UNIVERSITY OF KWAZULU-NATAL

COVARIATES AND LATENTS IN GROWTH MODELLING

By

SILESHI FANTA MELESSE

2014

COVARIATES AND LATENTS IN GROWTH MODELLING

By

SILESHI FANTA MELESSE

Submitted in fulfilment of the academic

requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

in the

School of Mathematics, Statistics and Computer Science

University of KwaZulu–Natal

Pietermaritzburg

2014

Dedication

To my late father, Fanta Melesse Negussie and

my mother Yeshi Haile Woldmariam

Declaration

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Sciences, University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Professor Temesgen Zewotir.

I, Sileshi Fanta Melesse, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

April, 2014

________________________ _______________

Sileshi Fanta Melesse Date

_______________________ ________________

Professor Temesgen Zewotir Date

Note

The following papers have been published from this thesis.

1. Melesse, S.F. and Zewotir, T. (2013 a) The effect of correlated climatic factors on the radial growth of eucalypt trees grown in coastal Zululand of South Africa. African Journal of Agricultural Research, 8(14):1233-1244 Available online at http://www.academicjournals.org /AJAR.

2. Melesse, S.F. and Zewotir, T. (2013 b) Path models approach to the study the effect of climatic factors and tree age on the radial growth of juvenile eucalyptushybrid clones. African Journal of Agricultural Research, 8(22): 2685-2695 Available online at http://www.academic journals.org/AJAR.

Acknowledgement

I was fortunate and privileged to have Prof. Temesgen Zewotir as my supervisor. He encouraged me to start my PhD study. He stimulated me to work with the Sappi Forests Research data, and introduced me to Dr. Valarie Grzeskowiak and Dr. Nicky Jones. He facilitated the process of getting the data from Sappi. It would have been impossible for me to reach the current stage if not for the all rounded support and encouragement of Prof. Temesgen Zewotir throughout my PhD training. He gave me excellent supprort and guidance and supported me in the days of difficulties.

My heartfelt thanks go to Dr.Valarie Grzeskowiak and Dr Nicky Jones who shared their ideas in every progress report presentation I made. Nicky's editorial correction to every paper I drafted for publication was so instrumental. I am grateful to Sappi Forests Research for providing the data set, without which the work may not have been completed.

I am so grateful to all members of the department of Statistics, Addis Ababa University, for supporting the decision I made to start my PhD. I have enjoyed closely working with Dr. M.K Sharma, when I was in Addis Ababa University. Dr. Sharma encouraged me so much to continue my study and openly told me that it is possible to do it. Several E-mails that I exchanged with Dr. Sharma are unforgettable and witnessed that he is indeed a true friend and colleague.

I wish to express my deep and genuine thanks to Professor Delia North, whose contribution to my study has been valuable. She supported me in all aspects especially during my difficult times she was the one to be contacted. Her immense support goes way beyond the academic support.

To friends and colleagues at UKZN, thank you to all of you. Many thanks to Pietermaritzburg staff for sharing some ideas with me for the last three years.

I would like to thank my parents, relatives, brothers and sisters whose support was the source of encouragement throughout my life.

Last, and by no means the least, the biggest vote of thanks goes to my delightful wife, Yodit Afework G/ Michael, for her love, care and tireless encouragement. My wife kept the family intact and helped me to concentrate on my research work.

I would like to apologize to my two daughters Megnot and Meron for the time not spent with them during the last three and half years and I am very grateful for their understanding.

Abstract

The growth curve models are the natural models for the increment processes taking place gradually over time. When individuals are observed over time it is often apparent that they grow at different rates, even though they are clones and no differences in treatment or environment are present. Neverthless the classical growth curve model only deals with the average growth and does not account for individual differences, nor does it have room to accommodate covariates. Accordingly we strive to construct and investigate tractable models which incorporate both individual effects and covariates.

The study was motivated by plantations of fast growing tree species, and the climatic and genetic factors that influence stem radial growth of juvenile *Eucalyptus* hybrids grown on the east coast of South Africa. Measurement of stem radius was conducted using dendrometres on eighteen sampled trees of two *Eucalyptus hybrid* clones (*E. grandis x E.urophylla, GU and E.grandis x E. Camaldulensis, GC).* Information on climatic data (temperature, rainfall, solar radiation, relative humidity and wind speed) was simultaneously collected from the study site.

We explored various functional statistical models which are able to handle the growth, individual traits, and covariates. These models include partial least squares approaches, principal component regression, path models, fractional polynomial models, nonlinear mixed models and additive mixed models. Each one of these models has strengths and weaknesses. Application of these models is carried out by analysing the stem radial growth data.

The partial least squares and principal component regression methods were used to identify the most important predictor for stem radial growth. Path models approach was then applied mainly to find some indirect effects of climatic factors. We further explored the tree specific effects that are unique

to a particular tree under study by fitting a fractional polynomial model in the context of linear mixed effects model. The fitted fractional polynomial model showed that the relationship between stem radius and tree age is nonlinear. The performance of fractional polynomial models was compared with that of nonlinear mixed effects models.

Using nonlinear mixed effects models some growth parameters like inflection points were estimated. Moreover, the fractional polynomial model fit was almost as good as the nonlinear growth curves. Consequently, the fractional polynomial model fit was extended to include the effects of all climatic variables. Furthermore, the parametric methods do not allow the data to decide the most suitable form of the functions. In order to capture the main features of the longitudinal profiles in a more flexible way, a semi-parametric approach was adopted. Specifically, the additive mixed models were used to model the effect of tree age as well as the effect of each climatic factor.

Table of Contents

List of Tables

List of Figures

Acronyms

AIC: Akaike Information Criteria

AMOS: Analysis of Momment Structure

AR: Autoregressive

ARMA: Autoregressive Moving Average

AM: Additive Models

AMM: Additive Mixed Models

BIC: Bayesian Information Criteria

CFA: Confirmatory factor analysis

CP: Conventional polynomial

CR: Critical ratio

EFA: Exploratory Factor Analysis

ECVI: Expected Cross Validation Index

FP: Fractional Polynomial

GC: Grandis Camaldulensis

GLS: Generalized Least Squares

GU: Grandis Urophylla

LME: Linear Mixed Effects

LMM: Linear Mixed Models

MA: Moving Average

MECVI: Modified Expected Cross Validation Index

MFP: Multivariate Fractional Polynomial

ML: Maximum Likelihood

MLE: Maximum Likelihood Estimation

NFI: Normed Fit Index

NLME: Nonlinear Mixed Effects

OLS: Ordinary Least Square

PCA: Principal Component Analysis

PCR: Principal Component Regression

PLS: Partial Least Squares

REML: Restricted Maximum Likelihood

RMSE: Root Mean Square Error

RMSEA: Root Mean Square Error of Approximation

RMSECV: Root Mean Square Error Cross Validation

RMSEP: Root Mean Square Error

RMSEP: Root Mean Square Error of Prediction

SEM: Structural Equation Modelling

SLCs: Standardized Linear Combination

ULS: Unweighted Least Squares

Chapter 1

Introduction

Usually growth is modelled as a function of time. The main goal of such modelling is to describe naturally occurring changes in the response over time. The other objective with respect growth could be a comparison of growth profiles for different groups.

In the absence of new computing facilities and readily available statistical software for analyzing correlated data, summary measure analysis of longitudinal data has obvious application. In the summary measures analysis the average for each individual is modelled using the standard statistical techniques. That is, the averages on different individuals are independent of one another.

Summary measure analysis can also be appealing when the sample sizes are not sufficiently large for estimation of the correlation among the repeated measures. However, despite the simplicity of the method, it does have a number of distinct drawbacks. One drawback is that it focuses on only one aspect of the repeated measures over time. When repeated measures are replaced by a single number summary, there must be some loss of information. Another problem of the summary measure approach is that the covariates must be time invariant covariates. Thus, if one of the key covariates is time varying, the method cannot be applied (Fitzmaurice, et al., 2004). Furthermore the individual variability is not taken into account. The summary method ignores the key characteristic of longitudinal data. That means the correlation between the observations on the same individual is ignored. On the other hand longitudinal studies can be used to directly study changes over time. Moreover, longitudinal methods can be used to evaluate factors that influence this change, as well as to evaluate the within-subject changes. Statistical estimates of individual changes can be used to comprehend heterogeneity in the population. Longitudinal methods can also help to understand the factors that affect growth and change at the individual level. Furthermore, in growth curve modelling, although one time-varying response may be of primary interest, the association between response and any other covariates can reveal an insightful understanding about the mechanism of change.

Notwithstanding the advantages of a longitudinal study, there are challenges in the analysis that must be addressed accordingly. In the presence of other categorical covariates (other than time), it is possible to make a separate modelling process for each level of the categorical variable. However, the challenge is to combine individual effect and covariates in the growth modelling. Measurements obtained from the same individual tend to be correlated. Measurements on the same individual close in time have a tendency to be more correlated than measurements far apart in time, and the variances of longitudinal data often change with time (Diggle et al., 2002; Fitzmaurice et al., 2004). These complicated patterns of correlation and variation may be even more complicated in the presence of more than one covariate. This complicated covariance structure must be taken into account in order to draw reliable conclusions from the data. Therefore, more complex statistical models have to be used to account for the complicated covariance structure. This calls for parameter estimation methods that can be computationally rigorous. The practical motivation of this problem emanated from the Sappi's climatic and genotype factors' study on Eucalyptus tree growth.

1.1 Motivational Background

Increasingly, eucalypts have become the most widely planted hardwood species in the world (Turnbull, 1999). At present, eucalypts provide sawn timber, mine props, pulp and paper, fiberboard, poles, firewood, charcoal, essential oils, nectar for honey, tannin, shade, and shelter. Most eucalypt plantations are established and managed for profit. The rate of growth is an important economic factor, and plantations with faster growth will be available for processing earlier compared with slower growth plantations. Tree growth and wood production is a product of the interaction between genetic (Kozlowski and Pallardy, (1997); Apiolaza et al., 2005; Zweifel et al., 2006) and silvicultural (Pallett and Sale, 2004). Some studies have found significant effects of environmental factors on wood property variation in Eucalyptus (Gallaham, 1962; February et al., 1995; Searson et al., 2004; Drew and Pammenter, 2006). Climatic factors such as temperature, humidity sunlight, rainfall (Eagleman, 1985; Miller, 2001) and wind speed (Wadsworth, 1959) contribute to the growth of plants. The knowledge of the relationships beween climatic variables and the pattern of stem growth may facilitate the prediction of wood properties for a given site. However, such studies are limited. Available studies commonly focus on growth rate pattern of growth as a function of age (Miehle et al., 2009; Crecente-Campo et al., 2010; Mateus and Tomé, 2011). Extensive literature on genetic factors affecting the growth of trees can be found in Kozlowski and Pallardy (1997). The most recent work by Downes et al. (2009) provides an excellent overview on measuring stem growth and wood formation. Other examples are those by Drew et al. (2009), which focused on differences in daily stem diametre variation and growth in two hybrid eucalypts, and Zweifel et al. (2006) who studied the intra-annual radial growth and water relations of trees and the implications on growth mechanisms.

In a study that considered the data extracted from the same database as used in this study, Drew et al. (2009) found the GU (Eucalyptus grandis x urophylla) clone had fewer days on which net growth occurred than did the GC (E. grandis x camaldulensis) clone. However, when growth did occur, the GU grew for longer each day and at a higher rate than did the GC. Thus, it still had an overall larger net stem increment during the study period. Drew et al. (2009) studied the relationship between stem radius and climatic factors using the correlation matrix. A number of post graduate researches were under taken on the data from the same data base. These are studies by Ayele (2010), Chauke (2008) and Eksteen (2012).

A study by Chauke (2008) did not consider the longitudinal nature of the data. The study by Ayele (2010), applied linear mixed model and nonlinearity is not assessed. Therefore, there are still rooms for the improvement of statistical methodlogy.

Weather variables such as temperature, solar radiation, rainfall, humidity, and wind speed all contribute to the growth of the tree. For instance, Downes et al. (1999) studied daily radial stem growth in irrigated Eucalyptus globulus and E. nitens in relation to climate over a 12-month period using multiple linear regression models. That study, which was conducted in southern Australia, showed that daily weather variations accounted for 40 to 50 percent of the variance in the daily increment of stem radius. Downes et al. (1999) also argued that understanding the relationship between weather and the rate and pattern of stem growth will facilitate the prediction of wood properties at a given site.

Our approach provides an alternative to the methods used by Downes et al. (1999) and post graduate researches conducted on the data so far. A study by Phipps (1982) presented a general discussion regarding problems inherent to developing climatically sensitive tree-ring chronologies from eastern North America. The same study by Phipps (1982) indicated that tree ring collections from eastern forests are typically not climatically sensitive as western collections. A general treatment of dendroclimatology can be found in Fritts (1976). Other studies such as those by D'Arrigo et al. (1992), Hofgaard et al. (1999) and Schweingruber et al. (1993) reported that late spring or summer temperatures had a positive effect on annual growth. Zweifel and Häsler (2001) showed that radius change could be determined by stem water content and wood bark growth, including the degradation of dead phloem cells. The water related fraction is a short-term effect lasting from a few hours to several weeks, and can either have positive or negative effects on stem radius, depending on the changing turgor of stem tissues (Zweifel and Häsler, 2001).

Most of the above studies used growth as a linear function of time/age. Nevertheless it is understandable that growth is not a simple linear function of age (Seber and Wild, 2003). The authors used a linear model because the linear model is the common model which can accommodate covariates. As the baseline growth is not linear (or the nonlinear curve is not linearized with some transformation techniques) the conclusion derived from such a linear model may not be trustworthy.

1.2. The Statistical Challenges

Usually growth curves are approximated by linear function of time. However, in reality the relationship between the response and time may not always be linear. In some cases where nonlinear relationship between time and reponse can be fitted, it is challenging to extend the model to capture the effects of other covariates. The contribution of each explanatory variable is often influenced by correlation existing among explanatory variables. However, studies that consider the effects of co-linearity are limited. Most studies commonly use the relationship between response and time as an indicator of growth rate and pattern. Moreover, in many circumstances, growth accounts only for the average response. It does not provide any information about how the responses of individuals change over time. Models, which account for within subject changes in response over time need to be considered in growth modelling. The focus of this study is to explore different models that account for the above statistical challenges and come up with a reasonably better model for the problem at hand.

1.3. The Objectives of The Study

The main objective of this thesis is to look for a reasonable model that can explain the dependence of stem radial growth on weather variables and tree age. Specifically, the thesis attempts to describe the effects of climatic variables on radial growth of *Eucalyptus grandis* × *E.urophylla (GU)* and E.grandis ×*E.camaldulensis (GC)* hybrid clones established in Zululand on eastern coast of South Africa. Moreover, the focus of this study is to determine the weather variables that may influence radial growth during juvenile (the first two years of age) stages of tree growth using some advanced modelling techniques. The study of juvenile tree growth is very important to have a productive matured tree. Identification of the relationship between natural climatic conditions and radial growth has an immense significance for eucalyptus plantation mangers. Inorder to mange resources effectively, it is important for tree growers to understand the properties of the material being produced. The findings of this study can also be useful in developing tools to identify genotypes with a better growth potential.

The rest of the thesis is organized as follows. In Chapter 2, a full description of the stem radius data is given together with the covariates, and exploratory work undertaken. In Chapter 3 we review classical growth curve models and we strive to fit baseline growth models to the Sappi data. Chapter 3 assesses the impact of climatic factors on the average stem radius growth using principal component regression and partial least square approaches. In Chapter 4, we present the structural equation models where the emphasis is on a path models approach. Chapter 5 presents a review of fractional polynomial models which account for individual tree and covariates effects. In Chapter 6, a review of nonlinear models with random effects and comparisons with fractional polynomial models was made. Chapter 7 presents the semi-parametric approaches and their applications on the problem at hand. Lastly, in Chapter 8 the discussions and conclusions are presented.

Chapter 2

Data and Exploratory Data Analysis

2.1. Study Design and Data

The data used in this study are secondary data from Sappi Forest Research Center in Tweedie. Sappi started the dendrometre trial project in July 2001. The research site is located near the town of KwaMbonambi in KwaZulu-Natal, South Africa, (28.530 S, 32.140 E, 55 M AMSL), approximately 200 km north-east of the city of Durban. On average, the site receives 1,000 mm of rainfall per annum and has a mean annual temperature of 21 degrees Celsius (Drew et al., 2009). The eucalyptus fibre research experiment was initiated in July 2001 and a huge database acquired. The experiment was designed to run over a nine-year period and was divided into separate phases. Each phase ended with the destructive sampling of study trees to measure anatomical characteristics of the wood. The results presented in this work are based on the data collected during the first of these phases, from April 2002 until August 2003. The data used by Drew et al., (2009) and this particular study are extracted from the same database put in place by Sappi (one of the leading suppliers of coated fine paper and chemical cellulose). However, the two data sets are not exactly the same. Two commercially deployed Eucalyptus hybrid clones, E. grandis x urophylla (GU) and E. grandis x camaldulensis (GC), were planted at the study site (Drew, 2004). According to the South African soil classification system, the soil was identified as Rhodic Ferralsol Hutton by a limited soil survey undertaken at the site (Schulze, 1997). The soil is medium grade sand with clay percent in the lower B horizon not exceeding 40%, and in A horizon not exceeding 10% with an average depth of A horizon 20 cm and total potential rooting depth in excess of 1.8m (Drew et al., 2009). Planting took place on 16 July 2001, prior to which in April 2001, stumps of trees from the previous rotation were treated with herbicide (to prevent coppicing), and harvest slash was burned. Each rooted cutting was planted between existing stumps, with approximately two litres of water and 125g granular fertilizer, the equivalent of 8 g N, 12 g P and 8 g K per plant. The two clones were planted in alternating rows seven trees wide each (Fig. 1), with spacing between trees of 3 metres (east to west) x 2.5 metres (north to south). These rows have been numbered from 1 to 6, starting at row (GC) closest to the entrance gate. Each row of clones consists of three plots of 12 trees each with two surrounding rows of trees (Fig. 2.1). This effectively separates each plot by four rows of trees, an important part of the design since periodic destructive sampling is required in the experiment. The plots were established as pairs, such that for any phase of the research, a GU and a GC plot could be measured simultaneously (Drew, 2004). From the 18 plots (Fig. 1), plots 9 and 10 were chosen for monitoring during project phase 1. Within a 12-tree plot, nine trees were selected from each clone for intensive monitoring of radial growth and other physiological characteristics (Drew, 2004). Measurements of stem radius were obtained from hourly dendrometre readings in the 18 sample trees. Automatic point dendrometres were mounted at nine months of age at 1.3 m above the ground on the north side of each tree to measure the radius of the main stem with a rod held against the outside surface by constant force. The tree growth data were initially recorded on an hourly basis. This makes the quantity of data for each phase large and difficult to manage.

Hourly measurements were made of total rainfall (mm), temperature (°C), relative humidity (%), wind speed (m/s) and total solar radiation (mJ/hr). Daily total rainfall and daily averages of the other weather variables were used in the analysis. Daily averages of stem radius were obtained by cumulating and averaging the hourly measurements. Accordingly daily meteorological data was obtained using an automatic weather station (MCSystems, Cape Town, South Africa) located approximately 300 m from the research trial site (Drew et al., 2009). The daily data for stem radius used in this study has 8640 observations from the two clones.

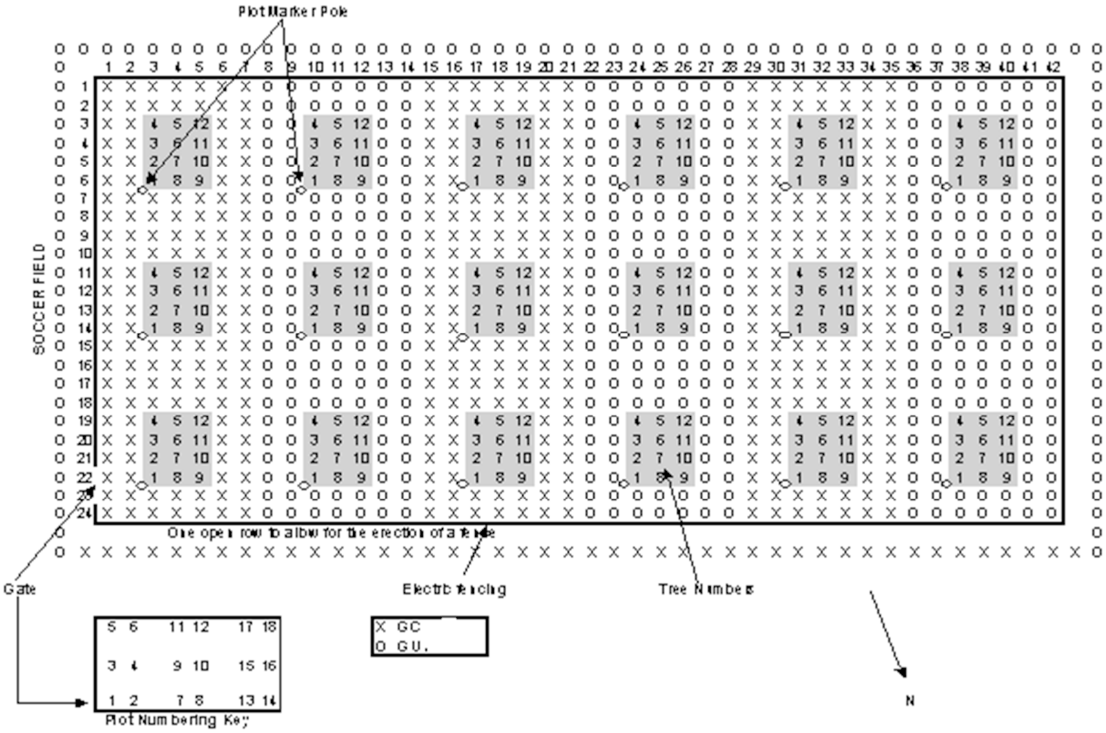

Figure 2.1 The layout of the experimental plots at the research site in eastern South Africa**.**

Half the data set is from the GU clone and the remaining half is from the GC clone. Daily measurements were used in some parts of our analysis. The observed minimum and maximum of stem radius as well as the mean and standard deviation is summarized in Table 2.1. The measurements for GU clone appear to be larger than the measurements for GC clone. The summary measures for the climatic variables are also presented in Table 2.2.

Table 2.1 Some descriptive measures for stem radius (in micro metres)

Clone	N	Mean	Standard deviation	Minimum	Maximum
GC	4320	14679.59	6424.01	12.47	31275.31
GU	4320	17371.82	8144.16	26.74	32649.92
Total	8640	16025.7	7456.78	12.47	32649.92

Table 2. 2 Some descriptive measures for climatic variables

Covariates	N	Minimum	Maximum	Mean	Standard deviation
Radius (micro-metre)	8640	26.74	32649.92	16025.71	7456.78
Temperature(°C)	8640	11.53	28.74	19.74	3.81
Rainfall (mm)	8640	0	72	1.85	6.53
Relative humidity (%)	8640	57.60	109.80	84.54	8.89
Solar radiation (mJ/hr)	8640	0.03	1.21	0.60	0.24
Wind speed (m/s)	8640	0.33	3.44	1.54	0.58

Usually longitudinal data consists of a large number of short time points (Diggle et al., 2002). In our case the data consists of large time points. Dealing with this daily data in the longitudinal context will have computational problems. Moreover, the weekly growth measurements are more meaningful than the daily measurements. Sizeable growth on each tree can be easily observed on weekly measurements as compared to daily measurements. Consequently, weekly measurements were obtained by cumulating and averaging the daily measurements. These weekly data are used to fit longitudinal growth models. A total of 1242 measurements are obtained for 18 trees each measured 69 times.

2.2 Exploratory Data Analysis

Exploratory data analysis encompasses techniques to visualize patterns of the data. Data analysis must begin by making displays that expose patterns pertinent to the scientific question. The best methods are capable of uncovering patterns which are unanticipated. In this regard graphical displays are so important. Most longitudinal studies address the relationship of a response with explanatory variables, often comprising time. In this chapter we looked at the following aspects of the data: individual profiles, the average evolution, the variance function and the correlation structure. Data exploration is very helpful in the selection of appropriate models.

The profile plot gives us an idea as to how the profile of the population evolves over time. The results of this exploration will be useful in order to choose a fixed effects structure for the mixed model. Figure 2.2 shows the plots of stem radial measurements of 18 juvenile trees against time. Some evidence of variability between and within individual trees is observed. The between tree variability is small at the early age of the tree and increases as the age of the tree increases. Trees did not maintain their relative size of stem radius over time. Trees that started with a large stem radius did not always retain the largest radial measure throughout the growth follow up period.

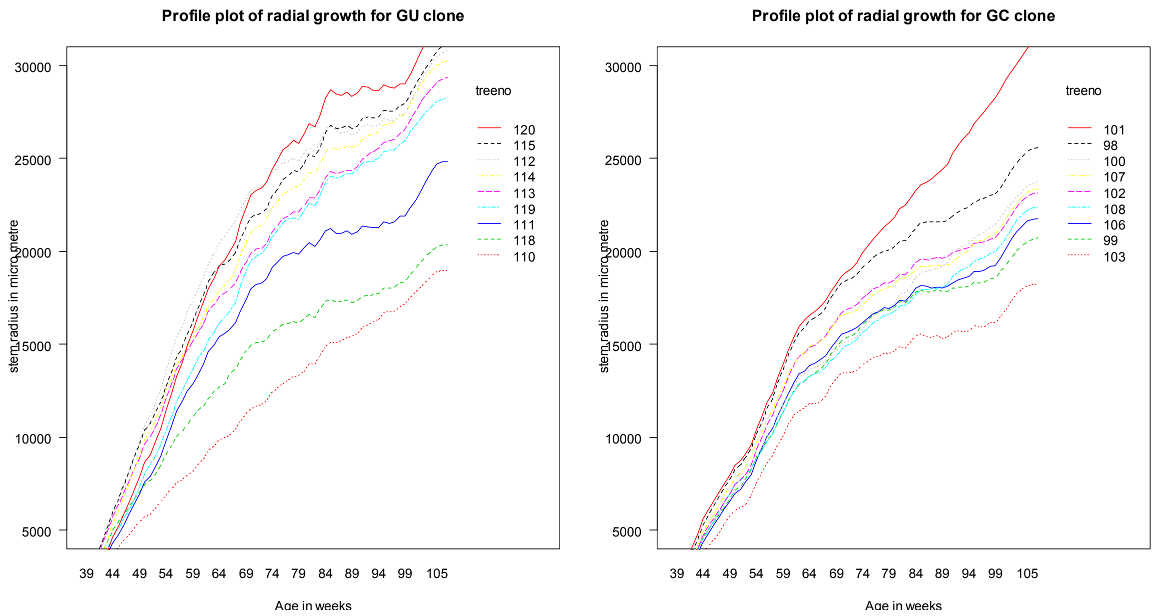

Figure 2. 2 Profile plot of stem radial measure (in micro metres) against tree age for the sampled trees of each clone, GU and GC.

The Loess smoothing technique by Cleveland (1979) is used to study the functional relationship between radial growth and tree age. Figure 2.3 shows that radial measurements were initiated at about the age of 40 weeks (when the dendrometres are attached to the trees without causing damage).

It shows a sharp increase in the estimated mean response profile of the stem radial growth from the beginning (39 weeks) up to the age of 70 weeks, and thereafter the rate of increase slows down for both clones. These curves suggest that the relationship between radial growth and age may be curvilinear (not linear). It also appears that the average profile of the GU clone is higher than that of the GC clone with the difference becoming very apparent after the age of 50 weeks.

The inferential focus of this study is on the mean response of the stem radial measure. In order to have a valid inference about the mean structure, the covariance structure must be incorporated into the statistical model. If the analysis does not take into account the correlation among repeated measures, incorrect standard errors will be produced. That means standard errors that are too large will be produced. Consequently, test statistics and p-values will also be incorrect, which leads to incorrect inferences about the parameters.

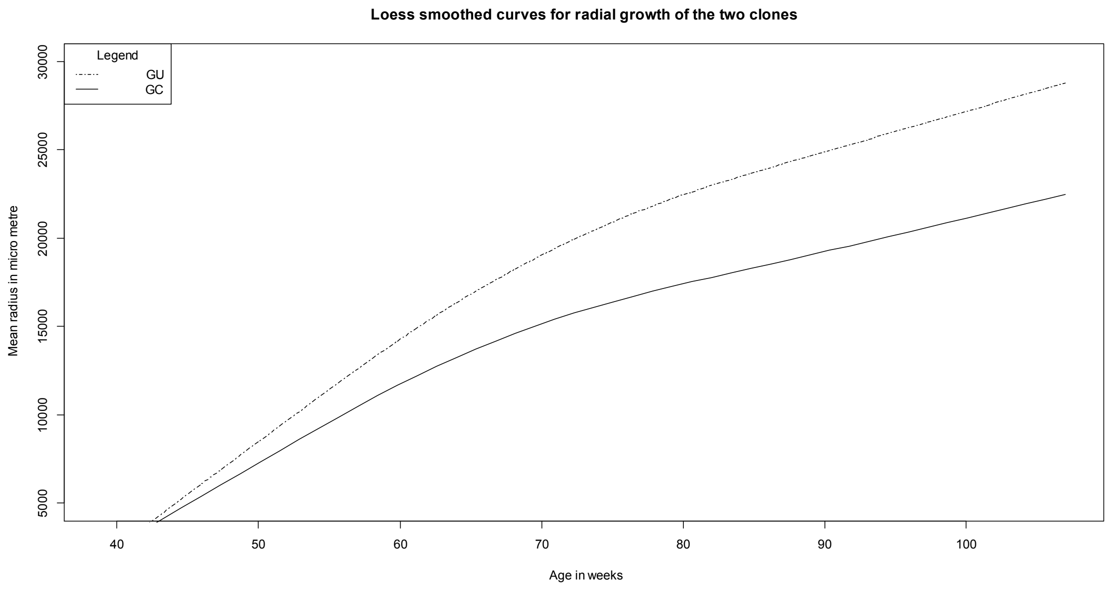

Figure 2.3 Loess smoothed curves of stem radial measure (in micro metres) against time for both clones.

In this type of longitudinal data there are at least three possible components of variability: random effects, serial correlation and measurement error (Diggle et al., 1994). Random effects are effects that arise from the characteristics of individual trees. Therefore, these effects explain the stochastic variation between trees. On the other hand, measurements of stem radius, on successive occasions of the same tree, are most likely to be serially dependent. Hence, we cannot extract as much information from these dependent observations as we could from the same number of independent measurements. That is, serial correlations mask part of the within tree variation in the data. The possibility of measurement error cannot be ignored. That is, during data collection, measurement error is expected. Therefore, these three sources of variability will be assessed in further analysis. Since the covariance structure usually accounts for all the variability in the data that cannot be explained the fixed effects, we start to explore the covariance structure by first removing all systematic trends (Verbeke and Molenberghs, 1997). Hence, residuals are obtained after regressing radial measure on time and square root of time. The estimated average evolution of the variance of the residuals at each time point for both clones is displayed in Figure 2.4. The plot indicated that the variance is not constant. It shows an increasing tendency with age for both GU and GC clones. To get more information on the nature of relationships among repeated measurements of stem radius within trees, the scatter plot matrix of the residuals for some time points was considered, as indicated in Figure 2.5. The scatter plot was made by discretizing time and selecting some time points. The upper panel of this figure shows a correlation matrix of Sresiduals for some time points. For instance, the first correlation coefficient 0.4(shown on the second panel) indicates the correlation between the residuals at time point 40 and time point 41. The correlation coefficient 0.1 (at top right corner of the graph) is the correlation between the residuals at time point is equal to 40 and time point is equal to 101. It seems that there is a decreasing tendency of correlation as the observations are moved further apart in time. This shows the presence of stronger serial correlation among residuals that are at closer time points.

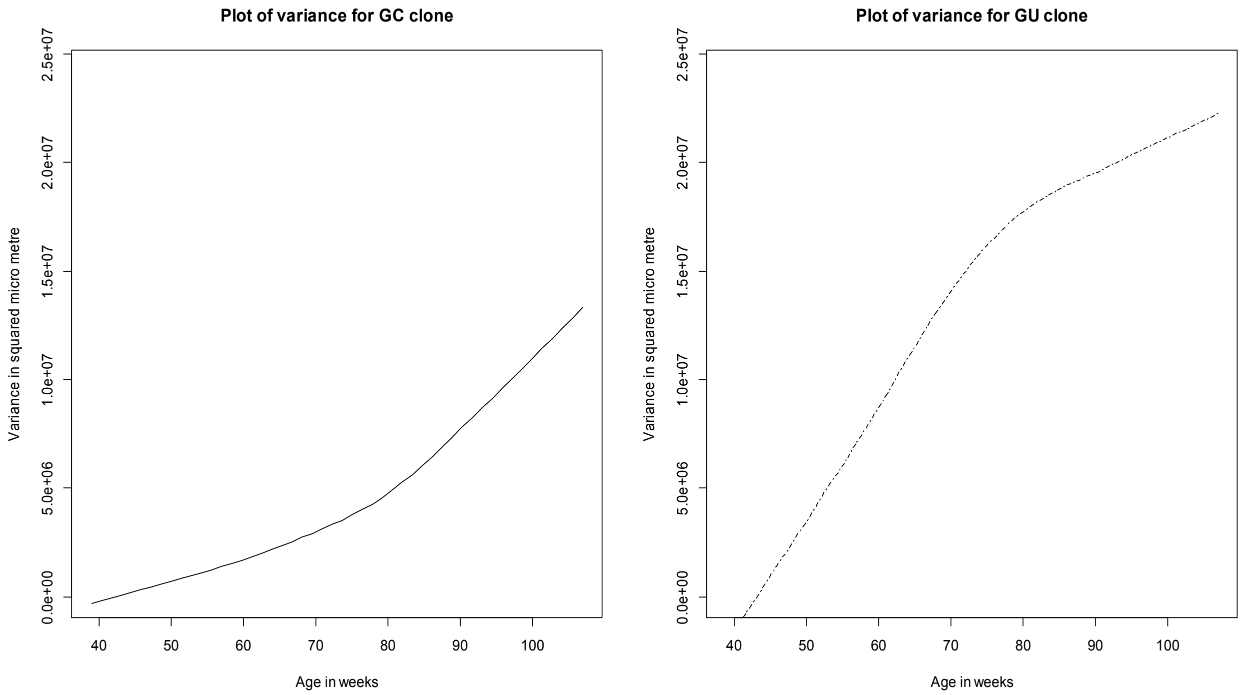

Figure 2. 4 Plot of the variance (square of micro metre) of residuals against tree age in weeks.

Mostly, in regression analysis, the coefficients are considered fixed. Actually it is somewhat useful, primarily because the inference is comparatively easy. Nevertheless, there are cases in which it makes sense to adopt some random coefficients. These cases characteristically happen in two circumstances.

- When the central concern is to make inference on the whole population which some levels are randomly sampled from.
- When the observations are correlated.

In many longitudinal studies, it is sensible to assume that correlations exist among the observations from the same individual or entity. Fixed effects are parameters associated with an entire population or with certain repeatable levels of experimental factors, while random effects are associated with individual experimental units drawn at random from the population.

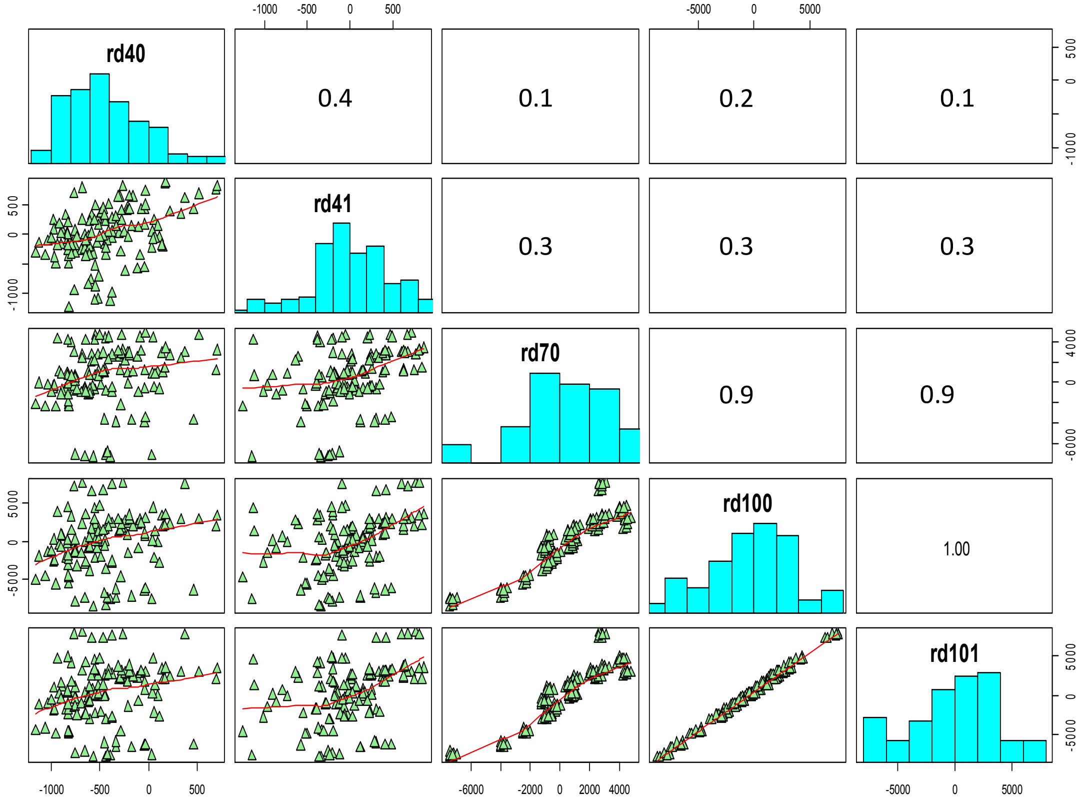

Figure 2.5 Scatter plot and correlation matrices of residuals for selected time points

In the following examples an attempt to clarify the importance of incorporating random effects in the model was made.

The ordinary least square (OLS) regression model of stem radius on tree age and square root of age was fitted and the residuals were examined. The box plots of these residuals by tree are indicated in Figure 2.6. The residuals corresponding to the same tree tend to have the same sign. This indicates the demand for a “tree effect” in the model, which is indeed the motivation for mixed effects models.

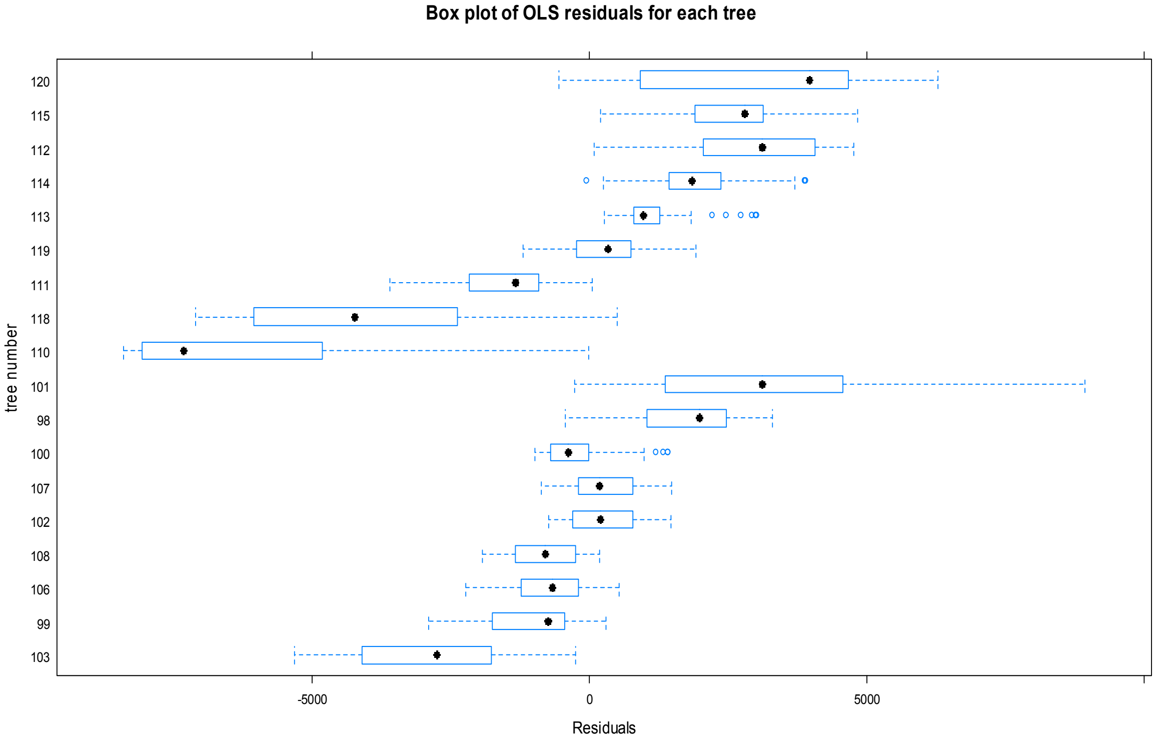

Figure 2. 6 Box plot of OLS residuals by tree for both clones

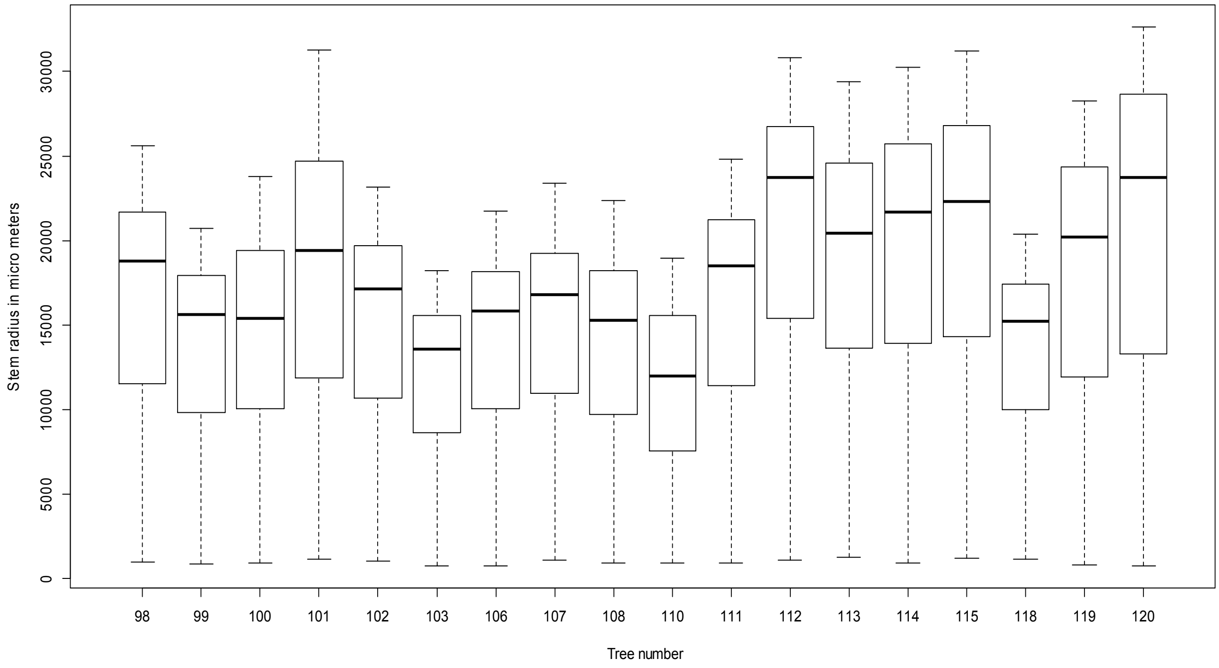

Figure 2.7 Box plot of stem radius expressed in micro metres for 18 trees.

Box plots of the stem radius with respect to each tree (the tree numbers are given during the experiment) are presented in Figure 2.7. It is evident that there is some variability in mean stem radius for different trees. The between tree variability is clearly seen from this plot. Moreover, the within tree variability is not the same for all trees. The modelling process needs to take into account all of the information obtained during the visualization process.

For balanced longitudinal data, the correlation structure can be studied through the correlation matrix, or a scatter plot matrix. In our case, we considered the weekly radial measure for some weeks to see how the correlations among repeated measurements of the data behave. The stem radial measures for weeks 39, 40, 41, 60, 70, 100, 101 and 102 were considered. The estimated correlation matrix for these selected time points is presented as follows.

$$\begin{bmatrix} 1 & 0.90 & 0.83 & 0.33 & 0.24 & 031 & 0.31 & 0.30 \\ 0.90 & 1 & 0.97 & 0.54 & 0.45 & 0.50 & 0.50 & 0.49 \\ 0.83 & 0.97 & 1 & 0.61 & 0.50 & 0.54 & 0.54 & 0.54 \\ 0.33 & 0.54 & 0.61 & 1 & 0.98 & 0.91 & 0.91 & 0.92 \\ 0.24 & 0.45 & 0.50 & 0.98 & 1 & 0.93 & 0.93 & 0.93 \\ 0.31 & 0.50 & 0.54 & 0.91 & 0.93 & 1 & 0.99 & 0.99 \\ 0.31 & 0.50 & 0.54 & 0.91 & 0.93 & 0.99 & 1 & 0.99 \\ 0.31 & 0.49 & 0.54 & 0.92 & 0.93 & 0.99 & 0.99 & 1 \end{bmatrix}$$

The correlation between measurements at week 39 and week 40 is 0.9 indicating a strong relationship between the measurements of week 39 and week 40. On the other hand the correlation between the measurements of week 39 and week 102 is only 0.3. This shows that there is a strong correlation between measurements that are at closer time points to each other. The correlation is dying as the length of time between two measurements increases.

2.3 Summary

The exploratory analyses suggest that the stem radial growth is increasing over time. However, the rate at which it is increasing is different for the two clones. Moreover, the exploration of the covariance structure shows that there is a clear indication for the between tree and within tree variability. It was also established that the stem radius data is balanced and free from the problem of dropout. This paves the way for justifiability of likelihood based analysis. Since commonly used longitudinal methods for continuous response are either the extension of linear regression or nonlinear models, it would have been logical to start with the discussion of linear models. However, from the data at hand all our covariates are correlated. The assumption for multiple regression approach failed. Therefore, a review of methods that overcome the problem of multicollinearity is provided in the next chapter. The next chapter mainly focuses on the use of latent variables in the modelling process. This may help to facilitate comparison of results obtained by different approaches.

Chapter 3

Principal Components and Partial Least Squares Approaches

3.1 Introduction

The simplest approach to being able to detect climatic effects (should they exist) is by the use of traditional regression or correlation methods. However, the effect measured from such approaches assumes the climatic variables are uncorrelated. This chapter therefore addresses several issues and questions. The primary question concerns the extent to which classical regression approaches are successful in detecting and estimating the effect of climatic conditions on stem radial growth. A second aim is to present latent variable modelling approaches, namely partial least squares and principal component regression, for better estimation and detection of the effects of climatic variables.

Principal component regression (PCR) and partial least square regressions are multivariate statistical techniques that have been applied to different sciences to obtain calibration models as an alternative to linear regression. These statistical methods have provided good predictive models for the simultaneous analysis of correlated ecological, pharmaceutical and other formulations (see for example, Rodriguez-Nogales, 2006; Dine et al., 2002; Fekedulegn et al., 2002; and Maitra and Yan, 2008).

3.2 Principal Component Regressions

Principal component analysis (PCA) is a multivariate method commonly used to reduce the number of predictive variables. By producing uncorrelated linear combinations of the predictive variables, it solves the multicollinearity problem. Principal component analysis considers a few uncorrelated linear combinations of the variables that can be used to summarize the data without losing much information in the data.

Let $\mathbf{X}_{nxp}$ denote the data matrix of explanatory variables, where each row denotes an observation on p different explanatory variables, $X_1 \; .. \; . \; X_p$. The problem at hand is to select a subset of the above columns that holds most of the information. Principal component analysis attempts to arrive at suitable standardized linear combinations (SLC) of the data matrix X based on Jordan decomposition of the variance covariance matrix, $\sum$ of X or equivalently based on the correlation matrix, Φ of X. The mean of the observations is denoted by μ_{1xp} . Let $X_{1xp} = (x_1, \; ... \; x_p)$ denote a random vector of observations in the data-matrix (i.e. any row of the n x p data matrix), with mean μ_{1xp} and covariance matrix $\sum$. A principal component is a transformation of x to w of the form

$$\mathbf{w}_{1xp} = (X - \mu)_{1xp} \, \Gamma_{pxp} \; ,$$

where Γ is obtained from the Jordan decomposition of $\sum$, i.e., $\Gamma' \sum \Gamma = \Lambda = diag(\lambda_{1,} \; \lambda_2 ... \lambda_p)$ with $\lambda_i s$ being the eigen values of the decomposition. Each element of $\mathbf{w}_{1xp}$ is a linear combination of the elements of X_{1xp}. Also each element of **w** is independent of the other. Thus, we obtain p independent principal components corresponding to the p eigen values of the Jordan decomposition of $\sum$.

Generally, only the first few principal components for a regression will be used. The principal component **w** has the following important properties. The mean for **wi** is zero and the variance for **wi** is λ_i. The covariance between any two principal components is zero, which shows the principal components are uncorrelated. The first principal component has the largest eigen value or variance, which is equal to λ_1 and no subsequent principal component has variance greater than λ_1 . Principal components capture the maximum of the variance of X and there is no standardized linear combination that can capture maximum variance without being one of the principal components. In the presence of a high degree of correlation among the original predictor variables, only the first few principal components are likely to capture the majority of the variance of the original predictor variables. The size of λ_i s provides the measure of variance captured by the principal components and employed to select the first few components for regression. After eliminating the least important components, the response variable is regressed on the reduced set of principal components using ordinary least squares regression (OLS). As the principal components are orthogonal, they are pair–wise independent and hence the OLS method is suitable. Once the regression coefficients for the condensed set of orthogonal variables have been obtained, they are transformed into a new set of coefficients that correspond to the actual or initial correlated set of variables. This transformation is briefly discussed as follows. In the context of multiple regression model of the form $Y = XB + \varepsilon$, the estimate of B is given by $\hat{B} = (X'X)^{-1} X'Y$. B is the regression coefficient for the original set of predictors. In PCR, the X matrix is decomposed into matices of orthogonal scores(T) and loadings (P) such that $X = TP$. After this decomposition, PCR regress Y on the first '*a*' columns of the scores T. Let us consider the model of Y on the scores (T) is given by $Y=Tb+\varepsilon$, where b is the vector of regression coefficients when the principal components are used as predictor variables. From the equation $X = TP$, we can get $T = XP'$. Therefore, using the relationships in the models $Y=Tb+\varepsilon$ and $Y = XB + \varepsilon$, we get $Y = XP'b + \varepsilon = XB + \varepsilon$. This last equation clearly shows that $B = P'b = P'(T'T)^{-1} T'Y$.

Principal components technique arrives at uncorrelated standardized linear combinations (SLCs), that capture only the characteristics of the X-vector or predictive variables. No significance is given as to how each predictive variable is related to the response variable. That means PCR creates components to explain the observed variability in the predictor variables (X-variance), without considering how they are related to the response variable at all. In a way it is an unsupervised dimension reduction technique (Maitra and Yan, 2008). When our key area of application is multivariate regression, there may be considerable improvements if we build SLCs of predictive variables to capture as much information in the raw predictive variables as well as in the relation between the predictive and target variables.

3.3 Partial Least Squares Approach

Partial least squares (PLS) allow us to achieve this balance and provide an alternate approach to the PCA technique (Maitra and Yan, 2008). Partial least squares is a variance based (component based) statistical method, which is often referred to as structural equation modeling (SEM). It was designed to replace multiple regression approach when the sample size is small and there is problem of multicollinearity or missing values. A comprehensive overview of this technique is given by Haenlein and Kaplan (2004).

Assume $\mathbf{X}$ is a n×p matrix and $\mathbf{Y}$ is a n×q matrix. The PLS technique works by sequentially extracting factors from both X and Ysuch that covariance between the extracted factors is maximized. That means PLS attempts to find a linear decomposition of both X and Y as described in the next paragraph. The PLS method can work with multivariate response variables (i.e. when $\mathbf{Y}$ is an $n \times q$ vector with (q >1). However, in the present study the response variable, $\mathbf{Y}$, is an n×1 vector.

Partial least squares tries to find a linear decomposition of X , and a linear decomposition of Y, $Y=UQ'+F$ such that the covariance between T and U is maximum. T and $Q$are called the scores or factors. There are multiple algorithms available to extract the scores. Each extracted score of X is of the form $t = X\mathbf{e}_1$ where $\mathbf{e}_1$ is the eigen vector corresponding to the first eigen value of $X'YY''X$. Similarly, the first extracted score of Y is $u = Y\mathbf{d}_1$, where $\mathbf{d}_1$ is the eigen vector corresponding to the first eigen value of $\mathbf{Y}'\mathbf{X}\ \mathbf{X}'\mathbf{Y}$. Once the first factors have been extracted we deflate the original values of X and Y as,

$$X_1 = X - tt'X \qquad and \qquad Y_1 = Y - u\,u'Y$$

The above process is then repeated with X_1 and Y_1 replacing X and Y respectively to extract the second partial least squares component. The process continues until all possible latent factors t and u have been extracted, i.e., when X is reduced to a null matrix. The number of latent factors extracted depends on the rank of$X$. It is known that linear regression achieves maximum correlation between the response Y and the explanatory variable X. Principal component regression captures maximum variance in X (X-variance) . Partial least squares regression tries to achieve both (maximum X-variance and maximum correlation) by maximizing the covariance between X and Y.

In the context PLSR, we have two sets of scores the X-scores matrix which is denoted by (T) and the Y scores matrix which is denoted by U. The Y score is not necessary to fit the regression model. Let $Y = XB + \varepsilon$, be the model with orginal set of predictor variables. The estimate of B is given by $\hat{B} = (X'X)^{-1}X'Y$. In PLSR, we use the model $Y = T\beta + \varepsilon$, where β is the regression coefficient when the columns of (T) are used as predictors. The relationship between matrix X and the scores matrix T is given by $T = XR$, where R is the matrix representing the weights in such a way that all coulmns of T relates to the original X matrix that is used before decomposition (Mevik and Wehrens, 2007).

Equating the models $Y = XB + \varepsilon$ and $Y = T\beta + \varepsilon$ and using the relationship $T = XR$, we have $Y = XB + \varepsilon = T\beta + \varepsilon = XR\beta + \varepsilon$. This equation shows that $\hat{B} = R\hat{\beta} = R(T'T)^{-1}T'Y$.

The predictive power of the models can be compared using root mean square error of prediction (RMSEP) and root mean square error of cross validation (RMSECV). To define RMSEP first we define mean square error of prediction (MSEP). MESP measures the squared difference between what the predictors predict for a particular value and the true value. Let y_i be the true value in the data and let $\hat{y}_i$ be the value predicted value by the model under consideration, the the MSEP is given by

$MSEP = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}$. The corresponding root mean square error of prediction is given by

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N}}$$

Regarding the cross validation of the models the leave one out cross validation approach is used in this thesis. For a data set with N samples, leave one out procedures fits model to $(N - 1)$ samples by leaving one sample for validation. The root mean square error cross validation (RMSECV) is described as follows.

Let y_i be the validation sample and let $\hat{y}_i$ be prediction of y_i based on the $(N - 1)$ remaining sample. The prediction error sum of squares (PRESS) is given by

$$PRESS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

The mean square error of cross validation (MSECV) is given by $MSECV = \frac{PRESS}{N}$. The corresponding root mean square error of cross validation is given by

$$RMSECV = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}$$

A model with smaller RMSECV is preferable.

3.4 Data Analysis

Statistical analysis was undertaken using R-statistical software. R is free software that can be downloaded from the R-project website R Core Team (2012). The simplest approach in detecting climatic effects is by the use of traditional regression methods. However, this traditional method assumes that the climatic variables are uncorrelated since one of the failures of regression methods is due to multi-collinearity. The problem of multi-collinearity arises when the predictors (in our case the climatic variables) are correlated. To overcome this, we applied principal component regression and partial least squares regression on daily measurement data. These methods were applied to the combined data set as well as to the data set for separate clones. Extensive discussions of these methods can also be found in Rodriguez-Nogales (2006); Dine et al. (2002); Fekedulegn et al. (2002); Maitra and Yan (2008); (Mevik and Cederkvist (2004); and Haenlein and Kaplan (2004).

3.5 Results of Fitting PCR and PLS Regressions

The variables included in the study are major climatic variables and one non-climatic variable (tree age) as described in Chapter II. The overall ordinary least squares (OLS) model was significant with an adjusted $R^2 = 0.79$ (Table 3.1). This indicates about 79% of the variation in stem radius is explained by the predictors (the five weather variables together with age of a tree) included in the model. An attempt to explore lags was made by considering lags up to 15 days. The use of five weather variables lagged by 15 days increased the variance explained by 0.3% only. Therefore, we did not consider the lags as an important issue at this age of the tree.

Table 3. 1 Summary of ordinary least square model

Predictors (climatic variables)	Estimate	Standard error	t-value	p-value
Intercept	-16558.67	550.61	-30.07	0.000
Temperature	23.73	12.65	1.88	0.061
Solar radiation	2865.35	222.01	12.91	0.000
Rainfall	2.57	6.21	0.41	0.679
Wind speed	1426.83	77.02	18.53	0.000
Tree age	313.22	2.21	142.05	0.000
$R^2 = 0.791$		$Adj \quad R^2 = 0.79$		

Table 3. 2 Correlation matrix of predictors

Variables	Temperature	Relative Humidity	Solar Radiation	Rainfall
Temperature	1			
Relative Humidity	-0.320(**)	1		
Solar Radiation	0.617(**)	-0.498(**)	1	
Rainfall	-0.107(**)	0.272(**)	-0.258(**)	1
Wind Speed	0.406(**)	-0.385(**)	0.374(**)	0.099(**)

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

Table 3. 3 The eigen value decompostion of the correlation matrix

Eigen values	Proportion of total	Cumulative proportion of total
2.375	0.396	0.396
1.252	0.209	0.605
1.083	0.181	0.786
0.625	0.104	0.890
0.412	0.069	0.959
0.253	0.042	1

The predictors included in the model are therefore important for determining radial tree growth. However, the individual t-ratios (estimated coefficient/standard error) for the coefficients of the most important climatic variables, that of rainfall and temperature, are non-significant (Table 3.1). This is an indication of the presence of multicollinearity among the predictors. From the correlation matrix of predictors (Table 3.2), temperature and solar radiation were highly correlated. The correlation coefficient was 0.62 and highly significant ($p < 0.001$). The correlation between wind speed and temperature was 0.41, which was also highly significant ($p < 0.001$). This shows the existence of significant multicollinearity among the explanatory climatic variables. Multicollinearity inflates the standard error of the regression coefficients, which results in low t-statistic values and a failure to reject the null hypothesis. The application of classical linear regression models therefore does not have a powerful inference on the regression coefficients. To address this problem, principal component regression and partial least square regression techniques were used. All predictors were treated as continuous variables with different units of measurements (for instance, rainfall in mm and temperature in °C). It might make more sense to standardize the predictors before trying principal components. This is equivalent to performing principal components analysis on the correlation matrix of predictor variables.

Table 3.3 shows the eigen value decomposition of the correlation matrix of the original or the covariance matrix of the standardized predictors. The first five principal components captured 95.9 % of the information in the correlation matrix. Table 3.4 shows the eigen vectors corresponding to each of the eigen values of Table 3.3. We constructed the principal components corresponding to each eigen value by linearly combining the standardized predictive variables using the corresponding eigen vector. Hence, the six principal components are computed as shown below.

$$PC1 = 0.49\ Z_1\ - 0.49\ Z_2\ + 0.55\ Z_3 - 0.21\ Z_4 + 0.41 Z_5 - 0.07\ Z_6$$
$$PC2 = -0.24 Z_1 - 0.42\, Z_2 - 0.14\, Z_3 - 0.26\, Z_4 - 0.28\, Z_5 - 0.77\, Z_6$$
$$\vdots$$
$$PC6 = 0.47 Z_1 - 0.59\, Z_2 - 0.54 Z_3 + 0.13\, Z_4 - 0.28\, Z_5 + 0.38\, Z_6$$

Where:

- Z1 is the standardized value of temperature
- Z2 is the standardized value of relative humidity
- Z3 is the standardized value of solar radiation
- Z4 is the standardized value of rainfall
- Z5 is the standardized value of wind speed
- Z6 is the standardized value of age

Table 3. 4 The eigen vectors associated with eigen values of Table 3.3

Eigen Vector 1	Eigen Vector 2	Eigen Vector 3	Eigen Vector 4	Eigen Vector 5	Eigen Vector 6
0.495	-0.239	-0.031	0.601	-0.463	0.347
-0.488	-0.415	0.085	0.301	-0.362	-0.593
0.546	-0.144	0.168	0.238	0.553	-0.539
-0.207	-0.255	-0.808	0.259	0.396	0.127
0.413	-0.280	-0.431	-0.594	-0.366	-0.279
-0.068	-0.774	0.354	-0.266	0.241	0.378

The principal components constructed above were used in a linear regression model. Stem radius was used as the dependent variable and the principal components as independent variables (Table 3.5). The rank of the predictive power did not line up with the order of the principal components. For instance, the first principal component was fewer explanatories (larger p-value) for the target than the second or the third, even though the first principal component contains more information on the six original explanatory variables. The principal components technique arrives at uncorrelated standardized linear combinations (SLCs) that capture only the characteristics of the X-vector or predictive variables. No significance is given as to how each predictive variable is related to the response variable. In a way, it is an unsupervised dimension reduction technique (Maitra and Yan, 2008) and therefore requires use of other analytical methods such as partial least squares.

Table 3.5 Summary of OLS model that uses principal components as predictors

Coefficients	Estimates	Std error	t-value	p-value
Intercept	16025.71	36.70	439.659	<2e-16 ***
PC1	60.83	23.82	-2.554	0.0107*
PC 2	-5402.82	32.80	-164.713	<2e-16 ***
PC 3	1987.07	35.27	56.34	<2e-16 ***
PC 4	-1742.90	46.42	-35.547	<2e-16 ***
PC 5	1330.27	57.18	-23.263	<2e-16 ***
PC 6	1425.38	72.99	19.530	<2e-16 ***

* shows significance at the 0.05 level; *** shows significance at the 0.001 level.

In comparing the importance of the constructed principal components, five components explained most of the variation in the predictors (95.9 %). The scree plot (not shown here) showed that almost all the variation in predictors (about 96%) was explained by the first five principal components. Therefore, a linear model that used the first five principal components as latent explanatory variables was fitted (Table3.6). The R^2 value 0.78 for the reduced model was close to the R^2 (0.79) value for the model with all six components. Once again, the rank of the predictive power did not correspond with the order of the principal components. In other words, principal component one appears to have less explanatory power (larger p-value) for the dependent variable as compared to other components. By transforming the principal components back to the original explanatory variables, the estimated coefficients of the original variables are given (Table 3. 7). That means first the principal components are obtained. These principal components are uncorrelated and an ordinary regression model was fitted using the principal components as explanatory variables. The five principal components appear to have significant effect on the radial measure (Table 3.6). The estimated coefficients for the original measured variables were obtained by transformation from the estimated coefficients for principal components. The estimated regression coefficients in Table 3.7 show that all predictors have a positive relationship with stem radial measure. Moreover, the five latent variables that produced the estimated coefficients are significant (see Table 3.6). This indicates the significant effect of climatic variables on stem radial measure. Separate estimates for GU and GC clones also show the positive effect of weather variables together with tree age (Table 3. 7).

Partial least square regression (PLS) can overcome the deficiencies of OLS regression in the case of highly collinear data. Moreover, partial least squares allow an analysis of the data in terms of independent latent variables or components. Applying PLS method to the data, the minimum root mean square error of prediction (RMSEP) is observed for five components model. The value of the X-variance for the model with five latent variables is 93.5 %. This means a model with five latent variables has explained 93.5 % of the variation in the original predictors. The variation explained in the response variable is 79.1 %. This is the same amount of variation explained by the ordinary least square regression. Therefore, the model formulated by five latent variables fits the data well with a high predictive power. The coefficients for the original set of variables when partial least square regression was applied to GC, GU and pooled data sets are indicated in Table 3. 8. It appears that the estimated coefficients for the original set of variables for the GC clone are smaller than that of the GU clone for all climatic variables. This indicates that the GU clone has on average a larger stem radius than the GC clone. The signs of the estimated coefficients for the GU clone and the signs for the estimated coefficients of the pooled data set are the same. However, the estimated coefficient of temperature is negative for the GC clone while it is positive for the GU clone and pooled data set. This indicates that the effect of temperature on stem radius goes in opposite directions for the two clones for this site and age class. The possible reason for this could be the difference in genetic makeup of the two clones. Moreover, the effect of weather variables may depend on the season of the year. The site difference cannot be a possible reason for this difference as site difference is controlled by the design. In the design the plots were established as pairs such that a GU and a GC plots are measured simultaneously (Fig. 2.1). For the rest of the climatic variables the effect follows the same direction for the two clones with some differences in magnitude.

In order to test whether the components that produced these coefficients are significant or not, latent variables or partial least square components were constructed while fitting the partial least square regression. After determining these latent variables, T1…T6 sequentially, the relationship between these latent constructs and the response was estimated by ordinary linear regression. The sample correlations between any pair of the latent constructs were zero. A linear model was then applied using the same radial measure as the dependent variable and the six partial least square components, T1…T6, as the independent variables.

A summary result for the model that uses the partial least square components as predictors is shown in Table 3. 9. The partial least square components were extracted in order of significance. The first five components were significant while the last component was not. The values of R^2 and adjusted R^2 for this model were 0.7908 and 0.7907 respectively.

Table 3. 6 Summary of OLS results for the model that uses the first five principal components

Coefficients	Estimates	Std error	t-value	p-value
Intercept	16025.71	37.50	427.35	<2e-16 ***
PC1	60.83	24.34	-2.50	0.0124
PC 2	-5402.82	33.52	-161.20	<2e-16 ***
PC 3	1987.07	36.04	55.14	<2e-16 ***
PC 4	-1742.90	47.43	-36.75	<2e-16 ***
PC 5	1330.27	58.43	-22.77	<2e-16 ***

* shows significance at the 0.05 level; *** shows significance at the 0.001 level.

Table 3. 7 The estimated coefficients of the original variables estimated using principal component regression

Predictors (Climatic variables)	Estimates for combined data	Estimates for GU clone	Estimates for GC clone
Intercept	-16558.67	-19048.26	-14069.07
Temperature	90.48	165.33	15.64
Relative humidity	581.14	680.05	482.29
Solar radiation	694.56	802.99	586.20
Rainfall	16.81	27.82	5.79
Wind speed	834.13	902.12	766.24
Tree age	6201.39	6764.65	5638.85

Table 3. 8 Estimated coefficients of the orginal set of climatic variables using the partial least squares method

Climatic variables	Estimates for both clones	Estimates for GU clone	Estimates for GC clone
Temperature	55.42	128.02	54.42
Relative humidity	596.58	696.94	596.58
Solar radiation	761.13	874.50	761.13
Rainfall	35.13	47.59	35.13
Wind speed	814.29	880.65	814.29
Tree age	6191.69	6754	6191.69

Table 3.10 shows the summary results for the model that involves only five partial least square components. From the results, it can be seen that all the coefficients listed in Tables 3.9 and 3.10 were the same for the first five components. This shows that the coefficients of the partial least square latent variables do not change by adding or dropping latent variables from the model. The results of the partial least squares model show that jointly all climatic variables had a significant effect on growth.

Table 3. 9 Summary of OLS results for the model that uses the PLS components as predictors

Coefficients	Estimates	Std error	t-value	p-value
Intercept	16025.71	36.70	436.64	<2e-16 ***
T1	5932.81	32.29	178.23	<2e-16 ***
T2	1193.6	45.41	26.28	<2e-16 ***
T3	318.38	30.45	10.46	<2e-16 ***
T4	299.85	40.22	7.46	9.83e-14 ***
T5	212.74	48.99	4.34	1.42e-05***
T6	78.66	58.87	1.336	0.182

*** shows significance at the 0.001 level.

Table 3. 10 Summary of OLS results for the model that uses the first five PLS components as predictors.

Coefficients	Estimates	Std error	t-value	p-value
Intercept	16025.71	36.70	436.64	<2e-16 ***
T1	5932.81	32.29	178.23	<2e-16 ***
T2	1193.6	45.41	26.28	<2e-16 ***
T3	318.38	30.45	10.46	<2e-16 ***
T4	299.85	40.22	7.46	9.83e-14 ***
T5	212.74	48.99	4.34	1.42e-05***

*** shows significance at the 0.001 level.

Table 3. 11 RMSE and RMSECV values for all prediction methods

	OLS	PCR	PLS
RMSE	3410.01	3484.53	3410.4
RMSECV	3414.39	3413	3413

With regard to the predictive powers of these models, a comparison was made based on root mean square error (RMSE) and the root mean square error of cross-validation (RMSECV, Table 3.11), a measure of the model's ability to predict new samples. The ordinary least square model had the smallest RMSE value (Table 3.11). The second smallest RMSE value belonged to the partial least square model. The RMSE for partial least square was actually very close to the RMSE for the ordinary least square model. However, this comparison was from the point of view of model fit. Under the condition of no multicollinearity, this might indicate that the ordinary least square model fitted the data better than the other two methods. For comparisons of models intended for prediction, it is inadequate to look just at model fit. The RMSECV obtained for PCR model with six components is the same as the RMSECV obtained for partial least square regression model with five components. As prediction is the objective, the partial least square and the PCR models that gave the lowest RMSECV value with smaller number of components is preferred. For the data set to which these models were applied, the partial least square model had the highest predictive ability with the lowest number of factors. In order to identify differences between clones, a separate partial least square model was fitted to data for each clone. For both clones, the optimum number of partial least square components was five. These five components were significant while the sixth component was not significant (Table 3.9). The percentage of total variation in radial measure captured by the optimal number of components for the GU clone is less (Table 3.12: 80% with p-value < 0.0001) than the amount of variation captured for the GC clone (Table 3.13: 87.21% with p-value < 0.0001). The percentage of total variation in climatic variables and tree age captured by the five components partial least squares model for the GU and GC clones is almost the same (93.5 %).

In order to determine the most important drivers of variation in short term stem radial measure (for the first two years of tree age) for the two clones, we applied standardized regression weights for both partial least squares and principal component regressions. This can be obtained by fitting the models on standardized variables. The factor with the highest coefficient in absolute value is the most important factor in explaining the variation in radial measure. The standardized regression weights (coefficients) for our predictors, when partial least square regression and principal components regression were applied to GC and GU data sets, are indicated in Table 3.14.

Table 3. 12 Percent of variance captured by partial least square components for GU clone

Components	Climatic variables and age		Radius	
	This component	Cumulative Total	This component	Cumulative Total
T1	20.53	20.53	77.53	77.53
T2	17.66	38.19	1.86	79.04
T3	30.25	68.44	0.35	79.39
T4	15.27	83.71	0.14	79.53
T5	9.8	93.51	0.04	79.57

Table 3. 13 Percent of variance captured by partial least square components for GC clone.

Components	Climatic variables and age		Radius	
	This component	Cumulative Total	This component	Cumulative Total
T1	20.47	20.47	84.74	84.74
T2	12.25	32.72	2.06	86.80
T3	25.85	58.57	0.25	87.05
T4	24.28	82.85	0.11	87.16
T5	10.68	93.53	0.05	87.21

It appears that tree age is the most important predictor of stem radius using both models and for both clones. Among climatic variables, it appears that wind speed, followed by solar radiation, is the most important driver of the variation in stem radius over the growth period of two years. However, the biological plausibility of these results is questionable. Moreover, we found the negative effect of temperature for GC clone. This might be due to the dependence of weather variables on season. The weather variables are likely to change over the year.

Table 3. 14 Table of standardized regression weights for both principal component regression and partial least square regression models

Predictors	PLS model		PCR model	
(climatic variables)	GU	GC	GU	GC
Temperature	0.016	-0.003	0.020	0.002
Relative humidity	0.086	0.078	0.083	0.075
Solar radiation	0.107	0.101	0.098	0.091
Rainfall	0.006	0.004	0.003	0.001
Wind speed	0.108	0.116	0.110	0.119
Tree age	0.829	0.876	0.830	0.878

This relative effect of weather variable might change from one season to the other. We analysed the same data by season in order to see the season effect. Summary results by season are shown in Table 3.15 and Table 3.16. In spring and summer, none of the weather variables has significant effect. The only variable that has significant effect on stem radius is tree age. In winter, all predictors have a significant effect on stem radius for GU clone while for GC clone all have a significant effect with the exception of rainfall. In autumn, solar radiation, wind speed and tree age have significant effects on the stem radius for both clones. In autumn, rainfall appears to have a significant effect on stem radius for GU clone while it has no significant effect on GC clone. The insignificant effect of rainfall in winter and autumn for GC clone might be due to a genetic factor, which needs further study. Temperature has a significant effect and is positively related to stem radius in winter for both clones (Table 3.16). In summer, autumn and spring, temperature has no significant effect on stem radius (Table 3.15 and 3.16). Therefore, the effect of weather variables on stem radius is dependent on the season.

Table 3. 15 Summary results of ordinary regression model for summer and autumn

Predictors	Summer			
	GC clone		GU clone	
	Estimate	p-value	Estimate	p-value
Intercept	2763.099	0.265	2695.785	0.588
Temperature	-2.143	0.963	-17.097	0.854
Relative humidity	5.088	0.781	9.983	0.786
Solar radiation	167.126	0.712	371.769	0.683
Rainfall	0.291	0.990	0.422	0.993
Wind speed	-47.827	0.813	-80.071	0.844
Tree age	185.506	0.000	231.252	0.000
	$R^2 = 0.107$		$R^2 = 0.045$	
Predictors	Autumn			
	GC clone		GU clone	
	Estimate	P-value	Estimate	P-value
Intercept	-11156.222	0.000	15921.22	0.000
Temperature	-12.152	0.578	28.38	0.377
Relative humidity	8.632	0.441	19.62	0.233
Solar radiation	1055.849	0.028	1907.87	0.007
Rainfall	13.029	0.550	23.89	0.029
Wind speed	378.068	0.011	476.58	0.029
Tree age	316.093	0.000	382.49	0.000
	$R^2 = 0.929$		$R^2 = 0.9$	

Table 3. 16 Summary results of ordinary regression model for winter and spring

Predictors	Winter			
	GC clone		GU clone	
	Estimate	p-value	Estimate	p-value
Intercept	-12364.279	0.000	-14159	0.000
Temperature	137.832	0.000	159.339	0.000
Relative humidity	39.106	0.000	46.699	0.000
Solar radiation	1980.674	0.000	1775.888	0.021
Rainfall	-5.541	0.442	-7.936	0.046
Wind speed	659.705	0.000	698.642	0.002
Tree age	266.982	0.000	312.839	0.000
	$R^2 = 0.896$		$R^2 = 0.841$	
Predictors	Spring			
	GC clone		GU clone	
	Estimate	P-value	Estimate	P-value
Intercept	-2217.472	0.077	-8561.296	0.002
Temperature	-20.944	0.366	-40.28	0.434
Relative humidity	-0.688	0.941	-2.816	0.893
Solar radiation	56.458	0.855	110.533	0.872
Rainfall	-1.488	0.870	-1.53	0.939
Wind speed	31.297	0.788	65.365	0.801
Tree age	262.869	0.000	403.825	0.000
	$R^2 = 0.282$		$R^2 = 0.158$	

Daily stem size variation is important as the net increment of a forest stand is ultimately determined by the accumulation of daily increment events (Drew et al., 2009). Several factors might affect the daily stem size of trees. For instance, the study by Zweifel et al. (2006) indicates that there is a strong dependence of radial growth on the current tree-water relations and only secondary dependence on the carbon-balance. The availability of soil water and the degree to which storage tissues were saturated were also factors affecting the diurnal course of stem radius changes (Zweifel and Häsler, 2001). Whitehead and Jarvis (1981) and Landsberg (1986) have suggested in theoretical approaches, that the diurnal stem radius fluctuations are coupled to tree-water relations by changing water potential gradients within the tree. Studies by Downs et al. (1999) and Deslauriers et al. (2003) consider the effect of weather on daily stem growth. Deslauriers et al. (2003) studied daily stem radial growth of balsam fir to show that total rainfall and maximum temperature were positively correlated with the stem radius. Climatic variables are highly inter-correlated, and the use of ordinary least squares to estimate the parameters of the response function results in instability and high variability of the regression coefficients. As a result, the regression coefficients become much larger than would seem reasonable physically or practically, and may fluctuate widely in sign and magnitude. Accordingly, it was observed that the ordinary regression estimates inflated the percentage of variation in the stem radial growth accounted for by climatic conditions. Ordinary regression inferences from such correlated climatic variables can result in misleading and confusing conclusions relating to variables of major interest to dendroecologists in terms of magnitude, sign, and standard error of the coefficients as well as R^2 (Fekedulegn et al., 2002).

Both principal component regression and partial least square regression methods have an advantage over ordinary least square regression because they do not require that the explanatory variables be orthogonal. The principal components are orthogonal, eliminating the multicollinearity problem. However, the problem of choosing an optimum subset of predictors remains. A possible strategy is to keep only a few of the first components. Nevertheless, the components are chosen to explain the independent (X) rather than the dependent (Y) and there are no guarantees that the principal components which explain the independent variable can be relevant to explain the dependent (Y). On the other hand, PLS regression finds components from X that are also relevant for Y. Partial least squares regression searches for a set of components that perform a simultaneous decomposition of X and Y with the constraint that these components explain much of the covariance between X and Y. The partial least squares approach is considered as a variance-based structural equation model. The alternative structural equation model (SEM) is a covariance-based structural equation model. Although both methods use a latent variable term, the latent variables used by the two methods are different. As indicated by Fornell and Bookstein (1982), the latent variables in partial least squares are estimated as exact linear combinations of their indicators. This shows that "latent" variables in partial least square are not true latent variables as defined in SEM, as they are not derived to explain the co-variation of their indicators, but only to approximate them (Mathes, 1993; McDonald, 1996). On the other hand, the latent variables in covariance-based SEMs are true latent variables. That is they are hypothetically existing entities or constructs. In other words, the covariance-based SEM latent variables cannot be found as weighted sums of manifest variables; they can only be estimated by such weighted sums (Schneeweiss, 1993). Arguably, partial least square has the advantage over the covariance based SEM, in that Jöreskog and Wold (1982) and Wold (1982; 1985) referred to partial least square technique as "soft modelling" because it did not require the "hard" distributional assumptions of maximum likelihood (ML) which is the core technique in SEM, and because it uses a suboptimal estimation technique that is faster to run than ML-SEM, which therefore allows for more user interaction.

