a * \bar{H}_{f,d_j} + b * \bar{H}^q_{f,d_j} \tag{3.24}$$

where $\bar{H}$ is the mean value of the LiDAR histogram, $a, b \in \mathbb{R}$ are calibration coefficients $a, b > 0$, and q is the suitable calibration power with $0.5 \leq q \leq 1.2$, and $h_{H_f,d_j} = [a \quad b]^T$.

The figure (3.6) shows the distribution of the mean values of H_f and H_f^q using the calibration equation.

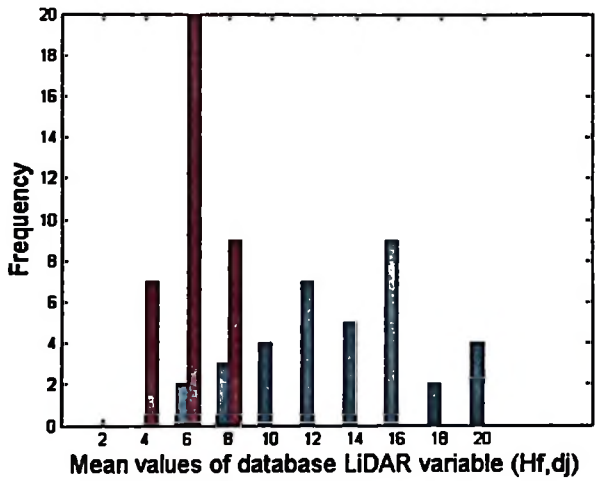

Figure 3.6: The mean value of the LiDAR histogram using the calibration equation

In figure (3.6), the blue bars represent the calibrated mean values as a result of the linear term $\bar{H}_{f,d_j}$ of the calibration equation while the red bars have been calibrated from the term $\bar{H}^q_{f,d_j}$.

Therefore, from the calibration process of section 3.3.2, the linear equation 3.11 is replace by the non-linear equation

$$\bar{X}_{H_f,\hat{c}_j} = h_{H_f,d_j}\begin{bmatrix}\bar{X}_{H_{f,d_{jc}}} & \bar{X}^q_{H_{f,d_{jc}}}\end{bmatrix} \tag{3.25}$$

where $h_{H_f,d_j} = [a \quad b]^T$.

Note that $\bar{H}_{f,d_j}$ is used when finding the mean values of the LiDAR histogram while $\bar{X}_{H_{f,d_{jc}}}$ is used during database calibration

Therefore, using equation 3.24, we have the following old and new sites information as compared in the figure (3.7). This is before calibration.

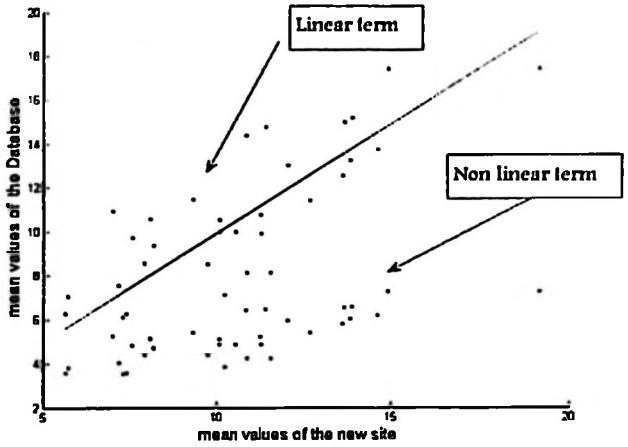

Figure 3.7: The New site and the uncalibrated database

As the figure (3.7) shows, the distribution of the values along the blue line resulting from the linear term $(\bar{H}_{f,n}, \bar{H}_{f,d_j})$ while the values below this line result from the non-linear term $(\bar{H}_{f,n}, \bar{H}^q_{f,d_j})$.

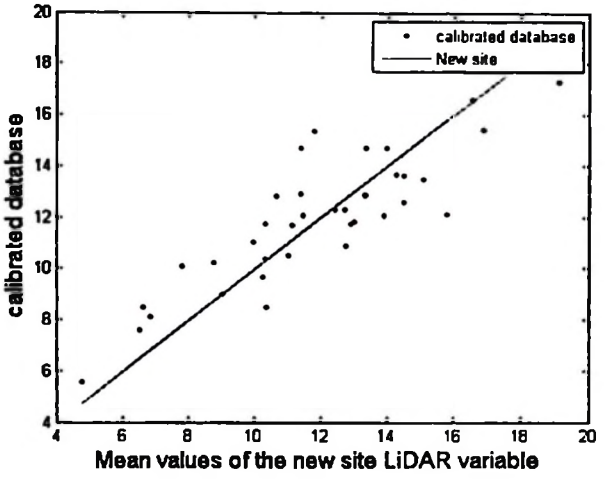

Figure 3.8: The new site and the calibrated database

Figure (3.8) provides the relationship between the new site and the calibrated database first pulse height.

