GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Master Thesis

Multivariate modelling of the dependency structure between article sales of a sportswear manufacturer

Author

Petros Christanas

from Nuremberg

Matriculation Number

11604278

Applied Statistics M.Sc.

Chair of Statistics and Econometrics

Supervisors

Prof. Dr. Thomas Kneib

Dipl.-Vw. Quant. Fabian H. C. Raters

Submitted February 18, 2020

Processing time of 20 weeks

Confidentiality Clause

Write in here the text for the confidentiality clause!

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

Signature

Petros Christanas

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

Write introduction here and "upper" subsections here (adidas, Motivation, etc...)

1.1 Data Sources

Throughout each season, transactional data are collected from online purchases of the sports brand's eCommerce website. Specifically, we are provided with weekly sales data for western European countries depicted in table 1.1.

Column	Description	Values
week_id	Calender week of a specific year (YYYYWW)	Factors: *201648, ..., 201852*
article_number	Unique article identification number (article ID)	Factors: *10669, 10, ...*
min_date_of_week	Minimum date of the respective week; always a Monday (YYYY-MM-DD)	Dates: *2016-11-28, ..., 2018-12-24*
art_min_price	Minimal recorded price of the article	Non-negative (integer) value
month_id	Calender month of a specific year (YYYYMM)	Factors: *201612, ..., 201812*
season	Season of year (format: SSYY) (Spring-Summer [SS]: December - May) Fall-Winter [FW]: June - November)	Factors: *SS17, FW17,* *SS18, FW18, SS19*
bf_w	Weekly "Black Friday" promotion intensity of the article	Between 0 and 1
ff_w	Weekly "Friends & Family" promotion intensity of the article	Between 0 and 1
ot_w	Weekly article promotion intensity of "Other" type	Between 0 and 1
gross_demand_quantity	Weekly amount of added articles to shopping cart	Non-negative (integer) value
base_price_locf	Retail price of the article without any discounts	Non-negative (integer) value
total_markdown_pct		
day_of_month	Day of the month	Integers: *1 - 31*
month_of_year	Month of the year	Factors: *January, ..., December*
year	Year	Integers: *2016, 2017, 2018*
week_of_year	Week of the year	Integers: *1 - 52*

Table 1.1: Transactional raw data description from online purchases of western European countries

Due to legal regulations of the company, some columns had to undergo anonymization in order for the data to be released. To ensure data protection and confidentiality, numeric variables (with exception of time-indicating columns) were transformed. As a consequence for the analysis part, most integer values were converted to float numbers. This fact should be kept in mind by the reader, since the above table serves as a reminder and reference point for the data documentation.

Another peculiarity of this setup is to be considered, too. We will often refer to the variable *gross demand quantity* as *sales*, even though it is obviously not

exactly the same. In the eCommerce environment, there are several stages before the purchase is complete, e.g. addition to cart, removal from cart, proceeding to checkout & even the return of bought articles. Targeting the articles added to cart, i.e. the (gross) demand quantity, provides the optimal data extraction for analytical purposes and is the closest to adequately model the dependency structure between net sales of articles.
Besides the transactional data, *article master data*, i.e. attributes of the articles, are provided and depicted in table 1.2.

Column	**Description**	**Values**
week_id	Calender week of a specific year (YYYYWW)	Factors: *201648, ..., 201852*
article_number	Unique article identification number (article ID)	Factors: *10669, 10, ...*
min_date_of_week	Minimum date of the respective week; always a Monday (YYYY-MM-DD)	Dates: *2016-11-28, ..., 2018-12-24*
art_min_price	Minimal recorded price of the article	Non-negative (integer) value
month_id	Calender month of a specific year (YYYYMM)	Factors: *201612, ..., 201812*
season	Season of year (format: SSYY) (Spring-Summer [SS]: December - May) Fall-Winter [FW]: June - November)	Factors: *SS17, FW17,* *SS18, FW18, SS19*
bf_w	Weekly "Black Friday" promotion intensity of the article	Between 0 and 1
ff_w	Weekly "Friends & Family" promotion intensity of the article	Between 0 and 1
ot_w	Weekly article promotion intensity of "Other" type	Between 0 and 1
gross_demand_quantity	Weekly amount of added articles to shopping cart	Non-negative (integer) value
base_price_locf	Retail price of the article without any discounts	Non-negative (integer) value
total_markdown_pct		
day_of_month	Day of the month	Integers: *1 - 31*
month_of_year	Month of the year	Factors: *January, ..., December*
year	Year	Integers: *2016, 2017, 2018*
week_of_year	Week of the year	Integers: *1 - 52*