Finally, the latent variable model approaches used in our study show that all climatic variables measured and tree age are positively correlated with stem radial measure for the pooled data of both clones. Moreover, all latent variables had significant effects on the radial measure. This was not the case when ordinary least square was applied. The effects of the two most important variables, rainfall and temperature, were not significant when the ordinary least square method was used (Table 3.1). This may be because the ordinary linear regression assumes that the predictors are uncorrelated while in our case the climatic variables are correlated (Table 3. 2). It may also be because the effect of weather variables changes with season. To overcome the problem of correlation among weather variables, two alternative methods (Principal component regression and partial least squares) were used. Principal component regression models were fitted for the GC and GU clones separately, resulting in a positive effect of climatic variables on stem radius for both clones. The weather data together with the age of a tree accounted for 79% of the variance in the stem radial growth for the combined data set. This is equivalent to R^2 in ordinary least square regression. The separate analysis of GC and GU clones showed that the weather variables and tree age explained 87% and 79.6% of the total variation in radial measure for the GC and GU clones respectively.

When comparing the partial least square model fitted for the GC clone and GU clone, the effect of climatic variables is similar for the two clones except for the effect of temperature. Temperature appears to have an opposite effect on the radial growth of the two clones. Moreover, 87% of the total variation in the stem radial measure is explained by the weather variables and tree age by using the PLS method for the GC clone and 79% of the variation is explained for the GU clone. This indicates that the amount of explained variation is larger for the GC clone than for the GU clone. The evaluation of the relationship between weather variables and stem radius is considered after separating the data by season. The effect of weather variables on stem radius was found different for different seasons. Tree age is the most important factor that influences change in stem radius. The importance of tree age in determining stem radius should be expected as growth is positively related to age most of the time. There is no significant effect of weather variables on stem radius during summer and spring for both GU and GC clones. In autumn, there is significant effect of some variables (tree age, solar radiation, wind speed) for both GU and GC clones. In winter, the variables temperature, relative humidity, solar radiation, wind speed and tree age have a significant positive relationship with stem radius for both clones (Table 3.16).

3.6 Summary

The PCR and PLS regression methods provided tools for assessing factors that affect stem radial growth. These statistical methods appear to be good in solving the problem of multicollinearity because they do not require that covariates are orthogonal. Although we intially suspected multicollinearity problems, with regard to the data at hand, it was not very severe. The results revealed that the relationships between the daily stem radius and weather variables is positive for both the GU and GC clones with the exception of temperature. The study indicates that tree age is the most important factor that influences stem radius during the juvenile stage of the tree (up to two years) in all seasons. In winter, temperature, relative humidity and wind speed appear to be more important than the other weather variables. Melesse and Zewotir (2013a) provide a detailed discussion of these results (attached in Appendix A).

The PLS approach is considered as a *variance-based* structural equation model. The alternative structural equation model (SEM) is a covariance-based structural equation model. PLS concentrates on maximizing the variance explained for the dependent variable in the model, whereas covariance-based SEM determines the model parameters required to come up with an empirically observed covariance matrix. PLS is based on least square approach while covariance-based approach is mainly based on maximum likelihood approach. The two latent variable modelling approaches (PCR and PLS) used in this chapter can produce the direct effect of each explanatory variable on the response. However, the indirect effect can only be studied if we consider covariance-based structural equation modelling. The next step is to review the alternative to partial least square approach namely the covariance-based structural equation modelling approach. Specifically, we begin by reviewing path models approach and use them to study the impact of climatic variables and tree age in chapter 4.

Chapter 4

Structural Equation Modelling (SEM)

4.1 Introduction

Structural Equation Modelling (SEM), also known as covariance structure analysis, covariance structure modelling, or causal modelling is a collection of related statistical techniques designed to model complex relationships between characteristics under investigation (Kline, 2005). SEM is one of the cutting-edge statistical techniques that assess a series of multiple dependent relationships simultaneously. SEM provides a chance to employ comprehensive methods for quantification and testing of theories of complex relationships, to explicitly take into account the measurement error, and to use latent variables as a cause and as an outcome. The fundamental hypothesis in SEM is that the covariance matrix of the observed variables is a function of a set of model parameters (Bollen, 1989), that is

$$\sum = \sum(\theta)$$

where

$\sum(sigma)$ = the population covariance matrix of observed variables.

θ is a vector that holds model parameters.

$\sum(\theta)$ = is the covariance matrix written as a function of θ.

Many well-known conventional statistical techniques such as regression analysis, correlation analysis, path modelling and factor analysis can be considered as special cases of SEM. During the development of SEM some basic terminologies have been developed. It is essential to give a review of a few key concepts associated with SEM methodology.

4.2 Basic Concepts Associated with SEM

Latent versus observed variable: In SEM, variables are mainly categorized into observed and unobserved. Observed variables are those variables that are measured directly whereas unobserved or latent variables are those that cannot be measured directly.

Latent variables also known as latent constructs are measured indirectly from multiple observed variables. Observed variables serve as indicators of the underlying construct or latent variables.

Exogenous versus Endogenous variables

Exogenous variables are synonymous with independent variables; they "cause" fluctuations in the values of other latent variables in the model. Changes in the values of exogenous variables are not explained by the model. Rather, they are considered to be influenced by other factors external to the model. Endogenous latent variables are synonymous with dependent variables and, as such, are influenced by the exogenous variables in the model, either directly or indirectly (Byrne, 2001). The values of the exogenous variables are determined outside the model while the values of endogenous variables are determined within the model.

Direct, Indirect and Total effects

Direct effect measures the impact of one variable on another that is not intervened by any other variable. The indirect effect measures the impact of an independent variable through all possible mediating variables. The sum of direct and indirect effect gives us total effect. For instance, consider the hypothetical relationship presented in Figure 4.1. In this figure, the captal letters (X , Y, Z and W) represent the varaibles and the small letters (x, y, z, w and u) stand for path coefficients.

- X and Y are correlated in a non-causal manner (also called unanalysed association to show that the explanations for the observed association are not examined or are not essential to consider within the context of the model).

- X and Y have a direct effect on Z. Similarly, Y and Z have a direct effect on W.
- Z has a mediating role in the relationship of X and W, along that of Y and W.
- There is no direct effect of X on W.
- X and Y have an indirect effect on W individually through Z. (These effects cannot be measured in ordinary regression).

With the two variables X and W from Figure 4.1, there is no direct effect of X on W. Nevertheless, there are three indirect effects and the sum of these two indirect effects will give us the total effects of X on W.

The magnitude of one indirect effect through Z is equal to $y \times u$.

The magnitude of another indirect effect through Y is equal to $x \times w$.

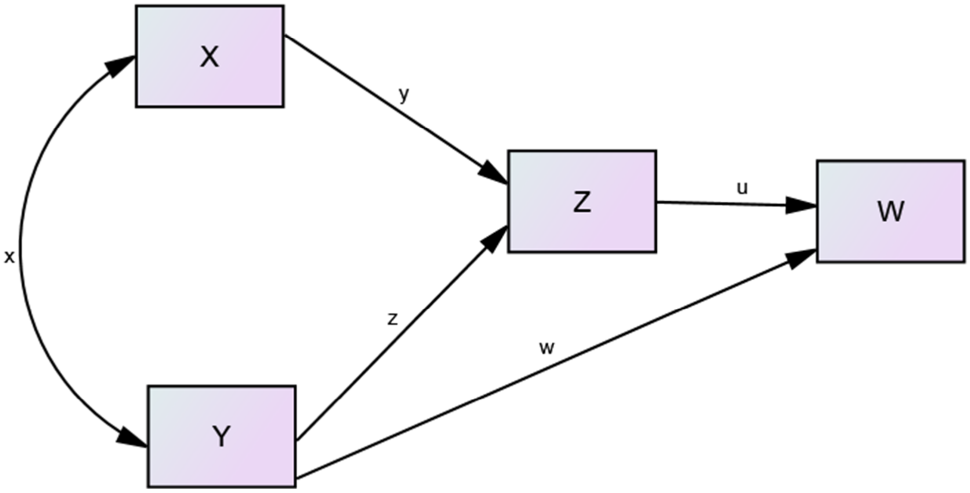

Figure 4. 1 Hypothesized causal model relationships between two exogenous variables (X, Y), one mediating variable (Z) and one outcome variable (W).

The magnitude of another indirect effect through Y and then through Z is equal to $x \times z \times u$.

Recursive and non-recursive models

Based on the manner in which variables are hypothesized to influence each other, one can identify the recursive models from non-recursive models. When one variable cannot influence a variable and at the same time be influenced by that variable in a given causal line, then the model is termed as recursive.

In a non-recursive model, variables possibly influence other variables (be an independent variable) and at the same time be influenced by the same variable (to be a dependent variable) in the same system of relational equations (reversed causality). Figure 4.2 is an example of a non-recursive model while Figure 4.1 can be considered as an example of a recursive model. Differentiating between recursive and non-recursive models has implications on the way the model is fitted to the data (Bollen, 1989).

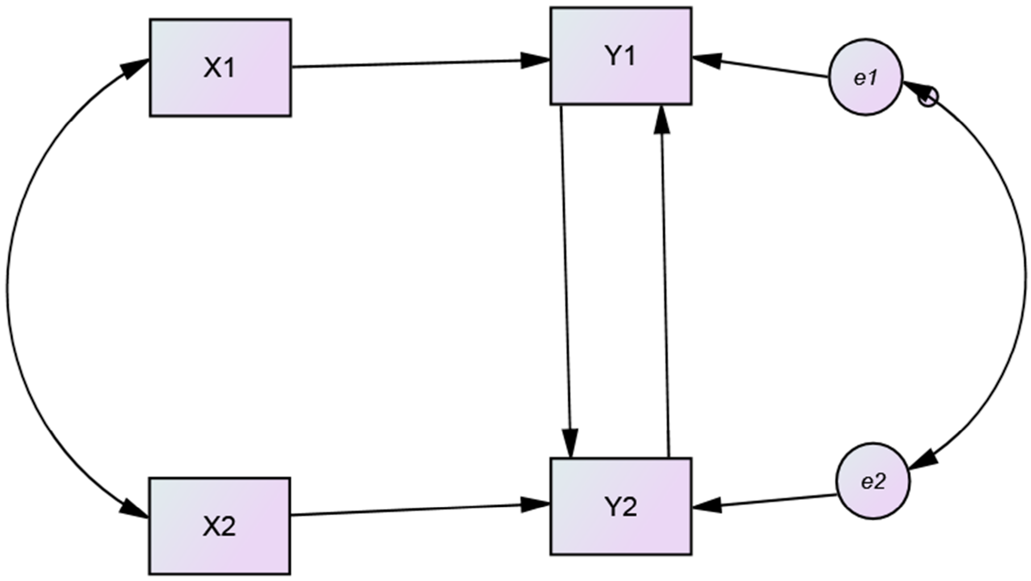

Figure 4. 2 An example of Non-recursive Structural Equation Model

Correlation and covariance:

These are measures of non-directional relationship between two measured variables and they play a pivotal role in SEM. For two continuous variables, the Pearson correlation coefficient (r) is obtained after standardizing the covariance of the two variables under investigation (Bollen, 1989). If we have two continuous variables namely X and Y each observed n times, then the Pearson correlation coefficient (r) is calculated as

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) \Big/ \left(n-1\right)}{\sqrt{\sum_{i=1}^{n}\frac{\left(x_i - \overline{x}\right)^2}{n-1} \times \sum_{i=1}^{n}\frac{\left(y_i - \overline{y}\right)^2}{n-1}}}$$

Hypothesising a cause and effect relationship is not mandatory to model the association of two variables using correlation analysis because the correction coefficient between the variables X and Y is the same as the correlation coefficient between Y and X.

4.3 Types of SEM Models

Structural equation modelling (SEM) has been developed over a long period of time in different disciplines. The direction of development has varied by the type of problem faced in each discipline. For instance, path analysis is first introduced in 1918 by Sewall Wright in his genetic work and it was fully described in the early 1920s (Wright, 1918; Wright, 1920; Wright, 1921; Wright, 1923). Wright (1934) developed the method of path analysis for estimating causal relations among variables based on correlation matrix of observed variables, stressing path coefficients (standardized regression coefficients) but also employing unstandardized coefficients. He also developed a graphical method of presenting causal relations using path diagrams, comprising variable labels connected by arrows for direct effects, double headed arrows for unanalysed correlations, and estimated path coefficients indicated over a single headed arrows.

In psychology, the interest in SEM is initiated in factor analysis. Factor analysis is a statistical technique for analysing a correlation matrix of the observed variables to identify a small number of factors, components, or latent variables that comprise much of the information in the original variables. This technique has two main parts namely, the exploratory factor analysis (EFA) which has been around for more than a century (Lovie and Lovie, 1993) and confirmatory factor analysis (CFA) which has been popularized since the mid-1960s (Brown, 2006). Simultaneous equation models are developed in economics to examine supply and demand.

Publications of different discipline-specific advances came together in the early 1970s and created the multi-disciplinary method currently known as SEM. The general approach to confirmatory maximum likelihood factor analysis by Jöreskog (1969), the work on treatment of unobservable variables in path analysis (Hauser and Goldberger, 1971), and generalized least square (GLS) results on unobservable independent variables (Zellner, 1970) are some of the examples that pave the way for the creation of a multidisciplinary SEM approach. The main emphasis of this chapter is on the applications of path models approach and therefore a discussion of path models is presented.

Path modelling approach

A brief description of path analysis and its relation to the classical regression model is given. Path analysis is the statistical technique used to examine causal relationships between two or more variables. It involves a set of simultaneous regression equations that theoretically establish the relationship among observed variables in the path model. Path analysis extends the idea of regression modelling and gives the flexibility of quantifying indirect and total causal effects in addition to the direct effect which is also possible in regression analysis (Bollen, 1989). In other words, regression analysis allows an independent variable to influence an outcome variable only directly. Path analysis however gives more flexibility and predictor variables are allowed to influence the outcome variable directly as well as indirectly through other mediating variables. Path analysis shares the following principles of regression analysis:

(i) The direction of influence in the relationship of variables should be specified from the theory behind the investigation.

(ii) Independent variables are assumed to be measured without error.

(iii) The relationship between target variables is linear.

(iv) Any outcome variable in the system of equations under investigation has an error term attached to it.

Path analysis is an extension of the regression model, which researchers use to test the fit of a correlation matrix with a causal model that has been, tested (Garson, 2004). The aim of path analysis is to provide estimates of the magnitude and significance of the hypothesized causal connections among sets of variables displayed through the use of path diagrams. There are three interrelated components in path analysis (Bollen, 1989):

(i) The translation of a conceptual problem into pictorial presentation, which shows the network of relationships;

(ii) Obtaining systems of equations that relate observed correlation and covariance to parameters; and

(iii) Decomposition of effects of one variable on another (i.e. direct, indirect and total effects) from the correlation of measured variables.

The pictorial presentation or path diagram assists in clarifying what is meant by a conceptually framed problem and leads to formulation of systems of mathematical equations that can be solved to give estimates of effects knowing the correlation or covariance of measured variables. A path diagram has been promoted as the easiest tool to conceptualize a “causal relationship “as well as to decompose the correlation between different variables into different sources (Wright, 1920; Wright, 1921; Wright, 1923).

4.4 Estimation in SEM

The estimation procedures originate from the relationship of the covariance matrix of the observed variables to the structural parameters. If the structural model is correct and the population parameters are known, then $\sum$ will equal $\sum(\theta)$ (Bollen, 1989). The unknown structural parameters are estimated so that the implied covariance matrix $\hat{\Sigma}$ is close to the sample covariance matrix S . To measure the closeness of the estimate a function must be minimized. The fitting functions $F(S, S(\theta))$ are based on S , the sample covariance matrix, and $\sum(\theta)$, the implied covariance matrix of structural parameters. If the estimates of θ are substituted in $\sum(\theta)$, this leads to the implied covariance matrix, $\hat{\Sigma}$. The value of the fitting function for $\hat{\theta}$ is $F(S, \hat{\Sigma})$. The fitting function has the following properties (Bollen, 1989).

(1) $F(S, \Sigma(\theta))$ is a scalar

(2) $F(S, \Sigma(\theta)) \geq 0$

(3) $F(S, \Sigma(\theta)) = 0 \;\; iff \;\; \sum(\theta) = S$

(4) $F(S, \Sigma(\theta))$ *is continuous in* S *and* $\sum(\theta)$.

Some functions are maximum likelihood (ML), un-weighted least squares (ULS) and generalized least squares (GLS). In this thesis, the maximum likelihood method is used and a brief description of this method is presented as follows.

Maximum Likelihood estimation (MLE):

The fitting maximum likelihood function is

$F_{ML} = Log\left|\sum(\theta)\right| + tr\left(S\,\Sigma^{-1}(\theta)\right) - \log\left|\,S\,\right| - (p+q)$ (4.1) where $(p+q)$ is the total number of observed variables .

$\Sigma(\theta)$ and S are assumed to be positive definite. Both the dependent (Y) and the independents (X) are assumed to have multivariate normal distributions and S has Wishart distribution. The derivation for 4.1 can be found in Bollen (1989). The function (4.1) is a complicated nonlinear function of the structural parameters, and easy solutions are not always available. As a result, iterative numerical procedures are applied to find the solution. For an overview of such numerical procedures one can refer to Bollen (1989). Some important asymptotic properties of likelihood estimators are:

(1) they are asymptotically unbiased;
(2) they are consistent;
(3) they are asymptotically efficient;
(4) the distribution of the estimator approximates a normal distribution as the sample size gets large;
(5) They are scale invariant.

From property (4), if we know the standard error of the estimator, the ratio of the estimator to its standard error can have a standard normal distribution for large sample. This ratio gives us the test statistic commonly known as critical ratio in SEM modelling. It is used to test if the parameter under consideration is significantly different from zero. Property (5) shows that the values of the fit function (4.1) are the same for correlation and covariance matrices, or more generally they are the same for any change of scale.

4.5 Evaluating Model Fit in SEM

The main interest in SEM is to find a meaningful explanation for the association of variables simultaneously. The association can be analysed using raw data or variance covariance matrix that has adequate information about the association. Several statistics have been proposed as a measure of the merit of the model. A focus has been made on a few, mainly based on the recommendation of Browne and Mels (1992), with the availability of the software also being taken into consideration. AMOS software (Amos Development Corporation) is employed to fit models. In AMOS, fit measures are reported for each model specified by the user and two additional models called the saturated model and the independence model.

In a saturated model, no constraints are placed on the population moments. The saturated model is the most general model possible. It is a vacuous model in the sense that it is guaranteed to fit any set of data perfectly. Any Amos model is a constrained version of the saturated model. On the other hand the independence model goes to the opposite extreme. In the independence model, the observed variables are assumed to be uncorrelated with each other. When means are being estimated or constrained, the means of all observed variables are fixed at zero. The independence model is so severely and implausibly constrained that they would expect it to provide a poor fit to any interesting set of data (Arbuckle, 2006).

One of the measures for assessing the goodness of fit of structural equation models is the chi-square statistic. Under the null hypothesis that $H_o : \Sigma = \Sigma(\theta)$, $(N-1)F_{ML}$ has an asymptotic chi-square distribution with degrees of freedom is equal to $\frac{1}{2}(p+q)(p+q+1)-t$, where $F_{ML}$the value of the fitting function defined in equation 4.1 evaluated at the final estimate, t is the number of free parameters and N is the total number of observations. The total number of observed variables is $(p+q)$. This test statistic is used to test the covariance structure hypothesis $H_o : \Sigma = \Sigma(\theta)$. Rejection of the null hypothesis suggests that at least one restriction is in error so that $\Sigma \neq \Sigma(\theta)$. It should be noted that the usage of chi-square statistic depends on a sufficiently large sample and on multivariate normality of the observed variables.

Normed fit index (NFI): The Bentler-Bonett (1980) normed fit index is a measure whose possible value lies between zero and one inclusive. It can be written as

$NFI = 1 - \frac{\hat{C}}{\hat{C}_b}$ where $\hat{C}$ is the minimum discrepancy of the model being evaluated and $\hat{C}_b$ is the discrepancy of the baseline model (Independence model). The NFI tests the hypothesized model against a reasonable baseline model and ideally should be 1·0. An NFI value of 0.9 and higher have been recommended to be used as an indicator of best fitting models.

Root Mean Square Error of Approximation (RMSEA): The RMSEA formula according to Bollen and Curran (2006) is

$$RMSEA = \sqrt{\frac{\chi_k^2 - df_k}{(N - 1) \times df_k}}$$

where

- χ_k^2 is the likelihood chi-square from the target model and df_k is its associated degrees of freedom.

- $\chi_k^2 - df_k$ is an asymptotically unbiased estimator of the non-centrality parameter for non-central chi-square distribution underlying χ_k^2.
- (N-1) in the denominator is used for adjusting the effect of sample size on the non-centrality parameter.
- df_k in the denominator is meant to provide a penalty for using model degrees of freedom.

A RMSEA of < 0·10 is considered a good fit and < 0·05 is very good, lower than 0.01 is considered as a beautiful fit (Steiger, 1990).

The single sample expected cross validation index (ECVI):

Browne and Cudeck (1989, 1993) developed a single sample expected cross validation index (ECVI) and also explained the use of ECVI in structural equation modeling. Except for a constant scale factor, ECVI is the same as the Akaike information criterion (Akaike, 1973, 1987). Arbuckle (2006, p. 542) reported the MECVI, which except for a scale factor is identical to the Brown-Cudeck Criterion (BCC). The BCC enforces a marginally greater penalty for model complexity than the AIC and it is a fit measure developed for the analysis of moment structures. These fit measures are planned for model comparisons and accordingly indicate "goodness of fit" with simple models that fit well receiving low values and poorly fitting models receiving high values. The ECVI is a function of chi-square and degrees of freedom. It is computed in AMOS as

$ECVI = \frac{AIC}{n}$ where $n = N - r$, with N the sample size and r the number of groups. Browne and Cudeck (1989, 1993) provided a confidence interval for ECVI. In AMOS the 90% lower and upper confidence limits C_L and C_u respectively, are given by

$\left[\frac{(\delta_L + d + 2q)}{n}, \qquad \frac{(\delta_u + d + 2q)}{n}\right]$, where δ_L and δ_U are the parameter estimates for the lower and upper limits respectively, d is the degree of freedom and q is the number parameters.

Path significance was based on the critical ratio (CR), with a CR > 2 in absolute value considered as significant (Arbuckle, 2006; Schumacker and Lomax, 2004).

4.6 Data Analysis

The statistical analyses were performed using AMOS software (Amos Development Corporation). Path analysis was conducted by considering the radial measure as dependent, climatic variables and age as independent factors explaining the radial growth. The chi-square statistic, the normed fit index (NFI), and Root Mean Square Error of Approximation (RMSEA) were used to check the goodness of model fit. The larger the probability associated with the chi-square, the better the fit of the model to the data (Bollen, 1989; Byrne, 2001). The NFI tests the hypothesized model against a reasonable baseline model and ideally should be 1·0. Model validity was assessed using the expected cross validation index (ECVI).

4.7 Results of Fitting Path Models

The independent variables included in the study were the five major climatic variables that were measured and the age of the trees. The association between the independent variables and the radial growth measurement of the clones is presented in Figure 4.3. The numbers displayed at the top of the diagram refer to the goodness of fit of the model. This fit statistic is the likelihood ratio chi-square test. The p-value associated with this measure is 0.894, which is far larger than 0.05 and indicates a non-statistical significance of the chi-square test. This implies the model is consistent with the data. The numbers displayed next to the double headed arrows are estimated correlation coefficients.

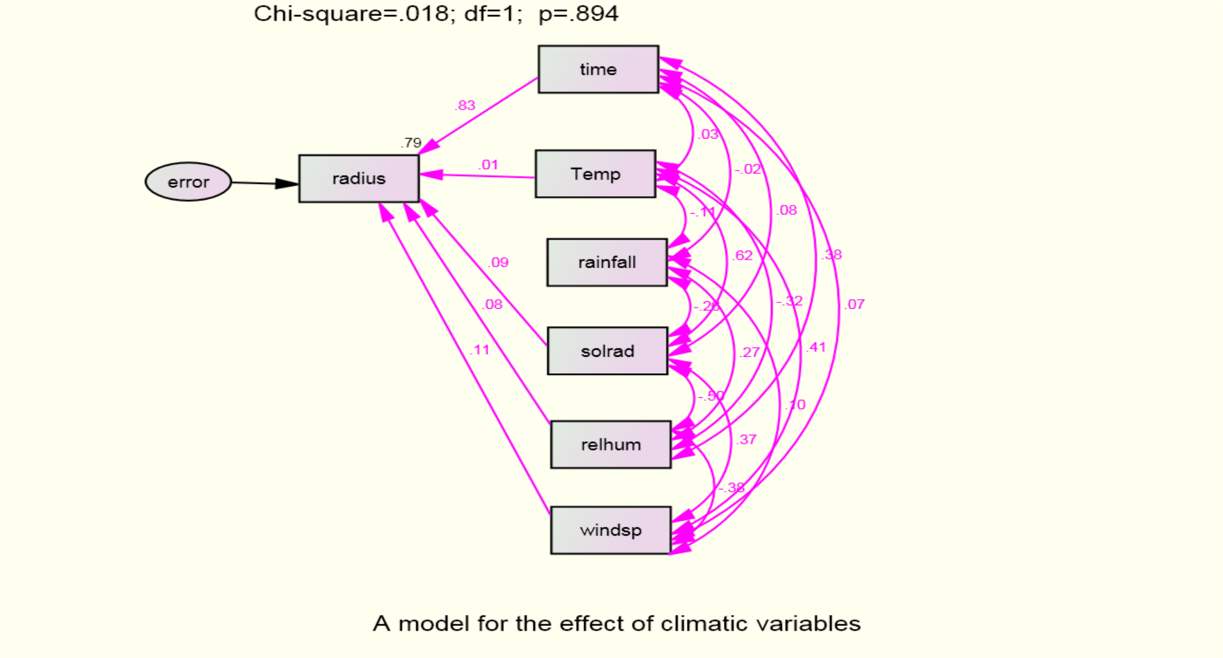

Figure 4. 3 Path diagram showing the effect of age and climatic variables on radius of Eucalyptus hybrid clones during the first measured phase of growth. Time = age; solrad = solar radiation; relhum = relative humidity; windsp = wind speed; Temp=temperature.

Various measures of fit (Table 4. 1) are presented for the fitted model, given in Figure 4.3, and include the saturated model, which is the ideal fit by including all possible paths. A model that can be defined as good is one that does not differ significantly from the saturated model despite omitting paths from the saturated model. On the other hand, the ordinary regression model or independent model fits by ignoring any potential relatedness between the independent variables thus considering all correlations among the independent variables as zero. The fit indexes for saturated model are very close to the fit indexes obtained for our model (Table 4.1) indicating that the model at hand can be considered good.

Table 4. 1 Different fit measures for the fitted model, saturated and ordinary regression models

Fit measure	Model		
	Fitted Model[1]	Saturated Model[2]	Ordinary Regression[3]
Chi square	0.02		1287.06
Chi square p-value	0.89		0
Normed fit index (NFI)	1	1	0
Root mean square error of approximation (RMSEA)	0		0.386
Expected cross-validation index (ECVI)	0.006	0.006	3.13
ECVI lower bound	0.006	0.006	3.068
ECVI upper bound	0.007	0.006	3.193
Modified expected cross validation index (MECVI)	0.006	0.006	3.131

[1] The model presented in Figure 4.3.

[2] Model that includes all possible paths.

[3] The independent model that assumes no correlation between the independent variables.

The statistical significance of individual parameter estimates for the paths in the fitted model (Figure 4.3) is one of the important criteria to be studied. The significance can be seen by computing the critical ratios, which are obtained by dividing the parameter estimates by their respective standard errors. The computed critical ratios together with the corresponding p-values are presented in Table 4.2. The regression weights for all variables were significant with the exception of rainfall, which was dropped from the model.

Table 4. 2 Regression weights indicating the relationship between radial growth and each independent variable for the combined data set.

Relationship	Maximum Likelihood Estimates	Standard Error	Critical Ratio	P-value
radius<---time	313.51	2.18	143.91	***
radius<--Temperature	23.74	12.64	1.88	0.06
radius<---solar radiation	2817.03	220.03	12.80	***
radius<---relative humidity	63.76	5.75	11.09	***
radius<---wind speed	1447.03	73.63	19.65	***

" *** " indicates the p-value is less than 0.001.

The other issue to consider at this stage is the magnitude and direction of the parameter estimates. In this particular model all the regression weights were positive indicating the existence of a positive relationship between radial growth and the climatic variables. The standardized regression coefficients are 0.832 (age of a tree), 0.012 (temperature), 0.092 (solar radiation), 0.076 (relative humidity) and 0.113 (wind speed). This suggests that the most important variable to explain radial growth is age of the tree. It is also estimated that the predictors of radius explain 79 % of its variance. In other words, the error variance of radius is approximately 20.9% of the variance of radius itself.

Although the goodness of fit measures indicate that the fitted model (Figure 4.3) is a good fit (refer Table 4.1), the parameter estimates show that rainfall has no direct influence on the radial growth. An attempt was made to modify the fitted model (Figure 4.3) by making rainfall a required variable in the model. Such a modification procedure is called specification search (Leamer, 1978). The objective of a specification search is to alter the original model in search of a model that is better fitting in some sense, and yields parameters having practical, and in this case biological significance and substantive meaning. The path diagram for the first attempt at modification is presented in Figure 4.4. For this path analysis model, a good 'goodness of fit' was obtained. The calculated value of the chi-square statistics was 3.194 with one degree of freedom and a p-value of 0.074. However, the goodness of fit for the second fitted model (Figure 4.4) was not as good as the model fit shown in Figure 4.3. The parameter estimates for the second fitted model (Figure 4.4) suggest that rainfall had no direct significant effect. Therefore, no additional information was gained by modifying the path diagrams from that of Figure 4.3 to that of Figure 4.4.

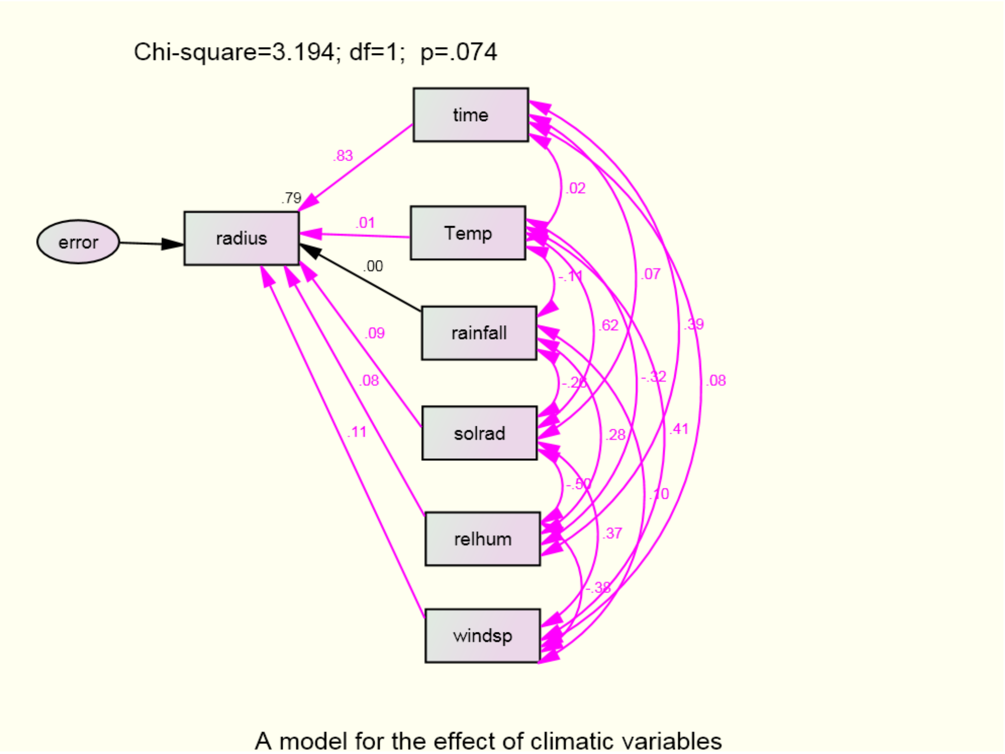

Figure 4. 4 Path diagram showing the effect of age and climatic variables on radius of *Eucalyptus* clones when rainfall is considered a required variable. Time = age; solrad = solar radiation; relhum = relative humidity; windsp = wind speed; Temp=temperature.

The third attempt at specification search was to consider a model fit for the second fitted model (Figure 4.4) that excluded wind speed as an explanatory variable (Figure 4.5). The model fit was good and parameter estimates were significant. The regression weight for rainfall in the prediction of radial growth was significantly different from zero at the 0.001 level (two-tailed, Figure 4.5). This indicates that rainfall has a significant effect on the radial growth of trees in the absence of wind speed. For this model, it is estimated that the predictors of radial growth explain 78.2% of its variance. This is very close to the value obtained for the first model (Figure 4.3), which includes all the predictors in the model. The standardized regression coefficients were 0.859 (age of a tree), 0.042 (temperature), 0.096 (solar radiation), 0.026 (relative humidity) and 0.03 (rainfall). These standard regression coefficients indicate that age of the tree is the most important variable in determining the stem radial growth.

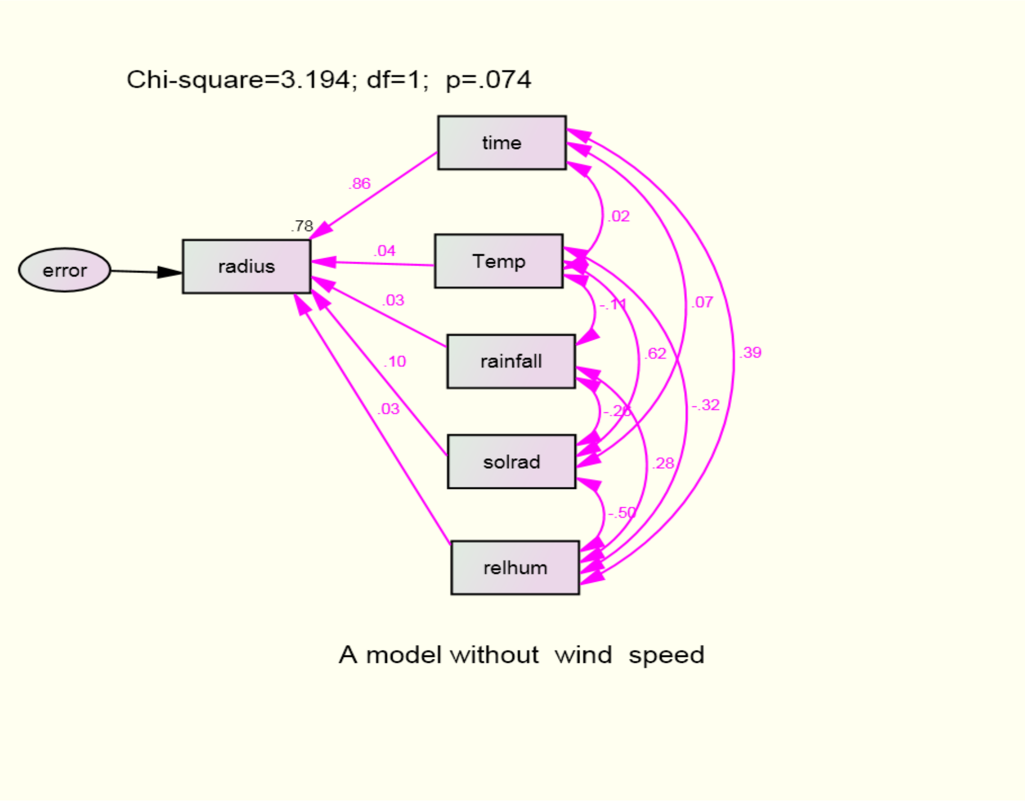

Figure 4. 5 Path diagram showing the effect of age and climatic variables on radius of *Eucalyptus* clones when wind speed is omitted as an explanatory variable. Time = age; solrad = solar radiation; relhum = relative humidity; Temp= temperature.

Models fitted without temperature or tree age as explanatory variables did not fit well. A model that excluded relative humidity fitted well and resulted in rainfall having a significant effect on radial growth. The significance of rainfall in the absence of relative humidity and solar radiation was possibly caused by multicollinearity or suppressor variables (where two or more predictor variables in a multiple regression model are highly correlated). The correlation among the climatic variables themselves is also significant. When only rainfall and wind speed were considered independent variables, the regression weight for rainfall became negative. The same occurred when only rainfall and relative humidity were treated as independent variables. This wrong sign of coefficients is an indication of possible multicollinearity. As a result, the effect of rainfall on radial growth cannot be completely ruled out, as its non-significance is possibly caused by multicollinearity. Some researchers noted that structural equation models are robust against multicollinearity (Malhotra et al., 1999), with some going as far as to explicitly state that Structural Equation Models (SEM) can remedy multicollinearity problems. For example, Maruyama (1998) argues that "structural equation approaches can help deal with some cases where the correlations among the predictors are large”. On the other hand, some researchers have warned that multicollinearity can lead to SEM estimates being far from the true parameters, as well as the occurrence of large standard errors of the estimates (Jagpal, 1982; Grapentine, 2000). A simulation study by Grewal et al. (2004) showed some conditions under which multicollinearity caused problems. The study showed that when multicollinearity is extreme, type II error rate (accepting the null hypothesis when it is false) is generally, unacceptably high. They also indicated that for multicollinearity levels of between 0.6 and 0.8, type II error rates can be substantial (greater than 50% and frequently above 80%), if composite reliability is weak, explained variance (R^2) is low and sample size is relatively small. When multicollinearity levels are between 0.4 and 0.5, type II error rates tend to be quite small except when reliability is weak, R^2 is low and the sample size is small. In the present study R^2 values were large and the multicollinearity level was not high. Estimates of regression weights for rainfall, which is important for growth, were inconsistent. Consideration of more complex models may improve results. In the path diagrams considered thus far only one dependent variable (radial growth) was used. Path analysis allows the simultaneous modelling of several related regression relationships. This means that path analysis can handle more than one dependent variable in the model. Moreover, a variable can be a dependent variable in one relationship and an independent variable in another relationship of the path model. An attempt was made to fit a model where two dependent variables, namely rainfall and temperature, mediated the effects of relative humidity, solar radiation and wind speed. In this model, it was hypothesized that tree age had a direct effect on radial growth. Solar radiation, relative humidity and wind speed were assumed to have an indirect effect. The fitted model is presented in Figure 4.6.

The value of the chi-square statistic is 862.7 with a p-value of zero. This indicates that the model does not fit the data well. However, the parameter estimates of the regression weights are all significant (Table 4.3). The magnitude of each effect is quantified by standardized regression coefficients. The standardized regression coefficients are 0.87 (age of the tree), 0.091 (temperature), and 0.018 (rainfall). From this it can be seen that the most important variable to explain radial growth is tree age. For the model in Figure 4.6, there are three structural equations, one for each of the three dependent variables: rainfall; temperature and radius. In terms of variable names, the structural equations are:

$$\begin{aligned}
&ra\inf all = relative\,humudity + solar\ radiation + wind\ speed + error1\\
&temperature = relative\,humudity + solar\ radiation + wind\ speed + error2\\
&radius = ra\inf all + temperature + time + error\,3
\end{aligned}$$

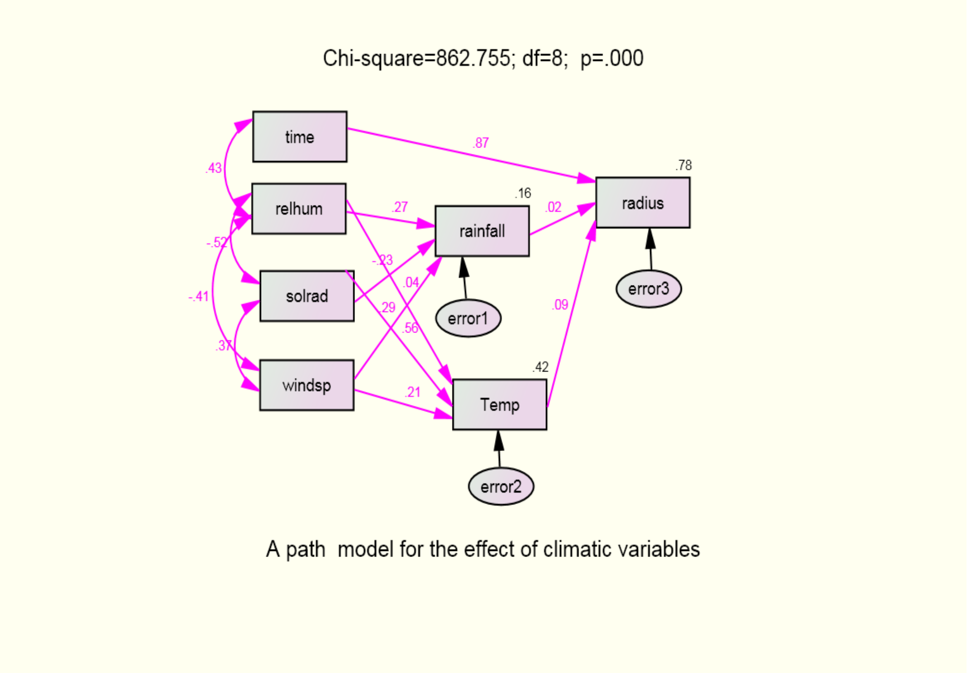

Figure 4.6 Path diagram showing the effect of multiple dependent variables (rainfall and temperature) on radial growth of *Eucalyptus* clones. Time = age; solrad = solar radiation; relhum = relative humidity.

This model includes direct effects (e.g. age of the tree on radial growth), indirect effects (e.g. effect of relative humidity through rainfall) and correlated explanatory variables (e.g. relative humidity, solar radiation and wind speed). The estimated model using AMOS statistical software is given by:

$$ra\inf all = 0.196\ relative\ \ humudity\ \ - \ 6.27\,solar\ \ radiation\ + \ 3.22\,wind\ \ speed$$

$$temperatur\ e = 0.017\,relative\ \ humudity\ \ + 8.77\ \ solar\ \ radiation\ + 1.39\,wind\ \ speed$$

$$radius\ \ = \ \ 20.73\,ra\inf all\ \ + 178.37\,temperatur\ e + 329.67\ \ time$$

From the above fitted model (Figure 4.6) the positive effect of the predictors, rainfall, temperature and tree age can be seen. The standardized regression weights for this model indicate that tree age, temperature and rainfall are respectively important determinants of radial growth.

The data set to which the above models were applied was a combined data set (for both *E. grandis* hybrid clones). In order to see if there was any difference between the two clones, a multiple group analysis was used. In this regard, two models (the model in Figure 4.3 and the model with multiple dependent variables (Figure 4.6)) were considered. The good fitting model of Figure 4.3 was fitted to the data set for GU clone alone. The model fitted the data very well. The value of the chi-square statistics was 0.06 with one degree of freedom and the corresponding p-value was 0.804. The next question to address was whether the same model fitted the data for the GC clone. Furthermore, the equality of the parameters needed to be tested. Instead of a separate group analysis, a single analysis that simultaneously estimated parameters and tested hypotheses about both groups was considered. This method provided a test for the significance of any differences found between the GU and GC clones. In addition, if there were no differences between the two clones, or if group differences concerned only a few model parameters, the simultaneous analysis of both groups would have provided more accurate parameter estimates than would have been obtained from separate single-group analyses. A test for pair wise path coefficient differences for the two clones was conducted. Some fit measures for various models were generated, together with fit measures for saturated and independence models (see Table 4.3).

The structural weight model specifies that the regression weights for predicting radial growth from the measured climatic variables and tree age were the same for the GU and GC clones. The unconstrained model is the model that assumes that all the parameters for the two groups are different. For the unconstrained model, the value of chi-square was 0.08 with the corresponding p-value equal to 0.96. This indicated that the unconstrained model fitted the data very well.

Table 4.3 Summary of fits for various models including the structural weight model

Model	Number of parameters	Chi-square	df	P-value	Chi-square / df
Unconstrained	54	0.08	2	0.96	0.04
Structural weights	49	364.59	7	0.00	52.09
Structural covariances	28	364.59	28	0.00	13.02
Structural residuals	27	1293.58	29	0.00	44.61
Saturated model	56	0.00	0		
Independent model	14	29255.12	42	0.00	696.55

df = Degrees of freedom

The structural weight model with a chi-square value of 364.59 and with seven degrees of freedom was rejected at any conventional significance level, suggesting that the regression weights of the two clones were significantly different. The assumption that the regression weights for the exogenous variables were the same for both clones was not supported. The estimated regression weights for the unconstrained model are summarized in Table 4.4 and Table 4.5 respectively for GU and GC clones. When comparing the regression weights for the two clones, they were all positive, indicating a positive effect of the climatic variables as well as tree age on radial growth. In addition, regression weights obtained for the GU clone were larger than those obtained for the GC clone, indicating that the GU clone grows faster than the GC clone. Regression weights of the GU and the GC clones, for the multiple dependent model in Figure 4.6 were also compared. The regression weights for the two clones were significantly different. The results of this model also show that the GU clone has a faster growth than the GC clone.

Table 4. 4 Regression weights for GU clone when the path model in Figure 4.3 was fitted to compare the two clones (unconstrained).

Relationship	Maximum Likelihood Estimates	Standard error	Critical ratio	P-value	Label
radius<---time	341.88	3.33	102.81	***	b1_1
radius<---temperature	43.34	19.30	2.25	0.025	b2_1
radius<---solar radiation	3253.04	335.85	9.69	***	b3_1
radius<---relative humidity	75.14	8.77	8.57	***	b4_1
radius<--- wind speed	1570.35	112.39	13.97	***	b5_1

“ *** ” indicates the p-value is less than 0.001.

The maximum likelihood estimates given in Tables 4.4 and 4.5 require the data to be of a continuous scale and have a multivariate normal distribution. The approximate standard errors used in the inference were therefore produced based on formulae that depend on normality assumptions. Non-normality can lead to spuriously low standard errors, with degrees of underestimation ranging from moderate to severe. The consequences are that because the standard errors are underestimated, the regression paths and factors / error covariances will be statistically significant, although they may not be so in the population (Byrne, 2001).

Table 4. 5 Regression weights for the GC clone when the path model in Figure 1 was fitted to compare the two clones (unconstrained).

Relationship	Maximum Likelihood Estimates	Standard error	Critical ratio	P-value	Label
radius<---time	285.14	2.075	137.436	***	b1_2
radius<---temperature	4.13	12.040	.343	0.732	b2_2
radius<---solar radiation	2381.02	209.543	11.363	***	b3_2
radius<---relative humidity	52.39	5.472	9.575	***	b4_2
radius <---wind speed	1323.72	70.119	18.878	***	b5_2

" *** " indicates the p-value is less than 0.001.

It is known that many data do not qualify for multivariate normality and the current data is no exception. Using AMOS statistical software the data was checked to see whether it had a multivariate normal distribution. The Mardia's (1970) coefficient of multivariate kurtosis was 57.31 with a critical ratio of 237.3, which highly favours multivariate non-normality of the data.

A possible approach to overcome the problem of the existence of multivariate non-normal data is to use a method known as "bootstrap" (West et al., 1995; Yung and Bentler, 1996). This technique enables us to create multiple subsamples from an original database. The importance of drawing these multiple samples is that we can examine parameter distributions relative to each of these newly produced samples. These distributions serve as a bootstrap sampling distribution and technically operate in the same way as the sampling distribution generally associated with parametric inferential statistics. In contrast to traditional statistical methods, however, the bootstrap sampling distribution is concrete and allows for comparison of parametric values over repeated samples that have been drawn (with replacement) from the original sample. The bootstrap method is free from the distributional assumptions and can be used to generate an approximate standard error for many statistics without having to satisfy the assumption of multivariate normality. With this beneficial feature in mind, the bootstrap method was applied to the good fitting model in Figure 4.3. In this process, 10,000 bootstrap samples were generated. The reported value of the chi-square was 0.018 with one degree of freedom. The bootstrap standard errors for regression weights are presented in Table 4. 6. The table lists the bootstrap estimate of the standard error for each independent variable in the model. Each value represents the standard deviation of the parameter estimates computed across the 10,000 bootstrap samples. These values were compared with the values of the approximate maximum likelihood estimates presented in Table 4.2. Some discrepancies between the two sets of standard error estimates were observed. The third column of Table 4.6, labelled SE-SE provides the approximate standard error of the bootstrap standard error itself. These values were very small indicating that the standard errors were estimated with a reasonable level of accuracy. Column four, labelled Mean, lists the mean parameter estimates computed across the 10,000 bootstrap samples. Arbuckle (2006) on page 301 emphasized that this bootstrap mean is not necessarily identical to the original estimate. Column five (Bias) represents the differences between the bootstrap mean estimates and the original estimates. These values are very small for most of the cases and positive values indicate that the estimates of the bootstrap samples are higher than the original maximum likelihood estimates. The low bias indicates that the maximum likelihood estimates and the bootstrap estimates are very close to each other.

Table 4. 6 Bootstrap standard errors for path model in Figure 4.3

Parameter (un-standardized)	SE	SE-SE	Mean	Bias	SE-Bias
radius<---time	2.35	0.017	313.52	0.010	.024
radius<---temperature	12.55	0.089	23.85	0.11	.125
radius<---solar radiation	220.36	1.56	2816.58	-0.451	2.204
radius<---relative humidity	5.89	0.042	63.75	-0.018	.059
radius<---wind speed	69.65	0.493	1446.07	-0.967	.697
Standardized Parameter					
radius<---time	.004	.000	.832	.000	.000
radius<---temperature	.006	.000	.012	.000	.000
radius<---solar radiation	.007	.000	.092	.000	.000
radius<---relative humidity	.007	.000	.076	.000	.000
radius<---wind speed	.006	.000	.113	.000	.000

The last column, labelled SE-Bias, reports the approximate standard error of the bias estimate. For the majority of the cases the estimated bias is smaller in magnitude than its standard error. This indicates that there is little evidence that the regression weights are biased.

The bootstrap confidence intervals are presented in Table 4.7. The bias-corrected confidence intervals are used because these intervals are considered to yield more accurate values than percentile confidence intervals (Efron and Tibshirani, 1993).

The confidence intervals for tree age, solar radiation, relative humidity and wind speed do not include zero. It can therefore be concluded that the regression weights of these independent variables are significantly different from zero. The value of p in the 'p' column of Table 4.7 indicates that a 100(1-p) percent confidence interval would have one of its end points at zero. In this sense, the p-value can be used to test the hypothesis that an estimate has a population value of zero. In this case the relationship between radius and temperature has a p-value 0.06, which means that a 94% confidence interval would have a lower boundary at zero. In other words, a confidence interval at any level less than 94% such as 90% or 92% would not include zero, and therefore reject the hypothesis that the regression weight is zero for a 90% confidence interval. For the relationship of radius with other independent variables the hypothesis at any conventional significance level such as 95% or 99% is rejected.

Table 4. 7 Ninety-five percent bootstrapped confidence intervals (bias corrected percentile method).

Regression Weights	Estimate	Lower	Upper	P
radius<---time	313.51	308.86	318.03	.000
radius<---temperature	23.74	-1.21	48.76	.060
radius<---solar radiation	2817.03	2392.34	3252.47	.000
radius<---relative humidity	63.76	52.27	75.19	.000
radius<---wind speed	1447.03	1314.33	1588.51	.000
Standardized regression weights				
radius<---time	0.832	0.824	0.841	.000
radius<---temperature	0.012	-0.001	0.025	.059
radius<---solar radiation	0.092	0.078	0.106	.000
radius<---relative humidity	0.076	0.063	0.090	.000
radius<---wind speed	0.113	0.103	0.124	.000

Therefore, by applying the bootstrap method, it can be seen that the independent variables had a significant effect on the radial growth of *Eucalyptus* trees. This result also agreed with the result obtained using the maximum likelihood method. It is also of interest to evaluate the appropriateness of the hypothesized model itself. Bollen and Stine (1993) provided a means of testing the null hypothesis that the specified model was correct. The Bollen-Stine bootstrap corrected p-value was 0.878. This corrected p-value indicates that the hypothesized model should not be rejected. This result is also in agreement with the maximum likelihood results. The other issue with the specified model was cross validation. To assess the validity of the model in Figure 4.3, expected cross validation index (ECVI) was applied. ECVI is proposed as a means to assess, in a single sample, the likelihood that the model cross-validates across similar size samples from the same population (Browne and Cudeck, 1989). It measures the discrepancy between the fitted covariance matrix in the analysed sample, and the expected covariance matrix that would be obtained in another sample of equivalent size. Application of ECVI assumes a comparison of models, whereby ECVI index is computed for each model and then all ECVI values are placed in rank order. The model having the smallest ECVI value exhibits the greatest potential for replication. There is no determined appropriate range of values for ECVI as it can assume any value (Byrne, 2001). In the present case the values of ECVI are presented in Table 4.1. In assessing the hypothesized model, its ECVI value of 0.006 was compared with that of the independence model (ECVI=3.13). The ECVI for the saturated model was also 0.006. The ECVI for the hypothesized model was less than that of the independence model. It can therefore be concluded that the hypothesized model represents the best fit to the data. Furthermore, a 95% confidence interval for ECVI is given by [0.006, 0.007]. This indicates that of the overall possible randomly sampled ECVI values, 95% of them will fall [0.006, 0.007], suggesting that the model cross validates over the independent model.

4.8. Summary

Classical methods, like ordinary regression models once the regression model is specified, do not permit any other relationships among the explanatory variables to be specified. This limits the potential of the variables to have direct, indirect and total effects on each other. In path analysis one can see the direct effect, indirect effect and total effects of variables. In path analysis a unique additional contribution of each variable can be studied using the standardized regression weights. Even though we can study the additional contribution of each variable in multiple regressions, this can work ideally only if all independent variables are highly correlated with the dependent variable and uncorrelated among themselves. In contrast, path models provide theoretically meaningful relationships in a manner not restricted to a multiple regression model (Schumacker, 1991). In path analysis, we can estimate parameters for more than one regression equation because this analysis can be considered as a series of regressions applied sequentially to the data. Structural Equation Models (SEM) are considered as path analysis involving latent variables. In the present case, latent variables were not included and hence path models were generated. Path analysis was employed mainly because the climatic variables were correlated and the unique, additional contribution of each climatic variable on radial growth of eucalypts was of interest.

The best fitting path model generated in this study showed that all climatic variables and age of the tree had a positive effect on stem radial growth for the pooled data of both clones. Furthermore, all except one variable (rainfall) had a significant, direct effect on radial growth. It was also observed that the age of the tree was the most important variable explaining stem radial growth. Although rainfall was not significant in the best fitting model, it was found to be significant for the model that excluded wind speed and for the model that omitted solar radiation. This revealed that the effect of rainfall on radial growth cannot be ruled out. To compare the effect of the explanatory variables on the radial growth of the GU and GC clones, a single analysis that estimated parameters and tested hypotheses about both groups simultaneously was considered. The regression weights for the two clones were significantly different. The regression weights were all positive indicating the positive effect of the climatic variables as well as tree age. In addition, the regression weights obtained for the GU clone were larger than the regression weights for the GC clone. This shows that the GU clone was growing faster than the GC clone which can easily be confirmed by looking at the growth of the two clones.

The main estimation method for path models, or any structural equation model (SEM) is maximum likelihood estimation. This method requires a distributional assumption, which the present data failed to satisfy. The bootstrap method was then applied to overcome the methodological failure due to non-normality. The estimated bias using the bootstrap method was very small showing that there was little evidence of bias in the estimates. The conclusion reached using the maximum likelihood method agreed with that of the bootstrap method. The expected cross-validation index obtained for the hypothesized model also showed that this model cross-validated over the independent model.

To sum up, the climatic variables measured in this study, together with tree age, had a positive effect on stem radial growth during the juvenile stage of development. Age of the tree was the most important variable in explaining stem radial growth. The growth of the GU clone was faster than the growth of the GC clone, possibly indicating that this clone has better genetic potential. However, this could also indicate that, compared to the GC clone, the GU clone is better adapted to the environmental conditions, or it is able to use the available resources more effectively. Melesse and Zewotir (2013b) provides a detailed discussion of these results (attached in Appendix). The models we have considered so far did not take into account the within-tree variability. The next step is to review some methods where the longitudinal aspect of the data is specifically taken into account. We begin by reviewing fractional polynomial models and use them to study the longitudinal growth of stem radius in chapter 5.

Chapter 5

Fractional Polynomial Models (FP)

5 .1 Introduction

Cross sectional study may allow comparison among subpopulations that happen to differ in age, but it does not provide any information about how individuals change over time. The assessment of within subject changes in response over time can only be achieved within a longitudinal study. A distinctive feature of longitudinal data is that observations on the same individual are correlated over time. Failure to account for the effect of correlation can result in an erroneous estimation of the variability of parameter estimates and hence in misleading inference. For example, if we want to estimate the change in mean response between two time points for N subjects, then the estimate of the change in the mean response between the two time points is given by

$d = \overline{y}_2 - \overline{y}_1$ where $\overline{y}_j = \sum_{i=1}^{N} \frac{y_{ij}}{N}$. To get the standard errors, we need to estimate the variance of the difference (d). The variance of the difference is given by

$$\operatorname{var}(d) = \operatorname{var}\left\{\overline{y}_2 - \overline{y}_1\right\} = \frac{1}{N}\left[\sigma_2^2 + \sigma_1^2 - 2\sigma_{12}\right]$$

The last term represents the covariance between the measurements of the two time points. Assuming that the two repeated measures are independent when there is strong positive correlation between them, would result in an incorrect estimate of variance. This will bring an overestimation of the variability of the difference in mean responses. Consequently, failure to account for correlation among repeated measures leads to incorrect standard errors. The incorrect standard will lead to incorrect test statistics and p-values. This will finally lead to incorrect inferences about the regression parameters (Fitzmaurice et al., 2004; Weiss, 2005). Therefore, the correlation among repeated measures necessitates a statistical analysis that appropriately accounts for the dependence among measurements within the same subject, which results in more precise and powerful statistical analysis.

This interdependence can be modelled using mixed models. The current data set consisted of repeated measurements of the same subjects over time; therefore, a mixed effects models approach was adopted in the analysis of the longitudinal data (Verbeke and Molenberghs, 1997; 2000; Fitzmaurice et al., 2004; Meng and Huang, 2010). Models for the analysis of such data recognize the relationship between serial observations on the same unit. Since change in stem radial growth, which is a continuous response variable, is the main object of the study, it is of interest first to study the mean effect of time (tree age). We also adopted the fractional polynomial (Royston and Altman, 1994; James, Wang and Zhu 2009) approach to the mixed model by using a polynomial regression model with parameters that are allowed to vary over individuals, and which are therefore called random effects or subject-specific regression coefficients. Their mean then reflects the average evolution in the population.

In any applied longitudinal data analysis the main objective is to fit a smooth curve over the time interval of data collection. When the relationship between the response and the independent variable (time) is believed to be linear, the shape of the smooth curve is not contested. The focus is mainly whether the straight line is horizontal or not. In contrast, when one believes the trend is not linear, then the smooth curve is commonly selected from some alternatives, such as orthogonal polynomials of a number of orders. Orthogonal polynomials are most closely associated with traditional methods that do not allow missing data. In models that allow missing data, such as the linear mixed model, correlated polynomials terms are often used. These so-called conventional polynomials (CPs) consists of power transformations of time metric with the integer exponents, $1, . . . , p$ are widely used as they have the advantage of simplicity, familiarity, invariance to change of origin, and the ability to approximate any nonlinear function (Long and Ryoo, 2010). Although the presence of curvature can be handled by using conventional polynomials, in most applications the choice is made between linear and quadratic terms, with cubic or higher order polynomials being rarely used or useful. It has long been recognized that conventional polynomials (which offer only few curve shapes) do not fit the data well. High order polynomials (sometimes even cubic polynomials) follow the data more closely but often fit badly at the extremes of the observed range of the independent. An extended family of curves called fractional polynomials, whose power terms are restricted to small predefined set of integer and non-integer values were proposed by Royston and Altman (1994). Fractional polynomials are analogous to conventional polynomials in that their time transformations are power functions, however, the exponents of fractional polynomial are not only integers but also negative numbers and fractions.

The paper by Long and Ryoo (2010) provides a unified framework for evaluating and selecting fractional polynomials in longitudinal data analysis. They discussed fractional polynomials within the context of linear mixed models. Parsimony, a wide variety of curve shapes for low order models and the ability to approximate asymptote are the attractive features of fractional polynomial models (Long and Ryoo, 2010). A brief introduction to fractional polynomial is given below.

5.2 Fractional Polynomial Models in the Context of Linear Mixed Models

Suppose that y_{ij} is the variable of interest for the i^{th} entity ($i = 1, 2 ... N$) at the j^{th} time point $(j = 1, 2, . . . n_i)$ and that y_i is an n_i dimensional response vector for one entity. The linear mixed model (Laird & Ware, 1982) is given by

$$y_i = X_i\beta + Z_i b + \varepsilon_i \qquad (5.1)$$
$$b_i \sim N(0, D)$$
$$\varepsilon_i \sim N(0, R_i)$$

where X_i is a design matrix of dimension $n_i \times (p+1)$, β *is a* $(p+1)\times 1$ vector containing fixed effects and Z_i is $n_i \times (q+1)$ known matrix linking b_i to y_i, b_i is the $(q+1)$ dimensional vector containing the random effects, ε_i is an n_i dimensional vector of residual components. Finally, D is a general $(q+1) \times (q+1)$ covariance matrix with $(i, j)^{th}$ element $d_{ij} = d_{ji}$ and $R_i = \sigma^2 I_{n_i}$, where I_{n_i} is a $(n_i\, x\, n_i)$ identity matrix which depends on i only through its dimension n_i. The random effects and the residual components are assumed independent. The diagonal elements of the matrix D are assumed non-negative. This latest assumption permits a hierarchical interpretation of the linear mixed models meaning both individual level and group level models are subsumed (Verbeke and Molenberghs, 2000).

Long and Ryoo (2010) consider models in which the covariates X_i and Z_i consist of only time transformations other than the first column of ones. They also assume $X_i = Z_i$ so that each fixed effect has the corresponding random effect (p=q). Under these conditions, the design matrix has the following form.

$$X_i = Z_i = \begin{bmatrix} 1 & f_1(t_{i1}) & . & . & . & f_p(t_{ip}) \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 1 & f_1(t_{in_i}) & . & . & . & f_p(t_i n_i) \end{bmatrix} \qquad (5.2)$$

where $f_b\left(t_{ij}\right)$ is fractional polynomial consisting of p-transformation in the design matrix. This fractional polynomial is defined as follows (Royston & Altman, 1994).

$$f_b(t_{ij}) = \begin{cases} t_{ij}^{(m_b)}, & if \quad m_b \neq m_{b-1} \\ \\ f_{b-1}(t_{ij}) \times \log\left(t_{ij}\right), & if \quad m_b = m_{b-1} \end{cases} \tag{5.3}$$

where $m_1 \leq m_2 \leq ... \leq m_p$ and $\log(t_{ij})$ indicates the natural log of t_{ij}. The round bracket notation, $t_{ij}^{(m_b)}$, represents the Box and Tidwell (1962) transformation,

$$t_{ij}^{(m_b)} = \begin{cases} t_{ij}^{m_b}, & if \quad m_b \neq 0 \\ \\ \log\left(t_{ij}\right), & if \quad m_b = 0 \end{cases} \tag{5.4}$$

with the constraint $t_{ij} > 0$ so that all transformations are defined (Long and Ryoo, 2010). Although many different combinations of powers (m_b) can be made in fractional polynomial models, it has been suggested that it is often adequate to choose powers from the restricted set $m_b = \{-2,\ -1, -0.5,\ 0,\ 0.5,\ 1,\ 2,\ \ 3\}$ for practical purposes. The inclusion of positive integers in the set indicates the conventional polynomials are special cases of fractional polynomials. In equation (5.3) p stands for the order of the polynomial. For instance, if $p = 1$, we have a fractional polynomial of order one. Threfore, the value of m_1 can be chosen from a set $m_1 = \{-2,\ -1, -0.5,\ 0,\ 0.5,\ 1,\ 2,\ \ 3\}$. This offers a wide variety of curve shapes such as square root $(m_1 = 0.5)$, linear $(m_1 = 1)$ and inverse $(m_1 = -1)$. For fractional polynomial of order two $(p = 2)$, the values of both m_1 and m_2 can be chosen from the set m_b. By choosing different values for the combination of m_1 and m_2 we can get different curve shapes. This set includes linear, reciprocal, square root, square and cubic transformations and their combinations. The best fit among the possible 36 combinations of such powers is defined as that which maximizes the likelihood function. Such second degree fractional polynomials offer considerably more flexibility and accommodate many functions with single turning points as well as j shaped relationships (see for example, Royston, Ambler and Sauerbrei (1999); Long and Ryoo (2010).

By making use of the design matrix in equation (5.2), the linear mixed model in the individual level for fractional polynomial of order p is

$$y_{ij} = \beta_0 + \sum_{b=1}^{p} \beta_b f_b\left(t_{ij}\right) + b_{i0} + \sum_{g=1}^{p} \beta_{ig} f_g\left(t_{ij}\right) + \varepsilon_{ij} . \qquad (5.5)$$

The first two terms of (5.5) represents the fixed effect part of the model. The third and fourth terms stands for the random effect part of the model and ε_{ij} is the residual component. The marginal model can be obtained by taking the expectation of (5.5) which is

$$E\left(y_{ij}\right) = \beta_0 + \sum_{b=1}^{p} \beta_b f_b\left(t_{ij}\right). \qquad (5.6)$$

The random effects and residual components are assumed to have zero mean (equation 5.1).

The marginal model in (5.6) can be used to study changes in nonlinear growth curves. The parameters of this model together with the fractional polynomials; $f_b\left(t_{ij}\right)$, represent the equation for the mean growth curve. The β parameters have the literal interpretations as the weights applied to the time transformation in determining the curve. Perhaps more attractive, the β parameters determine the instantaneous rate of change, which is a convenient means of studying changes with nonlinear growth curves. The instantaneous rate of change is the slope of the tangent line at the point

$(t_{ia} \quad E(y_{ia}))$, where t_{ia} is particular value of time and $E(y_{ia})$ is the mean growth evaluated at that particular value t_{ia}. The slope of the tangent line (γ_{tia}) indicates a change in $E(y_{ij})$ for a unit increase in t_{ij} evaluated at t_{ia} (Long and Ryoo, 2010).

For the marginal model (5.6) the general form the slope of the tangent is given by

$$\gamma_{t_{ia}} = \sum_{b=1}^{p} \beta_b \, f_b'(t_{ia}) \qquad (5.7)$$

where f' indicates the first derivative of the function f.

In the context of linear mixed models, the first order fractional polynomial model ($p = 1$) can be obtained from equation (5.3). Using the expected value notation, it is possible to write equation (5.3) as

$$E(y_{ij}) = \begin{cases} \beta_0 + \beta_1 \, t_{ij}^{m_1}, & if \quad m_1 \neq 0 \\ \beta_0 + \beta_1 \log(t_{ij}), & if \quad m_1 = 0 \end{cases} \qquad (5.8)$$

The slope of the tangent line for the first order FP model can be obtained by differentiating (5.8) with respect to t_{ij} and evaluating it at $t_{ij} = t_{ia}$.

$$\gamma_{t_{ia}} = \begin{cases} m_1 \beta_1 \, t_{ia}^{m_1 - 1}, & if \quad m_1 \neq 0 \\ \dfrac{\beta_1}{t_{ia}}, & if \quad m_1 = 0 \end{cases} \qquad (5.9)$$

For $m_1 = 1$, the slope of the tangent line is β_1. This indicates that we have straight line curve and the slope of the tangent line is the same as the slope of the function. For $m_1 \neq 1,$ time appears in the derivative indicating the nonlinearity of the first order FP.

The second order FP can be obtained from (5.3) by letting p=2. Applying the expected value notation to the response, we have

$$E\left(y_{ij}\right)=\begin{cases}\beta_0 + \beta_1\, t_{ij}^{(m_1)} + \beta_2\, t_{ij}^{(m_2)}, & if \quad m_1 < m_2 \\ \beta_0 + \beta_1\, t_{ij}^{(m_1)} + \beta_2\, t_{ij}^{(m_1)} \log\left(t_{ij}\right), & if \quad m_1 = m_2 \end{cases} \qquad (5.10)$$

The round bracket notation is the transformation in equation (5.4). The slope of the tangent line ($\gamma_{t_{ia}}$) is dependent on different values of m_1 and m_2. For instance, the slope of the tangent line ($\gamma_{t_{ia}}$) when $m_1 < 0 \quad and \quad m_2 = 0$ is

$$\gamma_{t_{ia}} = m_1\, \beta_1\, t_{ia}^{m_1 - 1} + \frac{\beta_2}{t_{ia}} \quad . \qquad (5.11)$$

Once $m_1 = m_2$, the slope of the tangent line is given by

$$\gamma_{tia} = \beta_1\, f_1'\left(t_{ia}\right) + \beta_2 \left[\frac{f_1\left(t_{ia}\right)}{t_{ia}} + f_1'\left(t_{ia}\right) \times \log\left(t_{ia}\right) \right] \quad . \qquad (5.12)$$

The slope of the tangent line is a function of time as seen in equations (5.9), (5.11) and (5.12). We may plot the slope of the tangent ($\gamma_{t_{ii}}$) against time to demonstrate how the tangent slope varies with time.

5.3 Selection of Fractional Polynomial Models

In section (5.2), we have seen that FPs can have different order. A number of possibilities are available for the selection of the exponents (m_b). The best fitting fractional polynomial model needs to be selected using some appropriate model selection criteria. There are numerous ways of choosing fractional polynomial models. The choice of a particular model may be due to the fact that

- the FP transformation has been used in previous research
- the model provides appropriate curve shape for the data under study
- the model is best fit model based on some fit indices

Selection of FPs in the context of the LMM is complicated by the fact that the fractional polynomials influence both the fixed effects structure and the random effects structure through X_i and Z_i respectively (Long and Ryoo, 2010). Consequently, one has to keep the random effects constant, while selecting the fixed effects part of the fractional polynomial. In this thesis, the selection of mean structure was made using **mpf** package in R. This package is a collection of R functions targeted at the use of fractional polynomial models for modelling the influence of continuous covariates on the outcome of regression models. It combines backward elimination with systematic search for a ‘suitable’ transformation to represent the influence of each continuous covariate on the response variable (Benner, 2010). The test procedure used by **mpf** package in R is called closed testing procedure.

5.4 Selection of Mean Structure and Model Formulation

The plot of an individual tree’s stem radius (Figure 2.2) and the Loess smoothed curve (Figure 2.3) suggest that the relationship between the radial measure and tree age is curved. The **mfp package**, discussed in section (5.3), was used to select the mean structure. The best fitting fractional polynomial curve for the current data is found to be the second order fractional polynomial with powers $m_1 = 0.5$ and $m_2 = 1$. That is the linear term plus a square root of time. The preliminary graphical analyses also indicated that the intercept and growth patterns were different for different trees (Figure 2.2). Therefore, having a different slope for each tree leads to subject-specific regression coefficients, which represent the random effect in the mixed model. The provisional model for stem radial growth as a function of time (tree age in weeks) is:

$$Y_{it'} = \beta_{0i} + \beta_{1i} t'^{1/2} + \beta_{2i} t' + \varepsilon_{it'} \qquad (5.13)$$

$$i = 1, 2, \ldots 18$$

The dendrometre measurements began when the tree was about 39 weeks of age. The actual age, t, of the tree differs from the dendrometre age, denoted by t', by 39 weeks. In other words, $t' = t - 39$, where: t' and t are the dendrometre and actual ages, respectively, of tree i. $Y_{it'}$ is the radial measure of the ith tree at age t'_i, β_{1i} is the coefficient of square root of time effect for the ith tree, β_{2i} is the coefficient of time effect for the ith tree, $\varepsilon_{it'}$ is the mean zero deviation which represents the within-tree variability, β_{0i} represents the mean radial size of tree i at the beginning of dendrometre measurements, that is when $t' = 0$.

The values β_0, β_1 and β_2 are the average intercept, coefficients of square root of time and linear time effects, respectively, of the population. After correcting for the effect of individual characteristics the individual coefficients can then be expressed as:

$$\beta_{0i} = \beta_0 + b_{0i}, \qquad \beta_{1i} = \beta_1 + b_{1i}, \qquad \beta_{2i} = \beta_2 + b_{2i}$$

For the ith tree, the terms b_{0i}, b_{1i} and b_{2i} represent the random deviations of the intercept, coefficients of square root of time and time, respectively, from the corresponding population parameters β_0, β_1 and β_2. Therefore, model (5.13) can be rewritten as:

$$Y_{it'} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) t'^{1/2} + (\beta_2 + b_{2i}) t' + \varepsilon_{it'} \qquad (5.14)$$

The matrix form of model (5.14) is

$$\mathbf{Y}_i = \mathbf{Z}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon} \qquad (5.15)$$

Where

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ . \\ . \\ . \\ Y_{in_i} \end{bmatrix} \qquad \mathbf{Z}_i = \begin{bmatrix} 1 & \sqrt{t_{i1}} & t_{i1} \\ 1 & \sqrt{t_{i2}} & t_{i2} \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & \sqrt{t}_{in_i} & t_{in_i} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \qquad \mathbf{b}_i = \begin{bmatrix} b_{0i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ . \\ \vdots \\ \varepsilon_{in_i} \\ \\ \end{bmatrix}$$

Model (5.15) is the same as the linear mixed model given in (5.1) and satisfies the assumptions indicated in (5.1). The random effects b_i are assumed to be normally distributed with mean vector $\mathbf{0}$ and (3×3) covariance matrix $\mathbf{D}$, where $\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \\ d_{31} & d_{32} & d_{33} \end{bmatrix}$. Likewise the vector of residuals $\boldsymbol{\varepsilon}_i$ is assumed to be normally distributed with mean vector $\mathbf{0}$ and covariance matrix $R_i = \sigma^2 I_{n_i}$. Assuming the random effects and error terms are independent, the marginal distribution for the vector of responses of the ith tree $\mathbf{Y}_i$ is normally distributed with mean vector $\mathbf{Z}_\mathbf{i}\,\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}_\mathbf{i} = \mathbf{Z}_\mathbf{i}\mathbf{D}\,\mathbf{Z}_\mathbf{i}' + R_i$.

5.5 Estimation for LME Models

Although, various methods of parameter estimation have been used for linear mixed models, the most commonly used methods are maximum likelihood and restricted maximum likelihood. Equation (5.1) can be expressed as

$$y_i\,|b_i \;\sim MVN\left(X_i\beta_i + Z_i\,b_i\,,\;R_i\;\right),\quad b_i \sim MVN\;\left(0,\;D\right) \qquad (5.16)$$

It is therefore, called a hierarchical model, in which a conventional density of y_i follows a multivariate normal. This model implies the marginal model given below (Verbeke and Molenberghs, 2000).

$$y_i \sim MVN\left(X_i\beta_i,\;Z_i\,D\,Z_i' + R_i\right) \qquad (5.17).$$

Let α denote the vector of all variance and covariance parameters (variance components) in $V_i = Z_i\,D Z_i' + R_i$, that is α contains all different elements in D matrix and all parameters in R_i. Suppose $\theta = (\beta\,',\;\alpha\,')'$ be the vector of all the parameters in the marginal model (5.16) for y_i .

The marginal likelihood function is given by

$$L_{ML}(\theta) = \prod\nolimits_{i=1}^{N}\left\{\,(2\,\pi)^{-\frac{n_i}{2}}\left|V_i(\alpha)\right|^{-\frac{1}{2}}\times \exp^{\left(-\;\frac{1}{2}(y_i\;-X_i\,\beta)'\;V_i^{-1}(\alpha)\;\;(y_i - X_i\,\beta)\;\right)}\right\} \qquad (5.18)$$

The classical inference approach is based on estimators obtained by maximizing (5.18) with repect to θ. There are two conditions about α, known and unknown.

Condition 1: Assume α to be known: Differentiating $\ln(L_{ML}(\theta))$ with respect to β gives

$$\frac{\partial \ln(L_{ML})}{\partial\beta} = \sum_{i=1}^{N}\left(X_i' V_i^{-1} y_i - X_i V_i^{-1} X_i \beta\right) \qquad (5.19)$$

Equating (5.19) to zero, and solving the resultant equation for β gives the maximum likelihood estimator of β as

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i' W_i y_i \qquad (5.20)$$

where $W_i = V_i^{-1}$.

Condition 2: Assume α is unknown: when an estimate of $\hat{\alpha}$ is available, we can set $\hat{V}_i = V_i(\hat{\alpha}) = \hat{W}_i^{-1}$, and estimate β by using (5.19) replacing W_i by $\hat{W}_i$.

Maximum likelihood (ML) and restricted maximum likelihood (REML) are the two commonly used methods for obtaining $\hat{\alpha}$.

The maximum likelihood estimator of α can be obtained by maximizing the $L_{ML}\left(\alpha,\ \hat{\beta}(\alpha)\right)$ given in (5.18) with respect to α, after β is replaced by (5.20).

The REML estimator for variance components α is obtained from maximizing the likelihood function of a set of error contrasts (Verbeke and Molenberghs, 2000), $U = K'Y$, where K is $n \times (n - p)$ full rank matrix with columns orthogonal to the columns of X matrix. Then we combine all models $y_i \sim N\left(X_i\beta, V_i\right)$ into one model $y \sim N\left(X\beta, V\right)$ where

$$y = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_N \end{pmatrix}, \qquad X = \begin{pmatrix} X_1 \\ . \\ . \\ . \\ X_N \end{pmatrix}, \qquad V(\alpha) = \begin{pmatrix} V_1 & . & . & . & 0 \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & V_N \end{pmatrix}$$ so we can obtain,

$$U = \begin{pmatrix} y_1 - y_2 \\ y_2 - y_3 \\ . \\ . \\ . \\ y_{N-2} - y_{N-1} \\ y_{N-1} - y_N \end{pmatrix} = K'Y \sim N\left(0,\ K'V(\alpha)K\right)$$

Then the MLE of α, which is based on U is called REML estimate, and denoted by $\hat{\alpha}_{REML}$. Similarly, resulting estimate $\beta(\hat{\alpha}_{REML})$ is for $\hat{\beta}_{REML}$.

