ON THE DEVELOPMENT AND APPLICATION OF INDIRECT SITE INDEXES BASED ON EDAPHO-CLIMATIC VARIABLES FOR COMMERCIAL FORESTRY IN SOUTH AFRICA

By

William Kevin Esler

This thesis is presented in partial fulfilment of the requirements for the degree of Master of Science in Forestry at the faculty of Agriscience, University of Stellenbosch.

Supervisor : Professor T. Seifert

January 2012

Declaration of content

By submitting this thesis electronically I hereby declare that the work submitted in this thesis is entirely my own, and has not been submitted in part or in entirety at any other university for a degree.

Acknowledgements

As with any work of this size there are always people who contributed to its completion, in particular I would like to specifically thank and acknowledge the following individuals and organisations:

- *Professor Thomas Seifert* for all the help, guidance, encouragement and support!
- Sappi , Mondi and the ICFR for allowing me access to their data, and more specifically *Nico Hattingh* (Sappi), *Johan Wiese* (Mondi), *Yvonne Fletcher* (Mondi) and *Trevor Morley* (ICFR) for compiling and supplying the data.
- *Anton Kunneke* for calculating the water balance data and extracting the agrohydrology data from the GIS.
- *Dr. Ben du Toit,* and *Cori Ham* for their valuable discussions.
- And finally my wife *Danene* for encouraging me to do this in the first place.

Table of Contents

Synopsis

Site Index is used extensively in modern commercial forestry both as an indicator of current and future site potential, but also as a means of site comparison. The concept is deeply embedded into current forest planning processes, and without it empirical growth and yield modelling would not function in its present form. Most commercial forestry companies in South Africa currently spend hundreds of thousands of Rand annually collecting growth stock data via inventory, but spend little or no money on the default compartment data (specifically Site Index) which is used to estimate over 90% of the product volumes in their long term plans. A need exists to construct reliable methods to determine Site Index for sites which have not been physically measured (the so-called "default", or indirect Site Index). Most previous attempts to model Site Index have used multiple linear regression as the model, alternative methods have been explored in this thesis: Regression tree analysis, random forest analysis, hybrid or model trees, multiple linear regression, and multiple linear regression using regression trees to identify the variables. Regression tree analysis proves to be ideally suited to this type of data, and a generic model with only three site variables was able to capture 49.44 % of the variation in Site Index. Further localisation of the model could prove to be commercially useful.

One of the key assumptions associated with Site Index, that it is unaffected by initial planting density, was tested using linear mixed effects modelling. The results show that there may well be role played by initial stocking in some species (notably *E. dunnii* and *E. nitens*), and that further work may be warranted. It was also shown that early measurement of dominant height results in poor estimates of Site Index, which will have a direct impact on inventory policies and on data to be included in Site Index modelling studies.

This thesis is divided into six chapters: Chapter 1 contains a description of the concept of Site Index and it's origins, as well as, how the concept is used within the current forest planning processes. Chapter 2 contains an analysis on the influence of initial planted density on the estimate of Site

Index. Chapter 3 explores the question of whether the age at which dominant height is measured has any effect on the quality of Site Index estimates. Chapter 4 looks at various modelling methodologies and compares the resultant models. Chapter 5 contains conclusions and recommendations for further study, and finally Chapter 6 discusses how any new Site Index model will effect the current planning protocol.

Keywords: *Indirect Site Index; Dominant Height; Initial planted density; Measurement age; Regression Trees; Random Forest; Hybrid model trees; Multiple Linear Regression.*

Opsomming

Hedendaagse kommersiële bosbou gebruik groeiplek indeks (Site Index) as 'n aanduiding van huidige en toekomstige groeiplek moontlikhede, asook 'n metode om groeiplekke te vergelyk. Hierdie beginsel is diep gewortel in bestaande beplanningsprosesse en daarsonder kan empiriese groei- en opbrengsmodelle nie in hul huidige vorm funksioneer nie. Suid-Afrikaanse bosboumaatskappye bestee jaarliks groot bedrae geld aan die versameling van groeivoorraad data deur middel van opnames, maar weinig of geen geld word aangewend vir die insameling van ongemete vak data (veral groeiplek indeks) nie. Ongemete vak data word gebuik om meer as 90% van die produksie volume te beraam in langtermyn beplaning. 'n Behoefte bestaan om betroubare metodes te ontwikkel om groeiplek indeks te bereken vir groeiplekke wat nog nie opgemeet is nie. Die meeste vorige pogings om groeiplek indeks te beraam het meervoudige linêre regressie as model gebruik. Alternatiewe metodes is ondersoek; naamlik regressieboom analise, ewekansige woud analise, hibriede- of modelbome, meervoudige linêre regressie en meervoudige linêre regressie waarin die veranderlike faktore bepaal is deur regressiebome. Regressieboom analise blyk geskik te wees vir hierdie tipe data en 'n veralgemeende model met slegs drie groeiplek veranderlikes dek 49.44 % van die variasie in groeiplek indeks. Verdere lokalisering van die model kan dus van kommersiële waarde wees.

'n Sleutel aanname is gemaak dat aanvanklike plantdigtheid nie 'n invloed op groeiplek indeks het nie. Hierdie aanname is getoets deur linêre gemengde uitwerkings modelle. Die toetsuitslag dui op 'n moontlikheid dat plantdigtheid wel 'n invloed het op sommige spesies (vernaamlik *E. dunnii* en *E. nitens*) en verdere navorsing kan daarom geregverdig word. Dit is ook bewys dat metings van jonger bome vir dominante hoogtes gee aanleiding tot swak beramings van groeiplek indekse. Gevolglik sal hierdie toestsuitslag groeivoorraad opname beleid, asook die data wat vir groeiplek indeks modellering gebruik word, beïnvloed.

Hierdie tesis word in ses hoofstukke onderverdeel. Hoofstuk een bevat 'n beskrywing van die

beginsel van groeiplek indeks, die oorsprong daarvan, asook hoe die beginsel tans in huidige bosbou beplannings prosesse toegepas word. Hoofstuk twee bestaan uit ń ontleding van die invloed van aanvanklike plantdigtheid op die beraming van groeplek indeks. In hoofstuk drie word ondersoek wat die moontlike invloed is van die ouderdom waarop metings vir dominante hoogte geneem word, op die kwaliteit van groeplek indeks beramings het. Hoofstuk vier verken verskeie modelle metodologieë en vergelyk die uitslaggewende modelle. Hoofstuk vyf bevat gevolgtrekkings en voorstelle vir verdere studies. Afsluitend, is hoofstuk ses ń bespreking van hoe enige nuwe groeiplek indeks modelle die huidige beplannings protokol kan beïnvloed.

Sleutelwoorde: *Indirekte groeiplek indeks, Dominante hoogte, Aanvanklike plantdigtheid, Opname ouderdom, Regressieboom analise, Ewekansige woud analise, Hibriede- of modelbome, Meervoudige linêre regressie.*

List of Tables

List of Equations

List of Figures

Chapter 1. SITE INDEX IN THE SOUTH AFRICAN PLANNING PROTOCOL

1.1. Introduction

The maximum productive capacity of any given site can be defined as the total biomass produced if the stand has fully utilised the available resources such as water, nutrients and solar radiation to produce tree growth. The concept is important because it allows for an estimate of the maximum amount of product (in this case wood fibre) that the site is capable of producing (West 2004). However since trees complete against one another for resources their individual sizes in the stand will differ, those that are more competitive will become larger and suppress the less competitive smaller trees. The degree of competition will be determined by the stand density, and the rate of growth of the larger more dominant trees. The dominant trees are therefore a reflection of the productive capacity of the site for that particular tree species. (West 2004)

Determining the productive capacity (or site quality) of a particular stand is important if one requires estimates of current and/or future production. It can also be used as a means of comparing actual production to potential, and to determine the correct species to be planted on the site.

Site quality can be determined by a number of methods (Loetsch et al. 1973):

- Using measured tree variables that are considered to be expressions of the effect of site on the tree (such as height).
- Using the natural vegetation and species mix as an indicator of site quality, and by
- Using soil, topographical and climatic features to determine site quality.

1.2. Origins of the concepts of Site and Site Index

There is a lack of conformity over the use and definition of the term “site” – it can be used in reference to the inherent features of the site (such as climate or soil), or to the growth of the trees on

the site. Since interest is in the crop rather than the land it is the second definition which is of more importance. Johnston et al. (1967) uses the terms "*site classifications*" and "*growth classifications*" to distinguish between the two types of classifications, other authors call these two methods "*Geocentric*" or earth based and "*Phytocentric*" or plant based (West 2004, Vanclay 1994). Skovsgaard & Vanclay (2008) use the terms "*Site quality"* and "*Site productivity"* to discriminate between the two concepts.

The actual number of true forest site classification methods is small since most do not reflect differences in tree growth potential, or cannot be expressed in terms of volume, most soil survey classifications for example cannot be used to define tree growth differences (Johnston et al. 1967). Johnston et al. (1967) give the following site and growth classifications:

Site classifications can be classed into :

Floristic Site Classifications : where the ground vegetation is correlated to tree growth. This is limited to areas that have been relatively undisturbed, and where there is little site and species variation. The method is most often used in the large (indigenous) coniferous forests of the northern hemisphere.

Environmental Site Classifications : particularly the use of soil variables and soil types as a method of site classification.

Climatic Site Classifications : where climatic variables such as temperature, rainfall, evapotranspiration, length of growing season etc. are correlated to growth. These indices seem to be useful on large scales such as countries or continents.

Growth classifications can be divided into:

Volume Site Classifications : where either mean annual increment (MAI_n) at a base age or more commonly the maximum MAI_{max} are used. However where the stand has been thinned, or where there has been heavy natural mortality the MAI_n or MAI_{max} becomes difficult to measure and interpret.

Basal Area Site Classifications: where basal area is used when the forest has reached a state of equilibrium (only useful for natural, or very old forests).

Height Site Classifications : where height (either mean height, or dominant height) are used at a some reference age to define the site classes.

These site classifications can be further divided into *direct* (e.g. directly measured tree volume, or height) and *indirect* (e.g. the use of ground vegetation, or soil type) methods (Skovsgaard & Vanclay 2008).

Since the interaction of edaphic and climatic effects on tree growth can be complex and these interactions are only partially understood, and in most cases the original ground vegetation has been removed or modified, the effect of the site on the crop can be used as a proxy for site quality. Originally this was based on the total standing volume or yield produced (and by implication MAI), however, since the introduction of silvicultural treatments such as thinning have a material effect on the yield it became necessary to find a measure of site quality that was less susceptible to forest operations, was easy to measure, and was highly correlated to the productive capacity of the site.

