GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Modelling the Diameter Distribution of Forests on Compartment Level based on an Airborne Laser Scanning-System

Authors

Clemens Haerder & Petros Christanas

Supervisor

Dr. Paul Magdon

Statistical Practical Training

September 11, 2019

Statutory Declaration

We declare that we have authored this thesis independently, that we have not used other than the declared sources / resources, and that we have explicitly marked all material which has been quoted either literally or by content from the used sources.

Clemens Haerder

Petros Christanas

_____

Signature

_____

Signature

Acknowledgments

We want to thank Gräflich Bernstorff´sche Betriebe for providing the LiDAR and Forest Inventory data which enabled this study in the course of a statistical practical training.

We owe further thanks to our supervisor Dr. Paul Magdon providing us insights into the forest structure, data bases and forest inventory.

Contents

1 Introduction

Forest inventories must be conducted in regular intervals to sustainably manage a forest. Forest inventories are traditionally conducted on different scales. On the larger enterprise level sample-based approaches are implemented. However, for decisions at the level of the management units (compartments) taxations are done by experienced foresters. Information on growth, mortality, age and stand structure and volume stock are collected in field by visits of each compartment. Tenants are especially interested in the growing stock of the forest as it presents the main source of incomes. The growing stock can be modeled by measuring the diameter of trees at breast height (1.3m). Such measurements are commonly done on small sample plots (locations where multiple measurements are conducted).

Alternative approaches to forest inventory are in development to be a more objective and cost efficient. One of those alternatives are airborne LiDAR (Light detection and ranging) systems. An airplane flies in a grid over the forest. The LiDAR system collects the distance to a target by illuminating the target with pulsed laser light and measuring the reflected pulses with a sensor. Differences in laser return times and intensities can then be used to obtain 3-D structural information of the target forest.

The 3-D model of the forest thus implies the height of trees from which a detailed canopy height model (CHM) can be generated, which represents tree / canopy height for each pixel. High trees are indicated with bright colors and objects with a height smaller then 5m will appear black. Using feature detection algorithms tree tops can be detected in the CHM and tree crowns delineated. Note that trees below the 5m threshold are of little relevance for growing stock estimates, as the diameter is simply too small.

Hence an estimate of the number of trees is obtained, including their associated height and crown area can be automatically extracted from the LiDAR point clouds analysis. However, two limitations hinder the direct integration into the forest inventory estimation setup:

1. To estimate growing stock and other relevant forest variables a diameter estimate is required which cannot be directly observed from the LiDAR data. Therefore, statistical models need to be developed.

2. Furthermore, the tree detection algorithms are unable to detect smaller trees which are partially covered by larger trees. Thus, a systematic bias of the diameter distribution is towards trees with larger diameters is expected and needs to be corrected. This correction must be performed on compartment level. A compartments level is a relatively homogeneous area within the forest, further described in 2.1 Forest Classification.

The objective of this study is to contribute to the development of LiDAR assisted forest inventories by developing a statistical sound approach to derive unbiased diameter distribution models from LiDAR data for homogeneous compartments of the forest.

2 Study Area & Data Sources

2.1 Forest Classification

The study area is a private forest enterprise in Gartow (Niedersachsen) Germany. It measures around 5674.2 ha in total (see Table **??**) and is a relatively homogenous forest consisting mostly of pine trees.

Table 2.1: Size of the different stratum and associated sampling grids. Stratum 2 and G have been merged to Location Class 2 which results in an identical sampling grid.

Stratum	Location Class	Area [ha]	Relative Area
1	1	338.6	0.06
2	2	1546.9	0.27
3	3	2129.3	0.38
4	4	1550.5	0.27
G	2	108.9	0.02
Total	-	5674.2	1

The forest itself is split into stratum to take site conditions, forest structure and thus natural variation of the areas into account (see Figure **??**). The assessment of variation was based on a forest inventory conducted 2008 (see Table **??**).

Table 2.2: Mean volume and sample variation estimates of the forest inventory 2008. Stratum 2 and 3 show little relative standard error (SE%), while stratum 1 inhibits more variation. Stratum G, which covers only 2% of the total area has a typical high variation

Stratum	Location Class	Area [ha]	Relative Area	Sample Size	Mean Volume / ha	Sample Variance	SE%
1	1	338.6	0.06	159	180.19	104.24	5.67
2	2	1546.9	0.27	805	246.75	22.75	1.93
3	3	2129.3	0.38	542	195.41	13.15	1.86
4	4	1550.5	0.27	134	131.90	30.45	4.18
G	2	108.9	0.02	55	271.26	734.08	9.99
Total	-	5674.2	1	1659	196.37	6.46	1.29

Main sources of the variation in the growing stock can be assigned by the varieties in the tree species and the age distribution of the trees. Young and therefore small trees have a smaller diameter. If an area has been cultivated around the same timespan with identical species, the

trees are expected to be centred on a certain diameter. On the other hand, a very diverse area in species and time will have naturally more variation (see Figure **??**)

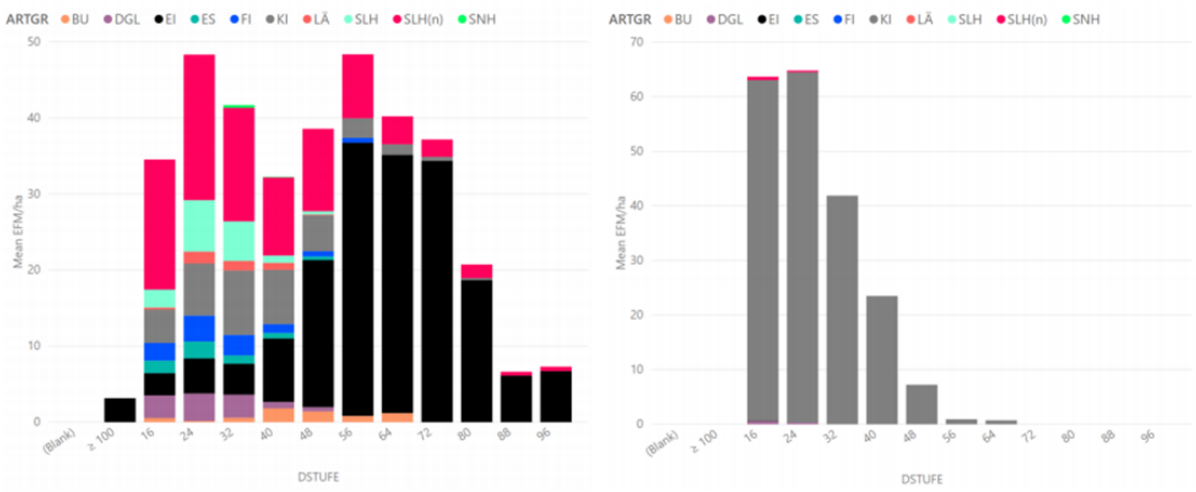


Figure 2.1: The bar plots (left to right: stratum 1, stratum 4). The bars indicate the mean volume per ha for different diameter classes [1].

Sampling activities are adjusting according to the inhibited variation of the stratum type (see Section **??**).

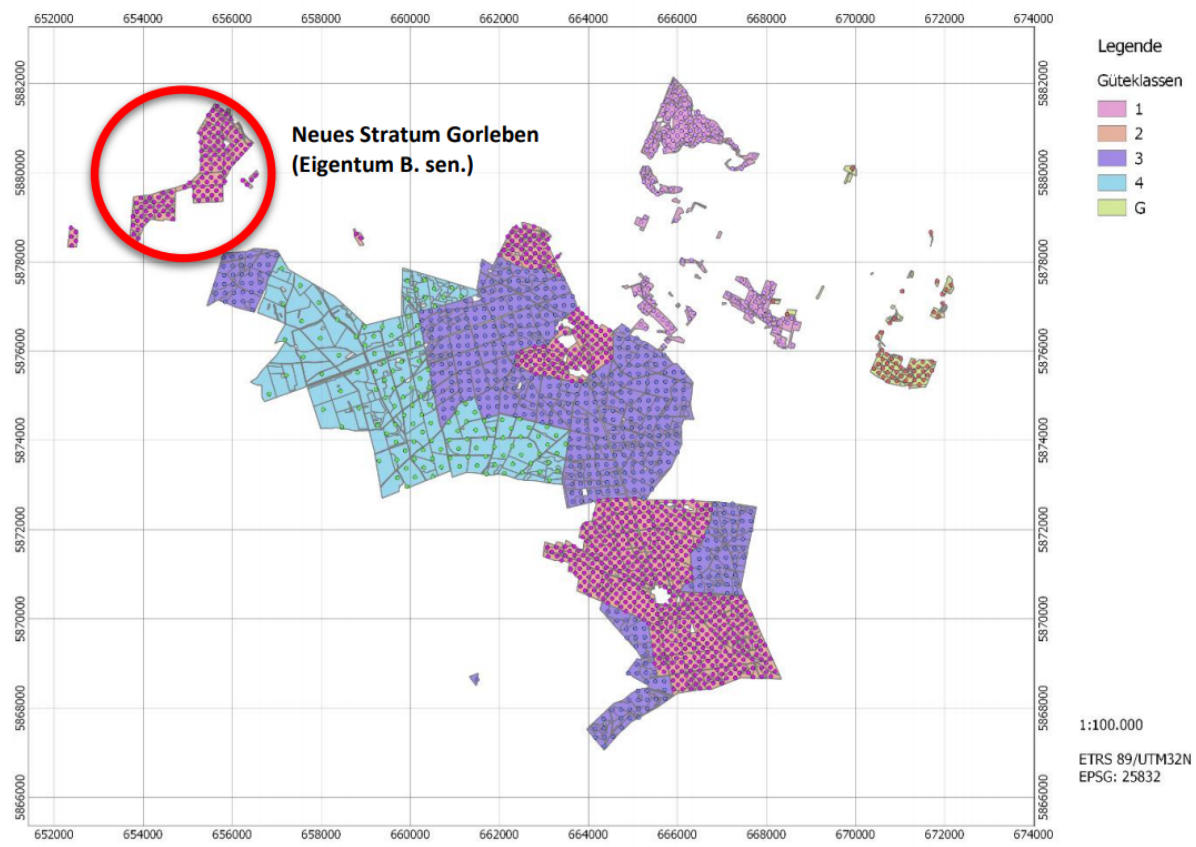


Figure 2.2: The forest of Gartow is divided into stratums based on past observed variation and ownership and subdivided into compartments indicated by grey lines. Each point within a section indicate a sampling point [1].