CHAPTER FOUR

Data Analysis and Methodology

4.1 Testing the accuracy of the calibrated database by predicting forest stand parameters

To verify the accuracy of the calibrated database, a sparse liner regression method introduced by Tipping, i.e., Tipping's Sparse Bayesian regression algorithm (Tipping 2001, 2004) is used to estimate forest stand parameter. The results for the database estimates can be verified against optimal case for which a large number of field measurements from the new area are gathered.

4.1.1 A sparse Bayesian model for Regression

Consider a linear approximation used to model the components of the target vector

$$t = \Phi w + \epsilon \tag{4.1}$$

with a linear function $y = \Phi w$ of independent measurements, where $w = [w_{k,1}, ...w_{k,p}]$ are the forest stand parameter specific weights of the model, and each forest parameter $t_{k,i}$ on each plot $i, i = 1, 2...p$ in the teaching set, $t = [t_{k,1}, ...t_{k,p}]$ is the forest stand parameters which are independent of each other, the matrix $\Phi = [\phi_1, ..., \phi_M]^T$, is the $P \times M$ design matrix of dependent variables (LiDAR variables) with M columns, and P rows on each plot such that $\Phi_{im} = \phi_m(x_i)$ and the model error $\epsilon = [\epsilon_{k,1,...,\epsilon_{k,p}}]^T$ which is assumed to be Gaussian distribution with mean zero and variance σ^2. That is $p(\epsilon|\sigma^2) = N(0, \sigma^2)$. For notational convenience, the index k is dropped.

Hence, in Bayesian regression, the likelihood of the complete data is given by the

product

$$p(t|x,w,\sigma^2) = \prod_{i=1}^{p} N(\Phi_i w, \sigma^2) \tag{4.2}$$

here the term $p(t|x,w,\sigma^2)$ is written as $p(t|w,\sigma^2)$ since we never seek to model the input vector x, hence

$$p(t|w,\sigma^2) = (2\pi\sigma^2)^{-p/2} exp\left\{ -\frac{||t-\Phi w||^2}{2\sigma^2} \right\} \tag{4.3}$$

where the mean of the likelihood is the estimate Φw and the variance is σ^2.
In order to avoid over fitting and to control the model complexity, a penalty term is introduced

$$p(w|\alpha) = \prod_{m=1}^{M} N(0, \alpha_m^{-1}) \tag{4.4}$$

$$p(w|\alpha) = \prod_{m=1}^{M} (\frac{\alpha}{2\pi})^{1/2} exp\{-\frac{\alpha}{2} w_m^2\} \tag{4.5}$$

Given the likelihood and the prior the posterior distribution over w via the Bayes' rule is given by

$$p(w|t,\alpha,\sigma^2) = \frac{likelihood \times prior}{normalizing factor} = \frac{p(t|w,\sigma^2)p(w|\alpha)}{p(t|\alpha,\sigma^2)} \tag{4.6}$$

As a result of combining the a Gaussian prior and a liner model within a Gaussian likelihood, the posterior is also Gaussian: $p(w|t,\alpha,\sigma^2) = N(\mu,\Sigma)$ with

$$\Sigma = (\sigma^{-2}\Phi^T\Phi + A)^{-1} \tag{4.7}$$

$$\mu = \sigma^{-2}\Sigma\Phi^T t \tag{4.8}$$

and all the hyperparameters are collected into the diagonal matrix $A = diag(\alpha_1, ...\alpha_M)$, hence

$$p(w|t,\alpha,\sigma^2) = (2\pi)^{-(P+1)/2}|\Sigma|^{-1/2}exp\{\frac{(w-\mu)^T\Sigma^{-1}(w-\mu)}{2}\} \tag{4.9}$$

From (4.6)

$p(t|\alpha,\sigma^2) = \int p(t|w,\sigma^2)p(w|\alpha)dw$

$$= (2\pi)^{-P/2}|\sigma^2 I + \Phi A^{-1}\Phi^T|^{-1/2}exp\{-\frac{1}{2}t^T(\sigma^2 I + \Phi A^{-1}\Phi^T)^{-1}t\} \tag{4.10}$$

$$p(t|\alpha,\sigma^2) = N(0,\sigma^2 I + \Phi A^{-1}\Phi^T) \tag{4.11}$$

Differentiating $\ln p(t|\alpha,\sigma^2)$, of equation (4.11) with respect to α and σ^2, setting to zero and rearranging we get the re-estimated hyperparameter