As early as 1765 Oettelt suggested stand height as the best indicator of site quality from the other easily measured stand characteristics (Loetsch et al. 1973). At around the same time (1788[1]) de Perthuis de Laillevault also proposed the use of height to assess site quality (Batho & García 2006). In 1841 Heyer identified the correlation between height and volume growth (Skovsgaard & Vanclay 2008). Later the mean height at a particular reference age became an obvious substitute and was successfully used in the construction of the original yield tables in Germany in the 1870's. Stands were classed into the various qualities using the "band method", or relative site classification, whereby a large number of stands with varying ages and productivities were measured. The upper and lower bounds of the variation in mean height over time were determined and the curves plotted. The difference between the upper and lower curves was then divided equally at the reference age into bands (generally five bands were used to define the site classes). The mean curves of these bands then defined the height quality classes. Stands that fell into a particular class were expected to have similar volumes at the same age (given similar stems per hectare), and the mean height of the stand would develop along the curve defined by the class. Eichorn formulated the so-called Eichorn rule in 1902 / 1904 which stated: *a given mean height of a stand delivers the same volume in all site classes*[2] (Skovsgaard & Vanclay 2008).

[1] Published posthumously in 1803 by his son.

[2] If not heavily thinned , and for a given tree species.

Later, dominant height was used together with mean height since it was possible during a heavy low thinning for a stand to artificially move from one site class to the next because the mean basal area had increased and therefore the mean height (Assmann 1971). Dominant height has the advantage over mean height in that it is less effected by thinnings where small or malformed trees are removed (García 1983).

When it is assumed that due to thinning or natural mortality that the trees with smaller than average basal area are the ones that are removed or die, additional variation is also introduced – a further reason for the introduction of dominant height (Pienaar 1965). This "**Site Index**" (dominant height at a reference age) replaced the previous **Site Class** (mean height at a reference age) Van Laar & Akça (1997).

Fairly early on attempts were made to describe the quantifiable relationships that were seen in yield tables as formulae. Attempts were also made to formulate universal growth "laws" - however, these proved to be too ambitious, and a general understanding that it is not possible to construct a single generally valid growth "law" was arrived at (Assmann 1971). With the advent of computer technology it became more practical to transform yield tables into empirical models, and the concept of Site Index was incorporated into these growth models. These empirical models form the core of forest planning, inventory and management systems today.

Typical empirical stand growth models are combinations of various mathematical functions that describe elements of

- **Stand growth** (e.g. dominant height, basal area, stems per hectare (SPHA), survival/mortality, basal area responses to thinning operations, and volume),
- **Stand Structure** (e.g. diameter distributions, average height) and
- **Product** (e.g. merchantable volume, log breakdowns).

The growth models themselves can either be calibrated with measured data obtained via inventory data (i.e. temporary sample plots or TSP's); or where the compartment has not been measured,

default data on the compartment; the regime; and the productivity of the site (in other words Site Index) are used to predict the future stand variable's.

Site Index can be seen as the integral of the site variables such as soil, radiation and rainfall on tree growth, one method to circumvent the use of Site Index is to incorporate the site variables directly into the growth model (Kaufmann & Ryan 1986). This method is, however, difficult to implement and highly data intensive.

1.3. Defining Site Index

The definition of dominant height, top height and Site Index can be problematic. The terms top height and dominant height are generally accepted as synonyms, however, the definitions of each are not standardised or universal (Philip 1994). Other terms such as total height and predominant height add to the confusion.

The definition of Site Index also has a material effect on the quality of the estimate[*], and can lead to statistically different estimates (Sharma et al. 2002). García (2010) states that the actual definitions of Site Index cannot in themselves be correct or incorrect, but that the statistical treatments will differ.

There are numerous definitions for top or dominant height, including :

- The average height of the dominants and co-dominants (the selection and definition of dominants and co-dominants can also be subjective).
- The average height of the dominants.
- The mean height of the 5; 30 ; 100 tallest trees per acre/hectare.
- The mean height of the 40; 100 largest (diameter) trees per acre/hectare.
- The average of heights greater than two standard deviations above the arithmetic mean.
- The regression height of the tree with a diameter equal to the mean plus two standard deviations of the diameter distribution.

[*] Sharma et al. (2002) defined the quality (or effectiveness) of the definition to mean : if the estimate of Site Index at base age was close to or the same as an estimate several years below of above the base age.

- The regression height of the tree with a diameter equal to the mean plus one and a half standard deviations of the diameter distribution.
- The average of the largest diameter trees within a certain distance from an inventory sample point.

(Philip 1994; Bredenkamp 1993; van Laar & Akça 1997; Husch et al. 2003; Van Laar 1978; Johnston et al. 1967)

Bredenkamp (1993) defined dominant (top) Height as the expected height of the largest diameter trees on a random 0.01 ha plot, however, for practical purposes he gave the following method of calculation, which is now the South African standard:

> ***Dominant height is calculated from the mean height of the top 20 % largest quadratic mean diameter trees. The height is based on the regression of the natural log of height, and the inverse of diameter at breast height based on a sample of at least 30 diameter/height pairs.***

Site Index can be viewed as either a property of the stand, in other words the actual dominant height achieved by the stand at the specific base age, or as a property of the site – in that the Site Index is seen as an average over a hypothetical stand which could be grown on that site, with Site Index being the most likely dominant height at the base age. García (2005; 2010) calls these two definitions the "**stand site index**" and the "**site site index**", the "site site index" in his view is more appropriate since it is in keeping with the original concept, and it also renders the base age irrelevant.

1.4. 'Direct' Site Index models

At this point some differentiation needs to be made between the Site Index models used to estimate Site Index, and to predict future height using measured data (i.e. direct methods), versus the models referred to in this thesis to estimate 'default' Site Index (i.e. indirect).

'Direct' Site Index models can take many forms but are generally based on regression equations which express height as a function of age (van Laar et al. 1997)[3]. Site Index is then calculated as a function of the measured dominant height, the age at which this height was measured and the base age. In other words the dominant height is projected to the base age for Site Index.

'Direct' Site Index models should fulfil the following requirements (Grey 1989):

- Provide unbiased estimates with equal precision across a range of ages.
- Base-age invariant, i.e. the model should produce the same results irrespective of the base age chosen, and height should equal zero when age equals zero.
- Individual Site Index curves should have individual and independent asymptotes.
- The models should be closed in form (i.e. not require iteration).

García (2004) gives three of the most common methods used to construct direct Site Index models based on either PSP (permanent sample plot) or stem analysis data[4]: Parameter prediction, mixed effects, and differential equations.

Further discussion of 'direct' Site Index models is limited to their use in the South African forest planning protocols, and unless stated, all references to Site Index models are to 'indirect' or 'default' models.

1.5. Problems and Limitations

There are a number of limitations associated with the use of Site Index within the growth model structure, these include:

The concept does not work well where there are either multiple species or multiple ages in a stand, or where the stand age is difficult to determine (Avery & Burkhart 2002; Husch et al. 2003). This is not really an issue for most South African commercial forestry companies if record keeping is kept

[3] Examples of difference form Site Index models can be found in section 4.3.3

[4] This is specifically with reference to empirical growth models – there are a number of other forest models forms such as process and hybrid models (Subasinghe 2008),

reliable, and if only single species have been used per stand. Historically there were two occasions when this may have presented a problem : At the start of the clonal programme when large numbers of single clones were not available and multiple clones were planted in single stands, and where a different species was used to blank[5] a compartment – both of these situations are now rare.

Site Index is not comparable between species (Kaufmann & Ryan 1986; Avery & Burkhart 2002; Husch et al. 2003), and definitely not between genera. It is possible, however, that this is a reflection of the growth model, or of the data used to construct the height model. Where data from multiple species in the same genus have been conglomerated to produce generic Site Index models (e.g. a generic *Eucalyptus* growth model for a specific geographic region), the issue may be less relevant.

Site Index is not a constant, it can change over time due to climatic; environmental; or management changes (Avery & Burkhart 2002; Skovsgaard & Vanclay 2008). Examples include research conducted by Spiecker (1999) which has shown a general increase in site productivity between successive rotations in many forest sites across Europe. Martín-Benito (2008) used the analysis of residuals from dominant height equations over time to detect trends in dominant height growth for *Pinus nigra* in three study areas in Spain, the author found reductions in dominant height growth over two decades (1960's & 1970's). Similar Site Index changes have been detected in spacing trials in South Africa over much shorter time periods due to abnormal rainfall (in this case a severe drought). Coetzee et al. (1996) found a change in the calculated Site Index by 2.68 m over as short a period as three years in a *Eucalyptus grandis* spacing trial at Kwambonambi in Zululand, South Africa. In the 3.0 x 3.5 m espacement or 952 SPHA plot, the Site Index was calculated as 26.76 m at age 5, and 24.08 m at age 8. Coetzee & Naicker (1998a) also found that the calculated Site Index had changed at the Kia-Ora *Eucalyptus grandis* spacing trial in Kwazulu Natal, South Africa. Over a four year period the Site Index had changed by as much as 2.5 m. (Plot 16 planted at 1666 SPHA, 3 m x 2 m. At age five it was calculated to be 16.68 m, at age nine it was 14.17 m). Coetzee & Naicker (1998b) gave a comparable example in the results of the Tanhurst *Eucalyptus grandis* spacing trial, where the Site Index had changed by 1.8 m in the 3 m x 2 m plot over five years. Coetzee (1994) pointed out that if such data is used in the development of Site Index guide curves,

[5] Blanking is a term used to denote the replacement of dead or missing seedlings soon after planting (generally this operation is carried out within three months of planting).

that the resultant functions would not reflect the development of height growth under normal growing conditions. He made an attempt to incorporate rainfall as an additional predictor variable, but it did not contribute significantly to the particular dominant height function he was building. Smith, Kassier and Cunningham (2005) pointed out in their summary of the 1986 trial series laid out to determine the effects of initial stand density on *Eucalyptus grandis*, that height growth varied widely due to drought effects during the course of the trials, resulting in differing estimates of Site Index.

Height can be one of the more difficult stand variables to measure accurately during inventory – especially during periods of wind, or where the crown is difficult to see. As a result Site Index can only be accurately determined when it is close to the base age (Sharma et al. 2002). The dominant height is inferred using the relationship between diameter and height calculated via a sub-sample during the inventory process. It is not measured directly. These regressions have been known to produce low correlations in plantation forestry – with R^2 values below 0.3 common unless the sample is taken systematically[6] (i.e. the sample is made up of selected large and small trees, with fewer "average" trees than would be the case if the diameter distribution was followed – this method breaks the random sample rule, but improves the regression between diameter and height).

Studies carried out on pine and spruce in Canada (Nigh & Love 1999) have shown that apparently undamaged trees selected for the measurement of Site Index had significantly more internal damage from frost and insects than anticipated when they were split open to measure height growth from the terminal bud scars. Over 50 % of the pine and 75 % of the spruce trees had damage which was not externally visible. There was evidence that this damage had effected the height growth of the pine trees.

Given these potential measurement issues it is possible to make substantial errors in the estimate of dominant height, and therefore of Site Index.