Both $\hat{\alpha}_{REML}$ and $\hat{\beta}_{REML}$ can be obtained from maximizing (5.21) with respect to all parameters simultaneously (α and β).

$$L_{REML}(\theta) = \left| \sum_{i=1}^{N} X_i' \; W_i(\alpha) \; X_i \right|^{-\frac{1}{2}} L_{ML}(\theta) \qquad (5.21)$$

Here, note that $L_{REML}(\theta)$ is not the likelihood of the original data (Wang, 2012).

Both ML and REML are based on the likelihood principle which leads to important properties such as consistency, asymptotic normality, and efficiency. However, REML yields less biased estimators for many special cases (Verbeke and Molenberghs, 1997).

5.6 Dealing with Heterogeneity in Linear Mixed Models

When the assumption $(R_i = \sigma^2 I_{n_i})$ in equation (5.1) is violated, the resulting model fit may not have correct standard errors. The F-statistics may no longer be distributed as F and the t-statistic also may not follow a t-distribution. This invalidates our statistical significance tests. The assumption of uncorrelated, homoscedastic within group errors can be relaxed by introducing heteroscedastic models. In this section, we will see how to fit the extended LMM by allowing heteroscedastic and correlated within group errors. The assumptions for model (5.1) will be modified to $\varepsilon_i \sim N(0,\ \Sigma_i)$ and $b_i \sim N(0, D)$.

Variance functions are used to model the variance structure of the within group errors using covariates. A detailed list of standard variance functions are presented in Pinheiro and Bates (2000). The description of some of these functions is given as follows.

The fixed variance structure (*varFixed*): This class represents a variance function with no parameter and a single variance covariate being used, with the within group variance known up to a proportionality constant. Suppose it is known that the within group variance increases linearly with time (t_{ij}) then the variance of the residuals is given by :

$$\text{var}\left(\varepsilon_{ij}\right) = \sigma^2 t_{ij}$$

This corresponds to the variance function, $g(t_{ij}) = \sqrt{t_{ij}}$. This variance structure allows larger variance for larger values of t_{ij} .

The VarIdent variance structure *(varIdent)*: This class represents a variance model with different variances for each level of the stratification variable s, taking values in the set $\{1, 2, \ldots\ S\}$, is $\text{var}\left(\varepsilon_{ij}\right) = \sigma^2 \delta^2{}_{ij}$.

This corresponds to the variance function, $g\,(S_{ij} \quad \delta\,) = \delta_{S\,ij}$. That means we assume a different spread per stratum$\left(\varepsilon_{ij} \sim N\left(0,\, \sigma_j^2\right)\right)$. This variance model uses S+1 parameters to represent S variances, and therefore is not identifiable. To achieve identifiably, we need to impose some restriction on the variance parameter, δ. We use$\delta = 1$, so thatδ_l, $l = 2,\ \ldots\ S$ represent the ratio between the standard deviations of the l^{th} stratum and the first stratum. By definition, $\delta_l > 0$, $l = 2,\ \ldots\ S$ (Pinheiro and Bates, 2000).

The varPower variance function *(varPower):* The variance model represented by this class is

$\operatorname{var}\left(\varepsilon_{ij}\right) = \sigma^2 \left|v_{ij}\right|^{2\delta}$, corresponding to the variance function $g\left(v_{ij},\ \delta\right) = \left|v_{ij}\right|^{\delta}$, which is the power of the absolute value of the variance covariate (v_{ij}). This class of variance function should not be used with variance covariates that may assume the value zero.

The varExp Variance structure *(varExp):* This structure models the variance of the residuals as σ^2multiplied by an exponential function of the variance covariate and unknown parameter δ.

The variance model represented by this class is

$\operatorname{var}\left(\varepsilon_{ij}\right) = \sigma^2 \exp^{\left(2\,\delta\, v_{ij}\right)}$, corresponding to the variance function $g\left(v_{ij},\ \delta\right) = \exp^{\left(\delta\, v_{ij}\right)}$ which is an exponential function of the variance covariates. The parameter δ is not restricted so that the variance function can model cases where the variance increases or decreases with the variance covariate. There are no restrictions on the variance covariate, which, in particular, may take the value zero (Pinheiro, and Bates, 2000).

The VarConstPower *(VarConstPower):* This variance structure is the constant plus power of the variance covariate function.

The variance model is defined as $\text{var}\left(\varepsilon_{ij}\right) = \sigma^{2}\left(\delta_{1} + \left|v_{ij}\right|^{\delta_{2}}\right)^{2}$, corresponding to the variance function

$g\left(v_{ij}, \delta\right) = \delta_{1} + \left|v_{ij}\right|^{\delta_{2}}$. The constant (δ_1) is restricted to be positive and δ_2 is unrestricted.

The varComb variance structure *(varComb):* This variance structure can be considered as a combination of *varIdent* and *varExp.* It can be given by

$\text{var}\left(\varepsilon_{ij}\right) = \sigma^{2}\delta_{1}^{2}, s_{ij} \exp\left(2\delta_{2}v_{ij}\right) = \sigma^{2} g_{1}^{2}\left(s_{ij}, \delta_{1}\right) g_{2}^{2}\left(v_{ij}, \delta_{2}\right)$, corresponding to variance functions $g_{1}\left(s_{ij}, \delta_{1}\right) = \delta_{1} s_{ij}$ and $g_{2}\left(v_{ij}, \delta_{1}\right) = \exp^{\left(\delta_{2} v_{ij}\right)}$ where s_{ij} and v_{ij} are variance covariates.

5.7 Correlation Structures for Modelling Dependence

In model (5.1), it was assumed that the within group error terms are independent and have constant variance. The assumption of constant variance can be relaxed by using different variance functions which were discussed in section (5.6). In this section different approaches of handling the dependence of the within error terms, were presented.

The dependence of within error terms is modelled using correlation structures. In time series data, serial correlations are used to model dependence. In the context of a linear mixed model, serial correlation captures the phenomenon that correlation structure within a subject depends on the time lag between two measurements. Jones (1993) discussed the serial correlation structures in detail for linear mixed effects models. The general serial correlation model is defined as

$cor\left(\varepsilon_{ij}, \ \varepsilon_{ij'}\right) = h(\rho)$, where $h(.)$ indexes autocorrelation function and ρ is a vector correlation. The description of some of the most common serial correlations used in practice is given below.

Compound Symmetry: This structure assumes equal correlation among all within group error of same subject (entity). It is the simplest serial correlation structure. The corresponding correlation model is

$cor(\varepsilon_{ij}, \varepsilon_{ik}) = \rho,$ for all $j \neq k$.

General Correlations Structure: This unstructured correlation structure. This structure represents the direct opposite in complexity to the compound symmetry structure. Each correlation is shown by a different parameter, the correlation function is

$$h(\rho) = \rho_k \ , \ k = 1, \ 2\dots .$$

The general correlation structure may be useful when we have few observations per subject.

Auto Regressive (AR): Box et al. (1994) defined the family of correlation structures which comprises diverse classes of linear stationary models. These are autoregressive models, moving average models, and a mixture of autoregressive-moving average models.

An auto regressive model of order p which is denoted by $AR(p)$ states that ε_t is the linear function of the previous "p" values of the series plus an error term (μ_t).

$\varepsilon_t = \phi_1 \varepsilon_{t-1} + \dots . + \phi_p \varepsilon_{t-p} + \mu_t \quad |\phi| < 1$ where ε_t stands for observation at time t, μ_t stands for a noise term with $E(\mu_t) = 0$ and assumed independent of the previous observations.

There are p-correlation parameters in AR(p) model, given by $\Phi = (\phi_1, \phi_2 \ldots \phi_p)$. The AR(1) model is the simplest and one of the most important autoregressive model. The correlation function of AR (1) model is given by

$$h(k, \phi) = \phi^{k}, \quad k = 0, 1 \ldots .$$

According to Pinheiro and Bates (2000) for autoregressive models of order greater than 1, the correlation function was defined as

$$h(k, \phi) = \phi_1 h(|k-1|, \phi) + \ldots + \phi_p h(|k-p|, \phi), \quad k = 1, 2, \ldots .$$

Moving Average Correlation

Moving average correlation models assume that the current observation is a linear function of independent and identically distributed noise terms.

$$\varepsilon_t = \theta_1 \alpha_{t-1} + \ . \ . \ . + \theta_q \alpha_{t-q} + \alpha_t$$

The moving average of order q is denoted by $MA(q)$. There are q correlation parameters in a $MA(q)$ model. The correlation function for a $MA(q)$ model is given by

$$h(k, \theta) = \begin{cases} \dfrac{\theta_k + \theta_1\theta_{k-1} + \ldots + \theta_{k-q}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2} & k = 1, \ldots q \\ 0 & k = q+1, q+2 \ldots \end{cases}$$

Mixed Autoregressive Moving Average Models (ARMA)

The combination of autoregressive model and moving average model gives us the ARMA models.

$$\varepsilon_t = \sum_{i=1}^{p} \phi_i \varepsilon_{t-i} + \sum_{j=1}^{q} \theta_j \alpha_{t-j} + \alpha_t$$

There are $p+q$ correlation parameters in ARMA (p, q) model. These are the p autoregressive parameters and the q moving average parameters. By convention ARMA (p, 0) is the same as AR (p) and ARMA (0, q) is MA (q). This shows that both autoregressive and moving average models are special cases of ARMA (p, q) models.

The likelihood ratio test cannot be used to differentiate between models with different covariance structure, if these are not nested to each other. On the other hand, information criteria can be used to select between such models. Two regularly used criteria are Akaike Information Criterion (AIC) [Sakamoto et al, 1986] and Bayesian Information Criteria (BIC) [Schwartz, 1978]. These are model comparison criteria evaluated as

$$AIC = -2 \log likelihood + 2n_{par},$$
$$BIC = -2 \log likelihood + n_{par} \log(N),$$

Where n_{par} stands for the total number of parameters in the model and N stands for total number of observations used to fit the model. We prefer the model with the smallest AIC, when comparing two or more models fitted to the same data. Similarly, when using BIC, we prefer the model with the lowest BIC.

5.8 Inference for Marginal Model Parameters

Usually, inference on the parameters of a fitted model is often a primary interest, due to the generalization of results from specific sample to general population from which the sample was taken (Verbeke and Molenberghs, 2000). As already seen in section (5.5), the vector β of fixed effects is estimated by

$$\hat{\beta}(\alpha) = \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \sum_{i=1}^{N} X_i' W_i y_i \qquad (5.22)$$

where $W_i = V_i^{-1}(\alpha)$, the unknown α of variance component is replaced by REML or ML estimate. Under the marginal model (5.17), and conditionally on α , $\hat{\beta}(\alpha)$ follows a multivariate normal distribution with mean vector β and variance covariance matrix

$$\begin{aligned}
\operatorname{var}(\hat{\beta}) &= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i \operatorname{var}(y_i) W_i X_i\right)\left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \\
&= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i V W_i X_i\right)\left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \\
&= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \left(\sum_{i=1}^{N} X_i' W_i X_i\right)\left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1} \qquad (5.23) \\
&= \left(\sum_{i=1}^{N} X_i' W_i X_i\right)^{-1}
\end{aligned}$$

To construct confidence intervals or to test hypotheses about β , we can use the ML estimate $\hat{\beta}$ and its estimated covariance matrix. The estimate of the covariance matrix in (5.23) can be obtained by replacing α by its ML or REML estimator. For individual parameter β_j in vector β , j= 1, 2 … p, the different confidence limits can be obtained from approximating the distribution of

$$\frac{(\hat{\beta}_j - \beta_j)}{\hat{S}.e(\hat{\beta}_j)}$$ by a standard normal distribution.

The test statistic $Z = \frac{\hat{\beta}_i}{\sqrt{\hat{V}\left(\hat{\beta}_j\right)}}$ can be used to test the null hypothesis

$H_o : \beta_j = 0$ for j=1, 2... p where $\hat{V}\left(\hat{\beta}_j\right)$ denotes the diagonal element for the estimator of (5.23) corresponding to β_j.

In general, it may be of interest to obtain confidence intervals and to test hypotheses about linear combinations of the components of β. For any vector or matrix of known weights L, a test for the hypothesis

$$H_o : L\beta = 0, \textit{ versus} \qquad H_a : L\beta \neq 0 \qquad (5.24).$$

The estimate of $L\beta$ is given by $L\hat{\beta}$. The sampling distribution of $L\hat{\beta}$ is multivariate normal with mean $L\beta$ and covariance $L \operatorname{var}(\hat{\beta})L'$. This implies

$\left(\hat{\beta} - \beta\right)' L'\left[L \ \hat{V}\left(\hat{\beta}\ \right)L' \quad \right]^{-1} L\left(\hat{\beta} - \beta\right)$ asymptotically follows a chi-square distribution with rank (L) degrees of freedom. Therefore, the test statistic,

$W^2 = \left(L\hat{\beta}\right)' \left[L\,\hat{V}\left(\hat{\beta}\right)L' \quad \right]^{-1} L\hat{\beta}$, which has a chi-square distribution with degrees of freedom rank (L) is used to test the hypothesis in (5.24). However, both the Wald test and chi-square tests are based on large sample properties of the sampling distribution of the ML estimate of β.

The Wald test statistics are based on estimated standard errors which underestimate the true variability in $\hat{\beta}$ because they do not take into account the variability introduced by estimating α.(Dempster et al., 1981; Verbeke and Molenberghs, 2000). This problem of downward bias can be solved by using approximate t-and F statistics for testing hypotheses about β. This can be done:

i) by approximating the distribution of, $\frac{\left(\hat{\beta}-\beta\right)}{\hat{S}.e\left(\hat{\beta}_j\right)}$ by an appropriate t-distribution

ii) by testing the general linear hypothesis in (5.24) using an F-approximation to the distribution of

$$F=\frac{\left(\hat{\beta}-\beta\right)'L'\left[L\left(\sum_{i=1}^{N}X_i'V_i^{-1}(\hat{\alpha})X_i\right)^{-1}L'\right]L\left(\hat{\beta}-\beta\right)}{rank\quad(L)}$$

The numerator degrees of freedom equals rank (L). The denominator degrees of freedom need to be estimated from the data. The degrees of freedom for t-distribution also need to be estimated from the data. Several estimation methods are available which might lead to different results. However, in longitudinal data analysis different individuals contribute independent information, which results in numbers of degrees of freedom which are typically large enough, whatever estimation method is used, to lead to very similar p-values (Verbeke and Molenberghs, 2000).

5.9. Results of Fitting the Fractional Polynomial Model

The ordinary least square (OLS) regression model of (5.13) was fitted and the residuals were examined. The box plots of these residuals by tree are indicated in Figure 2.6. The residuals corresponding to the same tree tend to have the same sign. This indicates the demand for a “tree effect” in the model, which is indeed the motivation for mixed effects models. The next step in the model building process is to choose which of the curve components (the intercept, time or square root of time) should have a random component to account for the between tree variation. The “lmList” function in R statistical software was used to fit the model to individual trees.

The 95% confidence interval for the parameters of the model was plotted at the individual tree level (Figure 5.1). The tree specific intervals do not overlap for the intercept or coefficient of time and nor for the coefficient of the square root of time. Therefore, these individual confidence intervals give a clear indication that a random effect is needed for tree to tree variability in the intercept, coefficients of time and square root of time. Moreover, to facilitate comparison among the distributions of intercepts, linear time and square root of time across the two clones, parallel boxplots for the coefficients were produced (Figure 5.2).

At the beginning of the trial, when measurements were first initiated, the GC clone showed a higher average radial growth compared to the GU clone (Figure 5.2). On the other hand, the average coefficients relating the stem radial measures to the linear effect of age and square root of age effect were larger for the GU clone than for the GC clone. Therefore, these graphical methods (Figures 5.1 and 5.2) suggest that a model with different intercept and time coefficients for each clone needs to be considered.

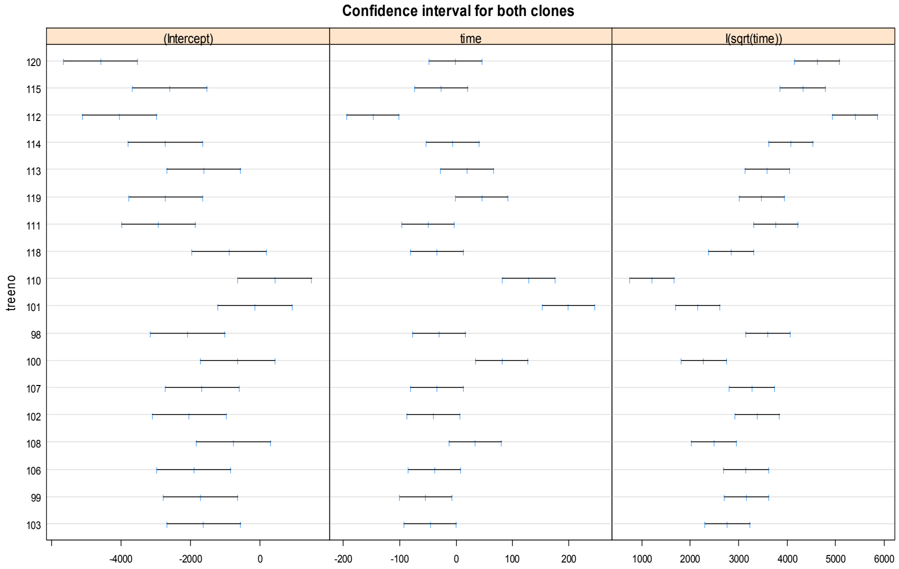

Figure 5 . 1 Ninety-five percent confidence intervals for intercepts, coefficient of time and the coefficient for the square root of time for each tree in the dataset**.**

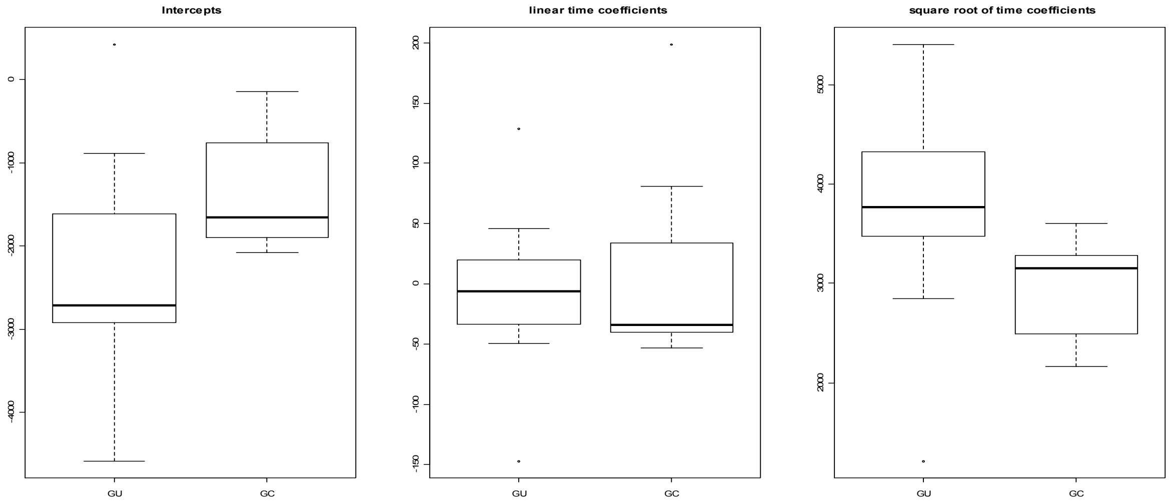

Figure 5. 2 Box plots of intercepts, coefficients of time and coefficients of the square root of time for the regressions of radial growth on age of a tree for GU and GC clones.

With the objective of selecting the best random effects model, a linear mixed model was fitted, assuming the diagonal elements in R_i are all equal and the off-diagonal elements are zero. Therefore, the variance of the response vector $Y_{it'}$ depends on time only through the component $\mathbf{Z}_\mathbf{i}\mathbf{D}\mathbf{Z}_\mathbf{i}'$. A hierarchical test procedure was followed to see if any of the random effects could be removed from the model. Hence the test begins with the inquiry as to whether or not the square root of time effect differs between trees. The formulation of the test of hypothesis at a specified α-level of significance is:

Ho: d13 = d23 =d33 =0 against the alternative H_a: at least one of the d_{i3} is different from 0, i=1, 2, 3.

In the above d13, d23, d33 are the covariance of random intercept and square root of time random effect, the covariance of time coefficients and square root of time coefficients and the variance of square root of time random coefficients respectively. The classical likelihood based inference cannot be applied for testing the above null hypothesis since the null hypothesis (d33 =0) is on the boundary of the parameter space. To avoid this boundary value problem the asymptotic mixture of chi-squared distributions for the likelihood ratio test statistics was applied. This statistic is the difference of minus twice the logarithm of the likelihoods under the null and the alternative hypothesis. A large value of this difference rejects the null hypothesis and favours the alternative hypothesis, that there is a significant improvement in the fit when the extra random effect parameters are included.

The following random effect models were considered for testing:

Model 1: Intercept, time, square root of time

Model 2: Intercept, time

Model 3: Intercept, square root of time

Model 4: Time, square root of time

Model 5: Intercept only

The likelihood ratio test statistics based on restricted maximum likelihood (REML) together with the corresponding p-values are displayed in Table 5.1. The observed values of the test statistics are very large and yield p-values less than 0.0001. We conclude that the covariance structure should not be simplified by deleting any of the random effects from the model. The general positive definite matrix was used and the estimated covariance matrix is

$$\hat{D} = \begin{bmatrix} 1127600 & -889950 & 53266 \\ -889950 & 779260 & -48320 \\ 53266 & -48320 & 5754 \end{bmatrix}.$$

Table 5. 1 Likelihood ratio test for random effects using restricted maximum likelihood estimation

Random effects	*-2loglikelihood*	LR test Statistics	Comparison	Chi-square $\chi^2_{k_1:k_2}$	P-value
Model 1	20450.30	-	-	-	-
Model 2	20637.56	187.26	1 vs 2	$\chi^2_{2:3}$	< 0.0001
Model 3	20580.20	129.90	1 vs 3	$\chi^2_{2:3}$	< 0.0001
Model 4	20497.16	46.86	1 vs 4	$\chi^2_{2:3}$	< 0.0001
Model 5	21720.86	1270.56	1 vs 5	$\chi^2_{1:3}$	< 0.0001

The p-value is calculated by giving equal weight to a mixture of two chi-squared distributions with k_1 and k_2 degrees of freedom. That is

$$p - value = P(\chi^2_{k_1 k_2} \geq LR) = 0.5P(\chi^2_{k_1} \geq LR) + 0.5P(\chi^2_{k2} \geq LR)$$

The assumptions that the within-tree errors are normally distributed, are centred at zero and have constant variance were assessed. Initially, the box plot of residuals by group (tree) was considered (Figure 5.3). The residuals have zero mean as all centres are close to the vertical line. The variability of residuals is not exactly constant. The box plots in the upper part of Figure 5.3 appeared to have higher variability than the box plots in the lower part of the figure. To obtain a better impression of this pattern the plot of standardized residuals versus fitted values, by clone, were examined.

The plot of standardized residuals for the ***homoscedastic*** model is presented in Figure 5.4. The residual variability for the GU clone is larger than for the GC clone. Some outlying values are observed for some trees. A more general model to represent the radial growth data that allows different

variances by clone for the within-tree error was applied. Based on this heteroscedastic model by clone, several variance functions discussed in section (5.6) and dependence model of section (5.7) were considered for the variance of the within-tree error.

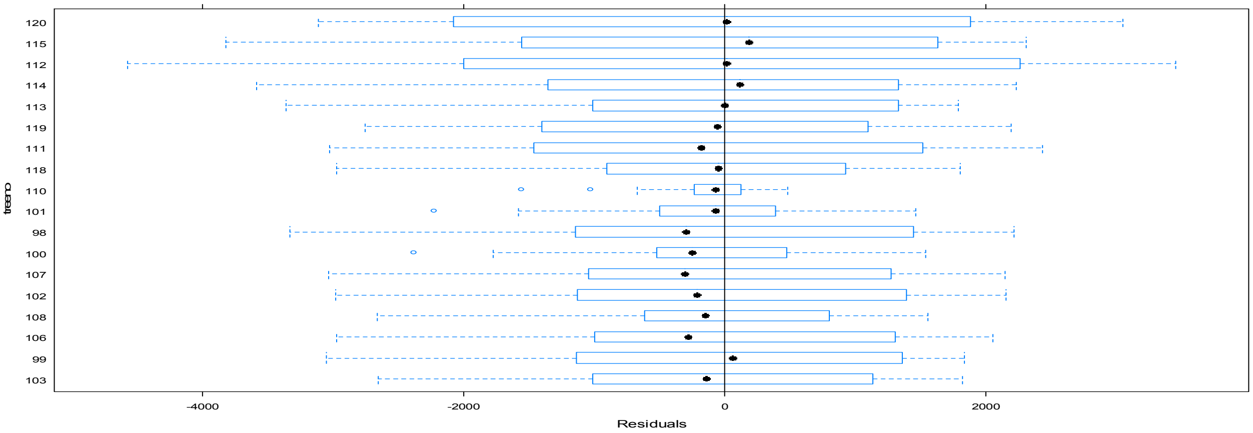

Figure 5. 3 Box plot of residuals by tree.

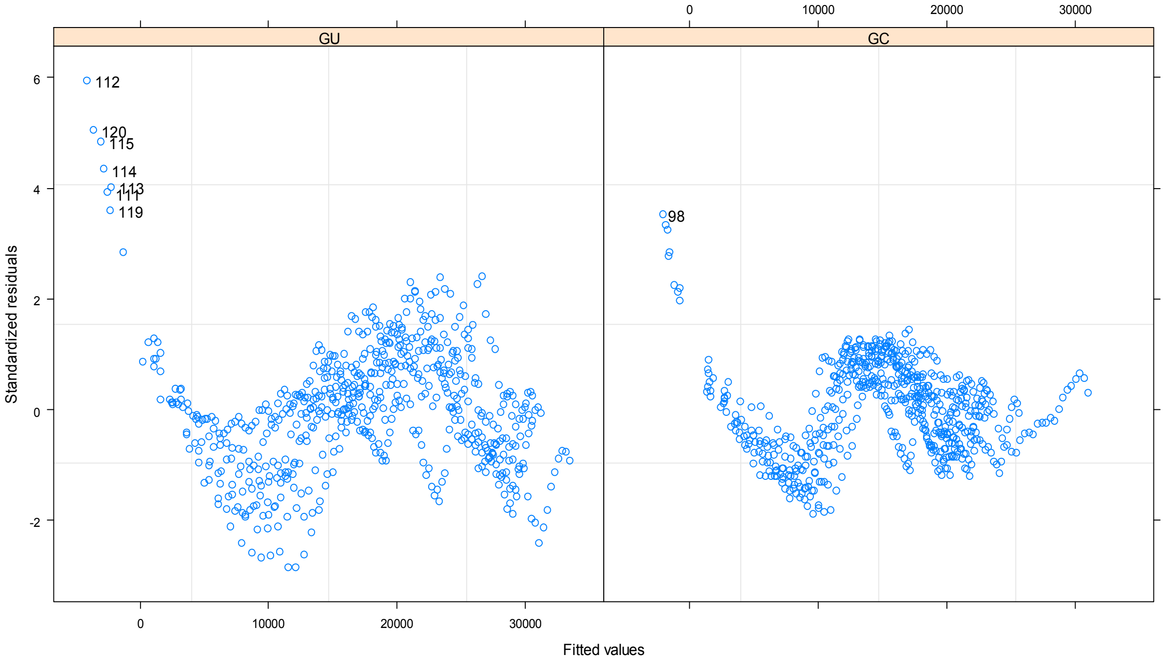

Figure 5. 4 Plot of standardized residuals for a *homoscedastic* model.

Among the models for which convergence is achieved, a variance which is an exponential function of time was found to be the best fit. That means the two clones had different variances and their variance function was a function of tree age. The estimated standard error for the GC clone is about 76% of that for the GU clone. The estimate for fixed effects is similar to the estimates of the ***homoscedastic*** model. The estimates of fixed effects are presented in Table 5.2.

Table 5. 2 Fixed effect estimates for heteroscedastic model

Effect	Value	Standard error	DF	t-value	P-value
Intercept	-5547.85	812.91	1220	-6.82	0.001
Time	-137.39	38.62	1220	-3.36	0.001
Clone(GC)	2738.96	1122.49	16	2.44	0.026
$\sqrt{time}$	5072.58	462.64	1220	10.96	0.001
$\sqrt{time}$ × Clone (GC)	-1514.95	648.76	1220	-2.33	0.019
Clone (GC) × time	84.26	54.17	1220	1.56	0.12

As seen from Table 5.2, the interaction of time effect with a clone was not significant and hence was removed from the model. The interaction between clone and the square root of time effect is significant. This indicates that the two clones have different coefficients for the square root of time. Therefore, the longitudinal growth of the GU clone is significantly higher than that of the GC clone.

The plots of the standardized residuals versus fitted values, by clone, were re-examined to assess the adequacy of the heteroscedastic model (Figure 5.5). The difference in variability of the residuals for the two clones has improved (less variability is observed). Some outlying observations are still observed for some trees.

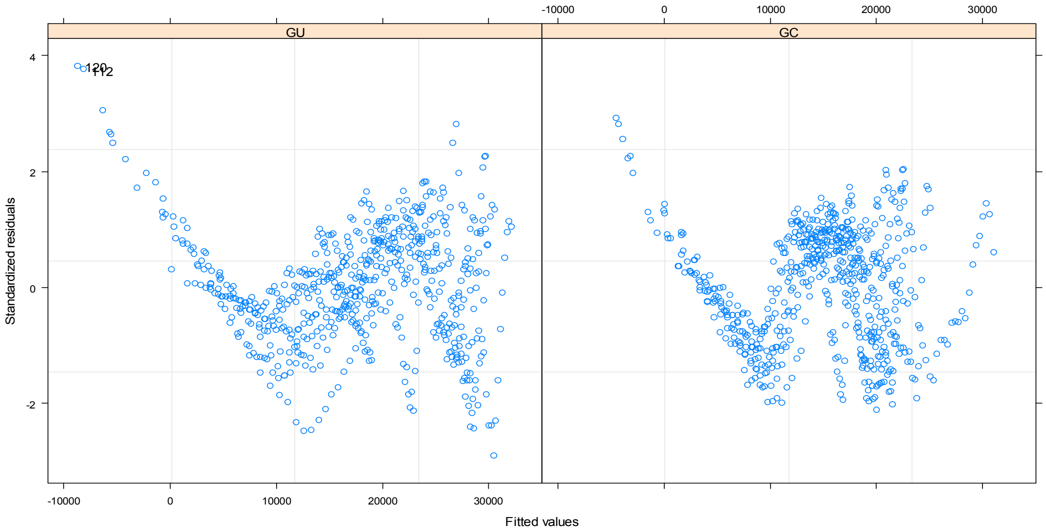

Figure 5.5 Plot of residuals versus fitted values by clone for a heteroscedastic model.

Overall the standardized residuals are small suggesting that the mixed effects model with heteroscedastic variance is successful in explaining the radial growth curves. The homoscedastic model and the heteroscedastic model are also compared using a formal test. The results of the formal tests are given in Table 5.3. The very small p-value of the likelihood ratio statistic confirms that the heteroscedastic model explains the data significantly better than the homoscedastic model.

Table 5.3 Test that compares homoscedastic model and heteroscedastic model

Model	df	AIC	LogLik test	Test	L.Ratio	P-values
Homoscedastic	11	20537.62	-10257.81			
Heteroscedastic	13	20069.17	-10021.58	1 vs 2	472.45	< 0.001

The assumption of normality for the within group errors was assessed using the normal probability plot of residuals. The normal probability plot of residuals is shown in Figure 5.6. Close examination of the behaviour of the two plots (see Zewotir and Galpin, 2004) shows that the normality assumption is plausible.

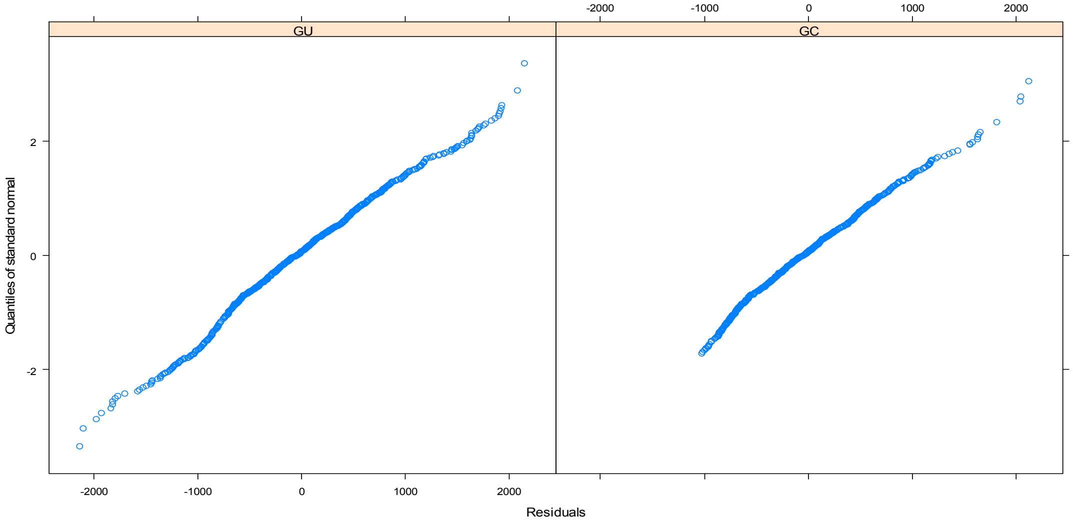

Figure 5. 6 Normal probability plot of residuals by clone.

The investigation of the marginal normality of the corresponding random effects was also made. The normal probability plot of the random effects is indicated in Figure 5.7. The assumption of normality seems reasonable for all three random effects.

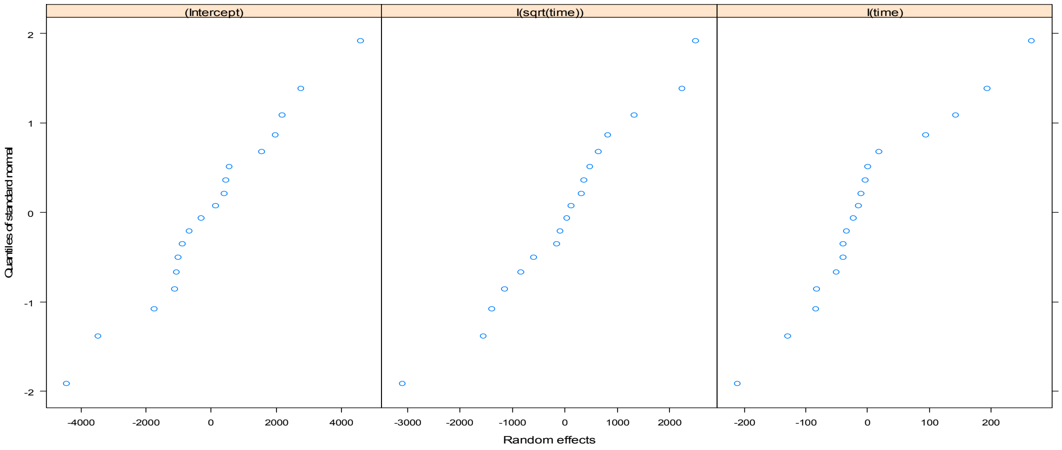

Figure 5 . 7 Normal probability plot of random effects.

The maximum likelihood estimates for the fixed effects as well as the variance components of the final heteroscedastic model are presented in Table 5.4.

Table 5. 4 Maximum likelihood estimates for the parameters of the fitted model

Effect	Parameter	Estimated Value
intercept	β_0	-4762.79
Time	β_1	-94.44
$\sqrt{time}$	β_2	4614.43
clone (GC)	β_3	1220.95
clone (GC) × $\sqrt{time}$	β_4	-618.06
var(b_0)	d_{11}	5212300
var(b_1)	d_{22}	1915800
var(b_2)	d_{33}	13544
cov(b_0, b_1)	$d_{12} = d_{21}$	-3084300
cov(b_0, b_2)	$d_{13} = d_{31}$	233920
cov(b_1, b_3)	$d_{23} = d_{32}$	-143460
$\hat{\sigma}^2$	GC clone	$2991626 \sqrt{t'_{ij}} \exp^{(-0.378 \sqrt{t'_{ij}})}$
	GU clone	$7188634 \sqrt{t'_{ij}} \exp^{(-0.438 \sqrt{t'_{ij}})}$

The fitted marginal model or the average profile of the radial measure at the age of t' for the two clones can be summarized as follows:

$$\hat{Y}_{t'} = -3541.84 - 94.44\, t' + 3996.37 \sqrt{t'} \quad \textit{for GC clone}$$

$$\hat{Y}_{t'} = -4762.79 - 94.44\, t' + 4614.43 \sqrt{t'} \quad \textit{for GU clone}$$

After fitting the selected model with proper covariance structure, evaluation of the final model is necessary. Therefore, in what follows we assess the goodness of fit of the final model. All trees included in the study were measured the same number of times. There was no dropout. The likelihood based analysis made in this study is justifiable. Some graphical techniques were applied to informally check whether the model fitted the data set well. Since the main objective of the study is to fit the mean structure of the data, it is necessary to compare the fitted and observed mean response profiles for radial growth. The Loess smoothing technique was applied to summarize the trend of average radial measure as a function of time. This technique estimates the underlying regression function without any restrictive parametric form. In addition to its use in assisting to choose the parametric models, it can also be used as a diagnostic tool by comparing the parametric and non-parametric fits. The superimposed fitted average profile on the smoothed Loess curves are indicated in Figure 5.8. The left panel of this figure compares Loess fit (smoothing parameter= 0.9) with the fitted average radial growth for the GC clone. In the plot we can see that the two fitted curves are very close to each other. The middle panel compares Loess fit (smoothing parameter=0.9) with the fitted average radial measure for the GU clone. The fitted average profile is very close to the smoothed Loess curve and the observed discrepancy is very minimal. The right hand panel of Figure 5.8 shows the fitted curve for both the GU and the GC clones. The plots indicate that the GU clone has a higher growth profile than the GC clone.

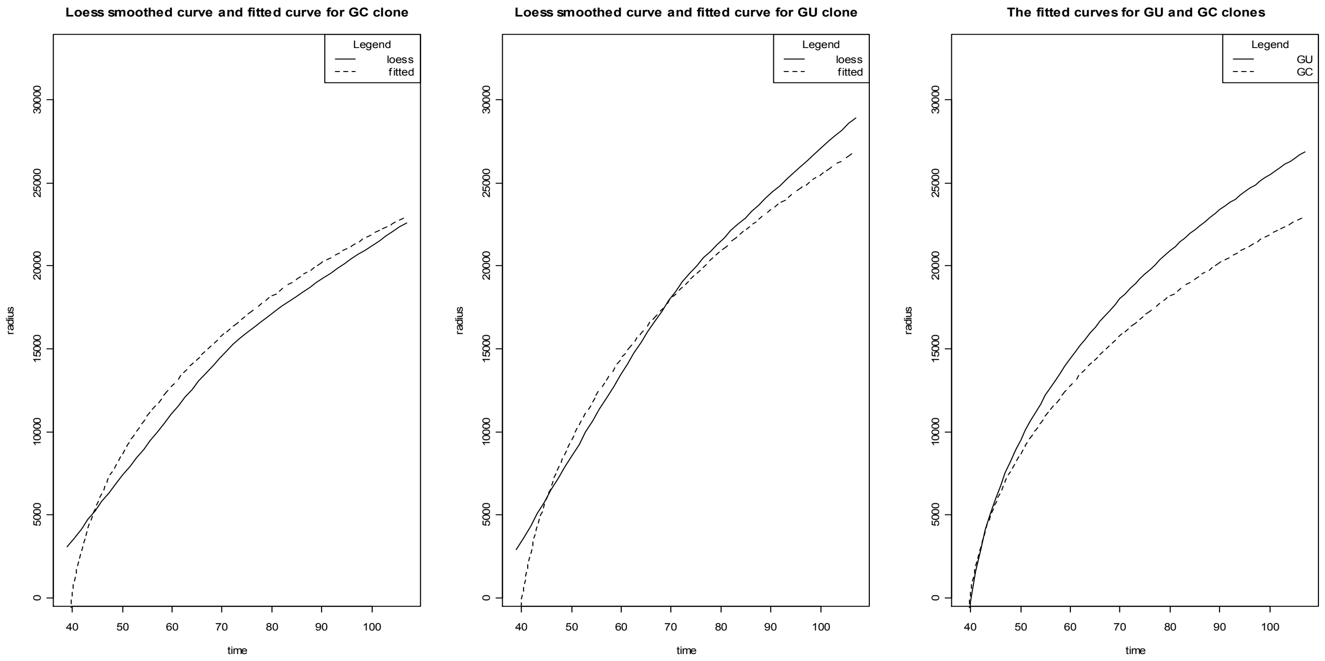

Figure 5. 8 The fitted average profiles of radial growth measures and the Loess smoothed curve (band width=0.9).

In addition, the adequacy of the fractional polynomial model, at individual tree level was checked. The plot of the augmented predictions, by tree, was used as an assessment for adequacy of the fractional polynomial model (Figure 5.9). The predicted values closely matched the observed radial growth measurements demonstrating the adequacy of the model.

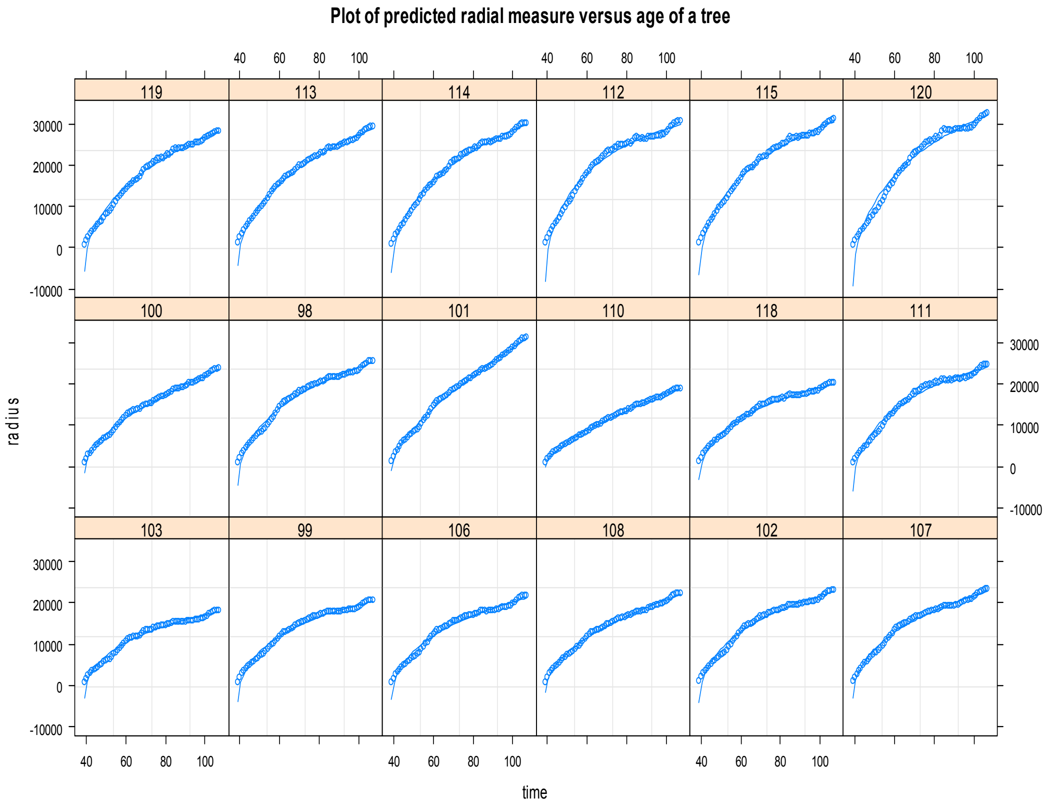

Figure 5. 8 Plot of predicted radial measure versus time by tree for the final model.

To assess overall measure for the goodness of fit of the first stage regression model, R^2_{meta} was used. R^2_{meta} is given by the following formula.

$$R^2_{meta} = \frac{\sum_{i=1}^{N} (SSTO_i - SSE_i)}{\sum_{i=1}^{N} SSTO_i}$$

Where $SSTO_i$ and SSE_i are the tree specific total and error sum of squares respectively. This quantity expresses what proportion of the total within subject variability can be explained by the first stage regression models (Verbeke and Molenberghs, 2000).

The overall coefficient R_{meta}^{2} of multiple determinations is equal to 0.99. This indicates that the model explains about 99% of the total within-tree variability. All tree specific coefficients, R_i^2, are greater than or equal to 0.98, suggesting that the first stage model fits the observed profiles reasonably well. From the fitted fractional polynomial model, given that time and the square root of time explain about 99% of the weekly stem radial growth, it will be interesting to study the impact of weekly climatic conditions on the weekly growth of the two clones. In the next section, the intriguing effect of climatic covariates on the current model will be presented.

5.10 The Effect of Climatic Variables

The fitted fractional polynomial models are extended to include the effect of climatic variables and their interaction with a clone. The effect of each climatic variable together with the interaction between clone and each climatic variable is considered in the modelling process. The results of the fixed effect estimates are presented in Table 5.5.

Table 5. 5 Fixed effect estimates for the model that includes the effect of weather variables.

Covariates	Value	Standard Error	t-value	p-value
Intercept	-5764.61	669.58	-8.6	0.0000
Time	45.45	24.26	1.87	0.0600
Clone (GC)	2085.67	538.27	3.87	0.0010
$\sqrt{time}$	3095.81	322.69	9.59	0.0000
Temperature	39.58	11.09	3.57	0.0004
Rainfall	3.72	0.96	3.86	0.0001
Relative humidity	15.40	4.98	3.09	0.0020
Solar radiation	2381.09	246.63	9.65	0.0000
Wind speed	818.52	67.27	12.17	0.0000
Clone $\times\sqrt{time}$	-612.89	281.70	-2.18	0.0298
Clone × Temperature	-48.91	14.18	-3.45	0.0006
Clone × Solar radiation	-669.98	302.85	-2.21	0.0271

The interaction effect of clone with each climatic variable is studied one by one. The hierarchical procedure is used to test for an additional parameter in the model. The clone is found to have significant interaction with temperature and solar radiation. The interaction of a clone with other climatic variables is not significant. Temperature appears to have an opposite effect on the radial growth of the two clones. The rest of the weather variables appear to have a positive effect on the stem radial growth. However, the above result has not considered the effect of season on the weather variables. The effect of weather variables might depend on season. The effect of weather variables on stem radius is considered after including season as one of the factors that determines stem radial growth. The results of the model that includes the effect of the season are presented in Table 5.6.

Table 5. 6 Fixed effect estimates for the model that includes the effect of the season.

Covariates	Value	Standard Error	t-value	p-value
Intercept	2623.54	1576.26	1.66	0.0963
Time	59.22	27.31	2.17	0.0303
Clone (GC)	2177.79	578.60	3.76	0.0017
$\sqrt{time}$	3014.48	357.36	8.44	0.0000
Temperature	33.46	24.57	1.36	0.1735
Rainfall	22.95	2.83	8.12	0.0000
Relative humidity	-50.15	10.83	-4.63	0.0000
Solar radiation	1274.96	285.45	4.47	0.0000
Wind speed	-371.84	163.67	-2.27	0.0233
Clone ›$\sqrt{time}$	-612.31	285.85	-2.14	0.0324
Clone × Temperature	-54.63	9.06	-6.03	0.0000
Clone × Solar radiation	-609.67	195.40	-3.12	0.0019
Season(Autumn)	-6983.10	1553.89	-4.49	0.0000
Season(Winter)	-13145.67	1537.28	-8.55	0.0000
Season(Spring)	-2281.52	2044.41	-1.12	0.2647
Temperature × Season(Autumn)	87.22	26.06	3.35	0.0008
Temperature × Season(Winter)	58.71	26.56	2.21	0.0270
Temperature × Season(Spring)	-8.41	28.26	-0.29	0.7658
Rainfall × Season(Autumn)	-16.84	3.60	-4.68	0.0000
Rainfall × Season(Winter)	-24.63	2.91	-8.47	0.0000
Rainfall × Season(Spring)	-21.31	3.26	-6.54	0.0000
Wind speed× Season(Autumn)	284.17	199.27	1.43	0.1541
Wind speed × Season(Winter)	730.49	180.05	4.06	0.0001
Wind speed × Season(Spring)	255.05	255.19	0.99	0.3178
Solar radiation× Season(Autumn)	-489.65	363.35	-1.35	0.1780
Solar radiation × Season(Winter)	6053.74	462.14	13.09	0.0000

Covariates	Value	Standard Error	t-value	p-value
Solar radiation × Season(Spring)	-459.05	386.76	-1.19	0.2355
Relative humidity × Season(Autumn)	50.57	12.33	4.10	0.0000
Relative humidity × Season(Winter)	92.67	11.79	7.85	0.0000
Relative humidity × Season(Spring)	36.24	18.23	1.99	0.0471

The results of Table 5.6, suggest that rainfall and solar radiation have a positive effect on stem radial growth during summer. The effect of temperature is negative for the GC clone while no significant effect of temperature is observed for the GU clone in summer. Wind speed and relative humidity appears to have a negative effect on the stem radial growth of both clones during summer.

In autumn, the effect of rainfall, temperature, relative humidity and solar radiation on stem radial growth appear positive for both clones. The effect of wind speed is negative on the stem radial growth for both clones in autumn. In winter temperature, relative humidity, solar radiation and wind speed have a positive effect on the stem radial growth of both clones. The effect of rainfall on stem radius appears negative for both clones during winter. In spring, rainfall and solar radiation have a positive effect on the stem radial growth for both clones. Relative humidity and wind speed have a negative effect on the stem radial growth for both clones in spring. The effect of temperature on stem radial growth is negative for the GC clone while no significant effect was observed for the GU clone in spring. From our results it is evident that some weather variables have a negative effect in one season and a positive effect in another season. For instance, temperature has a positive effect on stem radial growth of both clones in autumn and winter. On the other hand, a negative effect of temperature is observed in summer and spring for the GC clone, while no significant effect is observed for the GU clone.

The plots of the standardized residuals versus fitted values, by clone, were re-examined to assess the adequacy of the heteroscedastic model (Figure 5.10). The difference in variability of the residuals for the two clones has improved (less variability is observed). Overall the standardized residuals are small, suggesting that the mixed effects model with the effect of climatic covariates included is successful in explaining the radial growth curves.

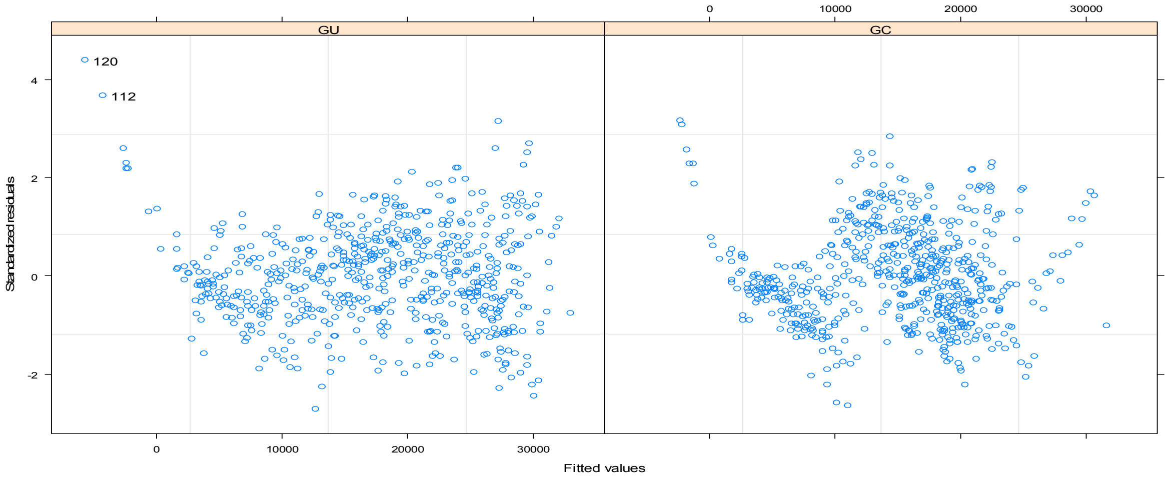

Figure 5 .9 Plot of residuals versus fitted values by clone for the final model

The assumption of normality for the within group errors was assessed using the normal probability plot of residuals. The normal probability plot of residuals is shown in Figure 5.11. Close examination of the behaviour of the two plots (see Zewotir and Galpin 2004) shows that the normality assumption is plausible.

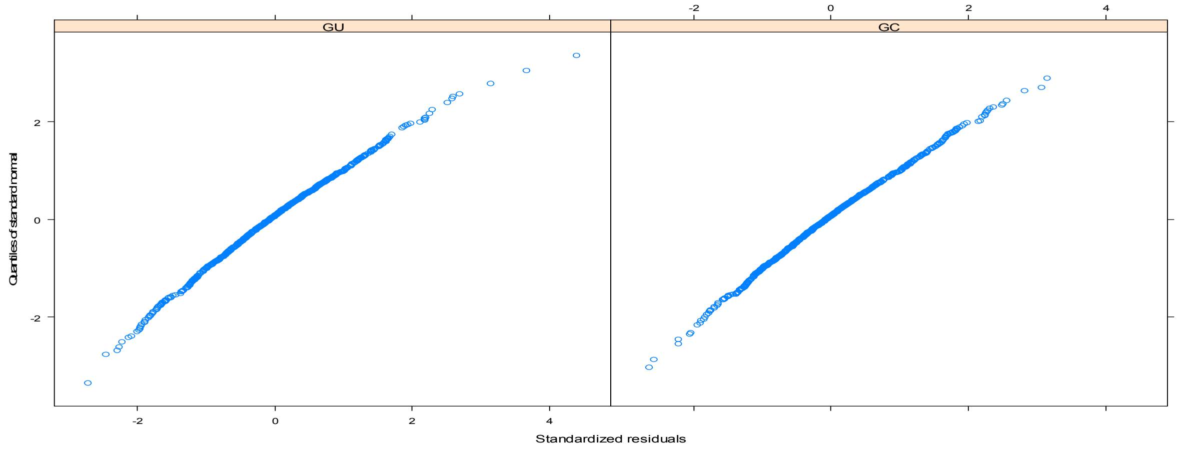

Figure 5. 10 Normal probability plots of residuals by clone for the final model.

The investigation of the marginal normality of the corresponding random effects was also made. The normal probability plot of the random effects is indicated in Figure 5.12. The assumption of normality seems reasonable for all three random effects.

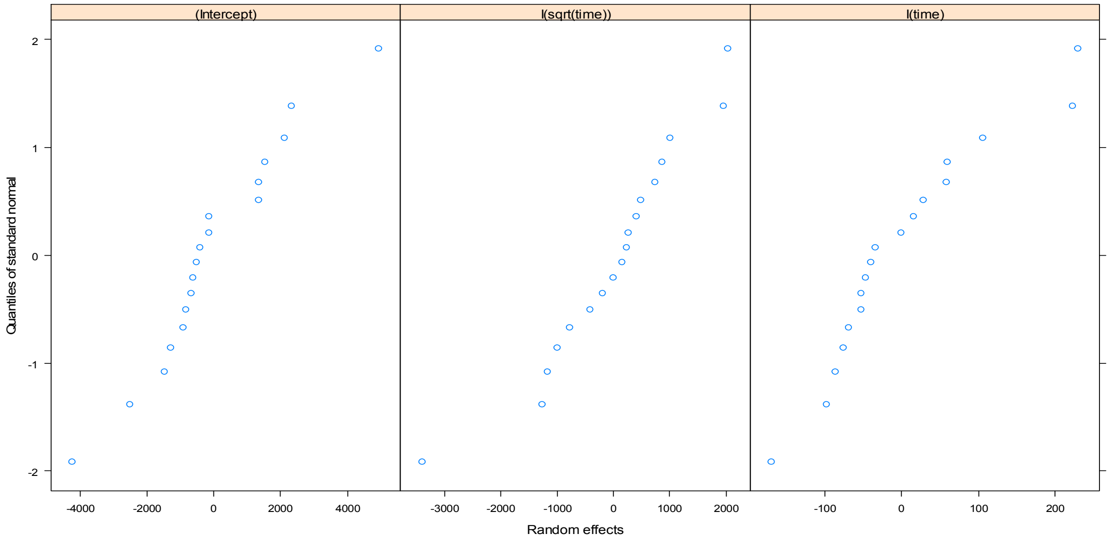

Figure 5. 11 Normal probability plots of random effects for the final model.

5.11 Summary

Based on descriptive and graphical exploratory analysis and using the **mfp** package in R, an appropriate preliminary mean growth model is identified as fractional polynomial model of order two. The selected preliminary mean structure shows that radial measure is a function of linear time and the square root of time. Following the selection of mean structure, the selection of random effects resulted in the significance of all three random effects (namely, intercept, coefficients of time, and coefficients of square root of time). While selecting the unstructured covariance as covariance structure of random terms, a search for best structure for the covariance of the error component was made. The search resulted in the heterogeneous variance, which varies by clone and exponential function of square root of time, as the best fit. Loess smoothing technique (Cleveland, 1979) attests that the selected model fits the data set well. Moreover, the non-parametric Loess fitted curves for both the GC and GU clones showed the plausibility of the fitted fractional polynomial models. The analyses showed that the GU clone has faster stem radial growth than the GC clone. The larger intercept for the GC clone showed at the initial stage, that the mean profile of the GC clone is higher than that of the GU(see Figure 2.3). The growth pattern of the two hybrid clones is similar during the juvenile stage. The rate of change of stem radial growth at the instantaneous change of time for both clones is a function of time, t' . However the rate of growth is different for the two clones. For the GC clone the growth rate at time t' is $\frac{d\hat{Y}_{t'}}{dt'} = \frac{1998.185}{\sqrt{t'}} - 94.44$ but for the GU clone it is $\frac{d\hat{Y}_{t'}}{dt'} = \frac{2307.215}{\sqrt{t'}} - 94.44$.

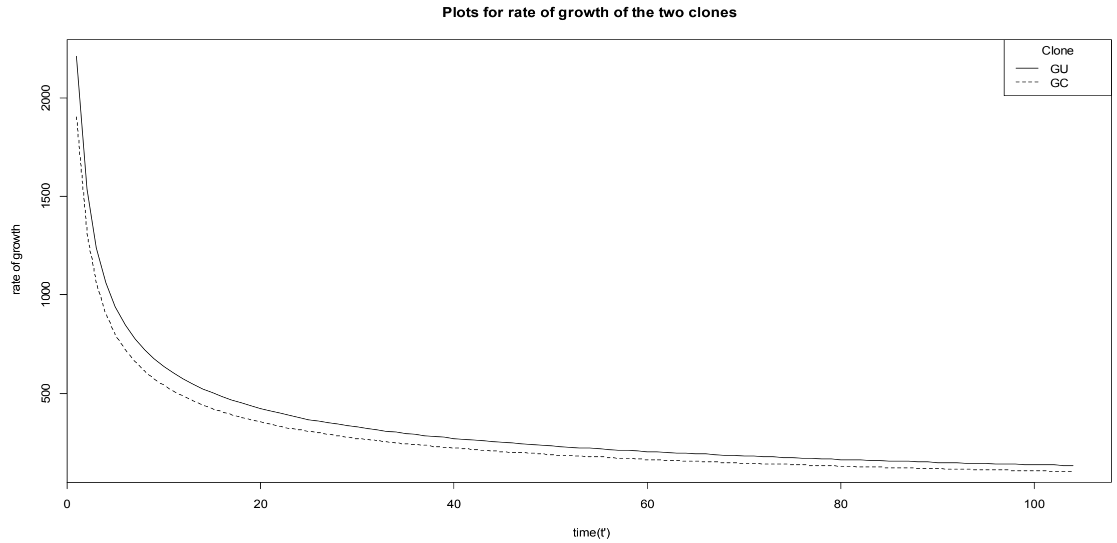

Figure 5. 12 Plot for the estimated rate of growth of the two clones.

The rate of growth for the two clones is presented in Figure 5.13. These increment rates are large at the initial stages and as t' increases the stem radial growth slows down and then tends to increase at a stable rate. The GU clone growth rate is larger than the GC clone during the entire juvenile stage. The faster growth characteristics of the GU clone points to improved genetics of this hybrid cross and to its potential ability to better exploit available resources, making it an economically viable hybrid cross as reported elsewhere (Galloway, 2003).

This fast growth shows that the GU clone has a genetic economic potential for rapid stem growth as compared to the GC clone. At time $t' = T$, the average radial growth advantage of the GU clone is $\sum_{t'=1}^{T}\left(\frac{309.03}{\sqrt{t'}}\right)$. For instance, after 52 weeks (one year) from the dendrometre installation age (i.e, 39 weeks), T=52, the average radial growth advantage of GU clone is 4027 μm.

The fractional models which were functions of tree age are extended to account for the effect of the climatic variables. Although tree age is the most important variable in determining the stem radial growth during the juvenile stage (up to two years), there is a significant effect of climatic variables on the stem radial change. Most of the climatic variables have a positive effect on the stem radius during the juvenile stage of tree development. It was found that temperature has an opposite effect on the radial growth of the two clones. The effect of temperature on the radial growth of GU clone is positive while it is negative for the GC clone. This could be primarily due to genetic variation between the two clones. Of course, this may entail further research in the area. The effect of weather variables depends on season. In winter, temperature, relative humidity, solar radiation and wind speed have a positive effect on the stem radial growth. In autumn, rainfall, temperature, relative humidity and solar radiation have a positive effect on the stem radial measure.

In this chapter we applied fractional polynomial models to model growth. We also extended the model so that it incorporates the effect of covariates. This model comprises a variety of curve shapes and can be easily modelled under the linear mixed model framework. The model can easily be extended to include the effect of other covariates. On the other hand, standard nonlinear growth models can be used to model growth curves. However, extending these models to handle the effect of other covariates can be very complicated. With the objective of comparing the fractional polynomial with that of nonlinear mixed models, a review of the nonlinear mixed model is presented in Chapter 6. Moreover, the comparison of the results obtained from fractional polynomial with that of standard nonlinear mixed models is made.

Chapter 6

Nonlinear Mixed Models and Comparison to Fractional Polynomial in the Context of Linear Mixed Models

6.1 Introduction

Mixed effects models are usually used to model repeated measures data. These methods are useful to flexibly model the within-group correlation commonly present in this type of data. Most of the work on methods of repeated measures data has focused on data that can be modelled by an expectation function that is linear in its parameters (e.g. Laird and Ware, 1982). Nonlinear mixed-effects models involve both fixed and random effects, in which some, or all, of the fixed and random effects occur nonlinearly in the model function.

Numerous nonlinear mixed-effects models have been proposed. These include Sheiner and Beal (1980); Mallet et al. (1988); Lindstrom and Bates (1990); Vonesh and Carter (1992); Davidian and Gallant (1992); and Wakefield et al. (1994).

Davidian and Giltinan (1995) and Vonesh and Chinchilli (1996) offered overviews along with general theoretical developments and some examples of nonlinear mixed models. Lindstrom and Bates (1990) proposed a general nonlinear mixed-effects model for repeated measures data and defined estimators for its parameters. These estimators are a combination of the least square estimators for Nonlinear fixed effects models and maximum likelihood (or restricted maximum likelihood) estimators for linear mixed-effects models. Pinheiro and Bates (2000) presented a slight generalization of the nonlinear mixed models proposed by Lindstrom and Bates (1990). This generalization allows the incorporation of “time varying” covariates in the fixed effects or the random effects for the model. This general formulation is implemented in R Statistical software (R Core Team, 2013). The implementation in R allows the use of nested random effects and also permits the within group error to be correlated (and/or) to have unequal variances. This general formulation was considered in this study. Nonlinear mixed models can be viewed as an extension of the linear mixed-model of Laird and Ware (1982) in which the conditional expectation of the response, given the random effects, is allowed to be a nonlinear function of the coefficients. It can also be regarded as an extension of nonlinear models for independent data (Bates and Watts, 1988) in which random effects are integrated in the coefficients to allow them to vary by group. Nonlinear mixed-models are becoming increasingly popular (Wolfinger, 1999). They are applied in many fields of study such as agriculture, forestry, biology, ecology and biomedicine. According to Pinheiro and Bates (2000), the main reasons for using a nonlinear mixed-model are interpretability, parsimony and validity beyond the observed range of data. By increasing the order of the polynomial model, it is possible to get increasingly accurate approximations to the true, usually nonlinear, regression function, within the range of the data. However, these higher order polynomial models may result in multicollinearity problems and they also provide no theoretical considerations about the underlying mechanism producing the data. Nonlinear models on the other hand are often mechanistic, that is based on a model for the mechanism producing the response. Consequently, the model parameters in nonlinear models generally have a natural physical interpretation. Even when derived empirically, nonlinear models usually incorporate known, theoretical characteristics of the data, such as asymptotes and monotonicity, and in these circumstances, can be considered as semi-mechanistic models. A nonlinear model generally uses fewer parameters than a competitor linear model, such as a polynomial, giving a more parsimonious description of the data. Nonlinear models also provide more reliable predictions for the response variable outside the observed range of the data than, say, polynomial models would (Pinheiro and Bates, 2000). The objectives of this chapter is to develop a stem radial increment model based on a nonlinear mixed model for two *Eucalyptus grandis* x *E. urophylla* hybrid clones comparing their growth potential with respect to the estimated parameters of the model.

6.2 Description of the General Nonlinear Mixed Model

The nonlinear mixed model can be viewed as a two stage model. In the first stage the j^{th} observation on the i^{th} individual is modelled as:

$$y_{ij} = f(\phi_{ij},\ X_{ij}) + \varepsilon_{ij} \qquad i = 1,\ 2\ \ldots M \quad and \quad j = 1,\ \ldots\ n_i \qquad (6.1)$$

where, y_{ij} is the j^{th} observation on the i^{th} individual, f is a nonlinear function of an individual specific parameter vector ϕ_{ij} and the predictor vector X_{ij} and ε_{ij} is the normally distributed within-group error term. M is the total number of individuals and n_i is the number of observations on the i^{th} individual. In the second stage the individual specific parameter vector (ϕ_{ij}) is modelled as:

$$\phi_{ij} = A_{ij}\beta + B_{ij}b_i \qquad b_i \sim N(0, \psi) \qquad (6.2)$$

where, β is a p-dimensional vector of fixed population parameters, and b_i is a q-dimensional random effects vector associated with the i^{th} individual (not varying with j), with variance covariance matrix ψ. The matrices A_{ij} and B_{ij} are design matrices for the fixed and random effects respectively. It is further assumed that observations made on different individuals are independent and that the within group errors ε_{ij} are independently distributed as $N(0, \sigma^2)$ and independent of the b_i. We can express (6.1) and (6.2) in matrix form (for the response vector of the i^{th} individual) as

$$\begin{aligned} y_i &= f(\phi_i \quad X_i) + \varepsilon_i \\ \phi_i &= A_i\beta + B_i \; b_i \\ &for \; i = 1, 2, \; \ldots \; M \end{aligned} \qquad (6.3)$$

where $y_i = \left[y_{i1} \quad y_{i2} \quad . \quad . \quad . \qquad\qquad y_{in_i} \right]^T$

$$\phi_i = \left[\phi_{i1} \quad \phi_{i2} \quad . \quad . \quad . \qquad\qquad \phi_{in_i} \right]^T$$

$$\varepsilon_i = \left[\varepsilon_{i1} \quad \varepsilon_{i2} \quad . \quad . \quad . \qquad\qquad \varepsilon_{in_i} \right]^T$$

$$X_i = \left[X_{i1} \quad X_{i2} \quad . \quad . \quad . \qquad\qquad X_{in_i} \right]^T$$

$$A_i = \left[A_{i1} \quad A_{i2} \quad . \quad . \quad . \qquad\qquad A_{in_i} \right]^T$$

$$B_i = \left[B_{i1} \quad B_{i2} \quad . \quad . \quad . \qquad\qquad B_{in_i} \right]^T$$

$$f_i(\phi_i, X_i) = \left[f_i(\phi_{i1}, X_{i1}) \quad f_i(\phi_{i2}, X_{i2}) \quad . \quad . \quad . \qquad\qquad f_i(\phi_{in_i}, X_{in_i}) \right]^T$$

Several methods for estimating the parameters of the nonlinear mixed models have been suggested. Our emphasis will be on two of them namely maximum likelihood and restricted maximum likelihood.

The evaluation of the log-likelihood function of the data is a complex numerical issue because it usually involves integral that does not have a closed-form expression.

The maximum likelihood estimation in (6.1) is based on the marginal density of given by:

$$p\left(y / \beta , \sigma^{2}, \psi\right) = \int p\left(y / b , \beta , \psi , \sigma^{2}\right) p(b)\, db \qquad (6.4)$$

Where,

$p\left(y / \beta, \sigma^{2}, \psi\right)$ = is the marginal density of y

$p\left(y / b, \beta, \sigma^{2}\right)$ is the conditional density of y given the random effect

$p(b)$ is the marginal distribution of b

In general, the integral in model (6.4) does not have a closed-form expression when the model function *f* is non-linear in random effects. Different approximations have been proposed for estimating it. Some of these methods are the LME approximation method suggested by Lindstrom and Bates (1990); the method by Sheiner and Beal (1980) and Vonesh and Carter (1992) that takes first order Taylor expansion of the model function, *f,* around the expected value of the random effects; a modified Laplacian approximation (Tierney and Kadane 1986) and Gaussian quadrature (Davidian and Gallant 1992). Pinheiro and Bates (1995) analysed several approximations to log-likelihood of non-linear mixed effects model and concluded that Lindstrom and Bates' (1990) approximation usually gives accurate results. In the section that follows the method suggested by Lindstrom and Bates (1990) to approximate the log-likelihood (6.4) is presented.

6.3 Approximations to The Likelihood in The Nonlinear Mixed Effects Model

Lindstrom and Bates (1990) suggest an alternating algorithm for estimating the parameters of model nonlinear mixed model. This estimation algorithm alternates between two steps, a penalized nonlinear least squares (PNLS) step and a linear mixed model (LME) step. The alternating algorithm for model (6.1) is a follows.

For the nonlinear mixed effects model (6.1), the random effects variance-covariance matrix can be expressed, in terms of the precision factor (Δ), so that $\psi^{-1} = \sigma^{-2} \Delta^T \Delta$.

In the PNLS step, the current estimate of Δ (the precision factor) is held fixed, and the conditional modes of the random effects b_i and the conditional estimates of the fixed effects β are obtained by minimizing a penalized nonlinear least squares objective function (Pinheiro and Bates, 2000)

$$\sum_{i=1}^{M} \left[\left\| y_i - f_i\left(\beta, b_i\right) \right\|^2 + \left\| \Delta b_i \right\|^2 \right] \qquad (6.5).$$

The LME step updates the estimate of Δ based on first order Taylor expansion of the model function f around the current estimates of β and the conditional modes the random effects b_i . The current estimate of β and the modes of the random effects are denoted by $\hat{\beta}^{(w)}$ and $\hat{b}_i^{(w)}$, respectively.

Using

$$\hat{X}_i^{(w)} = \left. \frac{\partial f_i}{\partial \beta^T} \right|_{\hat{\beta}^{(w)}, \hat{b}_i^{(w)}}, \qquad \hat{Z}_i^{(w)} = \left. \frac{\partial f_i}{\partial b_i^T} \right|_{\hat{\beta}^{(w)}, \hat{b}_i^{(w)}},$$

$$\hat{w}_i^{(w)} = y_i - f_i\left(\hat{\beta}^{(w)}, \hat{b}_i^{(w)}\right) + \hat{X}_i^{(w)} \hat{\beta}^{(w)} + Z_i^{(w)} b_i^{(w)} \qquad (6.6)$$

the approximate log-likelihood function used to estimate Δ is

$$l_{LME}\left(\beta,\sigma^2,\ \Delta \mid y\ \right) = \frac{-N}{2}\ \log\left(2\pi\sigma^2\right) - \frac{1}{2}\sum_{i=1}^{M}\Bigg\{\ \log\left|\ \textstyle\sum_i\ (\Delta)\ \right|$$
$$+\sigma^{-2}\left[\hat{w}_i^{(\,w)} - \hat{X}_i^{(\,w)}\beta\ \right]^T \textstyle\sum_i^{-1}(\Delta)\ \left[\hat{w}_i^{(\,w)} - \hat{X}_i^{(\,w)}\beta\right]\ \Bigg\} \tag{6.7}$$

where $\sum_i\left(\Delta\right) = I + \hat{Z}_i^{(\,w)}\,\Delta^{-1}\,\Delta^{-T}\,\hat{Z}_i^{(w)T}$.

This log-likelihood is identical to that of a linear mixed model in which the response vector is given by $\hat{w}^{(\,w)}$, and the fixed and random effects design matrices are given by $\hat{X}^{(w)}$ and $\hat{Z}^{(w)}$respectively. This greatly simplifies the optimization problem (Pinheiro and Bates, 2000).

Lindstrom and Bates (1990) also suggested a restricted maximum likelihood estimation method forΔ, which involves changing the log-likelihood in the LME step of the alternating algorithm by the following log-restricted-likelihood.

$$l_{LME}^{R}\left(\sigma^2,\ \Delta \mid y\right) =$$
$$l_{LME}\left(\hat{\beta}(\Delta), \sigma^2, \Delta \mid y\right) - \frac{1}{2}\sum_{i=1}^{M}\log\left|\sigma^{-2}\hat{X}_i^{(\,w)T}\textstyle\sum_i^{-1}(\Delta)\,\hat{X}_i^{(w)}\right| \tag{6.8}$$

If either the fixed effects or the random effects change, the penalty factor for the log-restricted likelihood (6.8) will also change. This is because of the fact that $\hat{X}_i^{(w)}$ depends on both $\hat{\beta}^{(w)}$ and $\hat{b}_i^{(w)}$. This implies that log-restricted-likelihoods from nonlinear mixed effects (NLME) models with different fixed or random effects are incomparable. The algorithm alternates between penalized nonlinear least squares (PNLS) and LME steps until a convergence is achieved. These alternating algorithms appear to be more efficient when the estimates of the variance-covariance components (Δ and σ^2) are not strongly correlated with the estimates of the fixed effects (β).

Only the LME step was used by Lindstrom and Bates (1990) to update the estimate of Δ. However, the LME step also produces updated estimates of

β and the conditional modes of b_i. Thus one can iterate LME steps by re-evaluating equations (6.6) and (6.7) or (6.8) for the log-restricted-likelihood at the updated estimates of β and b_i (Wolfinger, 1993, Pinheiro and Bates, 2000).

6.4 Inferences and Predictions for Nonlinear Mixed Models

The parameters of nonlinear mixed effects model are estimated via the alternating algorithm. Inference on these parameters is based on the LME approximation to the log-likelihood function defined in section 6.3. Under the LME approximation, for fixed effects, the distribution of the maximum likelihood or restricted maximum likelihood estimators ($\hat{\beta}$) is

$$\hat{\beta} \sim N\left(\beta,\ \sigma^2\left[\sum_{i=1}^{M} \hat{X}_i^T \sum_i^{-1} \hat{X}_i\right]^{-1}\right) \qquad (6.9),$$

where $\sum_i = I + \hat{Z}_i \Delta^{-1} \Delta^{-T} \hat{Z}_i$, with $\hat{X}_i$ and $\hat{Z}_i$ are defined as in (6.6).

The standard errors included in the **summary** method for **nlme** objects are obtained from the approximate variance-covariance matrix in (6.9). The t and F tests are reported in the **summary** method and the *anova* method for single argument are also based on (6.9).

Assume θ denote an unconstrained set of parameters that determine the precision factor Δ. The LME approximation is also used to offer an estimated distribution for REML or ML estimators $\left(\hat{\theta},\ \log\hat{\sigma}\right)^T$. Using $\log\sigma$ instead of σ^2 to provide an unrestricted parameterization for which the normal approximation tends to be more accurate.

$$\begin{bmatrix} \hat{\theta} \\ \log\ \hat{\sigma} \end{bmatrix} \sim N\left(\begin{bmatrix} \theta \\ \log\ \sigma \end{bmatrix},\ \ I^{-1}(\theta,\ \sigma)\right),$$

$$I(\theta, \sigma) = -\begin{bmatrix} \partial^2 l_{LME_P} / \partial\theta\partial\theta^T & \partial^2 l_{LME_P} / \partial\log\sigma\partial\theta^T \\ \partial^2 l_{LME_P} / \partial\theta\partial\log\sigma & \partial^2 l_{LME_P} / \partial^2\log\sigma \end{bmatrix} \quad (6.10)$$

where $l_{LME_P} = l_{LME_p}(\Delta, \sigma)$ denotes the LME approximation to the log-likelihood, profiled on the fixed effects, and I denotes the empirical information matrix. The identical approximate distribution is usable for the REML estimators with l_{LME_p} replaced by the log-restricted-likelihood l^R_{LME} defined in (6.8). In real-world, Δ and σ^2 are replaced by their respective REML or ML estimates in the expressions for the approximate variance-covariance matrices in (6.9) and (6.10). The approximate distributions for the REML or ML estimators are used to produce the confidence intervals reported in the intervals method for **nlme** objects (Pinheiro and Bates, 2000).

The fitted values and predictions for nonlinear mixed models can be obtained at different levels of nesting or at the population level. The prediction that estimate the expected value of the response by considering the random effects to have their mean value zero is called population level predictions. For instance, if the covariate X_h stands for a vector of fixed effects and V_h a vector of other model covariates, the corresponding population prediction for the response y_h estimates $f(X_h \beta, V_h)$.

The predictions at level k is obtained by adding together the contributions from the estimated fixed effects and the estimated random effects at levels $\leq k$ and evaluating the model function at the resulting estimated parameters. For instance, if $Z_h(i)$ stands for a vector of covariates corresponding to random effects associated with the i^{th} group at the first level of nesting, the level-1 predictions estimate $f\left(X_h^T \beta + Z_h(i)^T b_i, v_h\right)$.

The REML or ML estimates of the fixed effects and the conditional modes of the random effects, which are estimated Best Linear Unbiased Predictors (BLUPs) of the random effects in the LME approximate log-likelihood, are used to obtain predicted values for the response. For instance, the population and level-1 predictions for y_h are $\hat{y}_h = f\left(X_h^T \hat{\beta}, v_h\right)$ and $\hat{y}_h(i) = f\left(X_h^T + Z_h(i)^T \hat{b}_i, v_h\right)$ respectively (Pinheiro and Bates, 2000).

6.5 Extending the Nonlinear Mixed Model

The nonlinear mixed model formulation used in equation (6.3) conforms to the assumption that the within-group errors be independent and have constant variance. This model is called the basic NLME model. It provides an appropriate model for a wide range of applications. However, there are several practical cases in which this assumption of independence and constant variance may not work. In this section a brief discussion how to extend the basic nonlinear mixed model will be given.