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \tag{4.12}$$

$$(\sigma^2) = \frac{||t-\Phi\mu||^2}{N - \sum_{i=1}^{M}\gamma_i} \tag{4.13}$$

where$\gamma_i \in [0,1]$, μ_i is the mean component of the posterior of w_i and $\gamma_i = 1 - \alpha_i \sum_{ii}$,

predicting new target values

The estimates of the target vector are computed with the expected value of the weights, i.e., using the posterior distribution over the weights. The posterior is conditioned on those values of α_{MP}, σ^2_{MP} which maximize the likelihood (4.11)

The distribution that predicts the target vector when the datum Φ_* is given as the marginal likelihood

$$p(t_*|t,\alpha_{MP},\sigma^2_{MP}) = \int p(t_*|w,\sigma^2)p(w|t,\alpha_{MP},\sigma^2_{MP})dw, \tag{4.14}$$

the solution of this equation is Gaussian distribution

$$p(t_*|t,\alpha_{MP},\sigma^2_{MP}) = N(y_*,\sigma^2_*), \tag{4.15}$$

where $y_* = \mu^T \Phi_*$ and $\sigma_*^2 = \sigma_{MP}^2 + \Phi_* \Sigma \Phi_*^T$

This shows that the predicted target vector value is based on the mean value of the weight, and the error bars of the prediction consist of two variance components; estimated noise on the data and variance due to uncertainty of prediction of the weights.

4.1.2 Estimation of the forest Stand parameters:

According to Tipping, in ordinary regression estimates, the estimates for each forest stand parameter at hand are based on the linear equation

$$y = Xw + \epsilon, \tag{4.16}$$

where y is the vector, i.e., the forest stand parameter, X is the matrix containing candidate variables used in the estimation, w is the linear regression weight vector which is assumed to be sparse, and ϵ is the error vector with zero mean and variance σ^2. The ordinary regression estimate for the weight is

$$\hat{w} = (X^T X)^{-1} X^T y \tag{4.17}$$

In the Bayesian regression formulation, the likelihood of the data plot i is normally distributed, then we have

$$p(y_i|w, \sigma^2) = N(y_i|X_i w, \sigma^2) = (2\pi\sigma)^{1/2} exp\Big(- \frac{1}{2\sigma^2} ||y_i - X_i w||^2 \Big) \tag{4.18}$$

and the total likelihood of the sample s consisting of the N_c samples from the calibration set c together with the collection of the N_d samples from the available databases is

$$p(y_s|w, \sigma^2) = \prod_{i \in s} N(y_i|X_i w, \sigma^2) = N(y_s|X_s w, \sigma^2 I_N) \tag{4.19}$$

where N_c is the calibration set, N is the total likelihood of the sample.

4.1.3 Measuring error

when predicting forest stand parameters of the new site using calibrated databases error and bias are likely to occur. The objective is to achieve a smaller root mean square error, RMSE, in the new site verification set than when using only calibration set The predicted target vector values of verification plots are defined with model

$$Y_{P_s} = \Phi_{P_s}\mu_s, \tag{4.20}$$

where P_s is the set of plots in the verification set (stand). The stand level estimation of the target vector value is calculated as the mean of the estimated target vector values of the plots in the verification stand,

$$\bar{Y}_s = \frac{\sum_{P_s} Y_P}{P_s}. \tag{4.21}$$

the plot level error of the target vector values of the whole set of plots is expressed with the root mean square error (RMSE %) and bias (%)

$$RMSE_{plot} = \sqrt{\frac{\sum_P (t_P - Y_P)^2}{P}} \tag{4.22}$$

$$RMSE_{plot}(\%) = \frac{RMSE_{plot}}{\bar{t}_P} \times 100\% \tag{4.23}$$

$$BIAS_{plot} = \sqrt{\frac{\sum_P (t_P - Y_P)}{P}} \tag{4.24}$$

$$BIAS_{plot}(\%) = \frac{BIAS_{plot}}{\bar{t}_P} \times 100\% \tag{4.25}$$

4.2 The validation process

The procedure used to estimate the forest stand parameters was tested using the cross validation procedure using only the LiDAR variables. One site at a time was used as the new site and the others were left as database set. The calibration plots are complemented with plots from each of the four database from other missions on the other test site. In this study there were 20 forest stand parameters based on field measurements. The forest stand parameters consist of the total parameters median tree diameter (Dgm), median tree height (Hgm), stem number(N), basal area(G) and volume(V) supplemented with corresponding species-specific parameters Dgm1, Dgm2, Dgm3, Hgm1, Hgm2, Hgm3, N1, N2, N3, G1, G2, G3, and V1, V2, V3. Here the indices 1-3 refer to the species :1 for scots pine(pinus silvestris), 2 for Norway spruce(Picea Abies) and 3 for hardwoods treated as a group, but mostly comprising birch(betula pendula or Betula pubescent).