Since dominant height is considered to be relatively independent of variables such as mean diameter and mortality the prediction models are normally developed separately as a self contained sub-

[6] Based on personal experience.

model (García 1983). It is fairly common to have separate height and Site Index equations within the growth model configurations – one used to predict Site Index from a known height and age, the other to estimate height as a function of Site Index and age. This can result in incompatibility whereby the prediction of Site Index from known heights and ages, do not compare well with height predictions using the Site Index and age (Rose et al. 2003).

Coetzee (1994) pointed out that it is important to keep in mind the age of the trees used in the data sets to develop a Site Index equation, extrapolation beyond the age that was used can produce unreliable results. This is quite often seen in practice where abnormally old stands produce unrealistic results if predictions are made using the standard growth model configurations. Coetzee (1994) suggested the use of polymorphic growth curves which allow for different shaped curves for different ages and sites, rather than the single guide curve or anamorphic approach, which is proportionally shifted above and below depending on site quality. An alternative approach is the development of multiple anamorphic Site Index curves to suit the various data ranges. What is of more importance is that the data used to construct Site Index functions should be reflective of both the age ranges and growth conditions that the function will eventually be used for.

One of the key assumptions necessary to uphold the concept of Site Index is that it is unaffected by initial stand density, however, some research has pointed to the possibility that this does not always hold true. Since this is so important, it forms the basis of the first objective in this study.

Finally, since Site Index is defined as a dominant height at a specific point (the base age), it singles out this part of the growth curve, and does not fully describe the way height growth has developed, or will develop over the life span of the stand (Grey 1989).

Given all of these limitations, why is Site Index still favoured? The simple answer is that the alternatives (such as, stem volume, biomass, combined height and diameter etc.) are either too difficult to incorporate, or too expensive to measure. As long as the limitations are recognised and properly managed (by reducing known bias), Site Index is currently the only viable alternative.

1.6. The application of Site Index in South African forest planning protocols

As previously stated, the concept of Site Index has been fully incorporated into the stand growth models used by commercial forestry companies in South Africa today. In a typical pulpwood working circle model configuration, Site Index is used directly to calculate height and volume and indirectly (via the calculated height) in the calculation of basal area (see Figure 1) – so it is a key element in the determination of the future stand variables. The estimates of site productivity need to be accurate, since any bias introduced can affect all of the model results (Vanclay 1994).

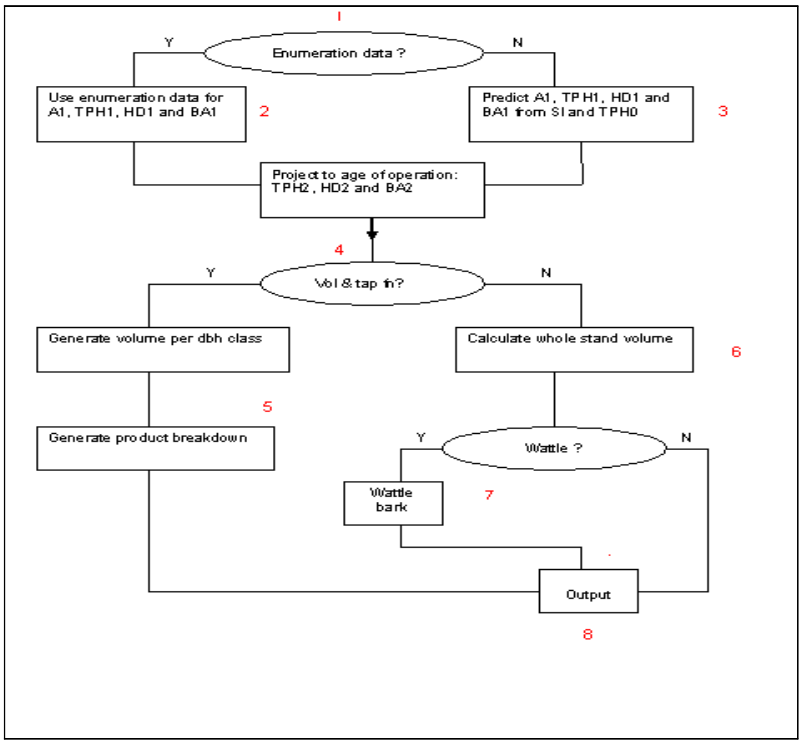


Figure 1: The process flow for the empirical growth model configurations usual to South African commercial forest companies (Adapted from Fletcher 2006).

Variable list for variables used in Figure 1

A1	Age at point of calibration (years)
TPH0	Planting stems per hectare
TPH1	Stems per hectare at point of calibration
TPH2	Stems per hectare at point of projection
HD1	Dominant height at point of calibration
HD2	Dominant height at point of projection

BA1	Basal area at point of calibration
BA2	Basal area at point of projection
SI	Site Index

During a typical simulation which takes place within the harvest scheduling system (HSS)*, the following sequence takes place:

- The compartment is checked for relevant inventory data.
- If the compartment has been enumerated, the SPHA, Basal area and dominant height at the age of inventory are projected to the future age of operation (Felling or Thinning).
- If there is no inventory data the compartment default Site Index, and the initial planted SPHA are used to predict the SPHA, Basal area and dominant height at the future age of operation.
- Once the future SPHA, Basal area and dominant height have been calculated, the volume is then calculated either by
- the use of volume and taper functions, which generate product breakdown and log volumes per diameter class (mostly required in the mining timber and saw timber regimes), or
- via whole stand volume equations which calculate the merchantable volume in the stand.
- In wattle compartments, bark volumes are also calculated.
- Finally the output for the stand is produced. This output is in cubic meters and has to be converted using a factor to the unit of production (usually tonnes in pulpwood working circles).

The critical use of the default Site Index in un-enumerated compartments (which form the majority of any long term plan) is obvious. If this default is poorly calculated the resulting volume predictions will also be poorly calculated.

1.6.1. The forest planning process

Forest planning can be separated into three distinct levels: **Strategic**, **Tactical** and **Operational** planning. Each of these planning levels have different objectives, methodologies, and time scales. Strategic planning is concerned mainly with long term (20 to 30 years) sustainability of production. Tactical planning is primarily concerned with resource balancing (roads; machines; contractors;

* HSS – Harvest Scheduling System by Syndicate database solutions (see : http://syndicate.co.za/files/hss.html) is currently the main strategic forest planning tool used in South Africa today – it is used to simulate the effects of management decisions on long term forest production. A growth and yield simulator is incorporated into the system which utilises empirical growth models to predict stand variables.

labour; capital expenditure etc. on a medium term scale of between 3 and 5 years), and operational planning is principally concerned with production (felling directions, safety, plant numbers etc. on an annual or monthly level). Operational planning itself has two levels – the annual plan of operations (or APO), and the compartment plan. Each planning level feeds into one another (see Figure 3) and the entire process follows a specific time line (see Figure 2 below).

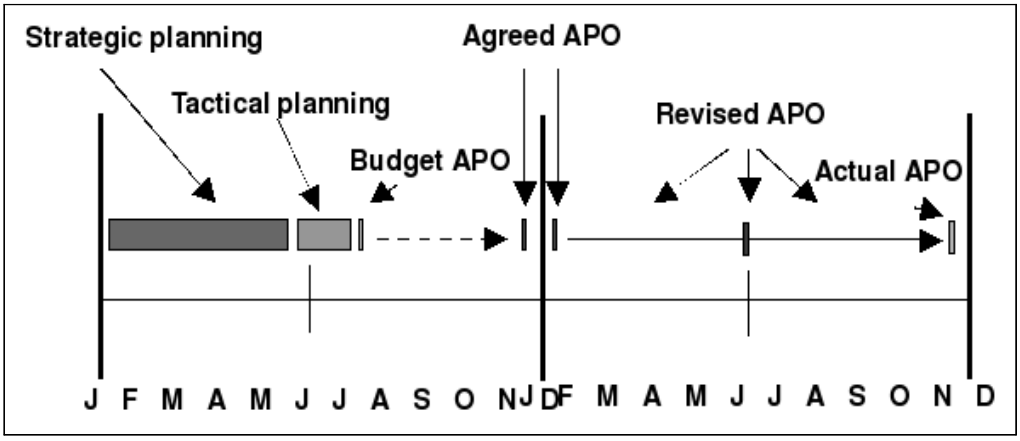


Figure 2: The planning time line (Morkel 2005).

The planning time line can be read as follows: during the initial months of the year the strategic plan is reconstructed – this is generally completed during April or May[7]. The tactical plan is then created from the first 3 to 5 years of the strategic plan and the first year of this plan in turn is used to put together the budget for the following year. Since the budget is usually put together in June or July (for approval in August/September), there is a time delay of approximately 5 months before the annual plan of operations (APO) takes effect. Changes and adjustments are made during this period, and the plan is continually reviewed during the year – the actual APO is audited and monitored, and these changes in turn are fed back into the input for the following Strategic Plan (see Figure 3).

[7] Time-lines will obviously differ between companies depending on the financial year

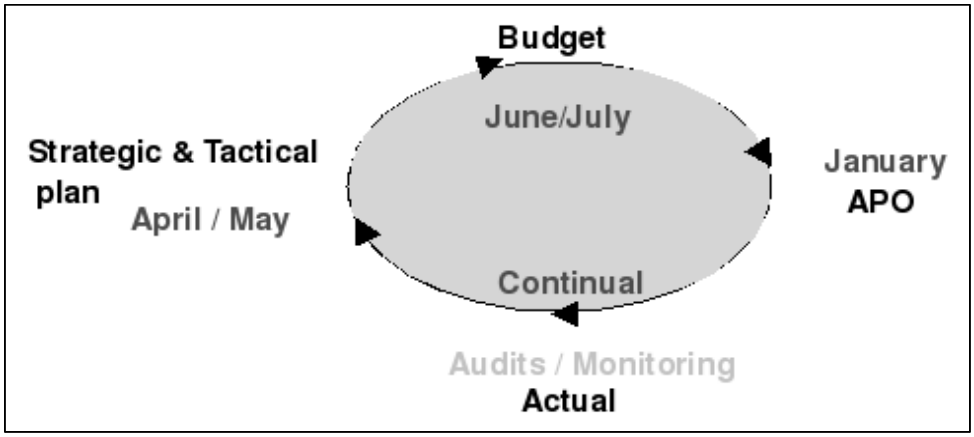


Figure 3: The planning loop.