The stratum themselves are subdivided into compartments with the intention to create homogeneous sub regions (see Figure **??**). The diameter distribution must be found for each of those compartments.

2.2 Inventory Data

In spring 2018, a sample-based forest inventory was carried out in Gartow. 942 sampling locations are defined which are spread over the forest based on a stratified sampling approach. This accounts for the past observed variation within the regions. Compartments of stratum 1 and 2 are sampled with a dense sampling grid, while 3 and 4 have a wider sampling grid (see Table **??** & Figure **??**).

At each sampling location (so called plots) several attributes of the trees within a certain circular area are measured. The parameters of primary interest in this study are the diameter, species and height. The diameter is measured at breast height (around 1.3 meters) with a measuring tape. Subsequently, the height is measured with varying, but established methodologies. Unlike the diameter, not every tree height is collected. In each plot, three main species trees (less if there are fewer trees) are measured. To cover the total range of values, a small, a medium sized and a large one is gauged. Additionally, one tree of every other species is measured to cover the variety of species. Table **??** provides an overview of total measured trees.

Table 2.3: Overview of number of measured trees for height and diameter per Stratum

Stratum	# Measured Trees
1	1287
2	3434
3	2734
4.1	616
4.2	792
G	523
GL	619
Total	10005

2.3 LiDAR

While the previously described data will be used for modelling, data captured by the airborne LiDAR is of main interested in this report, since the ability of innovating forest inventory is discussed.

LiDAR uses a laser scanning system to capture distances. In context, laser scanning is referred to the active emitting and sensing of light. Thus, Light Detection and Ranging is a suitable description of the mechanics, also known as LADAR (Laser Detection and Ranging). LiDAR is a more generalized definition, as instead of laser- light also xenon or flash lamps can be used [2]. A high-level definition of the functionality of LiDAR is as follows. Laser beams are continuously emitted of the LiDAR system, mounted on an airborne vehicle. The coordinates are throughout captured by a GPS (Global Positioning System) and IMU (Inertial Measurement Unit – used to capture adjust for e.g. inclined positioning, acceleration of the vehicle). Laser or xeon/flash light is emitted of an active sensor and the distance captured once it is traveled back to the scanner. As forests have a relatively turbulent surface and only little light can reach the surface of the forest, many systems only capture the first and last impulse [3]. The cloud of captured points can then be used to create a 3-D image of the forest (see Figure **??**).

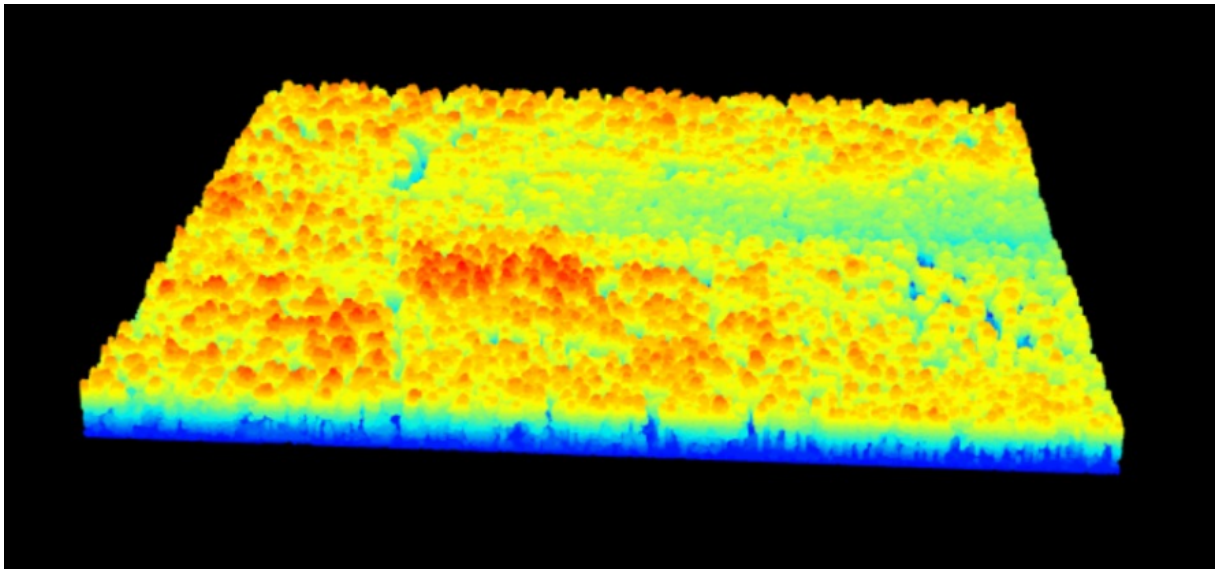


Figure 2.3: 3-D Image of a small area of the forest of Gartow made by the airborne LiDAR. The determined height is colorized. A dense group of height trees is found almost in the middle and directly behind an aisle of small trees.

Main benefits of LiDAR compared to other systems is the ability to capture data regardless of sun positioning, day or night and the ability to map through the highly dense areas (the canopies of the trees). The main benefit compared to the traditional way of forest inventory is

relative intuitive. A plane is capable of objectively measuring the forest subject to this report in under a week; while a forester must inspect every single hectare, providing a more subjective intuition of the forest inventory.

Table 2.4: Flight log of the airborne laser scanning of the forest of Gartow [12]

Flight Altitude	**Approx. 590m above ground**
Nominal point density (laser)	6 points / m^2
Ground resolution	4.3cm
Point density (to circumvent overlap)	12 points / m^2
Ground resolution	4.3cm

ForestEye Research GmbH & Co. KG provided the detected single tree location, tree species and canopy area based on LiDAR.

3 Methodology

3.1 Challenges

This study faces several major challenges which need to be resolved in the analysis:

1. As mentioned, the tree detection algorithm based on LiDAR imposes a systematic bias. Subdominant trees which are covered by dominant tree crowns cannot be detected (see Figure **??**). Unfortunately, the height of the covered trees cannot be determined. This leads to a systematic overestimation of the tree diameter distribution of each compartment. The covered tree height cannot be determined but must be within the interval 5m < x < height of covering tree. The main task of this study is to correct for this bias.

2. Although the diameter of over 10000 trees are collected, only 23 compartments have 40 or more measured trees. This problem is aggravated as ~700 of the 1642 compartments don't have any measurements. Meaning that a distribution fitting via Maximum Likelihood Estimation for compartments with too little or no data must be performed.

3. The measured data for the regression is unreliable. Tree species, crown area and height are detected by LiDAR for the inventory dataset. Whereby the tree species is detected by neural networks with an approximate 85% accuracy. The tree diameter differs depending on the species significantly (as seen in Section **??**), which results in additional bias in the prediction.

4. The fourth challenge was discovered in Section **??**. Dominant (large) trees cannot only cover subdominant (smaller) trees, but equally tall trees might be detected as only one tree (see Figure **??**). As a result, the crown area will be significantly overestimated in the inventory dataset which is used in the regression resulting in addtional bias in the prediction.

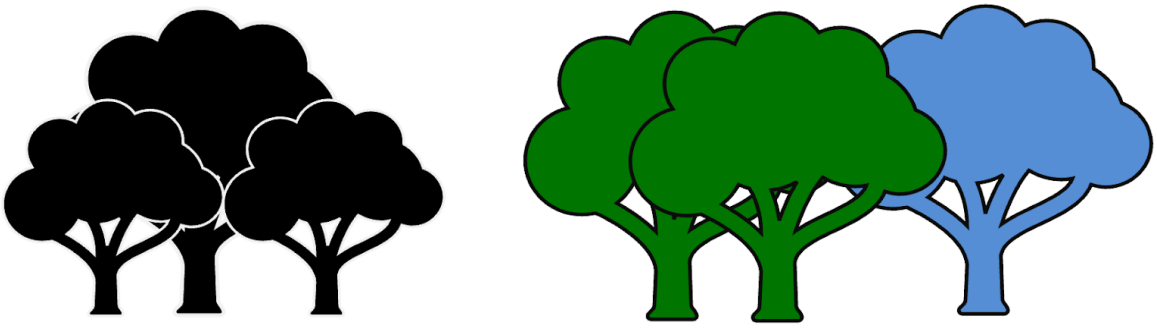


Figure 3.1: Visualization of detection problems.

Left: a dominant tree covers subdominant trees (challange 1).

Right: Close equally tall trees (green) are detected as one tree, causing overestimation of the crown area (challange 4).

3.2 Methodology

In a first step, summary statistics of the inventory data are presented to get an insight of the forest and correlations (see Section **??**). In a second step, a regression model is constructed to estimate the diameter of trees based on the height and crown area from the inventory data. Additionally, the effect of tree species is examined. Subsequently, the model is used to predict the diameter of the trees in the LiDAR dataset.

1. **The Log-normal Regression Model**

   The log-transformation (and back-transformation) is an established method introduced in 1941 by Finney (see [4]). The transformation of the response yields in a log-normal model

   $$ln(y) = X\beta$$

   where X is the design matrix. The random variable

   $$z = ln(y)$$

   is normally distributed with μ_z and σ_z^2. The back-transformed random variable

   $$y = e^z$$

   is log-normal distributed. It can be shown that

   $$\mu_y = e^{x_i'\beta+\sigma_z^2/2}.$$

   σ_z^2 is estimated by the MSE s_z^2 and $e^{\sigma_z^2}$ can be considered as a correction term, which due to its positivity always increases the back-transformation.

2. **Generalized Linear Models**

   Instead of transforming the response itself, Generalized Linear Models (GLM) are used to instead transform the mean. Thus, μ_i is connected to the linear predictor

$$\eta_i = x_i'\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

   through

$$\mu_i = h(\eta_i) = h(x_i'\beta) \quad \text{or} \quad \eta_i = g(\mu_i),$$

   where h is the response function and g the link function [5]. The Gaussian and gamma from the exponential family are chosen. The log function as link is chosen for the same reason as in the log-normal model. The transformation of the mean can result in substantially different result for the Gaussian model with log link compared to the log-normal model. The dispersion parameter (see Table **??**) of the Gaussian model is just the σ^2 and v^{-1} (inverse scale) for the gamma model [5].

Bias correction is performed on compartment level, where inventory data is sparse. K-means clustering will group compartments with similar tree structure allowing for a richer inventory set used for bias correction.

K-means clustering

K-means clustering is a relatively simple iterative clustering procedure. In short, k centroids are set randomly into the data space.