The results are compared with the optimal, dense field measurements based on Leave-out-one (LOO) method where one plot of the total measured plots in the new area at a time is treated as verification set while the rest of the plots as the teaching set.

For each site a repetition procedure of 50 iterations was performed with 50 calibration plots that were randomly selected from the new site using given selection criteria. However the stability of the results was tested for 100 plots and the results were recorded. Taking each test site as a new site at a time while others are treated as database sites. The optimal estimate for each of the verification set were the estimates derived using the leave-out-one procedure in the full set of plots in the new site, and these estimates were used as the benchmarks against which the use of databases was tested. The error of each forest stand parameter estimate was verified using the rest of the measured using the measured plots from the new site as verification plots. The procedure was repeated 50 times for each site to verify the robustness of results from the databases. The tables below show the median of the

$RMSE(\%)$ of estimation results in 50 and 100 repeated calibration plot selections using databases respectively.

	50 plots			100 plots		
RMSE (%)	c	c+all dbs	optimal	c	c+all dbs	optimal
Dgm	16.3	15.0	14.0	14.9	14.6	14.0
Hgm	10.2	9.9	8.8	9.0	9.1	8.8
N	32.9	30.0	27.6	29.9	28.9	27.6
G	19.7	17.6	16.2	17.8	17.2	16.3
V	24.7	21.5	19.8	22.5	21.2	20.0

Table 4.1: Median of RMSE (%) results in Matalansalo in 50 and 100 plots

	50 plots			100 plots		
RMSE (%)	c	c+all dbs	optimal	c	c+all dbs	optimal
Dgm	21.5	19.7	18.6	19.9	19.6	18.5
Hgm	10.4	10.5	9.0	9.9	10.4	9.1
N	47.0	44.6	42.7	44.6	44.1	42.8
G	27.3	26.3	22.0	24.7	25.9	22.0
V	34.9	333.5	28.1	31.3	32.8	28.1

Table 4.2: Median of RMSE (%)of karrtula in 50 and 100 plots

	50 plots			100 plots		
RMSE (%)	c	c+all dbs	optimal	c	c+all dbs	optimal
Dgm	18.3	17.8	16.0	17.3	17.1	16.1
Hgm	12.2	12.4	10.4	11.2	12.0	10.4
N	58.1	56.7	54.9	56.6	56.2	54.7
G	28.8	27.9	26.1	27.3	27.1	26.0
V	30.4	28.9	27.1	28.4	28.3	27.0

Table 4.3: Median of RMSE (%) results in Pello in 50 and 100 plots

	50 plots			100 plots		
RMSE (%)	c	c+all dbs	optimal	c	c+all dbs	optimal
Dgm	19.0	18.0	16.9	17.9	17.6	17.0
Hgm	11.0	10.6	9.7	10.2	10.3	9.7
N	33.4	33.0	30.7	32.0	32.0	30.7
G	19.9	19.1	17.1	18.5	18.4	17.1
V	22.8	21.3	18.7	20.9	20.3	18.6

Table 4.4: Median of RMSE (%) results in Juuka in 50 and 100 plots

	50 plots			100 plots		
RMSE (%)	c	c+all dbs	optimal	c	c+all dbs	optimal
Dgm	16.9	16.1	14.9	16.0	15.4	14.9
Hgm	13.0	12.8	11.5	12.1	12.0	11.5
N	41.5	40.7	38.7	39.8	39.1	38.5
G	25.8	24.5	23.2	24.2	23.7	23.2
V	30.3	29.1	26.8	28.3	27.8	26.9

Table 4.5: Median of RMSE (%) results in Loppi in 50 and 100 plots

4.3 Results.

The two calibration equations are tested for accuracy to see which equation gives better calibration results. The following results were obtained