The data generated by each planning process is used by each other planning process: so for example the APO is used as input into the Strategic plan, and visa versa. Obviously the timing for each of these processes will differ for each company depending on when their financial year starts and ends. If the financial year does not correspond to a normal calendar year the timing will be different (and in most cases will add to the complexity of management). Since new Site Index models will specifically affect the processes involved with the production of plans it is worth focusing in on these processes. Figure 4 below shows the current generalised process of plan production followed

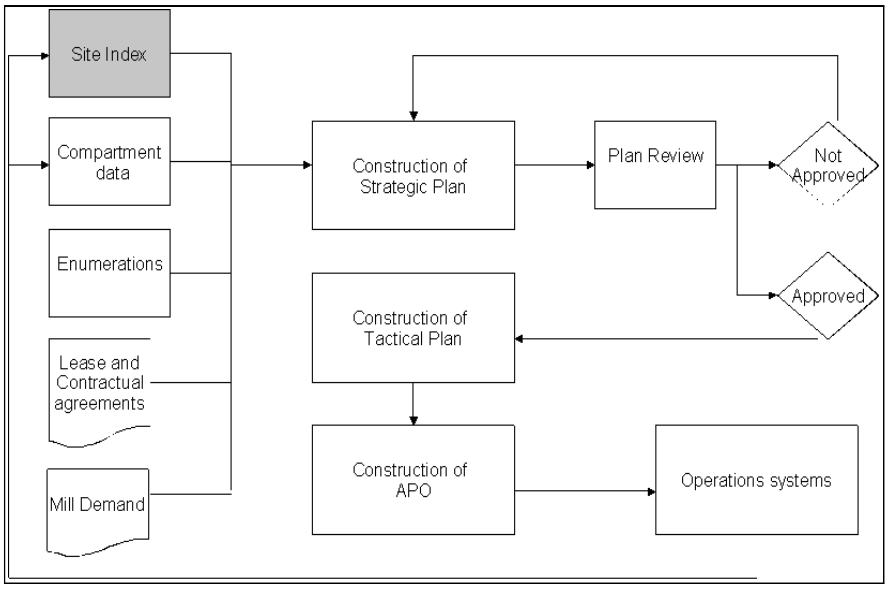


Figure 4: High level process flow of the process's required to produce Strategic, Tactical and Operational plans. Showing where the default Site Index fits into the process.

by most large commercial forestry companies in South Africa .

Throughout the year the operations systems record all operations that occur within a compartment (this includes any operation for which payment is required). These systems vary from company to company but are generally integrated into the financial systems and may or may not be part of the forestry database (the compartment register), however, they will invariably be linked in some form to the compartment data, either directly or indirectly. Operations which cause a status change to a compartment (e.g. if the compartment is felled or planted) are used to update the compartment database.

Currently actual production from compartments, in the form of tonnes or m^3 produced is used within the Site Index process to update the Site Indexes (see Figure 5). It is this process which will be directly affected by the introduction of an alternative methodology.

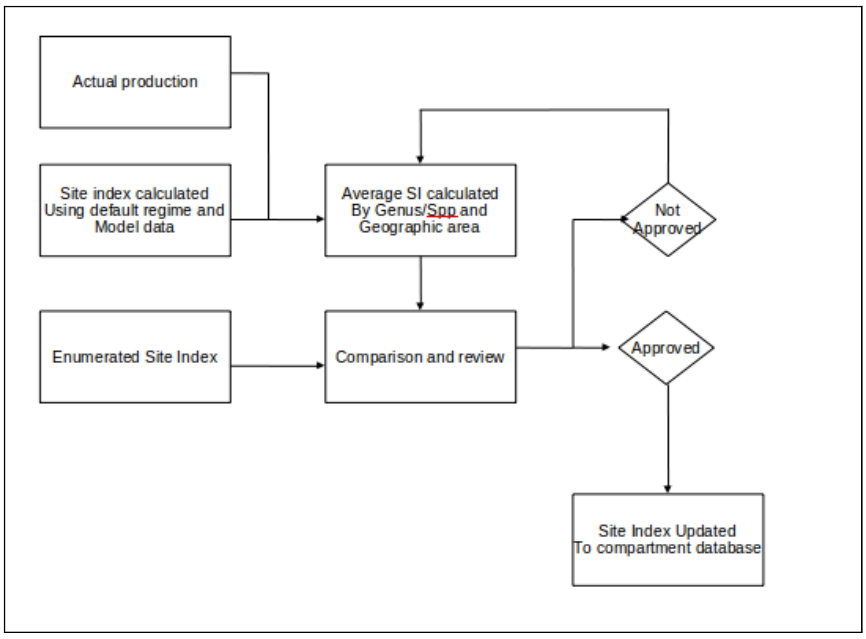


Figure 5: The current process followed to generate the default Site Index.

The Site Index data together with inventory data; lease and contractual information; and mill demand (both in the form of quantity and quality) are used to construct the strategic plan. Once the plan has been constructed it is reviewed in detail, and if approved the initial years of the plan are

used as input into the tactical plan. The APO is then constructed using the first year of the tactical plan, and as actual production takes place this is captured into the operations systems.

Various methods are currently used to generate the default (i.e. un-measured, or un-enumerated) Site Index which is utilised in the production of Strategic and Tactical plans. There are two common methods:

- The use of past inventory data, or using inventory data from adjacent compartments. This method can be time consuming and regime or species changes confound the calculation. However with the use of GIS technology it is possible to reduce the time it takes to do the calculations.
- The more common method is to "reverse engineer" the Site Index from the expected volume, it is this method which is described here.

Actual production in the form of tonnes or m^3 produced from compartments in the past is conglomerated via the operations systems. This data is used to estimate the future productive capacity of sites. The growth models are then used to calculate what Site Index is needed (given the default regime) to produce the given production. A spreadsheet is then generally used to calculate the average Site Index for the given genus, species, working circle[8] and geographic area. How this data is separated will depend on the amount of data available, and the business hierarchy. This "default" Site Index is then compared to the enumerated Site Index, and any further production information. A process of review together with the relevant harvesting forester will then either approve or not approve the default Site Index. If it is approved, this data is then updated to the compartment database.

As can be seen, this process is unscientific and has a number of problems associated with it:

- The method relies on considerable manual involvement on the part of the forest planner (in the form of maintaining and updating the Site Index spreadsheet).
- The method is not standardised across business units where there are different forest planners, or where the data levels are different.
- The method does not use the enumerated Site Index directly.

[8] Working circle is a reference to the product the compartment was established to produce – e.g. saw timber, mining timber, poles, pulp etc.

- It is not based on the compartments future productive capacity – it is simply a conglomerate of either expected capacity, or past production. And is therefore not a true reflection of potential future production.
- The method is particularly problematic where it is used on relatively newly afforested areas , quite often the better sites are planted first and more marginal sites later – this leads to a larger proportion of better sites in the older age classes. Using this data can lead to over expectations for the younger more marginal sites.
- The method is even more challenging if production has been measured in tonnes, since the further confounding effect of the conversion factor is introduced.
- It can be subject to manipulation and abuse. Since it is based on a view of the future, the plans produced using this method are biased by this pre-defined view – and reflect the subjective view held rather than a true and unbiased estimate.

The main forest site classification systems in use in South Africa today, are particularly useful for silvicultural practices such as site species matching (Kunz and Pallett 2000; Smith et al. 2005; Louw et al. 2011), however, they are of less direct use for forest planners who need empirical measures such as Site Index to incorporate into the planning systems. Although numerous attempts have been made in the past to relate abiotic elements directly to Site Index (e.g. Grey 1979a, Schafer 1988a, Schafer 1988b, Louw 1997, Louw et al. 2006) these have generally been on specific species and/or on regional or local levels, and/or have required expensive data collection. These studies have tended to have deceasing correlations as the geographic area has increased (Louw and Scholes 2002), and the majority of these studies internationally have used multiple linear regression as the predictive model.

1.7. Discussion

In may ways the use of Site Index as a productivity indicator is a trade-off between the simplicity of a single (understandable) measure, and the associated limitations of the use of a single variable to describe what is in essence the complex effect of an entire ecosystem on tree growth. Due to it's simplicity, Site Index has become a vital component not only of the empirical growth models, but also as a means of separating compartments and research treatments. However the concept comes

with potential problems which forest planners and researchers need to be cognisant of. It is important that forest planners understand the associated limitations of Site Index, that they are aware of the limitations of the datasets used to build the growth models which use Site Index and, in so doing, do not extrapolate beyond the capability of the models.

The methods used to calculate Site Index can either add or subtract from the level of accuracy. The current method of calculating the default Site Index by reverse engineering, is clearly inadequate. There is clearly a need to enable the calculation of the default Site Index using the actual drivers of forest growth, and although there have been previous attempts to do so in South Africa these have been on local levels and on specific species. These studies have also almost exclusively used multiple linear regression as the analytical model.

To quote Vanclay (1994): "*The status of indirect phytocentric methods is so inflated that some speak of direct and indirect methods, not of site productivity estimation, but of site index estimation. This appears to be an unhealthy situation; what began as an interim solution (site index) to a difficult problem (geocentric approach) should not now be called the solution to the original problem.*"

1.8. Thesis objectives

The current forestry site classification systems in use in South Africa generally do not produce estimates of Site Index, or are on a localised species specific level and are expensive with regards to data collection. Forest planners require Site Index as a key input to enable estimates of future production. Site Index comes in two states – measured or direct Site Index, and unmeasured or 'default' Site Index. Direct Site Index can be relatively easily calculated from measurements of dominant height. However direct measurement is not always possible or appropriate. Firstly, and most obviously, when the crop in question is not physically present (i.e. it is yet to be planted, or where the potential for a species which currently does not grow on the site is required), and secondly when the crop is too young to measure. An important question therefore is what is the appropriate age of measurement? This forms the basis of the second objective of this thesis. Prior to this, however, the main assumption associated with Site Index, vis that it is unaffected by initial

planted density, needs to be tested. This forms the basis of the first objective. Finally, since previous studies have generally been done locally and using multiple linear regression, various alternative and novel modelling methodologies have been explored in the third objective.

Two sets of data have been used to explore these issues :

- **Data set 1:** Espacement trial data. Dominant height measurements from 11 trials consisting of five *Eucalyptus* species and seven treatments ranging from 952 stems per hectare to 2222 SPHA.
- **Data set 2:** Temporary, and permanent sample plot (TSP & PSP) data. Measurements of dominant height from 5457 *Eucalyptus*, 4226 *Pinus* and 520 *Acacia* plots, each with 232 site associated predictor variables.

In conclusion this thesis explores the following main objectives (see Figure 6):

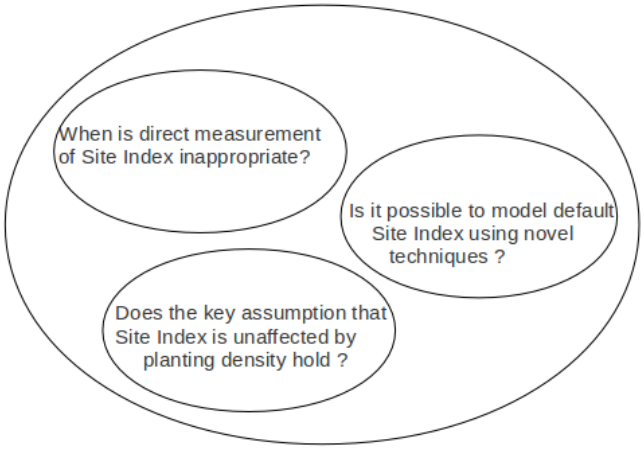


Figure 6: The three separate but related thesis objectives.