Table 1.2: Article master data

2 Theory & Methods

This chapter introduces various statistical methods used during the conduction of this thesis. It is assumed that basic understanding and knowledge of the reader regarding mathematical foundations of statistics (like linear algebra, probability theory, etc) already exists.

2.1 Generalized Linear Models

Generalized Linear Models (GLMs) are an extension of the *Classical Linear Regression Model (LM)*

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n$$

which in matrix notation can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where the response variable y_i can take values from several probability distributions (e.g. Poisson, Binomial, Gamma and others), which are members of the exponential family [Fahrmeir et al., 2003]. The linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

is passed through a *response function h* (a one-to-one, twice differentiable transformation), such that

$$E(y_i) = h(\eta_i)$$

i.e. the expected value of the response variable belongs to the value range of *h*. The inverse of the response function,

$$g = h^{-1}$$

is called the *link function*.

2.2 Generalized Additive Models

2.3 Mixed Effects Models

Linear Mixed Models (LMMs) are powerful tools when dealing with clustered data or data with a longitudinal structure (repeated measurements of individuals). As

in the classical LM, there are population-specific effects, namely the parameter vector of *fixed effects* $\boldsymbol{\beta}$, as well as the cluster- or individual-specific effects of such models called *random effects* [Fahrmeir et al., 2003]. In the following, we will refer to our clusters or individuals as "groups" for briefness. Mathematically speaking, the linear predictor $\eta_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta}$ is extended to

$$\eta_{ij} = \boldsymbol{x'}_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n_i, \quad \text{where}$$

- i is the number of groups
- j is the number of observations per group
- $\boldsymbol{\beta}$ is the vector of fixed effects
- $\boldsymbol{\gamma}_i$ is the vector of random effects
- $\boldsymbol{x}'_{ij}$ is the vector of covariates and
- $\boldsymbol{u}'_{ij}$ is a subvector of $\boldsymbol{x}'_{ij}$.

$\boldsymbol{x}'_{ij} = (1, x_{ij1}, \ldots, x_{ijk})$ & $\boldsymbol{u}'_{ij} = (1, u_{ij1}, \ldots, u_{ijk})$ are therefore the design vectors and ϵ_{ij} are the error terms of the *measurement model*

$$y_{ij} = \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \boldsymbol{u}'_{ij}\boldsymbol{\gamma}_i + \epsilon_{ij}, \quad \varepsilon_{ij} \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right) \tag{2.1}$$

or in matrix notation
WRITE THE MATRIX NOTATION HERE

3 Study Area & Data Sources

3.1 Forest Classification

The study area is a private forest enterprise in Gartow (Niedersachsen) Germany. It measures around 5674.2 ha in total (see Table 3.1) and is a relatively homogenous forest consisting mostly of pine trees.

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

Table 3.1: 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.

The forest itself is split into stratum to take site conditions, forest structure and thus natural variation of the areas into account (see Figure 3.2). The assessment of variation was based on a forest inventory conducted 2008 (see Table 3.2).