1 For each data point (compartment), the distance to the k centroids is calculated.

2 Each data point is assigned to the centroid with its minimal distance forming k-groups.

3 The mean of all data points within the group is calculated, which is essentially the update of the centroid.

This is repeated until there is no more change of the attribution of the compartments to the clusters (convergence is reached). As the initiation of the k-centroids is random, results can vary. The number of k-clusters must be defined prior to the start. Different numbers of cluster k should be tested, and decision can be made by visualization of the clusters and/or change in the total within sum squares used to pick the best (heuristic) k. Ultimately, the clusters should make sense in a way that experts can describe them. For more details see ref. [7].

For each of the resulting clusters of compartnents, parametric distributions are fitted on the values of the measured and predicted tree diameter. This leads to a posession of parameters of the (chosen kind of) distribution, such that tuning of such parameters allows for bias reduction.

Parametric Distribution Fitting

LiDaR as well as inventory diameters on a cluster level are subject to a fit of a preselected distribution. The chosen distribution needs to be reasonable in the sense that parameter or variable restrictions are well defined and the preselected distribution should properly reflect the observed distribution [13].

There are several methods for fitting a parametric distribution to the data.
In this case, Maximum Likelihood Estimation (MLE) is applied to estimate the parameters of the distribution, i.e. the Log-Likelihood function of the respective distribution is the objective of a maximization problem with respect to the distributional parameters [15].

Gamma Distribution

Let X be a random variable which follows a gamma distribution: $X \sim Ga(\alpha, \beta)$, where α is called the shape parameter and β is called the rate parameter.
The corresponding probability density function is

$$f(x;\alpha,\beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for} \quad x > 0 \quad \text{and} \quad \alpha, \beta > 0,$$

where $\Gamma(\alpha)$ is the gamma function [14].
"The gamma distribution is not only a good model for waiting times, but one for many non-negative random variables of the continuous type. For illustrations, the distribution of certain incomes could be modelled satisfactorily by the gamma distribution, since the two parameters α and β provide a great deal of flexibility." [14]

4 Analysis

Regression models are fitted on the whole inventory dataset to predict the tree diameter distribution of the LiDAR dataset. Later, this density distribution is bias corrected on compartment level.

The initial problem regarding this estimation lied within the sparsity of variables. A regression model can only use variables which exist in both datasets (LiDAR & inventory data).

During the first half of this study, only the tree height and geographical position of each tree existed within the LiDAR dataset. Further important variables were later acquired and provided by ForestEye Research GmbH & Co. KG. An overview can be found in Table **??**.

Therefore, two different approaches for the regression model were developed. The regression model tackling sparsity can be found in the Appendix, providing an alternative approach for studies which have only the height and geographical position of trees.

The final model is discussed in the following Sections.

4.1 General Overview of the Variables

Table 4.1: Variables used in the study and their sources

Variable	Data Type	Dataset
Diameter [cm]	Continuous	Inventory
Crown Area [cm^2]	Continuous	LiDAR
Height [cm]	Continuous	LiDAR
Species (Art_ grp)	Factor	LiDAR (Neural Networks)

First the general properties of the inventory data of 2018 are examined to obtain a first intuition of dependencies and structures for further analysis. As discussed in Section **??**, the diameter is predicted based on some variables. Previous research proposed to create models individually for each tree species, using height measurements, the crown area & the crown area of larger trees [6].

Figure **??** indicates that this proposition is also valid for the measured trees in Gartow, as the boxplots of the diameter and height differ per species. Similar patterns are seen for height and diameter, indicating high correlation.

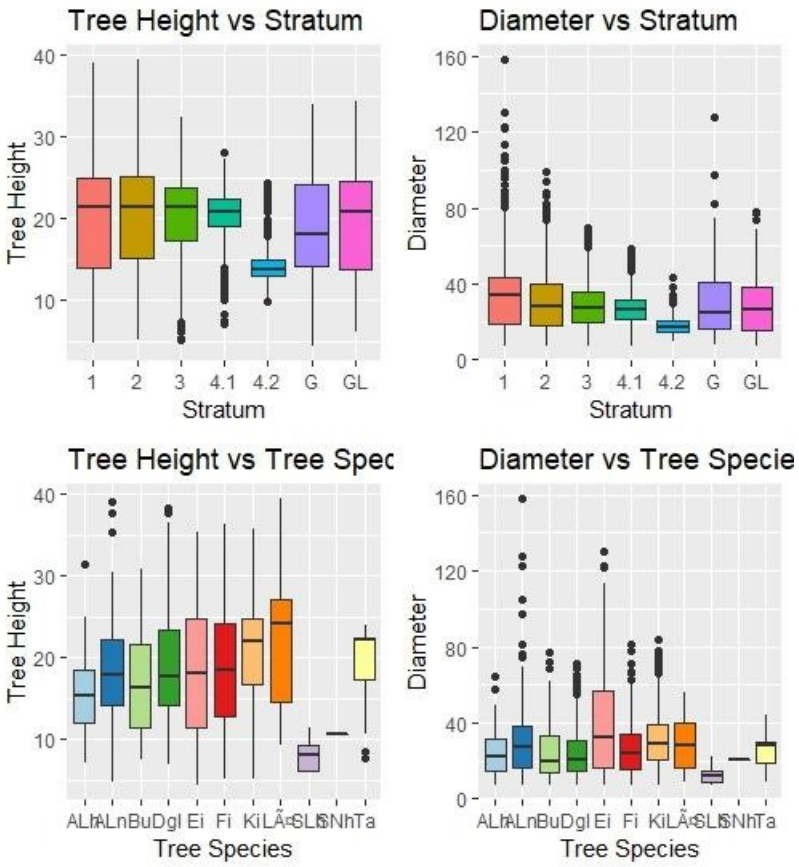


Figure 4.1: Boxplots based on Stratum (upper) and Tree Species (lower). Same patterns show potential correlation in height and diameter.

Figure **??** underlines this expected height correlation. A strong non-linear relationship is observed. This almost exponential curve can be explained by the natural growth of a tree. Once a tree species reaches its maximum height, only the diameter continues to grow up to a natural limit. The effect of tree specific is well observable comparing oaks (Ei) and pines (Ki). They are well separated between 20 and 30 meters of height.

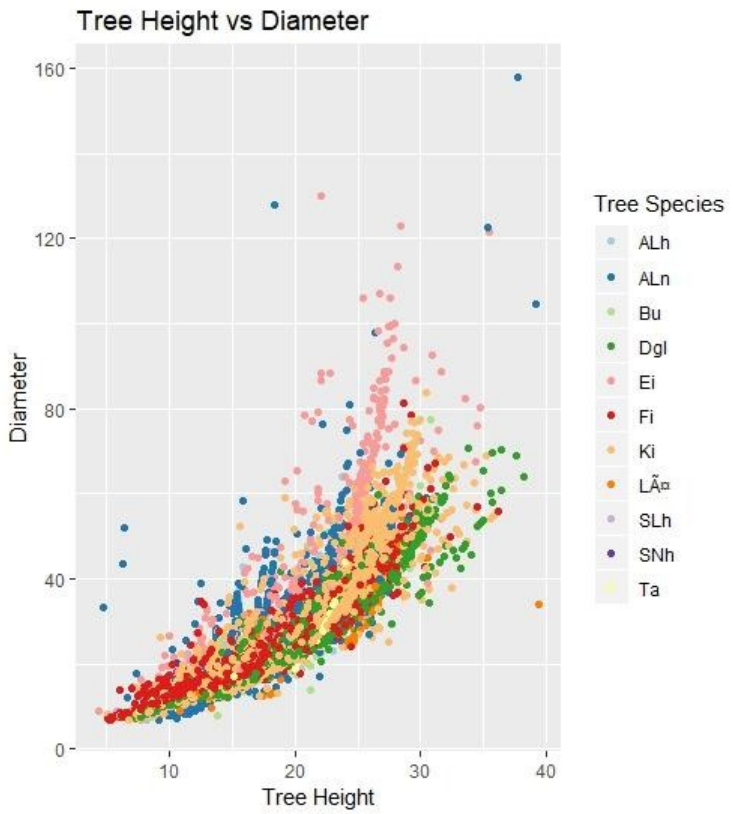


Figure 4.2: Relation between diameter and height. Tree species are visualized by colouring

A log transformation is imposed on the response diameter (see Figure **??**). The log transformation reduces the skewness of the data, while keeping a potential linear dependency. The diameter is always positive and a log transformation can be applied. The relationship between height and log(diameter) has improved regarding linearity.

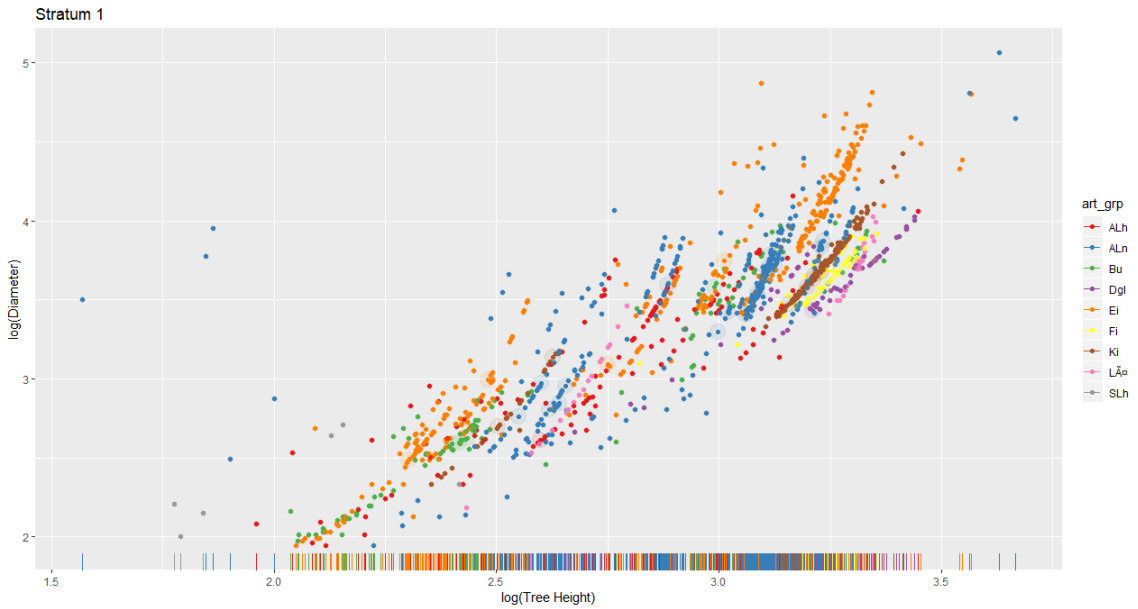


Figure 4.3: Scatterplot of log transformed diameter and height for Stratum 1.