Objective 1 : To investigate whether initial planted stems (stand density) has any influence on the estimate of Site Index, using data set one.

Objective 2 : To investigate whether the age at which dominant height is measured has any

influence on the estimate of Site Index, using both data sets one and two, and lastly :

Objective 3 : To model Site Index using readily available climatic and edaphic variables and to investigate various modelling approaches, using data set two. The intention of this objective is not to find a valid model, rather to compare alternative modelling methodologies.

Chapter 2. OBJECTIVE ONE: THE INFLUENCE OF INITIAL PLANTED STEMS ON SITE INDEX

2.1. Introduction

The assumption that dominant height is largely uninfluenced by the initial planting density has been shown to be incorrect in a number of studies – McFarlane, Green and Burkard (2000) found a negative correlation between the initial density and Site Index[9] for 184 *Pinus taeda* stands planted in four geographic locations in the southern United States (Virginia and North Carolina). The stands tested were between 14 and 16 years old, with a Site Index base age of 25, and the initial densities ranged from 747 to 6719 stems per hectare (SPHA). The authors speculate that the (relative) early age of measurement may have some influence on the results, and that the higher density sites may catch up with the lower density sites by the time of the base age. This seems unlikely since the correlation found was strong (significant with a two-tailed t test to $p < 0.0001$), it is also irrelevant if the rotation age is lower than 25 years.

Coetzee (1990) found that the early results of a *Eucalyptus grandis* spacing trial in Zululand showed that espacement had a noticeable effect on height growth (although not statistically significant), the author cautioned that Site Index calculations based on early observations in this case at 3 years should be treated with some care. This is somewhat disconcerting as the commonly used base age for *Eucalyptus* Site Indexes on the Zululand Coast is 5 years, and the rotation age is generally between 5 and 7 years, this means that the majority of enumerations (which are used to calculate Site Index) occur at between 4 and 6 years of age. The author observed that the higher density espacements (i.e. 2222 SPHA) had higher mean heights initially than the lower density espacements (833 SPHA), but that this difference reduced after 18 months. At 18 months the difference between the two treatments was as much as 1.2m, at 36 months this had reduced to 0.65m. Possible explanations for this behaviour were suggested by the author: Wider espacements allow for bigger branches which are retained for longer – this in effect reduces the amount of energy expended on

[9] Calculated from the seven tallest trees at each location.

height growth. In higher density stands the competition for light causes an increase in height growth earlier than in wider espacement stands, however, the larger canopies of the lower density stands eventually become more efficient than the smaller canopies of the higher density stands, resulting in improved height growth.

Van Laar and Bredenkamp (1979) also found that both mean and dominant height were related to the initial SPHA in their analysis of the Langepan *Eucalyptus grandis* correlated curve trend (CCT) spacing trial in Zululand. Bredenkamp (1987) found excellent correlations between the relatively dry Nyalazi CCT and the Langepan CCT for mean diameter (R^2 = 0.98), mean height (R^2 = 0.97) and mean tree volume (R^2 = 0.98), although dominant height was not included in the analysis, the correlations found would imply that the relationship between initial stems and dominant height found at Langepan is true across various sites.

In the analysis of the *P. patula* CCT trial at Mac-Mac Van Laar (1978) found a curvilinear relationship between dominant height and stand density, with the highest dominant height found in the plot with the medium density (755 spha at age 33).

Other research has pointed to little or no effect of initial espacement on height growth (West 2004; Avery & Burkhart 2002; Bernardo et al. 1998; Zumrawi 1986); many of these observations, however, were for sites in the northern hemisphere where the base age for Site Index is as old as 50 or even 100 years – by this time any effect that initial stand density would have had can potentially no longer be measured. Meredieu, Perret and Dreyfus (2003) postulated that stand density effects can change over the stands lifespan, that these effects are more pronounced in younger stands, and that in mature stands the effect is not significant (again this is in long rotation European conifer stands). The authors suggest the use of a correction for stand density effects.

Schönau and Coetzee (1989) concluded in their review of research into the effects of stand density, initial spacing and thinning in *Eucalyptus* plantations, that although the results of various spacing experiments seem contradictory, within the commercial stockings of 1000 to 2000 stems per hectare dominant height does not change but mean height increases with decreasing stand density. They also concluded that this relationship is further affected by species, site quality, and age.

2.2. Objective

The objective of the analysis is to determine whether the initial planted density plays any role in the development of height (specifically dominant height)[10].

2.3. Materials

The data used for this objective comes from 11 ICFR espacement trials laid out between 1994 and 1997, with on average 2 replications per treatment per trial. The trials are geographically spread from as far north as Tzaneen and as far south as Seven Oaks, however, the majority are concentrated in the Zululand area (see Figure 9 and Table 2). The trials consist of one genus (*Eucalyptus*) and five species[11] (*E. dunnii; E. nitens; E. grandis x E. tereticornis (E. g x t); E. grandis x E. urophylla (E. g x u); E. grandis x E. camaldulensis (E. g x c)*). Seven treatments (in this case initial planted stems per hectare (SPHA)) ranging from a low of 952 SPHA to a high of 2222 SPHA are represented. Dominant height has been repeatedly measured in each of the trials, giving an age range from 0.5 years to 13.3 years (see Figure 8). A summary of the data is given in Table 1.

Table 1: Data summary of the espacement trial data set

	Age	TPH0	HDom	SI	TPH
Minimum	0.5	952	0.88	0.43	488
1st Quantile	2.92	1111	12.35	16.32	1111
Median	5.01	1667	19.95	19.12	1466
Mean	5.44	1507	18.93	18.57	1469
3rd Quantile	8.08	1667	25.54	21.41	1667
Maximum	13.33	2222	36.1	30.32	4266

Where:

Age - Measurement age.

TPH0 - initial planted Stems per Hectare (SPHA).

HDom - Dominant height.

SI - Site Index.

TPH - current Stems per Hectare.

Since the trial lifespans essentially all cover the same time period[12], any covariance due to macro

[10] And therefore by definition Site Index.

[11] And hybrid crosses.

[12] April 1994 to February 2008

climatic influences such as long drought or high rainfall periods is probably minimal. Figure 7 below shows the lifespans of the trial data.

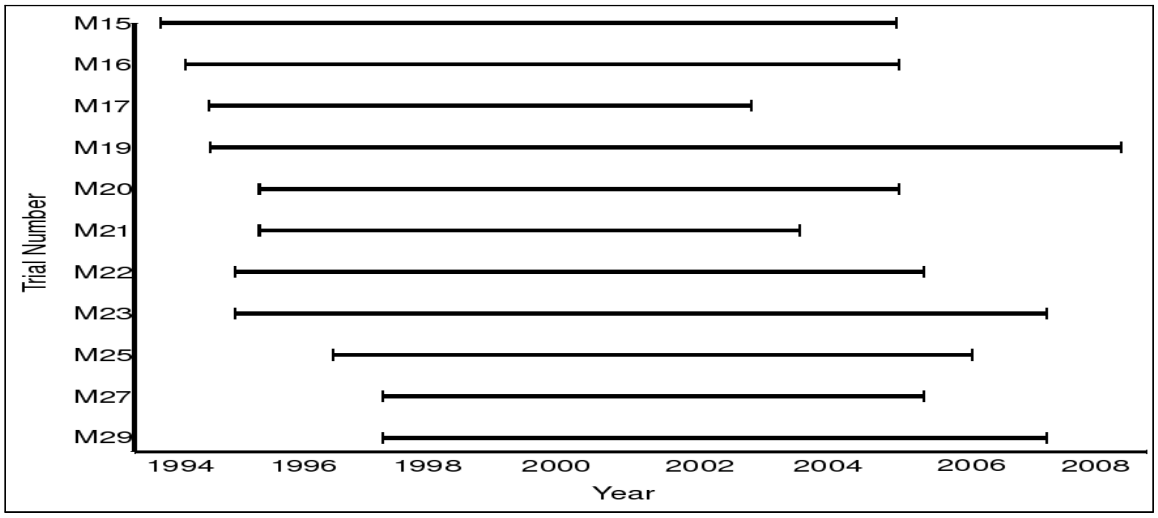


Figure 7: Time period covered by the espacement trial lifespans – lines represent the start and end dates.

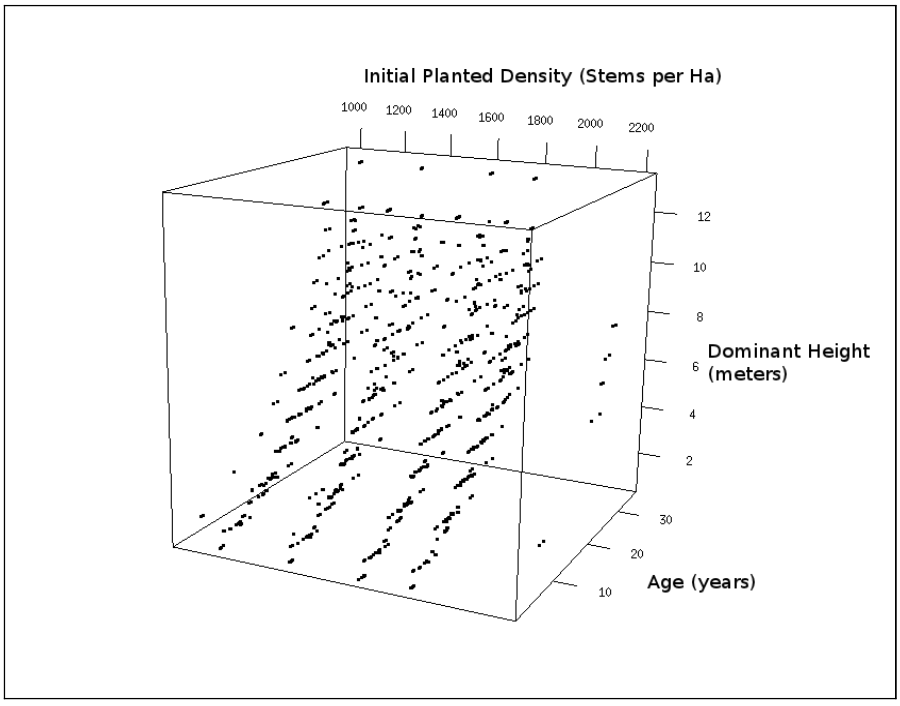


Figure 8: 3D representation of the espacement trial data.