The model in equation (6.3) assumes that the within-group errors ε_i are independent $N\left(0, \sigma^2 I\right)$ random vectors. The extended nonlinear mixed model relaxes this assumption by allowing heteroscedastic and correlated within-group errors and can be expressed as

$$y_i = f_i(\phi_i, v_i) + \varepsilon_i, \qquad \phi_i = A_i \beta_i + B_i b_i$$
$$b_i \sim N(0, \psi), \qquad \varepsilon_i \sim N\left(0, \sigma^2 \Lambda_i\right) \qquad (6.11)$$

The Λ_i are positive definite matrices. The within-group errors ε_i are independent of the random effects b_i. As in the LME models, the variance covariance structure of the within-group errors can be decomposed into two independent components: a variance structure and correlation structure. The variance function models described in section (5.6) and the correlation models described in section (5.7) can also be applied for extending the nonlinear mixed model.

The estimation procedures and all inference procedures can be applicable for the extended model (6.11) because of the following transformation.

The matrix Λ_i in (6.11) is positive definite. It admits an invertible square root $\Lambda_i^{1/2}$ (Thisted, 1988; Pinherio and Bates, 2000), with inverse $\Lambda_i^{-1/2}$ such that

$\Lambda_i = \Lambda_i^{-T/2} \Lambda_i^{1/2}$ and $\Lambda_i^{-1} = \Lambda_i^{-1/2} \, \Lambda_i^{-T/2}$. Using the transformation $y_i^* = \Lambda_i^{-T/2} y_i$, $f_i^*(\phi_i, v_i) = \Lambda_i^{-T/2} f_i(\phi_i, v_i)$, $\varepsilon_i^* = \Lambda_i^{-T/2} \varepsilon_i$

$E(\varepsilon_i^*) = E(\Lambda_i^{-T/2} \varepsilon_i) = \Lambda_i^{-T/2} E(\varepsilon_i) = 0$ and

$\operatorname{var}(\varepsilon_i^*) = \operatorname{var}(\Lambda_i^{-T/2} \varepsilon_i) = \Lambda_i^{-T/2} \operatorname{var}(\varepsilon_i) (\Lambda_i^{-T/2})^T = \sigma^2 \Lambda_i^{-T/2} \Lambda_i \Lambda_i^{-1/2} = \sigma^2 I$.

This implies that, $\varepsilon_i^* \sim N(0, \sigma^2 I)$. As a result, it is possible to rewrite (6.11) as

$$\begin{aligned} &y_i^* = f_i^*(\phi_i, v_i) + \varepsilon_i^*, \qquad && \phi_i = A_i \beta_i + B_i b_i \\ &b_i \sim N(0, \psi), && \varepsilon_i^* \sim N(0, \sigma^2 I) \end{aligned} \qquad (6.12)$$

The equation in (6.12) is the same basic nonlinear mixed effects model given in equation (6.3). The log-likelihood function $l(\beta, \sigma^2, \Delta, \lambda / y^*)$corresponds to the basic NLME model with model function f_i^* and, therefore, the approximations presented in previous sections can be applied to it. The results presented for inference and predictions also remain applicable.

6.6 Selection of the Nonlinear Function and NLME Model Reformulation

Due to the large number of possible nonlinear functions that can be used in a nonlinear model, the determination of the appropriate function is not always easy. Scientific knowledge about the phenomena under study is important in determining the appropriate model. Historical knowledge from previous studies of functions that fit similar data well in the past might be helpful in selecting the proper function for data. Sometimes the plot of the data suggests a well-known function. Probably, the best way to select an initial model is to plot the data. Based on the exploratory data analysis of Chapter 2 and the shapes of different functions, three nonlinear growth functions were selected as candidates for stem radial growth modelling. These three growth curves were used to replace the function *f in* model (6.1).

I. **Three parameter logistic regression**: The first growth curve introduced is the three parameter logistic regression. This model can be expressed as

$$f(x,\phi) = \frac{\phi_1}{\left\{1+\exp\left[-(x-\phi_2)\big/\phi_3\right]\right\}} \qquad (6.13)$$

The parameters of this model have a physical interpretation. ϕ_1 refers to the asymptotic stem radius. ϕ_2 refers to the time at which the tree reaches half of the asymptotic stem radius. ϕ_3 is the time elapsed for the tree to reach between half and three fourths of its asymptotic stem radius.

The nonlinear mixed model corresponding to the logistic function 6.13, with the random effects for all parameters, is

$$y_{ij} = \frac{\phi_{1i}}{1+\exp\left[-\left(t_{ij}-\phi_{2i}\right)\big/\phi_{3i}\right]} + \varepsilon_{ij} \qquad (6.14)$$

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} = \beta + b_i$$

$$b_i \sim N(0, \psi), \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Where y_{ij} is the stem radius for tree i at t_{ij} weeks after planting. The fixed effects, β represent the mean value of the individual parameters, ϕ_i, in the population of eucalyptus trees and the random effects , b_i , represent the deviations of the ϕ_i from their mean values.

II. **The asymptotic regression model:** this is given by the formula

$$f(x, \phi) = \phi_1 + (\phi_2 - \phi_1)\left(\exp\left[(-\exp(\phi_3))\, x\right]\right) \qquad (6.15)$$

ϕ_1 is the asymptote as x approaches infinity. ϕ_2 is the value y when x is zero. ϕ_3 is the logarithm of the rate constant. The corresponding nonlinear mixed effects model for the radial measure y_{ij} and tree i at t_{ij} weeks after planting is

$$y_{ij} = \phi_{1i} + (\phi_{2i} - \phi_{1i})\left(\exp\left[(-\exp(\phi_{3i}))\, t_{ij}\right]\right) + \varepsilon_{ij} \qquad (6.16)$$

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} = \beta + b_i$$

$$b_i \sim N(0, \psi), \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

III. **The Gompertz growth function:** The three parameter Gompertz function can be expressed as

$$y = \alpha \ \exp\left(-\beta \times \exp(-\gamma t)\right) \quad \alpha > 0,\ \beta > 0,\ \gamma > 0 \qquad (6.17)$$

The limiting value as t approaches infinity is α. The starting value of y at t=0 is $\alpha \exp(-\beta)$, and with the restrictions on the parameters $0 < \alpha \exp(-\beta) < \alpha$.

The representation of the Gompertz function in R Statistical Software is

$$y = \phi_1 \ \exp\left(-\phi_2 \times \phi_3^t\right) \qquad (6.18)$$

The relationship between the parameters of model (6.17) and (6.18) is that

$\alpha = \phi_1$,

$\beta = \phi_2$

$\exp(-\gamma) = \phi_3$.

The representation in R is used, because the R Statistical Software is employed for fitting the model. R is free software that can be downloaded from the R project website R core team (2013).

The parameters of this model have physical interpretation. ϕ_1 refers to the asymptotic stem radius. The starting value of the stem radius at (t=0) is $\phi_1 \times \exp^{(-\phi_2)}$ with the restrictions on the parameters $0 < \phi_1 \exp(-\phi_2) < \phi_1$.

ϕ_3 is the exponent of the negative of the shape parameter. This indicates the parameter $(-\ln(\phi_3))$ models the shape of the function. Differentiating model (6.18), with respect to t, we have:

$$\frac{dy}{dt} = \phi_1 (-\ln(\phi_3))\, \phi_2\, \phi_3^t \, \exp\left(-\phi_2 \times \phi_3^t\right) = (-\ln(\phi_3))\, y\, \phi_2\, \phi_3^t \qquad (6.19)$$

$$\frac{d^2y}{dt^2} = (-\ln(\phi_3))^2 \; y \; \phi_2 \, \phi_3^t (\phi_2 \, \phi_3^t - 1) \qquad (6.20)$$

From model (6.21) there is a point of inflection when

$$t = \frac{-\ln(\phi_2)}{\ln(\phi_3)} \qquad (6.21)$$

The relative growth rate as a function time (t) is

$$\frac{1}{y}\frac{dy}{dt} = -\ln(\phi_3) \; \phi_2 \; \phi_3^t \qquad (6.22)$$

The nonlinear mixed model corresponding to the Gompertz function (6.18), with the random effects for all three parameters, is:

$$y_{ij} = \phi_{1i} \; \exp\left(-\phi_{2i} \times \phi_{3i}^{t_{ij}}\right) + \varepsilon_{ij} \qquad (6.23)$$

$$\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \end{bmatrix} = \beta + b_i$$

where

$$b_i \sim N(0, \psi), \qquad \varepsilon_{ij} \sim N(0, \sigma^2)$$

y_{ij}: The stem radius at time j for i^{th} tree (μm)

t_{ij} : The age at time j for the i^{th} tree (weeks).

The fixed effects, β represent the mean value of the individual parameters, ϕ_i , in the population of eucalyptus tree and the random effects , b_i , represent the deviations of the ϕ_i from their mean values.

6.7 Results of Fitting Three Parameter Logistic Model

It is necessary to consider the questions of determining which parameters in the model should have a random component and whether the variance covariance matrix of the random effects can be structured in simpler form with fewer parameters. The first question that should be addressed in the analysis is choosing which parameters should be random effects and which are purely fixed effects. A separate fit for each tree was made and inter-tree variability was assessed using the individual confidence intervals. Since several repeated measurements were considered for each tree, the data include sufficient observations to have meaningful parameter estimates in the individual fits.

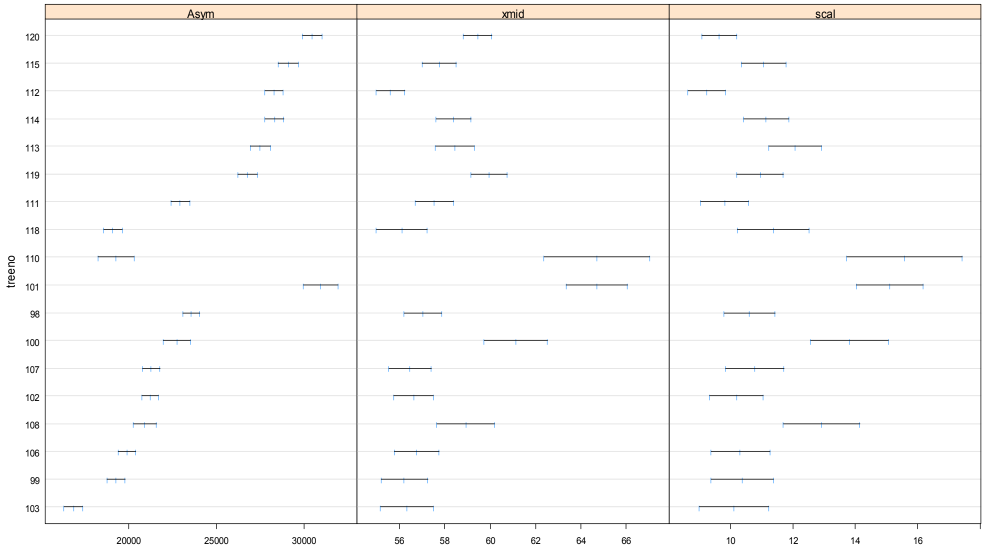

Figure 6. 1 Ninety five percent confidence intervals on the parameters of logistic model (6. 14) based on individual tree fit. The parameters in this graph are related to the parameters in the logistic model as follows ($\phi_1 = Asym$, $\phi_2 = xmid$ and $\phi_3 = scal$)

The approximate 95% confidence intervals on parameters of model (6.14) are given in Figure 6.1 for each tree. It is clear that for each parameter, all confidence intervals do not overlap. This suggests that random effects for all three parameters may be necessary. An alternative approach is to fit different prospective models and compare nested models using the likelihood ratio tests or information criterion statistics, such as the Akaike Information Criterion (AIC) (Sakamoto et al.,1986). This alternative approach was considered for the parameters of model (6.14). The model was fitted with each of ϕ_1, ϕ_2, and ϕ_3 as mixed effects, called model I. The resulting AIC was 18478.85. From a reduced form of model I, with only ϕ_1 and ϕ_2 as mixed, we get an AIC of 18608.32. This was model II. The model, that considers ϕ_1 and ϕ_3 as mixed effects, was also fitted. The resulting AIC was 18692.69. This was model III. Finally, the model with ϕ_2 and ϕ_3 as mixed effects, was considered and the resulting AIC was equal to 19481.16. This was termed model IV. The AIC of the models that considered each of ϕ_1, ϕ_2 and ϕ_3 at a time as fixed effects were 19481.16, 18692.69 and 18608.32 respectively. All these values were larger than the AIC of model I. This gives a clear indication that the elimination of any of these random effects has a huge impact on the quality of the fit. The comparison of model I with any of the three reduced models (Models II, III and IV) using likelihood ratio test, produced a *p*-value, which was less than 0.0001 for all comparisons. It was concluded that the covariance structure should not be simplified by deleting any of the random effects of model (6.14). This is consistent with the conclusions of the individual fits analysis discussed, using the approximate confidence intervals. Therefore, a model with random effects for all three parameters was considered.

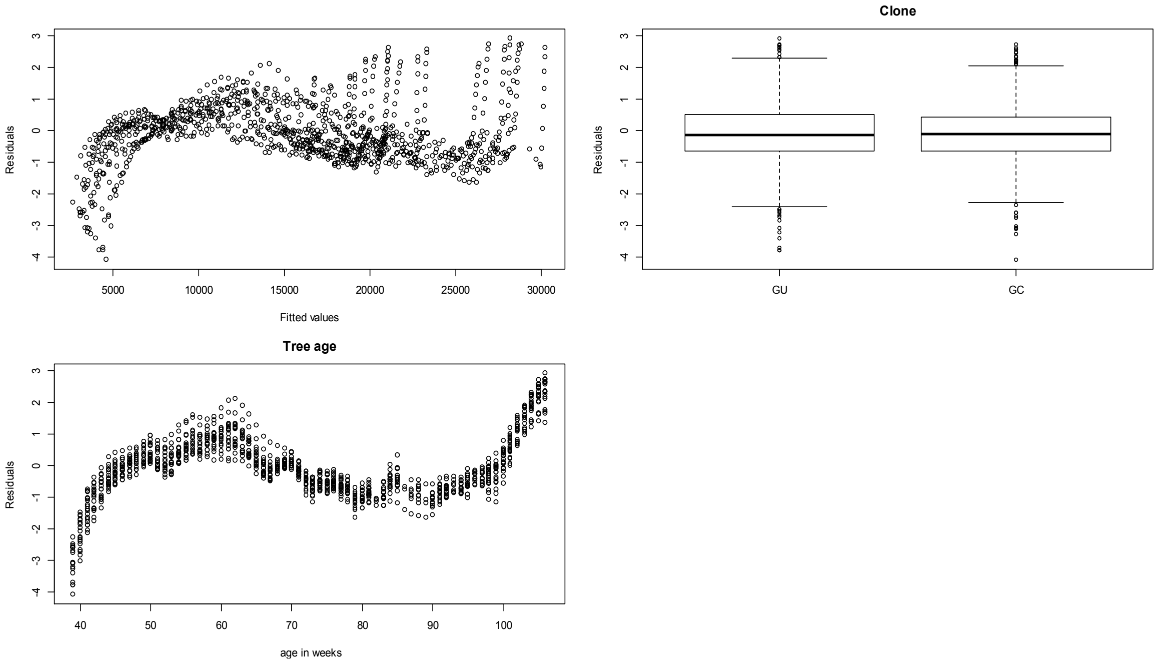

Figure 6. 2 Model validation graphs for the model with all three parameters as mixed effects**.**

The residuals versus fitted values, clone and tree age are shown in Figure 6.2. The graphs show clear violation of the assumption of homogeneity of variances. The plot of residuals versus fitted values shows a clear pattern of variability for the within-group errors. The residuals also fluctuate with tree age and the variance of residuals is not the same for the two clones.

The within-group heterogeneity was modelled using different variance functions and different correlation structures as discussed in section 5.6 and 5.7. The model with different variance of residuals for each time point appears to be the best fit among those models for which convergence was achieved. The AIC of this model was 17115.96 which is the smallest value of all the models fitted. The adequacy of this model was assessed by plotting the standardized residuals against the fitted values, tree age and clone as shown in Figure 6.3. There was a huge improvement in the validation graphs. There was no clear indication for the departure from nonlinear mixed model assumption. The model with different variances for each time point adequately fits the within-group heteroscedasticity.

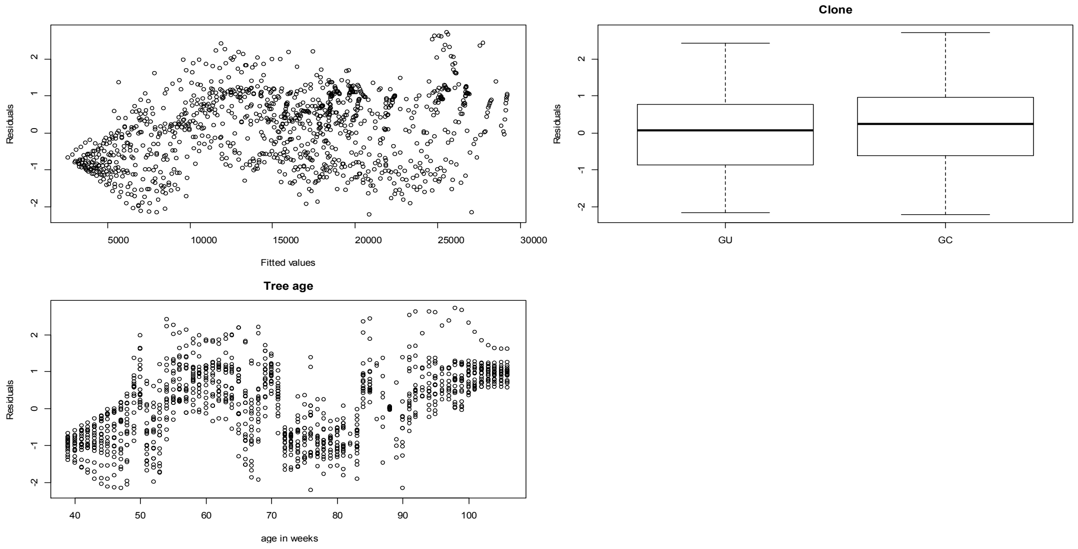

Figure 6. 3 Model validation graphs for the extended model with different variance for each time point.

The primary question of interest for the data at hand was whether there was a pattern between the growth in stem radius and the type of clone. The plots for the estimates of random effects by clone are given in Figure 6.4. In Figure 6.4 two of the parameters seem to vary with clone. It appears that the asymptote (Asym) and the time at which half of the asymptotic radius is attained (xmid) are larger for the GU clone than for the GC clone.

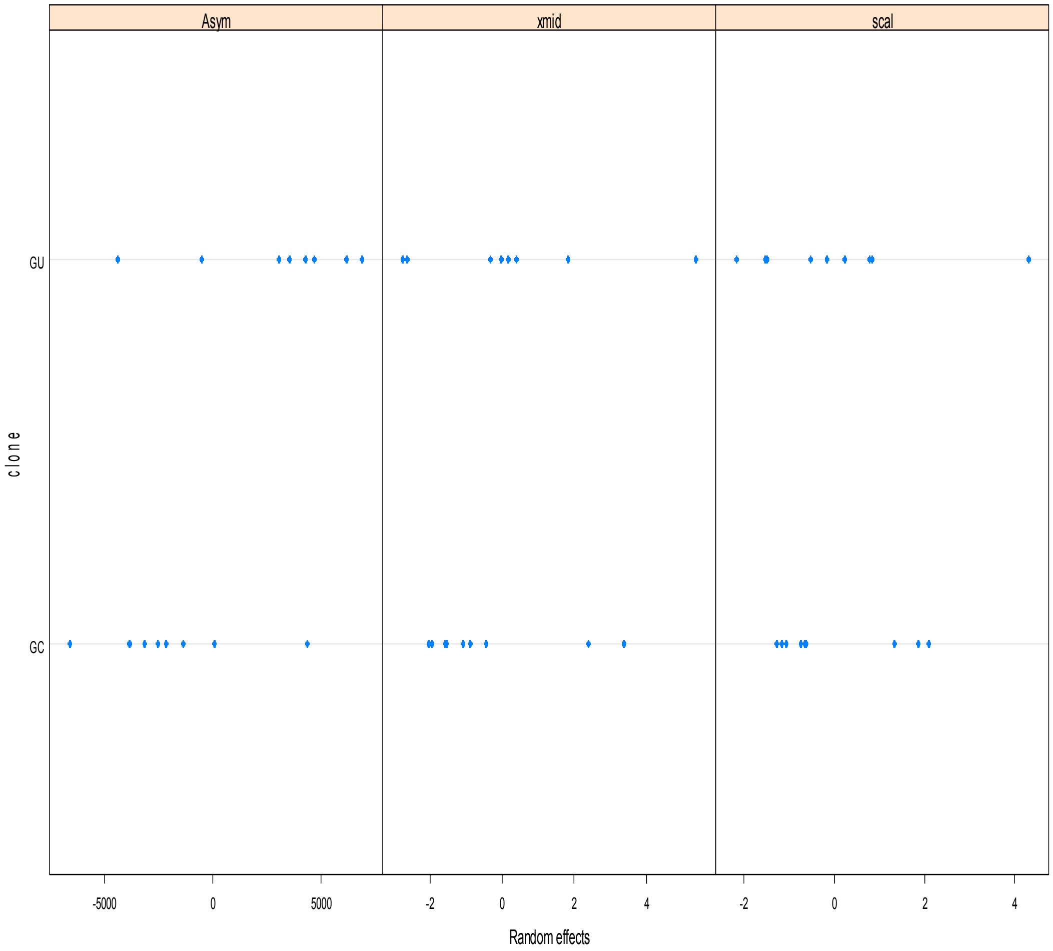

Figure 6.4 Estimates of random effects by clone in the model different variance by tree age.

The dependence of all three parameters on the clone was modelled. The significance of the clone for fixed effects was also assessed by comparing the models with and without the clone effect. The clone had a significant effect on the asymptote (ϕ_1) of the model (p-value is equal to 0.03). The fitted three parameter logistic model is given by:

$$radius = \frac{24263.93}{1 + \exp\left[-\left(\frac{t - 56.67}{10.98}\right)\right]} \quad for \quad GU \quad clone$$

$$radius = \frac{20868.44}{1 + \exp\left[-\left(\frac{t - 56.67}{10.98}\right)\right]} \quad for \quad GC \quad clone$$

Table 6. 1 ANOVA table for the fitted nonlinear mixed effects logistic model

Source of variation	NDF^a	DDF^a	F-value	P-value
Asymptote-intercept (ϕ_1)	1	1096	1291.71	< 0.0001
Asymptote-clone (slope)	1	1096	4.95	0.03
Inflection point (ϕ_2)	1	1096	18676.97	< 0.0001
Scale parameter (ϕ_3)	1	1096	856.57	< 0.0001

a) NDF, Numerator degrees of freedom. DDF, Denominator degrees of freedom.

The ANOVA table (Table 6.1) for the fitted model suggests that the clone has a significant effect on the asymptotes of the logistic curves. The parameter estimate suggests that the average stem radius of each tree reached the inflection point about 57 weeks after the first measurement was taken. Another 11 weeks after the inflection point was reached (i.e., 68 weeks after first measurement), the average stem radius reached about 75% of the growth asymptote for each experimental tree. The overall average stem radius at the end of the juvenile stage of the tree was 24263.96 and 20868.45 for the GU clone and GC clones respectively. Clone had a significant negative slope (Table 6.2), which indicates the asymptote for the GU clone is larger than that of the GC. This is in agreement with results obtained in Figure 2.3 and is an indication that the GU clone has a better genetic potential for growth than the GC clone.

Table 6. 2 Summary of the fixed effects parameter estimation results for the fitted logistic mixed effects model.

Fixed effects	LCL^a	Estimated	UCL^a
Asymptote-intercept (ϕ_1)	22093.31	24263.96	26434.61
Asymptote-clone (slope)	-6272.98	-3395.51	-518.04
Inflection point (ϕ_2)	55.68	56.67	57.66
Scale parameter (ϕ_3)	10.24	10.98	11.71

a) LCL, approximated 95% lower confidence limit; UCL, approximated 95% upper confidence limit.

The assumption of normality for the within group errors was assessed using the normal probability plot of residuals (Figure 6.5). Close examination of the behaviour of the two plots (see Zewotir and Galpin, 2004) shows that the normality assumption is plausible.

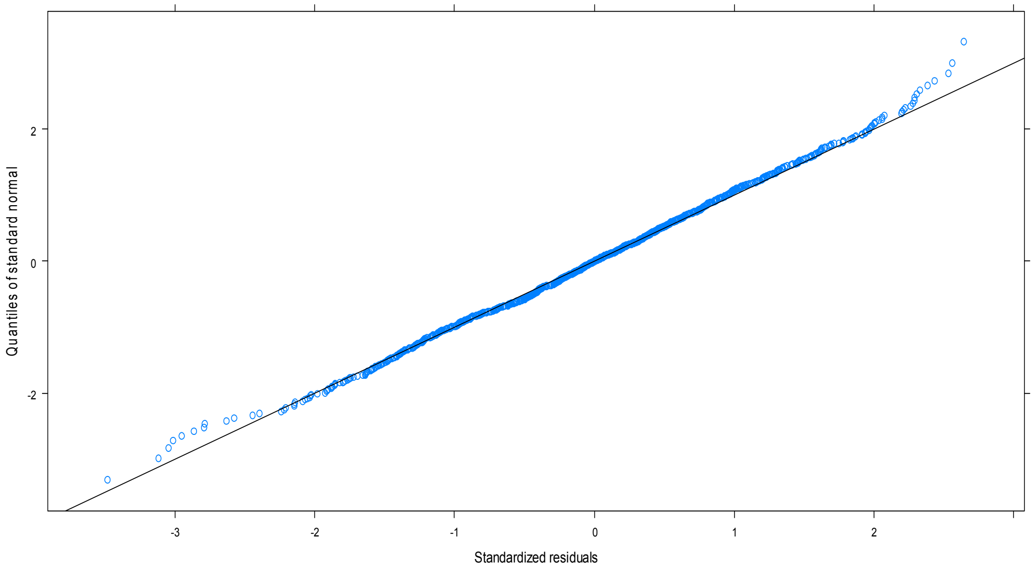

Figure 6. 5 Normal probability plot of residuals.

Investigation of the marginal normality of the corresponding random effects was also made. The normal probability plots of the random effects are indicated in Figure 6.6. The assumption of normality seems reasonable for all three random effects.

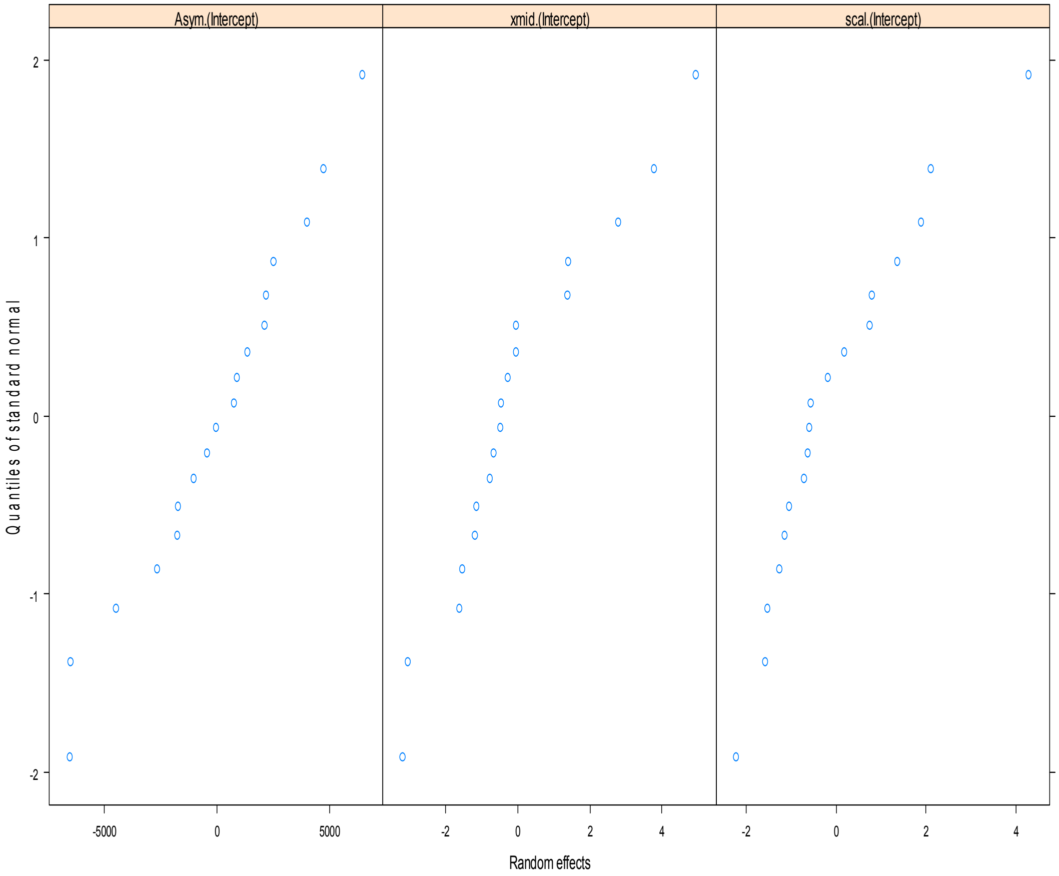

Figure 6.6 Normal probability plot of random effects

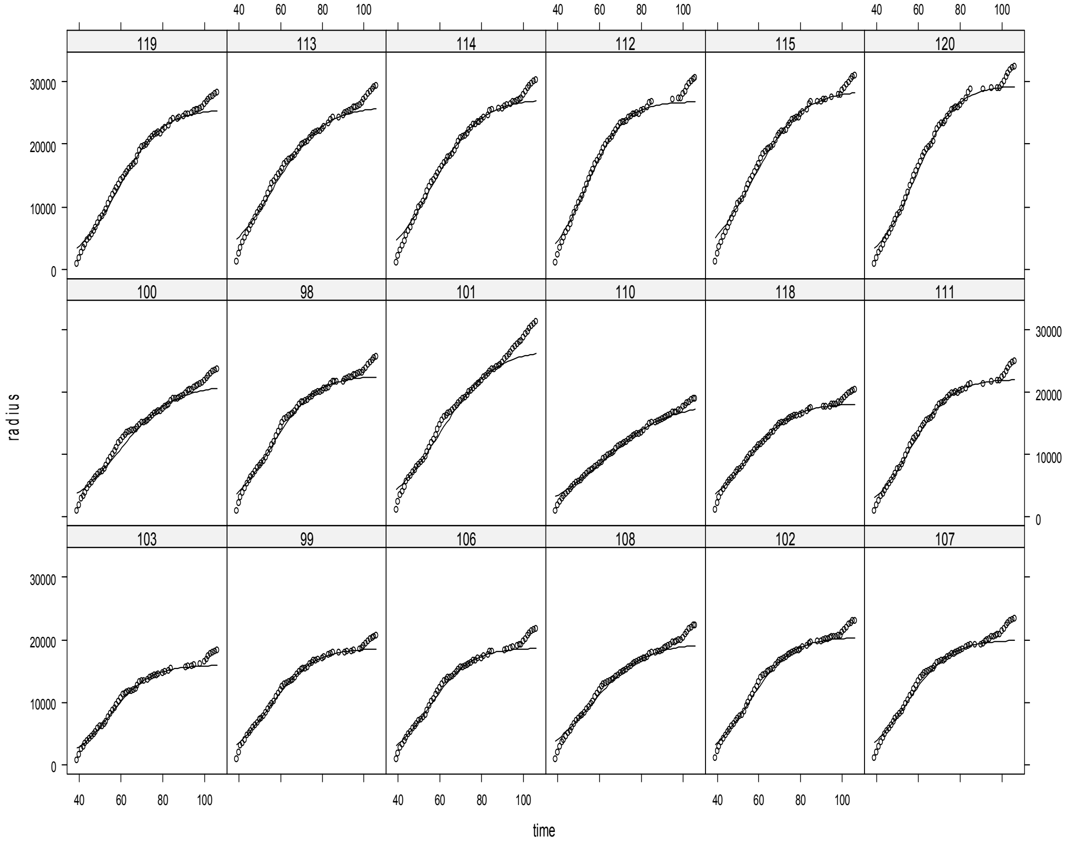

Figure 6. 7 Plots of the fitted model and observed values for each tree using the three parameter logistic model.

The adequacy of the three parameter logistic model, at individual tree level, was checked. The plot of the augmented predictions, by tree, was used as an assessment for adequacy of the logistic growth model (Figure 6.7). The predicted values closely matched the observed radial growth measurements demonstrating the acceptability of the model. Moreover, the linear regression between the observed and fitted values, which had an $R^2 = 0.98$, suggested that the overall model fit was satisfactory (Figure 6.8).

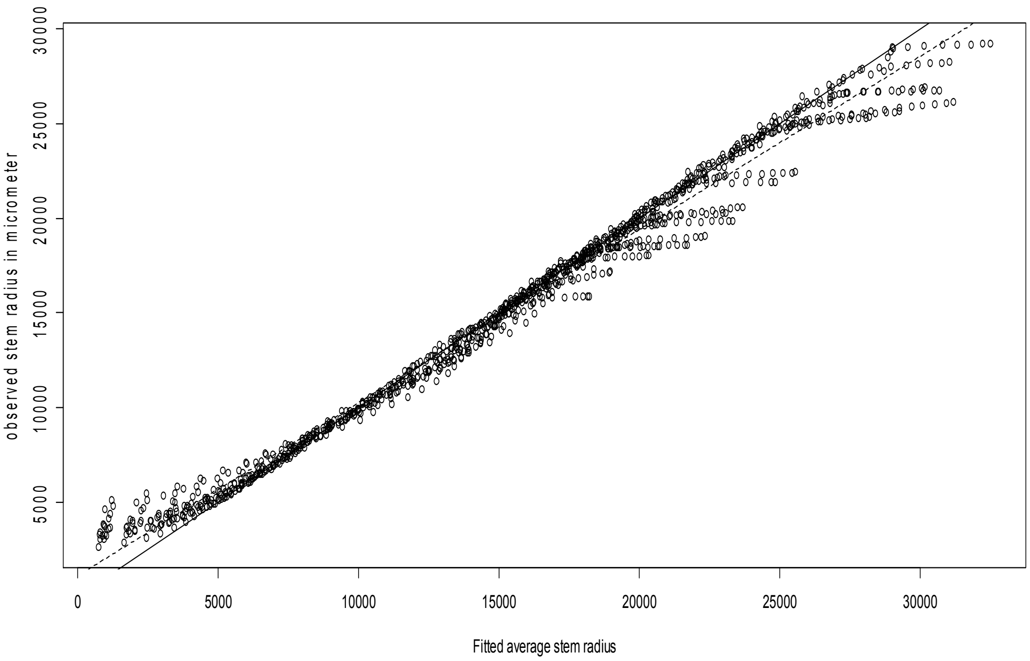

Figure 6.8 Scatter plot of the fitted versus observed average stem radius. The dashed line is the estimated regression line between the observed and fitted values. (Fitted =1109+0.915 observed) and the solid line is the 1: 1 line.

6.8 Results of Fitting the Asymptotic Regression Model

Figure 6.9 gives the approximate 95% confidence intervals on parameters of model (6.16) for each tree. It is clear that for each parameter, all confidence intervals do not overlap. This suggests that the random effects for all tree parameters may be necessary for asymptotic regression model. Models with different structures are fitted and compared using the likelihood ratio tests. The AIC of the models that consider each of ϕ_1, ϕ_2, and ϕ_3 at a time as fixed effects are respectively 18342.28, 17850.49 and 17751.62. All these values are larger than the AIC (17405.92) of the model which consider all three parameters as mixed effects. The comparison of the model with all three parameters as mixed effects, with any of the three reduced models using likelihood ratio test, produced a p-value which is less than 0.0001 for all comparisons. We conclude that the covariance structure should not be simplified by deleting any of the random effects of model (6.16).

Figure 6.10, shows residuals versus fitted values, clone and tree age. The graphs show clear violation of the assumption of homogeneity of variances. The plot of residuals versus fitted values shows a clear pattern of variability for the within-group error. The residuals also fluctuate with tree age and the variance of residuals is not the same for the two clones.

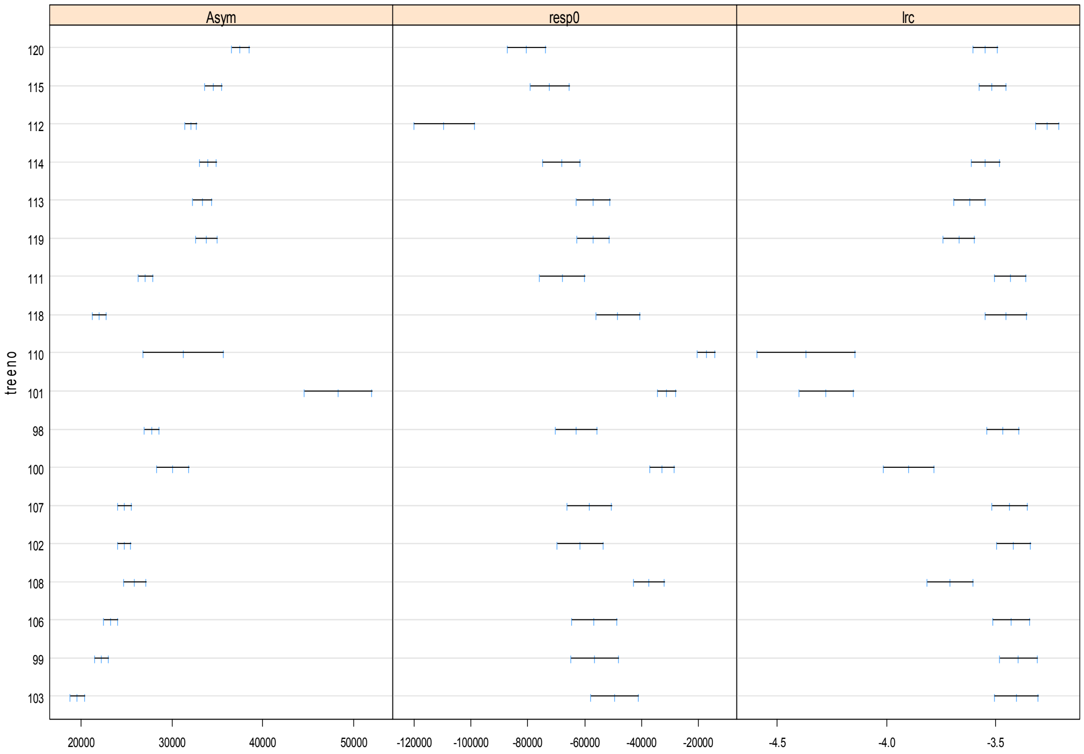

Figure 6. 9 Ninety five percent confidence intervals on the asymptotic regression model (6.16) based on individual tree fit. The parameters in this graph are related to the parameters of asymptotic regression model as follows ($\phi_1 = Asym$, $\phi_2 = resp0$ and $\phi_3 = lrc$).

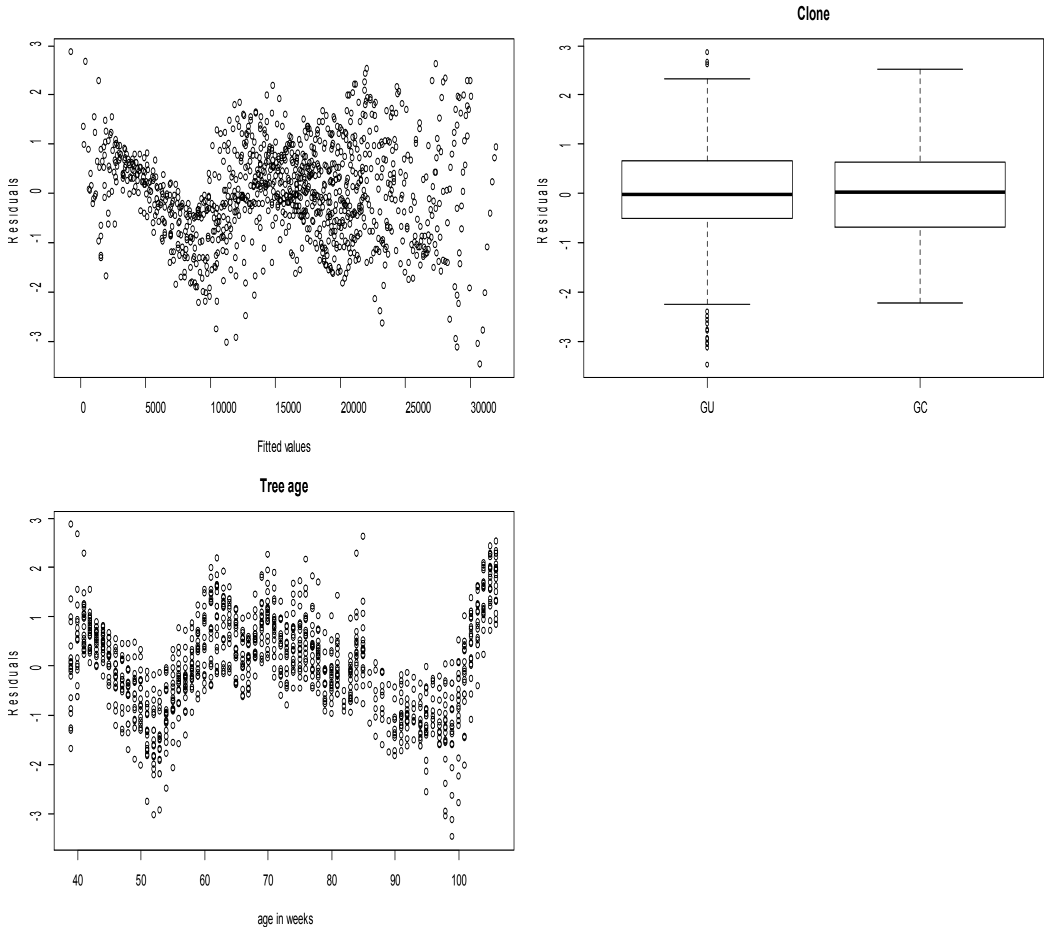

Figure 6.10 Model validation graphs for the model with all three parameters as mixed effects for the asymptotic regression model (6.16).

The within group heterogeneity was modelled using different variance functions and different correlation structures as discussed in sections 5.6 and section 5.7. The model with the different variance of residuals for each time point appears to be the best fit. The AIC of this model is 16983.87 which is the smallest of all the models fitted so far for asymptotic regression model. We assessed the adequacy of this model by plotting the standardized residuals against the fitted values, tree age and clone as shown in Figure 6.11. There is no clear indication for the departure from nonlinear mixed model assumption. The model with a different variance for each time point adequately fits the within-group heteroscedasticity.

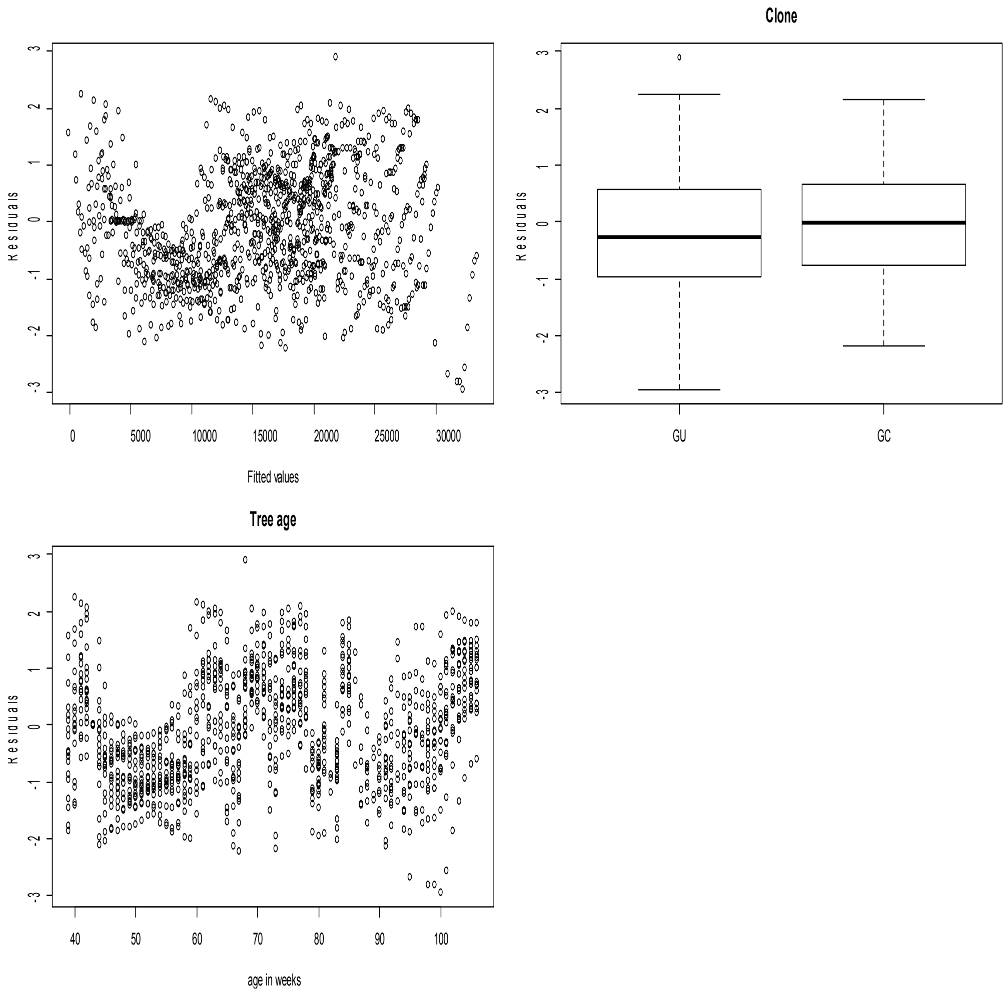

Figure 6.11 Model validation graphs for the extended model with different variance for each time point for asymptotic regression model**.**

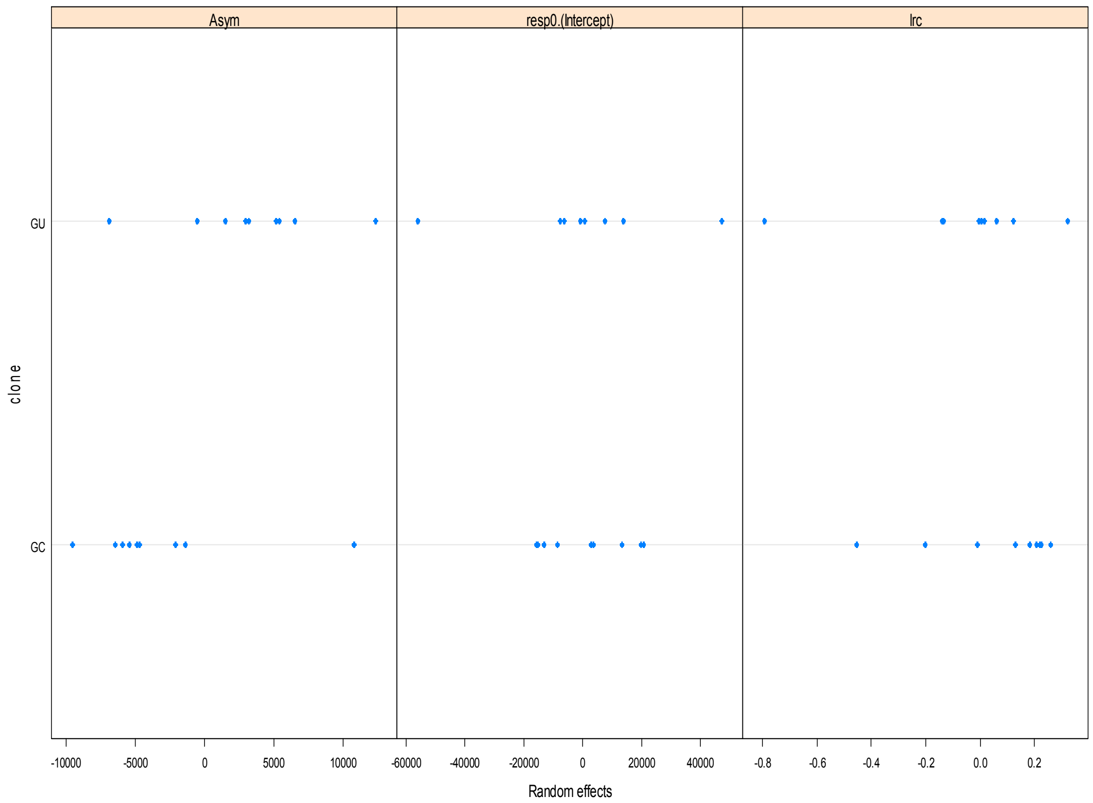

Figure 6. 12 Estimates of random effects by clone in the model different variance by tree age for asymptotic regression model.

In Figure 6.12 the asymptote parameter (Asym) seems to vary with clone. It appears that the asymptote (Asym) is larger for GU clone than for the GC clone. The dependence of each parameter on clone is modelled and tested. Clone has a significant effect on ϕ_1 and ϕ_2 of the model (p-value is less than 0.0001). The fitted model is given by

$$radius = 31023.28 + \left(-69191.06 - 31023.28\right)\exp\left[\left(-\exp(-3.52)\right) time\right] \quad for\ GU$$

$$radius = 25524.55 + \left(-53546.28 - 25524.55\right)\exp\left[\left(-\exp(-3.52)\right) time\right] \quad for\ GC$$

The assumption of normality for the within group errors was assessed using the normal probability plot of residuals. The normal probability plot of residuals is shown in Figure 6.13. Close examination of the behaviour of the two plots (see Zewotir and Galpin, 2004) shows that the normality assumption is somewhat plausible. The Shapiro-Wilk normality test (W = 0.9974, p-value=0.07) also suggests there is no violation in the assumption of normality.

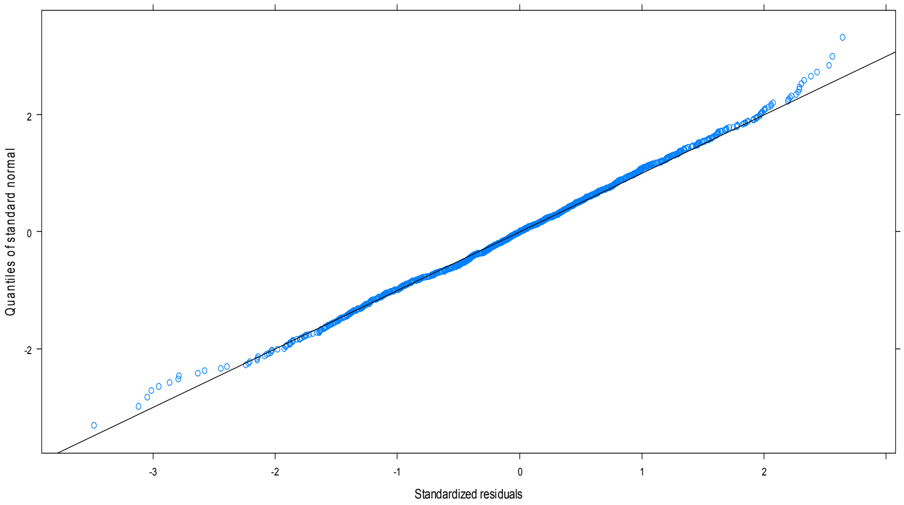

Figure 6.13 Normal probability plot of residuals by clone for model 6.16.

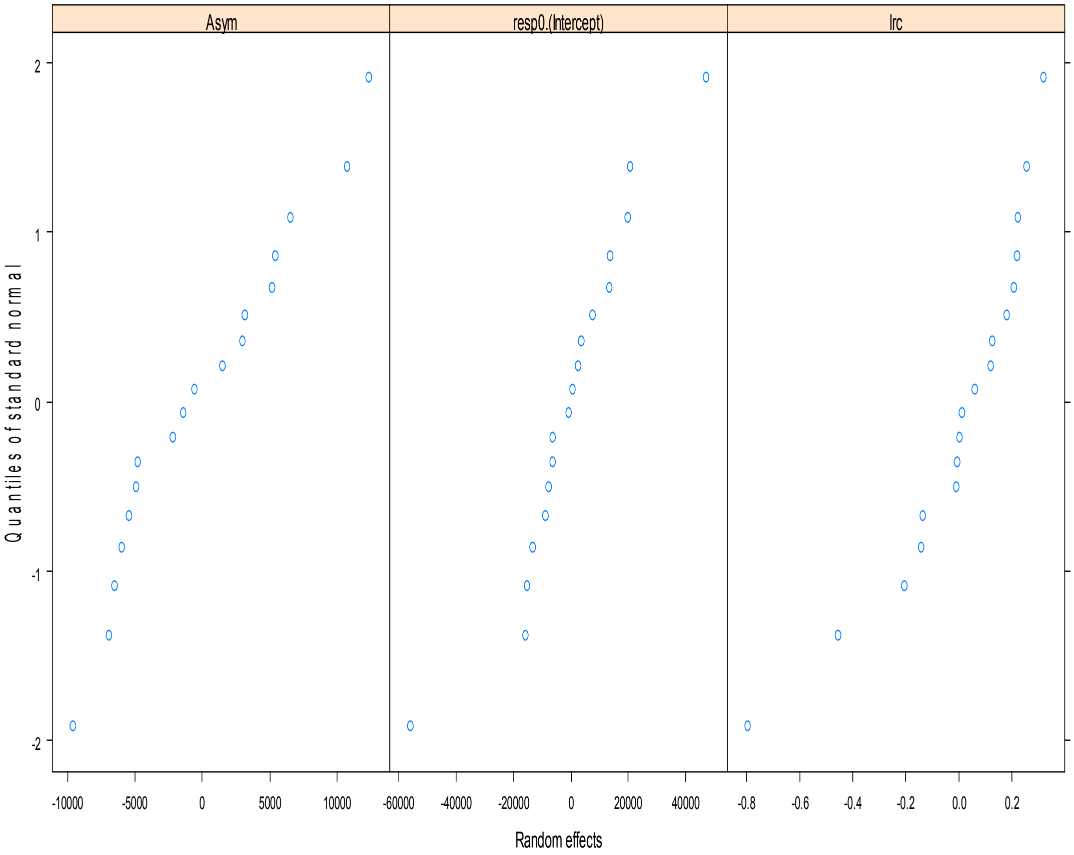

Figure 6.14 Normal probability plot of random effects for model 6.16.

The investigation of the marginal normality of the corresponding random effects was also made. The normal probability plots of the random effects are indicated in Figure 6.14. The assumption of normality seems reasonable for all three random effects. The p-values reported for the Shapiro Wilk test are 0.4, 0.16 and 0.1 respectively for random effect associated with ϕ_1, ϕ_2 and ϕ_3 of model (6.16) .

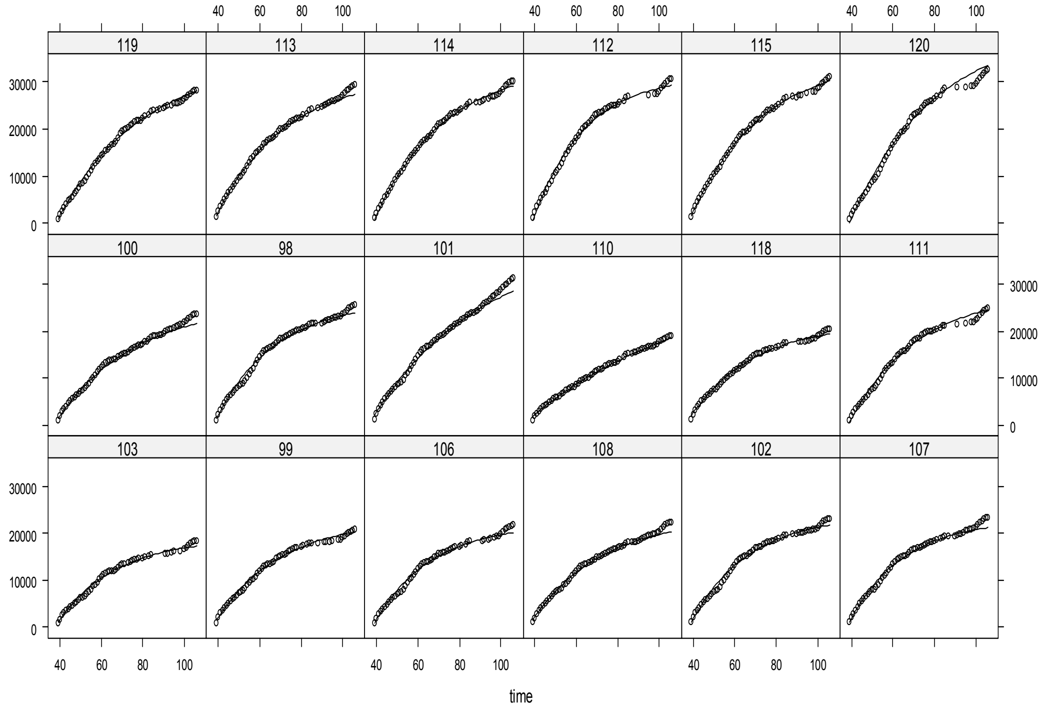

Figure 6. 15 Plots of the fitted model and observed values for each tree using the asymptotic regression model (6.16).

The adequacy of asymptotic regression model, at individual tree level, was checked. The plot of the augmented predictions, by tree, was used as an assessment for adequacy of the logistic growth model (Figure 6.15). The predicted values closely matched the observed radial growth measurements demonstrating the appropriateness of the model. Moreover, the linear regression between the observed and fitted values, which had a $R^2 = 0.9936$, suggested that the overall model fit was good (Figure 6.16).

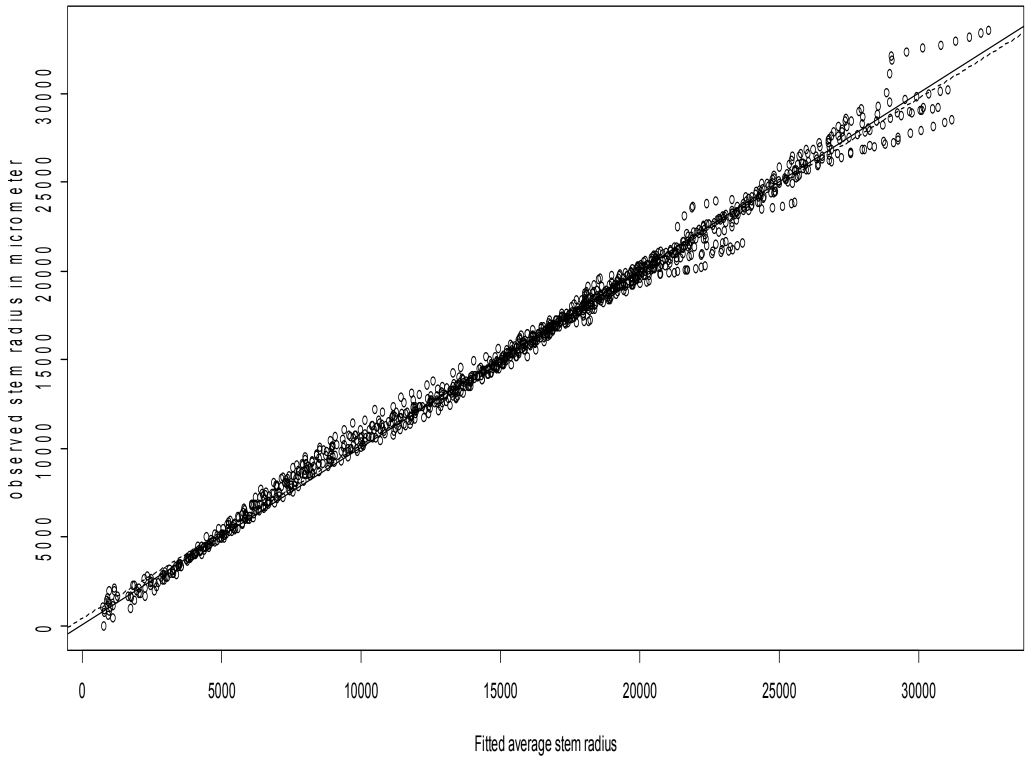

Figure 6. 16 scatter plot of the fitted versus observed average stem radius. The dashed line is the estimated regression line between the observed and fitted values. (Fitted =446.4+0.976 observed) and the solid line is the 1: 1 line.

Table 6. 3 Fixed effects parameter estimates for the asymptotic regression model (6.16).

Fixed effects	Estimate	Standard error	Degree of freedom	t-value	P-value
Asymptote-intercept (ϕ_1)	31023.28	1656.56	1095	18.73	0.000
Asymptote-clone (slope)	-5498.73	2175.38	1095	-2.53	0.010
resp0-intercept (ϕ_2)	-69191.06	5700.30	1095	-12.14	0.000
resp0-clone ϕ_2	15644.78	5523.57	1095	2.83	0.005
Lrc Scale parameter (ϕ_3)	-3.52	0.065	1095	-54.34	0.000

The summary (Table 6.3) for the fitted model suggests that clone did have a significant effect on the asymptotes of the asymptotic regression model. There is no significant effect of clone on the growth rate parameter (ϕ_3). The overall average stem radius at the end of the juvenile stage of the tree is 31023.28 with the 95% confidence interval [27780.17, 34266.39] for the GU clone. The estimate for GC clone is 25524.55 with the corresponding 95% confidence interval [18022.62, 33026.48] (Table 6.4). The asymptote for the GU clone is larger than the asymptote for the GC clone.

Table 6. 4 Summary of the fixed effects parameter estimation results for the asymptotic regression model (6.16).

Fixed effects	LCL^a	Estimated	UCL^a
Asymptote-intercept (ϕ_1)	27780.17	31023.28	34266.39
Asymptote-clone (slope)	-9757.55	-5498.73	-1239.91
resp0-intercept(ϕ_2)	-80350.75	-69191.06	-58031.37
resp0-clone ϕ_2	4831.09	15644.78	26458.47
Lrc Scale parameter (ϕ_3)	-3.65	-3.52	-3.39

a) LCL, approximated 95% lower confidence limit; UCL, approximated 95% upper confidence limit.

6.9 Results of Fitting the Gompertz Curve

The first question to be addressed in the analysis is which parameters should be treated as random effects and which were purely fixed effects. A separate fit for each tree was made and inter-tree variability was assessed using the individual confidence intervals. Since several repeated measurements were considered for each tree, the data have sufficient observations to have meaningful parameter estimates in the individual fits.

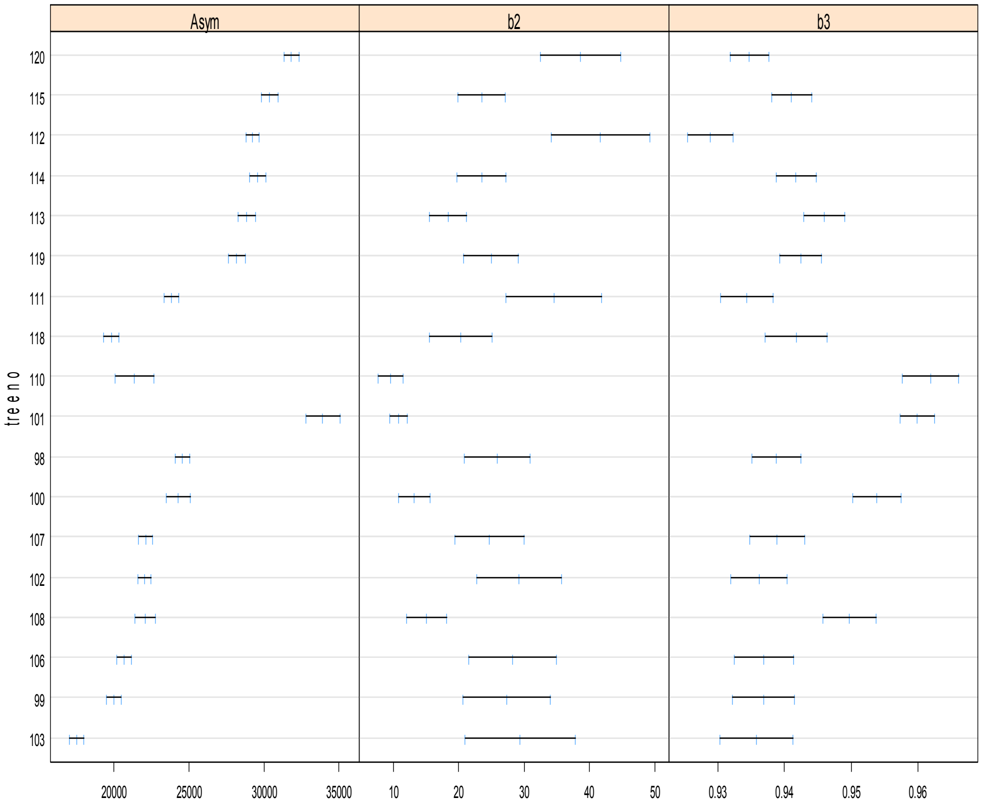

Figure 6.17 Ninety five percent confidence intervals on the parameters of the Gompertz model based on individual tree fit. The parameters in the graph are related to the parameters in the Gompertz model as follows ($\phi_1 = Asym, \phi_2 = b_2$ and $\phi_3 = b_3$).

The approximate 95% confidence intervals for parameters of model (6.23) for each tree are presented in Figure 6.17. It was clear for each parameter that all confidence intervals did not overlap. This suggests that the random effects for all three parameters may be necessary. Using the alternative approach, different prospective nested models were fitted and compared using the the Akaike Information Criterion (AIC) (Sakamoto et al., 1986). This alternative approach was considered for the parameters of model (6.23). The first model was fitted with each of ϕ_1 , ϕ_2, and ϕ_3 as mixed effects, called model I. The resulting AIC was 17777.77. From a reduced form of model I, with only ϕ_1 and ϕ_2 as mixed, an AIC of 18027.40 was obtained. This is model II. The model with ϕ_1 and ϕ_3 as mixed effects was also considered. The resulting AIC was 17982.21. This is model III. Finally, the model with ϕ_2 and ϕ_3 as mixed effects was considered and its AIC was equal to 19041.99. This represented model IV. The AIC of the models that considered each of ϕ_1, ϕ_2, and ϕ_3 at a time as fixed effects were 19041.99, 17982.21 and 18027.40, respectively. All these values are greater than the AIC of model I. This gave a clear indication that the elimination of any of these random effects has an enormous influence on the quality of the fit. The comparison of model I with any of the three reduced models (Models II, III and IV) using the likelihood ratio test, resulted in a p-value which was less than 0.0001 for all comparisons. It was established that the covariance structure should not be streamlined by deleting any of the random effects of model (6.23). This is in agreement with the conclusions of the individual fits analysis discussed using the approximate confidence intervals. A model with random effects for all three parameters was selected for random effect covariance structure. In the plots of residuals versus fitted values,clone and tree age, noticeable violation of the assumption of homogeneity of variances is realised (Figure 6.18). The residuals also fluctuate with tree age and the variance of residuals is not the same for the two clones.

The within group heterogeneity was modelled using different variance functions and different correlation structures as discussed in section 5.6 and 5.7. The model with the different variance of residuals for each time point appears to be the best fit among those models for which convergence is achieved. The AIC of this model is 16496.5, which is the smallest value of all the models fitted. The adequacy of this model was assessed by plotting the standardized residuals against the fitted values, tree age and clone as shown in Figure 6.19.

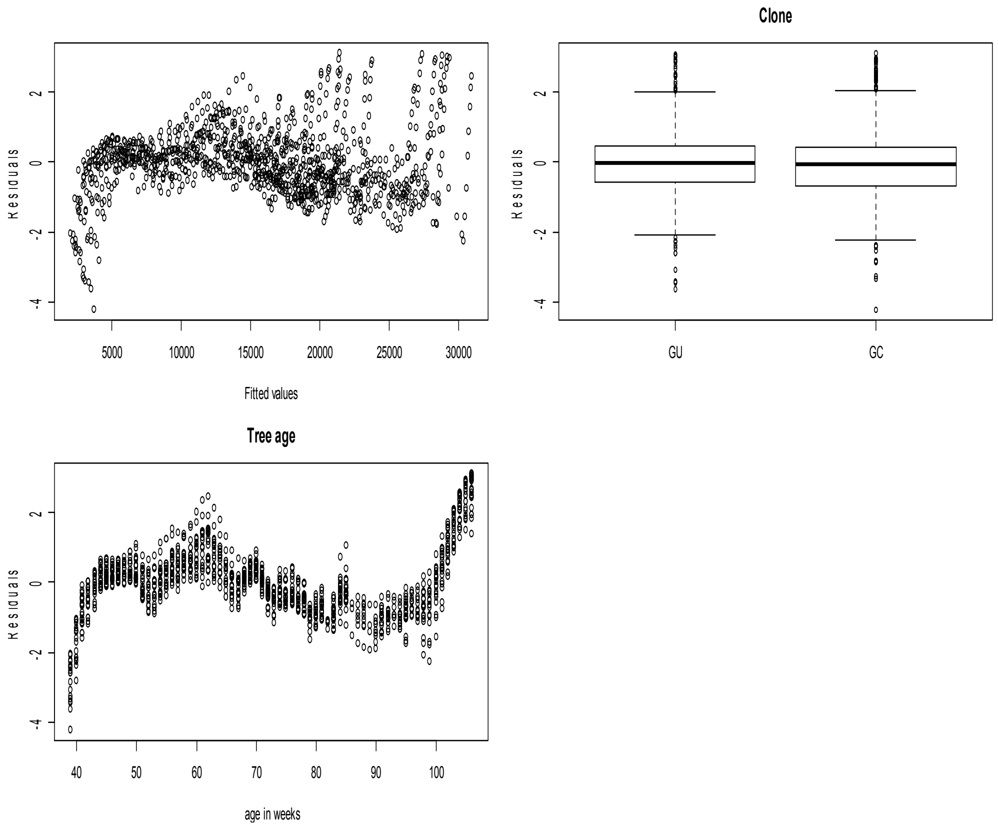

Figure 6.18 : Model validation graphs for the model with all tree parameters as mixed effects for Gompertz model.

There is a huge improvement of the validation graphs. There is no clear indication for the departure from the nonlinear mixed model assumption. The model with different variance for each time point adequately fits the within-group heteroscedasticity. The primary question of interest for the data at hand is the possible pattern between the growth in stem radius and the type of clone.

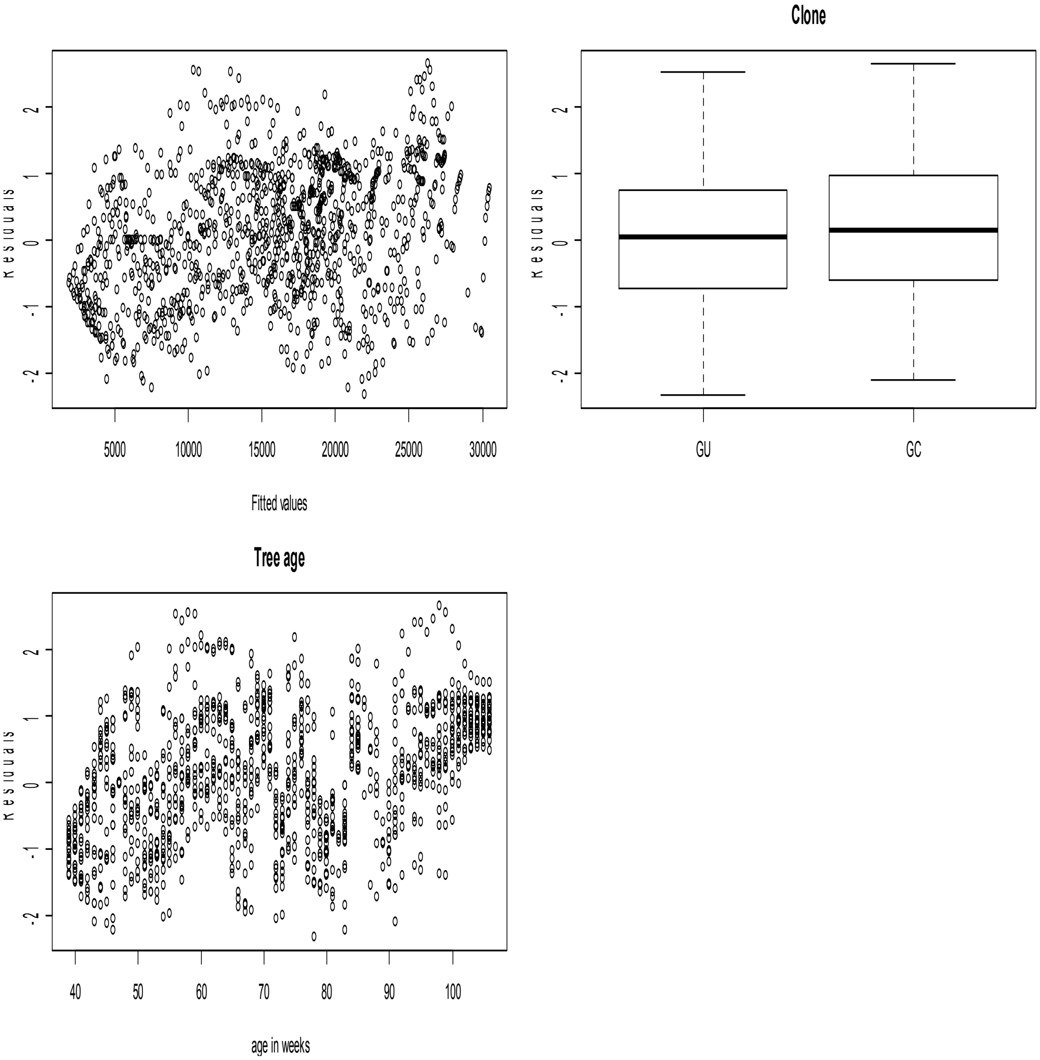

Figure 6.19 Model validation graphs for the extended model with different variances at each time point (Gompertz model).

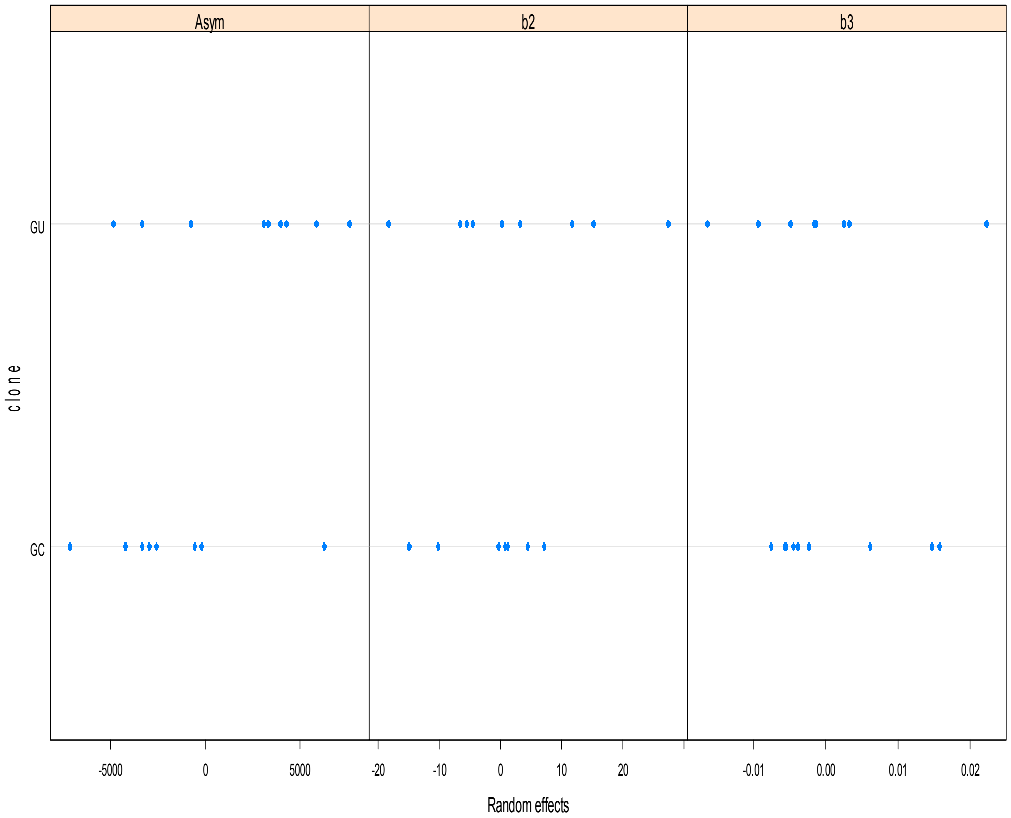

Figure 6.20 Estimates of random effects by clone for the model with different variances by age (Gompertz model). The parameters in the graph are related to the parameters in Gompertz model as follows ($\phi_1 = Asym,\ \phi_2 = b_2$ and $\phi_3 = b_3$).

From the plot of the estimates of random effects by clone, it appears that the asymptote (Asym) for the GU clone is larger than that of the GC clone (Figure 6.20). Some differences between the GU and the GC clones is observed for the remaining parameters ϕ_2 and ϕ_3. The dependence of all three parameters on clone is modeled. The significance of clone for fixed effects is also assessed by comparing the models with clone effect and without clone effect using the likelihood ratio statistics. Clone has significant effect on the asymptote (ϕ_1) of the model (p-value is equal to 0.014). Moreover, the ANOVA table (Table 6.5) for the fitted model suggests that clone had a significant effect on the asymptotes of the Gompertz curve while, no significant effect of clone is observed for the remaining parameters.

Table 6. 5 ANOVA table for the fitted nonlinear Gompertz curve

Parameter	Estimate	Standard error	Degree of freedom	t-value	P-value
Asymptote-intercept(ϕ_1)	25938.38	1252.68	1095	20.71	0.000
Asymptote-clone(slope)	-4326.60	1765.02	1095	-2.45	0.015
b2.(Intercept) (ϕ_2)	29.40	2.75	1095	10.68	0.000
b2-clone(slope)	-3.33	1.84	1095	-1.81	0.070
b3.(Intercept) (ϕ_3)	0.94	0.002	1095	421.16	0.000

The fitted Gompertz model is given by:

$$radius = 25938.32 \exp^{\left(-29.4 \times 0.94^{t}\right)} \quad for \quad GU \quad clone$$

$$radius = 21611.78 \exp^{\left(-29.4 \times 0.94^{t}\right)} \quad for \quad GC \quad clone$$

From model (6.19) the estimated rate of growth in stem radius for the two clones (where “y” stands for stem radius) is given by:

$$\frac{d\hat{y}}{d\hat{t}} = 47185.46 \times (0.94)^{t} e^{-29.4 \times (0.94)^{t}} \; for \; GU \;\; clone$$

$$\frac{d\hat{y}}{d\hat{t}} = 39314.79 \times (0.94)^{t} e^{-29.4 \times (0.94)^{t}} \; for \; GC \;\; clone$$

The estimated rates of growth curves indicated that the GU clone grows faster than the GC clone during the entire juvenile stage. This suggests that the GU clone has a faster growth potential than the GC clone in the specific area and environment where growth took place. The overall asymptotic average stem radius towards the end of the juvenile stage of the tree was 25938.32 and 21611.78 for GU clone and GC clones, respectively. Statistical significance of the fixed effect parameters of the final nonlinear mixed Gompertz model was also determined by evaluating the 95% asymptotic confidence intervals of the estimated parameters (Table 6.6).