4.2 Regression Models

To predict the diameter of a tree using regression models is straight forward. We chose the later described regression models based on following criteria.

1. *Interpretability* is required so experts can understand the results to continue to work with the proposed models. Inference on all covariates must be feasible, as large standard errors (i.e. high uncertainty in the predicted diameter) causes the model to be unusable.

2. *Generalization* (i.e. the goodness of fit to new data) must be high to assure sufficient prediction of the diameter of trees detected by LiDAR. While it is likely that regression splines might result in an overall better fit to the data, relying too much on the sampling data might cause substantial over-fitting. As described in Section **??**, over-fitting must be prevented due to the unreliability in the data. Regularization methods such as ridge regression could be applied. Yet, based on the results in the applied models it was not further investigated.

3. *Complexity* should be kept moderate to again make the models more accessible. The model is only used in this study to predict the diameter. Using Generalized Additive Models for Location Scale and Shape could be an interesting addition to this study to further investigate the uncertainty at different in the estimates but is potentially too complex for the task of a simple prediction.

Based on those criteria, a log-linear model (based on the observations in Section **??**) and generalized linear models are compared.

4.3 The Regression Outputs

All regression models are fitted using all available covariates:

$$\text{Diameter} \sim \text{Tree Height} + \text{Crown Area} + \text{factor(Species)}$$

Table 4.2: Diameter prediction models: log-linear, Gaussian and gamma

Coefficients	Log-linear Model		GLM Gaussian (Link: Log)		GLM gamma (Link: Log)	
-	Estimate & Std. Error		Estimate & Std. Error		Estimate & Std.Error	
Intercept	1.685	0.010	1.498	0.014	1.709	0.010
Height	0.081	0.000	0.091	0.001	0.081	0.000
Crown Area	0.001	0.000	0.001	0.000	0.001	0.000
Art_ grpDgl	-0.296	0.008	-0.415	0.008	-0.305	0.008
Art_ grpEi	0.183	0.010	0.222	0.008	0.174	0.010
Art_ grpFi	-0.185	0.009	-0.259	0.009	-0.195	0.009
Art_ grpKi	-0.115	0.007	-0.139	0.007	-0.125	0.007
Art_ grpLä	-0.250	0.015	-0.323	0.015	-0.254	0.016
Overdispersion Φ	-	-	12.5	-	0.01	-

The estimates in all models (see Table **??**) are significantly different from 0. The untransformed confidence intervals are visualized in Figure **??**. The transformed confidence intervals for the gamma model are found in Table **??**.

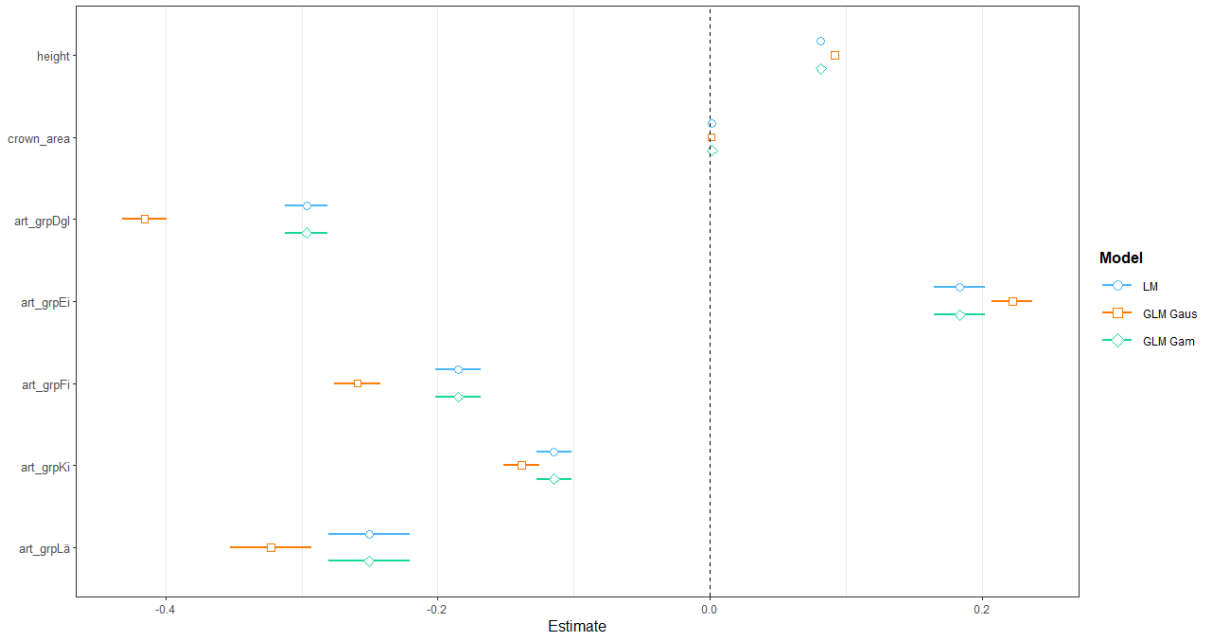


Figure 4.4: While the estimates and confidence intervals of the gamma and log-linear model tend to be similar. The Gaussian model estimates are larger for values greater 0 and smaller for values smaller then 0.

Table 4.3: Transformed confidence intervals (CI) and width of the gamma model in cm

Coefficient	2.5% CI Lower bound	97.5% CI Upper bound	CI Width
Intercept	5.4170	5.6320	0.2150
Height	1.0831	1.0851	0.0020
Crown Area	1.0013	1.0016	0.0003
Art_ grpDgl	0.7250	0.7488	0.0238
Art_ grpEi	1.1666	1.2135	0.0469
Art_ grpFi	0.8085	0.8375	0.0290
Art_ grpKi	0.8704	0.8943	0.0239
Art_ grpLä	0.7515	0.8002	0.0487

Model selection is done by three criterions: R^2 , AIC & BIC and Residual Sum of Squares (RSS)

$$\sum (y - y_{\hat{predict}})^2.$$

All values can be found in Table **??**.

Table 4.4: Model selection criterion for Log-linear, Gaussian and gamma

-	LM	GLM Gaussian		GLM gamma	
R^2 / Pseudo R^2 (Cragg-Uhler)		Null Deviance	654900	Null Deviance	533.9
		Residual Deviance	46260	Residual Deviance	35.72
	0.937	0.929		0.933	
AIC	-	19903.18		19102.88	
BIC	-	19959.15		19158.84	
RSS	60136.49	46257.25		53569.47	
RSE ($\sqrt{RSS/n}$)	4.0315	3.53581		3.80503	

The models have a (pseudo) R^2 of almost 93% while the Log-linear model is favorable with 93.7% explained variation. The AIC & BIC of the gamma model is better compared to the Gaussian model. This cannot be compared to the Linear Model, of which the AIC is constructed based on the RSS and not a likelihood. Finally, the Gaussian model has the lowest Sum Squares and therefore the smallest Residual Standard Error with approximately 3.53 cm in the diameter.

All models have an exceptional fit, whereby we finally decide to use the gamma model. A bias correction is introduced by fitting parametric distributions in Section **??**. Thus, the model with the highest likelihood (AIC & BIC) is finally favoured.

Figure **??** visualizes the good fit of the model. Most points by species are well described by the panes, whereby we see for both tree species an underestimation at large value. This is expected, as trees have natural high limits. Once it is reached, only the diameter increases, resulting in the underestimation.

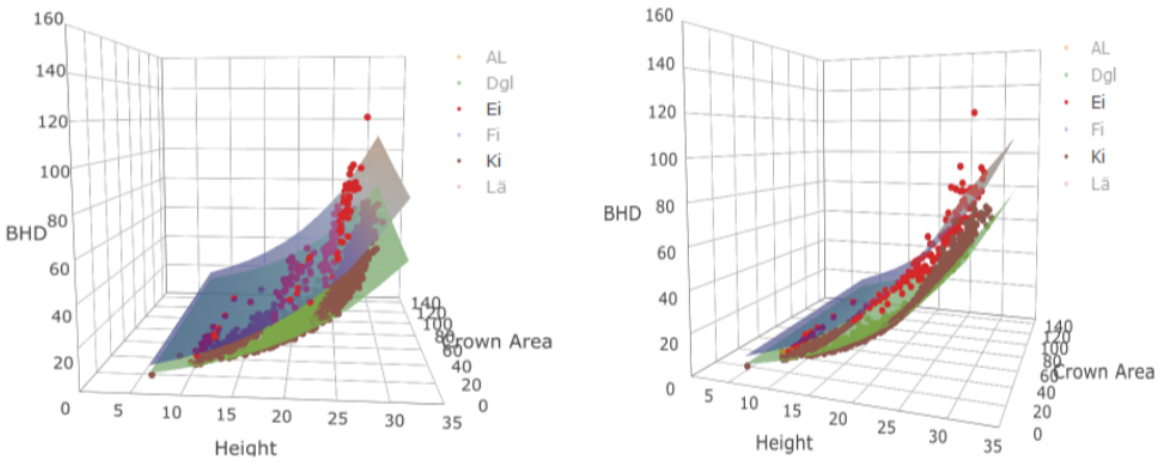


Figure 4.5: Predicted pane of the gamma model for oaks (Ei red) and pines (Ki brown)

Figure **??** displays the residual plots of the gamma model. Some unwanted behavior is seen in the scale location plot, which is used to identify violations of homoscedasticity. A somewhat linear behavior between 2.5 to 3 followed by a curvature to 4 results in potential mild heteroskedasticity for larger values as indicated by the red line. 5 unique observations in over 3700 observations are identified as outliers. This leads to the conclusion that the GLM gamma fits well and is acceptable regarding the underlying assumptions.