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

Table 3.2: 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

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 3.1)

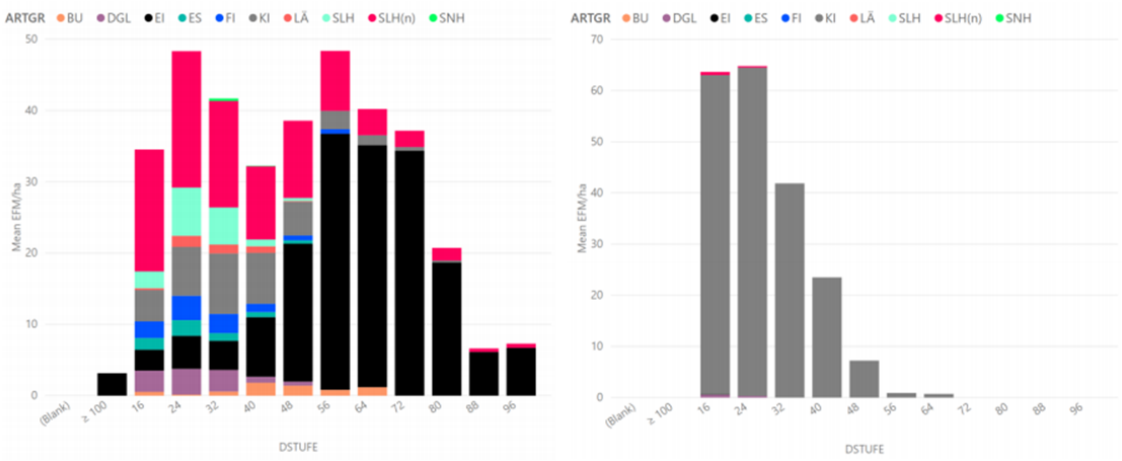

Figure 3.1: The bar plots (left to right: stratum 1, stratum 4). The bars indicate the mean volume per ha for different diameter classes [1].

Sampling activities are adjusting according to the inhibited variation of the stratum type (see Section 3.2).