Table 6. 6 Summary of the fixed effects parameter estimation results for the fitted Gompertz together with 95% confidence interval.

Fixed effects	LCL^a	Estimated	UCL^b
Asymptote-intercept(ϕ_1)	23485.968	25938.382	28390.796
Asymptote-clone(slope)	-7782.049	-4326.600	-871.152
b2.(Intercept) (ϕ_2)	24.010	29.396	34.782
b2-clone(slope)	-6.931	-3.330	0.271
b3.(Intercept) (ϕ_3)	0.934	0.938	0.943

[a] LCL, approximated 95% lower confidence limit; [b] UCL, approximated 95% upper confidence limit.

The null hypothesis that the parameter $H_o : \phi_j = 0$ will be rejected when the 95% asymptotic confidence interval of ϕ_j does not include zero. Clone has a significant negative slope for the asymptote, which indicates the asymptote for the GU clone is larger than that of the GC clone. This translates to the better productive capacities of the GU clone compared to the GC clone. For the other two parameters the 95% confidence interval for the slope of clone includes zero which indicates that there is no significant difference between the two clones with regard to these parameters. By applying equation (6.22) to the parameter estimates, the result revealed that the average stem radius (for both clones) reached the inflection point about 55 weeks after the first measurement was taken. Applying equation (6.23), the relative growth rate for both clones is estimated by

$$relative\ \ growth = -\ln(0.94) \times 29.4 \times 0.94^{t}.$$

This meant that for both clones, the relative growth rate decreased with time and that the two clones grew in a similar manner.

The assumption of normality for the within group errors was assessed using the normal probability plot of residuals (Figure 6.21). Close examination of the behaviour of the two plots (see Zewotir and Galpin, 2004) showed that the normality assumption is plausible.

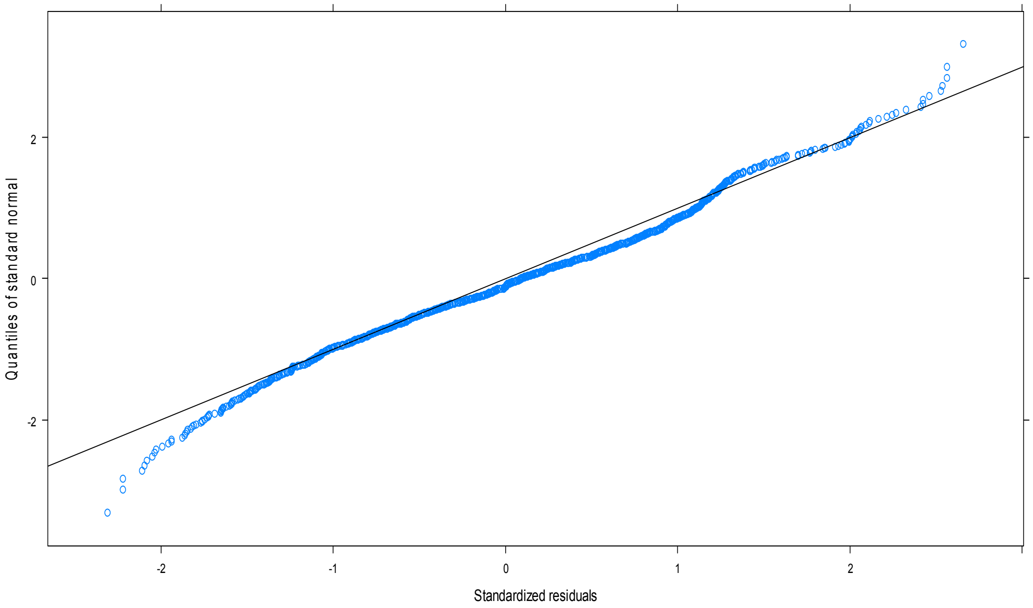

Figure 6.21 Normal probability plot of residuals for the Gompertz model.

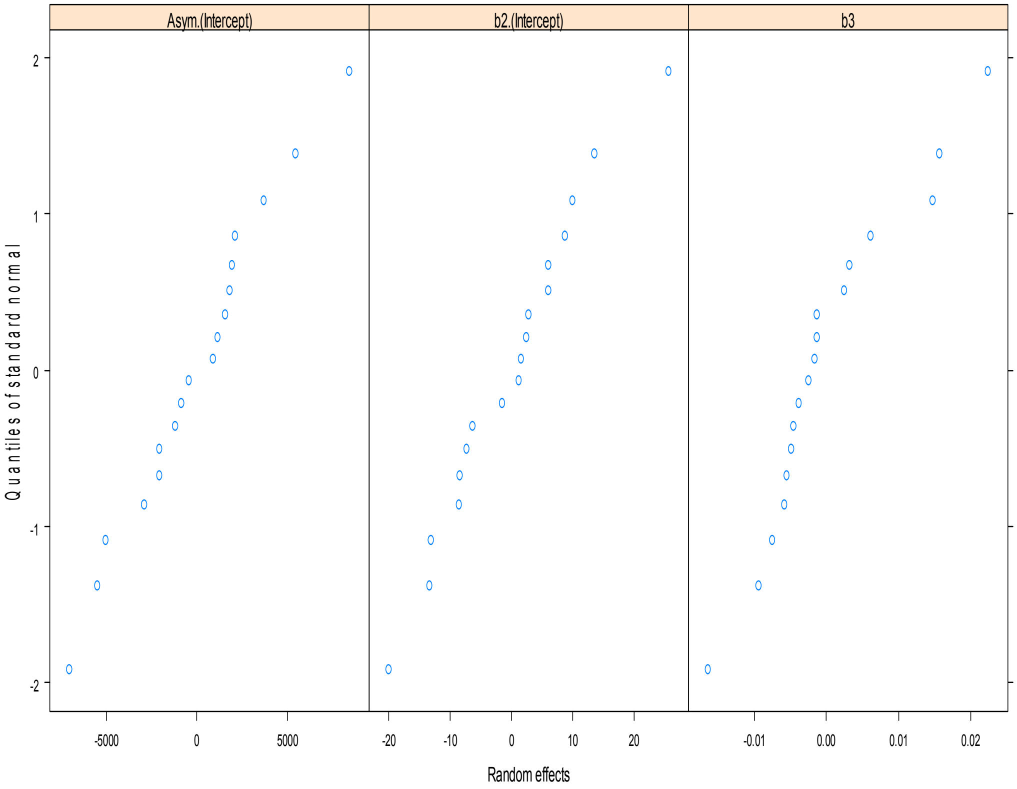

Figure 6.22 Normal probability plot of random effects for the Gompertz model

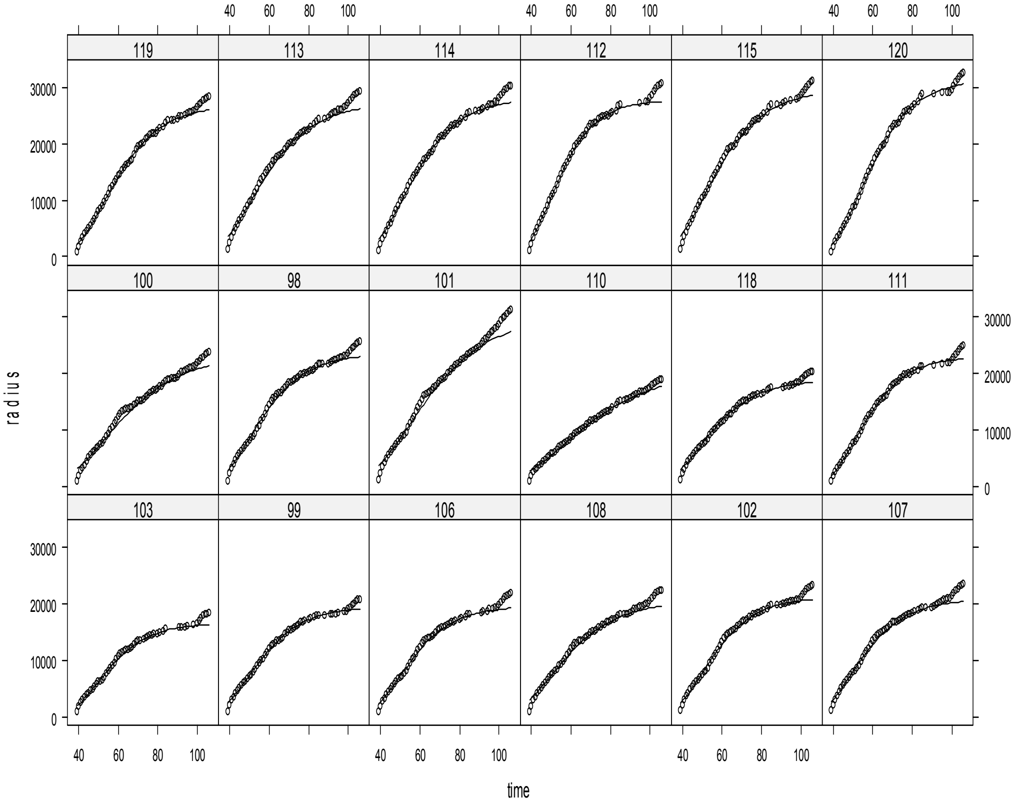

Figure 6.23 Plots of the fitted values of the Gompertz model and observed values for each tree.

The investigation of the marginal normality of the corresponding random effects was also made. The normal probability plots of the random effects are indicated in Figure 6.22. The assumption of normality seemed reasonable for all three random effects. The adequacy of the three parameter Gompertz model, at individual tree level, was checked. The plot of the augmented predictions, by tree, was used as an assessment for adequacy of the Gompertz growth model (Figure 6.23). The predicted values closely matched the observed radial growth measurements, demonstrating the acceptability of the model. Moreover, the linear regression between the observed and fitted values, which had an $R^2 = 0.99$, suggested that the overall model fit was good (Figure 6.24).

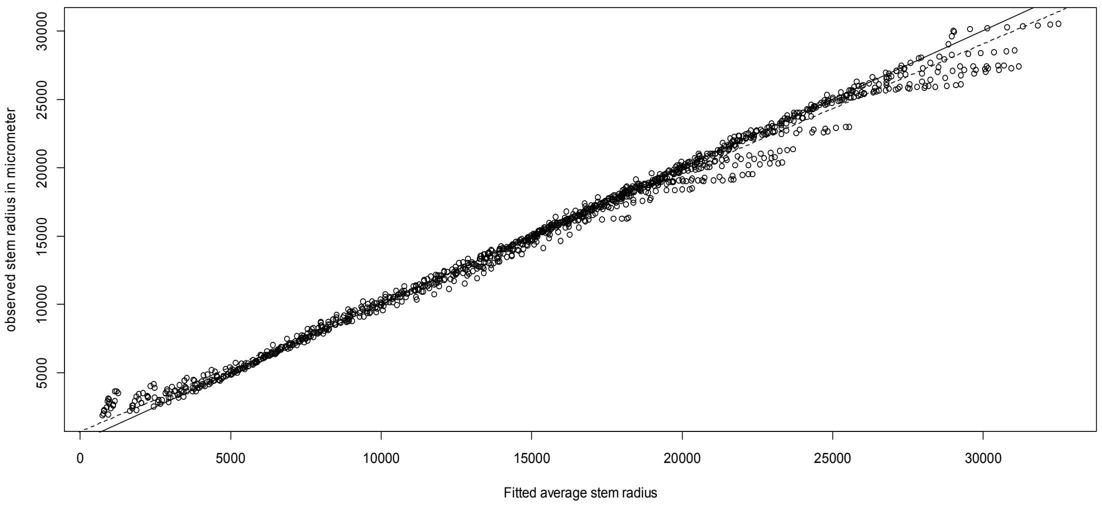

Figure 6.24 Scatter plot of the fitted versus observed average stem radius. The dashed line is the estimated regression line between the observed and fitted values (Fitted = 708.8+0.943 observed) and the solid line is the 1:1 line.

All three nonlinear growth curves fit the data well. The effect of clone on the parameters of each growth curve was studied. This analysis suggests that the GU clone has a larger stem radial measure than the GC clone during the entire juvenile stage. Although only one clone from each hybrid cross was tested in this study, the faster growth characteristics of the GU clone points to improved genetics of this hybrid cross and to its potential ability to better exploit available resources, making it an economically viable hybrid cross as reported elsewhere (Galloway, 2003). In addition to being able to describe the data well, the nonlinear growth curves used in this study were also biologically meaningful.

6.10 Comparison of Results of Nonlinear Mixed Models and Fractional Polynomial Model.

Conventional polynomial models can be used to approximate the nonlinear growth curves. However, the approximation can only be valid within the observed range of data. On the other hand, nonlinear models provide more reliable predictions for the response variable outside the observed range of the data. The model parameters in nonlinear growth models generally have a natural physical interpretation. Nonlinear growth models generally use fewer parameters than the competitor linear model, such as a polynomial, giving a more parsimonious description of the data. The flexibility of these nonlinear models does not come without cost. Because the random effects are allowed to enter the model nonlinearly, the marginal likelihood function, obtained by integrating the joint density of the response and the random effects, with respect to the random effects, does not have a closed-form expression, as in the linear mixed effects model. This computation will be even more complicated when the effects of more than one covariate is introduced in the modelling process. Consequently, an approximate likelihood function needs to be used for the estimation of the parameters, leading to more computationally intensive estimation algorithms and *less reliable inference results* (Pinheiro and Bates, 2000). The extended fractional polynomials are similar to conventional polynomials in that their time transformations are power functions, but the exponents are not restricted to integers and can be negative numbers and fractions. Moreover, they might be useful to model nonlinear growth trends with smooth curves. Compared to nonlinear mixed models, fractional polynomials have less computational difficulty. The estimation of parameter and inference for fractional polynomials can be performed under linear mixed models framework with less computational difficulty. Some interesting features of fractional polynomials include parsimony, a wide variety of curve shapes for low order models, and the ability to approximate asymptotes (Long and Ryoo, 2010). Fractional polynomials are applied within the context of the linear mixed models as this model has a number of positives, such as the accommodation of missing data and parsimony of covariance structure (Fitzmaurice et al., 2004, chap. 8). In this thesis fractional polynomial and nonlinear mixed models are used.

A Loess smoothed curve was used to compare the fit of the fractional polynomial model (as a function of tree age and clone) with that of nonlinear mixed model. Figures 6.25 to 6.27 compare the fractional polynomial model to the nonlinear model. By superimposing the Loess smoothed curve in each case, a comparison between the nonlinear models and fractional polynomial model was made. The fractional polynomial model fit is almost as close as the nonlinear growth curves to the Loess smoothed curve. This implies that the fractional polynomial model performs almost equally and even better for some parts of the data. This indicates that the fractional polynomial model offers a good fit to the data.

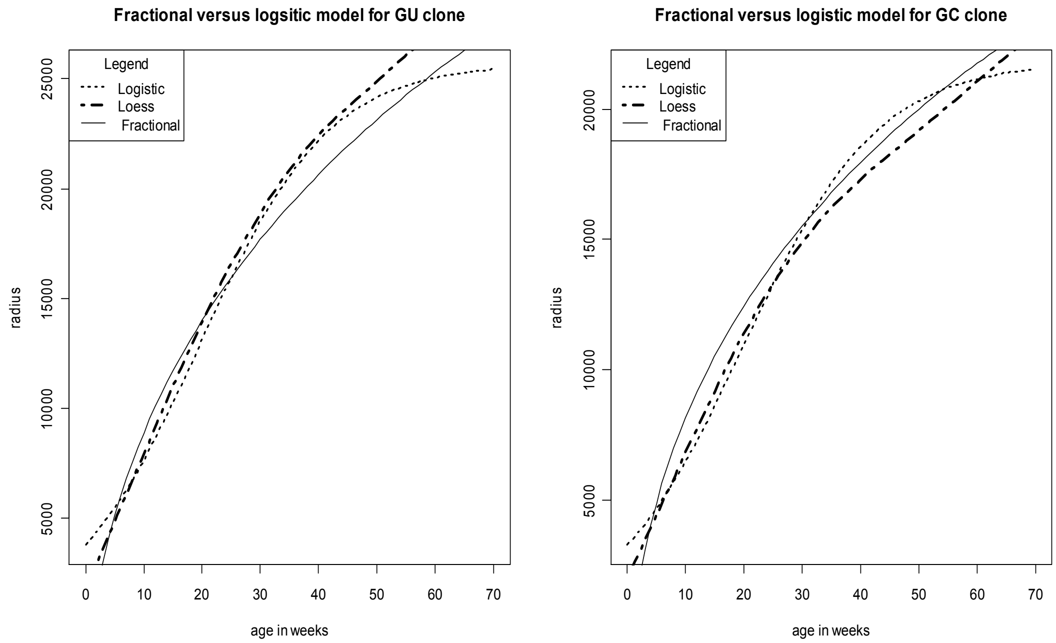

Figure 6.25 Plot of fitted three parameter logistic versus fractional polynomial model.

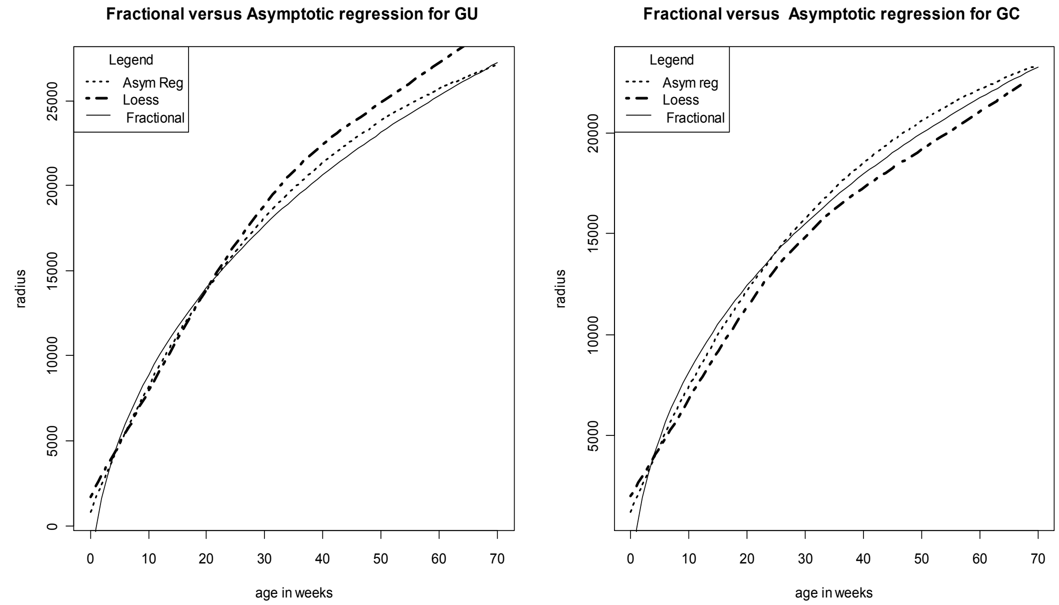

Figure 6.26 Plot of asymptotic regression versus fractional polynomial model.

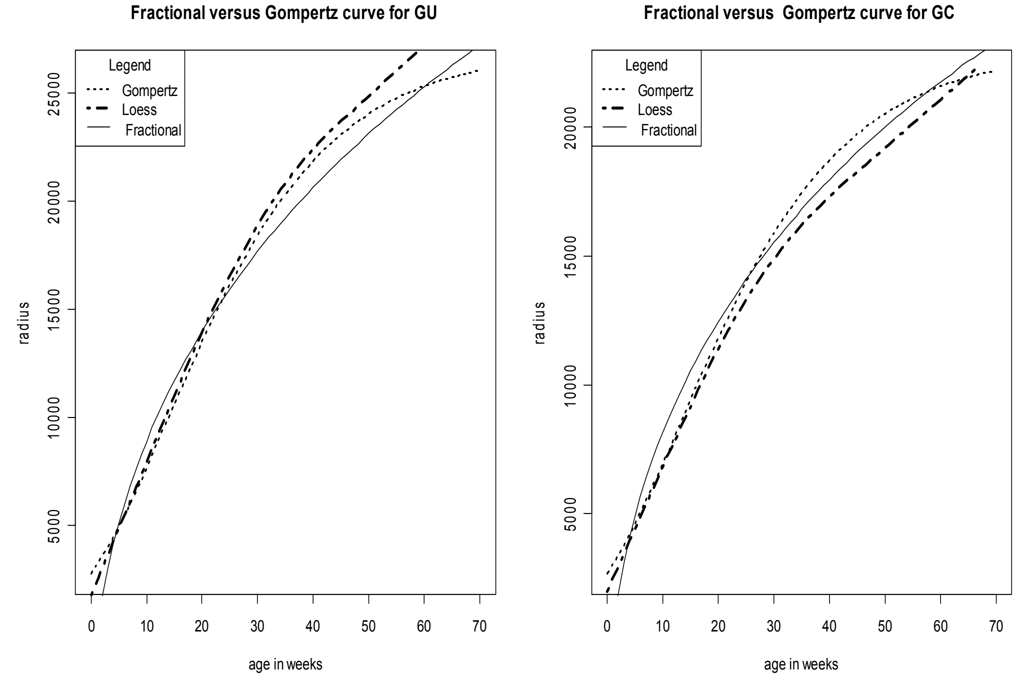

Figure 6.27 Plot of Gompertz curve versus fractional polynomial model.

6.11 Summary

Based on descriptive and graphical exploratory analysis, appropriate nonlinear growth functions were identified. The nonlinear growth curves were fitted to individual trees under consideration and the presence of random effects for each parameter of the (three parameter logistic, asymptotic regression and Gompertz) growth curves were assessed graphically. Following the graphical assessment, the selection of random effects was made by fitting different prospective models and comparing these nested models using likelihood ratio tests or information criterion statistics. These resulted in the significance of all three random effects for all growth curves. Model validation graphs showed that the within-group errors were heteroscedastic in all three cases. The extended nonlinear mixed effects models with heteroscedastic, correlated within group error, were fitted for all three growth curves. The models with the heterogeneous variance that varies with tree age were found to be the best fitting models. A comparison of the nonlinear model's fit, to the fit of the fractional polynomial model was made. The Loess smoothing technique (Cleveland, 1979) was used to compare the nonlinear growth fit with the fractional polynomial model. It was found that the fractional polynomial model was almost as good as that of the nonlinear model in fitting the data. This indicates that the fractional polynomial is as competent as the nonlinear model. This performance of fractional polynomials coupled with less computational difficulty suggests that they might be more useful when the objective is to model nonlinear growth.

All the models considered from Chapter 3 to Chapter 6 are parametric models. All these models deal with only global effects. It may be interesting to consider more flexible models that reflect both global and local effects. Consequently, the application of the semi-parametric models is reviewed and discussed in Chapter 7.

Chapter 7

Semi-Parametric Mixed Models

7.1 Introduction

Statistical methods like normal regression models, the logistic regression model for binary data and Cox's proportional hazards model for survival data assume a linear, or some parametric form, for the covariate effects. However, in several applications, this assumption of linear dependence of the response on the predictors is not appropriate. In the last two chapters, we reviewed and fitted stem radius data using parametric regression methods for longitudinal data. These parametric models provide a powerful tool for modelling the relationship between the responses and the covariates. However, parametric models suffer from inflexibility in modelling complicated relationships between the responses and covariates. In parametric methods, the form of the underlying relationship must be known in advance except for the values of a finite number of parameters. That means the relationship between the mean of the longitudinal response and the covariates is fully parametric.

The main drawback of parametric modelling is that it may be too restrictive or limited for many practical cases. This limitation has motivated a demand for developing nonparametric regression methods for analysis of longitudinal data. These methods can help to estimate a more flexible functional form between the responses and covariates from the data. Consequently, complicated relationships between longitudinal responses and covariates can possibly be captured from the data. The main idea behind the nonparametric approach is to let the data decide the most suitable form of the functions. According to Wu and Zhang, (2006), nonparametric and parametric regression methods should not be regarded as competitors, instead they complement each other. In some situations, nonparametric techniques can be used to validate or suggest a parametric model. A combination of both nonparametric and parametric methods is more powerful than any single method in many practical applications.

Although parametric models may be restrictive for some applications, nonparametric models may be too flexible to make concise conclusions in comparison with parsimonious parametric models. Semi-parametric models are good compromises and retain nice features of both the parametric and non-parametric models (Fan and Li, 2004).

Significant changes in non-parametric and semi-parametric regression methods for longitudinal data have taken place in the past 15 years. The presence of the within-subject correlation among repeated measures over time presents major challenges in developing kernel and spline smoothing methods for longitudinal data (Lin and Carroll, 2008). As a result, the extension of classical local likelihood based kernel methods and their natural local estimating equation fails to account for the within-subject correlation. This leads to the development of a non-local kernel estimator. Some advanced kernel and spline based methods for longitudinal data, have been developed recently. One such method is the extension of spline smoothing to longitudinal data. This extension entails clearly accounting for the within-subject correlation in building the penalized likelihood function. In this thesis, the focus is on a class of splines referred to as penalized splines. The motivation for focusing on penalized splines is:

i) that penalized splines are direct extensions of linear models.

ii) that they are closely connected with linear mixed models.

iii) their mixed model representation makes their extension to the longitudinal setting relatively straightforward.

Ruppert, Wand and Carroll (2003) described a very flexible semi-parametric regression approach using the linear mixed model representation of penalized splines. The generalized additive models (Hastie and Tibshirani, 1986, 1990) are among those widely used nonparametric methods for independent data. The generalized additive models (GAM) can be represented using penalized regression splines. GAM models with continuous response are called additive models. Additive models replace the linear relationship between the response and covariates to a relationship between the response and sum of smooth functions of covariates.

References on additive models or more generally on generalized additive models (GAM) are Hastie and Tibshirani, 1986; Keele, 2008; Faraway, 2006; Wood, 2006a; and Wood, 2011. The underlying assumption on GAM models is that the data are independent, which is not the case for longitudinal data. The extended form of GAM is called the generalized additive mixed mode (GAMM). A GAMM model with a Gaussian response is called additive mixed model (AMM). The aim of this chapter is to review AMM models and fit them to stem radius data. In order to develop a better understanding of AMM, a brief overview of generalized additive models (GAMs) for independent data is provided.

7.2 Smoothing Functions

To begin with the simplest smooth function, we considered a model containing one smooth function of one covariate,

$$y_i = f(x_i) + \varepsilon_i \qquad (7.1)$$

Where y_i is a response variable, x_i is a covariate, f a smooth function and the ε_i are independent and identically distributed random variables with mean zero and constant variance.

To estimate f in the linear modelling context, it is necessary to choose a basis, defining the space of functions of which f (or a close approximate of it) is an element. The function f can be approximated by the linear combination of basis functions $b_j(x)$ as

$$f(x) = \sum_{j=1}^{q} b_j(x)\ \beta_j \qquad (7.2)$$

for some unknown parameter β_j. The issue of controlling the roughness or "Wiggliness" of the estimated function can be achieved by adding a "Wiggliness" penalty to the least squares fitting objective (Wood, 2006a).

That means instead of fitting the model by minimizing $(Y - X\beta)^T\ (Y - X\beta)$, the model is fitted by minimizing the following criteria.

$$(Y - X\beta)^T (Y - X\beta) + \lambda \int [f''(x)]^2\ dx \qquad (7.3)$$

The second part of equation (7.3) is a penalty and that is why the names penalized least squares and penalized smoothers are used. It contains λ and an integral over the second derivatives. The smoothness of the curve is measured by the second derivative. A high value of second derivative (f'') means that the smoother f is highly nonlinear, whereas a zero value of second derivative indicates a straight line or the perfect smooth curve. The smoothing parameter, λ, controls the trade-off between model fit and model smoothness. For λ close to ∞ the minimization of (7.3) gives a linear fit and letting λ close to zero gives un-penalized regression spline estimate. These considerations reveal that the choice of λ plays a great role in the estimation.

Since f is linear in the parameters, β_j, it can be shown that the penalty in (7.3) can be written as a quadratic form of β,

$$\int [f''(x)]^2 dx = \beta^T S \beta \quad ,$$

Where the matrix $S = \int d(x)\ d(x)^T dx$ is a matrix of known coefficients and

$$d(x) = \begin{bmatrix} b_1''(x) \\ b_2''(x) \\ . \\ . \\ . \end{bmatrix}$$

This leads us to the argument that penalized regression spline fitting problem is equivalent to minimizing

$$(Y - X\beta)^T (Y - X\beta) + \lambda\beta^T S\beta \qquad (7.4)$$

The degree of smoothness of the model is estimated by the parameter λ.

By rewriting (7.4) as

$$(Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T (X^T X + \lambda S)\beta$$

and differentiating with respect to β and equating to zero, the penalized least square estimator of β (for a given λ) can be obtained as

$$\hat{\beta} = (X^T X + \lambda S)^{-1} X^T Y$$

In principle, λ can be set by hand and the penalized likelihood maximization can be used to estimate the parameter, β. It is also possible to choose λ in a data driven way. Two basic approaches are useful: when scale parameter is known attempting to minimize the expected mean square error leads to estimation by Mallow's Cp/UBRE (Unbiased Risk Estimator); when the scale parameter is unknown then attempting to minimize prediction error leads to ordinary cross validation or generalized cross validation (GCV) (Wood, 2006a). Ruppert et al. (2003, Chap.5) deliberate on several procedures for choosing the smoothing parameter λ, and Wand (1999) derives a closed form approximation to the optimal value of λ ·

Rupert and Carroll (2000) consider spatially varying penalties and Ruppert (2002) provides recommendations for selecting the knots. Long and Wand (2004) showed that smoothing methods that use basis function can be formulated as fits in mixed model framework.

The other issue that needs discussion is the amount of smoothing for smoothing splines.

If the model has two smoothers, say

$$y_i = f_1(X_i) + f_2(Z_i) + \varepsilon_i \qquad (7.5)$$

then these two smoothers have the form

$$f_1(x_i) = \sum_{j=1}^{p} b_j(x_i)\,\beta_j \qquad f_2(z_i) = \sum_{j=1}^{p} b_j(z_i)\,\gamma_j$$

Using two smoothers in place of one smoother has an effect on the definitions of the Y , X and β in equation (7.4), but the general form remains the same. The optimization criterion with the penalty for the wiggliness becomes

$$(Y - X\beta)^T(Y - X\beta) + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta \,. \qquad (7.6)$$

This allows different amounts of wiggliness per smoothing spline. That means some smoothers can be smooth (large λ_j), whereas others may not be smooth (small λ_j). This indicates that the values of $\lambda_j s$ determines the amount of smoothing.

To get the λ_j s, the objective in (7.6) can be written as

$$(Y - X\beta)^T (Y - X\beta) + \beta^T S \beta \qquad (7.7)$$

by defining $S = \lambda_1 S_1 + \lambda_2 S_2$.

The amount of smoothing of a smoother is not expressed in terms of the λ_j s but expressed as effective degrees of freedom for a smoother.

A high value (8-10 or higher) means that the curve is highly nonlinear, whereas a smoother with 1 degree of freedom is a straight line. Technically, the matrix S , which depends on the λs , is involved in determining the effective degrees of freedom (edf) and it mirrors the algebra underpinning linear regression (Zuur et al, 2009) .

7.3 Additive Models

The additive model can be formulated by admitting the smooth function (7.1) in the classical linear regression model.

$$y = X^* \alpha + \sum_{j=1}^{p} f_j\left(x_j\right) + \varepsilon; \qquad \varepsilon \sim N(0,\ \sigma^2 I) \qquad (7.8)$$

Where $X*$ is a model matrix for the parametric components of the model, α is the corresponding parameter vector and the $f_j(\ .)$ is a smooth arbitrary function of a covariate x_j , ε is the vector of random errors. The assumptions of the additive model are the same as the assumptions in the linear model except for the assumption of linearity. These are

(1) Homoscedasticity: The error variance is the same whatever is the value of the explanatory variable.
(2) Normality: The error is normally distributed
(3) Independence: The errors are uncorrelated.

7.4 Additive Mixed Models:

The inclusion of the random effects into the additive model gives us the additive mixed model.

$$y = X^{*}\alpha + \sum_{j=1}^{p} f_j\left(x_j\right) + Zb + \varepsilon; \qquad (7.9)$$

where Z is the design matrix for random effects b.

ε is a vector of random error which is independent of b and $\varepsilon \sim N(0, R)$

$b \sim N(0, G_\theta)$. Both covariance matrices R *and* G_θ are positive definite. These matrices are also assumed to depend on a parsimonious set of covariance parameters.

The additive mixed model (AMM) that is allowed to have non-normal response will be the generalized additive mixed model (GAMM). A GAMM has the following structure

$$G(y) = X^{*}\alpha + \sum_{j=1}^{p} f_j\left(x_j\right) + Zb + \varepsilon; \qquad (7.10)$$

where G (.) is a monotonic differentiable function. A GAMM represents the model with higher flexibility and complexity, where mixed effects, smooth terms and non-normal responses are included (Lin and Zhang, 1999). These models can be viewed as additive extensions of the generalized linear mixed models.

7.5 Inference in Generalized Additive Mixed Models

Statistical inference in generalized additive mixed models comprises estimations of the non-parametric functions $f_j(.)$, the smoothing parameters, λ, and all the variance components. In the case of Gaussian response and identity link function, the estimation of non-parametric functions, smoothing and variance parameters in the context of GAMM is achieved using Restricted Maximum Likelihood (REML).

For non-Gaussian response, PQL (Penalized Quasi Likelihood) (Breslow and Clayton, 1993) and DPQL (Double Penalized Quasi Likelihood) are used to estimate the parameters and non-parametric function (Lin and Zhang, 1999). Both PQL and DPQL take their origin from maximum likelihood (ML) technique. The ML has direct application only in fixed models with Gaussian response. The maximum likelihood approach is also used in linear mixed models; however the maximum likelihood estimators (MLE) of variance are, in general, biased. First ML and REML estimation methods of linear mixed models are briefly introduced. Following the introduction of ML and REML the PQL methodology, which can be used to estimate GAMM parameters for non-normal response, is presented.

Maximum Likelihood Estimation (MLE)

Consider the following Gaussian model

$$y = X\alpha + Zb + \varepsilon .$$

The distribution of the response

$$y \sim N(X\alpha, V) \quad with \quad V = R + Z^T G_\theta Z \qquad (7.11).$$

The log-likelihood is given by

$$l(y;\ \alpha, \theta) = c - \frac{1}{2}\log(|V|) - \frac{1}{2}(y - X\alpha)^T V^{-1} (y - X\alpha) \qquad (7.12)$$

where c is a constant and θ is the vector of variance components involved in V .

The partial derivative of $l(y;\alpha,\theta)$ with respect to the parameters, θ and α can be obtained

$$\frac{\partial l}{\partial \alpha} = X^T V^{-1} y - X^T V^{-1} X\alpha \qquad (7.13)$$

$$\frac{\partial l}{\partial \theta_r} = \frac{1}{2}(y - X\alpha)^T V^{-1} \frac{\partial V}{\partial \theta_r} V^{-1}(y - X\alpha) - tr(V^{-1}\frac{\partial V}{\partial \theta_r}) \qquad (7.14)$$

where θ_r is the r-th component of θ of dimension q . Assuming that α has dimension p and $rank(X) = p$, then the MLE is obtained by solving equations (7.13) and (7.14). The MLE of α is

$$\hat{\alpha} = \left(X^T \hat{V}^{-1} X\right)^{-1} X^T \hat{V}^{-1} y \qquad (7.15)$$

This requires the estimation of V and of its components θ . The estimate, $\hat{V}$ is obtained by solving

$$y^T P \frac{\partial V}{\partial \theta_r} Py = tr(V^{-1}\frac{\partial V}{\partial \theta_r}) \qquad (7.16)$$

where $P = V^{-1} - V^{-1} X(X^T V^{-1} X)^{-1} X^T V^{-1}$. Then $\hat{\alpha}$ is obtained by plugging $\hat{V}$ into equation (7.15).

Restricted Maximum Likelihood Estimation (REML)

The maximum likelihood estimates of the variance components are biased. In contrast to the ML estimation method, REML can produce unbiased estimates of variance components. The REML estimation procedure applies transformation to the data to eliminate the fixed effects, and then uses the transformed data to estimate the variance components.

Assume $rank\,(X) = p$ and let A be and $n \times (n-p)$ matrix such that $rank\,(A) = n - p$. Then, define $z = A^T y$ where $z \in N(0,\ A^T VA)$. It follows that the log-likelihood based on z, that is the restricted log-likelihood, is given by

$$l_R\left(z;\theta\right) = c - \frac{1}{2}\log\left(\left| A^T VA \right|\right) - \frac{1}{2} z^T \left(A^T VA\right) z \qquad (7.17)$$

By differentiating the $l_R(z,\theta)$, one obtains in terms of y

$$\frac{\partial l_R}{\partial \theta_i} = \frac{1}{2}\left(y^T P \frac{\partial V}{\partial \theta_i} P y - tr(P \frac{\partial V}{\partial \theta_i}) \right) \qquad (7.18)$$

where $P = A\left(A^T VA\right)^{-1} A^T$ *and* $i = 1, \ldots q$. Although the REML estimator is defined through a transformation matrix A, it does not depend on A. That means the estimator does not depend on the transformation matrix. The restricted log-likelihood function (7.17) is a function of θ only, which means the REML method is a method of estimating θ and not α, since the fixed term α is removed before the estimation. However, once the REML estimator of θ is obtained, α is usually estimated in the same way as the ML, that is, by equation (7.15) where $V = V(\hat{\theta})$ with $\hat{\theta}$ being the REML estimator (Valeria, 2011).

Both ML and REML are based on the assumption that the response is normally distributed. The assumption of normality is often easily violated in practice making the likelihood inference difficult. In the absence of the random effects and errors distributions, the likelihood function cannot be available. Even in the presence of non-normal distributions of the random effects and errors with some unknown parameters, the likelihood function can involve quite formidable difficulty in calculation and may not have an analytic appearance. Moreover, the distributional assumptions for any non-normal distribution may not hold in practice. These problems have led to the attention of methods other than maximum likelihood. One such method is the quasi-likelihood also known as Gaussian likelihood approach. The computational difficulty of the maximum likelihood method can be avoided by using quasi-likelihood. The REML estimates can be derived from a quasi-likelihood (Heyde, 1994). Therefore, the Gaussian REML estimation can be considered as a method of quasi-likelihood.

Laplace approximation

When the exact likelihood function is computationally intractable, there are no simple solutions to get the parameter estimates. One possible option is to use numerical integration techniques. Some of these are Gaussian quadrature, numerical integration like Markov chain, Monte Carlo algorithms, stochastic approximations algorithms and penalized quasi-likelihood (Zuur et al., 2009). Penalized likelihood estimation has been proposed as a computationally simple alternative to methods based on numerical quadrature, especially when the number of random effects is relatively large (Fitzmaurice et al., 2004). The key concept in quasi-likelihood is Laplace's approximation which is described below.

Suppose it is necessary to approximate an integral of the form,

$$\int \exp^{\{-q(x)\}} dx \qquad (7.19)$$

where $q(.)$ achieves its minimum value at $x = \tilde{x}$ with $q'(\tilde{x}) = 0$ and $q''(\tilde{x}) > 0$.

The quantities q' and q'' denote the gradient (that is the vector of derivatives) and Hessian (that is the matrix of second derivatives) of q, respectively. Then we have

$$\int \exp^{\{-q(x)\}} dx \approx c \mid q''(\tilde{x}) \mid^{\frac{-1}{2}} \exp^{\{-q(\tilde{x})\}} \qquad (7.20)$$

where c is a constant depending only on the dimension of the integral and $\mid q''(\tilde{x}) \mid$ denotes the determinant of the Hessian.

Penalized Quasi-likelihood Estimation

By employing Laplace approximation, an approximated maximum likelihood can be obtained instead of the exact likelihood. Such approximated likelihood is called Penalized Quasi-Likelihood (PQL). Penalized Likelihood is essential in the case non-normal models. Following the estimation procedure by Lin and Zhang (1999), Valeria (2011) gave the following discussion. According to Lin and Zhang (1999), for a given λ and θ, the spline estimator of $f_j(.)$ maximizes the following penalized log-quasi-likelihood

$$l_{PQ}\{y;\alpha, f_1, f_2 \ldots f_{p,}\ \theta) - \frac{1}{2}\sum_{i=1}^{p} \lambda_j \int_{t_j^0}^{t_j^n} f_j''(x)^2\, dx =$$

$$l_{PQ}\{y;\alpha, f_1 \ldots f_{p,}\ \theta\} - \frac{1}{2}\sum_{j=1}^{p} \lambda_j f_j^T K_j f_j \qquad (7.21)$$

where $(t_j^0 \quad t_j^n)$ defines the range of the j^{th} covariate and K_j is the non-negative definite penalty matrix of f_j(see Green and Silverman (1994)).

Differentiating equation (7.21) with respect to α, smoothing functions and b respectively, yields to a system of equations that can be solved by Fisher scoring algorithm with working vectors of response and estimated (centred) smooth functions. Lin and Zhang (1999) proposed an alternative to (7.21), since it still requires numerical integration, the DPQL approximation (see Lin and Zhang (1999)) for more details.

These estimators can be obtained by iteratively fitting a working generalized linear mixed model (GLMM) to an updated response. The basic idea of this approach is to re-parameterize a GAMM as GLMM. In fact, the GAMM in (7.10) can be reformulated as a GLMM as follows (Valeria, 2011):

$$G(\mu^b) = X\beta + Ua + Zb \tag{7.22}$$

This is achieved by assuming that the smooth function estimation can be split into fixed and random components. That means we have $f_j = X_j \beta_j + U_j a_j$, where β_j represents the fixed effects while a_j stands for the random effects. In particular if B_k is a set of spline bases with k=1, 2…r, then the model is specified by $X = (B_K)_{k=1,2}$ and U, such a transformed matrix of remaining base matrix $B = (B_k)_{3 \le k \le r}$. But the estimation of smoothing functions, $f(.)$, needs the previous estimation of λ and θ.

The smoothing parameters, λ, and the variance components, θ, can be jointly estimated by using the marginal quasi-likelihood by extending the REML approach of Wahba (1985). They can be obtained by fitting a working linear mixed model (LMM) and REML, with $\tau = \left(\frac{1}{\lambda_1} \quad \cdots \quad \frac{1}{\lambda_p} \right)$ treated as extra–variance components in addition to θ. Then the GLMM can be fitted iteratively. Hence a marginal quasi-likelihood of

(τ, θ), $l_m PQ(y; \tau, \theta)$, can be constructed (eq. 21 in Lin and Zhang (1999)). The l_m reduces to REML under AMM (Valeria, 2011).

Equation (21) in Lin and Zhang (1999) sometimes has serious numerical problems and it must be approximated using methods like Laplace's approximation. This approximation corresponds to the REML log-likelihood under LMM

$$\mu^b = X\beta + Ua + Zb \qquad (7.23)$$

where a and b are random effects. It follows that τ and θ can be easily estimated by iteratively fitting model (7.23) using REML. After estimating τ and θ, it is possible to use the Best linear unbiased prediction (BLUP) estimators of β_j and a_j to construct approximate spline estimators $\hat{f}_j$ by PQL (or DPQL) (Valeria, 2011).

7.6 The Software for GAMM

Although several R packages (R core team, 2013) are developed to fit GAMM, the most versatile that can handle modelling the correlation structure is the package **mgcv** (Wood, 2006b). This uses the **nlme** implementation of nonlinear mixed models. It also fits non-Gaussian responses by calling **MASS**'s generalized linear mixed model penalized quasi-likelihood (glmmPQL). The main advantage of this package is that it is possible to include serial and/or spatial correlation structures of the random effects. The package **mgcv** (Wood, 2006b) is used to fit the additive mixed models.

7.7 Results of Fitting AMM Using One Covariate at a Time to Stem Radius Data

At the beginning the AMM that involves only tree age as an explanatory variable is considered. The estimated smoothed curve together with its 95% confidence interval is shown in Figure 7.1.

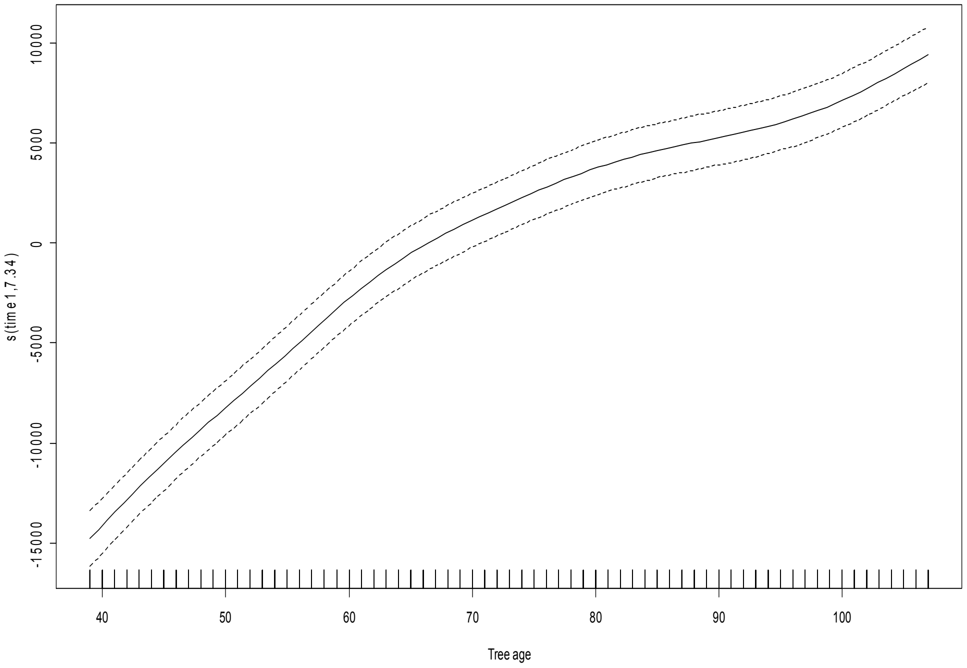

Figure 7.1 Estimated smoothing curve for the simplest AMM model (the solid line is the smoother and the dotted lines are 95% confidence intervals).

The plot of the tree age effect (Figure 7.1) indicates that the relationship between stem radius and tree age is nonlinear. Moreover, the estimated effective degree of freedom is 7.3 confirming the non-linearity of the relationship.

The non-linearity was tested using a formal test by comparing a model specifying the smooth term with a model specifying a linear trend. The difference between the two models (linear trend versus smooth terms) is statistically significant (p-value less than 0.0001). Similar results are obtained for other covariates (refer to Table 7.1).

In Table 7.1, each individual covariate is separately considered as an explanatory variable. The results of this table indicate that all covariates have a nonlinear relationship with the stem radius. The plots that indicate the relationship of the stem radius with each of the covariates are also shown from Figure 7.1 to Figure 7.6. In each of these plots the stem radius is expressed in mean deviation form, the smooth terms $(s_j(x_j)$, where x_j stands for each covariate) is centred and hence each plot represents how stem radius change relative to its mean, with change in covariate under consideration. The interpretation of the scale of the graphs is as follows: The value of zero on the vertical axis is the mean of stem radius. As the line moves away from zero in a negative direction we subtract the distance from the mean to determine the fitted value. If the line moves in a positive direction, we add the similar distance. For instance, in order to get the fitted value for radius (using Figure 7.1) when tree's age is 46 weeks we need to add the mean radius (16240.27) and a value of the smooth when age is 46 weeks (-10000) which will give us 6240.27 (micro metres). The fitted value will be around 21240 micro metres when the tree age is about 90 weeks.

Table 7. 1 Comparison of models with linear trend and models with smooth terms

Variable	Model	Degree of freedom	Log-Likelihood	Likelihood .ratio test statistic	P-value	Effective Degree of Freedom (edf)
Tree age	Linear trend	4	-11363.66			
	Smooth term	5	-10866.56	994.21	<.0001	7.34
Temperature	Linear trend	4	-12751.85			
	Smooth term	5	-12663.37	176.94	<.0001	8.64
Rainfall	Linear trend	4	-12761.92			
	Smooth term	5	-12680.62	162.59	<.0001	7.78
Relative Humidity	Linear trend	4	-12587.87			
	Smooth term	5	-12577.78	20.17	<.0001	8.22
Solar radiation	Linear trend	4	-12723.05			
	Smooth term	5	-12719.06	7.98	0.0047	4.4
Wind speed	Linear trend	4	-12706.21			
	Smooth term	5	-12622.40	167.63	<.0001	7.80

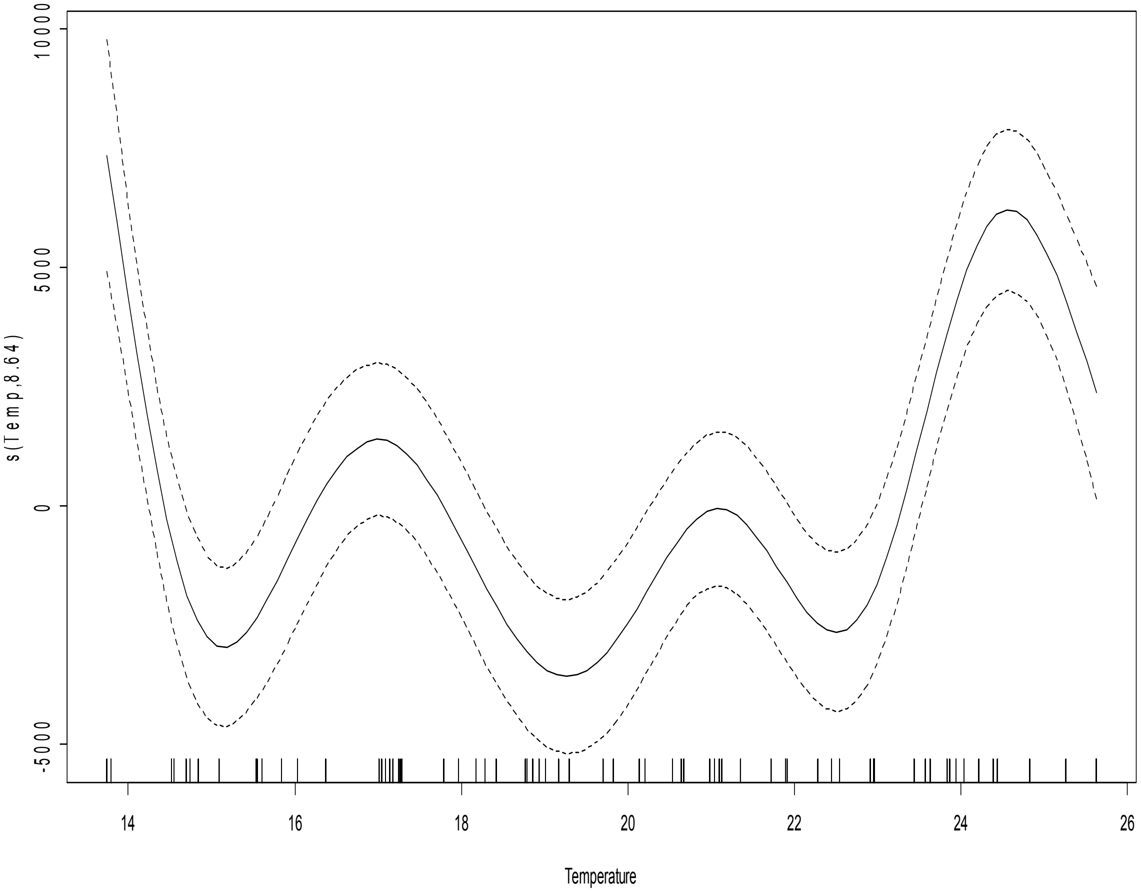

Figure 7.2 Estimated smoothing curve for the simplest GAMM model that uses only temperature as an explanatory variable (the solid line is the smoother and the dotted lines are 95% confidence intervals).

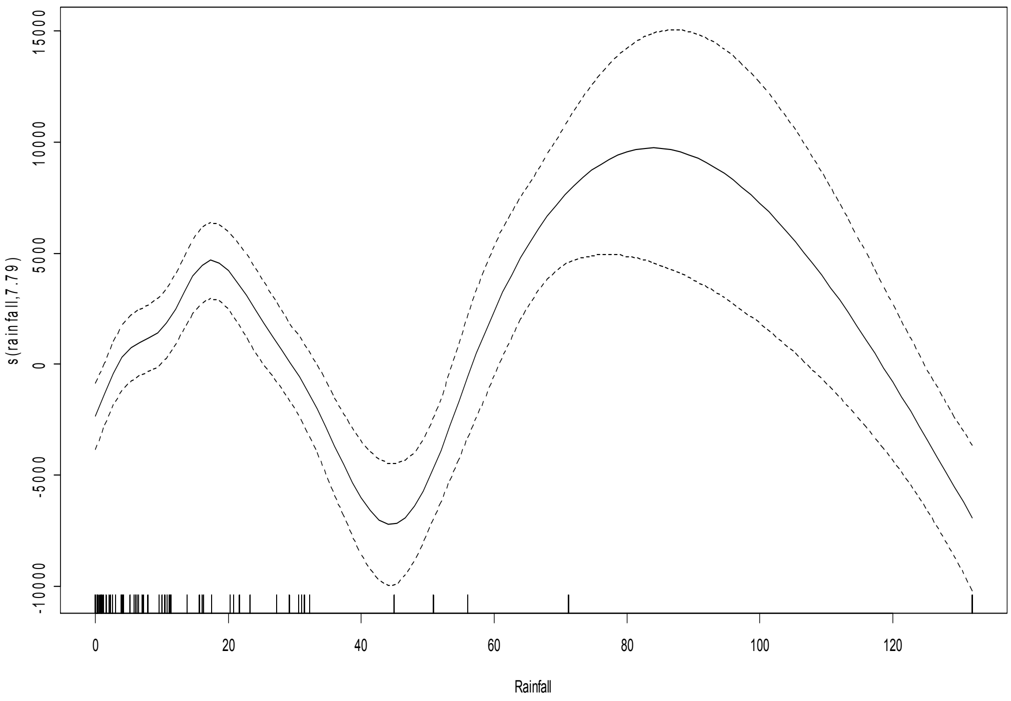

Figure 7. 3 Estimated smoothing curve for the simplest GAMM model that uses only rainfall as an explanatory variable (the solid line is the smoother and the dotted lines are 95% confidence intervals).

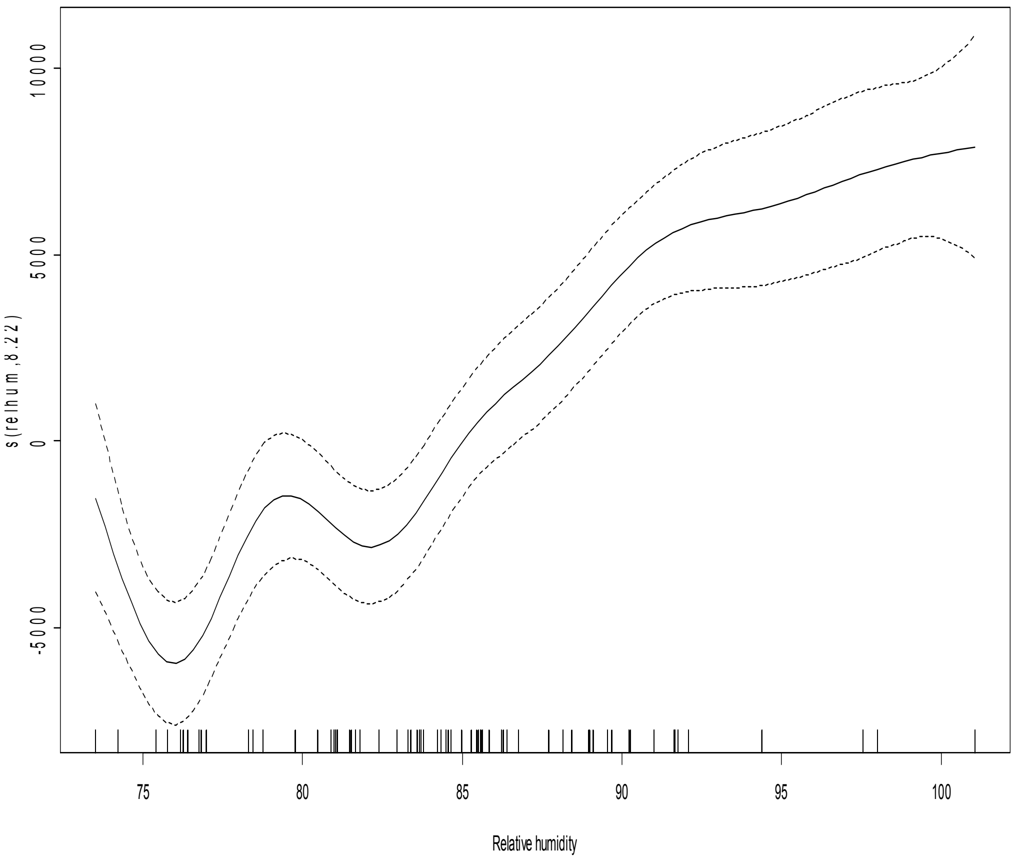

Figure 7. 4 Estimated smoothing curve for the simplest GAMM model that uses only relative humidity as an explanatory variable (the solid line is the smoother and the dotted lines are 95% confidence intervals).

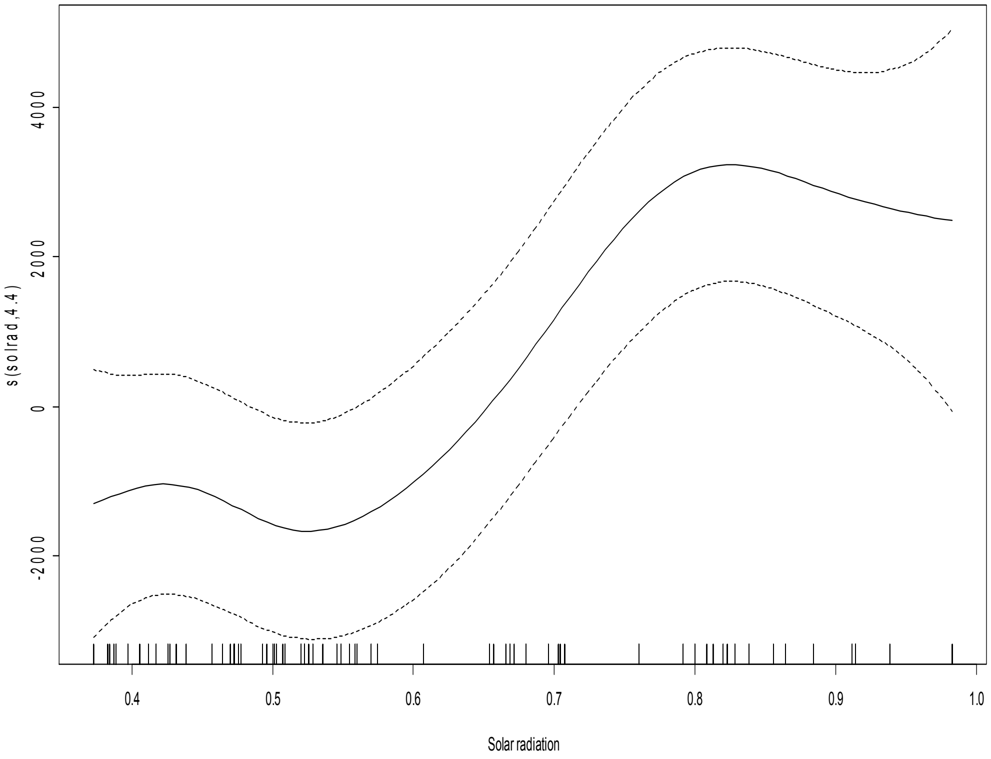

Figure 7. 5 Estimated smoothing curve for the simple GAMM model that uses only solar radiation as an explanatory variable (the solid line is the smoother and the dotted lines are 95% confidence intervals).

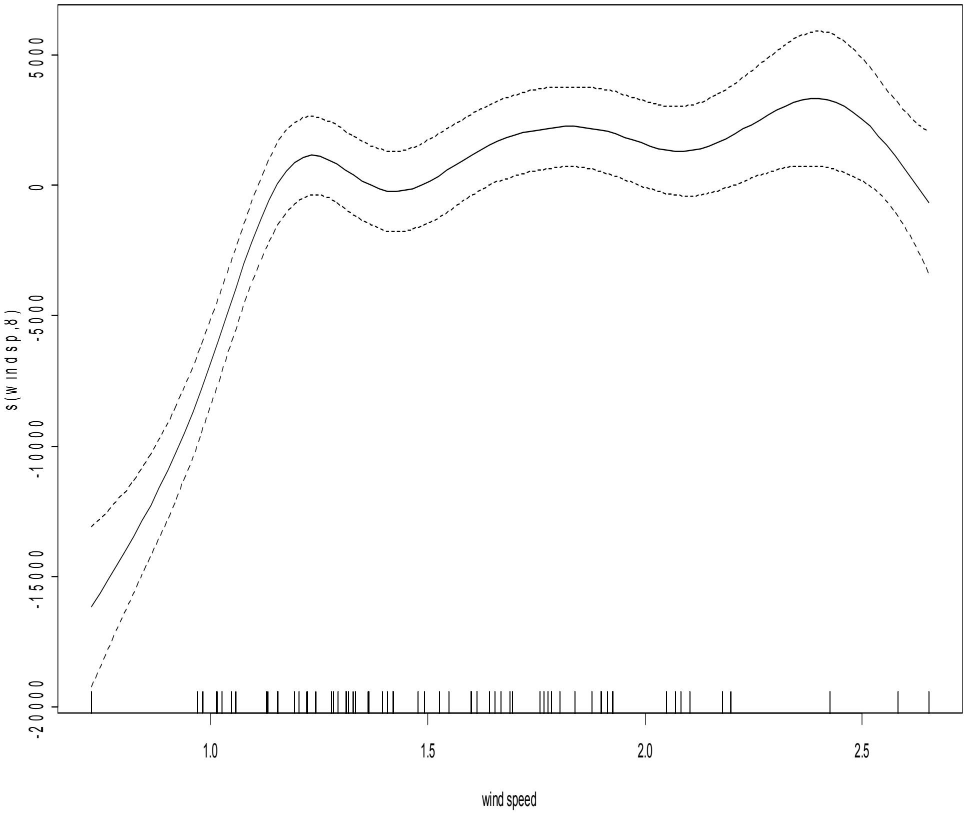

Figure 7. 6 Estimated smoothing curve for the simple GAMM model that uses only wind speed as an explanatory variable (the solid line is the smoother and the dotted lines are 95% confidence intervals).

7.8 Modelling The Effect of Tree Age for Each Clone

The effect of each covariate on stem radius may vary with clone or season. Instead of applying one smoother for both clones, a model with two smoothers (one smoother for each clone) is fitted to study the effect of tree age on stem radius. The model with clone added is better judged by likelihood ratio test statistics (255.7, df=2 and P value < 0.0001). Therefore, a model with one smoother per clone is preferable to the model with one smoother for both clones. The results of the fitted additive mixed model with two different smoothers (one per clone) are presented on Figure 7. 7 and Table 7.2.

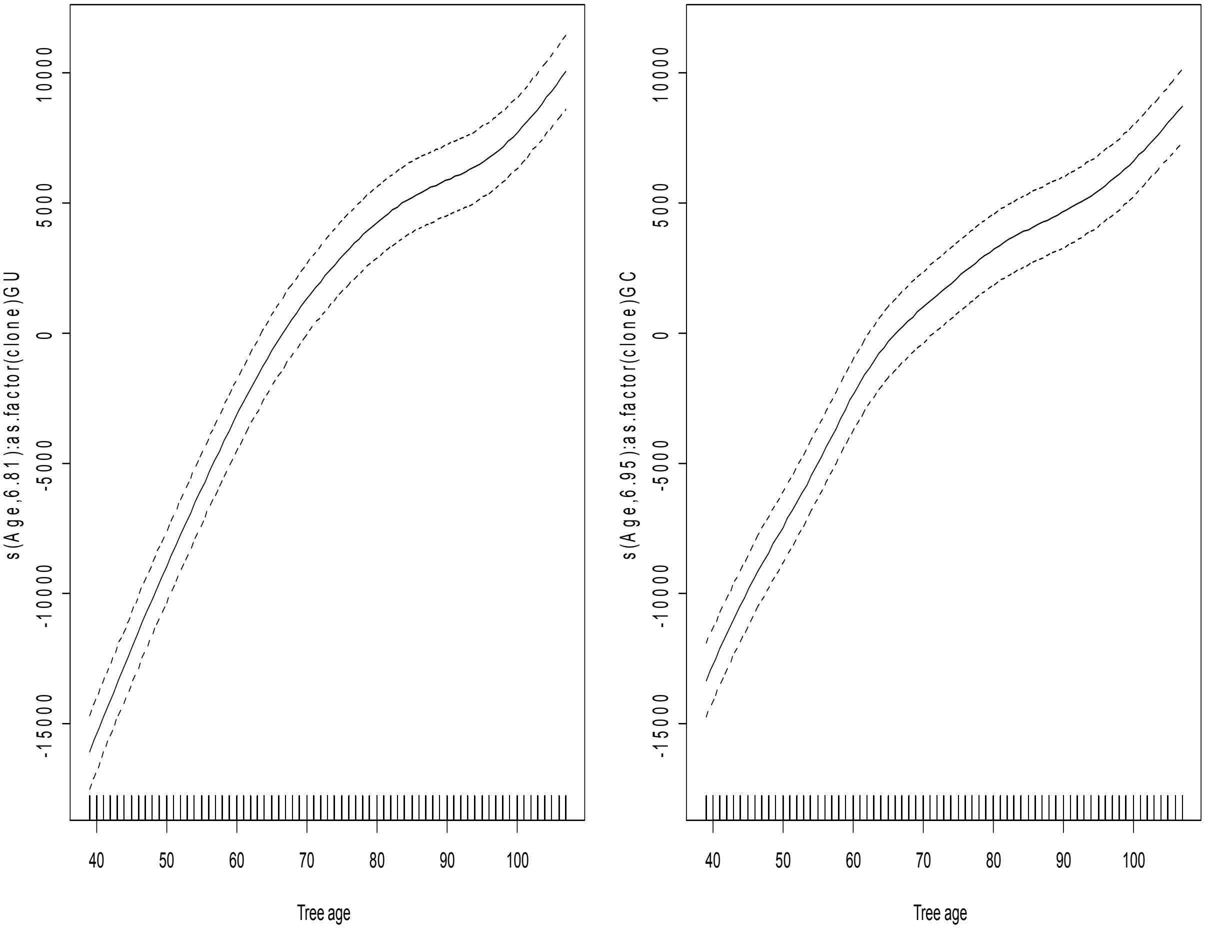

Figure 7. 7 Estimated smoothing curve for the GAMM model that uses tree age by clone as an explanatory variable (the solid line is the smoother and the dashed lines are 95% Bayesian credible intervals).

The effect of tree age is estimated as smooth curves with 6.806 and 6.954 effective degrees of freedom for GU and GC clones respectively. The p-values for both smoothed terms is very small (p-value < 2e-16) and very large value of F (see table 7. 2). This indicates that the relationship between tree age and stem radius remains nonlinear after adding the clone to the model. The adjusted R^2(more or less the square of the correlation coefficient between observed and fitted values) is 0.821. This indicates that there is a strong relationship between observed and fitted values of the model.

Table 7. 2 The fitted additive mixed model with one smoother of tree age per clone (Maximum likelihood estimates)

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	16240.3	671.6	24.18	< 2e-16 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(Age, clone=GU)	6.806	6.806	2925	<2e-16 ***
s(Age, clone=GC)	6.954	6.954	1951	<2e-16 ***
R-sq.(adj) = 0.821				
"*** " indicates significance at 0.0001				

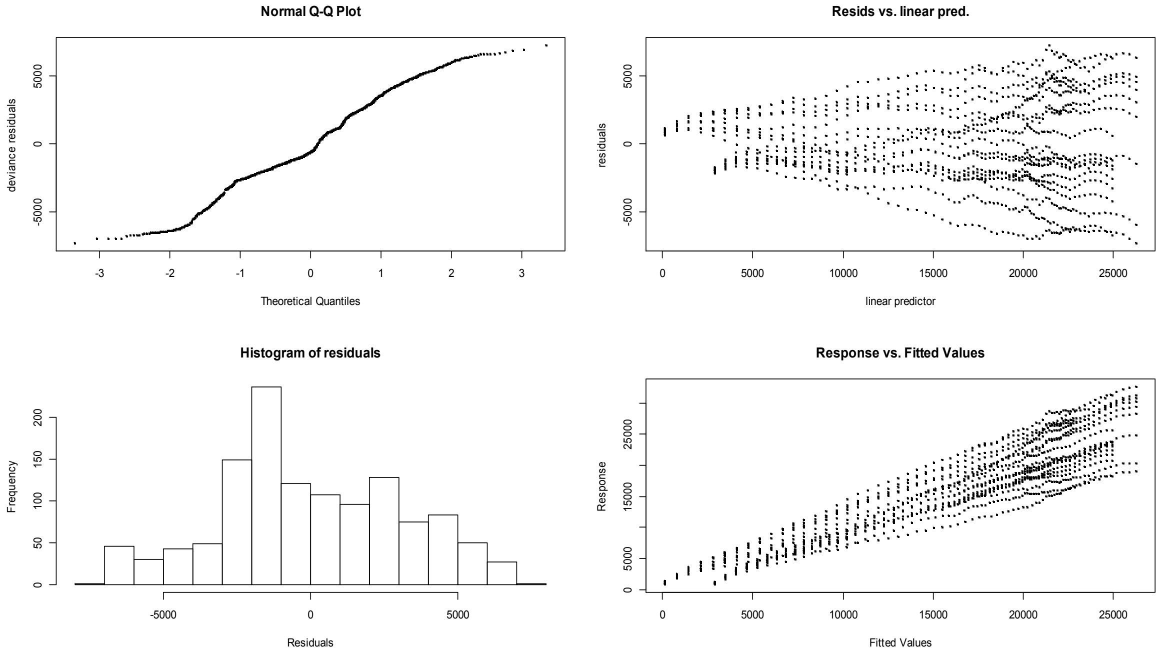

Figure 7.8 Model validation graphs for the additive mixed model that has two smooth curves of tree age (one per clone).

The QQ plot and the histogram of residuals show some non-normality (Figure 7.8). The residuals versus predictor plot shows that there is a clear violation of homogeneity of variance. The plot of the response against fitted value shows that there is a strong linear relationship between the observed response and the fitted value. Before fitting more complicated models (e.g. additive mixed models with more complex covariance structure), an attempt to extend the current model with the effect of more than one covariate was made.

7.9 Modelling The Effect of Tree Age by Season and Clone

An attempt to fit eight smoothers (one for each clone and season combination) was not successful due to numerical problems encountered. Instead a model with four smoothers for each clone is fitted after separating the data into two, namely the data for GU and the data GC clone.

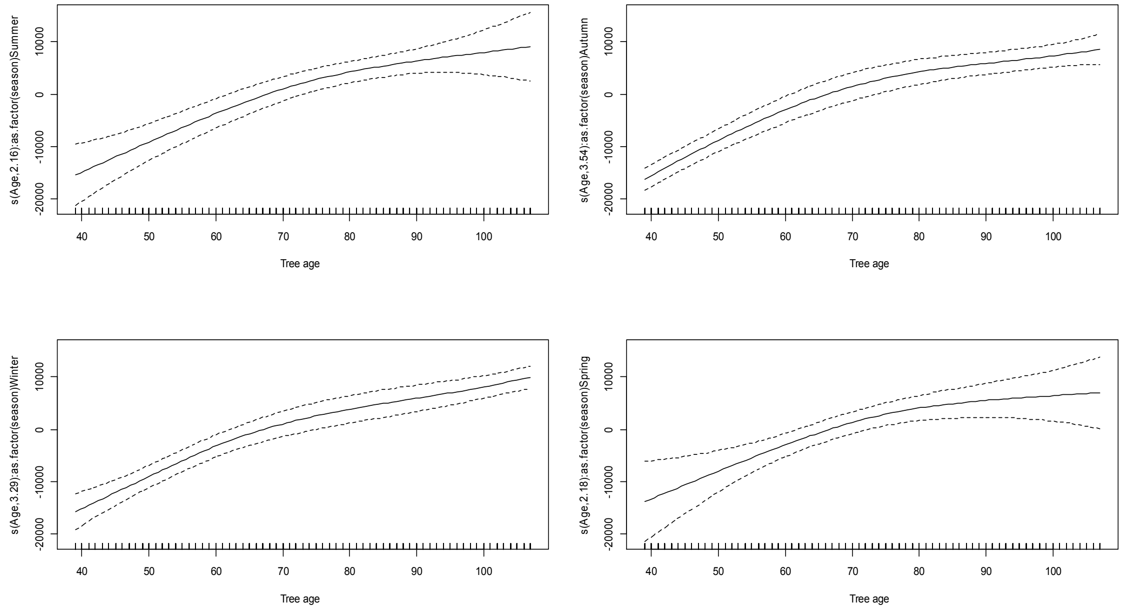

Figure 7. 9   Estimated smoothing curves and 95% confidence bands for the GAMM model that uses tree age by season for GU clone.

Table 7.3 The fitted additive mixed model with four different smoothers of tree age (one per season) for the GU clone (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	17563	1073	16.37	< 2e-16 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(Age, season = summer)	2.162	2.162	103.45	<2e-16 ***
s(Age, season = Autumn)	3.541	3.541	2469.08	<2e-16 ***
s(Age, season = Winter)	3.286	3.286	1343.93	<2e-16 ***
s(Age, season = Spring)	2.183	2.183	53.16	<2e-16 ***
R-sq.(adj) = 0.818				

The smoothers for all seasons have very small (p-value < 2e-16) and the values of the test statistic F are very large. This indicates the relationship between tree age and stem radius appears to be nonlinear for all seasons with a slight variation in the values of effective degrees of freedom (edf) (Table 7.3).

For GC clone, the smoothers for all seasons are significant (p-value < 2e-16) (see Table 7.4). This shows that the two clones grow in a similar manner which means, in both cases, the relationship between tree age and stem radius is nonlinear.

Table 7. 4 The fitted additive mixed model with four different smoothers of tree age (one per season) for the GC clone (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	15101.9	641.1	23.55	< 2e-16 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(Age, season = summer)	2.086	2.086	79.98	<2e-16 ***
s(Age, season = Autumn)	3.888	3.888	3150.49	<2e-16 ***
s(Age, season = Winter)	3.886	3.886	1715.73	<2e-16 ***
s(Age, season = Spring)	2.092	2.092	48.44	<2e-16 ***
R-sq.(adj) = 0.899				

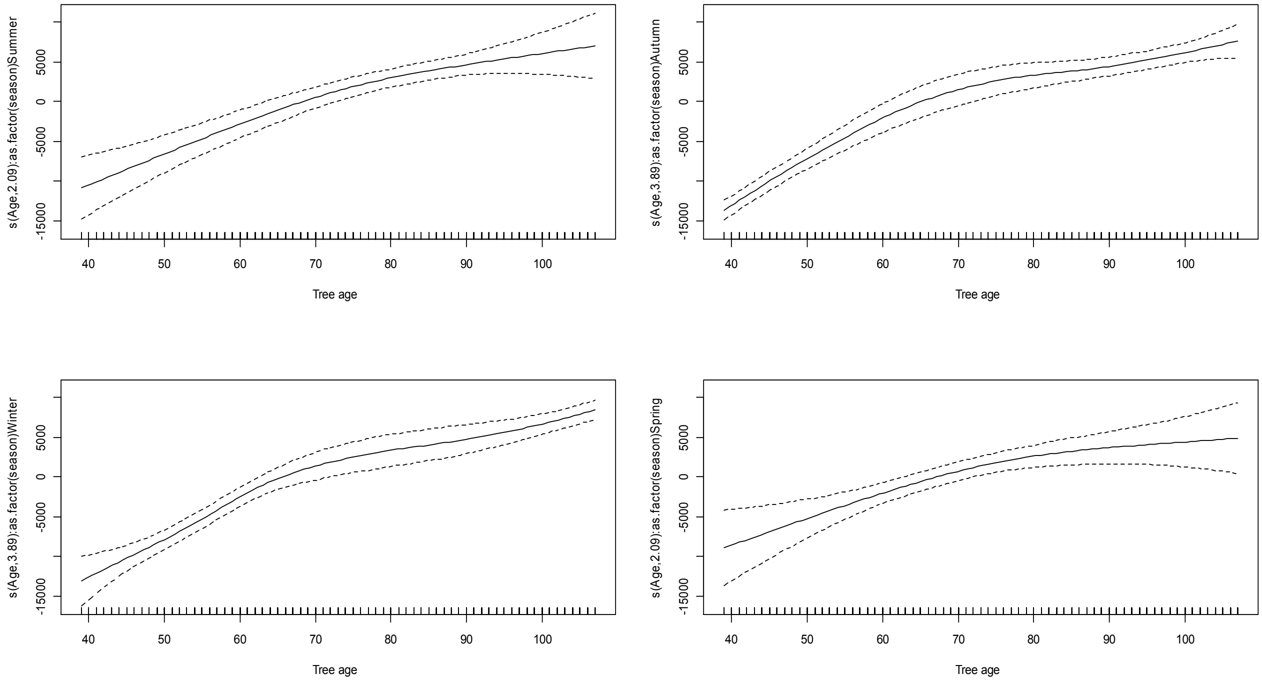

Figure 7. 10 Estimated smoothing curves and 95% confidence bands for the AMM model that uses tree age by season for the GC clone.

7.10 Modelling The Effect of Tree Age by Including the Interaction of Season and Clone in The Parametric Part of The Model

An attempt to fit the model with four smoothers of age (one for each season) was made by including the interaction between clone and season on the parametric part of the additive mixed model. The results of the model fit show that all parametric coefficients and the smooth terms are significant. For summer and spring the smoothers have an effective degree of freedom equal to one, essentially fitting a straight line (Table 7.5). This shows the relationship between stem radius and tree age is linear in summer and spring by taking into account the parametric effect of clone and season. Figure 7.10 also confirms that the type of relationship between stem radius and tree age depends on season.