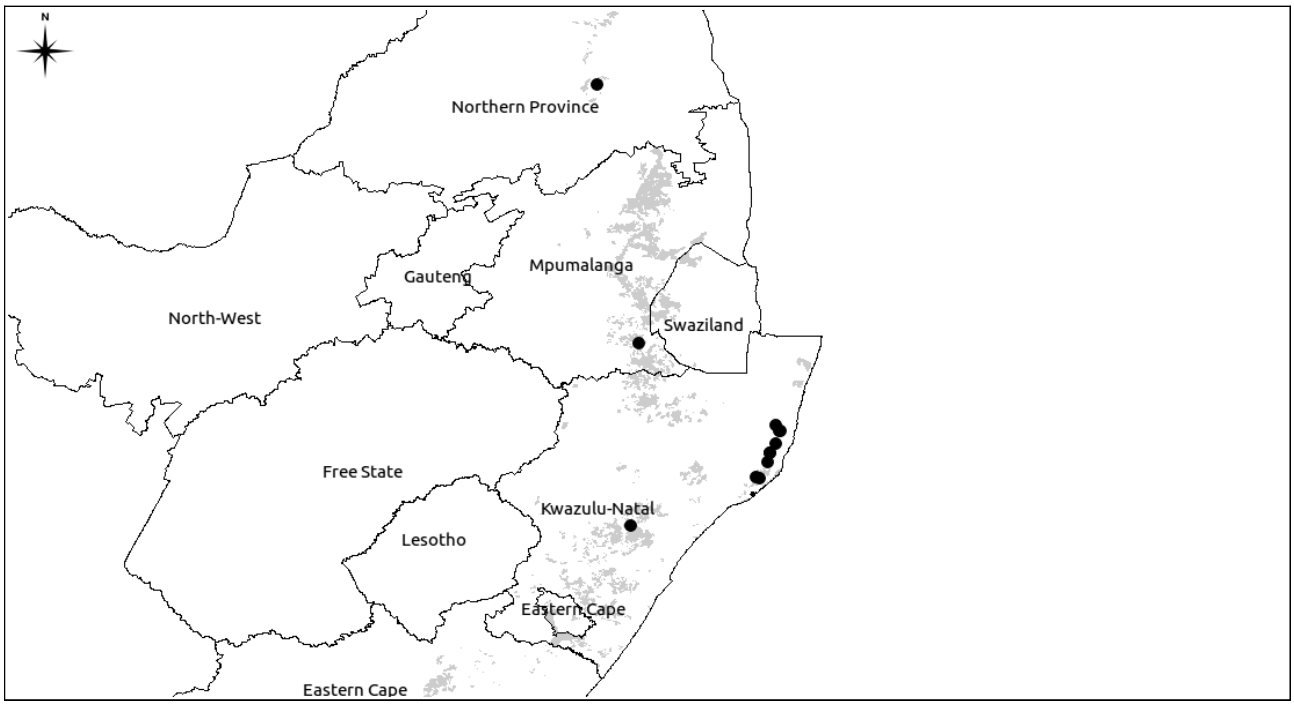


Figure 9: Map showing the geographic distribution of the espacement trial data within South Africa.

Table 2: Co-ordinates of the espacement trials.

TRIAL NAME	PLANTATION	SPECIES	LONGDITUDE	LATITUDE
M15	PALMRIDGE	*E. g x u*	32.26000	-28.30667
M16	NYALAZI	*E. g x t*	32.33400	-28.19700
M17	BUSHLAN	*E. g x t*	32.39000	-28.03000
M19	RIVERBEND	*E. nitens*	30.65900	-26.95100
M20	K.T.	*E. g x u*	32.14250	-28.62667
M21	NYALAZI	*E. g x c*	32.23333	-28.41667
M27	MANAAN	*E. g x c*	30.14300	-23.76600
M29	AMANGWE	*E. g x u*	32.10000	-28.60000
M22	BUSHLANDS	*E. g x c*	32.33100	-27.96000
M23	SEVEN OAKS	*E. dunnii*	30.56300	-29.21000
M25	FUTULULU	*E. g x u*	32.26700	-28.31100

2.3.1. Initial data analysis

The purpose of this initial analysis is to view the data with the intention of understanding the scope, content, and distribution of the data as well as specifically identifying the following :

- Outliers
- Data errors

- Underlying relationships
- Skewed or unusual distributions and relationships
- Potential transformations of the response and predictor variables

A combination of graphical and numerical summaries of the data have been pursued.

2.3.2. Identification of outliers

An initial view of the data in Figure 10 below shows little or no change overall of dominant height

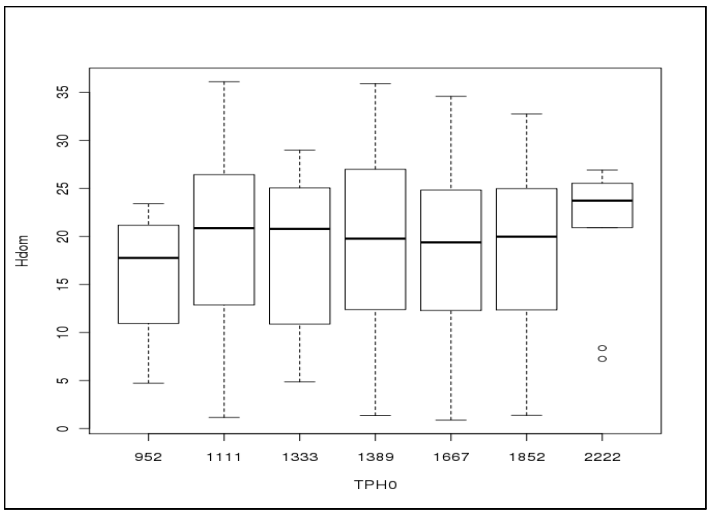


Figure 10: Box plot of Initial stems (`TPH0`) and dominant height (`Hdom`).

over initial planted stems, this may, however, not be a true reflection since the influence of site and age are not included. The figure does show two outliers in the 2222 `TPH0` treatment. These have been left since the treatment has a low sample size in comparison to the other treatments.

Figure 11 below similarly shows little or no interaction between `TPH0` and dominant height over age, however, when the age predictor is log[13] transformed, some interaction is potentially visible (Figure 12).

[13] Natural logarithm.

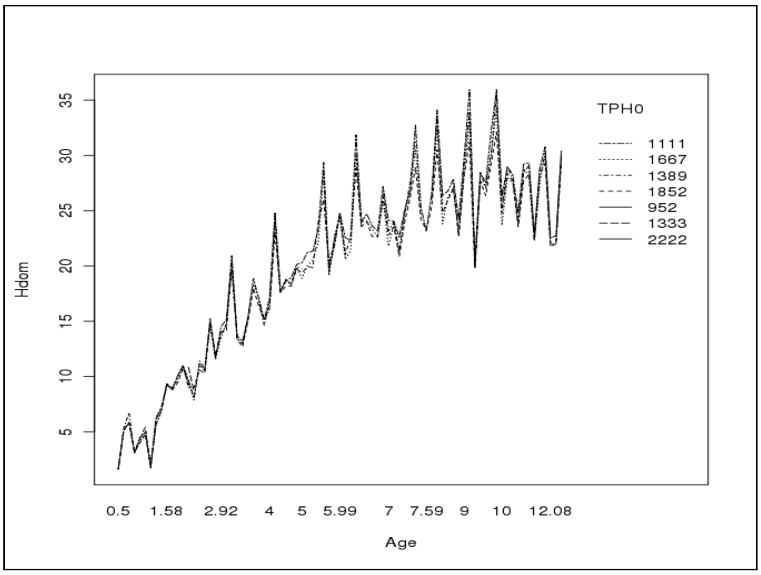


Figure 11: Interaction plot between age and dominant height.

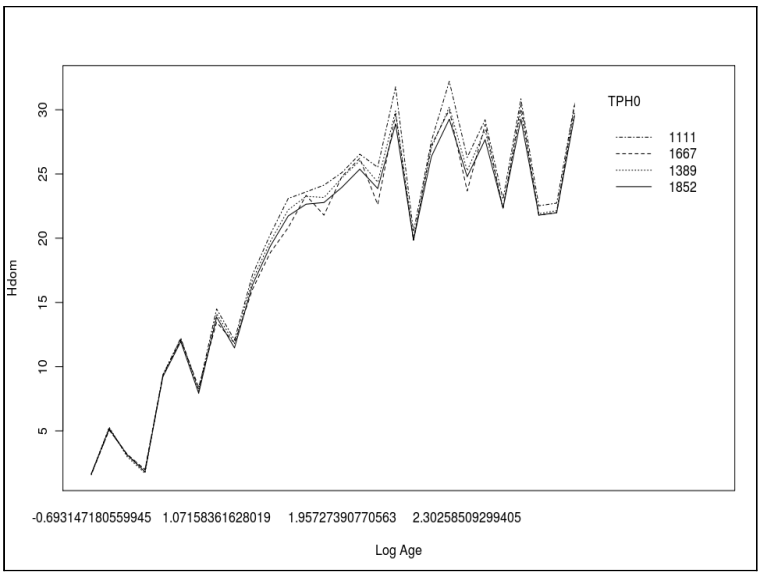


Figure 12: Interaction plot between the natural log of age and dominant height.

From the pairwise plot matrix below (Figure 13) it can be seen that something “odd” has occurred

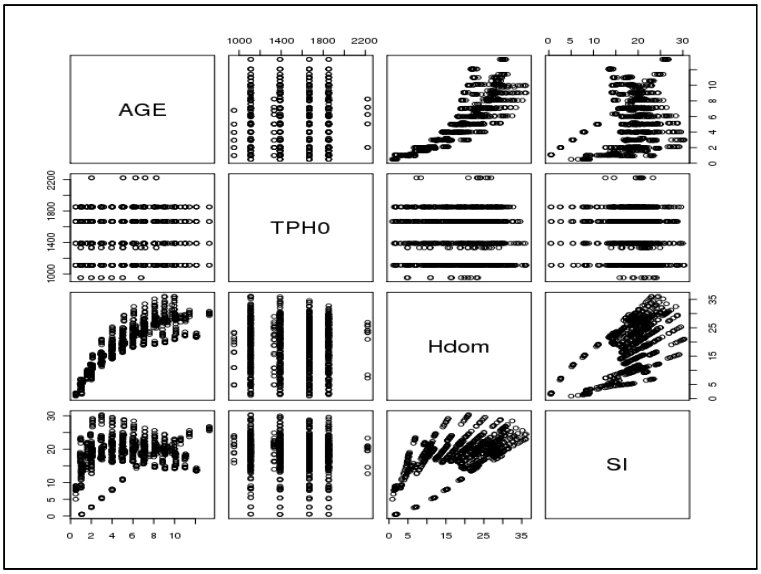


Figure 13: Pairwise plots of the espacement trial data.

between Site Index and age, as well as between dominant height and Site Index – this is likely to be due to errors and or differences in the modelling of Site Index. In order to ensure that Site Index models do not confuse or distort the analysis, dominant height was used as the response variable rather than Site Index.

2.3.3. Transformation

Initial examination of the dominant height and age data suggests that transformation would be advantageous. The following transformations were tested to see what effect they had on the data distribution as well as to linearise the relationship between dominant height and age:

- Transformation of the response (dominant height)
- Transformation of the predictor (age)
- Transformation of both

Only the natural log transformation of the age predictor proved to be useful:

As can be seen from the following figures (14 & 15), the natural log transformation of age has a notable effect on the linearisation of the relationship between dominant height and age across all the initial stems (`TPH0`). What is also noticeable is the increasing variation over age (this may,

however, be the increasing influence of site on dominant height over age).