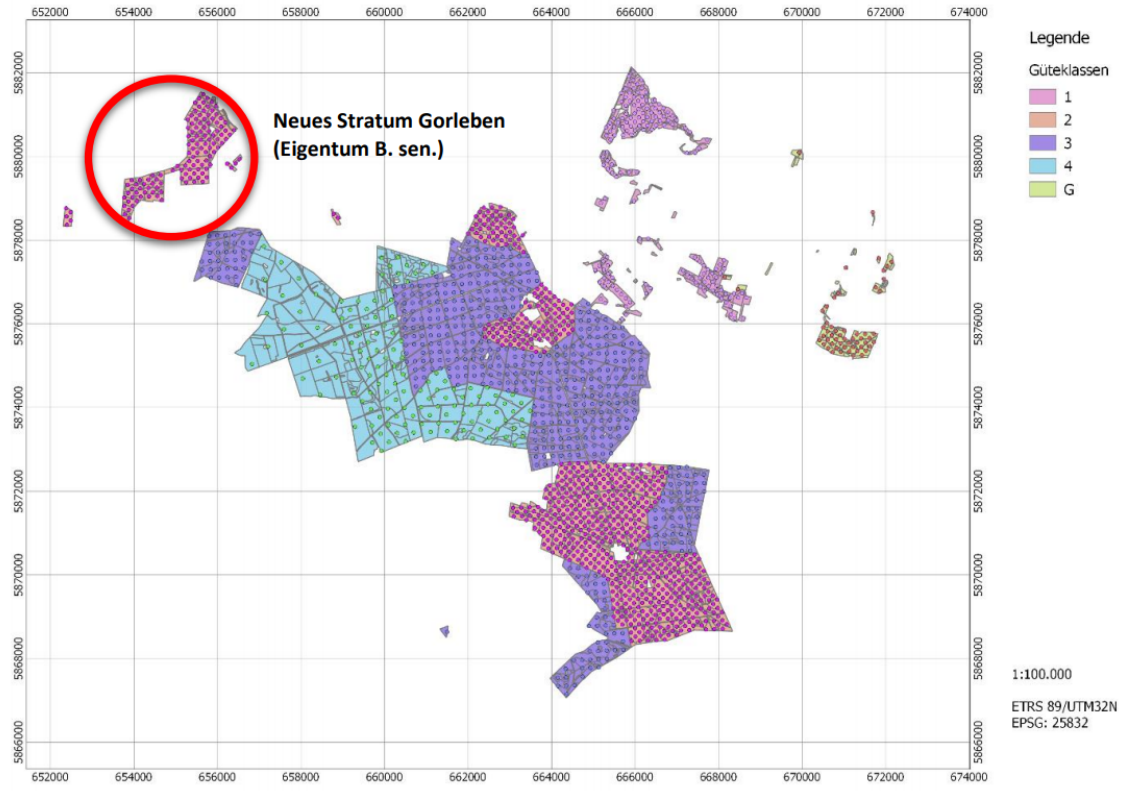

Figure 3.2: The forest of Gartow is divided into stratums based on past observed variation and ownership and subdivided into compartments indicated by grey lines. Each point within a section indicate a sampling point [1].

The stratum themselves are subdivided into compartments with the intention to create homogeneous sub regions (see Figure 3.2). The diameter distribution must be found for each of those compartments.

3.2 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 3.2 & Figure 3.2).

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 3.3 provides an overview of total measured trees.

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

Table 3.3: Overview of number of measured trees for height and diameter per Stratum

3.3 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 3.3).

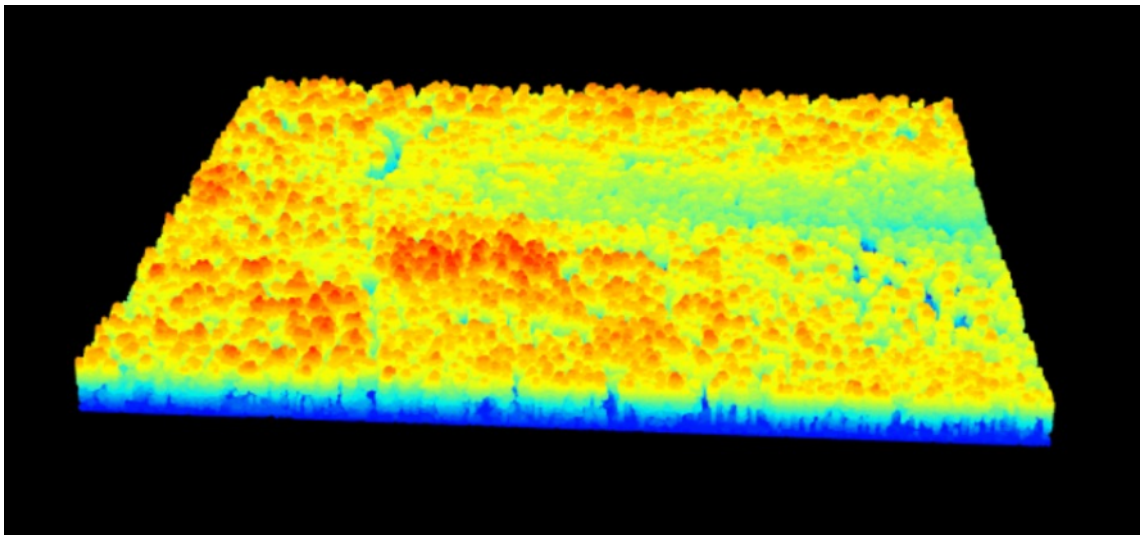

Figure 3.3: 3-D Image of a small area of the forest of Gartow made by the airborne LiDAR. The determined height is colorized. A dense group of height trees is found almost in the middle and directly behind an aisle of small trees.

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.

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

Table 3.4: Flight log of the airborne laser scanning of the forest of Gartow [12]

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

4 Data Exploration

Let's begin with the data exploration

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left((y - X\beta)'(y - X\beta) + \lambda \sum_{j} \beta_j^2 \right)$$

jkhnvcv

5 Conclusion

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.