The upper left and the lower right panels of Figure 7.10 show the relationship between tree age and stem radius in summer and spring respectively**.** From the plot it appears the relationship is linear. The upper right and the lower left panels of Figure 7.10 show the relationship between tree age and stem radius in autumn and winter respectively. It appears that the relationship is clearly nonlinear for autumn and winter. A similar model, but without the interaction effect of clone and season in the parametric part is fitted for comparison with the current model under consideration. The likelihood ratio test statistic shows the model with interaction is better (the value of test statistic is 43.91 with 3 degrees of freedom and p-value =<0.0001). Therefore, we cannot further simplify the model whose output is presented in Table 7.5.

Table 7. 5 The fitted additive mixed model with four different smoothers of tree age (one per season) with the interaction between season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	20338.8	868.0	23.43	< 2e-16 ***
Clone(GC)	-3796.9	1194.7	-3.18	0.001519 **
Season(Autumn)	-3291.7	407.9	-8.07	1.66e-15 ***
Season(Winter)	-2722.6	468.3	-5.81	7.81e-09 ***
Season(Spring)	-652.4	299.9	-2.18	0.029816 *
Clone(GC) × Season(Autumn)	1478.2	233.2	6.34	3.25e-10 ***
Clone(GC) × Season(Winter)	1327.5	239.4	5.54	3.61e-08 ***
Clone(GC) × Season(Spring)	1025.0	263.6	3.89	0.000106 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1	1	70	<2e-16 ***
s(Age, season = Autumn)	3.321	3.321	4823.6	<2e-16 ***
s(Age, season = Winter)	3.307	3.307	2559.2	<2e-16 ***
s(Age, season = Spring)	1	1	175.4	<2e-16 ***
R-sq.(adj) = 0. 85				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

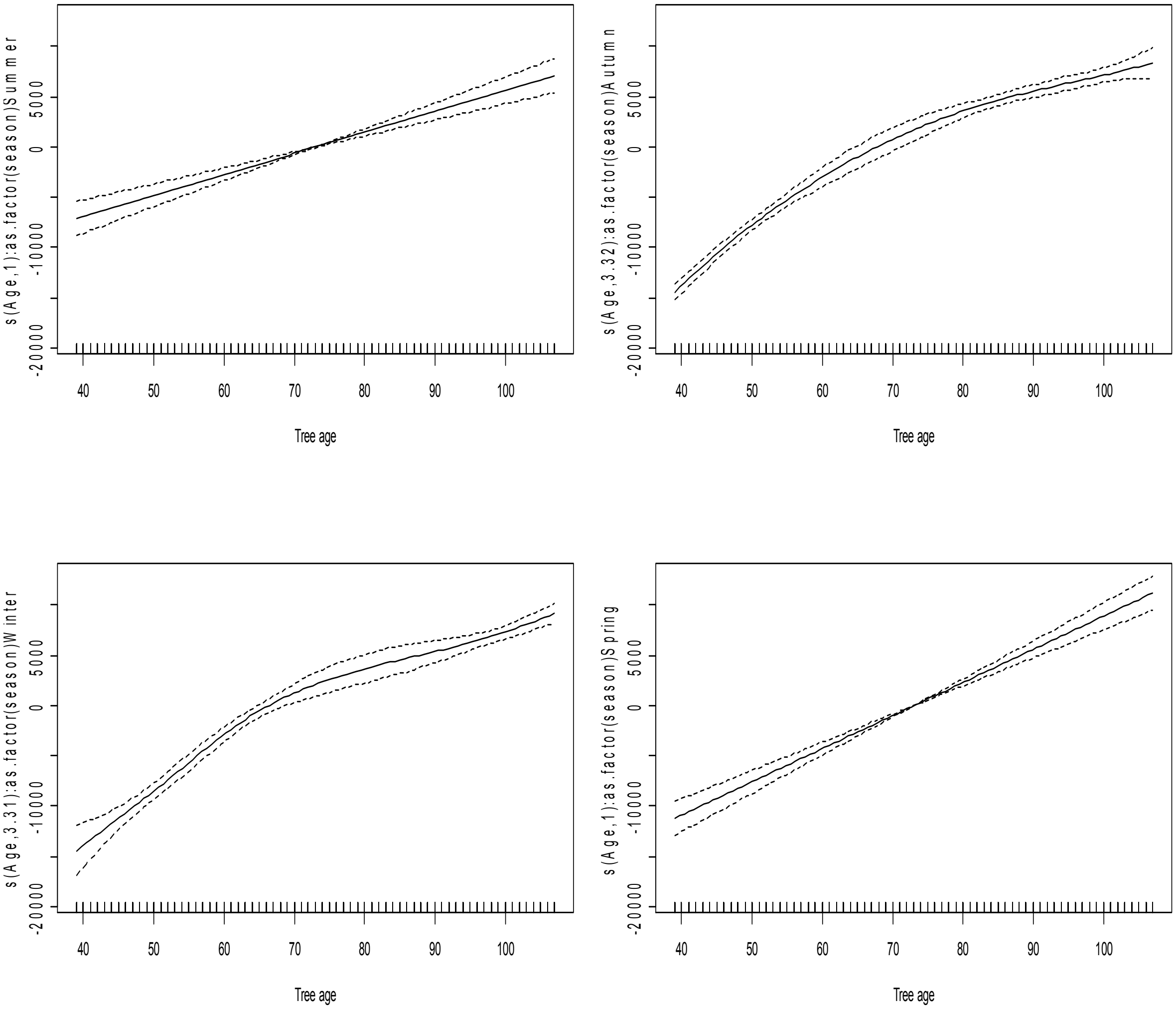

Figure 7.11 Estimated smoothing curves and 95% confidence bands for the GAMM model that uses four smoothers of tree age and includes the interaction of season by clone in the parametric part.

7.11 Model for The Effect of Temperature by Including Clone and Season in Parametric Part of The Model

Additive mixed model (AMM) with four smoothers of temperature (one for each season) is fitted by including the interaction between clone and season in the parametric part. However, the interaction between season and clone was not significant. Moreover, the likelihood ratio test comparing the model with interaction and the model without interaction term in the parametric part of the model is 3.09 with p-value =0.38. Therefore, a model without the interaction effect of clone and season on the parametric part of the additive mixed model is fitted. The results for the effect of temperature show that there is a nonlinear relationship between stem radius and temperature in autumn and winter. The smoothers for the effect of temperature appear linear in summer and spring. Moreover, the effect of temperature is not significant in either summer (p-value=0.904) or spring (p-value =0.30633). The adjusted R-square ($R^2 = 0.358$) also shows that the effect of temperature on stem radius is not as strong as the effect of tree age (Tables 7.5 & 7.6).

Figure 7.11 shows the estimated smoothers for temperature by season. The temperature smoothers in summer and spring form a horizontal band around zero. This indicates that the effect of temperature on stem radius is not significant for the two seasons. On the other hand the temperature smoothers for autumn and winter show that there is a nonlinear relationship between stem radius and temperature. Both the parametric coefficients and non-parametric approximate smooth terms are significant in autumn (Table 7.6).

Table 7. 6 Parameter estimates of the additive mixed model with four different smoothers of temperature (one per season) with season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	21694	2007	10.807	< 2e-16 ***
Clone(GC)	-2726	1183	-2.305	0.0213 *
Season(Autumn)	-5498	2698	-2.037	0.0418 *
Season(Winter)	-8074	9242	-0.874	0.38250
Season(Spring)	-3963	1878	-2.11	0.0351 *
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(temperature, season = Summer)	1	1	0.014	0.9040
s(temperature, season = Autumn)	8.222	8.222	59.39	< 2e-16 ***
s(temperature, season = Winter)	5.02	5.02	16.68	4.91e-16 ***
s(temperature, season = Spring)	1.000	1.000	1.66	0.198
R-sq.(adj) = 0.358				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

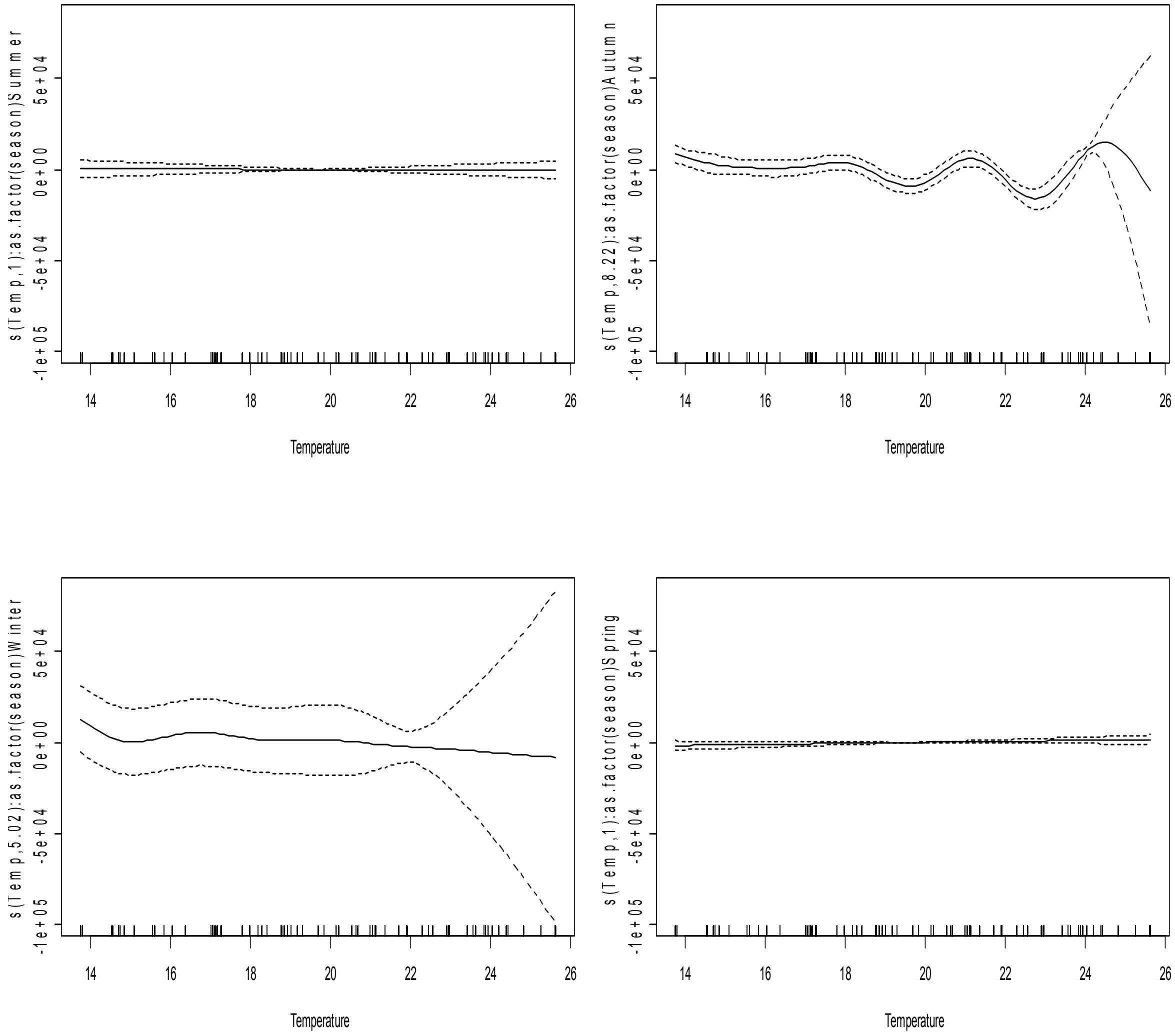

Figure 7. 12 Estimated smoothing curves and 95% confidence bands for the GAMM model that uses four smoothers of temperature with season and clone in the parametric part.

7.12 Model for The Effect of Rainfall by Including Clone and Season in Parametric Part of The Model

Additive mixed model (AMM) with four smoothers of rainfall (one for each season) is fitted by including the interaction between clone and season in the parametric part. However, the interaction between season and clone was not significant. Moreover, the likelihood ratio test comparing the model with interaction and the model without an interaction term in the parametric part of the model is 2.74 with p-value =0.43. Therefore, a model without the interaction effect of clone and season on the parametric part of the additive mixed model is selected as a better model. The estimates of the parametric coefficients show there is significant difference between the two clones. The effective degree of freedom of the smooth terms for both summer (p-value = 0.575) and spring (p-value =0.895) are not significant. The smooth terms for autumn (edf=4.638, p-value = <2e-16) and winter (edf= 7.37, p-value =<2e-16) are significant. The parametric part of the model shows that the coefficients for winter and spring are significant. This shows that the relationship between rainfall and stem radius is nonlinear in autumn and winter.

Table 7. 7 Parameter estimates of the additive mixed model with four different smoothers of rainfall (one per season) with the effects of season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	21462.1	902.7	23.8	< 2e-16 ***
Clone(GC)	-2726.3	1182.7	-2.3	0.02130 *
Season(Autumn)	-6436.6	5814.2	-1.1	0.268500
Season(Winter)	-3737.6	508.1	-7.4	3.45e-13 ***
Season(Spring)	-3454.9	535.2	-6.5	1.56e-10 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(rainfall, season = Summer)	1	1	0.30	0.575000
s(rainfall, season = Autumn)	4.64	4.64	39.40	< 2e-16 ***
s(rainfall, season = Winter)	7.37	7.37	24.10	<2e-16 ***
s(rainfall, season = Spring)	1.00	1.00	0.02	0.895000
R-sq.(adj) = 0.291				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.13 Model for The Smoothed Effect of Relative Humidity by Including The Effect of Clone and Season in Parametric Part of The Model

The AMM with four smoothers (one per season) for relative humidity including the interaction between season and clone in the parametric part is fitted. The likelihood ratio statistics that compare this model with a model without interaction, favours the model without interaction (p-value = 0.132). Relative humidity smoothers for autumn (edf=8.458) and winter (edf= 8.586) are also significant (Table 7.8). Relative humidity smoothers for summer (edf=1) and spring (edf=1) are not significant. The adjusted R^2 for this model is 0.593 which is larger than the adjusted R^2 when either temperature or rainfall is used in the model (see Tables 7.6, 7.7 and 7. 8).

Table 7. 8 Parameter estimates of the additive mixed model with four different smoothers of relative humidity (one per season) with the effects of season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	21510.0	869.3	24.74	< 2e-16 ***
Clone(GC)	-2726.3	1182.4	-2.31	0.0213 *
Season(Autumn)	-5744.6	506.4	-11.34	< 2e-16 ***
Season(Winter)	-4197.1	877.2	-4.78	1.9e-06 ***
Season(Spring)	-3494.8	412.5	-8.47	< 2e-16 ***
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(relative humidity, season = Summer)	1.000	1.000	1.98	0.159000
s(relative humidity, season = Autumn)	8.458	8.458	170.25	<2e-16 ***
s(relative humidity, season = Winter)	8.586	8.586	55.80	<2e-16 ***
s(relative humidity, season = Spring)	1	1	0.000	0.995000
R-sq.(adj) = 0.593				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.14 Model for The Effect of Smoothed Solar Radiation by Including The Effects of Clone and Season in Parametric Part of The Model

The additive mixed model with four smoothers for solar radiation (one per season) that includes the effect of clone and season on the parametric part is fitted.

The likelihood ratio statistics that compare the model without interaction with the model that includes the interaction of season and clone in parametric part, favours the model without interaction (p-value= 0.4929). The smoothed solar radiation for summer (edf=1) and spring (edf=1) appear to be linear and it is not significant (p-values =0.6 and 0.131 respectively for summer and spring).

The smoothed solar radiation for autumn (edf=7.81) shows that the relationship between stem radius and solar radiation is nonlinear. All the parametric coefficients of this model are significant (Table 7.9.)

Table 7. 9 Parameter estimates of the additive model with four different smoothers of solar radiation (one per season) with season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	21992	1339	16.43	< 2e-16 ***
Clone(GC)	-2726	1185	-2.30	0.02156700 *
Season(Autumn)	-114244	33710	-3.39	0.000724 ***
Season(Winter)	-5984	1688	-3.54	0.000409 ***
Season(Spring)	-4497	1183	-3.80	0.000151 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(solar radiation, season = Summer)	1.00	1.00	0.28	0.600000
s(solar radiation, season = Autumn)	7.81	7.81	18.77	< 2e-16 ***
s(solar radiation, season = Winter)	2.14	2.14	1.95	0.138000
s(solar radiation, season = Spring)	1.00	1.00	2.29	0.131000
R-sq.(adj) = 0.291				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

7.15 Model for The Smoothed Effect of Wind Speed by Including Clone and Season in Parametric Part of The Model

The additive mixed model with four smoothers for wind speed (one per season) that includes the effects of clone and season on the parametric part is fitted. A likelihood ratio test is used to test for the presence of interaction between the clone and season in the parametric part of the model. The test favours the model without interaction (p-value=0 .3434). The smoothed wind speed for summer (edf=1) and spring (edf=1) appears to be linear and it is not significant (p-values = 0.0826 and 0.3162 respectively for summer and spring). The smoothed wind speed for autumn (edf= 2.498) shows that the relationship between stem radius and wind speed is nonlinear. The smoothed wind speed for the winter season (edf= 7.637) shows that the relationship between stem radius and wind speed is nonlinear (p-value= <2e-16). The parametric coefficients for spring is highly significant with a small p-value = 5.1e-09 (Table 7. 10).

Table 7.10 Parameter estimates of the AMM with four different smoothers of wind speed (one per season) with season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	22121.9	965.6	22.91	< 2e-16 ***
Clone(GC)	-2726.3	1183.6	-2.30	0.02140 *
Season(Autumn)	-1782.9	946.9	-1.88	0.060000
Season(Winter)	-1831.8	2395.2	-0.77	0.444600
Season(Spring)	-4598.9	781.2	-5.89	5.1e-09 ***
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(solar radiation, season = Summer)	1.00	1.00	3.02	0.082600
s(solar radiation, season = Autumn)	2.50	2.50	104.36	< 2e-16 ***
s(solar radiation, season = Winter)	7.64	7.64	49.37	<2e-16 ***
s(solar radiation, season = Spring)	1.00	1.00	1.01	0.316200
R-sq.(adj) = 0.291				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.16 AMM Fitted with The Smoothers Tree Age and Temperature

In the previous sections we applied a series of additive mixed models on stem radius for various covariates separately. Tests for the presence of interaction between clone and season were made for different models considered. It was observed that the smoothers for tree age and each climatic variable also depend on season. The analysis made so far may help to see the effect of individual covariates on stem radius. In order to see the effect of more than one covariate on stem radius, it is essential to fit models that involve the smoothers for two or more covariates. This demands the application of model selection procedures. It is known that model selection with mixed models is complicated by the presence of fixed effects and random effects. The fixed effect structure and the random effect structure are dependent on each other and the selection of one affects the other. There are two strategies that are commonly used in a model selection process. These are the top-down strategy (Diggle et al., 2002) and the step-up strategy (West et al., 2006). In the step-up strategy one starts with a limited model (e.g., few fixed and random effects) and then additional fixed effects and random effects are added based on statistical tests. In the top-down procedure, the initial model has one random intercept but with a model where the fixed component contains all explanatory variables and as many interactions as possible. This is called the beyond optimal model. Using the beyond optimal model, one can find the optimal component of the random component (Zuur et al., 2009). The beyond optimal model is sometimes unrealistic due to a large number of explanatory variables, interactions or numerical problems. In this thesis we followed the step-up approach.

Both tree age and temperature were smoothed to see their effect on stem radius. The smoothed temperature for all four seasons is not significant (Table 7.11). It appears in the presence of the smoothed tree age effect in the model, the smoothed effect of temperature is not significant. An attempt to include temperature with one smoother for all four seasons in the model also shows that the smoothed temperature is not significant (edf=1, p-value=0.76). Instead of using temperature as a smoothed component of the AMM, an attempt to use temperature in the parametric part of the AMM was made. A likelihood ratio test was applied by including the interaction of temperature with season and the interaction of temperature with clone in the parametric part of the additive mixed model. In both cases the interaction effect of temperature is not significant (p-value = 0.8 and 0.9 for the interaction with clone and season respectively). A likelihood ratio test was applied by including temperature in the parametric part of the additive mixed model, the result shows that temperature is not important in explaining stem radial growth in the presence of the smoother for tree age in the model (p-value = 0.75).

Table 7. 11 Parameter estimates of the additive mixed model with four different smoothers of age and temperature (one per season in each case) with the interaction between season and clone included in parameteric part (Maximum liklihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	20365.0	993.3	20.50	0.0001 ***
Clone(GC)	-3796.9	1194.7	-3.18	0.002 **
Season(Autumn)	-3309.2	635.1	-5.21	0.0001 ***
Season(Winter)	-2506.5	692.6	-3.62	0.0004 ***
Season(Spring)	-627.9	580.5	-1.08	0.279611
Clone(GC) × Season(Autumn)	1478.2	233.2	6.34	0.0001 ***
Clone(GC) × Season(Winter)	1327.5	239.5	5.54	0.0001 ***
Clone(GC) × Season(Spring)	1025.0	263.6	3.89	0.0001 ***
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1.000	1.000	70.38	0.0001 ***
s(Age, season = Autumn)	3.064	3.064	5081.76	0.0001 ***
s(Age, season = Winter)	2.999	2.999	2778.96	0.0001 ***
s(Age, season = Spring)	1.000	1.000	151.69	0.0001 ***
s(Temperature, season = Summer)	1.000	1.000	0.003	0.957
s(Temperature, season = Autumn)	1.000	1.000	0.086	0.770
s(Temperature, season = Winter)	1.000	1.000	1.012	0.315
s(Temperature, season = Spring)	1.000	1.000	0.186	0.666
R-sq.(adj) = 0.848				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

7.17 AMM Fitted with The Smoothers of Tree Age and Rainfall

The additive mixed model (AMM) fitted using smoothed tree age and rainfall shows that the smoothed rainfall is not significant (Table 7.12) for all seasons. The AMM that uses four smoothers of tree age and one smoother for rainfall also shows that the smoothed rainfall is not significant (p-value= 0.508). The likelihood ratio tests used to compare a model without any effect of rainfall with the models that have the interaction of rainfall with clone or season show that the interaction of rainfall is not significant (p-value = 0.8036 and 0.7495 for the interaction with clone and season respectively). A likelihood ratio test was also applied by including rainfall without any interaction in the parametric part of the additive mixed model. The result shows that there is not enough evidence from this data to show the importance of rainfall (p-value = 0.9492) in the presence of tree age in the model.

Table 7. 12 Parameter estimates of the additive mixed model with four different smoothers of age and rainfall (one per season in each case) with the interaction between season and clone incuded in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	20319.1	869.8	23.36	< 2e-16 ***
Clone(GC)	-3796.9	1194.7	-3.18	0.001520 **
Season(Autumn)	-3265.1	413.6	-7.89	6.43e-15 ***
Season(Winter)	-2691.8	459.8	-5.86	6.14e-09 ***
Season(Spring)	-636.5	305.7	2.08	0.037528 *
Clone(GC) × Season(Autumn)	1478.2	233.1	6.34	3.22e-10 ***
Clone(GC) × Season(Winter)	1327.5	239.4	5.55	3.59e-08 ***
Clone(GC) × Season(Spring)	1025.0	263.5	3.89	0.000106 ***
Approximate significance of smooth terms				
	Edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1.000	1.000	65.63	1.24e-15 ***
s(Age, season = Autumn)	3.26	3.26	4899.16	< 2e-16 ***
s(Age, season = Winter)	3.23	3.23	2574.142	< 2e-16 ***
s(Age, season = Spring)	1.000	1.000	175.30	< 2e-16 ***
s(rainfall, season = Summer)	1.000	1.000	0.119	0.7300000
s(rainfall, season = Autumn)	1.000	1.000	0.049	0.8250000
s(rainfall, season = Winter)	1.000	1.000	0.797	0.3720000
s(rainfall, season = Spring)	1.000	1.000	0.061	0.8040000
R-sq.(adj) = 0.848				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.18 AMM Fitted with The smoothers of Tree Age and Relative Humidity

The additive mixed model fitted using smoothed tree age and relative humidity by season shows that all the smoothed terms are linear (with edf=1). The smoothed relative humidity in winter is significant (p-value =0.047). Smoothers of relative humidity for the rest of the season are not significant (Table 7.13). An attempt to include the relative humidity in the parametric part of the additive mixed model was made. The model that has the effect of relative humidity in the parametric part is compared with the model without any effect of relative humidity using the likelihood ratio test. The result favours the model without any effect of relative humidity (p-value = 0.6527). An additive mixed model that includes the interaction of relative humidity with season and a model without any effect of relative humidity is compared using the likelihood ratio test. The likelihood ratio test favours the model without the effect of relative humidity (p-value= 0.9).

Table 7. 13 Parameter estimates of the additive mixed model with four different smoothers of age and relative humidity (one per season in each case) with the interaction between season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t- value	p-value
Intercept	20400.4	880.2	23.18	0.0001 ***
Clone(GC)	-3796.9	1196.8	-3.17	0.002 **
Season(Autumn)	-4499.8	303.1	-14.84	0.0001 ***
Season(Winter)	-3286.3	304.9	-10.78	0.0001 ***
Season(Spring)	-745.2	350.1	-2.13	0.033486 *
Clone(GC) × Season(Autumn)	1478.2	254.8	5.80	0.0001 ***
Clone(GC) × Season(Winter)	1327.5	261.7	5.07	0.0001 ***
Clone(GC) × Season(Spring)	1025.0	288.1	3.56	0.0004 ***
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1	1	46.23	0.0001 ***
s(Age, season = Autumn)	1	1	4572.14	0.0001 ***
s(Age, season = Winter)	1	1	5664.52	0.0001 ***
s(Age, season = Spring)	1	1	147.14	0.0001 ***
s(relative humidity, season = Summer)	1	1	0.339	0.5605
s(relative humidity, season = Autumn)	1	1	2.092	0.1483
s(relative humidity, season = Winter)	1	1	3.958	.0469 *
s(relative humidity, season = Spring)	1	1	0.248	0.6186
R-sq.(adj) = 0.841				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.19 Additive Mixed Models (AMM) Fitted with The Smoothers of Tree Age and Solar Radiation

The additive mixed model fitted using smoothed tree age and solar radiation also shows that all the smoothed terms are linear (edf=1). The smoothed solar radiation for winter is significant (p-value =5.98e-12). The smoother for autumn is also significant (p-value =0.00059). Solar radiation smoothers for the rest of the seasons (summer, p value=0.982, and spring p-value=0.608) are not significant (Table 7.14). An attempt to include solar radiation in the parametric part was made. A model without any effect of solar radiation was compared with the model that includes solar radiation in the parametric part. The likelihood ratio statistic favours the model without solar radiation (p-value= 0.3417). A model that includes the interaction of solar radiation with season and a model that includes the interaction between solar radiation and clone in the parametric part are each compared with a model without any effect of solar radiation. In both cases the likelihood ratio test favours a model without the effect of solar radiation (p-value for the interaction with clone = 0.1016 and p-value for interaction with season= 0.678).

Table 7.14 Parameter estimates of the additive mixed model with four different smoothers of age and solar radiation (one per season in each case) with the interaction between season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	20331.8	926.7	21.94	0.0001 ***
Clone(GC)	-3796.9	1196.2	-3.17	0.002 **
Season(Autumn)	-4301.3	420.5	-10.23	0.0001 ***
Season(Winter)	-2205.5	445.1	-4.95	0.0001 ***
Season(Spring)	-729.3	477.0	-1.53	0.1267
Clone(GC) × Season(Autumn)	1478.2	249.4	5.93	0.0001 ***
Clone(GC) × Season(Winter)	1327.5	256.1	5.18	0.0001 ***
Clone(GC) × Season(Spring)	1025.0	281.9	3.64	0.0001 ***
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1	1	57.24	7.30e-14 ***
s(Age, season = Autumn)	1	1	13812.833	< 2e-16 ***
s(Age, season = Winter)	1	1	7337.006	< 2e-16 ***
s(Age, season = Spring)	1	1	116.502	< 2e-16 ***
s(solar radiation, season = Summer)	1	1	0.001	0.982033000
s(solar radiation, season = Autumn)	1	1	11.870	0.000589 ***
s(solar radiation, season = Winter)	1	1	48.241	5.98e-12 ***
s(solar radiation, season = Spring)	1	1	0.263	0.608376000
R-sq.(adj) = 0.843				
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.20 Additive Mixed Models (AMM) Fitted with The Smoothers of Tree Age and Wind Speed

The additive mixed model fitted using smoothed tree age and wind speed also shows that all the smoothed terms related to wind speed are linear (edf=1) (Table 7.15). Wind speed smoothers for all seasons (summer, p value=0.185, autumn p-value=0.539, winter p-value=0.766 and spring –p value=0.643) are not significant.). An attempt to include wind speed in the parametric part of the additive mixed model was made. A model without any effect of wind speed was compared with the model that includes wind speed in the parametric part. The likelihood ratio statistic favours the model without wind speed (p-value= 0.6967).

A model that includes the interaction of wind speed with season in the parametric part of the additive mixed model was compared with a model without any effect of wind speed. The likelihood ratio statistic favours a model without wind speed (p-value= 0.6558).

A model that includes the interaction between wind speed and clone in the parametric part of the additive mixed model was compared to a model without any effect of wind speed. The likelihood ratio statistic favours the model with the interaction of wind speed and clone (p-value= 9e-04).

Table 7.15 Parameter estimates of the additive mixed model with four different smoothers of age and wind speed (one per season in each case) with the interaction between season and clone included in parameteric part (Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	20730.2	916.6	22.62	< 2e-16 ***
Clone(GC)	-3796.9	1194.7	-3.18	0.001519 **
Season(Autumn)	-3656.3	502.2	-7.28	5.91e-13 ***
Season(Winter)	-3091.6	549.8	-5.62	2.33e-08 ***
Season(Spring)	-1122.0	453.0	-2.48	0.013402 *
Clone(GC) × Season(Autumn)	1478.2	233.0	6.34	3.16e-10 ***
Clone(GC) × Season(Winter)	1327.5	239.2	5.55	3.53e-08 ***
Clone(GC) × Season(Spring)	1025.0	263.4	3.89	0.0001 ***
Approximate significance of smooth terms				
	edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1	1	31.934	1.97e-08 ***
s(Age, season = Autumn)	3.290	3.290	3710.531	< 2e-16 ***
s(Age, season = Winter)	3.136	3.136	2745.244	< 2e-16 ***
s(Age, season = Spring)	1	1	162.415	< 2e-16 ***
s(wind speed, season = Summer)	1	1	1.763	0.185 000
s(wind speed, season = Autumn)	1	1	0.377	0.539 000
s(wind speed, season = Winter)	1	1	0.089	0.766000
s(wind speed, season = Spring)	1	1	0.215	0.643000
R-sq.(adj) =	0.848			
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

7.21 Additive Mixed Models with Additive Effect of More Than Two Covariate Smoothers

The AMM that uses the smoothers age and any one of the climatic variables resulted in significant smoothers for age in all cases. The smoothers for temperature, rainfall and wind speed did not appear to be significant. However, the smoothers for relative humidity for winter (p-value=0.047) and the smoothers for solar radiation for winter (p-value < 0.00001) and autumn (p-value= 0.0006) are significant. An additive mixed model that includes the smoothers of tree age, wind speed and solar radiation was fitted. In this additive model all smoothers appear to have the estimated effective degrees of freedom equal to 1 (Table 7.16). The smoothers of tree age for all seasons (summer, autumn, winter and spring are significant with very small respective p-values (p-value < 0.00001). The smoothers for solar radiation are significant in autumn and winter with respective p-values (0.00171 and 1.95e-14). In all of the above models, random intercept for each tree is used in combination with the assumption that residuals are normally distributed with mean 0 and constant variance. In an attempt to validate the last model (that includes smoothers of tree age, solar radiation and relative humidity) the model validation graphs are plotted (Figure 7.12). The lower right panel of the graphs show a strong relationship between fitted and observed values of stem radius. The upper right panel shows that the assumption of constant variance is violated. The upper left and lower left panels show that there is some deviation from normality. The plots of normalized residuals against covariates (tree age, solar radiation and relative humidity, clone and season) are plotted as part of the model validation process. There is no clear pattern as to the dependence of residuals on any of the covariates of tree age, solar radiation and relative humidity (Figure 7.13). However, the spread of residuals depend on tree age. The spread of residuals also clearly depends on clone and season (Figure 7.14.). This indicates that the variation in the data differ between seasons and clone. It was also observed that there is more variation in autumn and winter than in summer and spring which violates the homogeneity assumption. Therefore, the assumption that the residuals are normally distributed with mean zero and constant variance is relaxed. Moreover, an attempt to use random slope instead of random intercept was made.

Table 7.16 Parameter estimates of the additive mixed model with four smoothers of age and relative humidity and solar radiation (one per season in each case) with the interaction between season and clone included in parametric part(Maximum likelihood estimates).

Parametric coefficients	Estimate	Standard error	t-value	p-value
Intercept	20331.0	925.6	21.966	< 2e-16 ***
Clone(GC)	-3796.9	1196.0	-3.175	0.001538 **
Season(Autumn)	-4301.5	417.8	-10.295	< 2e-16 ***
Season(Winter)	-1700.3	460.7	-3.691	0.00023 ***
Season(Spring)	-709.4	487.8	-1.454	0.14614800
Clone(GC) × Season(Autumn)	1478.2	247.8	5.965	3.2e-09 ***
Clone(GC) × Season(Winter)	1327.5	254.4	5.217	2.1e-07 ***
Clone(GC) × Season(Spring)	1025.0	280.1	3.659	0.000264 ***

… Table 7.16 Approximate significance of smooth terms

	edf	Ref. df	F-value	p-value
s(Age, season = Summer)	1	1	48.49	5.3e-12 ***
s(Age, season = Autumn)	1	1	3614.49	< 2e-16 ***
s(Age, season = Winter)	1	1	5019.97	< 2e-16 ***
s(Age, season = Spring)	1	1	83.14	< 2e-16 ***
s(relative humidity, season = Summer)	1	1	0.41	0.52007000
s(relative humidity, season = Autumn)	1	1	0.07	0.79623000
s(relative humidity, season = Winter)	1	1	15.28	9.8e-05 ***
s(relative humidity, season = Spring)	1	1	0.03	0.86855000
s(solar radiation, season = Summer)	1	1	0.06	0.81294000
s(solar radiation, season = Autumn)	1	1	9.88	0.00171 **
s(solar radiation, season = Winter)	1	1	59.96	0.00171 **
s(solar radiation, season = Spring)	1	1	0.03	0.85950000
R-sq.(adj) = 0.843				

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

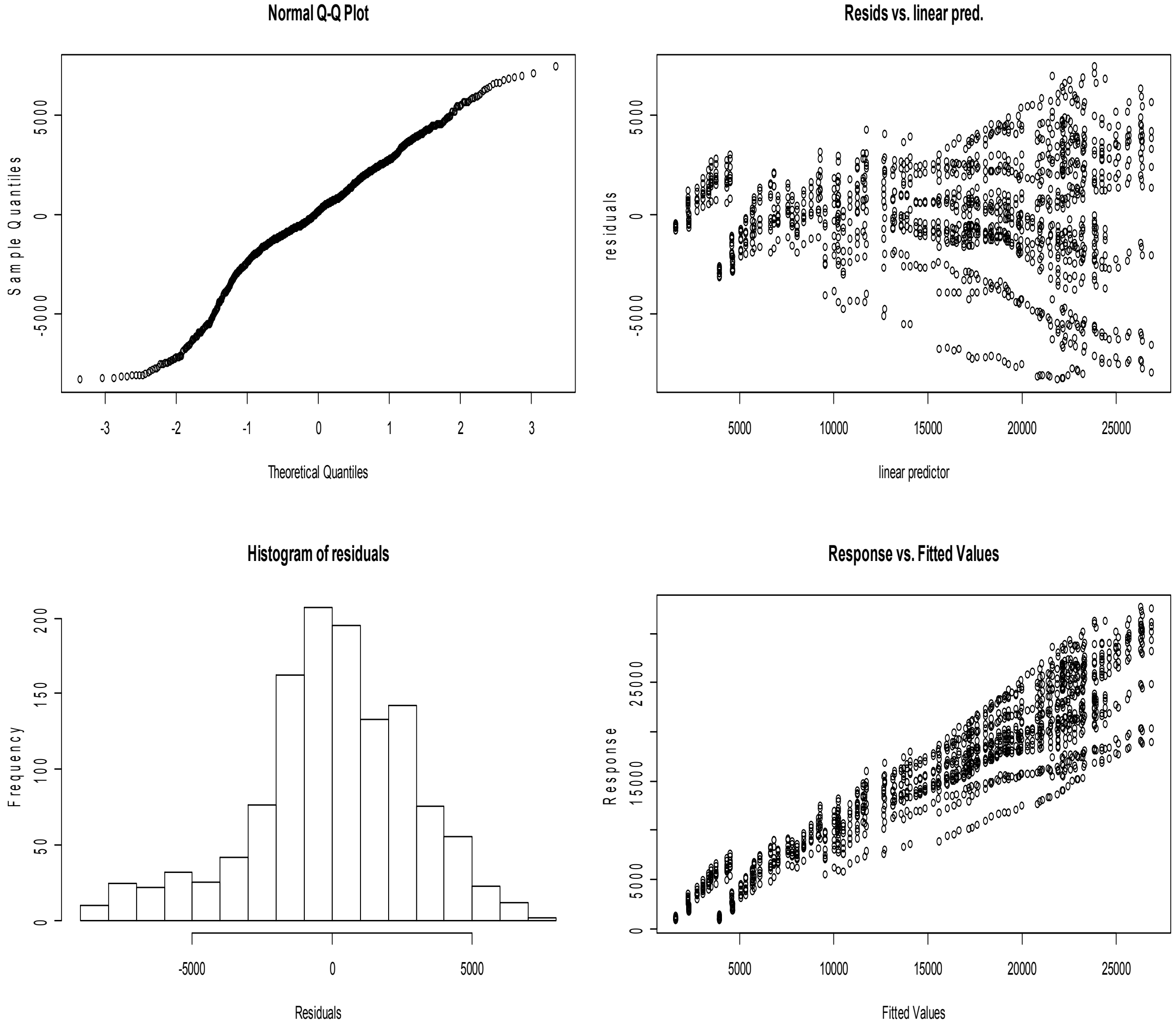

Figure 7.13 Basic model checking plots for the additive model with the smoothers of tree age solar radiation and relative humidity.

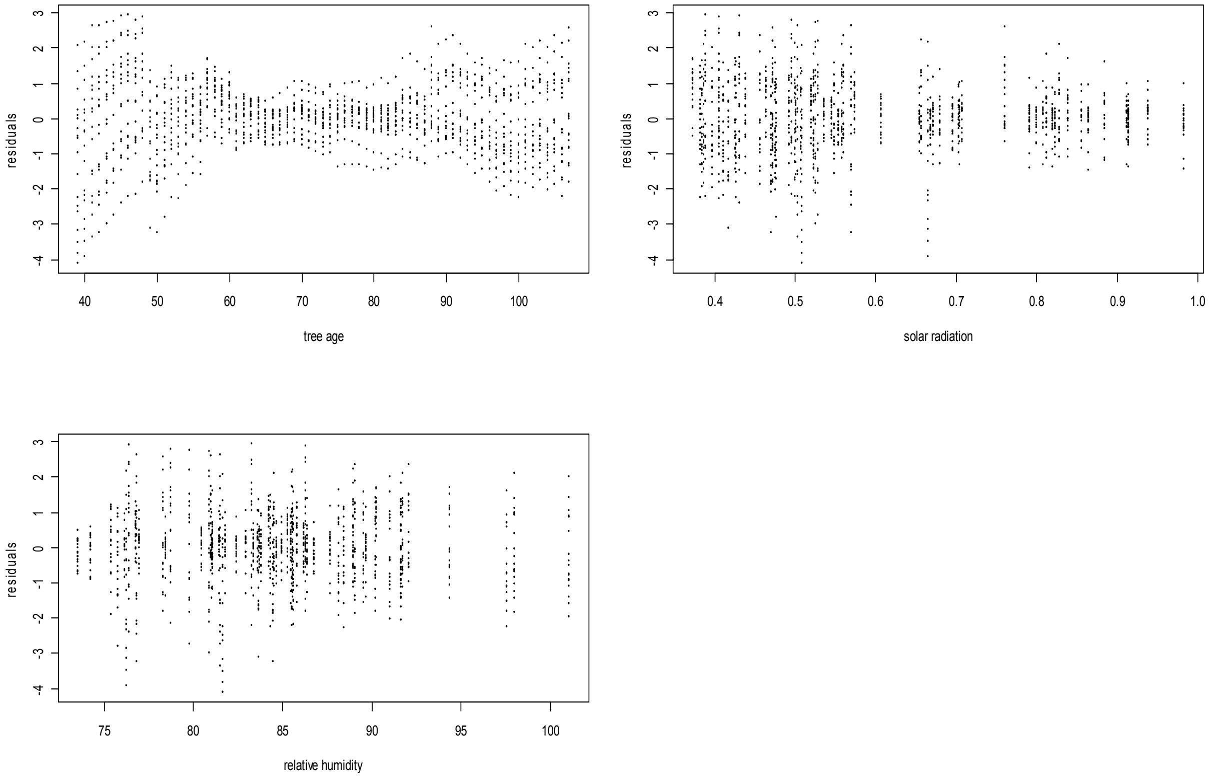

Figure 7.14 Plot of residuals versus tree, tree age, solar radiation and relative humidity.

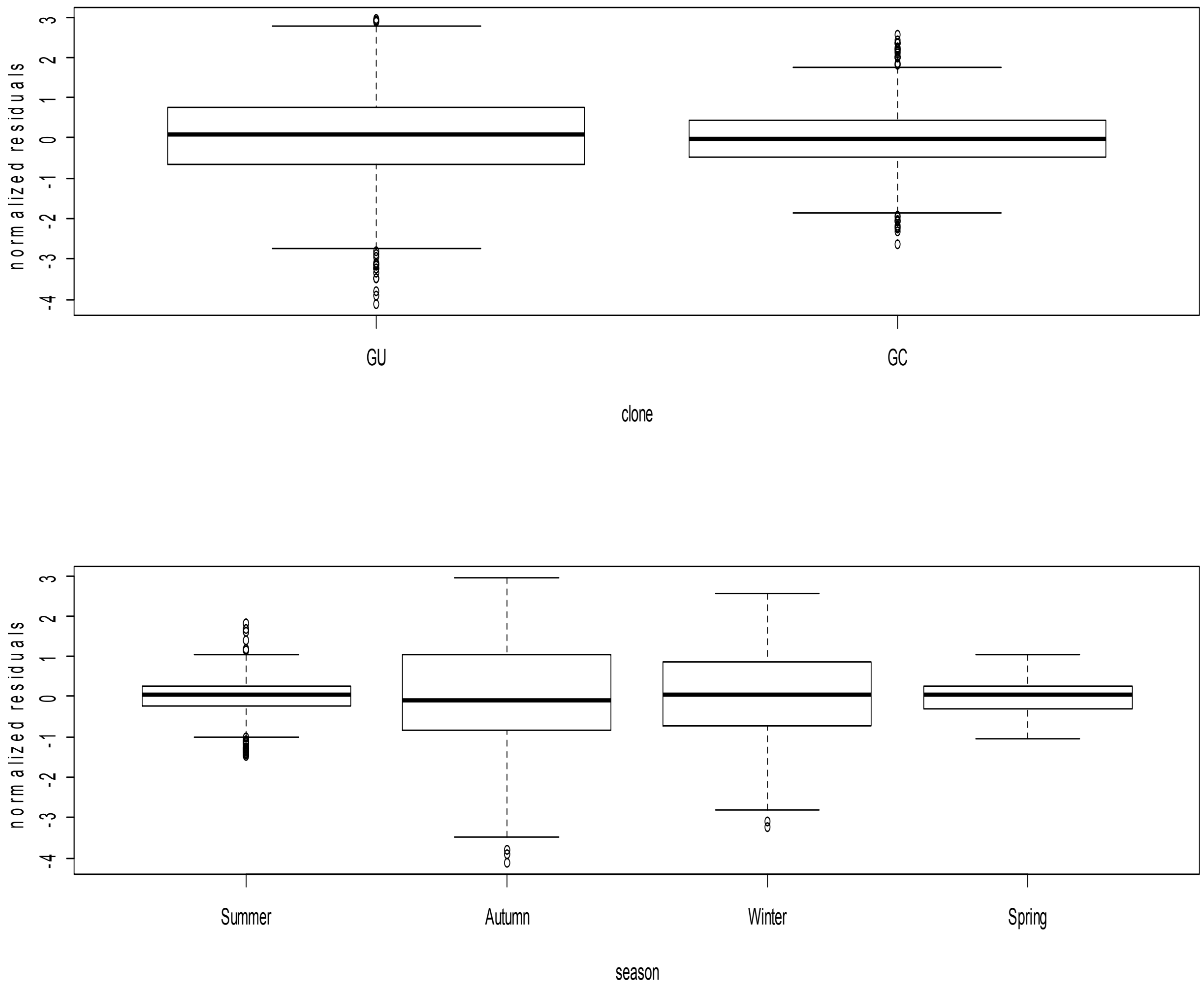

Figure 7.15 Normalized residuals plotted versus clone and season for the model that considers the smoothers of tree age, solar radiation and relative humidity.

All models are fitted using the REML estimation procedure and model comparison is made using the Akaike Information Criteria (AIC).

Table 7.17 The AIC for models with different variance and correlation structure

Variance structure	AIC	BIC
Model with random intercept and constant residual variance	21869.84	22044.07
Model with random intercept and residual variance that varies with clone and season combination	20836.39	] 21045.82
Model with random slope and constant residual variance	19289.85	19473.75
Model with random slope and residual variance that varies with clone and season combination	18899.35	19119

The model with random slope and different residual variance for each combination clone and season has the smallest AIC and BIC (Table 7.17). The validation graph did not show any variation between seasons or between clones (Figure 7.16). However, the plot of normalized residuals versus the fitted values showed that there is still a certain degree of heterogeneity in the residuals. The last aspect of the modelling process was to allow for spatial or temporal correlation in the residuals. However, the attempt was not successful due to the complexity of the model.

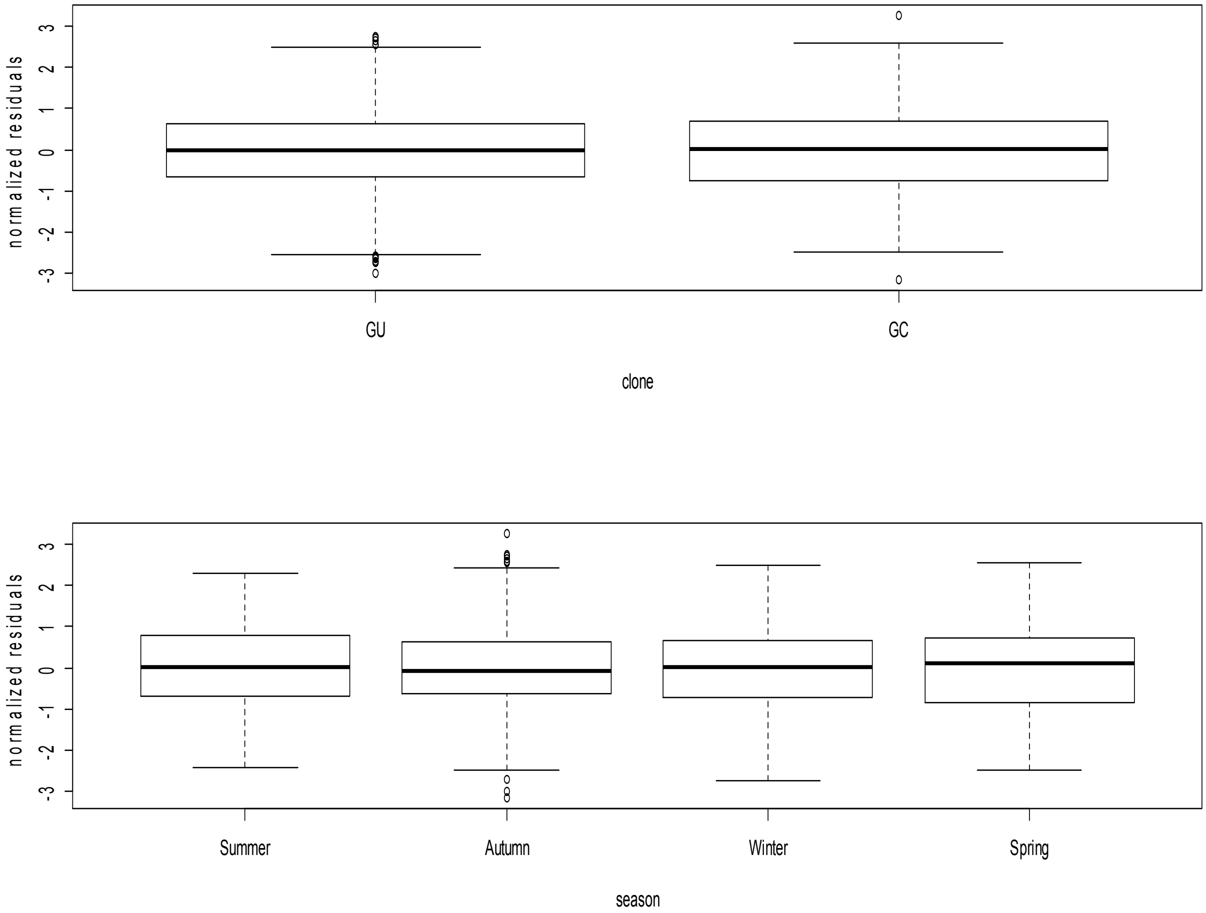

Figure 7. 16 Normalized residuals plotted versus clone and season for the model with random slope and residual variance that varies with clone and season combination.

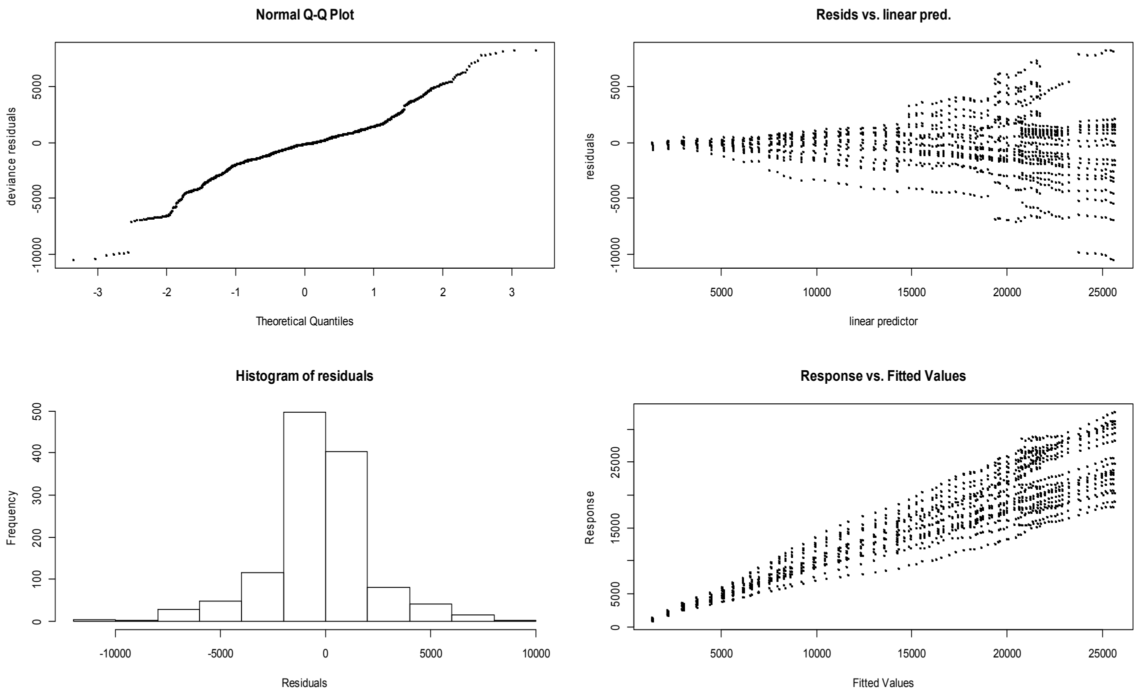

Figure 7.17 Basic model checking plots for the additive model with random slope and residual variance that varies with clone and season combination.

7.22 Summary

The variables included in the study are major climatic variables and a non-climatic variable tree age. Moreover, the clone of the tree and the season are additional factors included in the data. Classical techniques such as simple and multiple parametric regression analyses assume the linearity of the relationship between response and independent variables. Moreover, these classical techniques rely on the rigid assumptions of constant variance and the independent and identical distributions of the error terms. In many instances the assumption of linear relationship between the covariates and responses is very idealistic. The data under consideration is no exception. Consequently, a semi-parametric approach was employed to explore the type of relationship between stem radius and the various covariates. The relationship between stem radius and each of the predictors was fitted using the **mgcv** package. Both ML and REML estimation procedures are attempted and the penalized-spline is used as the basis function for smoothing.

Different additive mixed models (AMM) that range from one explanatory variable to multiple explanatory variables are fitted and several forms of relationships are investigated. For models with only one explanatory variable at a time, it was observed that all covariates have a nonlinear relationship with stem radius. The effect of each covariate on stem radius is found to depend on the clone of the tree. It was also observed that the two clones grow in similar fashion and in both cases the relationship between tree age and stem radius is nonlinear. The interaction between clone and season was found significant when the smoothers of tree age and temperature are used in the model. The effects of all covariates (temperature, relative humidity, rainfall, solar radiation and wind speed) depend on season. The results of the AMM with nonparametric smoothers of covariates validate the results obtained in previous chapters and generally gave more insight regarding the stem radial growth.

Chapter 8

Conclusion

This work has focused on statistical methods aimed at modelling the growth data. Explicitly, we have been concerned with statistical methods for continuous response data which are common in many research areas, in particular in agricultural and biological studies. Growth measurements occur when two or more observations of a response variable are achieved at different moments for each subject under study. The bulk of the work on methods for growth data has concentrated on data that can be modelled as a nonlinear function of time. The methodologies for nonlinear functions with covariates are less developed compared to the methods for linear expectation functions. Accordingly, even with current software developments, data analysts still face a huge challenge in fitting the most appropriate growth models with covariates.

In this thesis, we strived to give more insight into the different approaches to incorporate covariates and latent variables in the growth models. The proposed methodologies have been reviewed and their practicality is examined in-depth.

The study was motivated by a multitude of Sappi data to assess the climatic factors affecting the growth of juvenile eucalyptus trees. Data reduction and latent variable modelling was crucial for the growth modelling with covariates. Accordingly the latent variable modelling approaches, namely principal component regression and partial least squares regressions are used on daily stem radius data. The study on daily averages of stem radius show that tree age is the most important factor that influences stem radius during the juvenile stage (up to 2 years). The results also revealed that the climatic variable on stem radius depends on season. That means the effects vary from one season to another. The analysis by season shows that there is no relationship between weather variables (temperature, relative humidity, solar radiation, wind speed and rainfall) and stem radius for two seasons (summer and spring). In winter, there is a positive relationship between each of the variables (tree age, temperature, relative humidity, solar radiation and wind speed) and stem radius. In autumn, the relationship between stem radius and variables (solar radiation, wind speed and tree age) is positive for both clones. In autumn and winter, the effect of rainfall on stem radius is significant for the GU clone while it is not significant for the GC clone. This could be mainly due to genetic differences between the two clones. This may need further research in the area. The type of relationship between stem radius and climatic varaibles needs to be confirmed by further research using different set of data.

In an attempt to account for both direct and indirect effects of covariates on growth, a path modelling approach was used. The best fitting path model to the data was identified and this showed that all climatic variables and tree age had positive effects on stem radial growth for the pooled data of both clones. Furthermore, all except one variable (rainfall) had significant direct effects on radial growth. Although rainfall was not significant in the best fitting model, it was found to be significant for the model that excluded wind speed and for the model that omitted solar radiation. This shows that the effect of rainfall on radial growth cannot be ruled out. To compare the effect of the explanatory variables on the radial growth of the GU and GC clones, a single analysis that estimated parameters and tested hypotheses about both groups simultaneously was considered. The regression weights for the two clones were significantly different. The regression weights were all positive indicating the positive effect of the climatic variables and tree age.

In addition, the regression weights obtained for the GU clone were larger than the regression weights for the GC clone. It was confirmed that the GU clone grows at a faster rate than the GC clone. The main estimation method for path models, or any structural equation model (SEM) is maximum likelihood estimation. This method requires a distributional assumption, which the present data failed to satisfy. The bootstrap method was then applied to overcome the methodological failure due to non-normality. The estimated bias using the bootstrap method was very small showing that there was little evidence of bias in the estimates. The conclusion reached using the maximum likelihood method agreed with that of the bootstrap method. The expected cross-validation index obtained for the hypothesized model also showed that this model cross-validated over the independent model.

Following the application of path models, a review of some methods where the longitudinal aspects of the data can be taken into account was made. The weekly averages of stem radius were considered as the response variable. The fractional polynomial model in the context of the linear mixed model was formulated and fitted on the weekly data. The functional relationship between stem radius and tree age is identified and the parameters are estimated.

Based on descriptive and graphical exploratory analysis and using the **mfp** package in R, it was found that stem radial measure is a function of linear time plus the square root of time. The selection of random effects resulted in the significance of all three random effects (namely, intercept, coefficients of time, and coefficients of square root of time). The search for the best covariance structure of error component suggested that heterogeneous variance, which varies by clone and exponential function of the square root of time, as the best fit.

It was found that the growth pattern of the two hybrid clones is similar during the juvenile stage. However, the rate of growth for the GU clone is faster than the rate growth for the GC clone. This result supports the results obtained by the previous methods. The fractional polynomial models were extended to account for the effect of the climatic variables. Although tree age is the most important variable in determining the stem radial growth during the juvenile stage (up to two years), there is a significant effect of climatic variables on the stem radial change. Most of the climatic variables have a positive effect on the stem radius during the juvenile stage of tree development. In general, the results obtained using fractional polynomials supports the results obtained by the previous methods described in Chapter 3 and Chapter 4.

Following the fractional polynomial models, nonlinear mixed effects modelling approaches were reviewed, mainly to compare the performance of fractional polynomial models with that of the standard nonlinear growth curves. Although several different methods for estimating the parameters in nonlinear mixed effects models have been proposed, the practical consideration mainly focuses on two of them. These are maximum likelihood and restricted maximum likelihood. The difficulty in evaluating the loglikelihood of the data has a limiting aspect and was evident in the computational phases of fitting nonlinear mixed models where intensive computing times were experienced with very large data sets. In some instances, there were convergence problems with more complex models. This indicates the need to further investigate the performance of possibly simplified methods which would require less powerful computational resources. Frational polynomial models are relatively computationally simple and can be used as an alternative to the the standard nonlinear growth curves. With this idea in mind, nonlinear mixed models are fitted to three selected standard growth curves and their performance is compared with that of fractional polynomials.

All three nonlinear growth functions are fitted to the weekly data. All these three nonlinear mixed models fit the data almost equally well. The assessments of model fit for both fractional polynomial and nonlinear models were made. It was found that the fractional polynomial model was almost as good as the nonlinear models in fitting the data.

For all parametric methods, the form of the underlying relationship between the response and the covariates must be known in advance. Only a few numbers of parameters have to be estimated to get the relationship between the response and covariates. The semi-parametric methods can provide a chance for the underlying relationships to be estimated in a data driven way. That means the type of relationship between the variables is decided by the data rather by intuitions.

Therefore, the application of the semi-parametric models is reviewed and discussed. It was found that the relationship between stem radius and each covariate (tree age, temperature, rainfall, solar radiation, wind speed and relative humidity) can be better explained by a nonlinear relationship. The effect of each covariate on stem radius varies with season. The adjusted R^2 used as a measure of the relationship between the observed and fitted values shows the relationship between tree age and stem radius is the strongest (R^2=0.82).

The AMM that uses the smoothers of tree age and any one of the climatic variables resulted in significant smoothers for tree age in all cases. The smoothers for temperature, rainfall and wind speed did not appear to be significant when tree age is included in the model. This indicates that none of these climatic variables has a significant effect on the growth of stem radius in the presence of tree age.

However, the smoothers for relative humidity for winter (p-value=0.047) and the smoothers for solar radiation for winter (p-value < 0.00001) and autumn (p-value= 0.0006) are significant. Moreover, a model that includes the interaction between wind speed and clone in the parametric part of the additive mixed model was compared to a model without any effect of wind speed. The likelihood ratio statistics favour the model with the interaction of wind speed and clone (p-value= 9e-04).

An additive mixed effects model that includes the smoothers of tree age, wind speed and solar radiation was fitted. The smoothers for tree age and solar radiation appear to be significant. The conclusions made in the semi-parametric methods are in agreement with that of the parametric methods.

This work demonstrates that with suitable statistical modelling of real life data, taking into account the longitudinal nature of the data and scientific backgrounds, a worthwhile contribution to the knowledge and literature in areas of particular application can be made. For example, the findings of this study identified that tree age is the most important variable in explaining stem radial growth at the juvenile stage of the tree. The relationship between stem radius and all covariates can be better explained by a nonlinear relationship.

In summary, six different types of modelling techniques were reviewed and applied in modelling the growth in stem radius. All the analyses demonstrated that these models are useful in the study of factors affecting the longitudinal growth of stem radius. Furthermore the thesis highlighted that fractional polynomials in the framework of linear mixed models can be an alternative to the more complicated ones of nonlinear mixed effects models in modelling growth. There are opportunities for further work in this research. The most important is validating the models with data of matured trees. The applications of similar techniques to adult trees and comparison of the results deserves further research.

In conclusion this work demonstrates that with suitable statistical modelling of real life data, taking into account the longitudinal nature of the data and scientific backgrounds, a worthwhile contribution to the knowledge /literature in areas of particular application can be made. The findings of this study identified that tree age is the most important variable in explaining stem radial growth at juvenile stage of the two hybrid clones. The relationship between stem radius and all weather variables can better be explained by nonlinear relationship. Although only one clone from each hybrid cross is tested in the study, the faster growth features of the GU clone points to enhanced genetics of this hybrid cross and its potential ability to better exploit existing resources, making it an economically feasible hybrid cross as reported elsewhere(Galloway, 2003). Moreover, the study indicated that the effect of weather variables on stem radial growth vary from season to season.

One possible limitation associated with this study is that most of the parameters are estimated using maximum likelihood and restricted maximum likelihood methods. The difficulty in evaluating the loglikelihood of the data has a limiting aspect and was evident in the computational phases of fitting nonlinear mixed models where intensive computing times were experienced with very large data set. In some instances there were also convergence problems with more complex models. The other limitation may be from the data itself. In this analysis only one set climatic variables is used to each time point (tree age). If planting was made over time so that we would have had trees of the same age (example one year) , that would have experienced different values of climatic variables and then the impact of climatic variable could have been determined better.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Proceedings of the Second International Symposium on Information Theory, B.N. Petrov and F. Csaki, (eds). Budpest: Akademiai Kiado. 267-281.

Akaike, H. (1987). Factor Analysis and AIC. *Psychometrika* 52, 317-332.

Apiolaza, L.A., Raymond, C.A., & Yeo, B.J. (2005). Genetic variation of physical and chemical wood properties of *Eucalyptus globulus*. *Silvae Genetica* 54, 160-166.

Arbuckle, J. L. (2006). *Amos 7.0 User's Guide*. Chicago: SPSS.

Ayele, D. (2010). Longitudinal analysis of the effect of climatic factors on the wood anatomy of two eucalyptus clones. M.Sc. thesis in *Statistics.* Pietermartizburg , university of Kwazulu Natal.

Bates, D.M., & Watts, D.G. (1988). *Nonlinear Regression Analysis and its Applications*. New York: John Wiley and Sons.

Benner, A. (2010). *Multivariable Fractional Polynomials* (http://cran.r-project.org/web/packages/mfp/vignettes/mfp.pdf)

Bentler, P.M., & Bonett, D.G. (1980). Significance tests and goodness of fit in analysis of covariance structures. *Psychological Bulletin* 88, 588-606.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K.A., & Curran, P. J. (2006). *Latent Curve Models: A structural equation perspective.* Hoboken, New Jersey: John Wiley & Sons.

Bollen, K.A., & Stine, R.A. (1993). *Bootstrapping goodness-of-fit measures in structural equation modelling.* In: Bollen, K.A., & Long, J.S. (eds). *Testing Structural Equation Models* (pp. 111-135). Newbury Park, CA: Sage.

Box, G.E., Jenkins, G.M., & Reinsel, G.C. (1994). *Time Series Analysis: Forecasting and Control.* ed. San Francisco: Holden-Day.

Breslow, N.E., & Clayton, D.G. (1993). *Approximate inference in generalized linear mixed models. Journal of the American Statistical Association* 88 (421), 9-25.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research.* New York: The Guilford Press.

Browne, M.W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioural Research* 24: 445-455.

Browne, M.W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In: Bollen, K.A., & Long, J.S.(eds). *Testing Structural Equation Models* (pp. 136-162). Newbury Park, CA: Sage.

Browne, M.W., & Mels, G. (1992). *RAMONA User's Guide*. Columbus, OH: Ohio State University.

Byrne, B.M. (2001). *Structural Equation Modelling with AMOS: Basic Concepts, Applications, and Programming*. Mahwah, New Jersey: Erlbaum Associates.

Callaham, R.Z. (1962). Geographic variability in growth of forest trees. In: Kozlowski, T. (ed.), *Tree Growth*. New York: The Ronald Press Company pp. 311-325.

Carroll, R.J. & Ruppert, D. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal of Statistics* 42(2), 205-223.

Chauke, M. (2008). Modelling Environmental factors for two eucalyptus clones. M.Sc. thesis in *Statistics.* Pietermartizburg, University of Kwazulu Natal.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association* 74(368), 829-836.

Crecente-Campo, F., Tome, M., Soares, P., & Dieguez-Aranda, U. (2010). A generalized nonlinear mixed-effect height-diametre model for *Eucalyptus globulus* L. in northern western Spain**.** *Forest Ecological Management* 259, 943–952.

D'Arrigo, R.D., Jacoby, G.C., & Free, R.M. (1992). Tree-ring width and maximum latewood density at the North American tree line: parameters of climate change. *Canadian Journal of Forestry Research* 22, 1290-1296.

Davidian, M., & Gallant, A.R. (1992). Smooth nonparametric maximum likelihood estimation for populations. *Journal of Pharmacokinetics and Biopharmaceuticals,* 20, 529-556.

Davidian, M., & Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measurement Data.* New York: Chapman & Hall.

Dempster, A.P., Rubin, R.B., & Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association,* 76, 341-353.

Deslauriers, A., Morin, H., Urbinati, C., & Carrer, M. (2003). Daily weather response of balsam fir [*Abies balsamea* (L.) Mill.] stem radius increment from dendrometre analysis in the boreal forests of Quebec (Canada). *Trees* (Berl) 17, 477-484.

Diggle, P.J., Heagerty, P., Liang, K.Y., & Zeger, S.L. (2002). The Analysis of Longitudinal Data. 2nd ed. Oxford: Oxford University Press.

Diggle, P.J., Liang, K.Y., & Zeger, S.L. (1994). *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

Dine, E., Yücesoy, C., & Onur, F. (2002). Simultaneous spectrophotometric determination of mefenamic acid and paracetamol in a pharmaceutical preparation using ratio spectra derivative spectrophotometry and chemo metric methods. *Journal of Pharmaceutical and Biomedical Analysis.* 2, 1091–1100.

Downes, G., Beadle, C., & Worledge, D. (1999). Daily stem growth patterns in irrigated Eucalyptus globules and *E.nitens* in relation to climate. *Trees* 14, 102-111.

Downes, G., Drew, D., Battaglia, M., & Schulze, D. (2009). Measuring and modelling stem growth and wood formation: an overview. *Dendrochronologia* 27, 147-157.

Drew, D.M. (2004). *Dendrometre trial phases one technical report.* Report No. EFR092T. Division of Water, Environment and Forestry Technology, CSIR.

Drew, D., Downes, G., Grzeskowiak, V., & Naidoo, T. (2009). Differences in daily stem size variation and growth in two hybrid eucalypt clones. *Trees - Structure and Function* 23, 585-595.

Drew, D.M., & Pammenter, N.W. (2006). Vessel frequency, size and arrangement in two eucalypt clones growing at sites differing in water availability. *New Zealand Journal of Forestry* 51, 23-28.

Eagleman, J.R. (1985). *Meteorology, the Atmosphere in Action.* Belmont, California: Wadsworth Publishing Co. pp.17-284.

Efron, B., & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall. pp. 184-187.

Eksteen, A. (2012). Growth characteristics of three eucalyptus clonal hybrids in response to drought stress: the underlying physiology. Phd thesis in Biological and Conservation Sciences. Durban University of Kwazulu Natal.

Fan, J., & Li, R. (2004). New estimation and model selection procedures for semi-parametric modelling in longitudinal data analysis. *Journal of the American Statistical Association* 99, 710-723.

Faraway, J.J. (2006). *Extending Linear Model with R*. London: Chapman and Hall/CRC.

February, E.C., Stock, W.D., Bond, W.J., & Le Roux, D.J. (1995). Relationships between water availability and selected vessel characteristics in *Eucalyptus grandis* and two hybrids. *IAWA Journal* 16, 269-276.

Fekedulegn, B.D., Colbert, J.J., Hicks, R.R., & Schucker, M.E. (2002). *Coping with multicolinerarity: an example on application of Principal Components Regression in Dendroecology. Research Paper NE-721*. United States Department of Agriculture.

Fitzmaurice, G.M., Laird, N.M., & Ware, J.H. (2004). *Applied longitudinal analysis*. New York: John Wiley & Sons.

Fornell, C., & Bookstein, F. (1982). Two structural equation models: LISREL and PLS applied to Consumer Exit-Voice Theory. *Journal of Marketing Research* 19, 440-452.

Fritts, H.C. (1976). *Tree Rings and Climate.* New York: Academic Press pp. 28-54.

Gallaham, R.Z. (1962). Geographic variability in growth of forests. In: Kozolowski, T. (ed), *Tree Growth*. The Ronald Press Company pp. 311-325.

Galloway, G. (2003). Comparison of the performance of various sources of *E. grandis* and operational clones under operational conditions. Sappi Forests Research File Note ESST002aK.

Garson, G.D. (2004). *Path analysis.* Retrieved February 20, 2012 from http://faculty.chass.ncsu.edu/garson/pa765/path.htm.

Grapentine, T. (2000). Path analysis and Structural Equation Modelling. *Marketing Research* 12, 12-20.

Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Mixed Models*. London: Chapman and Hall.

Haenlein, M., & Kaplan. A.M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics* 3: 283-297.

Hastie, T.J. & Tibshirani, R.J. (1986). Generalized additive models. *Statistical Science* 1, 297-318.

Hastie, T.J. & Tibshirani, R.J. (1990).*Generalized Additive Models*. London: Chapman and Hall.

Hauser, R.M., & Goldberger, A.S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology* 3: 81-117.

Hofgaard, A., Tardif, J., & Bergeron, Y. (1999). Dendroclimatic response of *Picea mariana* and *Pinus banksiana* along a latitudinal gradient in the eastern Canadian boreal forest. *Canadian Journal of Forest Research* 29, 1333-1346.

Heyde, C.C. (1994). A Quasi-likelihood approach to the REML estimating equations. *Statistics & Probability letters* 21, 381-384.

Jagpal, H.S. (1982). Multicollinearity in structural equation models with unobservable variables. *Journal of Marketing Research* 19, 199-218.

James, G. M., Wang, J., & Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics* 37(5A): 2083-2108.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34(2): 183-202.

Jöreskog, K.G., & Wold, D. (1982). The ML and PLS techniques for modelling with latent variable historical and comparative aspects. In: Jöreskog, K.G., & Wold, H. (eds). *Systems under indirect observation: causality, structure, prediction*. Amsterdam: North Holland. pp. 263-270.

Keele, L.J. (2008). *Semi-parametric Regression for Social Sciences*. Chichester: John Wiley and Sons.

Kline, R.B. (2005). *Principles and Practices of Structural Equation Modelling*. 2nd ed. New York: Guilford Press.

Kozlowski, T.T., & Pallardy, S.G. (1997). *Physiology of Woody Plants*. 2nd ed. San Diego: Academic Press. pp. 1-6.

Laird, N.M., & Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* 38, 963-974.

Lin, X. & Carroll, R.J. (2008). Non-parametric and semi-parametric regression methods for longitudinal data. In: *Longitudinal Data Analysis: A handbook of modern statistical methods*. Fitzmaurice, G., Davidian, M., Verbeke, G., & Molenberghs, G. (Eds). London: Chapman & Hall/CRC.

Lin, X., & Zhang, D. (1999). Inferences in generalized additive mixed models using smoothing splines. *Journal of the Royal Statistical Society*, Series B 61, 381-400.

Lindstrom, M. J., & Bates, D.M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* 46, 673-687.

Leamer, E. (1978). *Specification Searches: Ad Hoc Inference with Non-Experimental Data.* New York: John Wiley and Sons.pp.87-106.

Long, J., & Ryoo, J. (2010). Using fractional polynomials to model nonlinear trends in longitudinal data. *British Journal of Mathematical and Statistical Psychology* 63(1): 177–203.

Long, N., & Wand, M.P. (2004). Smoothing with mixed model software, *Journal of Statistical Software* 9(1): 1-56.

Lovie, A.D., & Lovie, P. (1993). *Charles Spearman, Cyril Burt, and the origins of factor analysis.* Journal of the History of the Behavioural Sciences 29(4): 308-321.

Maitra, S., & Yan, J. (2008). *Principal component analysis and partial least squares: two dimension deduction techniques for regression.* Casualty Actuarial Society Discussion Paper Program. pp 79-90.

Malhotra, N.K., Peterson, M., & Kleiser, S. (1999). Marketing research: A state of the art review and directions for the twenty first century. *Journal of the Academy of Marketing Science* 27(2): 160-182.

Mallet, A., Mentre, F., Steimer, J.L., & Lokiek, F. (1988). Nonparametric maximum likelihood estimation for population pharmacokinetics, with application to Cyclosporine. *Journal of Pharmacokinetics and Biopharmaceuticals* 16: 311-327.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometreika* 57, 519-530.

Maruyama, G. M. (1998). *Basics of Structural Equation modelling.* Thousand Oaks, CA: Sage. Pp.60-61.

Mateus, A., & Tomé, M. (2011). Modelling the diametre distribution of *Eucalyptus* plantations with Johnson's S_B probability density function: parameters recovery from a compatible system of equations to predict stand variables. *Annals of Forest Science* 68(2): 325-335.

Mathes, H. (1993). Global optimization criteria of the PLS algorithm in recursive path models with latent variables. In: Haagen, K., Bartholomew, D.J., & Deistler, M. (eds). *Statistical modelling and latent variables.* Amsterdam: Elsevier. pp 229-248.

McDonald, R.P. (1996). Path analysis with composite variables. *Multivariate Behavioural Research* 31, 239-270.

Melesse, S.F. & Zewotir, T. (2013a). The effect of correlated climatic factors on the radial growth of eucalypt trees grown in coastal Zululand of South Africa. *African Journal of Agricultural Research.* 8(14): 1233-1244 Available online at http://www.academicjournals.org/AJAR.

Melesse, S.F. & Zewotir, T. (2013b). Path models approach to the study the effect of climatic factors and tree age on radial growth of juvenile Eucalyptus hybrid clones. *African Journal of Agricultural Research.* 8(14), 2685-2695 Available online at http://www.academicjournals.org/AJAR.

Meng, S.X. & Huang, S. (2010). Incorporating correlated error structures into mixed forest growth models: prediction and inference implications. *Canadian Journal of Forest Research* 40(5): 977-990.

Mevik, B.H., & Cederkvist, H.R.(2004). Mean Squared Error of Prediction (MSEP) estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). *Journal of Chemometrics* 18, 422-429.

Mevik, B.H., Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. Journal of Statistical Software 18(2). 1-24.

Miehle, P., Battaglia, M., Sands, P.J., et al. (2009). A comparison of four process-based models and statistical regression model to predict growth of *Eucalyptus globulus* plantations. *Ecological Modelling* 220, 734-746.

Miller, G.J. (2001). Environmental Science: Working with the Earth. 8th ed. Pacific Grove, CA: Brooks/Cole. Pp.84-104.

Pallett, R.N., & Sale, G. (2004). The relative contributions of tree improvement and cultural practice towards productivity gains in *Eucalyptus* pulpwood stands. *Forest Ecology and Management* 193, 33-43.

Phipps, R.L. (1982). Comments on interpretation of climatic information from tree rings, eastern North America. *Tree Ring Bulletin* 42, 11-22.

Pinheiro, J.C., & Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics:* 4(1): 12-35.

Pinheiro, J.C. & Bates, D. M. (2000). *Mixed Effects Models in S and S-Plus*. New York: Springer.

R Core Team. (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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

Rodriguez-Nogales, J.M. (2006). Approach to the quantification of milk mixtures by partial least-squares, principal component and multiple linear regression techniques. *Food Chemistry* 98, 782-789.

Royston, P. & Altman, D.G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). *Journal of the Royal Statistical Society. Series C* (*Applied Statistics)* 43, 429-467.

Royston, P., Ambler, G. & Sauerbrei, W. (1999). *T*he use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 28(5):964-974.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *J Comput. Graph Stat* 11: 735-757.

Ruppert, D. Wand M.P & Carroll, R.J. (2003). *Semi-parametric Regression.* Cambridge, UK: Cambridge University Press.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. Holland: D. Reidel Publishing Company.

Schneeweiss, H. (1993). Consistency at large in models with latent variables. In: Haagen, K., Bartholomew, D.J., & Deistler, M. (eds). *Statistical modelling and latent variables.* pp 299-320. Amsterdam: Elsevier.

Schulze, R.E. (1997). *South African atlas of agro- hydrology and climatology.* South African water research commission (WRC). Pretoria soil classification working group (1991) soil classification : a taxonomic system for South Africa . ARC –Institute for soil, climate and water, Pretoria.

Schumacker, R.E. (1991). Relationship between multiple regression, path, factor and LISREL analysis. *Multiple Linear Regression Viewpoints.*18, 28-46.

Schumacher, R.E. & Lomax, R.G. (2004). A Beginner's Guide to Structural Equation Modelling. 2nd ed. Lawrence Erlbaum Associates Inc. pp.328-329.