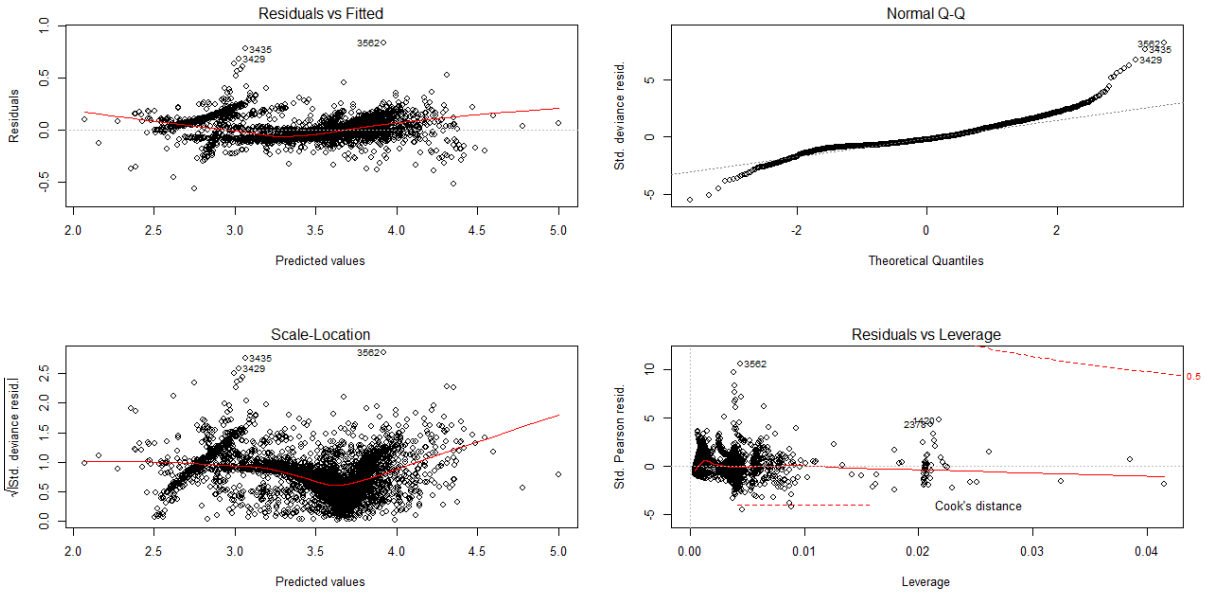


Figure 4.6: Residual plots of GLM gamma. Data points with an index are considered outliers.

4.4 Clustering

A crucial part of this study is the cluster analysis of the forest compartments. Therefore, we perform simple k-means clustering on the compartments to resolve a key issue: The single compartments themselves do not have enough observations to perform any kind of distribution fitting via Maximum Likelihood Estimation. Clustering relatively similar compartments could resolve this issue by creating k clusters, whereby each cluster then has enough observations to perform distribution fitting.

The following auxiliary variables are assigned to each of the compartments to calculate distances. The thought process of those variables to describe the systematic bias is further outlined in Table **??**.

Table 4.5: Auxiliary variables for k-mean clustering

Variable	Potential to explain Bias
Forest Cover Ratio Area / # of Detected Trees	A densely forested region will have more dominant subdominant structures.
Tree crown coverage Area surface - Total Crown Area	
Variation of crown area	Large variation in the Tree Crowns and overall large crowns will have more dominant subdominant structure
0.75 quantile of crown area	
Variation of height	If the cultivation time is similar, the trees should be of same height, resulting in less covering

All variable combinations were tested by a time-consuming trial-and-error procedure. Meaning that for each generated cluster output, a Principle Component Analysis is performed (if > 2 variables), visualizing the goodness of separation from each group (Figure **??**). Subsequently 3-D plots from the LiDAR dataset for several compartments of each cluster are created. Clustering is considered successful, once the groups can be well explained by the 3-D images and the separation of the clusters is reasonable.
This is presented with the final model used in this study.

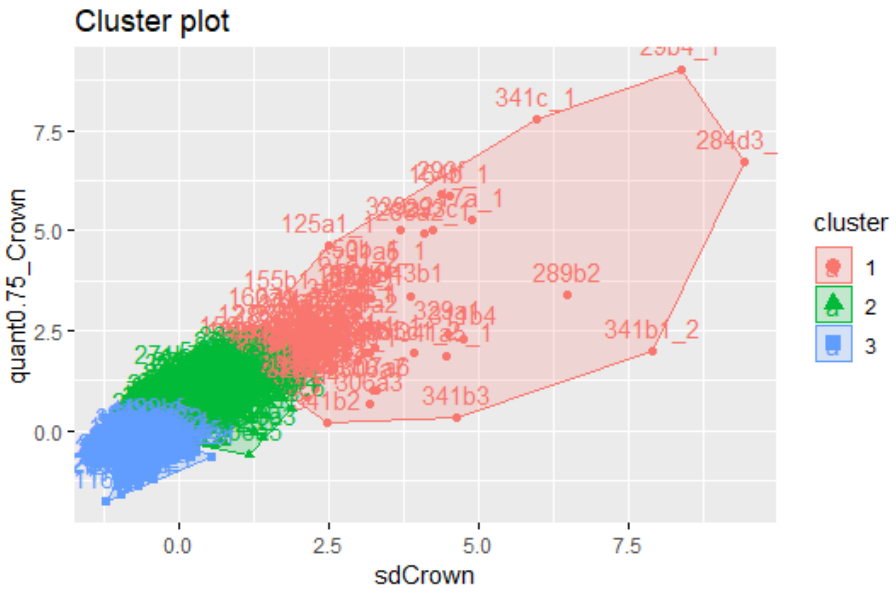


Figure 4.7: Scatterplot of the sections with the grouping based on k-means clustering

The chosen final auxiliary variables are the 0.75 quantile of the crown area and the variation of the crown area. Figure **??** shows high correlation. This is additionally depicted in Figure **??**.

The desired property of a clear separation of the clusters is not fully given. We render this as a minor problem. The three cluster are large enough so that enough samples from inventory data are in each cluster. Cluster 1 336, cluster 2 4180, cluster 3 5471.

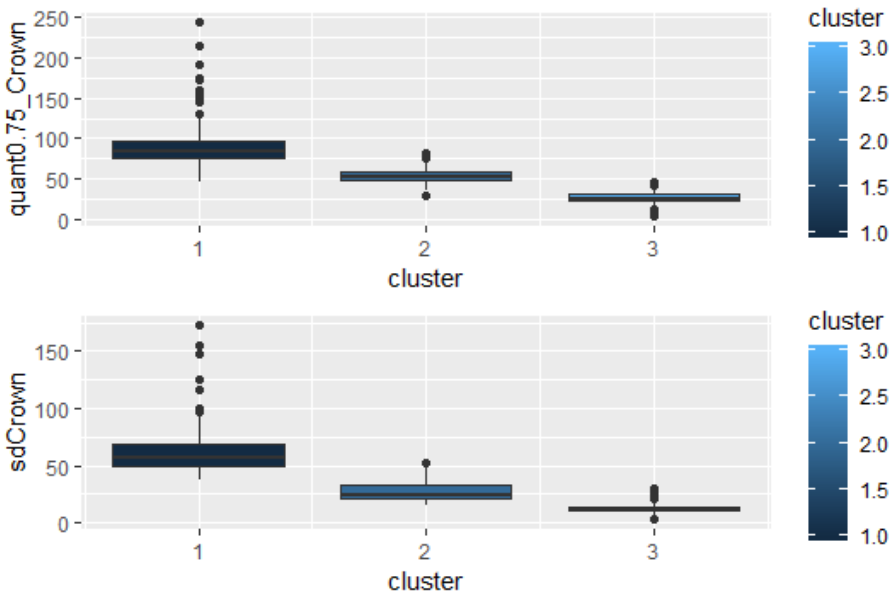


Figure 4.8: Boxplots for the used variables - clusters

The 3-D plots of some compartments for each cluster offer insights into the goodness of the clustering. Sections in cluster 1 appear to be very dense areas as the forests soil is barely visible. Interestingly, the tree height appears to be rather equal, resulting in homogeneous patches. This is well depicted by compartment 261a2 (Figure **??**). The compartment in cluster 2 are sparser wooded (Figure **??**). Blue patches (forest soil) are visible in every compartment. This is even stronger pronounced in the compartments of cluster 3 (Figure **??**).