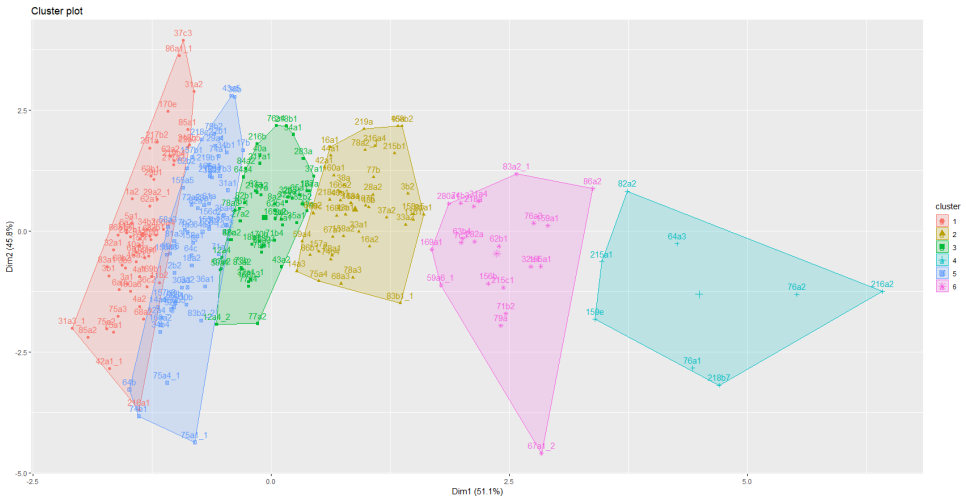

Figure 5.1: Clustered sections with respect to mean and variance of both height and diameter using principal components.

Figure 5.1 provides a visual representation using the first two Principle Components and then draw an ellipse around each cluster (using R package factoextra[8]).

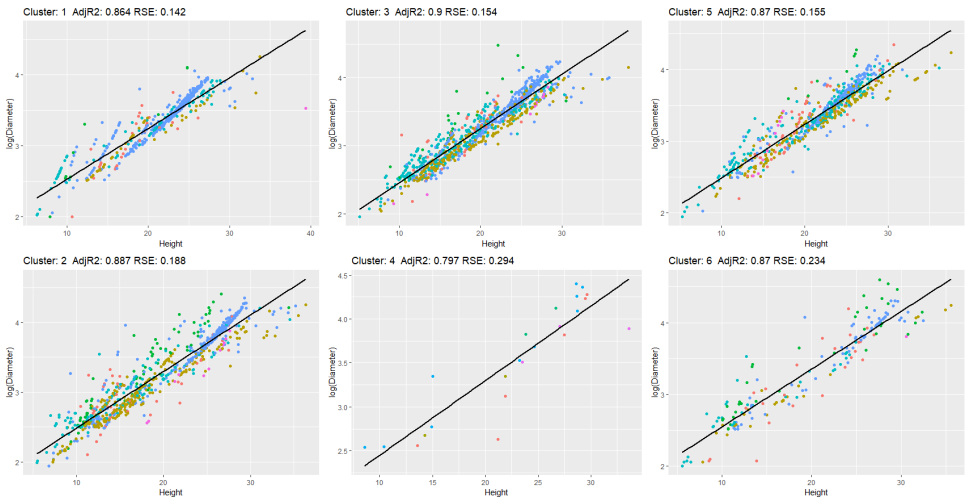

Figure 5.2: : Log-linear regression models log(diameter) ~height for each cluster. The color indicates the tree species.

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

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

Regression models on each cluster showed satisfying results with adjusted R^2 around 0.87 for the largest cluster (see Figure 5.2). After acquiring the additional information on crown area and species group, this approach was neglected.

List of Figures

List of Tables

List of Abbreviations

BIC Bayesian Information Criterion

GLM Generalized Linear Model

LM Linear Model

LMM Linear Mixed Model

References

[Fahrmeir et al., 2003] Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2003). Regression; Models, Methods and Applications. 2013.

[Lütkepohl, 2005] Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.