50 plots	Equation 1			Equation 2		
RMSE (%)	c	c+all dbs	optimal	c	c+all dbs	optimal
Dgm	16.9	16.1	14.9	16.7	25.0	14.9
Hgm	13.0	12.8	11.5	13.0	22.6	11.5
N	41.5	40.7	38.7	41.8	42.6	38.6
G	25.8	24.5	23.2	25.1	28.7	23.1
V	30.3	29.1	26.8	29.9	41.8	26.8

Table 4.6: Comparing the RMSE % results for equations 1 and 2 using 50 plots

From the table (4.6), the RMSE% for equations 1 and 2 show that equation 1 has smaller RMSE% compared to equation 2 except for species 3 whose RMSE% seems to be higher in in equation 1 than 2. This behavior was tested for different values of q, i.e., $0.5 \leq q \leq 1.2$ and the optimal results in both cases was obtained when $q = 0.7$.

Since the first equation gives better calibration results than the second, it is used to calibrate database LiDAR histogram.

The results for different test sites as new areas are given by tables (4.1) for mata-lansalo, table (4.2) for karttula, table (4.3) for pello, table (4.4) for Juuka, and table (4.5) for Loppi. The tables show the repetitions of the Median RMSE% of each test site with calibration set only, calibration set together with the selected plots from all databases, and with the whole set of the measured plots in the new site. It is

observed that the median RMSE % for the calibration plots and all the databases is close to the median RMSE % using a dense set of field sample plots from the new site (optimal case). The RMSE % of the calibration set only decreases as the database plots are added to it. However as the bias% is close to zero almost in all test sites however, to some test sites the bias % slightly varies. This variation is due to the size of the calibration sets. It is also observed that the accuracy of the results increases slightly as the number of the calibration plots are increased from 50 plots to 100 plots. This means that 50 calibration plots from the new site are enough to justify the new site stand characteristics hence be used for the calibration of the databases for the estimation of the stand parameters

CHAPTER FIVE

Conclusion and Recommendations

5.1 Conclusion

The results from the test sites show that it is possible to calibrate histogram databases, and the calibrated LiDAR histograms can be used in new inventory areas for the forest stand parameters estimation. This process is cost effective since instead of scanning the whole new inventory area, only randomly selected set of plots (approximately 50) can be scanned, and the information can be complemented with calibrated database plots for prediction of the forest stand characteristics. Also, the use of calibrated databases can be a useful method in areas which do not support airborne laser scanning due to geographical factors like mountainous areas. In such area only field work can be performed to few heuristically selected set of plots, calibrated plots from databases can be added and hence used in prediction of the forest stand parameters.

5.2 Recommendation

The calibration methodology and estimation procedures used in this study are not the best way to get better results for operational purposes. Different LiDAR histogram database calibration methods as well as distributions used in plots selection should be employed to test the robustness of the results obtained. The aim is to get the best LiDAR histogram calibration method which gives higher accuracy in stand parameters estimation and should be cost effective.

Bibliography

[1] Junttila, V., M. Maltano, and T.Kauranne. 2008. Sparse Bayesian estimation of forest stand characteristics From airborne Laser scanning. *Journal of forest science*. 54(5): 543-552.

[2] Junttila, V., L. Vesa, and T.Kauranne.2010. Estimation of Forest Stand Parameters from Airborne Laser Scanning using calibrated plot databases *Journal of Forest Science*. 56(3): 543-552..

[3] Mardia, K.V., J.T. Kent, and J. M. Bibby. 1980. **Multivariate analysis**. Academic Press Inc, London. Chap.2.5

[4] Maltamo, M., and Kangas, A. (1998). Methods based on k-nearest neighbour regression in the estimation of basal area diameter distribution. *Canadian Journal of Forest Research. 28. 1107-1115.*

[5] Næsset, E. (1997). Estimating timber volume of forest stands using airborne laser scanner data. *Remote Sensing of Environment.* 51, 246-253.

[6] Næsset, E. (2002). Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data. *Remote Sensing of Environment.* 80-89.

[7] Packaln, P., and Maltamo, M. (2006). Predicting the plot volume by tree species using airborne laser scanning and aerial photographs. *Journal of Forest Science. 52(6), 611-622.*

[8] Maltamo, M., Peuhkurinen, J., Malinen, J., Vauhkonen, J., Packaln, P. and Tokola, T. 2009. Predicting tree attributes and quality characteristics of Scots pine using airborne laser scanning data. *Silva Fennica.* 43(3): 507521.

[9] Tipping, M.E., 2001. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research.* 1,211-244.

[10] 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: 543553..*

[11] Suvanto, A. and Maltamo, M.(2010). Using mixed estimation for combining airborne laser scanning data in two different forest areas. *Silva Fennica.* 44(1):91-107.