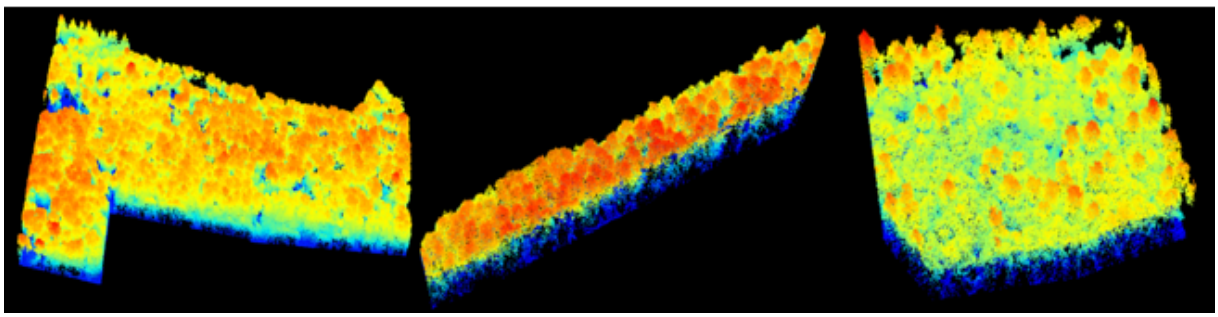


Figure 4.9: Cluster 1 section ID's from left to right: 155a1, 311b1, 261a2

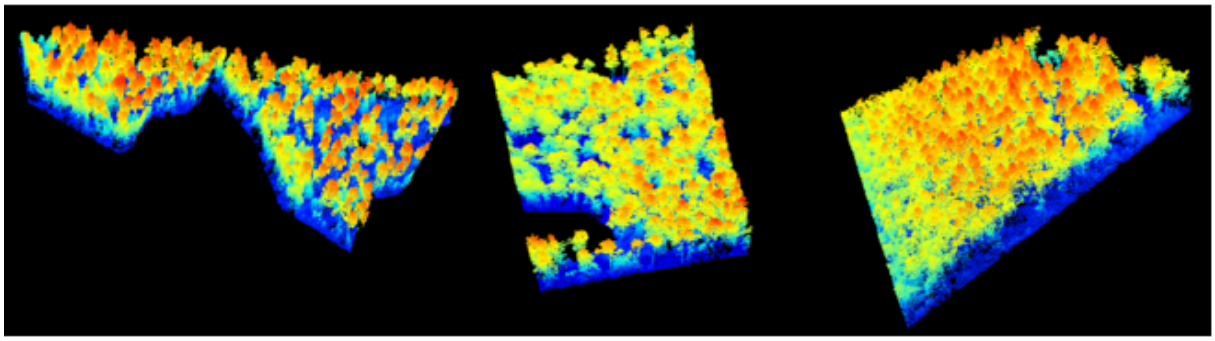


Figure 4.10: Cluster 2 section ID's from left to right: 37a2, 71b5, 309a2

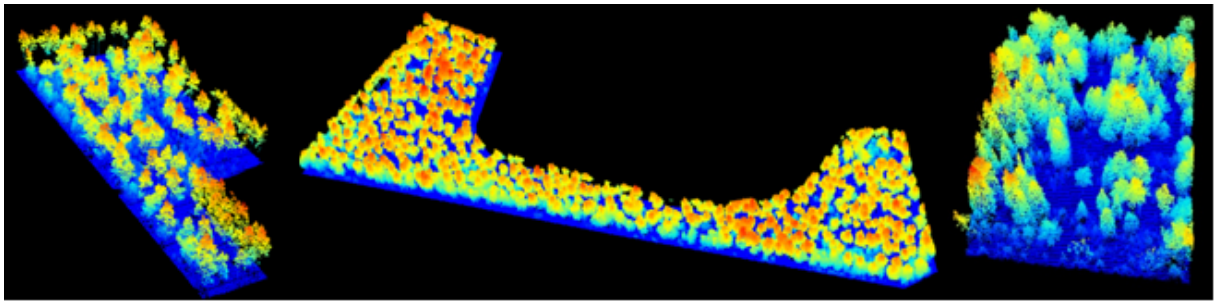


Figure 4.11: Cluster 3 section IDs from left to right: 314a2, 207a2, 36c1

Using the variable tree crown to describe the clusters is difficult. No actual difference in the variation and quantile of the crown area between the clusters can be observed visually. Anyhow, dense and sparse forested region are well separated.

It might be explained as following. Dense compartments show regions of equally height trees. The tree detection algorithm is unable to detect all trees, as tree crowns of several trees are detected as one (see Section **??**). This leads to unnatural high variation in the crown area. The quantile of the crown area further captures the extent of this effect.

4.5 Distribution Engineering

After predicting the tree diameter of the LiDAR data, a systematic bias occurs due to the downsides of the LiDAR system difficulties mentioned in Section **??**. To overcome such kind of obstacles of the main goal, i.e. approximating an ideally unbiased diameter distribution of the forest on a compartment level, we introduced a technique which they refer to as "distribution engineering". It takes advantage of some features and diagnostics of parametric distributions. The usefulness of the clusters is also presented in this Section.

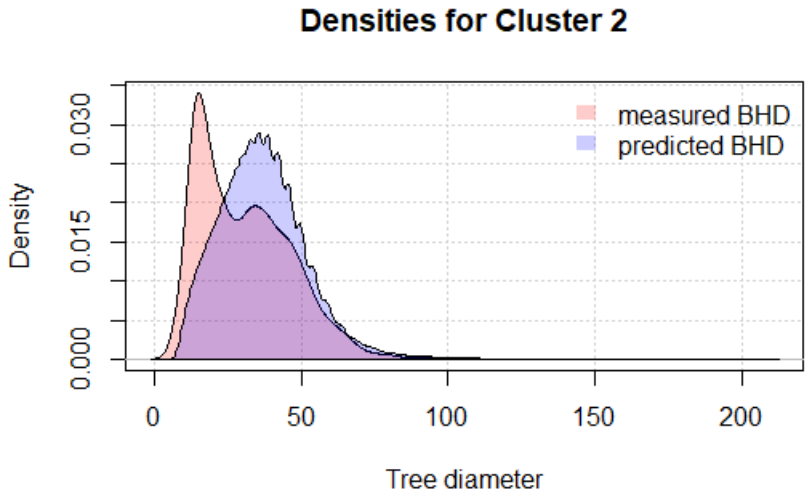
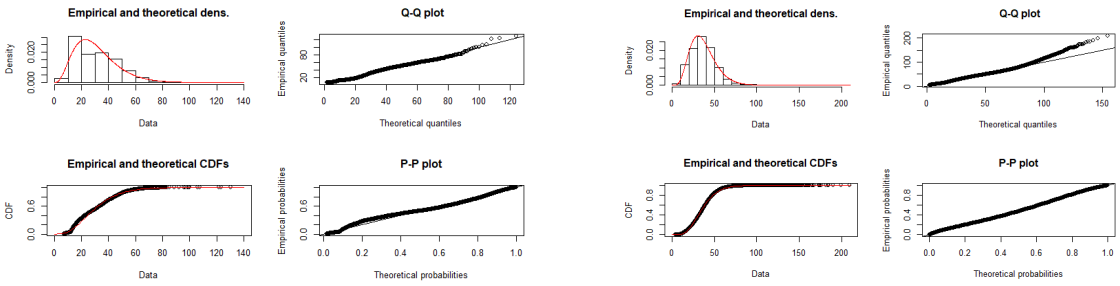


Figure 4.12: Measured vs predicted densities of the tree diameter for compartments of the 2^{nd} cluster

In a first step, the diameter values of the LiDAR data as well as the inventory data are subject to a parametric distribution fit. This initial fit is employed on a cluster level, which provides us with the respective parameters of those distributions. In particular but not exclusively, a gamma distribution is fitted in this case (using R-Package: fitdistrplus [16]). The reasoning behind that is to retain coherence, since a GLM gamma regression model was applied to predict the tree diameter. An example for cluster 2 with further details of the fits can be observed in Figure **??**.



(a) Inventory (b) LiDAR

Figure 4.13: Graphical fit diagnostics for cluster 2 - Tree diameter

Looking at the QQ-Plot of Figure **??**, one can clearly see the outliers of the predicted tree diameter on the upper quanitles (LiDAR), which is (amongst other things) attributed to the over-estimation of the crown area discussed previously in Sections **??** & **??**.

The density function of a gamma distribution is composed of a function containing two parameters, the shape parameter $\alpha_{i,j}$ and the rate parameter $\beta_{i,j}$, where i is the cluster index and j

indicates the data source (Inventory or LiDAR).

A summary of the distribution fits can be found in Tables **??** - **??**.

Table 4.6: Summary of gamma distribution fit on tree diameter of inventory data by MLE

Cluster	Shape	Rate	Log-Likelihood	AIC	BIC
1	6.012	0.225	-20512.63	41029.25	41042.47
2	3.234	0.238	-1498.457	3000.914	3008.548
3	3.810	0.124	-17076.31	34156.61	34169.29

Table 4.7: Summary of gamma distribution fit on predicted tree diameter of LiDAR data by MLE

Cluster	Shape	Rate	Log-Likelihood	AIC	BIC
1	9.013	0.380	-4844594	9689191	9689216
2	4.595	0.109	-122176.7	244357.3	244373.8
3	5.925	0.161	-17076.31	2880304	2880325

This leads to the second step of this procedure, i.e. calculating the correction terms

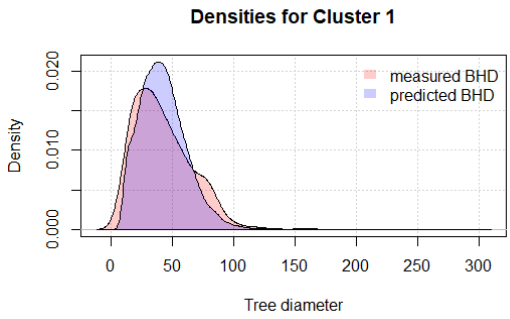
$$c_{\alpha_i} = \frac{\alpha_{i,invent}}{\alpha_{i,lidar}} \quad \text{and} \quad c_{\beta_i} = \frac{\beta_{i,invent}}{\beta_{i,lidar}}$$

for every cluster i.

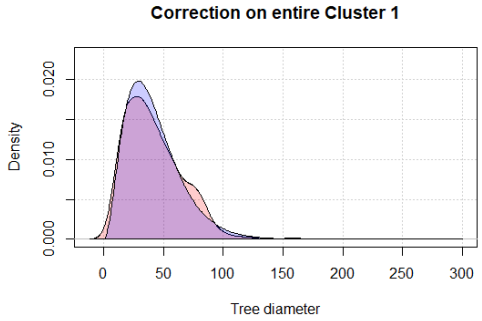
By multiplying the correction terms with the respective shape and rate parameters of the LiDAR based fit, we obtain the new (corrected) parameters:

$$\alpha_i^* = c_{\alpha_i} * \alpha_{i,lidar} \quad \text{and} \quad \beta_i^* = c_{\beta_i} * \beta_{i,lidar}$$

To demonstrate that the corrected parameters return a new and significantly less biased distribution as we would expect, Figures **??** - **??** are depicted to compare the old and new densities of the LiDAR diameters with respect to the diameter densities of the inventory. It shall be emphasized that the densities from the inventory data serve as a reference for how the actual densities may look like.