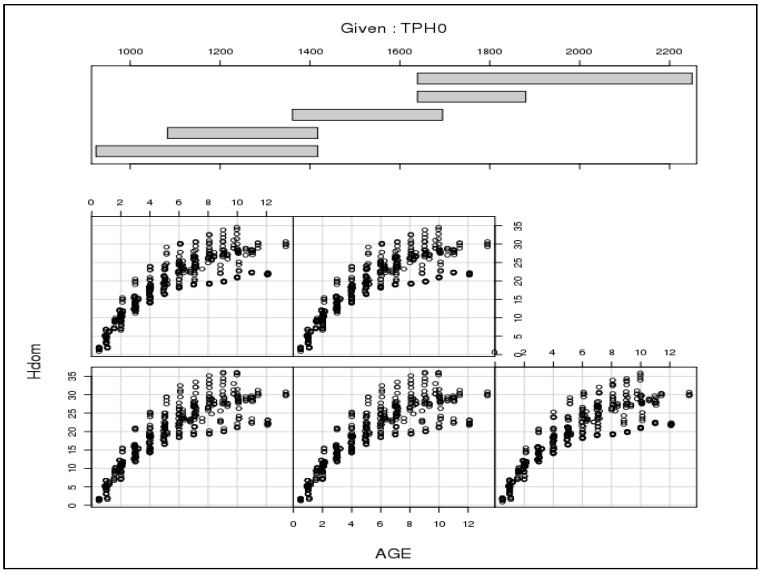


Figure 14: Co-plot of dominant height on `TPH0` and age.

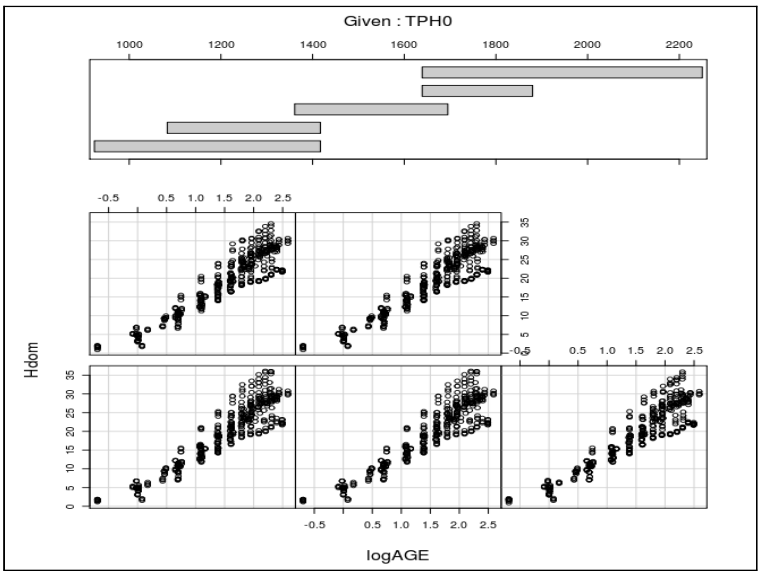


Figure 15: Co-plot of dominant height on `TPH0` and natural log-transformed age (`logAGE`).

2.3.4. Species Differences

From Figure 16 below it is clear that there is some level of species differentiation – this is

particularly noticeable in the case of *E. dunnii.* The analysis therefore needs to take this into account.

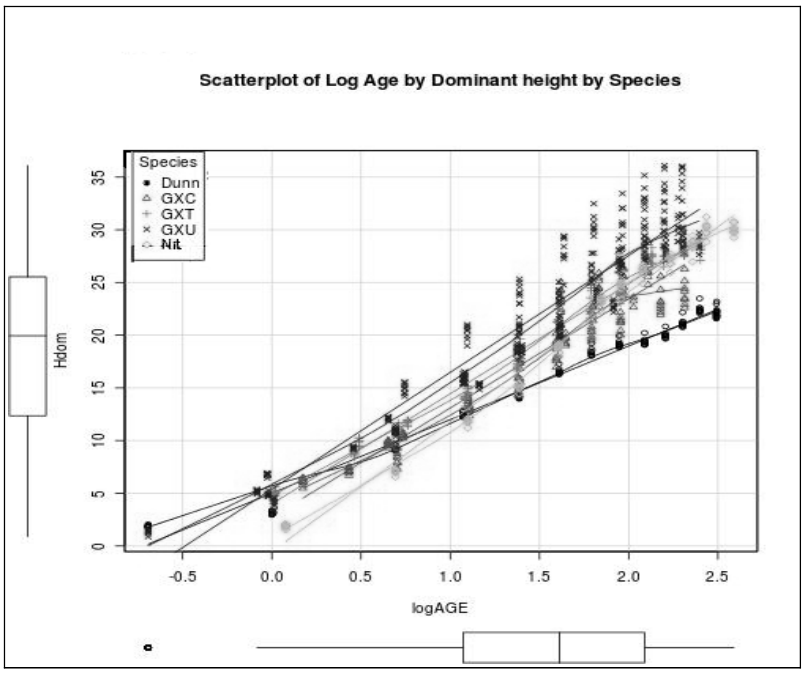


Figure 16: Scatter plot of natural log of age by dominant height, separated by species.

2.3.5. Data treatment

One final issue to determine before analysis, is whether to treat specific variables as categorical or continuous. The response variable (dominant height) will obviously be treated as a continuous variable, and species is obviously a categorical variable, however, age and `TPH0` are not as obvious to class. Age can be treated as a continuous variable since there are numerous age points for each trial, it does however introduce the assumption that dominant height increases smoothly over time – which, from the initial analysis does not appear to be an unreasonable assumption to take. Since there are only 7 initial plot treatments `TPH0` could be treated as a categorical variable, however, since the treatments are not evenly spaced it was decided to treat `TPH0` as a continuous variable.

2.4. Method

As the espacement trial data is longitudinal in nature (i.e. repeated measures over time) it needs to be analysed with this in mind (Bates 2010; Nakai & Ke 2009). Since we have several measurements on the same plot, and two observations from the same plot are likely to be correlated to one another due to factors that are specific to that plot. Maindonald and Braun (2007) give a summary of the approaches that have been traditionally used to analyse repeated measures data – these include:

- Using summary statistics for each subject and then to use these for summary analysis – this is, however, inappropriate here since there is a clear trend of increasing dominant height over time, and summary statistics would therefore not make sense.
- Analysis of variance (ANOVA) can in principle be used when variance is the same (homoscedastic) over all time periods, and the correlation between results is the same for each pair of times. This also implies that the variance of the difference is the same for all pairs of time points. This method allows for the analysis of variance between subjects as well as between times. However the assumptions of equal variance over time are unrealistic.
- Adaptations of the ANOVA method which allow for the potential of heteroscedasticity (unequal variance) between time differences. Maindonald and Braun (2007) state that these should be avoided since good alternatives to ANOVA exist.
- Multivariate models which compare all possible correlations between time points.
- Repeated measures models which aim to reflect the changes over time in the fixed, and random effects as well as in the correlation structure – also called '*mixed effects models*'.

It is this final method which has been applied here.

The theory of repeated measures modelling revolves around the fact that there are at least two levels of variation – *between subjects* and *within subjects*[14] (Maindonald & Braun 2007; Fox 2002).

The name 'mixed effects model' comes from the fact that the models incorporate two types of effects[15]: *fixed* and *random* effects. These names are somewhat misleading since they refer to the properties of the levels of the covariate rather than the effects associated with them:

- If the levels observed represent a random sample of the population of levels (for example the

[14] Plus measurement error

[15] Parameters associated with the levels of the model covariates are sometimes called “effects” .

site) these are referred to as *random effects* .

- If the set of levels is fixed (for example Genus or Species), and or reproducible (such as initial planted stems), the parameters are referred to as *fixed effects*.

Mixed effect models are models which incorporate both of these effects (Bates 2010, Bates 2005).

The linear mixed effects models were fitted using the **lmer** function (within the lme4 package of R (Bates et al. 2011).

The **lmer** function is similar to the normal linear modelling function **lm** with the addition of a random term which identifies the source of the repeated measurements, in this case the plot (Bates 2010; Everitt & Hothorn 2010; Sakar 2008; Bates 2005). According to Fox (2002) this can be specified in the following general mathematical form:

$$y_i = \beta_1 x_{1ij} + \ldots + \beta_p x_{pij} + b_{i1} z_{1ij} + \ldots b_{iq} z_{qij} + \varepsilon_{ij}$$

Equation 1: Linear mixed effects form (Fox 2002)

with the distributional assumptions of

$b_{ik} \sim N\left(0, \psi_k^2\right), Cov\left(b_k, b_{k'}\right) = \psi_{kk'}$, and

$\varepsilon_{ij} \sim N\left(0, \sigma^2 \lambda_{ijj}\right), Cov\left(\varepsilon_{ij}, \varepsilon_{ij'}\right) = \sigma^2 \lambda_{ijj'}$

where:

y_{ij} is the value of the response variable for the *jth* of n_i observations in the *ith* of *M* groups or clusters. In R this is coded as follows :

lmer(response ~ fixed effects + (grouping factor | Random effects), data)

2.4.1. Parameter estimation method

In all cases maximum likelihood (ML) has been used to estimate the parameters of the models since although restricted maximum likelihood (REML) is generally preferred over maximum likelihood and has less bias (Sheather 2009), maximum likelihood is the method for the calculation of *p*-values for specific terms suggested by Bates (2009)[16]. The outcome was verified using REML and produced the same result.

[16] The author of the R package used.

2.4.2. Random effects

To test whether the random effects term (i.e. the site – represented in this case by the `Plotid`) is necessary we first test the following model :

lmer(Hdom ~ 1 + (1|Plotid), data)

The only fixed effect term in the models is a constant of 1. This model allows us to estimate whether the amount of between group variation (i.e. between sites) is sufficient to warrant incorporating it into the model. The model essentially shows the intercept only (which represents the mean level of the response). A standard deviation of zero for the intercept would indicate that the random effects term is not necessary (it does not imply that there is no variation between sites) – and therefore a more traditional linear modelling approach could be used, models such as these can be described as 'degenerate' (Bates 2010). Here each species was modelled separately[17].

Table 3: Results of the intercept only mixed effects model by species.

E. g x c

Formula : Hdom ~ 1 + (1 | Plotid) For : *E. g x c*				
AIC	BIC	lOGlIK	Deviance	REMLdev
1115	1125	-554.5	1109	1109
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.000	0.000	
Residual		43.209	6.5733	
Number of Obs : 168, groups: PlotID , 26				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		17.6215	0.5071	34.75

E. nitens

Formula : Hdom ~ 1 + (1 | Plotid) For : *E. nitens*				
AIC	BIC	lOGlIK	Deviance	REMLdev
701.5	709.2	-347.8	695.5	693.8
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.00	0.0000	
Residual		82.016	9.0563	
Number of Obs : 96, groups: PlotID , 8				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		20.3644	0.9243	22.03

[17] This was attempted using dummy variables, but gave an error since the matrix X'X is not positive definite.