Schweingruber, F.H., Briffa, K.R., & Nogler, P. (1993). A tree-ring densitometric transect from Alaska to Labrador. *International Journal of Biometeorology* 37, 151-169.

Searson, M.J., Thomas, D.S., Montagu, K.D., & Conroy, J.P. (2004). Wood density and anatomy of water limited eucalypts. *Tree Physiology* 24, 1295-1302.

Seber, G.A.F., & Wild, C.J. (2003). *Nonlinear Regression.* New Jersey: John Wiley and Sons.

Sheiner, L.B., & Beal, S.L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters. I.Michaelis-Menton Model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetic and Biopharmaceutics.* 8(6): 553-571.

Steiger, J.H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioural Research* 25, 173-180.

Thisted, R.A. (1988). *Elements of Statistical Computing.* London: Chapman and Hall.

Tierney, L., & Kadane, J.B. (1986). Accurate approximations for posterior moments and densities. *Journal of the American Statistical Association.* 81(393): 82-86.

Turnbull, J.W. (1999). Eucalyptus plantations. *New Forests* 17, 37-52.

Valeria, M. (2011). Additive Mixed Models applied to the study of red shrimp landings: comparison between frequentist and Bayesian perspectives. (http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_393.pdf)

Verbeke, G., & Molenberghs, G. (1997). *Linear Mixed Models in Practice: a SAS oriented approach. Lecture notes in Statistics* 126. New York: Springer-Verlag.

Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Vonesh, E.F., & Carter, R.L. (1992). Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics* 48, 1-18.

Vonesh, E.F., & Chinchilli, V.M. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wadsworth, R.M. (1959). An optimum wind speed for plant growth. *Annals of Botany* 23, 195 -199.

Wahba, G. (1985) A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Annals of Statistics* 13, 1378-1402.

Wakefield, J.C., Smith, A.F.M., Racine-Poon, A., & Gelfand, A.E. (1994). Bayesian analysis of linear and nonlinear population models using Gibbs Sampler. *Appl. Statist.* 43: 201-221.

Wand, M.P. (1999). On the optimal amount of smoothing in penalized spline regression. *Biometreika* 86, 936-940.

Wang, T. (2012). Linear Mixed Effects Model for a Longitudinal Genome Wide Association Study of Lipid Measures in Type 1 Diabetes. *Open Access Dissertations and Theses.* Paper 7468.

Weiss, R. E. (2005). *Modelling Longitudinal Data*. New York: Springer.

West, B., Welch, K.B., & Galecki, A.T. (2006). *Linear Mixed Models: A Practical Guide Using Statistical Software*. New York: Chapman & Hall /CRC.

West, S.G., Finch, J.F. & Curran, P.J. (1995). Structural equation models with non-normal variables: Problems and remedies. In: Hoyle, R.H. (ed). *Structural Equation Modelling: Concepts, Issues, and Applications*. Thousands Oakes, CA: Sage. pp. 56-75.

Whitehead, D., & Jarvis, P.G. (1981). Coniferous forests and plantations. In: Kozlowski, T.T. (ed). *Water Deficits and Plant Growth*. New York: Academic Press. pp 50-153.

Wold, H. (1982). Soft modelling: the basic design and some extensions. In: Jöreskog, K.G., & Wold, H. (eds). *Systems under Indirect Observations: Causality, Structure, Prediction.* Amsterdam: North Holland Publishing. pp. 1-54.

Wold, H. (1985). Systems analysis by partial least squares. In: Nijkamp, P., Leitner, H., & Wrigley, N. (eds). *Measuring the Unmeasurable*. Boston: Martinus Nijhoff. pp 221-251.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika* 80: 791-795.

Wolfinger, R. (1999). Fitting nonlinear mixed models with the new NLMIXED procedure. Proceedings of the 24th Annual SAS Users Group International Conference. SAS Institute Inc., Cary, NC. pp: 287.

Wood, S.N. (2006a). *Generalized Additive Models: An introduction with R.* Boca Raton Fl: CRC Press.

Wood, S.N. (2006b). mgcv: Multiple Smoothing Parameter Estimation by GCV or UBRE. URL: http://www.maths.bath.ac.uk/~sw283/.

Wood, S.N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society* (B) 73(1): 3-36.

Wright, S. (1918). On the nature of size factors. *Genetics* 3(4): 367-374.

Wright, S. (1920). The relative importance of heredity and environment in determining the piebald pattern of guinea pigs. *Proceedings of the National Academy of Sciences* 6, 320-332.

Wright, S. (1921). Correlation and causation. Part I: methods of path coefficients. *Journal of Agricultural Research* 20, 557-585.

Wright, S. (1923). The theory of path coefficients. A replay to Niles's criticism. *Genetics* 8(3): 239-255.

Wright, S. (1934). The method of path coefficients. *The Annals of Mathematical Statistics* 5(3): 161-215.

Wu, H., & Zhang, J.T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis.* Hoboken, New Jersey: John Wiley and Sons.

Yung, Y-F., & Bentler, P.M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In: Marcoulides, G.A. and Schumacker, R.E. (eds). *Advanced Structural Equation Modelling: Issues and Techniques.* Mahwah, N.J: Lawrence Erlbaum Associates pp.195-226.

Zellner, A. (1970). Estimation of regression relationships containing unobservable variables. *International Economic Review* 11, 441-454.

Zewotir, T., & Galpin, J.S. (2004). The behaviour of normality under non-normality for mixed models. *South African Statistical Journal* 38, 115-138.

Zuur, A.F., Leno, E.N., Walker, N.J., Savliev, A.A., & Smith, G.M. (2009). *Mixed effects models and extension in ecology with R.* Springer.com http://link.springer.com/book.

Zweifel, R., & Häsler, R. (2001). Dynamics of water storage in mature subalpine *Picea abies*: temporal and spatial patterns of change in stem radius. *Tree Physiology* 21, 561–569.

Zweifel, R., Zimmerman, L., Zeugin, F., & Newbery, D.M. (2006). Intra-annual radial growth and water relations of trees: implication towards a growth mechanism. *Journal of Experimental Botany* 57, 1445-1459.

academicJournals

Vol. 8(14), pp. 1233-1244, 18 April, 2013
DOI: 10.5897/AJAR12.1671
ISSN 1991-637X ©2013 Academic Journals
http://www.academicjournals.org/AJAR

African Journal of Agricultural Research

Full Length Research Paper

The effect of correlated climatic factors on the radial growth of eucalypt trees grown in coastal Zululand of South Africa

Sileshi F. Melesse* and Temesgen Zewotir

School of Mathematics Statistics and Computer Science, University of KwaZulu-Natal, South Africa.

Accepted 28 December, 2012

Understanding the relationship between stem radial growth and climatic conditions in plantation productivity is important to identify the climatic factors that most influence tree growth. This study aims to determine the climatic factors that most influence the stem radial growth of eucalypt trees plantation in the coastal Zululand area of South Africa. Daily stem radius was measured using automated point dendrometers located on 18 sample trees of *Eucalyptus grandis* × *Eucalyptus urophylla* (GU) and *E. grandis* × *Eucalyptus camaldulensis* (GC) hybrid clones. Daily averages of climatic data (temperature, solar radiation, relative humidity and wind speed) and total rainfall were also obtained from the site over the study period. Several statistical models that cope with the issue of multicollinearity were applied. Weather variables, together with tree age, explained a substantial amount of the variation (87% for GC clone and 79% for GU clone) in the daily stem radius. This study indicates that tree age is the most important factors that influence stem radius during the juvenile stage (up to 2 years) in all seasons. In winter, temperature, relative humidity and wind speed appear to be more important than the other weather variables.

Key words: Tree radial growth, latent variables, multicollinearity, ordinary least squares, partial least squares, principal component regression, plantation.

INTRODUCTION

Increasingly, eucalypts have become the most widely planted hardwood species in the world (Turnbull, 1999). At present, eucalypts provide sawn timber, mine props, pulp and paper, fiberboard, poles, firewood, charcoal, essential oils, nectar for honey, tannin, shade, and shelter. Most eucalypt plantations are established and managed for profit. The rate of growth is an important economic factor, and plantations with faster growth will be available for processing earlier compared with slower growth plantations. Tree growth and wood production are product of the interaction between genetic and environmental factors (Callaham, 1962). Some studies have found significant effects of environmental factors on wood property variation in *Eucalyptus* (February et al., 1995; Searson et al., 2004; Drew and Pammenter, 2006). Extensive literature on genetic factors affecting the growth of trees can be found in the work of Kozlowski and Pallardy (1997). The most recent work by Downes et al. (2009) provides an excellent overview on measuring stem growth and wood formation. Other examples are those by Drew et al. (2009), which focussed on differences in daily stem diameter variation and growth in

*Corresponding author. E-mail: melesse@ukzn.ac.za.

two hybrid eucalypts, and Zweifel et al. (2006) who studied the intra-annual radial growth and water relations of trees and the implications on growth mechanisms.

In a study that considered the data extracted from the same database as used in this study, Drew et al. (2009) found that the *Eucalyptus grandis* × *Eucalyptus urophylla* (GU) clone had fewer days on which net growth occurred than did the *E. grandis* × *Eucalyptus camaldulensis* (GC) clone. However, when growth did occur, the GU grew for longer each day and at a higher rate than did the GC. Thus, it still had an overall larger net stem increment during the study period. Drew et al. (2009) studied the relationship between stem radius and climatic factors using the correlation matrix.

Weather variables such as temperature, radiation, rainfall, humidity, and wind speed all contribute to the growth of the tree. For instance, Downes et al. (1999) studied daily radial stem growth in irrigated *Eucalyptus globulus* and *Eucalyptus nitens* in relation to climate over a 12-month period using multiple linear regression models. The study, which was conducted in southern Australia, showed that daily weather variation accounted for 40 to 50% of the variance in the daily increment of stem radius. Downes et al. (1999) also argued that understanding the relationship between weather and the rate and pattern of stem growth will facilitate the prediction of wood properties at a given site. Our approach provides an alternative one to the methods used by Downes et al. (1999). A study by Phipps (1982) presented a general discussion regarding problems inherent to developing climatically sensitive tree-ring chronologies from eastern North America. The same study by Phipps (1982) indicated that tree ring collections from eastern forests are typically not climatically sensitive as western collections. A general treatment of dendroclimatology can be found in the work of Fritts (1976). Other studies such as those by D'Arrigo et al. (1992), Hofgaard et al. (1999) and Schweingruber et al. (1993) reported that late spring or summer temperatures had a positive effect on annual growth. Zweifel et al. (2001) showed that radius change could be determined by stem water content and wood bark growth, including the degradation of dead phloem cells. The water related fraction is a short-term effect lasting from a few hours to several weeks, and can either have positive or negative effects on stem radius, depending on the changing turgor of stem tissues (Zweifel et al., 2001).

The contribution of each climatic variable is often influenced, by correlation, with one or more other climatic variables. However, studies that consider the effects of colinearity into account are limited. Studies commonly use diameter at a given tree age as an indicator of growth rate and pattern. Most eucalypt plantations are limited by rainwater for growth, therefore identification of the relationship between natural climatic conditions and radial increment is important for eucalypt plantation managers. In order to manage resources effectively, it is important for tree growers to understand the properties of the material being produced. This paper describes the effects of climatic variation on radial growth of GU and GC hybrid clones established in Zululand on the eastern coast of South Africa. The focus of this study is to determine the climatic factors that influence radial growth during the juvenile (the first 2 years of age) stages of tree growth. This is mainly because these data are the data collected on phase one of the data collection process. Moreover, the study of juvenile trees is very important, to have a productive matured tree. The primary question addressed by this study concerns the extent to which classical regression approaches are successful in detecting and estimating the effects of climatic conditions on stem radial growth. A secondary aim is to present latent variable modeling approaches, namely partial least squares (PLS) and principal component regression, for better estimation and detection of effects of the climatic variables.

MATERIALS AND METHODS

Study design

The research site is located near the town of KwaMbonambi in KwaZulu-Natal, South Africa, (28.53° S, 32.14° E, 55 M a.m.s.l), approximately 200 km north-east of the city of Durban. On average, the site receives 1,000 mm of rainfall per annum and has a mean annual temperature of 21°C (Drew et al., 2009). The *Eucalyptus* fiber research experiment was initiated in July, 2001 and a huge database acquired. The experiment was designed to run over a 7-year period and was divided into separate phases. Each phase ended with the destructive sampling of study trees to measure anatomical characteristics of the wood. The results presented in this paper are based on the data collected during the first of these phases, from April, 2002 until August, 2003. The data were used by Drew et al. (2009) and this particular study is extracted from the same database put in place by Sappi (One of the leading suppliers of coated fine paper and chemical cellulose). However, the two data sets are not exactly the same. Two commercially deployed *Eucalyptus* hybrid clones, GU and GC, were planted at the study site (Drew, 2004). According to the South African soil classification system, the soil was identified as Rhodic Ferralsol Hutton by a limited soil survey undertaken at the site (Soil classification workshop group, 1991). The soil is medium grade sand with clay percent in the lower B-horizon not exceeding 40%, and in A-horizon not exceeding 10% with an average depth of A-horizon 20 cm and total potential rooting depth in excess of 1.8 m (Drew et al., 2009). Planting took place on 16 July, 2001, prior to which in April, 2001, stumps of trees from the previous rotation were treated with herbicide (to prevent coppicing), and harvest slash was burned. Each rooted cutting was planted between existing stumps, with approximately 2 L of water and 125 g granular fertilizer, the equivalent of 8 g nitrogen, 12 g phosphorus and 8 g potassium per plant. The two clones were planted in alternating rows seven trees wide each (Figure 1), with spacing between trees of 3 m (east to west) × 2.5 m (north to south). These rows have been numbered from 1 to 6, starting at row (GC) closest to the entrance gate. Each row of clones consists of three plots of 12 trees each with two surrounding rows of trees (Figure 1). This effectively separates each plot by four rows of trees, an important part of the design since periodic destructive sampling is required in the experiment. The plots were established as pairs, such that for any phase of the

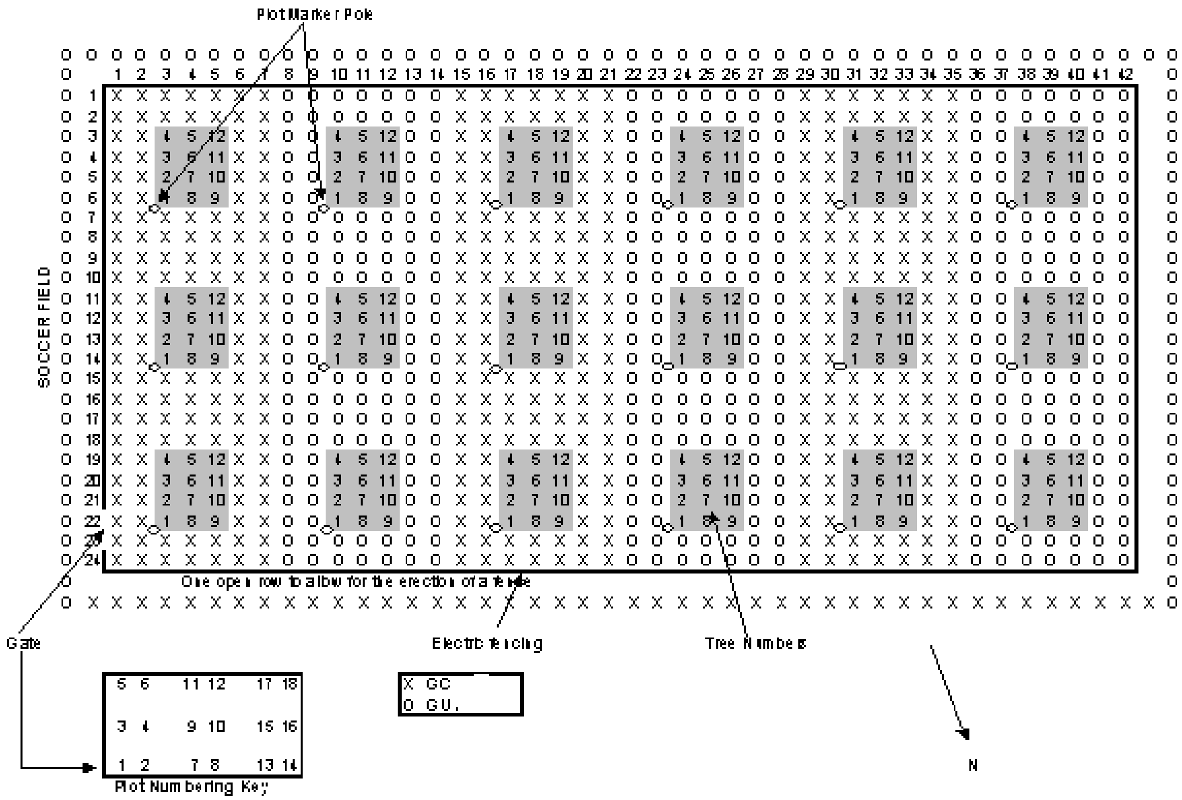

Figure 1. The layout of the experimental plots at the research site in eastern South Africa.

research, a GU and a GC plot could be measured simultaneously (Drew, 2004). From the 18 plots (Figure 1), plots 9 and 10 were chosen for monitoring during project Phase 1. Within a 12-tree plot, nine trees were selected from each clone for intensive monitoring of radial growth and other physiological characteristics (Drew, 2004). Measurements of stem radius were obtained from hourly dendrometer readings in the 18 sample trees. Automatic point dendrometers were mounted at 9 months of age at 1.3 m above the ground on the north side of each tree to measure the radius of the main stem with a rod held against the outside surface by constant force. The data for stem radius used in this paper has 8640 observations from the two clones. Half the data set is from the GU clone and the remaining half is from the GC clone. Daily measurements were used in our analysis. Daily averages of stem radius were obtained by cumulating and averaging the hourly measurements. Meteorological data was obtained using an automatic weather station (MCSystems, Cape Town, South Africa) located approximately 300 m from the research trial site (Drew et al., 2009). Hourly measurements were made of total rainfall (mm), temperature (ºC), relative humidity (%), wind speed (m/s) and total solar radiation (mJ/h). Daily total rainfall and daily averages of the other weather variables were used in the analysis.

Data analysis

Statistical analysis was undertaken using R-statistical software. R is a free software that can be downloaded from the R-project website R Core Team (2012). The simplest approach in detecting climatic effects is by the use of traditional regression methods. However, this traditional method assumes that the climatic variables are uncorrelated since one of the failures of regression methods is due to multicollinearity. Multicollinearity problem arises when the predictors (in our case the climatic variables) are correlated. To overcome this, we applied principal component regression and PLS regression. These methods were applied to the combined data set as well as to the data set for separate clones. Extensive discussion of these methods can also be found in Rodriguez-Nogales (2006), Dine et al. (2002), Fekedulegn et al. (2002), Maitra and Yan (2008), Mevik and Cederkvist (2004), and Haenlein and Kaplan (2004).

RESULTS AND DISCUSSION

The variables included in the study are major climatic variables and one non-climatic variable (tree age) as previously described. The overall ordinary least squares (OLS) model was significant with an R^2 = 0.791 and adjusted R^2 = 0.79 (Table 1). This indicates that about 79% of the variation in stem radius is explained by the predictors (the five weather variables together with age of a tree) included in the model. An attempt to explore lags was made by considering lags up to 15 days. The use of five weather variables lagged by 15 days increased the variance explained by 0.3% only. Therefore, we did not consider the lags as an important issue at this age of the tree.

The predictors included in the model are therefore important for determining radial tree growth. However,

Table 1. Summary OLS model.

Predictor (climatic variables)	Estimate	Standard error	t-value	p-value
Intercept	-16558.67	550.61	-30.07	0.000
Temperature	23.73	12.65	1.88	0.061
Solar radiation	2865.35	222.01	12.91	0.000
Rainfall	2.57	6.21	0.41	0.679
Wind speed	1426.83	77.02	18.53	0.000
Tree age	313.22	2.21	142.05	0.000
R^2 = 0.791		Adj R^2 = 0.79		

Table 2. Correlation matrix of predictors.

Variable	Temperature	Relative humidity	Solar radiation	Rainfall
Temperature	1			
Relative humidity	-0.320**	1		
Solar radiation	0.617**	-0.498**	1	
Rainfall	-0.107**	0.272**	-0.258**	1
Wind speed	0.406**	-0.385**	0.374**	0.099**

*Correlation is significant at the 0.05 level (2-tailed). **Correlation is significant at the 0.01 level (2-tailed).

Table 3. The eigen value decomposition of the correlation matrix.

Eigen values	Proportion of total	Cumulative proportion of total
2.375	0.396	0.396
1.252	0.209	0.605
1.083	0.181	0.786
0.625	0.104	0.890
0.412	0.069	0.959
0.253	0.042	1

the individual t-ratios (estimated coefficient/standard error) for the coefficients of the most important climatic variables, that of rainfall and temperature, are non-significant (Table 1). This is an indication of the presence of multicollinearity among the predictors. From the correlation matrix of predictors (Table 2), temperature and solar radiation were highly correlated. The correlation coefficient was 0.62 and highly significant ($p < 0.001$). The correlation between wind speed and temperature was 0.41, which was also highly significant ($p < 0.001$). This shows the existence of significant multicollinearity among the independent climatic variables. Multicollinearity inflates the standard error of the regression coefficients, which results in low t-statistic values and a failure to reject the null hypothesis. The application of classical linear regression models therefore does not have a powerful inference on the regression coefficients. To address this problem, principal component regression and PLS regression techniques were used. All predictors were treated as continuous variables with different unit of measurements [for instance, rainfall (mm) and temperature (°C)]. It might make more sense to standardize the predictors before trying principal components. This is equivalent to performing principal components analysis on the correlation matrix of predictor variables. Table 3 shows the eigen value decomposition of the correlation matrix of the original or the covariance matrix of the standardized predictors. The first five principal components captured 95.9% of the information in the correlation matrix. Table 4 shows the eigen vectors corresponding to each of the eigen values of Table 3. We constructed the principal components corresponding to each eigen value by linearly combining the standardized predictive variables using the corresponding eigen vector. Hence, the six principal components are computed as follows:

Table 4. The eigen vectors associated with the eigen values of Table 3.

Eigen vector 1	Eigen vector 2	Eigen vector 3	Eigen vector 4	Eigen vector 5	Eigen vector 6
0.495	-0.239	-0.031	0.601	-0.463	0.347
-0.488	-0.415	0.085	0.301	-0.362	-0.593
0.546	-0.144	0.168	0.238	0.553	-0.539
-0.207	-0.255	-0.808	0.259	0.396	0.127
0.413	-0.280	-0.431	-0.594	-0.366	-0.279
-0.068	-0.774	0.354	-0.266	0.241	0.378

Table 5. Summary of OLS model that uses principal components as predictors.

Coefficient	Estimates	Standard error	t-value	p-value
Intercept	16025.71	36.70	439.659	<2e-16***
PC_1	60.83	23.82	-2.554	0.0107*
PC_2	-5402.82	32.80	-164.713	<2e-16***
PC_3	1987.07	35.27	56.34	<2e-16***
PC_4	-1742.90	46.42	-35.547	<2e-16***
PC_5	1330.27	57.18	-23.263	<2e-16***
PC_6	1425.38	72.99	19.530	<2e-16***

*Significance at the 0.05 level. ***significance at the 0.001 level.

$$PC1 = 0.49\ Z_1\ - 0.49\ Z_2\ + 0.55\ Z_3\ - 0.21\ Z_4\ + 0.41\ Z_5\ - 0.07\ Z_6$$
$$PC2 = - 0.24 Z_1 - 0.42\ Z_2\ - 0.14\ Z_3\ - 0.26\ Z_4 - 0.28\ Z_5\ - 0.77\ Z_6$$
$$\vdots$$
$$PC6 = 0.47 Z_1 - 0.59\ Z_2\ - 0.54 Z_3 + 0.13\ Z_4 - 0.28\ Z_5\ + 0.38\ Z_6$$

where Z_1 is the standardized value of temperature, Z_2 is the standardized value of relative humidity, Z_3 is the standardized value of solar radiation, Z_4 is the standardized value of rainfall, Z_5 is the standardized value of wind speed, and Z_6 is the standardized value of age

The principal components constructed above were used in a linear regression model. Stem radius was used as the dependent variable and the principal components as independent variables (Table 5). The rank of the predictive power did not line up with the order of the principal components. For instance, the first principal component was less explanatory for the target than the second or the third, even though the first principal component contains more information on the six original explanatory variables. The principal components technique arrives at uncorrelated standardized linear combinations (SLCs) that capture only the characteristics of the X-vector or predictive variables. No significance is given as to how each predictive variable is related to the response variable. In a way, it is an unsupervised dimension reduction technique (Maitra and Yan, 2008) and therefore requires the use of other analytical methods such as PLS.

In comparing the importance of the constructed principal components, five components explained most of the variation in the predictors (95.9%). The scree plot (not shown here) showed that almost all the variation in predictors (about 96%) was explained by the first five principal components. Therefore, a linear model that used the first five principal components as latent explanatory variables was fitted (Table 6). The R^2 value 0.78 for the reduced model was close to the R^2 value for the model with all six components (R^2 = 0.79). Once again, the rank of the predictive power did not correspond with the order of the principal components. In other words, principal component one appears to have less explanatory power for the dependent variable as compared to other components. By transforming the principal components back to the original explanatory variables, the estimated coefficients of the original variables are given in Table 7. That means, firstly, the principal components were obtained. These principal components are uncorrelated and an ordinary regression model was fitted using the principal components as explanatory variables. The five principal components appear to have significant effect on the radial measure (Table 6). The estimated coefficients for the original measured variables were obtained by transformation from the estimated coefficients for principal components. The estimates of the regression coefficients in Table 7 show that all predictors have a positive relationship with stem radial measure. Moreover, the five latent variables that produced the above estimated coefficients are significant (Table 6). This indicates the significant effect of climatic variables on radial measure. Separate estimates for GU and GC clones also show the positive

Table 6. Summary of OLS results for the model that uses the first five principal components.

Coefficient	Estimates	Standard error	t-value	p-value
Intercept	16025.71	37.50	427.35	<2e-16***
PC_1	60.83	24.34	-2.50	0.0124*
PC_2	-5402.82	33.52	-161.20	<2e-16***
PC_3	1987.07	36.04	55.14	<2e-16***
PC_4	-1742.90	47.43	-36.75	<2e-16***
PC_5	1330.27	58.43	-22.77	<2e-16***

*Significance at the 0.05 level. ***Shows significance at the 0.001 level.

Table 7. The estimated coefficients of the original climatic variables estimated by using principal component regression.

Predictors (Climatic variables)	Estimates for combined data	Estimates for GU clone	Estimates for GC clone
Intercept	-16558.67	-19048.26	-14069.07
Temperature	90.48	165.33	15.64
Relative humidity	581.14	680.05	482.29
Solar radiation	694.56	802.99	586.20
Rainfall	16.81	27.82	5.79
Wind speed	834.13	902.12	766.24
Tree age	6201.39	6764.65	5638.85

Table 8. Estimated coefficients of the original set of climatic variables using PLS method.

Climatic variable	Estimates for both clones	Estimates for GU clone	Estimates for GC clone
Temperature	55.42	128.02	54.42
Relative humidity	596.58	696.94	596.58
Solar radiation	761.13	874.50	761.13
Rainfall	35.13	47.59	35.13
Wind speed	814.29	880.65	814.29
Tree age	6191.69	6754	6191.69

effect of weather variables together with tree age (Table 7). Partial least square regression (PLS) can overcome the deficiencies of OLS regression in the case of highly collinear data. Moreover, partial least squares allow an analysis of the data in terms of independent latent variables or components. Applying PLS method to the data, the minimum root mean square error of prediction (RMSEP) is observed for five components model. The value of the X-variance for the model with five latent variables is 93.5 %. This means a model with five latent variables has explained 93.5 % of the variation in the original predictors. The variation explained in the response variable is 79.1 %. This is the same amount of variation explained by the ordinary least square regression. Therefore, the model formulated by five latent variables fits the data well with a high predictive power. The coefficients for the original set of variables when partial least square regression was applied to GC, GU and pooled data sets are indicated in Table 8. It appears that the estimated coefficients for the original set of variables for the GC clone are smaller than that of the GU clone for all climatic variables. This indicates that the GU clone has on average a larger stem radius than the GC clone. The signs of the estimated coefficients for the GU clone and the signs for the estimated coefficients of the pooled data set are the same. However, the estimated coefficient of temperature is negative for the GC clone while it is positive for the GU clone and pooled data set. This indicates that the effect of temperature on stem radius goes in opposite directions for the two clones for this site and age class. The possible reason for this could be the difference in genetic makeup the two clones. Moreover, the effect of weather variables may depend on the season of the year. The site difference cannot be a possible reason for this difference as site difference is controlled by the design. In the design the plots were established as pairs such that a GU and a GC plots are measured simultaneously (Figure 1). For the rest of the climatic variables the effect follows the same direction for the two clones with some differences in magnitude.

In order to test whether the components that produced these coefficients are significant or not, latent variables or

Table 9. Summary of OLS results for the model that uses the PLS components as predictors.

Coefficient	Estimates	Standard error	t-value	p-value
Intercept	16025.71	36.70	436.64	<2e-16***
T_1	5932.81	32.29	178.23	<2e-16***
T_2	1193.6	45.41	26.28	<2e-16***
T_3	318.38	30.45	10.46	<2e-16***
T_4	299.85	40.22	7.46	9.83e-14***
T_5	212.74	48.99	4.34	1.42e-05***
T_6	78.66	58.87	1.336	0.182

***Significance at the 0.001 level.

Table 10. Summary of OLS results for the model that uses the first five PLS components as predictors.

Coefficient	Estimates	Standard error	t-value	p-value
Intercept	16025.71	36.70	436.64	<2e-16***
T_1	5932.81	32.29	178.23	<2e-16***
T_2	1193.6	45.41	26.28	<2e-16***
T_3	318.38	30.45	10.46	<2e-16***
T_4	299.85	40.22	7.46	9.83e-14***
T_5	212.74	48.99	4.34	1.42e-05***

***Significance at the 0.001 level.

Table 11. RMSE and RMSECV values for all prediction methods.

Parameter	OLS	PCR	PLS
RMSE	3410.01	3484.53	3410.4
RMSECV	3414.39	3413	3413

PLS components were constructed while fitting the PLS regression. After determining these latent variables, $T_1 \ldots T_6$ sequentially, the relationship between these latent constructs and the response was estimated by ordinary linear regression. The sample correlations between any pair of the latent constructs were zero. A linear model was then applied using the same radial measure as the dependent variable and the six PLS components, $T_1 \ldots T_6$, as the independent variables. Summary results for the model that uses the PLS components as predictors is shown in Table 9. The PLS components were extracted in order of significance. The first five components were significant, while the last component was not. The values of R^2 and adjusted R^2 for this model were 0.7908 and 0.7907, respectively. Table 10 shows the summary results for the model that involves only five PLS components. From the results, it can be seen that all the coefficients listed in Tables 9 and 10 were the same for the first five components. This shows that the coefficients of the PLS latent variables do not change by adding or dropping latent variables from the model. The results of the PLS model show that all climatic variables had a significant effect on growth.

With regard to the predictive powers of these models, a comparison was made based on RMSE and the RMSE of cross-validation (RMSECV, Table 11), a measure of the model's ability to predict new samples. The OLS model had the smallest RMSE value (Table 11). The second smallest RMSE values belong to the PLS model.

The RMSE for PLS was actually very close to the RMSE for the OLS model. However, this comparison was from the point of view of model fit. Under the condition of no multicollinearity, this might indicate that the OLS model fitted the data better than the other two methods. For comparisons of models intended for prediction, it is inadequate to look just at model fit. As prediction is the objective, the PLS model that gave the lowest RMSECV value is preferred. For the data set to which these models were applied, the PLS model had the highest predictive ability with the lowest number of factors. In order to identify differences between clones, separate PLS model was fitted to data for each clone. For both clones, the optimum number of PLS components was five. These five components were significant, while the sixth

Table 12. Percent of variance captured by PLS components for GU clone.

Component	Climatic variables and age		Radius	
	This component	Cumulative total	This component	Cumulative total
T_1	20.53	20.53	77.53	77.53
T_2	17.66	38.19	1.86	79.04
T_3	30.25	68.44	0.35	79.39
T_4	15.27	83.71	0.14	79.53
T_5	9.8	93.51	0.04	79.57

Table 13. Percent of variance captured by PLS components for GC clone.

Component	Climatic variables and age		Radius	
	This component	Cumulative total	This component	Cumulative total
T_1	20.47	20.47	84.74	84.74
T_2	12.25	32.72	2.06	86.80
T_3	25.85	58.57	0.25	87.05
T_4	24.28	82.85	0.11	87.16
T_5	10.68	93.53	0.05	87.21

Table 14. Standardized regression weights for both principal component regression and PLS regression models.

Predictor	PLS model		PCR model	
(climatic variables)	GU	GC	GU	GC
Temperature	0.016	-0.003	0.020	0.002
Relative humidity	0.086	0.078	0.083	0.075
Solar radiation	0.107	0.101	0.098	0.091
Rainfall	0.006	0.004	0.003	0.001
Wind speed	0.108	0.116	0.110	0.119
Tree age	0.829	0.876	0.830	0.878

component was not significant (Table 9). The percentage of total variation in radial measure captured by the number of components for the GU clone is less (Table 12: 80% with p-value < 0.0001) than the amount of variation captured for the GC clone (Table 13: 87.21% with p-value < 0.0001). The percentage of total variation in climatic variables and tree age captured by the five components PLS model for the GU and GC clones is almost the same (93.5%).

In order to determine the most important drivers of variation in short-term stem radial measure (for the first 2 years of tree age) for the two clones, we applied standardized regression weights for both PLS and principal component regressions. This can be obtained by fitting the models on standardized variables. The factor with the highest coefficient in absolute value is the most important factor in explaining the variation in radial measure. The standardized regression weights (coefficients) for our predictors, when PLS regression and principal components regression were applied to GC and GU data sets, are indicated in Table 14. It appears that tree age is the most important predictor of stem radius using both models and for both clones. Among climatic variables, it appears that wind speed, followed by solar radiation, is the most important driver of the variation in stem radius over the growth period of 2 years. However, the biological plausibility of these results is questionable. Moreover, we found the negative effect of temperature for GC clone. This might be due to the dependence of weather variables on season. The weather variables are likely to change over the year. This relative effect of weather variable might change from one season to the other. We analyzed the same data by season in order to see for the season effect. Summary results by season are shown in Tables 15 and 16. In spring and summer, none of the weather variables has significant effect. The only variable that has significant effect on stem radius is tree age. In winter, all predictors have significant effect on stem radius for GU clone, while for GC clone all have significant effect with the exception of rainfall. In autumn,

Table 15. Summary results of ordinary regression model for summer and autumn.

Predictor	Summer			
	GC clone		GU clone	
	Estimate	p-value	Estimate	p-value
Intercept	2763.099	0.265	2695.785	0.588
Temperature	-2.143	0.963	-17.097	0.854
Relative humidity	5.088	0.781	9.983	0.786
Solar radiation	167.126	0.712	371.769	0.683
Rainfall	0.291	0.990	0.422	0.993
Wind speed	-47.827	0.813	-80.071	0.844
Tree age	185.506	0.000	231.252	0.000
	R^2 = 0.107		R^2 = 0.045	
Predictor	Autumn			
	GC clone		GU clone	
	Estimate	P-value	Estimate	P-value
Intercept	-11156.222	0.000	15921.22	0.000
Temperature	-12.152	0.578	28.38	0.377
Relative humidity	8.632	0.441	19.62	0.233
Solar radiation	1055.849	0.028	1907.87	0.007
Rainfall	13.029	0.550	23.89	0.029
Wind speed	378.068	0.011	476.58	0.029
Tree age	316.093	0.000	382.49	0.000
	R^2 = 0.929		R^2 = 0.9	

Table 16. Summary results of ordinary regression model for winter and spring.

Predictor	Winter			
	GC clone		GU clone	
	Estimate	p-value	Estimate	p-value
Intercept	-12364.279	0.000	-14159	0.000
Temperature	137.832	0.000	159.339	0.000
Relative humidity	39.106	0.000	46.699	0.000
Solar radiation	1980.674	0.000	1775.888	0.021
Rainfall	-5.541	0.442	-7.936	0.046
Wind speed	659.705	0.000	698.642	0.002
Tree age	266.982	0.000	312.839	0.000
	R^2 = 0.896		R^2 = 0.841	
Predictor	Spring			
	GC clone		GU clone	
	Estimate	P-value	Estimate	P-value
Intercept	-2217.472	0.077	-8561.296	0.002
Temperature	-20.944	0.366	-40.28	0.434
Relative humidity	-0.688	0.941	-2.816	0.893
Solar radiation	56.458	0.855	110.533	0.872
Rainfall	-1.488	0.870	-1.53	0.939
Wind speed	31.297	0.788	65.365	0.801
Tree age	262.869	0.000	403.825	0.000
	R^2 = 0.282		R^2 = 0.158	

solar radiation, wind speed and tree age have significant effects on the stem radius for both clones. In autumn, rainfall appears to have significant effect on stem radius for GU clone, while it has no significant effect on GC clone. The insignificant effect rainfall in winter and autumn for GC clone might be due to genetic factor, which needs further study. Temperature has significant effect and positively related to stem radius in winter for both clones (Table 16). In summer, autumn and spring, temperature has no significant effect on stem radius (Tables 15 and 16). Therefore, the effect of weather variables on stem radius is dependent on season.

Daily stem size variation is important as the net increment of a forest stand is ultimately determined by the accumulation of daily increment events (Drew et al., 2009). Several factors might affect the daily stem size of trees. For instance, the study by Zweifel et al. (2006) indicates that there is a strong dependence of radial growth on the current tree-water relations and only secondary dependence on the carbon-balance. The availability of soil water and the degree to which storage tissues were saturated were also factors affecting the diurnal course of stem radius changes (Zweifel et al., 2001). Whitehead and Jarvis (1981) have suggested in theoretical approaches, that the diurnal stem radius fluctuations are coupled to tree-water relations by changing water potential gradients within the tree. Studies by Downs et al. (1999) and Deslauriers et al. (2003) consider the effect of weather on daily stem growth. Deslauriers et al. (2003) studied daily stem radial growth of balsam fir to show that total rainfall and maximum temperature were positively correlated with the stem radius. Climatic variables are highly inter-correlated, and the use of OLS to estimate the parameters of the response function results in instability and high variability of the regression coefficients. As a result, the regression coefficients become much larger than would seem reasonable physically or practically, and may fluctuate widely in sign and magnitude. Accordingly, it was observed that the ordinary regression estimates inflated the percentage of variation in the stem radial growth accounted for by climatic conditions. Ordinary regression inferences from such correlated climatic variables can result in misleading and confusing conclusions relating to variables of major interest to dendroecologists in terms of magnitude, sign, and standard error of the coefficients as well as R^2 (Fekedulegn et al*.,* 2002).

Both principal component regression and PLS regression methods have an advantage over OLS regression because they do not require that the explanatory variables be orthogonal. The principal components are orthogonal, eliminating the multicollinearity problem. However, the problem of choosing an optimum subset of predictors remains. A possible strategy is to keep only a few of the first components. Nevertheless, the components are chosen to explain the independent (X) rather than the dependent (Y) and there are no guarantees that the principal components which explain the independent variable can be relevant to explain the dependent (Y). On the other hand, PLS regression finds components from X that are also relevant for Y. PLS regression searches for a set of components that perform a simultaneous decomposition of X and Y with the constraint that these components explain much of the covariance between X and Y. The PLS approach is considered as a variance-based structural equation model (SEM). The alternative SEM is a covariance-based SEM. Although both methods use a latent variable term, the latent variables used by the two methods are different. As indicated by Fornell and Bookstein (1982), the latent variables in PLS are estimated as exact linear combinations of their indicators. This shows that “latent” variables in PLS are not true latent variables as defined in SEM, as they are not derived to explain the co-variation of their indicators, but only to approximate them (Mathes, 1993; McDonald, 1996). On the other hand, the latent variables in covariance-based SEMs are true latent variables. That is they are hypothetically existing entities or constructs. In other words, the covariance-based SEM latent variables cannot be found as weighted sums of manifest variables; they can only be estimated by such weighted sums (Schneewiss, 1993). Arguably, PLS has the advantage over the covariance based SEM, in that Jöreskog and Wold (1982) and Wold (1982, 1985) referred to PLS technique as “soft modeling”, because it did not require the “hard” distributional assumptions of maximum likelihood (ML) which is the core technique in SEM, and because it uses a suboptimal estimation technique that is faster to run than ML-SEM, which therefore allows for more user interaction.

Finally, the latent variable model approaches used in our study show that all climatic variables measured and tree age are positively correlated with stem radial measure for the pooled data of both clones. Moreover, all latent variables had significant effects on the radial measure. This was not the case when OLS was applied. The effects of the two most important variables, rainfall and temperature, were not significant when the OLS method was used (Table 1). This may be because the ordinary linear regression assumes that the predictors are uncorrelated, while in our case the climatic variables are correlated (Table 2). It may also be because the effect of weather variables changes with season. To overcome the problem of correlation among weather variables, two alternative methods (Principal component regression and PLS) were used. Principal component regression models were fitted for the GC and GU clones separately, resulting in a positive effect of climatic variables on stem radius for both clones. The weather data together with the age of a tree accounted for 79% of the variance in the stem radial growth for the combined data set. This is equivalent to R^2 in OLS regression. The separate analysis of GC and GU clones showed that the weather variables and tree age explained 87 and 79.6% of the total variation in radial measure for the GC and GU clones, respectively.

When comparing the PLS model fitted for the GC clone and GU clone, the effect of climatic variables is similar for the two clones except for the effect of temperature. Temperature appears to have an opposite effect on the radial growth of the two clones. Moreover, 87% of the total variation in the stem radial measure is explained by the weather variables and tree age by using the PLS method for the GC clone and 79% of the variation is explained for the GU clone. This indicates that the amount of explained variation is larger for the GC clone than for the GU clone. The evaluation of the relationship between weather variables and stem radius is considered after separating the data by season. The effect of weather variables on stem radius was found different for different seasons. Tree age is the most important factors that influences change in stem radius. The importance of tree age in determining stem radius should be expected as growth is positively related to age most of the time. There is no significant effect of weather variables on stem radius during summer and spring for both GU and GC clones. In autumn, there is significant effect of some variables (tree age, solar radiation, wind speed) for both GU and GC clones. In winter, the variables temperature, relative humidity, solar radiation, wind speed and tree age have significant positive relationship with stem radius for both clones (Table 16).

Conclusions

The study demonstrated that the relationships between the daily stem radius and weather variables is positive for both the GU and GC clones with the exception of temperature. This conclusion was drawn without considering season. The analysis by season shows that there is no relationship between weather variables (temperature, relative humidity, solar radiation, wind speed and rainfall) and stem radius for two seasons (summer and spring). In winter, there is a positive relationship between each of the variables (tree age, temperature, relative humidity, solar radiation and wind speed) and stem radius. In autumn, the relationship between stem radius and variables (solar radiation, wind speed and tree age) is positive for both clones. In autumn and winter, the effect rainfall on stem radius is significant for GU clone, while it is not significant for GC clone. This could be mainly due to genetic difference between the two clones. This may need further research in the area. The study also helps not only to see the role of climatic variables on the radial growth but also can be an example of an analysis of the effect of correlated predictors on the growth of plants in general. Regarding the statistical methods used in this study, PLS method appears to be best in solving the problem of multicollinearity. However, it is advisable to analyze the data using different alternative methods before we embark on conclusion. From this study, the lesson learnt is that the consideration of seasonal effect is indispensable, to study the effect of weather variables on tree growth.

In conclusion, the climatic variables, together with tree age, explained a substantial amount of variation (79%) in the stem radius. Tree age is the most important factor that influences change in stem radius. The importance of weather variables depends on season. In autumn, solar radiation and wind speed appears to be more important than the other weather variables. In winter, temperature, relative humidity and wind speed are more important than other weather variables in determining stem radius. This study is based on data collected at the juvenile stage of *Eucalyptus* trees. The application of the same techniques to adult trees and comparison of the results shall be the subject of future work.

ACKNOWLEDGEMENTS

The authors are grateful to Dr Valerie Grzekowiak and Dr Nicky Jones for their valuable comments and suggestions.

REFERENCES

Callaham RZ. (1962). Geographic variability in growth of forest trees. In: Kozlowski T (ed.), *Tree Growth*. New York: The Ronald Press Company. pp. 311-325.

D'Arrigo RD, Jacoby GC, Free RM (1992). Tree-ring width and maximum latewood density at the North American tree line: parameters of climate change. Canadian J. Forest. Res. 22:1290-1296.

Deslauriers A, Morin H, Urbinati C, Carrer M (2003). Daily weather response of balsam fir [*Abies balsamea* (L.) Mill.] stem radius increment from dendrometer analysis in the boreal forests of Quebec (Canada). *Trees* (Berl) 17:477–484.

Dine E, Yücesoy C, Onur F (2002). Simultaneous spectrophotometric determination of mefenamic acid and paracetamol in a pharmaceutical preparation using ratio spectra derivative spectrophotometry and chemometric methods. J. Pharm. Biomed. Anal. 2:1091–1100.

Downes G, Beadle C, Worledge D (1999). Daily stem growth patterns in irrigated Eucalyptus globules and *E.nitens* in relation to climate. *Trees* 14:102-111.

Downes G, Drew D, Battaglia M, Schulze D (2009). Measuring and modeling stem growth and wood formation: an overview. *Dendrochronologia* 27:147-157.

Drew DM (2004). Dendrometer trial phase one technical report. Report No. EFR092T. Division of Water, Environment and Forestry Technology, CSIR.

Drew D, Downes G, Grzeskowiak V and Naidoo T (2009). Differences in daily stem size variation and growth in two hybrid eucalypt clones. *Trees – Stru. Function.* 23:585-595.

Drew DM, Pammenter NW (2006). Vessel frequency, size and arrangement in two eucalypt clones growing at sites differing in water availability. *New Zealand J. Forest.* 51:23-28.

February EC, Stock WD, Bond WJ, Le Roux DJ (1995). Relationships between water availability and selected vessel characteristics in *Eucalyptus grandis* and two hybrids. *IAWA J.* 16:269-276.

Fekedulegn BD, Colbert JJ, Hicks RR, Schucker ME (2002). *Coping with multicolinerarity: an example on application of Principal Components Regression in Dendroecology. Research, Paper NE-721*. United States Department of Agriculture.

Fornell C, Bookstein F (1982). Two structural equation models: LISREL and PLS applied to Consumer Exit-Voice Theory. *J. Mark. Res.* 19:440-452.

Fritts HC (1976). *Tree rings and climate.* New York: Academic Press. pp. 28-54.

Haenlein M, Kaplan AM (2004). A beginner's guide to partial least squares analysis. Understanding Stat. 3:283-297.

Hofgaard A, Tardif J, Bergeron Y (1999). Dendroclimatic response of *Picea mariana* and *Pinus banksiana* along a latitudinal gradient in the eastern Canadian boreal forest. *Canadian J. Forest Res.* 29: 1333-1346.

Jöreskog KG, Wold D (1982). The ML and PLS techniques for modeling with latent variable historical and comparative aspects. In: Jöreskog KG, Wold H (eds.), *Systems under indirect observation: causality, structure, prediction*. Amsterdam: North Holland. pp. 263-270.

Kozlowski TT, Pallardy SG (1997). *Physiology of woody plants*. 2nd edn. San Diego: Academic Press. pp. 1-6.

Maitra S, Yan J (2008). *Principal component analysis and partial least squares: two dimension deduction techniques for regression*. Casualty Actuarial Society Discussion Paper Program. pp. 79-90.

Mathes H (1993). Global optimization criteria of the PLS algorithm in recursive path models with latent variables. In: Haagen K, Bartholomew DJ, Deistler M. (eds.), *Statistical modeling and latent variables*. Amsterdam: Elsevier. Pp. 229-248.

McDonald RP (1996). Path analysis with composite variables. *Multivariate Behav. Res.* 31:239-270.

Mevik BH, Cederkvist HR (2004). Mean Squared Error of Prediction (MSEP) estimates for Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). J.Chemom.18:422-429.

Phipps RL (1982). Comments on interpretation of climatic information from tree rings, eastern North America. Tree Ring Bull. 42:11-22.

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rodriguez-Nogales JM (2006). Approach to the quantification of milk mixtures by partial least-squares, principal component and multiple linear regression techniques. *Food Chem.* 98:782-789.

Schneewiss H (1993). Consistency at large in models with latent variables. In: Haagen K, Bartholomew DJ, Deistler M. (eds.), *Statistical modeling and latent variables*. Amsterdam: Elsevier. pp. 299-320.

Schweingruber FH, Briffa KR, Nogler P (1993). A tree-ring densitometric transect from Alaska to Labrador. Int. J. Biometeorol. 37:151-169.

Searson MJ, Thomas DS, Montagu KD, Conroy JP (2004). Wood density and anatomy of water limited eucalypts. Tree Physiol. 24:1295-1302.

Turnbull JW (1999). Eucalyptus plantations. *New Forests* 17:37-52.

Whitehead D, Jarvis PG. (1981). Coniferous forests and plantations. In: Kozlowski TT (ed.), *Water deficits and plant growth*. New York: Academic Press. pp. 50-153.

Wold H (1982) . Soft modeling: the basic design and some extensions. In: Jöreskog KG, Wold H. (eds.), *Systems under indirect observations : causality , structure, prediction.* Amsterdam: North Holland. pp. 1-54.

Wold H (1985). Systems analysis by partial least squares. In: Nijkamp P, Leitner H, Wrigley N(eds.), *Measuring the unmeasurable*. Boston : Martinus Nijhoff. pp. 221-251.

Zweifel R, Ha¨sler R (2001). Dynamics of water storage in mature subalpine *Picea abies*: temporal and spatial patterns of change in stem radius. Tree Physiol. 21:561–569.

Zweifel R, Zimmerman L, Zeugin F, Newbery DM.(2006). Intra-annual radial growth and water relations of trees: implication towards a growth mechanism. J. Exp. Bot. 57:1445-1459.

academicJournals

Vol. 8(22), pp. 2685-2695, 13 June, 2013
DOI: 10.5897/AJAR2013.6920
ISSN 1991-637X ©2013 Academic Journals
http://www.academicjournals.org/AJAR

African Journal of Agricultural Research

Full Length Research Paper

Path models-approach to the study of the effect of climatic factors and tree age on radial growth of juvenile *Eucalyptus* hybrid clones

Sileshi F. Melesse* **and Temesgen Zewotir**

School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, South Africa.

Accepted 30 May, 2013

Due to increasing wood consumption and pulp and paper demands, plantations of fast growing tree species, have a growing importance for the sustainability of industrial wood raw material. Consequently, the efficient utilization of fast growing plantations can have a large impact on productivity. Adequate management requires good understanding of factors affecting tree growth. This study aimed to determine the factors that influence stem radial growth of juvenile *Eucalyptus* hybrids grown in the east coast of South Africa. Measurement of stem radius was conducted using dendrometers on sampled trees of two *Eucalyptus* hybrid clones (*Eucalyptus grandis* × *Eucalyptus urophylla*, GU and *E. grandis* × *Eucalyptus camaldulensis*, GC). Daily averages of climatic data (temperature, solar radiation, relative humidity and wind speed) were simultaneously collected with total rainfall from the site. In this study, path analysis was employed. The joint effect of the climatic variables as well as the direct effect of each climatic variable was studied. Bootstrap estimation procedures, which relax the distributional assumption of the maximum likelihood estimation method, were used. It is found that all variables had a positive effect on stem radial growth. The study showed that tree age is the most important determinant of radial measure.

Key words: Bootstrap, cross-validation, dendrometer, maximum likelihood, path analysis, standardized regression weights.

INTRODUCTION

Eucalyptus has increasingly become the most widely planted, hardwood genus in the world (Turnbull, 1999). Eucalypts provide sawn timber, mine props, paper, pulp, fiberboard, poles, firewood, charcoal, essential oils, honey and tannin products. Eucalypt plantation growth rate is an important economic factor as fast growing trees will be available for processing earlier compared to slower growing trees. Tree growth and the ultimate production of wood is a product of the interaction of genetic (Kozlowski and Pallardy, 1997; Apiolaza et al., 2005; Zweifel et al., 2006), silvicultural (Pallett and Sale, 2004) and environmental factors (Gallaham, 1962; February et al., 1995; Searson et al., 2004; Drew and Pammenter, 2006).

Climatic factors such as temperature, humidity, sunlight, rainfall (Eagleman, 1985; Miller, 2001) and wind speed (Wadsworth, 1959) contribute to the growth of plants. Growth generally occurs under a broad range of climatic variables, but ideal growth occurs during optimum climatic conditions. The net contribution of each climatic variable is, however, often masked or influenced by one or more other climatic variables. Understanding

*Corresponding author. E-mail: Melesse@Ukzn.ac.za.

the relationships between climatic variables and the pattern of stem growth would facilitate the prediction of wood properties for a given site. However, such studies are limited. Available studies commonly focus on growth rate and pattern of growth as a function of age (Miehle et al., 2009; Crecente-Campo et al., 2010; Mateus and Tomé, 2011). Downes et al. (1999) studied the effects of climatic variation on radial growth of irrigated eucalypts in Australia. The work of Downes et al. (1999) focused on daily stem growth patterns in irrigated *Eucalyptus globulus* and *E. nitens* in relation to climate. Applying multiple regressions, they have shown that weather variables accounted for 40 to 50% of the variance in stem radial increment. Downes et al. (2009) gave an excellent overview on measuring and modeling stem growth and wood formation. Since most eucalypt plantations rely on natural conditions for growth (no irrigation), assessments of the effects of the natural environment is useful to begin to understand what the potential impact of drought or even climate change may have, not only on growth, but potentially also on wood properties. Drew et al. (2009) studied the relationship between stem radius and climatic factors using the correlation matrix. The methods used by both Downes et al. (1999) and Drew et al. (2009) do not permit any other relationships among the independent variables to be specified. This limits the potential of the variables to have direct, indirect and total effects on each other. The path models approach used in this study can overcome these limitations. This paper describes the effects of tree age and climatic variation on radial growth of *Eucalyptus grandis* × *E. urophylla* (GU) and *E. grandis* × *E. camaldulensis* (GC) hybrid clones established in Zululand on the eastern coast of South Africa. The particular emphasis of this paper is on determining the climatic factors that most influence radial growth of *Eucalyptus* hybrid clones during the juvenile stages of growth.

MATERIALS AND METHODS

Study design

A dendrometer trial, which focused on the growth of two *Eucalyptus* hybrid clones was established on Sappi landholdings at KwaMbonambi (28.53° S, 32.140 E, 55 m MASL) on the Zululand coast in the eastern part of South Africa. On average, the site receives 1,000 mm of rainfall per annum and has a mean annual temperature of 21°C (Drew et al., 2009). The experiment was designed to extend over a seven-year period divided into separate phases of growth. Each phase ended with the destructive sampling of study trees to facilitate measurement of wood anatomical characteristics. The results presented in this study are based on the data collected only during the first of these phases of growth. This phase ran for 16 months from April 2002 until August 2003. Two *Eucalyptus* hybrid clones, *E. grandis* × *E. urophylla* (GU) and *E. grandis* × *E. camaldulensis* (GC), which were commercially deployed at the time, were established in the trial (Drew, 2004).

Planting took place on 16 July 2001. Prior to planting, in April 2001, stumps of the trees from the previous rotation on the site were treated with herbicide (to prevent coppicing) and slash from harvest was burnt. Each rooted cutting was planted in a planting pit between existing stumps, with approximately two liters of water. The two clones were planted in alternating blocks (three repeats) of 7 × 24 trees at a spacing of 3 m (E-W) × 2.5 m (N-S). Within each block of a particular clone, three plots of 12 (3×4) trees, each with two surrounding rows of trees were identified. The plots were established as pairs, such that for any phase of the research, a GU and a GC plot could be measured simultaneously. Within a 12 tree plot, nine trees were selected from each clone for intensive monitoring of radial growth and other physiological characteristics during Phase 1 (Drew, 2004). Radial growth (μm) was measured using 18 electric point dendrometers (AEC) mounted on nine trees per clone in adjacent plots. One dendrometer was mounted on the north side of each sampled tree, at breast height (1.3 m), from when trees were nine-months-old. In addition to radial growth, an automatic weather station was installed at a distance of approximately 200 m from the trial to record hourly temperature (°C), relative humidity (%), solar radiation (mJ/h), rainfall (mm) and wind speed (m/s). Later on the daily total rainfall and the daily average of other variables were obtained from the hourly data. The data set used in this study has a total of 8,640 observations for the two clones which is the daily data. Half the data set pertains to the GU clone and the remaining half to the GC clone.

Data analysis

The statistical method employed to analyze the data is path analysis. A brief description of path analysis and its relation to the classical regression model is given. Path analysis is the statistical technique used to examine causal relationships between two or more variables. It involves a set of simultaneous regression equations that theoretically establish the relationship among observed variables in the path model. Path analysis extends the idea of regression modeling and gives the flexibility of quantifying indirect and total causal effects in addition to the direct effect which is also possible in regression analysis (Bollen, 1989). In other words, regression analysis allows an independent variable to influence an outcome variable only directly. Path analysis however gives more flexibility and predictor variables are allowed to influence the outcome variable directly as well as indirectly through other mediating variables. Path analysis shares the following principles of regression analysis:

1. The direction of influence in the relationship of variables should be specified from the theory behind the investigation;
2. Independent variables are assumed to be measured without error.
3. The relationship between target variables is linear.
4. Any outcome variable in the system of equations under investigation has an error term attached to it.

Path analysis is an extension of the regression model, which researchers use to test the fit of a correlation matrix with a causal model that has been, tested (Garson, 2004). The aim of path analysis is to provide estimates of the magnitude and significance of the hypothesized causal connections among sets of variables displayed through the use of path diagrams. There are three interrelated components in path analysis (Bollen, 1989):

1. The translation of a conceptual problem into pictorial presentation, which shows the network of relationships;
2. Obtaining systems of equations that relate observed correlation and covariance to parameters; and
3. Decomposition of effects of one variable on another (that is, direct, indirect and total effects) from the correlation of measured variables.

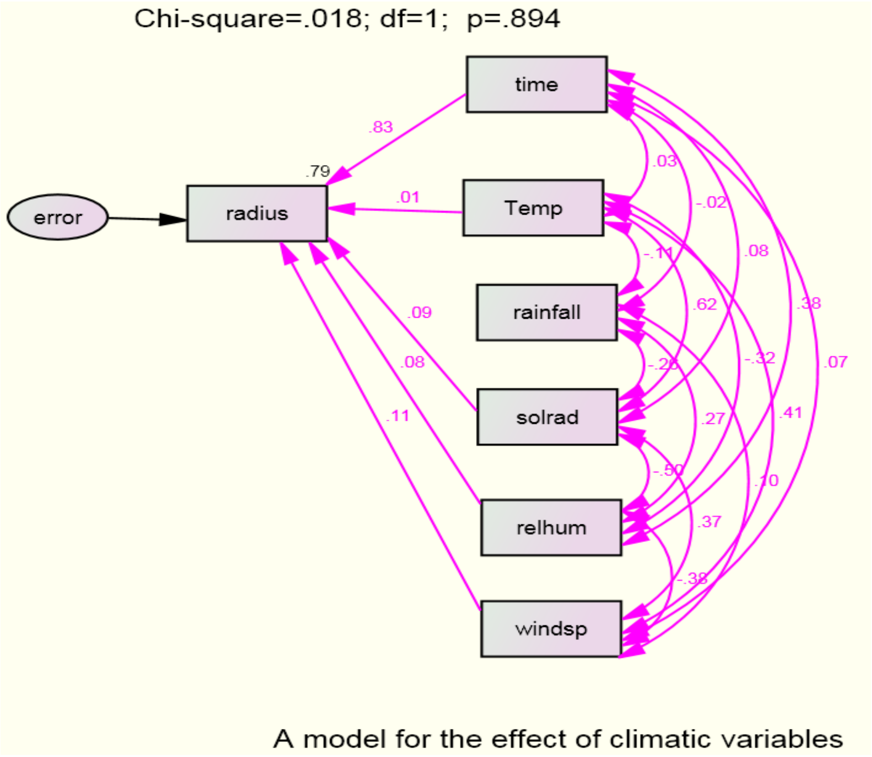

Figure 1. Path diagram showing the effect of age and climatic variables on radius of *Eucalyptus* hybrid clones during the first measured phase of growth. Time = age; solrad = solar radiation; relhum = relative humidity; windsp = wind speed.

The statistical analyses were performed using AMOS software (Amos Development Corporation). Path analysis was conducted by considering the radial measure as dependent climatic variables and age as independent factors explaining the radial growth. The chi-square statistic, the normed fit index (NFI), and root mean square error of approximation (RMSEA) were used to estimate model fit. The larger the probability associated with the chi-square, the better the fit of the model to the data (Bollen, 1989; Byrne, 2001). The NFI tests the hypothesized model against a reasonable baseline model and ideally should be 1·0. A RMSEA of < 0·10 is considered a good fit and < 0·05 is very good and lower than 0.01 is considered as beautiful fit (Steiger, 1990). Model validity was assessed using the expected cross validation index (ECVI). Path significance was based on the critical ratio (CR), with a CR > 2 in absolute value considered as significant (Arbuckle, 2006; Schumacker and Lomax, 2004).

RESULTS AND DISCUSSION

The independent variables included in the study were the five major climatic variables that were measured and the age of the trees. The association between the independent variables and the radial growth measurement of the clones is presented in Figure 1. The numbers displayed at the top of the diagram refer to the goodness of fit of the model. This fit statistic is the likelihood ratio chi-square test. The p-value associated with this measure is 0.894, which is by far larger than 0.05 and indicates a non-statistical significance of the chi-square test. This implies the model is consistent with the data. The numbers displayed next to the double headed arrows are estimated correlation coefficients.

Various measures of fit (Table 1) are presented for the fitted model, given in Figure 1, and include the saturated model, which is the ideal fit by including all possible paths. A model that can be defined as good is one that does not differ significantly from the saturated model despite omitting paths from the saturated model. On the other hand, the ordinary regression model or independent model fits by ignoring any potential relatedness between the independent variables thus considering all correlations among the independent variables as zero.

The statistical significance of individual parameter estimates for the paths in the fitted model (Figure 1) is one of the important criteria to be studied. The significance can be seen by computing the critical values, which are obtained by dividing the parameter estimates by their respective standard errors. The computed critical values together with the corresponding p-values are presented in Table 2. The regression weights for all variables were significant with the exception of rainfall, which was dropped from the model.

The other issue to consider at this stage is the magnitude and direction of the parameter estimates. In this particular model all the regression weights were

Table 1. Different fit measures for the fitted model, saturated and ordinary regression models.

Fit measure	Model		
	Fitted model[1]	Saturated model[2]	Ordinary regression[3]
Chi square	0.02		1287.06
Chi square p-value	0.89		0
Normed fit index (NFI)	1	1	0
Root mean square error of approximation (RMSEA)	0		0.386
Expected cross-validation index (ECVI)	0.006	0.006	3.13
ECVI lower bound	0.006	0.006	3.068
ECVI upper bound	0.007	0.006	3.193
Modified expected cross validation index (MECVI)	0.006	0.006	3.131

[1]The model presented in Figure 1. [2]Model that includes all possible paths. [3]The independent model that assumes no correlation between the independent variables.

Table 2. Regression weights indicating the relationship between radial growth and each independent variable for the combined data set (Maximum Likelihood Estimates).

Relationship	Maximumlikelihood estimates	Standard error	Critical ratio	P-value
Radius<---time	313.51	2.18	143.91	***
Radius<---temperature	23.74	12.64	1.88	0.06
Radius<---solar radiation	2817.03	220.03	12.80	***
Radius<--- relative humidity	63.76	5.75	11.09	***
Radius<---wind speed	1447.03	73.63	19.65	***

*** the p-value is less than 0.001.

positive indicating the existence of a positive relationship between radial growth and the climatic variables. The standardized regression coefficients are 0.832 (age of a tree), 0.012 (temperature), 0.092 (solar radiation), 0.076 (relative humidity) and 0.113 (wind speed). This suggests that the most important variable to explain radial growth is age of the tree. It is also estimated that the predictors of radius explain 79% of its variance. In other words, the error variance of radius is approximately 20.9% of the variance of radius itself.

Although the goodness of fit measures indicate that the fitted model (Figure 1) is a good fit (Table 1), the parameter estimates show that rainfall has no direct influence on the radial growth. An attempt was made to modify the fitted model (Figure 1) by making rainfall a required variable in the model. Such a modification procedure is called specification search (Leamer, 1978). The objective of specification search is to alter the original model in search of a model that is better fitting in some sense, and yields parameters having practical, and in this case biological significance, and substantive meaning. The path diagram for the first attempt at modification is presented in Figure 2. For this path analysis model, a good ‘goodness of fit’ was obtained. The calculated value of the chi-square statistics was 3.194 with one degree of freedom and a p-value of 0.074. However, the goodness of fit for the second fitted model (Figure 2) was not as good as the model fit shown in Figure 1. The parameter estimates for the second fitted model (Figure 2) suggest that rainfall had no direct significant effect. Therefore, no additional information was gained by modifying the path diagrams from that of Figure 1 to that of Figure 2.

The third attempt at specification search was to consider a model fit for the second fitted model (Figure 2) that excluded wind speed as an explanatory variable (Figure 3). The model fit was good and parameter estimates were significant. The regression weight for rainfall in the prediction of radial growth was significantly different from zero at the 0.001 level (two-tailed, Figure 3). This indicates that rainfall has a significant effect on the radial growth of trees in the absence of wind speed. For this model, it is estimated that the predictors of radial growth explain 78.2% of its variance. This is very close to the value obtained for the first model (Figure 1), which includes all the predictors in the model. The standardized regression coefficients were 0.859 (age of a tree), 0.042 (temperature), 0.096 (solar radiation), 0.026 (relative humidity) and 0.03 (rainfall). These standard regression coefficients indicate that age of the tree is the most important variable in determining the stem radial growth. Models fitted without temperature or tree age as

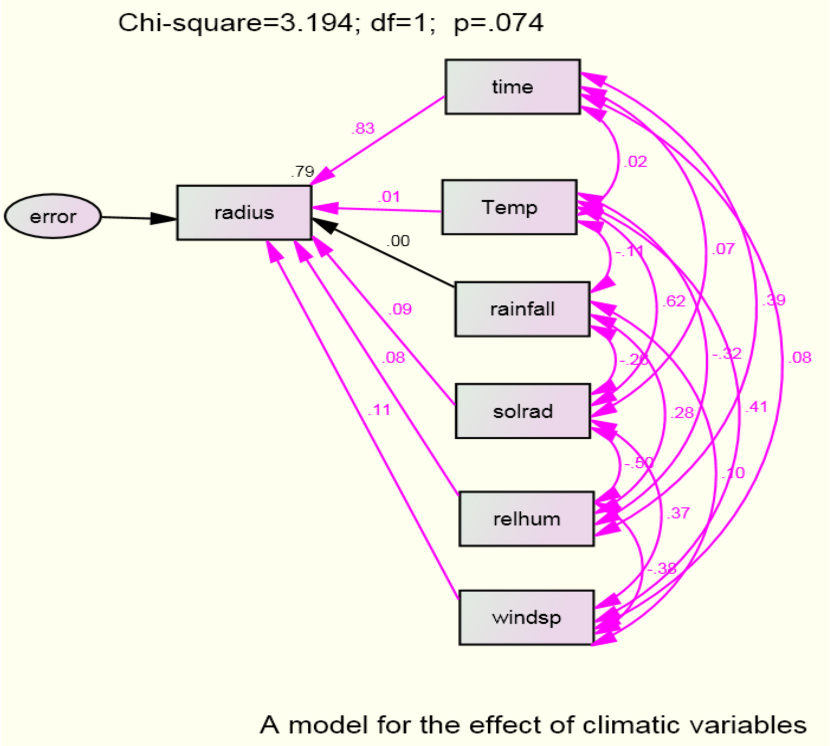

Figure 2. Path diagram showing the effect of age and climatic variables on radius of *Eucalyptus* clones when rainfall is considered a required variable. Time = age; solrad = solar radiation; relhum = relative humidity; windsp = wind speed.

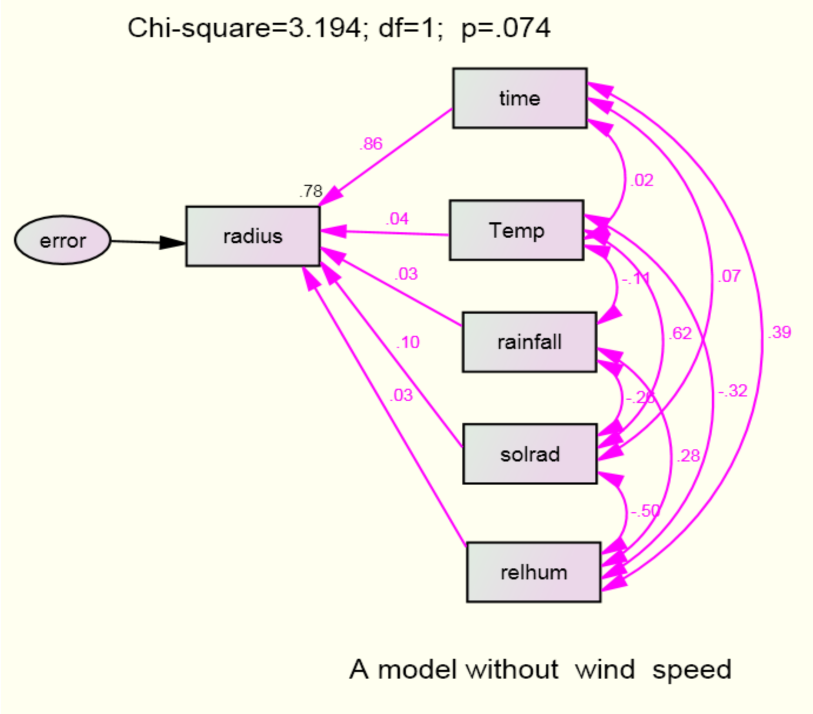

Figure 3. Path diagram showing the effect of age and climatic variables on radius of *Eucalyptus* clones when wind speed is omitted as an explanatory variable. Time = age; solrad = solar radiation; relhum = relative humidity.

explanatory variables did not fit well. A model that excluded relative humidity fitted well and resulted in rainfall having a significant effect on radial growth. The significance of rainfall in the absence of relative humidity and solar radiation was possibly caused by multicollinearity (where two or more predictor variables in a multiple regression model are highly correlated). The correlation among the climatic variables themselves is also significant. When only rainfall and wind speed were considered independent variables, the regression weight for rainfall became negative. The same occurred when only rainfall and relative humidity were treated as independent variables. This wrong sign of coefficients is an indication of possible multicollinearity. As a result, the effect of rainfall on radial growth cannot be completely ruled out, as its non-significance is possibly caused by multicollinearity. Some researchers noted that structural equation models are robust against multicollinearity (Malhotra et al., 1999), with some going as far as to explicitly state that Structural Equation Models (SEM) can remedy multicollinearity problems. For example, Maruyama (1998) argues that "structural equation approaches can help deal with some cases where the correlations among the predictors are large". On the other hand, some researchers have warned that multicollinearity can lead to SEM estimates being far from the true parameters, as well as the occurrence of large standard errors of the estimates (Jagpal, 1982; Grapentine, 2000). A simulation study by Grewal et al. (2004) showed some conditions under which multicollinearity caused problems. The study showed that when multicollinerity is extreme, type II error rate (accepting the null hypothesis when it is false) is generally, unacceptably high. They also indicated that for multicollinearity levels of between 0.6 and 0.8, type II error rates can be substantial (greater than 50% and frequently above 80%), if composite reliability is weak, explained variance (R^2) is low and sample size is relatively small. When multicollinearity levels are between 0.4 and 0.5, type II error rates tend to be quite small except when reliability is weak, R^2 is low and the sample size is small. In the present study R^2 values were large and the multicollinearity level was not high.

Estimates of regression weights for rainfall, which is important for growth, were inconsistent. Consideration of more complex models may improve results. In the path diagrams considered thus far only one dependent variable (radial growth) was used. Path analysis allows the simultaneous modeling of several related regression relationships. This means that path analysis can handle more than one independent variable in the model. Moreover, a variable can be a dependent variable in one relationship and an independent variable in another relationship of the path model. An attempt was made to fit a model where two dependent variables, namely rainfall and temperature, mediated the effects of relative humidity, solar radiation and wind speed. In this model, it was hypothesized that the age of a tree had a direct effect on radial growth. Solar radiation, relative humidity and wind speed were assumed to have an indirect effect. The fitted model is presented in Figure 4.

The value of the chi-square statistic is 862.7 with a p-value of zero. This indicates that the model does not fit the data well. However, the parameter estimates of the regression weights are all significant (Table 5). The magnitude of each effect is quantified by standardized regression coefficients. The standardized regression coefficients are 0.87 (age of the tree), 0.091 (temperature), and 0.018 (rainfall). From this it can be seen that the most important variable to explain radial growth is tree age. For the model in Figure 4 there are three structural equations, one for each of the three dependent variables: rainfall; temperature and radius. In terms of variable names, the structural equations are:

$$
\begin{aligned}
&ra\inf\ all = relative\ \ humudity\ \ + solar\ \ radiation\ \ + wind\ \ speed\ \ + error\ 1\\
&temperatur\ \ e = relative\ \ humudity\ \ + solar\ \ radiation\ \ + wind\ \ speed\ \ + error\ 2\\
&radius\ \ = ra\inf\ all\ + temperatur\ \ e + time\ \ + error\ \ 3
\end{aligned}
$$

This model includes direct effects (e.g. age of the tree on radial growth), indirect effects (e.g. effect of relative humidity through rainfall) and correlated independent variables (e.g. relative humidity, solar radiation and wind speed). The estimated model using AMOS statistical software is given by:

$$
\begin{aligned}
&ra\inf all = 0.196\ relative\ humudity\ - 6.27 solar\ radiation + 3.22 wind\ speed\\
&temperature = 0.017 relative\ humudity\ + 8.77\ solar\ radiation + 1.39 wind\ speed\\
&radius\ = 20.73 ra\inf all\ + 178.37 temperature + 329.67\ time
\end{aligned}
$$

From the fitted model (Figure 4) the positive effect of the predictors, rainfall, temperature and tree age can be seen. The standardized regression weights for this model indicate that tree age, temperature and rainfall are respectively important determinants of radial growth.

The data set to which the above models were applied was a combined data set (for both *E. grandis* hybrid clones). In order to see if there was any difference between the two clones, a multiple group analysis was used. In this regard, the good fitting model produced in Figure 1 and the model with multiple dependent variables (Figure 4) was considered. The good fitting model of Figure 1 was fitted to the data set for GU clone, alone. The model fitted the data very well. The value of the chi-square statistics was 0.06 with one degree of freedom and the corresponding p-value was 0.804. The next question to address was whether the same model fitted the data for the GC clone. Furthermore, the equality of the parameters needed to be tested. Instead of a separate group analysis, a single analysis that simultaneously estimated parameters and tested hypotheses about both groups was considered. This method provided a test for the significance of any differences found between the GU and GC clones. In addition, if there were no differences between the two

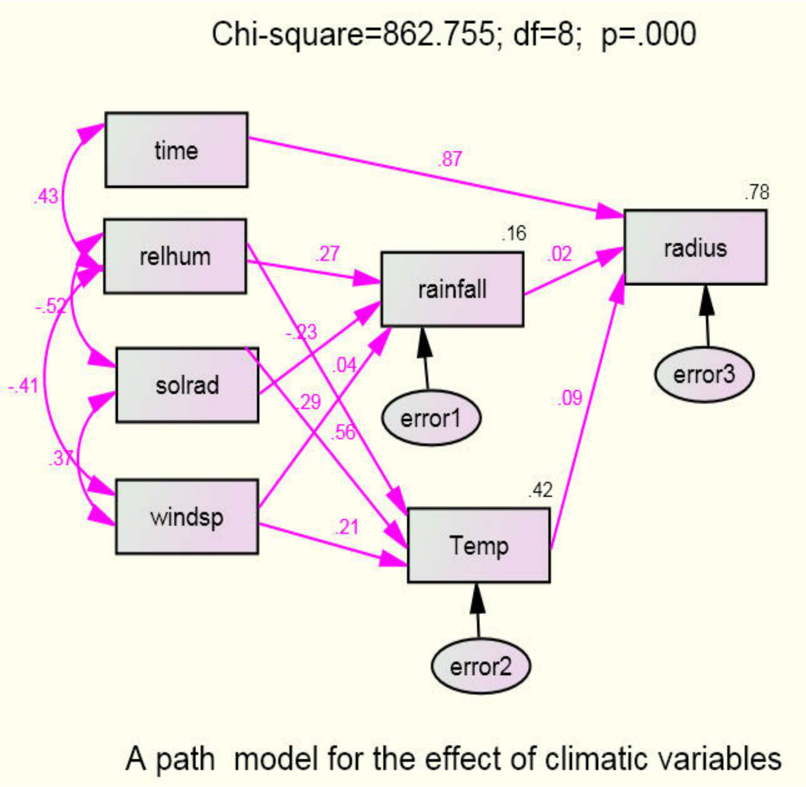

Figure 4. Path diagram showing the effect of multiple dependent variables (rainfall and temperature) on radial growth of *Eucalyptus* clones. Time = age; solrad = solar radiation; relhum = relative humidity.

clones, or if group differences concerned only a few model parameters, the simultaneous analysis of both groups would have provided more accurate parameter estimates than would have been obtained from separate single-group analyses. A test for pair wise path coefficient differences for the two clones was conducted. Some fit measures for various models were generated, together with fit measures for saturated and independence models are shown in Table 3.

The structural weight model specifies that the regression weights for predicting radial growth from the measured climatic variables and the age of tree were the same for the GU and GC clones. The unconstrained model is the model that assumes that all the parameters for the two groups are different. For the unconstrained model, the value of chi-square was 0.08 with the corresponding p-value equal to 0.96. This indicated that the unconstrained model fitted the data very well. The structural weight model with a chi-square value of 364.59 and with seven degrees of freedom was rejected at any conventional significance level, suggesting that the regression weights of the two clones were significantly different. The assumption that the regression weights for the exogenous variables were the same for both clones was not supported. The estimated regression weights for the unconstrained model are summarized in Table 4 and Table 5. When comparing the regression weights for the two clones, these were all positive, indicating a positive effect of the climatic variables as well as tree age on radial growth. In addition, regression weights obtained for the GU clone were larger than those obtained for the GC clone, indicating that the GU clone grows faster than the GC clone. Regression weights of the GU and the GC clones, for the multiple dependent model in Figure 4, were also compared. The regression weights for the two clones were significantly different. The results of this model also show that the GU has faster growth than the GC clone.