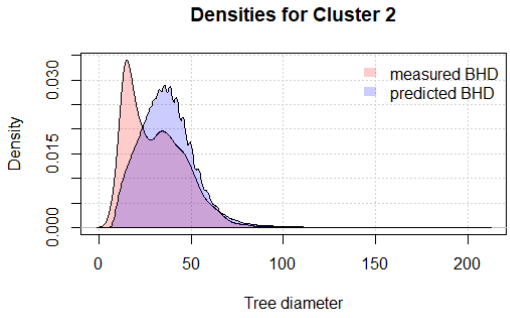


(a) Densities before applying correction terms

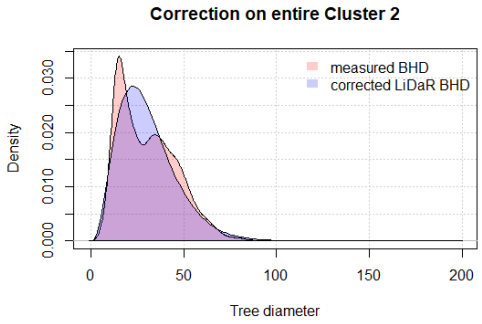


(b) Densities after applying correction terms

Figure 4.14: Measured vs predicted densities of the tree diameter for the 1^{st} cluster - Comparison

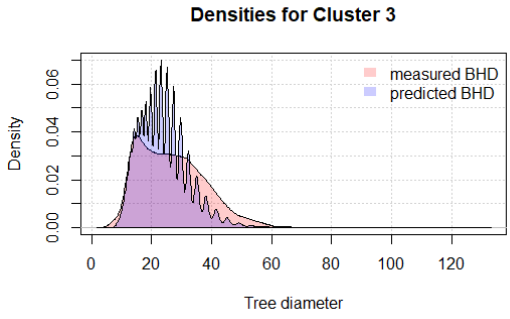


(a) Densities before applying correction terms

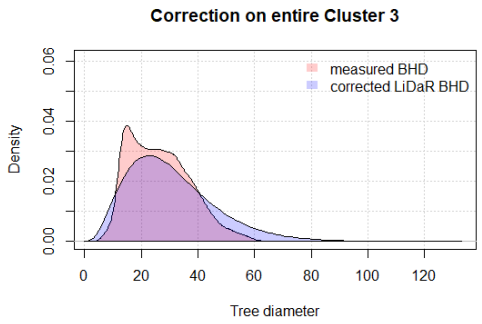


(b) Densities after applying correction terms

Figure 4.15: Measured vs predicted densities of the tree diameter for the 2^{nd} cluster - Comparison



(a) Densities before applying correction terms



(b) Densities after applying correction terms

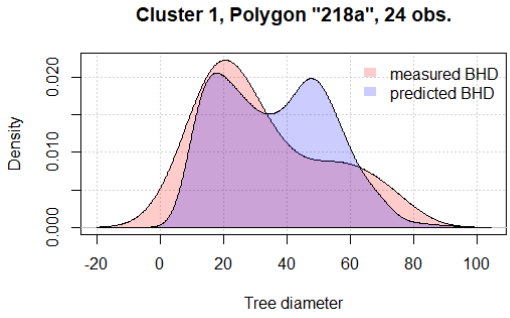
Figure 4.16: Measured vs predicted densities of the tree diameter for the 3^{rd} cluster - Comparison

Nevertheless, the aim of this technique is to reduce bias on a lower hierarchical level, namely the compartments. Therefore, for every single compartment a gamma distribution is obtained with (corrected) shape and rate parameters

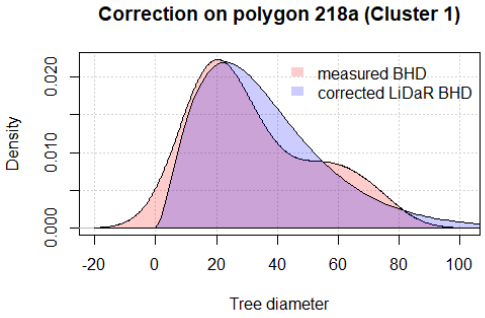
$$\alpha^*_{i,k} = c_{\alpha_i} * \alpha_{i,k} \quad \text{and} \quad \beta^*_{i,k} = c_{\beta_i} * \beta_{i,k},$$

where i is the cluster allocation of compartment k, $\alpha_{i,k}$ and $\beta_{i,k}$ are the shape and rate parameters of the LiDAR based fit on compartment k.

Some examples can be seen in Figures **??** - **??**.

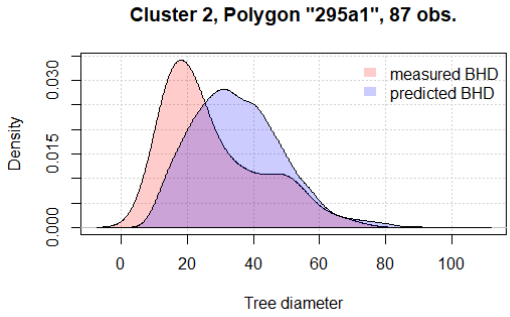


(a) Densities before applying correction terms

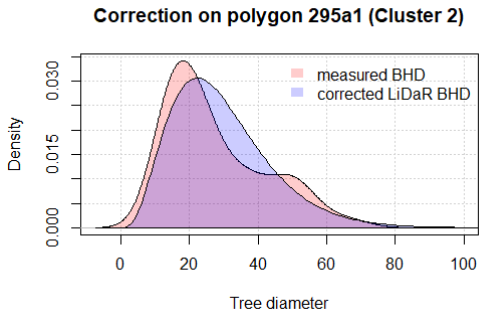


(b) Densities after applying correction terms

Figure 4.17: Measured vs predicted densities of the tree diameter for compartment 218a

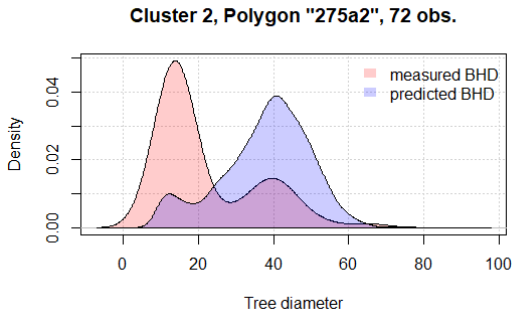


(a) Densities before applying correction terms

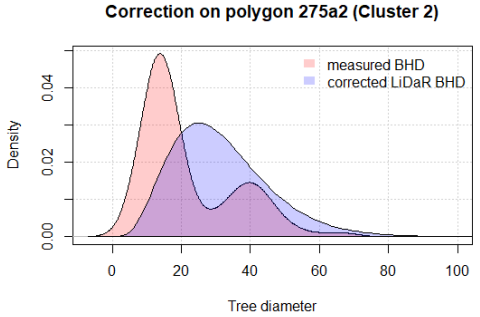


(b) Densities after applying correction terms

Figure 4.18: Measured vs predicted densities of the tree diameter for compartment 295a1

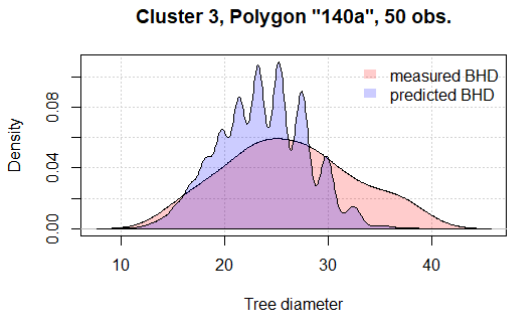


(a) Densities before applying correction terms

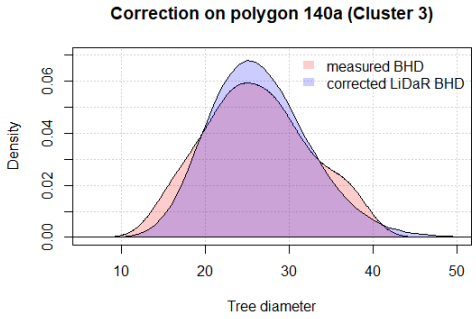


(b) Densities after applying correction terms

Figure 4.19: Measured vs predicted densities of the tree diameter for compartment 275a2



(a) Densities before applying correction terms



(b) Densities after applying correction terms

Figure 4.20: Measured vs predicted densities of the tree diameter for compartment 140a

Advantageously, corrected diameter densities from LiDAR data can be estimated even if there was no sampling carried out on a specific compartment. All that is needed is the compartment's allocation to a certain cluster. Another resolved issue using this method is that the corrected densities do no longer show long tails, meaning that overestimation is dealt with efficiently.

4.6 Validation of the Results

The methodology is validated by comparing the predicted mean diameter to the mean diameter from the inventory data on compartment level. Figure **??** depicts the results, whereby the predicted mean radius is obtained by sampling 1000 diameters from the corrected density distribution of the appropriate compartment. Only compartments with 40 or more inventory data points are considered. Table **??** depicts the summary statistics of the differences between sampled and observed values.

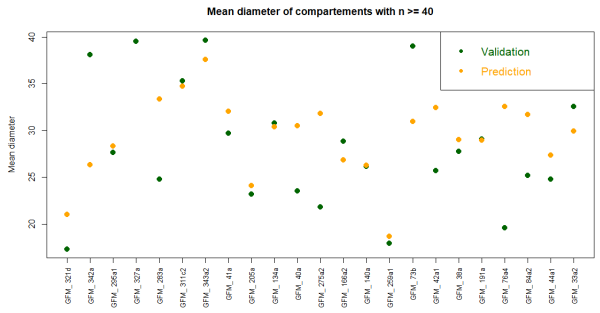


Figure 4.21: Sampled mean diameter vs observed mean diameter for compartments with sufficient samples.

Table 4.8: Summary Statistics of the observed diameter per compartment subtracted the mean sampled diameter

Median	Mean	RSE
-2.8218	-2.4574	5.487

The mean diameter is overestimated in the median and mean by 2.82cm and 2.46cm respectively. The residual standard error is 5.487cm.

5 Discussion

The objective of this study is to contribute to the development of LiDAR assisted forest inventories by developing a statistical sound approach to derive unbiased diameter distribution models from LiDAR data for homogeneous compartments of the forest. Prediction is successfully performed by a generalized linear regression model using the gamma distribution with log as a link function. Clustering the compartments into three groups led to enough samples per cluster to perform distribution engineering. Clustering is not meant to find groups upon a decision for more or less correction is made, but to find an appropriate correction for similar compartments. Clustering further uncovered the issue of not detecting equally height trees, which introduced additional bias in the tree diameter prediction. Bias correction is achieved by fitting a gamma distribution on the diameter distribution of each cluster from the inventory dataset and subsequently the predicted diameter distribution. A correction factor is calculated based on the ratio of the shape and scale parameter of both fitted distributions for each cluster. High confidence is thus given to the inventory dataset. Subsequently, the correction factor is applied for each section.

This approach could solve all present challenges, without relying on any heuristic methods apart from choosing an appropriate amount of clusters and variables. Hence the objective of finding a statistical sound approach is achieved.

We could not find studies with similar approaches to the objective. The achieved residual standard error of the mean diameter of the corrected compartment distribution of 5.49cm (see Section **??**) is satisfying. To compare and assess the modelling of the distribution, a inventory dataset of fully sampled compartments is necessary. The RSE could be further reduced by improving the tree species detection rate and likely crown area estimation. Comparing different amount of k cluster could be an interesting extension. The residual standard error of the regression model with 3.81cm is compared to another study. G. Liu, J. Wang, P. Dong, Y. Chen, Z. Liu (2018) [17] achieved a significantly better diameter residual standard error of 1.28 cm using solely LiDAR data. *"Octree segmentation, connected component labelling and random Hough transform are comprehensively used to identify trunks and extract DBH of trees in sample plots."*

Nevertheless, this sophisticated approach can only be applied on plot level (small sampling location in a forest) and likely not scaled up on a whole forest. Ultimately, a residual standard

error of below 5cm is satisfying.

The advantage of the presented approach is that an extension of the estimation of the unbiased tree diameter distribution based on just the LiDAR scanning system can be achieved, even though the majority of the area did not undergo any manual sampling activities.
Additionally, by making use of this bias correction attempt by fitting and adjusting parametric distributions instead of just relying on the predicted diameter distribution, outlier occurrences at the outer quantiles are no longer an issue (due to overestimated crown areas), meaning that overestimation of the tree diameter is also prevented.

Appendix

Sparse Data Case

As mentioned before, at the beginning of this study the availability of attributes in the LiDAR dataset was limited to just the height and location of the detected trees. A problem, as only regression models using those parameters could be included. Expert knowledge and further research highly advised using tree species and potentially the crown-area.

Consequently, the entire forest must be rearranged in such a way that sections with a similar structure are clustered together. The clustering of the sections can be highly advantageous. Important information of variables can be explained by the clusters. For example, there might be two areas with different dominant tree species. Those two species differ significantly in height and diameter. Clustering based on some variables could then detect those areas to be different and thus separate them.

Subsequently, for each cluster of multiple sections a regression model is created and later applied on the LiDAR data, resulting in better predictions.

To cluster sections, each of them requires numerical values. As discussed, the height and diameter are intuitive variables which can be used. Thus, the mean and variance for both variables are calculated for each tree in the individual sections and then assigned to them, allowing to apply common clustering methods.

In conclusion, each section four variables are assigned describing the structure of its tree (mean diameter, mean height, variance diameter, variance height). They are then clustered with the goal that subsequent regression models describe as much variation as if important variables (tree species, crown area) would exist.

We compared k-means and hierarchical clustering (single & complete linkage). Hierarchical methods group two data points iteratively, until one cluster containing all observations exist. A group is built with two variables which have the minimal distance compared to all other data points. After grouping one pair, the distance from each variable to the newly grouped variables is calculated and the next grouping step begins. The distance can be calculated with different linkage methods. Minimum distance for single linkage and maximum distance for complete

linkage was compared.

Hierarchical clustering will therefore group outliers at the very end, resulting in several clusters with only one section (outlier). This is undesirable, as one should consider that the auxiliary variables used for clustering are based on few samples. Applying regression models on sparse sections will likely lead to overfitting. Thus, hierarchical clustering is discarded.

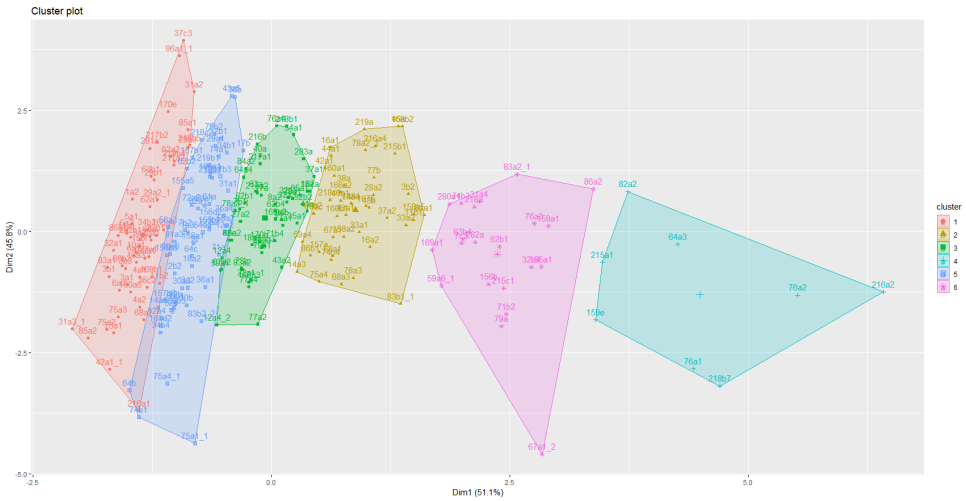


Figure 5.1: Clustered sections with respect to mean and variance of both height and diameter using principal components.

Figure **??** provides a visual representation using the first two Principle Components and then draw an ellipse around each cluster (using R package factoextra[8]).

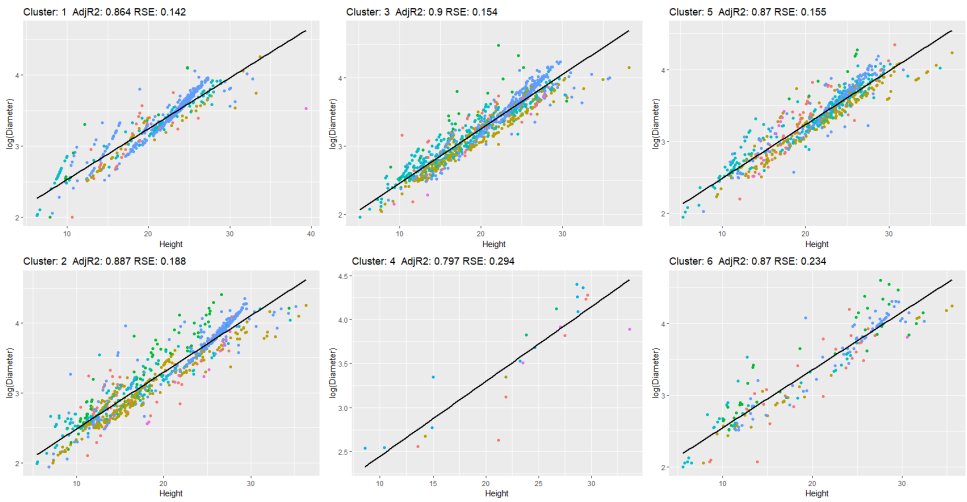


Figure 5.2: : Log-linear regression models log(diameter) ~height for each cluster. The color indicates the tree species.

Table 5.1: Estimates of the log-linear models based on k-means clustering

Cluster	Intercept	Height
1	1.7428	0.0745
2	1.6768	0.0808
3	1.5997	0.0849
4	1.6607	0.0795
5	1.7386	0.0805
6	1.8159	0.0713

Regression models on each cluster showed satisfying results with adjusted R^2 around 0.87 for the largest cluster (see Figure **??**). After acquiring the additional information on crown area and species group, this approach was neglected.

References

[1] ForestEye Research GmbH und ARGUS Forstplanung, 2018: Handanweisung zur Betriebsinventur und Forsteinrichtung in den Gräflich Bernstorff'schen Betrieben, Forstamt Gartow, Version 23.04.2018

[2] Airborne laser scanning - an introduction and overview, 1999: Aloysius Wehr, Uwe Lohr, ISPRS Journal of Photogrammetry and Remote Sensing. S. 68 - 82

[3] Das Laserscanning. Eine neue Datenquelle zur Erfassung der Topographie, 2004: Karl Kraus, Paul Dorninger, Wiener Schriften zur Geographie und Kartographie. S. 312 - 318.

[4] On the distribution of a variate whose logarithm is normally distributed, 1941, Finney, D. J., Journal Royal Statiscal Society, v.7, p.155-161, 1941. Supplement.

[5] Regression: Models, Methods and Applications. Berlin: Springer-Verlag. Ludwig Fahrmeir, Thomas Kneib, Stefan Lang, Brian Marx (2013)

[6] Matching remotely sensed and field-measured tree size distributions. Jari Vauhkonen, Lauri Mehtätalob. Canadian Journal of Forest Research, 2015, 45(3): 353-363

[7] Applied Multivariate Statistical Analysis (6th Edition), Richard A. Johnson, Dean W. Wichern. (02 April 2007)

[8] factoextra: Extract and Visualize the Results of Multivariate Data Analyses, Alboukadel Kassambara and Fabian Mundt, 2017, R package version 1.0.5
https://CRAN.R-project.org/package=factoextra

[9] jtools: Analysis and Presentation of Social Scientific Data, Jacob A. Long, 2018, R package version 1.1.1
https://cran.r-project.org/package=jtools

[10] plotly: Create Interactive Web Graphics via plotly.js. R package version 4.7.1. Carson Sievert, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec

and Pedro Despouy (2017).
https://CRAN.R-project.org/package=plotly

[11] ggplot2: Elegant Graphics for Data Analysis. H. Wickham. Springer-Verlag New York, 2016

[12] Projektbericht: Orthobild-Mosaik und Airborne Laserscanning Forsteinrichtung Gartow, Marko Pilger, GEOCART, 06.11.2018, Version 1.0

[13] Mathematical Methods of Statistics. Cramer, H. (1946) Princeton University Press, Princeton.

[14] Introduction to Mathematical Statistics, 4th edition. R. V. Hogg and A. T. Craig (1978) New York: Macmillan

[15] R.A. Fisher and the making of maximum likelihood 1912-1922. Aldrich, John. (1997). Stat Sci. 12. 10.1214/ss/1030037906.

[16] fitdistrplus: An R Package for Fitting Distributions. Marie Laure Delignette-Muller, Christophe Dutang (2015). Journal of Statistical Software, 64(4), 1-34.
https://cran.r-project.org/web/packages/fitdistrplus/vignettes/paper2JSS.pdf

[17] MDPI and ACS Style. Estimating Individual Tree Height and Diameter at Breast Height (DBH) from Terrestrial Laser Scanning (TLS) Data at Plot Level. Liu, G.; Wang, J.; Dong, P.; Chen, Y.; Liu, Z. Forests 2018, 9, 398.

List of Figures

List of Tables

List of Abbreviations

CHM Canopy Height Model

DBH Tree diameter (measured at 1.3m)

GAMLSS Generalized Additive Model for Location Scale and Shape

GIS Geographical Information System

LM Linear Model

GLM Generalized Linear Model

LiDAR Light Detection and Ranging

LADAR Laser Detection and Ranging

GPS Global Positioning System

RSE Residual Standard Error

SE% Relative Standard Error

IMU Inertial Measurement Unit

CI Confidence Interval

RSS Sum of Squared Residuals

MLE Maximum Likelihood Estimation

AIC Akaike's Information Criterion

BIC Bayesian Information Criterion

Tree Species

Bu Beech

Dgl Douglas fir

Ei Oak

Fi Spruce

Ki Pine

Lä Larch

SLh Other Hardwood

SNh Other Softwood

Ta Fir