The maximum likelihood estimates given in Tables 4 and 5 require the data to be of a continuous scale and have a multivariate normal distribution. The approximate standard errors used in the inference were therefore produced based on formulae that depend on normality assumptions. Non-normality can lead to spuriously low standard errors, with degrees of underestimation ranging from moderate to severe. The consequences are that, because the standard errors are underestimated, the regression paths and factors / error covariances will be statistically significant, although they may not be so in the population (Byrne, 2001).

It is known that many data do not qualify for multivariate normality and the current data is no

Table 3. Summary of fits for various models including the structural weight model.

Model	Number of parameters	Chi-square	df	P-value	Chi-square / df
Unconstrained	54	0.08	2	0.96	0.04
Structural weights	49	364.59	7	0.00	52.09
Structural covariance s	28	364.59	28	0.00	13.02
Structural residuals	27	1293.58	29	0.00	44.61
Saturated model	56	0.00	0		
Independent model	14	29255.12	42	0.00	696.55

df = Degrees of freedom.

Table 4. Regression weights for the GU clone when the path model in Figure 1 was fitted to compare the two clones (Unconstrained).

Relationship	Maximumlikelihood estimates	Standard error	Critical ratio	P-value	Label
Radius<---time	341.88	3.33	102.81	***	b1_1
Radius<---temperature	43.34	19.30	2.25	0.025	b2_1
Radius<---solar radiation	3253.04	335.85	9.69	***	b3_1
Radius<---relative humidity	75.14	8.77	8.57	***	b4_1
Radius<--- wind speed	1570.35	112.39	13.97	***	b5_1

***indicates the p-value is less than 0.001.

Table 5. Regression weights for the GC clone when the path model in Figure 1 was fitted to compare the two clones (Unconstrained).

Relationship	Maximumlikelihood estimates	Standard error	Critical ratio	P-value	Label
Radius<---time	285.14	2.075	137.436	***	b1_2
Radius<---temperature	4.13	12.040	.343	0.732	b2_2
Radius<---solar radiation	2381.02	209.543	11.363	***	b3_2
Radius<---relative humidity	52.39	5.472	9.575	***	b4_2
Radius <---wind speed	1323.72	70.119	18.878	***	b5_2

*** indicates the p-value is less than 0.001.

exception. Using AMOS statistical software the data was checked to see whether it had a multivariate normal distribution. The Mardia's (1970) coefficient of multivariate kurtosis was 57.31 with a critical ratio of 237.3, which highly favours multivariate non-normality of the data.

A possible approach to overcome the problem of the existence of multivariate non-normal data is to use a method known as "bootstrap" (West et al., 1995; Yung and Bentler, 1996). This technique enables us to create multiple subsamples from an original data base. The importance of drawing these multiple samples is that we can examine parameter distributions relative to each of these newly produced samples. These distributions serve as a bootstrap sampling distribution and technically operate in the same way as the sampling distribution generally associated with parametric inferential statistics. In contrast to traditional statistical methods, however, the bootstrap sampling distribution is concrete and allows for comparison of parametric values over repeated samples that have been drawn (with replacement) from the original sample. The bootstrap method is free from the distributional assumptions and can be used to generate an approximate standard error for many statistics without having to satisfy the assumption of multivariate normality. With this beneficial feature in mind, the bootstrap method was applied to the good fitting model in Figure 1. In this process, 10,000 bootstrap samples were generated. The reported value of the chi-square was 0.018 with one degree of freedom. The bootstrap standard errors for regression weights are presented in Table 6. The table lists the bootstrap estimate of the standard error for each independent variable in the model. Each value represents the standard deviation of the parameter estimates computed across the 10,000 bootstrap samples. These values were compared with the values of the approximate maximum likelihood estimates presented in Table 2. Some discrepancies between the two sets of standard error estimates were observed. The third column of Table 6, labeled SE-SE provides the approximate standard

Table 6. Bootstrap standard errors for the path model in Figure 1.

Parameter (un-standardized)	SE	SE-SE	Mean	Bias	SE-Bias
Radius<---time	2.35	0.017	313.52	0.010	0.024
Radius<---temperature	12.55	0.089	23.85	0.11	0.125
Radius<---solar radiation	220.36	1.56	2816.58	-0.451	2.204
Radius<---relative humidity	5.89	0.042	63.75	-0.018	0.059
Radius<---wind speed	69.65	0.493	1446.07	-0.967	0.697
Standardized parameter					
Radius<---time	0.004	0.000	0.832	0.000	0.000
Radius<---temperature	0.006	0.000	0.012	0.000	0.000
Radius<---solar radiation	0.007	0.000	0.092	0.000	0.000
Radius<---relative humidity	0.007	0.000	0.076	0.000	0.000
Radius<---wind speed	0.006	0.000	0.113	0.000	0.000

Table 7. Ninety-five percent bootstrapped confidence intervals (bias-corrected percentile method).

Regression weights	Estimate	Lower	Upper	P
Radius<---time	313.51	308.86	318.03	0.000
Radius<---temperature	23.74	-1.21	48.76	0.060
Radius<---solar radiation	2817.03	2392.34	3252.47	0.000
Radius<---relative humidity	63.76	52.27	75.19	0.000
Radius<---wind speed	1447.03	1314.33	1588.51	0.000
Standardized regression weights				
Radius<---time	0.832	0.824	0.841	0.000
Radius<---temperature	0.012	-0.001	0.025	0.059
Radius<---solar radiation	0.092	0.078	0.106	0.000
Radius<---relative humidity	0.076	0.063	0.090	0.000
Radius<---wind speed	0.113	0.103	0.124	0.000

error of the bootstrap standard error itself. These values were very small indicating that the standard errors were estimated with a reasonable level of accuracy.

Column four, labeled mean, lists the mean parameter estimates computed across the 10,000 bootstrap samples. Arbuckle (2006) on page 301 emphasized that this bootstrap mean is not necessarily identical to the original estimate. Column five (Bias) represents the differences between the bootstrap mean estimates and the original estimates. These values are very small for most of the cases and positive values indicate that the estimates of the bootstrap samples are higher than the original maximum likelihood estimates. The low bias indicates that the maximum likelihood estimates and the bootstrap estimates are very close to each other. The last column, labeled SE-Bias, reports the approximate standard error of the bias estimate. For the majority of the cases the estimated bias is smaller in magnitude than its standard error. This indicates that there is little evidence that the regression weights are biased.

The bootstrap confidence intervals are presented in Table 7. The bias-corrected confidence intervals are used because these intervals are considered to yield more accurate values than percentile confidence intervals (Efron and Tibshirani, 1993). The confidence intervals for tree age, solar radiation, relative humidity and wind speed do not include zero. It can therefore be concluded that the regression weights of these dependent variables are significantly different from zero. The value of p in the 'p' column of Table 7 indicates that a 100(1-p)% confidence interval would have one of its end points at zero. In this sense, the p-value can be used to test the hypothesis that an estimate has a population value of zero. In this case the relationship between radius and temperature has a p-value 0.06, which means that a 94% confidence interval would have a lower boundary at zero. In other words, a confidence interval at any level less than 94% such as 90% or 92% would not include zero, and therefore reject the hypothesis that the regression weight is zero for a 90% confidence interval. For the relationship of radius with other independent variables the hypothesis at any conventional significance level such as 95 or 99% is rejected. Therefore, by applying the bootstrap method, it can be seen that the dependent variables had a significant effect on the radial growth of *Eucalyptus* trees. This result also agreed with the result obtained using the maximum likelihood method. It is also of interest to evaluate the appropriateness of the hypothesized model itself. Bollen and Stine (1993) provided a means of testing the null hypothesis that the specified model was correct. The Bollen-Stine bootstrap corrected p-value was 0.878. This corrected p-value indicates that the hypothesized model should not be rejected. This result is also in agreement with the maximum likelihood results. The other issue with the specified model was cross validation. To assess the validity of the model in Figure 1, expected cross validation index (ECVI) was applied. ECVI is proposed as a means to assess, in a single sample, the likelihood that the model cross-validates across similar size samples from the same population (Browne and Cudeck, 1989). It measures the discrepancy between the fitted covariance matrix in the analyzed sample, and the expected covariance matrix that would be obtained in another sample of equivalent size. Application of ECVI assumes a comparison of models, whereby ECVI index is computed for each model and then all ECVI values are placed in rank order. The model having the smallest ECVI value exhibits the greatest potential for replication. There is no determined appropriate range of values for ECVI as it can assume any value (Byrne, 2001). In the present case the values of ECVI are presented in Table 1. In assessing the hypothesized model, its ECVI value of 0.006 was compared with that of the independence model (ECVI=3.13). The ECVI for the saturated model was also 0.006. The ECVI for the hypothesized model was less than that of the independence model. It can therefore be concluded that the hypothesized model represents the best fit to the data. Furthermore, a 95% confidence interval for ECVI is given by [0.006, 0.007]. This indicates that of the overall possible randomly sampled ECVI values, 95% of them will fall [0.006, 0.007], suggesting that the model cross validates over the independent model.

Conclusions

Classical methods, like ordinary regression models once the regression model is specified, do not permit any other relationships among the independent variables to be specified. This limits the potential of the variables to have direct, indirect and total effects on each other. In path analysis one can see the direct effect, indirect effect and total effects of variables. In path analysis a unique additional contribution of each variable can be studied using the standardized regression weights. Even though we can study the additional contribution of each variable in multiple regressions, this can work ideally only if all independent variables are highly correlated with the dependent variable and uncorrelated among themselves. In contrast, path models provide theoretically meaningful relationships in a manner not restricted to a multiple regression model (Schumacker, 1991). In path analysis, we can estimate parameters for more than one regression equation because this analysis can be considered as a series of regressions applied sequentially to the data. Structural Equation Models (SEM) are considered as path analysis involving latent variables. In the present case, latent variables were not included and hence path models were generated. Path analysis was employed mainly because the climatic variables were correlated and the unique, additional contribution of each climatic variable on radial growth of eucalypts was of interest.

The best fitting path model generated in this study showed that all climatic variables and age of the tree had a positive effect on stem radial growth for the pooled data of both clones. Furthermore, all except one variable (rainfall) had a significant, direct effect on radial growth. It was also observed that the age of the tree was the most important variable explaining stem radial growth. Although rainfall was not significant in the best fitting model, it was found to be significant for the model that excluded wind speed and for the model that omitted solar radiation. This revealed that the effect of rainfall on radial growth cannot be ruled out. To compare the effect of the explanatory variables on the radial growth of the GU and GC clones, a single analysis that estimated parameters and tested hypotheses about both groups simultaneously was considered. The regression weights for the two clones were significantly different. The regression weights were all positive indicating the positive effect of the climatic variables as well tree age. In addition, the regression weights obtained for the GU clone were larger than the regression weights for the GC clone. This shows that the GU clone was growing faster than the GC clone which can easily be confirmed by looking at the growth of the two clones.

The main estimation method for path models, or any structural equation model (SEM) is maximum likelihood estimation. This method requires a distributional assumption, which the present data failed to satisfy. The bootstrap method was then applied to overcome the methodological failure due to non-normality. The estimated bias using the bootstrap method was very small showing that there was little evidence of bias in the estimates. The conclusion reached using the maximum likelihood method agreed with that of the bootstrap method. The expected cross-validation index obtained for the hypothesized model also showed that this model cross-validated over the independent model.

To sum up, the climatic variables measured in this study, together with tree age, had a positive effect on stem radial growth during the juvenile stage of development. Age of the tree was the most important variable in explaining stem radial growth. The growth of the GU clone was faster than the growth of the GC clone, possibly indicating that this clone has better genetic potential. However, this could also indicate that, compared to the GC clone, the GU clone is better adapted to the environmental conditions, or it is able to use the available resources more effectively.

ACKNOWLEDGMENTS

The authors are grateful to Dr Valerie Grzekowiak and Dr Nicky Jones for their valuable comments and suggestions.

REFERENCES

Apiolaza LA, Raymond CA, Yeo BJ (2005). Genetic variation of physical and chemical wood properties of *Eucalyptus globulus*. Silvae Genetica 54:160-166.

Arbuckle JL (2006). Amos 7.0 User's Guide. Chicago: SPSS pp. 30-40.

Bollen KA (1989). Structural Equations with Latent Variables. New York: Wiley. pp. 32-39.

Bollen KA, Stine RA (1993). Bootstrapping goodness-of-fit measures in structural equation modeling. In Bollen KA, Long JS (eds.), Testing Structural Equation Models. Newbury Park, CA: Sage. pp. 111-135.

Browne MW, Cudeck R (1989). Single sample cross-validation indices for covariance structures. Multivar. Behav. Res. 24:445-455.

Byrne BM (2001). Structural Equation Modeling with AMOS: Basic concepts, Applications, and Programming. Mahwah, New Jersey: Erlbaum Associates. pp. 86-90.

Crecente-Campo F, Tome M, Soares P, Dieguez-Aranda U (2010). A generalized nonlinear mixed-effect height-diameter model for *Eucalyptus globulus* L. in northern western Spain. For. Ecol. Manage. 259:943-952.

Downes G, Beadle C, Worledge D (1999). Daily stem growth patterns in irrigated *Eucalyptus globulus* and *E. nitens* in relation to climate. Trees 14:102-111.

Downes G, Drew D, Battaglia M, Schulze D (2009). Measuring and modeling stem growth and wood formation: An overview. Dendrochronologia 27:147-157.

Drew DM (2004). Dendrometer trial phase one technical report. Report No. EFR092T. Division of Water, Environment and Forestry Technology, CSIR, Pretoria, South Africa.

Drew DM, Pammenter NW (2006). Vessel frequency, size and arrangement in two eucalypt clones growing at sites differing in water availability. New Zealand J. For. 51:23-28.

Drew D, Downes G, Grzeskowiak V, Naidoo T (2009). Differences in daily stem size variation and growth in two hybrid eucalypt clones. Trees Struct. Funct. 23:585-595.

Eagleman JR (1985). Meteorology, the Atmosphere in Action. Belmont, California: Wadsworth Publishing Co. pp. 17-284.

Efron B, Tibshirani RJ (1993). An introduction to the bootstrap. New York: Chapman and Hall. pp. 184-187.

February EC, Stock WD, Bond WJ, Le Roux DJ (1995). Relationships between water availability and selected vessel characteristics in *Eucalyptus grandis* and two hybrids. IAWA J. 16:269-276.

Gallaham RZ (1962). Geographic variability in growth of forest. In Kozolowski T (ed.), Tree Growth. The Ronald Press Company pp. 311-325.

Garson GD (2004). Path analysis. Retrieved February 20, 2012 from http://faculty.chass.ncsu.edu/garson/pa765/path.htm.

Grapentine T (2000). Path analysis and Structural Equation Modeling. Marketing Research 12: 12-20.

Grewal R, Cote A, Baumgartner H (2004). Multicollinearity and Measurement Error in Structural Equation Models: Implications for Theory Testing, Mark. Sci. 23(4):519-29.

Jagpal HS (1982). Multicollinearity in structural equation models with unobservable variables. J. Mark. Res. 19:199-218.

Kozlowski TT, Pallardy SG (1997). Physiology of Woody Plants. Second edition. San Diego: Academic Press. pp. 1-6.

Leamer E (1978). Specification Searches: Ad Hoc Inference with Non-experimental Data New York: John Wiley and Sons. pp. 87-106.

Malhotra NK, Peterson M, Kleiser S (1999). Marketing research: A state of the art review and directions for the twenty first century. J. Acad. Mark. Sci. 27(2):160-182.

Mardia KV (1970). Measures of multivariate skewness and kurtosis with applications. Biometerika 57:519-530.

Maruyama GM (1998). Basics of Structural Equation Modeling. Thousand Oaks, CA: Sage. pp. 60-61.

Mateus A, Tomé M (2011). Modelling the diameter distribution of *Eucalyptus* plantations with Johnson's S_B probability density function: parameters recovery from a compatible system of equations to predict stand variables. Ann. For. Sci. 68(2):325-335.

Miehle P, Battaglia M, Sands PJ, Forrester DI, Feikema P, Livesley SJ, Morris JD, Arndt SK (2009). A comparison of four process-based models and statistical regression model to predict growth of *Eucalyptus globulus* plantations. Ecol. Model. 220:734-746.

Miller GJ (2001). Environmental Science: Working With the Earth. 8th ed. Pacific Grove, CA: Brooks/Cole. pp. 84-104.

Pallett RN, Sale G (2004). The relative contributions of tree improvement and cultural practice towards productivity gains in *Eucalyptus* pulpwood stands. Fort. Ecol. Manage. 193:33-43.

Schumacker RE (1991). Relationship between multiple regression, path, factor and LISREL analysis. Multiple Linear Regression Viewpoints. 18:28-46.

Schumacher RE, Lomax RG (2004). A beginner's guide to structural equation modeling. (2nd ed.). Second edition. Lawrence Erlbaum Associates Inc. pp. 328-329.

Searson MJ, Thomas DS, Montagu KD, Conroy JP (2004). Wood density and anatomy of water limited eucalypts. Tree Physiol. 24:1295-1302.

Steiger JH 1990). Structural model evaluation and modification: an interval estimation approach. Multivar. Behav. Res. 25:173-180.

Turnbull JW (1999). Eucalyptus plantations. New For. 17:37-52.

Wadsworth RM (1959). An optimum wind speed for plant growth. Ann. Bot. 23:195-199.

West SG, Finch JF, Curran PJ (1995). Structural equation models with non normal variables: Problems and remedies. In Hoyle RH (ed.), Structural Equation Modeling: Concepts, Issues, and Applications. Thousands Oakes, CA: Sage. pp. 56-75.

Yung Y-F, Bentler PM (1996). Bootstrapping techniques in analysis of mean and covariance structures. In Marcoulides GA, Schumacker RE (eds.), Advanced Structural Equation Modeling: Issues and Techniques. Mahwah, NJ: Lawrence Erlbaum Associates pp. 195-226.

Zweifel R, Zimmerman L, Zeugin F, Newbery DM (2006). Intra-annual radial growth and water relations of trees: implication towards a growth mechanism. J. Exper. Bot. 57:1445-1459.

Appendix B : Partial R-code used in the thesis

```
##############################################################################
 ## R code for fitting the selected fractional polynomial models ############
 ################################################################################
library(nlme)
library(lattice) ## will attach library lattice ##
library(foreign)
mygeno<- read.spss(file="C:\\summ98.sav")
mygeno<-as.data.frame(mygeno)
mygeno<-as.data.frame(mygeno)
attach(mygeno)
myg1<-groupedData(radius ~ time|treeno, data = mygeno, outer = ~ clone)
attach(myg1)
dataGu<-myg1[clone=='GU',]
dataGc<-myg1[clone=='GC',]
xyplot(radius ~ time|treeno, mygeno, groups=clone, type="l",
xlab=" Age in weeks ", main="Profile plot of Individual Trees",
ylab="  radial growth ")
interaction.plot(time, clone, radius, fun=mean, col=2:14,
 xlab= "Age in weeks",ylab= " mean radius",
main="Mean profile of radial growth by hybrids",las=1)
interaction.plot(time, as.factor(treeno), radius, fun= var,
col=2:14, xlab= "Age in weeks", ylab= " mean radius",
main=" Profile plot of radial growth ",   las=1)
interaction.plot(time, as.factor(treeno), radius, fun= mean,
  col=2:14, xlab= "Age in weeks",ylab= " mean radius",
  main=" Profile plot of radial growth ",   las=1)
        ## Figure 2.2 ##
  par(mfrow=c(1,2))
  attach(dataGu)
  interaction.plot(time, treeno, radius, fun= mean, col=2:14,
  ylim=c(5000, 30000), xlab= "Age in weeks",ylab= " stem radius in micro metre",
  main=" Profile plot of radial growth for GU clone ",   las=1)
  attach(dataGc)
  interaction.plot(time, treeno, radius, fun= mean, col=2:14,
  ylim=c(5000, 30000), xlab= "Age in weeks",ylab= " stem radius in micro metre",
  main=" Profile plot of radial growth for GC clone ",   las=1)
########### Loess smoothed curves by clone Figure 2.3 ##############################
 attach(dataGu)
 plot(time, radius, type="n", ylim=c(5000, 30000), ylab=" Mean radius in micro metre",
 xlab= " Age in weeks", main="Loess smoothed curves for radial growth of the two clones")
 lines(loess.smooth(time, radius, span=0.6), lty=4)
 attach(dataGc)
 lines(loess.smooth(time, radius, span=0.6),lty=1)
 temp <- legend("topleft", legend = c(" ", " "),
      text.width = strwidth("1,000,000"),
      lty = c(4,1), xjust = 1, yjust = 1,
      title = "Legend")
text(temp$rect$left + temp$rect$w, temp$text$y,
  c( " GU", " GC"), pos=2)
```

```
    ##############################################################################
############lot  of variances of radius  #########################
    attach(myg1)
    attach(dataGu)
    variance1<-tapply(dataGu$radius, time, var)
    plot(unique(time), variance1, type="n", main=" Plot variance functions for GU and GC clones ",
    xlab='Age in weeks', ylab='Variance')
    lines(loess.smooth(unique(time), variance1, span=0.6), lty=4)
    attach(dataGc)
   variance2<-tapply(dataGc$radius, time, var)
   lines(loess.smooth(unique(time), variance2, span=0.6),lty=1, xlab='Age in weeks', ylab='Variance')
        temp <- legend("topleft", legend = c(" ", " "),
             text.width = strwidth("1,000,000"),
             lty = c(4, 1), xjust = 1, yjust = 1,
             title = "Legend")
    text(temp$rect$left + temp$rect$w, temp$text$y,
      c(" GU",  "GC"), pos=2)
   ## Figure 2.4 ##
    par(mfrow=c(1,2))
    attach(myg1)
    plot(unique(time), variance1, type="n", main=" Plot of variance for GC clone ",
    xlab='Age in weeks',ylab='Variance in squared micro metre')
    attach(dataGc)
    variance2<-tapply(dataGc$radius, time, var)
    lines(loess.smooth(unique(time), variance2, span=0.6),lty=1,
 xlab='Age in weeks', ylab='Variance')


attach(dataGu)

variance1<-tapply(dataGu$radius, time, var)

plot(unique(time), variance1, type="n", ylab='Variance in squared micro metre',

main=" Plot of variance for GU clone ", xlab='Age in weeks', )

lines(loess.smooth(unique(time), variance1, span=0.6), lty=4

   ############linear model that contains all covariates  ########################

mygeno1<- read.spss(file="C:\\p2commod.sav")

mygeno<-as.data.frame(mygeno1)

mygeno1<-as.data.frame(mygeno1)

attach(mygeno1)

myg11<-groupedData(radius ~ time|treeno, data = mygeno, outer = ~ clone)

attach(myg11)

dataGu1<-myg11[clone=='GU',]

dataGc1<-myg11[clone=='GC',]

attach(dataGu1)

mod1<-lm(radius~time+ I(sqrt(time))+Temp+rainfall+relhum+solrad+windsp, data=dataGu1)

variance1<-tapply(mod1$residuals, time, var)

par(mfrow=c(1,2))

plot(unique(time), variance1, type="n", ylim=c(0, 2500000),xlab='Age in weeks',

main=" Plot of variance of residuals for GU clone ", ylab='Variance')

lines(loess.smooth(unique(time), variance1, span=0.6), lty=4)

 attach(dataGc1)

mod2<-lm(radius~time+ I(sqrt(time))+Temp+rainfall+relhum+solrad+windsp, data=dataGc1)
```

```
variance2<-tapply(mod2$residuals, time, var)
plot(unique(time), variance2, type="n", ylim=c(0, 2500000), xlab='Age in weeks',
main=" Plot of variance of residuals for GC clone ", ylab='Variance')
lines(loess.smooth(unique(time), variance2, span=0.6), lty=2)
 attach(dataGu1)
mod1<-lm(radius~time+ I(sqrt(time))+Temp+rainfall+relhum+solrad+windsp, data=dataGu1)
variance1<-tapply(mod1$residuals, time, var)
plot(unique(time), variance1, type="n", xlab='Age in weeks',
main=" Plot of variance of residuals for GU and GC clones ",  ylab='Variance')
lines(loess.smooth(unique(time), variance1, span=0.6), lty=4)
 attach(dataGc1)
mod2<-lm(radius~time+ I(sqrt(time))+Temp+rainfall+relhum+solrad+windsp, data=dataGc1)
variance2<-tapply(mod2$residuals, time, var)
lines(loess.smooth(unique(time), variance2, span=0.6), lty=1)
temp <- legend("topleft", legend = c(" ", " "),
        text.width = strwidth("1,000,000"),
        lty = c(4, 1), xjust = 1, yjust = 1,
        title = "Legend")
  text(temp$rect$left + temp$rect$w, temp$text$y,
  c(" GU",  "GC"), pos=2)
   ## Code to plot Figure 2.5 ####################################
  #########################################################
 attach(myg11)
 lmod<-lm(radius~time+ I(sqrt(time))+Temp+rainfall+relhum+solrad+windsp, data=myg11)
res<-cbind(myg11$time, lmod$residuals, myg11$clone)
```

```
 res11<-as.data.frame(res)

 attach(res11)

 rrd<-myg11$radius

 rd40<-res11$V2[res11$V1 ==40]

 rd41<-res11$V2[res11$V1 ==41]

 rd70<-res11$V2[res11$V1==70]

  rd100<-res11$V2[res11$V1==100]

 rd101<-res11$V2[res11$V1==101]

   radius1<-cbind(rd40, rd41, rd70, rd100, rd101)

cor(radius1)

 panel.hist <- function(x, ...)

 {

  usr <- par("usr"); on.exit(par(usr))

  par(usr = c(usr[1:2], 0, 1.5) )

    h <- hist(x, plot = FALSE)

 breaks <- h$breaks; nB <- length(breaks)

   y <- h$counts; y <- y/max(y)

rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)

  }

       panel.cor <- function(x, y, digits=2, prefix="", cex.cor, ...)

       {

        usr <- par("usr"); on.exit(par(usr))

         par(usr = c(0, 1, 0, 1))

        r <- abs(cor(x, y))

      txt <- format(c(r, 0.123456789), digits=digits)[1]
```

```
     txt <- paste(prefix, txt, sep="")
     if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
      text(0.5, 0.5, txt, cex = cex.cor * r)
        }
   pairs(radius1, panel=panel.smooth, cex = 1.5, pch = 24,bg="light green",diag.panel=panel.hist,
    upper.panel=panel.cor,cex.labels = 2, font.labels=2)
########################################################################
#################################  Code Figure 2.6  #######################
     myg11.lm<-lm(radius~time+ I(sqrt(time))+Temp+rainfall+relhum+solrad+windsp, data=myg11)
       fm3Orth.lm <- update( myg11.lm, formula = . ~ . +clone )
       fm4Orth.lm <- update( myg11.lm, formula = . ~ . +clone+clone*time+clone*I(sqrt(time)) )
       fm5Orth.lm <- update( fm4Orth.lm, formula = . ~ . -relhum )
       fm6Orth.lm <- update( fm5Orth.lm, formula = . ~ . -windsp )
       fm7Orth.lm <- update( fm6Orth.lm, formula = . ~ . -clone )
      fm8Orth.lm <- update( fm7Orth.lm, formula = . ~ . -(clone*time)+time )
      summary(fm8Orth.lm)
     bwplot(getGroups(myg11)~resid(fm8Orth.lm), xlab='residuals',
     ylab='tree number', main='Boxplot of OLS residuals for each tree')
          ################  Figure 5.1 and Figure 5.2   #######
      Gu.list<-lmList(radius~time+ I(sqrt(time))|treeno, subset=clone=='GU',
     data=myg1, na.action= drop)
      Gc.list<-lmList(radius~time+I(sqrt(time))|treeno, subset=clone=='GC',
     data=myg1,na.action= drop )
      GcGU.list<-lmList(radius~time+I(sqrt(time))|treeno,
     data=myg1,na.action= drop )
      old<-par(mfrow=c(1,3))
```

```
                    ########   Figure 5.1  ############## ##

   plot(intervals(Gu.list), main= 'Confidence interval for GU clone')

     ## plot(intervals(Gc.list), main= 'Confidence interval for GC clone')##

     ## plot(intervals(GcGU.list), main= 'Confidence interval for both clones')##

       par(old)

     ## Figure 5.2 ##

      Gu.coef<-coef(Gu.list)

      Gu.coef[1:5,]

      Gc.coef<-coef(Gc.list)

      Gc.coef[1:5,]

     old<-par(mfrow=c(1,3))

     boxplot(Gu.coef[,1], Gc.coef[,1], main='Intercepts',names=c('GU', 'GC'))

     boxplot(Gu.coef[,2], Gc.coef[,2], main=' coefficient of time ',names=c('GU', 'GC'))

     boxplot(Gu.coef[,3], Gc.coef[,3], main=' coefficient of square root of time ',names=c('GU', 'GC'))

     par(old)

                #### ## Selecetion of random effects  for fractional polynomial model ### ##

   sqrfc1.lme<-lme(radius~ as.factor(clone)*I(time-39) + as.factor(clone)*I(sqrt(time-39))

    ,control= lmeControl(msMaxIter=100,

  data = myg1,returnObject=TRUE), method= 'REML',random = ~I(sqrt(time-39))+I(sqrt(time-
39)^2)|treeno)  ##

    sqr.lmeI <- lme(radius~ as.factor(clone) *I(time-39)+as.factor(clone)*I(sqrt(time-39)),

  data = myg1,       method= 'REML', random =  ~1|treeno)  ## Model 5##

    sqr.lme<-lme(radius~ as.factor(clone)*I(time-39) + as.factor(clone)*I(sqrt(time-111)),

   data = myg1,    method= 'REML',random = ~ I(time-39)|treeno) ## Model 2##

    sqr.lmes<-lme(radius~ as.factor(clone)*I(time-39) + as.factor(clone)*I(sqrt(time-39)),

   data = myg1,    method= 'REML', random = ~ I(sqrt(time-39))|treeno)## model 3##
  sqrfc.lme<-lme(radius~ as.factor(clone)*I(time-39) + as.factor(clone)*I(sqrt(time-39)),
  data = myg1,  method= 'REML', random = ~-1+I(time-39)+I(sqrt(time-39))|treeno)
   sqrfcmod.lme<-lme(radius~ I(time-39)+ as.factor(clone)*I(sqrt(time-39)),
 data = myg1,   method= 'REML', random = ~-1+I(time-39)+I(sqrt(time-39))|treeno)
   summary(sqrfcmod.lme)
  summary(sqrfc.lme)   ## Table 2 ##
  plot( sqrfc.lme, treeno~resid(.), abline = 0 )
  plot(sqrfc.lme, resid(., type = "p") ~ fitted(.) | clone, id = 0.0005, adj = -0.3 )
  anova(sqrfc.lme)
## models with Different variance functions ##
 sqrfcml.lme<-lme(radius~ as.factor(clone)*I(time-39) + as.factor(clone)*I(sqrt(time-39)),
data = myg1,  method= 'ML',random = ~-1+I(time-39)+I(sqrt(time-39))|treeno)
sqrd.lmes<-update(sqrfc.lme,  method='ML', weights = varIdent(form = ~I( sqrt(time-39))|clone) )
 sqrd1.lmes<-update(sqrfc.lme,  method='ML', weights = varIdent(form = ~1|clone) )
 sqrd11.lme<-update(sqrfc.lme,  data = myg1, method= 'REML',
  weights = varIdent(form = ~I(time-39)|clone))
 anova(sqrfc.lme, sqrd11.lme)
 anova(sqrfc.lme)
      sqrd12.lme<-update(sqrd11.lme, fixed=~as.factor(clone)+I(time-39) + I(sqrt(time-39)),
data = myg1, method= 'ML',    weights = varIdent(form = ~I(time-39)|clone))
  sqrd13.lme<-update(sqrd12.lme, fixed=~as.factor(clone)+I(time-39) , data = myg1,
method= 'ML',  weights = varIdent(form = ~I(time-39)|clone))
  sqrd14.lme<-update(sqrd12.lme, fixed=~as.factor(clone)+I(time-39) , data = myg1,
method= 'REML',  weights = varIdent(form = ~I(time-39)|clone))
     anova(sqrfcml.lme, sqrd13.lme)
     anova(sqrd.lmes, sqrd1.lmes, sqrd11.lme, sqrd13.lme)

      plot( sqrd13.lme, treeno~resid(.), abline = 0 )

   sqrd1ff.lme<-update(sqrd12.lme, fixed=~as.factor(clone)*I(sqrt(time-39))+I(time-39) ,

 data = myg1,   method= 'REML',weights = varIdent(form = ~I(time-39)|clone))

  sqrd1ffi.lme<-update(sqrd12.lme, fixed=~as.factor(clone)+I(time-39) ,

data = myg1,   method= 'ML', weights = varIdent(form = ~I(time-39)|clone))

     vf1fixed<-varFixed(~I(sqrt(time)))

     vfifixed<-Initialize(vf1fixed, data=myg1)

    sqrexp.lme <- update(sqrd11.lme, weights = varExp(form=~I(sqrt(time-39))|clone ))

   ## It fitted well and also better than the constant variance ##

     anova(sqrd11.lme, sqrexp.lme)  ## sqrd11 is choosen because of simplicity ##

  sqrexptime.lme <- update(sqrd11.lme, weights = varExp(form=~I(time-39)+

      I(sqrt(time-39))|clone ))

      anova(sqrd11.lme,sqrexptime.lme)

        sqrexptimeml.lme <- update(sqrd11.lme, method='ml',

   weights = varExp(form=~I(time-39)+I(sqrt(time-39))|clone ))

      vf7 <- varComb(varIdent(form =~ I(sqrt(time-39))| clone) ,varExp(form =~I(sqrt(time-39) )))

   vf77 <- varComb(varIdent(form =~ I(time-39)| clone),varExp(form =~I(time-39) ))

   vf8<- varComb(varIdent(form =~ I((time-39))| clone) ,varExp(form =~I((time-39 ) )|clone))

   sqrcom.lme <- update(sqrd11.lme, weights = vf77)

   anova(sqrd11.lme, sqrcom.lme, sqrexptime.lme)

   sqrcomvf8.lme <- update(sqrd11.lme, weights = vf8)

   anova(sqrd11.lme, sqrcom.lme, sqrexptime.lme, sqrcomvf8.lme)

  ##sqrexptim.lme or sqrcom.lme are better ##

   summary(sqrexptimeml.lme)    ## table 2 ##
   sqrd1ffi.lme<-update(sqrexptime.lme, fixed=~as.factor(clone)+I(time-39) , data = myg1,
  method= 'ML', weights = varIdent(form = ~I(time-39)|clone))  ## selected model ##
   anova (sqrexptime.lme, sqrfc.lme)   ## Table 3 ##
   summary(sqrd1ffi.lme)   ## table 4 ##
    ### ########## Code for Nonlinear mixed models   ##################
        library(nlme)
          library(lattice) ## will attach library lattice ##
          library(foreign)
         mygeno<- read.spss(file="C:\\summ98.sav")
         mygeno<-as.data.frame(mygeno)
         attach(mygeno)
        myg1<-groupedData(radius ~ time|treeno, data = mygeno, outer = ~ clone)
        attach(myg1)
        plot(myg1, outer = ~ clone, legend="FALSE" )
         summary(myg1$clone)
       myGU<-myg1[myg1$clone=="GU",]
     plot(myg1)  ### gives graph of stem radius by time ##
     plot(myGU)
     myg12<-na.omit(myg1)
 ## #########  Fitting separate model to GU and GC ################
    myg12GU<-myg12[clone=="GU",]
    myg12GU<-na.omit(myg12GU)
    myg12GC<-myg12[clone=="GC",]
      myg12GC<-na.omit(myg12GC)
fm1radGU.nls <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myg12GU )
```

```
summary(fm1radGU.nls)
plot(fm1radGU.nls, main="Plot of residuals versus the fitted ")
attach(myg12GU)
plot(fm1radGU.nls, treeno~resid(.), abline=0, main="Box plot of residuals by tree")

fm1rad.lis <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myg12 )
summary(fm1rad.lis)
 plot(intervals(fm1rad.lis), layout=c(3,1))
  plot(fm1rad.lis, treeno~resid(.), abline=0)
   pairs(fm1rad.lis, id=0:1)
 fm1rad.nlme <- nlme(fm1rad.lis)
  fm2rad.nlme <- update( fm1rad.nlme, random= Asym+xmid~1 )
  fm3rad.nlme <- update( fm1rad.nlme, random= Asym+scal~1 )
  fm4rad.nlme <- update( fm1rad.nlme, random= xmid+scal~1 )
   summary(fm1rad.nlme)
      summary(fm2rad.nlme)
       summary(fm3rad.nlme)
       summary(fm4rad.nlme)
      xv<-seq(40, 107, 0.5)
   plot(time, radius, pch=16, col=as.numeric(treeno))
    sapply(1:18,function(i)lines(xv,predict( fm1radextar.nlme,list(treeno=i,time=xv)),lty=2))
 ##  ########## Model 1 Three parameter logistic Regression ##################
   fm1rad.nls <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myeno )
  summary(fm1rad.nls)
  plot(fm1rad.nls, main="Plot of residuals versus the fitted ")
  plot(fm1rad.nls, treeno~resid(.), abline=0, main="Box plot of residuals by tree")
   fm1rad.lis <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myg12 )
   summary(fm1rad.lis)
    plot(intervals(fm1rad.lis), layout=c(3,1))
    plot(fm1rad.lis, treeno~resid(.), abline=0)
    pairs(fm1rad.lis, id=0:1)
    fm1rad.nlme <- nlme(fm1rad.lis)
    fm2rad.nlme <- update( fm1rad.nlme, random= Asym+xmid~1 )
    fm3rad.nlme <- update( fm1rad.nlme, random= Asym+scal~1 )
    fm4rad.nlme <- update( fm1rad.nlme, random= xmid+scal~1 )
    summary(fm1rad.nlme)
    summary(fm2rad.nlme)
     summary(fm3rad.nlme)
       summary(fm4rad.nlme)
      fm1radarma.nlme <- update(fm3rad.nlme, corr = corARMA(p=1, q=1))
     anova(fm1rad.nlme, fm2rad.nlme)
     intervals(fm1rad.nlme, which="var-cov")
     intervals(fm1rad.nlme)
     E2<-resid(fm1rad.nlme, type="normalized")
     F2<-fitted(fm1rad.nlme)
     op<-par(mfrow=c(2,2), mar=c(4,4, 3,2))
     myYlab<-"Residuals"
    plot(x=F2, y=E2, xlab="Fitted values", ylab=myYlab)
    boxplot(E2~clone, data=myg12, main="Clone", ylab=MyYlab)
    plot(x=myg12$time, y=E2, ylab=myYlab, main="Tree age", xlab=" age in weeks")
   par(op)

   plot(augPred(fm1rad.nlme, level=0:1))

  plot( ACF(fm1rad.nlme, maxLag = 15, resType = "n"), alpha = 0.05 )

  ## ########  extending the variance structure of the model ########### ##

    vf1<-varFixed(~time)

    vf2<-varIdent(form=~1|time)

    vf3<-varExp(form=~time)

    vf4<-varComb(varIdent(form=~1|time), varExp(form=~time))

    vf5<-varComb(varConstPower(power=0.1))

   fm1radVI.nlme <- update(fm1rad.nlme,weights=vf2 )

   fm1radVE.nlme <- update(fm1rad.nlme,weights=vf3 )

   fm1radVC.nlme <- update(fm1rad.nlme,weights=vf4 )

   fm1radVCP.nlme <- update(fm1rad.nlme,weights=vf5 )

     ## Fitting model 1 by clone ###

     radFix <- fixef(fm1rad.nlme )

     options( contrasts = c("contr.treatment", "contr.poly") )

     fm1radclone.nlme <- update(fm1rad.nlme, fixed = Asym + xmid + scal ~ clone,

     start = c(radFix[1], 0,  radFix[2], 0,  radFix[3], 0) )

     anova(fm1radclone.nlme, fm1rad.nlme)

     vf3<-varExp(form=~time)

     fm1radcloneex.nlme <- update(fm1radclone.nlme, weight=vf3)

       xv<-seq(40, 107, 0.5)

    plot(time, radius, pch=16, col=as.numeric(treeno))

    sapply(1:18,function(i)lines(xv,predict(fm1radextar.nlme,list(treeno=i,time=xv)),lty=2))

    summary(fm1rad.nlme)
```

```
    plot(fm1rad.nlme, id = 0.005, adj = -1, form = ~ clone )

    plot(augPred(fm1rad.nlme))

   fm3.nlme <- update(fm1rad.nlme, weights = varExp(form=~time))

   anova(fm1rad.nlme, fm3.nlme)   ## no significant difference observed ##

    qqnorm(fm3.nlme )

    plot(augPred(fm3.nlme, level = 0:1) )

    plot( augPred(fm1rad.nlme, level = 0:1) )

     plot( augPred(fm1rad.lis, level = 0:1) )

     plot(ranef(fm1rad.nlme, augFrame = T), form = ~ clone, layout = c(3,1))

    plot(ranef(fm1rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))

   radFix <- fixef(fm1rad.nlme )

   options( contrasts = c("contr.treatment", "contr.poly") )

   fm3rad.nlme <- update(fm1rad.nlme, fixed = Asym + xmid + scal ~ clone,

   start = c(radFix[1], 0,  radFix[2], 0,  radFix[3], 0) )

   anova(fm1rad.nlme,  fm3rad.nlme)

   fm3rad.nlme   ## Score and Xmid are highly correlated ##

   fm333rad.nlme<-update(fm3rad.nlme, random=Asym+xmid~1)

  ## a model without scale random effect ##

   fm33rad.nlme<-update(fm3rad.nlme, random=Asym+scal~1)

  ## a model without xmid random effect ##

   fm332rad.nlme<-update(fm3rad.nlme, random=xmid+scal~1)

   ## a model without Asym random effect ##




## Extending model1 by using variance function ##
```

```
  fm1radext.nlme <- update(fm1rad.nlme , weights = varConstPower(power = 0.1) )

   E2<-resid(fm1radext.nlme, type="normalized")

   F2<-fitted(fm1radext.nlme)

   op<-par(mfrow=c(2,2), mar=c(4,4, 3,2))

   myYlab<-"Residuals"

   plot(x=F2, y=E2, xlab="Fitted values", ylab=myYlab)

   boxplot(E2~clone, data=myg12, main="Clone", ylab=MyYlab)

   plot(x=myg12$time, y=E2, ylab=myYlab, main="Tree age", xlab=" age in weeks")

   par(op)

   fm3radext.nlme <- update( fm3rad.nlme , weights = varConstPower(power = 0.1) )

   fm32radext.nlme <- update( fm3rad.nlme , weights = varConstPower () )

   plot( ACF(fm32radext.nlme , maxLag = 10), alpha = 0.05 , resType="n")

   fm32radextar.nlme <- update(fm1radext.nlme, corr = corARMA(p=0, q=4),
  control=list(niterEM=100))

   corMatrix(fm32radextar.nlme)

   plot( ACF(fm32radextar.nlme , maxLag = 10), alpha = 0.05 , resType="n")

   fm32radextar.nlme <- update(fm1radext.nlme, corr = corARMA(p=0, q=2),
control=list(niterEM=100))

   ####   ## Model 2 The Asymptotic Regression Model ###################

    fm221rad.lis <- nlsList( radius ~ SSasymp(time, Asym, resp0, lrc), data = myg12 )

    plot(intervals(fm221rad.lis) )

    fm221rad.nlme <- nlme(fm221rad.lis )

  summary(fm221rad.nlme)

   plot( augPred(fm221rad.lis, level = 0:1) )

 ## gives plot of augumented prediction ##

   qqnorm(fm221rad.nlme )
```

```
## gives the normal probability plot of residuals ##

   plot(fm221rad.nlme)

   plot(ranef(fm221rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))

   radFix1 <- fixef(fm221rad.nlme )

   options( contrasts = c("contr.treatment", "contr.poly") )

   fm223rad.nlme <- update(fm221rad.nlme, fixed = Asym + resp0 + lrc ~ clone,

   start = c(radFix1[1], 0,  radFix1[2], 0,  radFix1[3], 0) )

   anova(fm223rad.nlme)

   anova(fm223rad.nlme , Terms=c(2, 4,6))

   fm223rad1.nlme<-update(fm223rad.nlme, random=Asym+resp0~1)

  ## a model without scale random effect ##

   fm223rad2.nlme<-update(fm223rad.nlme, random=Asym+lrc~1)

 ## a model without xmid random effect ##

   fm223rad3.nlme<-update(fm223rad.nlme, random=Asym+lrc~1)

  ## a model without Asym random effect ##

   ############### Extending model2 by using variance function ##################

      fm22radext.nlme <- update( fm221rad.nlme , weights = varConstPower(power = 0.1) )

    fm221radext.nlme <- update(fm223rad.nlme, weights = varConstPower(power = 0.1) )

    fm222radext.nlme <- update( fm221rad.nlme , weights = varConstPower () )

    ###  ########Model 3    Asymptotic Regression with an Offset #############

     fm331rad.lis <- nlsList(radius ~ SSasympOff(time, Asym, lrc, c0), data = myg12 )

     plot(intervals(fm331rad.lis) )

     fm331rad.nlme <- nlme(fm331rad.lis )

     fm331rad.nlme

     summary(fm331rad.nlme)
  plot( augPred(fm331rad.nlme, level = 0:1) )

  plot( augPred(fm331rad.lis, level = 0:1) )

  qqnorm(fm331rad.nlme )

  plot(fm331rad.nlme)

  plot(ranef(fm331rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))

  radFix2 <- fixef(fm331rad.nlme )

  options( contrasts = c("contr.treatment", "contr.poly") )

  fm323rad.nlme <- update(fm331rad.nlme, fixed = Asym + lrc+c0 ~ clone,

  start = c(radFix2[1], 0,  radFix2[2], 0,  radFix2[3], 0) )

  fm323rad.nlme

  summary(fm323rad.nlme)

## Extending model3 by using variance function ##

   fm331radext.nlme <- update( fm331rad.nlme , weights = varConstPower(power = 0.1) )

 fm323radext.nlme <- update(fm323rad.nlme,  weights = varConstPower(power = 0.1),
control=list(niterEM=100))

fm333radext.nlme <- update( fm331rad.nlme ,  weights = varConstPower (),
control=list(niterEM=100) )

  ## Model 4  Gompertz model##

  fm431rad.lis <- nlsList(radius ~ SSgompertz(time, Asym, b2, b3), data = myg12)

 plot(intervals(fm431rad.lis) )

 fm431rad.nlme <- nlme(fm431rad.lis )

plot( augPred(fm431rad.nlme, level = 0:1) )

plot( augPred(fm431rad.lis, level = 0:1) )

qqnorm(fm431rad.nlme )

plot(fm431rad.nlme)

plot(ranef(fm431rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))
```

```
radFix2 <- fixef(fm431rad.nlme )

options( contrasts = c("contr.treatment", "contr.poly") )

fm423rad.nlme <- update(fm431rad.nlme, fixed = Asym + b2+b3 ~ clone,

 start = c(radFix2[1], 0,  radFix2[2], 0,  radFix2[3], 0) )

anova(fm1rad.nlme, fm221rad.nlme ,fm331rad.nlme, fm431rad.nlme )


  ##  code additive mixed models  ##############

   library(nlme)

   library(lattice) ## will attach library lattice ##

   library(foreign)

    mygeno<- read.spss(file="C:\\summ98.sav")

    mygeno<-as.data.frame(mygeno)

    attach(mygeno)

    myg1<-groupedData(radius ~ time|treeno, data = mygeno, outer = ~ clone)

    attach(myg1)

     plot(myg1, outer = ~ clone, legend="FALSE" )

     summary(myg1$clone)

    myGU<-myg1[myg1$clone=="GU",]

    plot(myg1)  ### gives graph of stem radius by time ##

    plot(myGU)

    myg12<-na.omit(myg1)

## ######Fitting separate model to GU and GC  ####### ##

    myg12GU<-myg12[clone=="GU",]

    myg12GU<-na.omit(myg12GU)

    myg12GC<-myg12[clone=="GC",]
    myg12GC<-na.omit(myg12GC)

   fm1radGU.nls <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myg12GU )

   summary(fm1radGU.nls)

   plot(fm1radGU.nls, main="Plot of residuals versus the fitted ")

   attach(myg12GU)

   plot(fm1radGU.nls, treeno~resid(.), abline=0, main="Box plot of residuals by tree")

   fm1rad.lis <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myg12 )

   summary(fm1rad.lis)

   plot(intervals(fm1rad.lis), layout=c(3,1))

   plot(fm1rad.lis, treeno~resid(.), abline=0)

   pairs(fm1rad.lis, id=0:1)

   fm1rad.nlme <- nlme(fm1rad.lis)

   fm2rad.nlme <- update( fm1rad.nlme, random= Asym+xmid~1 )

   fm3rad.nlme <- update( fm1rad.nlme, random= Asym+scal~1 )

   fm4rad.nlme <- update( fm1rad.nlme, random= xmid+scal~1 )

   summary(fm1rad.nlme)

   summary(fm2rad.nlme)

   summary(fm3rad.nlme)

   summary(fm4rad.nlme)

## #####   Model 1 Three parameter logistic Regression ###############

     fm1rad.nls <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myeno )

   summary(fm1rad.nls)

   plot(fm1rad.nls, main="Plot of residuals versus the fitted ")

   plot(fm1rad.nls, treeno~resid(.), abline=0, main="Box plot of residuals by tree")

      fm1rad.lis <- nlsList(radius ~ SSlogis(time, Asym, xmid, scal), data = myg12 )
```

```
  summary(fm1rad.lis)

  plot(intervals(fm1rad.lis), layout=c(3,1))

  plot(fm1rad.lis, treeno~resid(.), abline=0)

  pairs(fm1rad.lis, id=0:1)

  fm1rad.nlme <- nlme(fm1rad.lis)

  fm2rad.nlme <- update( fm1rad.nlme, random= Asym+xmid~1 )

  fm3rad.nlme <- update( fm1rad.nlme, random= Asym+scal~1 )

  fm4rad.nlme <- update( fm1rad.nlme, random= xmid+scal~1 )

  summary(fm1rad.nlme)

  summary(fm2rad.nlme)

  summary(fm3rad.nlme)

  summary(fm4rad.nlme)

 fm1radarma.nlme <- update(fm3rad.nlme, corr = corARMA(p=1, q=1))

 anova(fm1rad.nlme, fm2rad.nlme)

 intervals(fm1rad.nlme, which="var-cov")

 intervals(fm1rad.nlme)

 E2<-resid(fm1rad.nlme, type="normalized")

 F2<-fitted(fm1rad.nlme)

 op<-par(mfrow=c(2,2), mar=c(4,4, 3,2))

myYlab<-"Residuals"

plot(x=F2, y=E2, xlab="Fitted values", ylab=myYlab)

boxplot(E2~clone, data=myg12, main="Clone", ylab=MyYlab)

plot(x=myg12$time, y=E2, ylab=myYlab, main="Tree age", xlab=" age in weeks")

par(op)

plot(augPred(fm1rad.nlme, level=0:1))
```

```
#############   extending the variance structure of the model  ########### #

vf1<-varFixed(~time)

vf2<-varIdent(form=~1|time)

vf3<-varExp(form=~time)

vf4<-varComb(varIdent(form=~1|time), varExp(form=~time))

vf5<-varComb(varConstPower(power=0.1))

fm1radVI.nlme <- update(fm1rad.nlme,weights=vf2 )

fm1radVE.nlme <- update(fm1rad.nlme,weights=vf3 )

fm1radVC.nlme <- update(fm1rad.nlme,weights=vf4 )

fm1radVCP.nlme <- update(fm1rad.nlme,weights=vf5 )

  ## Fitting model 1 by clone ###

radFix <- fixef(fm1rad.nlme )

options( contrasts = c("contr.treatment", "contr.poly") )

fm1radclone.nlme <- update(fm1rad.nlme, fixed = Asym + xmid + scal ~ clone,

start = c(radFix[1], 0,  radFix[2], 0,  radFix[3], 0) )

anova(fm1radclone.nlme, fm1rad.nlme)

vf3<-varExp(form=~time)

fm1radcloneex.nlme <- update(fm1radclone.nlme, weight=vf3)

xv<-seq(40, 107, 0.5)

plot(time, radius, pch=16, col=as.numeric(treeno))

sapply(1:18,function(i)lines(xv,predict( fm1radextar.nlme,list(treeno=i,time=xv)),lty=2))

summary(fm1rad.nlme)

plot(fm1rad.nlme, id = 0.005, adj = -1, form = ~ clone )

plot(augPred(fm1rad.nlme))
```

```
    fm3.nlme <- update(fm1rad.nlme, weights = varExp(form=~time))

    anova(fm1rad.nlme, fm3.nlme)   ## no significant difference observed ##

    qqnorm(fm3.nlme )

    plot( augPred(fm1rad.lis, level = 0:1) )

    plot(ranef(fm1rad.nlme, augFrame = T), form = ~ clone, layout = c(3,1))

    plot(ranef(fm1rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))

    radFix <- fixef(fm1rad.nlme )

   options( contrasts = c("contr.treatment", "contr.poly") )

   fm3rad.nlme <- update(fm1rad.nlme, fixed = Asym + xmid + scal ~ clone,

     start = c(radFix[1], 0,  radFix[2], 0,  radFix[3], 0) )

   anova(fm1rad.nlme,  fm3rad.nlme)

   fm3rad.nlme   ## Score and Xmid are highly correlated ##

   fm333rad.nlme<-update(fm3rad.nlme, random=Asym+xmid~1)

  ## a model without scale random effect ##

   fm33rad.nlme<-update(fm3rad.nlme, random=Asym+scal~1)

  ## a model without xmid random effect ##

   fm332rad.nlme<-update(fm3rad.nlme, random=xmid+scal~1)

 ## a model without Asym random effect ##

    ## Extending model1 by using  different variance functions
########################################

     fm1radext.nlme <- update(fm1rad.nlme , weights = varConstPower(power = 0.1) )

    E2<-resid(fm1radext.nlme, type="normalized")

    F2<-fitted(fm1radext.nlme)

    op<-par(mfrow=c(2,2), mar=c(4,4, 3,2))

    myYlab<-"Residuals"

    plot(x=F2, y=E2, xlab="Fitted values", ylab=myYlab)
```

```
boxplot(E2~clone, data=myg12, main="Clone", ylab=MyYlab)
plot(x=myg12$time, y=E2, ylab=myYlab, main="Tree age", xlab=" age in weeks")
par(op)
fm3radext.nlme <- update( fm3rad.nlme , weights = varConstPower(power = 0.1) )
fm32radext.nlme <- update( fm3rad.nlme , weights = varConstPower () )
plot( ACF(fm32radext.nlme , maxLag = 10), alpha = 0.05 , resType="n")
redfm1radextclone.nlme<-update(fm1radextclone.nlme, random=Asym+xmid~1)
red1fm1radextclone.nlme<-update(fm1radextclone.nlme, random=Asym+scal~1)
red2fm1radextclone.nlme<-update(fm1radextclone.nlme, random=scal+xmid~1)
anova(fm1radextclone.nlme)      ### final model ##
plot(augPred(fm1radextclone.nlme, level=0:1))
qqplot(ranef(fm1radextclone.nlme))
plot(fm1radextclone.nlme, treeno~resid(., type="normalized"), abline=0,
main="Box plot of residuals by tree")
anova( fm1radextar.nlme, fm1rad.nlme, fm3radextma2.nlme, fm1radextma10.nlme)
 plot(fm1radextclone.nlme, resid(., type="normalized") ~ fitted(.) | clone,
panel = function(x, y, ...) {
panel.grid()
panel.xyplot(x, y)
panel.loess(x, y, lty = 2)
panel.abline(0, 0)
 } )
plot(augPred(fm1radextclone.nlme, level=0:1))
par(mfrow=c(2,2))
plot(profile(fm1radextclone.nlme))
```

```
## ########### Model 2 The Asymptotic Regression Model  ###########

  fm221rad.lis <- nlsList( radius ~ SSasymp(time, Asym, resp0, lrc), data = myg12 )

  plot(intervals(fm221rad.lis) )

  fm221rad.nlme <- nlme(fm221rad.lis )

  summary(fm221rad.nlme)

  plot( augPred(fm221rad.nlme, level = 0:1) )

  plot( augPred(fm221rad.lis, level = 0:1) )

  qqnorm(fm221rad.nlme )

  plot(fm221rad.nlme)

  plot(ranef(fm221rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))

  radFix1 <- fixef(fm221rad.nlme )

  options( contrasts = c("contr.treatment", "contr.poly") )

  fm223rad.nlme <- update(fm221rad.nlme, fixed = Asym + resp0 + lrc ~ clone,

  start = c(radFix1[1], 0,  radFix1[2], 0,  radFix1[3], 0) )

  anova(fm223rad.nlme)

  anova(fm223rad.nlme , Terms=c(2, 4,6))

  fm223rad1.nlme<-update(fm223rad.nlme, random=Asym+resp0~1)

 ## a model without scale random effect ##

  fm223rad2.nlme<-update(fm223rad.nlme, random=Asym+lrc~1)

 ## a model without xmid random effect ##

  fm223rad3.nlme<-update(fm223rad.nlme, random=Asym+lrc~1)

## a model without Asym random effect ##

 ## Extending model2 by using variance function##


 fm22radext.nlme <- update( fm221rad.nlme , weights = varConstPower(power = 0.1) )
```

```
fm221radext.nlme <- update(fm223rad.nlme, weights = varConstPower(power = 0.1) )
fm222radext.nlme <- update( fm221rad.nlme , weights = varConstPower () )
plot( ACF(fm222radext.nlme , maxLag = 10), alpha = 0.05 , resType="n")
  #### ###   Model 3    Asymptotic Regression with an Offset by clone  ########
  fm331rad.lis <- nlsList(radius ~ SSasympOff(time, Asym, lrc, c0), data = myg12 )
  plot(intervals(fm331rad.lis) )
 fm331rad.nlme <- nlme(fm331rad.lis )
 fm331rad.nlme
 summary(fm331rad.nlme)
 plot( augPred(fm331rad.nlme, level = 0:1) )
 plot( augPred(fm331rad.lis, level = 0:1) )
 qqnorm(fm331rad.nlme )
 plot(fm331rad.nlme)
 plot(ranef(fm331rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))
 radFix2 <- fixef(fm331rad.nlme )
 options( contrasts = c("contr.treatment", "contr.poly") )
 fm323rad.nlme <- update(fm331rad.nlme, fixed = Asym + lrc+c0 ~ clone,
  start = c(radFix2[1], 0,  radFix2[2], 0,  radFix2[3], 0) )
 fm323rad.nlme
 summary(fm323rad.nlme)
## ######  Extending model3 by using variance function #################
 fm331radext.nlme <- update( fm331rad.nlme , weights = varConstPower(power = 0.1) )
 fm323radext.nlme <- update(fm323rad.nlme, weights = varConstPower(power = 0.1),
 control=list(niterEM=100) )
 fm333radext.nlme <- update( fm331rad.nlme , weights = varConstPower (),
       control=list(niterEM=100) )  ##  did not converge  ##
            ## ####  Model 4  Gompertz model  by clone  ######
       fm431rad.lis <- nlsList(radius ~ SSgompertz(time, Asym, b2, b3),
       data = myg12)  ## Model 4 did  converge ##
       plot(intervals(fm431rad.lis) )
       fm431rad.nlme <- nlme(fm431rad.lis )
       plot( augPred(fm431rad.nlme, level = 0:1) )
       plot( augPred(fm431rad.lis, level = 0:1) )
       qqnorm(fm431rad.nlme )
       plot(fm431rad.nlme)
       plot(ranef(fm431rad.nlme, augFrame = T), form = ~ clone , layout = c(3,1))
      radFix2 <- fixef(fm431rad.nlme )
      options( contrasts = c("contr.treatment", "contr.poly") )
      fm423rad.nlme <- update(fm431rad.nlme, fixed = Asym + b2+b3 ~ clone,
      start = c(radFix2[1], 0,  radFix2[2], 0,  radFix2[3], 0) )
      anova(fm1rad.nlme, fm221rad.nlme ,fm331rad.nlme, fm431rad.nlme )


#########  Code for additive mixed models   ###################
     library(nlme)
       library(lattice) ## will attach library lattice ##
       library(foreign)
       mygenoad<-read.spss(file="C:\\Sclimate.sav")
      mygenoad<-as.data.frame(mygenoad)
      attach(mygenoad)
      myg1<-groupedData(radius ~ time1|treeno, data = mygenoad, outer = ~ clone)
```

```
attach(myg1)
 plot( myg1, outer = ~ clone )
summary(myg1$clone)
myGU<-myg1[myg1$clone=="GU",]
myGC<-myg1[myg1$clone=="GC",]
plot(myg1)  ### gives graph of stem radius by time ##
plot(myGU)
myg12<-na.omit(myg1)
#############  ## Additive model code ##############################
  library(lattice)
  op<-par(mfrow=c(3,2),mar=c(5,4,1, 2))
  plot(myGU$time, myGU$radius, type="p", xlab='time', ylab='radius')
 plot(myGC$time, myGC$radius, type="p", xlab='time', ylab='radius')
 library(splines)
 library(gam)
 M1<-gam(radius~s(I(time1-39),3), span=0.5, data=myGU)
 M2<-gam(radius~s(I(time1-39), 3), span=0.5, data=myGC)
 par(mfrow=c(1,2))
 plot(M1, se=TRUE,  main="Additive model for GU clone ")
 plot(M2, se=TRUE,   main=" Additive Model for GC clone")
 par(mfrow=c(1,2))
 M11<-predict(M1, se=TRUE)
 plot(myGU$time, myGU$radius,  type='p', ylab='radius', xlab=' Tree age in weeks')
I1<-order(myGU$time)
lines(myGU$time[I1], M11$fit[I1], lty=1)
```

```
  lines(myGU$time[I1], M11$fit[I1]+2*M11$se[I1], lty=2)
  lines(myGU$time[I1], M11$fit[I1]-2*M11$se[I1], lty=2)
  M21<-predict(M2, se=TRUE)
  plot(myGC$time, myGC$radius,  type='p', ylab='radius', xlab='Tree age' )
 I1<-order(myGC$time)
 lines(myGC$time[I1], M21$fit[I1], lty=1)
 lines(myGC$time[I1], M21$fit[I1]+2*M21$se[I1], lty=2)
 lines(myGC$time[I1], M21$fit[I1]-2*M21$se[I1], lty=2)
 par(op)
library(mgcv)
 plot(myGU$time, myGU$radius,  type='p', ylab='radius', xlab=' Tree age in weeks')
 M3<-gam(radius~s(I(time-39), fx=FALSE, k=-1,bs='cr'), data=myGU)
 M31<-gam(radius~s(I(time-39), fx=FALSE, k=-1,bs='cr'), data=myGU, method='GACV.Cp')
 M32<-gam(radius~s(I(time-39), fx=FALSE, k=1,bs='cr'), data=myGU, method='P-ML')
 M33<-gam(radius~s(I(time-39), fx=FALSE, k=-1,bs='cr'), data=myGU, method='P-REML')
 I1<-order(I(time-39))
 M4<-gam(radius~s(I(time-39), fx=FALSE, k=-1,bs='cr'), data=myGC)
 par(mfrow=c(1,2))
 plot(M3, se=TRUE, main=" The cubic regression spline model to GU clone phase I ",
  xlab ="Tree age", ylab="radius" )
 plot(M4, se=TRUE, main=" The cubic regression spline model to GC clone phase I ",
   xlab ="Tree age", ylab="radius" )
 summary(M3)
 M34<-gam(radius~s(I(time-39)+factor(clone), fx=FALSE, k=-1,bs='cr', data=myg1))
 gam.vcomp(M34,rescale=TRUE,conf.lev=.95)
```

```
    ## Additive mixed modeling ##
    library(mgcv)
    M3M<-gamm(radius~clone+s(time), random=list(clone=~1), data=myg1)
    M3Mtree<-gamm(radius~clone+s(time1), random=list(treeno=~1), data=myg1)
    plot(M3Mtree$gam, xlab="tree age",rug=FALSE, se=TRUE, pages=1,
    too.far=1000, n=10000,pers=TRUE)
    M3Mtemp<-gamm(radius~clone+ s(time1)+s(Temp)+s(rainfall)+s(relhum)+s(windsp)+s(solrad),
    random=list(treeno=~1), data=myg1)
  M3Mtemptt<-gamm(radius~clone+s(time1)+s(Temp)+s(rainfall)+s(relhum)+s(windsp)+s(solrad),
    random=list(treeno=~1+time1), data=myg1)  ## Good moodel for gam ##
  M3Mtempss<-gamm(radius~clone+
season+s(time1)+s(Temp)+s(rainfall)+s(relhum)+s(windsp)+s(solrad),
     random=list(treeno=~1+time1), data=myg1)
      ## Radius and ecah climatic variable  ##
    ## Temperature ##
     Temp1r<-gamm(radius~ s(Temp), random=list(treeno=~1), data=myg1)
     plot(Temp1r$gam, xlab="Temperature",rug=FALSE, se=TRUE, pages=1,too.far=1000,
     n=10000,pers=TRUE)
     gam.check(Temp1r$gam)  ## Validation graph ##
     summary(Temp1r$gam)
     summary(Temp1r$lme)
       ## Rainfall ##
      rainfallr<-gamm(radius~ s(rainfall), random=list(treeno=~1), data=myg1)
      plot(rainfallr$gam, xlab="rainfall",rug=FALSE, se=TRUE, pages=1,too.far=1000,
      n=10000,pers=TRUE)
      gam.check(rainfallr$gam)  ## Validation graph ##
    summary(Temp1r$gam)
    summary(rainfallr$lme)
     ## Relative humidity ##
 relhumr<-gamm(radius~ s(relhum), random=list(treeno=~1), data=myg1)
 plot(relhumr$gam, xlab=" relative humidity",rug=FALSE, se=TRUE, pages=1,
 too.far=1000, n=10000,pers=TRUE)
 gam.check(relhumr$gam)  ## Validation graph ##
 summary(relhumr$gam)
 summary(relhumr$gam)
 ## wind speed ##
windspr<-gamm(radius~ s(windsp), random=list(treeno=~1), data=myg1)
plot(windspr$gam, xlab=" wind speed",rug=FALSE, se=TRUE, pages=1,
too.far=1000, n=10000,pers=TRUE)
gam.check(windspr$gam)  ## Validation graph ##
summary(windspr$gam)
summary(windsp$lme)
## solar radiation ##
solradr <-gamm(radius~ s(solrad), random=list(treeno=~1), data=myg1)
plot(solradr$gam, xlab=" Solar radiation",rug=FALSE, se=TRUE, pages=1,
too.far=1000, n=10000,pers=TRUE)
gam.check(solradr$gam)  ## Validation graph ##
summary(solradr$gam)
summary(solradr$lme)


 ## time (tree age) ##
```

```
    timer <-gamm(radius~ s(time1), random=list(treeno=~1), data=myg1)

    plot(timer$gam, xlab=" Tree age",rug=FALSE, se=TRUE, pages=1,too.far=1000, n=10000,pers=TRUE)

    gam.check(timer$gam)  ## Validation graph ##

    summary(timer$gam)

    summary(timer$lme)

     par(mfrow=c(3,2))

     plot(Temp1r$gam, xlab="Temperature",rug=FALSE, se=TRUE,main="additive mixed model fit ",

     too.far=1000, n=10000,pers=TRUE)

     plot(rainfallr$gam, xlab="rainfall",rug=FALSE, se=TRUE, too.far=1000, n=10000,pers=TRUE)

     plot(relhumr$gam, xlab=" relative humidity",rug=FALSE, se=TRUE, too.far=1000,
n=10000,pers=TRUE)

     plot(windspr$gam, xlab=" wind speed",rug=FALSE, se=TRUE, too.far=1000, n=10000,pers=TRUE)

     plot(solradr$gam, xlab=" Solar radiation",rug=FALSE, se=TRUE, too.far=1000, n=10000,pers=TRUE)

     plot(timer$gam, xlab=" Tree age",rug=FALSE, se=TRUE, too.far=1000, n=10000,pers=TRUE)

### #############   Models by clone ##############################

      mmad$Age<-mmad$time1

      ageclone<-radius~s(Age, by= as.factor(clone))

      Tempclone<-radius~s(Temp, by= as.factor(clone))

      mod1clone<-gamm(ageclone, random=list(treeno=~1), data= mmad)

      plot.gam( xlab= " Tree age   ", mod1clone$gam, pers= TRUE, pages=1,  seWithMean=TRUE)

      anova(mod1$lme, mod1clone$lme)

      par(mfrow=c(1,2))

      plot(mod1clone$gam,residuals=TRUE,pch=19) ## calls plot.gam

      summary(mod1clone$gam)

    gam.check(mod1clone$gam,  pch=19, cex=.3)

   ageclone1<-radius~ as.factor(clone)+s(Age, by= as.factor(clone))
  mod1clone121<-gamm(ageclone1, random=list(treeno=~1), data= mmad)

  gam.check( mod1clone121$gam,  pch=19, cex=.3)

    #####################  Models that smooth by season for tree age by

        including clone and season in the parameteric part     #############

   ageclone1sea<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))

   ageclone1add<-radius~ as.factor(clone)+as.factor(season)+s(Age, by= as.factor(season))

   modadd<-gamm( ageclone1add, random=list(treeno=~1), data= mmad)

   modseason<-gamm(ageclone1sea, random=list(treeno=~1), data= mmad)

   summary(modseason$gam)

   plot(modseason$gam,residuals=TRUE,pch=19, pages=1, xlab="Tree age")

   gam.check( modseason$gam,  pch=19, cex=.3)

     vis.gam(modseason$gam,theta=-35,color="heat")

  vis.gam(modseason$gam, view=c("season", " Age"), theta=-35,color="heat", type="response",
ticktype="detailed")

  #####################  Models that smooth by season for Temp #############

    Tempclone1sea<-radius~ as.factor(clone)+as.factor(season)+s(Temp, by= as.factor(season))

    Tempcloneadd<-radius~ as.factor(clone)*as.factor(season)+s(Temp, by= as.factor(season))

   modTempseason<-gamm(Tempclone1sea, random=list(treeno=~1), data= mmad)

   modTempadd<-gamm( Tempcloneadd, random=list(treeno=~1), data= mmad)

    anova(modTempseason$lme, modTempadd$lme)

   summary(modTempseason$gam)

   plot(modTempseason$gam,residuals=TRUE,pch=19, pages=1, xlab=" Temperature")

   gam.check( modTempseason$gam,  pch=19, cex=.3, pages=1)

  vis.gam(modTempseason$gam,theta=-35,color="heat")

vis.gam(modTempseason$gam, view=c("season", "Temp"), theta=-35,color="heat", type="response",
ticktype="detailed")
```

```
  #####################  Models that include rainfall smooth by season  #############

  rainfallpclone1sea<-radius~ as.factor(clone)+as.factor(season)+s(rainfall, by= as.factor(season))

  rainfallinter<-radius~ as.factor(clone)*as.factor(season)+s(rainfall, by= as.factor(season))

   modrainfallseason<-gamm(rainfallpclone1sea, random=list(treeno=~1), data= mmad)

   modinter<-gamm(rainfallinter, random=list(treeno=~1), data= mmad)

   anova( modrainfallseason$lme,  modinter$lme)

   summary(modrainfallseason$gam)

   plot(modrainfallseason$gam,residuals=TRUE,pch=19, pages=1, xlab=" rainfall")

   gam.check( modrainfallseason$gam,  pch=19, cex=.3, pages=1)

  vis.gam(modrainfallseason$gam,theta=-35,color="heat")

vis.gam(modrainfallseason$gam, view=c("season", "rainfall"), theta=-35,color="heat", type="response",
ticktype="detailed")

#####################  Models that smooth by season for relative humidity #############

  relhumclone1sea<-radius~ as.factor(clone)*as.factor(season)+s(relhum, by= as.factor(season))

  relhumcloneadd<-radius~ as.factor(clone)+as.factor(season)+s(relhum, by= as.factor(season))

   modrelhumseason<-gamm(relhumclone1sea, random=list(treeno=~1), data= mmad)

   modrelhumadd<-gamm( relhumcloneadd, random=list(treeno=~1), data= mmad)

  anova(modrelhumseason$lme, modrelhumadd$lme)

   summary(modrelhumseason$gam)

   plot(modrelhumseason$gam,residuals=TRUE,pch=19, pages=1, xlab=" relative humidity")

   gam.check( modrelhumseason$gam,  pch=19, cex=.3, pages=1)

 vis.gam(modrelhumseason$gam,theta=-35,color="heat")

vis.gam(modrelhumseason$gam, view=c("season", "relhum"), theta=-35,color="heat", type="response",
ticktype="detailed")

#####################  Models that smooth by season for solar radiation #############

   solradintera<-radius~ as.factor(clone)*as.factor(season)+s(solrad, by= as.factor(season))
  solradclone1sea<-radius~ as.factor(clone)+as.factor(season)+s(solrad, by= as.factor(season))

 modsolradint<-gamm( solradintera, random=list(treeno=~1), data= mmad)

 anova( modsolradint$lme, modsolradseason$lme)

   modsolradseason<-gamm(solradclone1sea, random=list(treeno=~1), data= mmad)

   summary(modsolradseason$gam)

   plot(modsolradseason$gam,residuals=TRUE,pch=19, pages=1, xlab=" solar radiation")

   gam.check(modsolradseason$gam,  pch=19, cex=.3, pages=1)

  vis.gam(modsolradseason$gam, view=c("season", "solrad"), theta=-35,color="heat",

  type="response", ticktype="detailed")

####################  Models that smooth by season for wind speed #############

  windspclone1sea<-radius~ as.factor(clone)+as.factor(season)+s(windsp, by= as.factor(season))

   windspcloneinter<-radius~ as.factor(clone)*as.factor(season)+s(windsp, by= as.factor(season))

    modwindspseason<-gamm(windspclone1sea, random=list(treeno=~1), data= mmad)

      modwindinter<-gamm(windspcloneinter, random=list(treeno=~1), data= mmad)

   anova(modwindspseason$lme, modwindinter$lme)

   summary(modwindspseason$gam)

   plot(modwindspseason$gam,residuals=TRUE,pch=19, pages=1, xlab=" solar radiation")

   gam.check(modwindspseason$gam,  pch=19, cex=.3, pages=1)

  vis.gam(modwindspseason$gam, view=c("clone", "windsp"), theta=-35,color="heat",

   type="response", ticktype="detailed")

  vis.gam(modwindspseason$gam, view=c("season", "windsp"), theta=-35,color="heat",
    type="response", ticktype="detailed")

 ######## Smoothing two covaiates at a time  Age and Temperature  ############################

  ageTempsea<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+

  s(Temp, by= as.factor(season))

  ageTempone<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+s(Temp)
   ageTem00<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))
   ageTemp11<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+Temp
   ageTemp11clone<-radius~ as.factor(clone)*as.factor(season)+
    s(Age, by= as.factor(season))+Temp*as.factor(clone)
   ageTemp11seas<-radius~ as.factor(clone)*as.factor(season)+
    s(Age, by= as.factor(season))+Temp*as.factor(season)
      modAgeTemp00<-gamm(ageTem00, random=list(treeno=~1), data= mmad)
    modAgeTemp11<-gamm(ageTemp11, random=list(treeno=~1), data= mmad)
  modAgeTemp11clon<-gamm( ageTemp11clone, random=list(treeno=~1), data= mmad)
    modAgeTemp11seas<-gamm( ageTemp11seas, random=list(treeno=~1), data= mmad)
    modAgepar<-gamm(ageTemppar, random=list(treeno=~1), data= mmad)
    anova( modAgepar$lme,  modAgeone$lme)
    ageTemp11<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))
    modAgeone<-gamm(ageTemp11, random=list(treeno=~1), data= mmad)
   modcloneintsea<-gamm(ageTempcloseas, random=list(treeno=~1), data= mmad)
    ageTemp11<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))
    modAgeTemp<-gamm(ageTempsea, random=list(treeno=~1), data= mmad)
    modAgeTemp11<-gamm(ageTemp11, random=list(treeno=~1), data= mmad)
   anova( modAgeTemp$lme, modAgeTemp11$lme)
  ##  A model with smoothed temperature is not significantly better ###
   summary(modAgeTemp$gam)
   plot(modAgeTemp$gam,residuals=TRUE,pch=19, pages=1 )
   gam.check(modAgeTemp$gam,  pch=19, cex=.3, pages=1)
    vis.gam(modAgeTemp$gam, view=c("Temp", "Age"), theta=-35,color="heat",
    type="response", ticktype="detailed")
```

```
    vis.gam(modAgeTemp$gam, view=c("Temp", " Age"), theta=-35,color="heat", ticktype="detailed")

         ###########    Use temperature  in the parameteric part   ######

  ageTemppar<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+Temp
 *as.factor(season) ## Model that uses temperature in parameteric part ##

  ageTemppar1<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+Temp

   modAgeTemppar<-gamm(ageTemppar, random=list(treeno=~1), data= mmad)

   modAgeTemppar1<-gamm(ageTemppar1, random=list(treeno=~1), data= mmad)

   summary(modAgeTemppar$gam)

   summary(modAgeTemppar1$gam)

  ######## Smoothing two covaiates at a time  Age and rainfall ############################

   agerainfall<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+

     s(rainfall, by= as.factor(season))

    agerainfallone<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))+s(rainfall)

      agerainfall22<-radius~ rainfall+as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))

       agerainfall33<-radius~ as.factor(clone)*as.factor(season)+s(Age, by= as.factor(season))

       modAgerainfallone<-gamm(agerainfallone, random=list(treeno=~1), data= mmad)

    modAgerainfall<-gamm( agerainfall, random=list(treeno=~1), data= mmad)

    modAgerainfall22<-gamm(agerainfall22, random=list(treeno=~1), data= mmad)

     modAgerainfall33<-gamm(agerainfall33, random=list(treeno=~1), data= mmad)

      anova(modAgerainfall22$lme,  modAgerainfall33$lme)

    anova( modAgerainfall$lme, modAgerainfall11$lme)

   ##  A model with smoothed rainfall is not significantly better ###

     summary(modAgerainfall$gam)

     plot(modAgeTemp$gam,residuals=TRUE,pch=19, pages=1, xlab= "Tree age" )

     gam.check(modAgeTemp$gam,  pch=19, cex=.3, pages=1)

     vis.gam(modAgeTemp$gam, view=c("Temp", "Age"), theta=-35,color="heat",
type="response", ticktype="detailed")

 vis.gam(modAgeTemp$gam, view=c("Temp", " Age"), theta=-35,color="heat", ticktype="detailed")
```