Table 3 continued. E. g x t

Formula : Hdom ~ 1 + (1 | Plotid) For : *E. g x t*				
AIC	BIC	lOGlIK	Deviance	REMLdev
1080	1089	-536.8	1074	1073
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		1.8233e-11	0.00000427	
Residual		5.7040e+01	7.55248084	
Number of Obs : 156, groups: PlotID , 16				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		18.2243	0.6047	30.14

E. g x u

Formula : Hdom ~ 1 + (1 | Plotid) For : *E. g x u*				
AIC	BIC	lOGlIK	Deviance	REMLdev
2394	2405	-1194	2388	2387
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		1.5215e-08	0.00012335	
Residual		8.3125e+01	9.11730801	
Number of Obs :329, groups: PlotID , 34				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		20.8441	0.5027	41.47

E. dunnii

Formula : Hdom ~ 1 + (1 | Plotid) For : *E. dunnii*				
AIC	BIC	lOGlIK	Deviance	REMLdev
919.9	928.6	-456.9	913.9	913.2
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.000	0.0000	
Residual		44.002	6.6334	
Number of Obs : 138, groups: PlotID , 10				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		15.7726	0.5647	27.93

This model has `Plotid` as the only random effect and shows a non zero standard deviation in only the *E. g x t* and *E. g x u* models, which would indicate that the random effects term (`Plotid`) is in fact necessary for these models. Interestingly the models for *E. nitens, E. dunnii* and *E. g x c* all proved to be 'degenerate'. The same results were obtained using restricted maximum likelihood. The reason for this is unclear. Since the outcome is unaffected whether traditional linear modelling or mixed effect modelling is used, all species were analysed using mixed effects.

Additional models where the random effects were grouped using `TPH0` (**lmer(Hdom ~ 1 + (TPH0| Plotid), data)** were also tested, but proved not to add value, and were not significantly different to the first models. The way the random effects term is specified can also add value to the model, however, to keep the analysis as simple as possible this was not pursued. A list of alternative specifications of the random effects term can be found in Appendix 1.

2.4.2.1. Model 1 - The relationship between age and dominant height

The first step in the modelling process is to determine a model for dominant height and age. This first model has one fixed effect parameter, the natural logarithm of age, and one random effects term (the site) generating a simple scalar random effect for each site:

Table 4: Results of the mixed effects Model 1 - dominant height as a function of age, by species.

E. g x c

Formula : Hdom ~ logAGE + (1 | Plotid) For : *E. g x c*				
AIC	BIC	lOGlIK	Deviance	REMLdev
664.7	677.2	-328.4	656.7	658.9
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		1.4267	1.1945	
Residual		2.2808	1.5102	
Number of Obs :168 , groups: PlotID , 26				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		3.2808	0.3892	8.43
LogAGE		10.1212	0.1996	50.72
Correlation of Fixed Effects:				
		(Intercept)		
LogAGE		-0.737		

E. nitens

Formula : Hdom ~ logAGE + (1 | Plotid) For : *E. nitens*				
AIC	BIC	lOGlIK	Deviance	REMLdev
296.1	306.3	-144.0	288.1	292.1
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.061879	0.24875	
Residual		1.128227	1.06218	
Number of Obs : 96, groups: PlotID , 8				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		-0.5284	0.2880	-1.83
LogAGE		12.3115	0.1485	82.93
Correlation of Fixed Effects:				
		(Intercept)		
LogAGE		-0.875		

Table 4 continued. E. g x t

Formula : Hdom ~ logAGE + (1 | Plotid) For : *E. g x t*				
AIC	BIC	lOGlIK	Deviance	REMLdev
491.7	503.9	-241.8	483.7	487.9
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.26672	0.51645	
Residual		1.15317	1.07386	
Number of Obs : 156, groups: PlotID , 16				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		4.0924	0.2270	18.03
LogAGE		10.3174	0.1233	83.66
Correlation of Fixed Effects:				
		(Intercept)		
LogAGE		-0.727		

E. g x u

Formula : Hdom ~ logAGE + (1 | Plotid) For : *E. g x u*				
AIC	BIC	lOGlIK	Deviance	REMLdev
1216	1231	-603.8	1208	1211
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		4.9781	2.2312	
Residual		1.6200	1.2728	
Number of Obs : 329, groups: PlotID , 34				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		5.39922	0.40888	13.21
LogAGE		10.91938	0.09095	120.06
Correlation of Fixed Effects:				
		(Intercept)		
LogAGE		-0.304		

E. dunnii

Formula : Hdom ~ logAGE + (1 | Plotid) For : *E. dunnii*				
AIC	BIC	lOGlIK	Deviance	REMLdev
361.8	373.5	-176.9	353.8	360.4
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		2.6391e-11	5.1373e-06	
Residual		7.6021e-01	8.7190e-01	
Number of Obs : 138, groups: PlotID , 10				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		5.00515	0.14240	35.15
LogAGE		6.98034	0.07879	88.60
Correlation of Fixed Effects:				
		(Intercept)		
LogAGE		-0.853		

2.4.3. Fixed effects

It is now possible to add the additional fixed effect parameter of initial stems per hectare (`TPH0`).

2.4.3.1. Model 2 - Including fixed effects for initial planting density and natural log of age

The model to be fitted is a linear model with fixed effects terms for `TPH0`, and `logAGE`, the random effect associated with the site is a simple additive shift :

lmer (Hdom ~ logAGE + TPH0 + (1|Plotid), data)

Table 5: Results of the mixed effects Model 2 - dominant height as a function of age, including fixed effects for initial planted density, by species.

E. g x c

Formula : Hdom ~ logAGE + TPH0 + (1 | Plotid) For : *E. g x c*				
AIC	BIC	lOGlIK	Deviance	REMLdev
666.7	682.3	-328.4	656.7	671..3
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		1.4267	1.1945	
Residual		2.2808	1.5102	
Number of Obs : 168, groups: PlotID , 26				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		3.287e+00	1.287e+00	2.55
LogAGE		1.012e+01	1.996e-01	50.71
TPH0		-3.799e-06	7.929e-04	0.00
Correlation of Fixed Effects:				
		(Intercept)	LogAGE	
LogAGE		-0.211		
TPH0		-0.953	-0.013	

E.nitens

Formula : Hdom ~ logAGE + TPH0 + (1 | Plotid) For : *E.nitens*				
AIC	BIC	lOGlIK	Deviance	REMLdev
292.9	305.7	-141.4	282.9	301.3
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.0000	0.0000	
Residual		1.1147	1.0558	
Number of Obs : 96, groups: PlotID , 8				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		0.9429974	0.6386121	1.48
LogAGE		12.3115496	0.1475658	83.43
TPH0		-0.0009778	0.0003838	-2.55
Correlation of Fixed Effects:				
		(Intercept)	LogAGE	
LogAGE		-0.392		
TPH0		-0.904	0.000	

Table 5 continued. E.g x t

Formula : Hdom ~ logAGE + TPH0 + (1 | Plotid) For : *E.g x t*				
AIC	BIC	lOGlIK	Deviance	REMLdev
492.4	507.7	241.2	482.4	500
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.23531	0.48509	
Residual		1.15403	1.07426	
Number of Obs : 156, groups: PlotID , 16				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		4.9962906	0.8280171	6.03
LogAGE		10.3249423	0.1232671	83.76
TPH0		-0.0006061	0.0005324	-1.14
Correlation of Fixed Effects:				
		(Intercept)	LogAGE	
LogAGE		-0.184		
TPH0		-0.963	-0.015	

E. g x u

Formula : Hdom ~ logAGE + TPH0 + (1 | Plotid) For : *E. g x u*				
AIC	BIC	lOGlIK	Deviance	REMLdev
1217	1236	-603.6	1207	1222
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		4.9183	2.2177	
Residual		1.6201	1.2728	
Number of Obs : 329, groups: PlotID , 34				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		6.5678953	1.9228302	3.42
LogAGE		10.9198929	0.0909490	120.07
TPH0		-0.0007956	0.0012794	-0.62
Correlation of Fixed Effects:				
		(Intercept)	LogAGE	
LogAGE		-0.058		
TPH0		-0.977	-0.007	

E. dunnii

Formula : Hdom ~ logAGE + TPH0 + (1 | Plotid) For : *E.dunnii*				
AIC	BIC	lOGlIK	Deviance	REMLdev
357.8	372.5	-173.9	347.8	369.0
Random Effects:				
Groups		Variance	Std.Dev.	
Plotid (Intercept)		0.00000	0.00000	
Residual		0.72816	0.85332	
Number of Obs : 138, groups: PlotID , 10				
Fixed Effects:				
		Estimate	STD error	t value
(Intercept)		6.0612896	0.4506241	13.45
LogAGE		6.9789608	0.0771101	90.51
TPH0		-0.0006865	0.0002786	-2.46
Correlation of Fixed Effects:				
		(Intercept)	LogAGE	
LogAGE		-0.271		
TPH0		-0.951	0.007	

2.4.4. Adding interaction terms

Interaction[18] between the fixed effect terms is possible: By way of illustration models specifying interaction can been tested using interaction terms between Species, `logAge` and `TPH0`: (In R interaction terms in the lmer function can be specified in the same way interaction terms are specified in linear models: `a x b` is the same as `a + b + a:b` , `a x b` is equivalent to `b x a`). For example:

lmer(Hdom ~ logAGE + Spp*TPH0 + (1|Plotid), data)

However, since we are not interested at this stage in the true nature of the effect the terms have on dominant height, in other words a predictive model, but simply whether `THP0` has an effect, the addition of interaction terms would simply serve to complicate the analysis.

2.5. Results

There is currently no reliable method to calculate *p*-values in mixed effects models for specific terms – the suggested method is as follows: Fit the mixed effects model including the term using maximum likelihood, fit it again without the term and compare the results using the function anova(). The likelihood ratio statistic will then be compared to a chi-squared distribution to get a *p*-value (Bates 2009). The *t* statistic can also be used to show the significance of each effect : a *t*-value between 2 and -2 implies no significance at a 95 % level (Faraway 2006). Both the *E. nitens* and *E. dunnii* models have *t*-values below -2 implying that `TPHO` may be significant.

We have the following models :

Model 1 : Hdom ~ logAGE + (1 | Plotid)

Model 2 : Hdom ~ logAGE + TPH0 + (1 | Plotid)

In order to find out what effect the TPHO term has had we simply compare the models as follows (Table 6):

[18] e.g. does `TPH0` have the same effect across all species

Table 6: Comparison between Models 1 and 2 by Species.

E.g x c

	Df	AIC	BIC	LogLik	Chisq	Chi Df	Pr(>Chisq)
Model 1	4	664.71	677.20	-328.35			
Model 2	5	666.71	682.33	-328.35	0	1	0.9962

E. nitens

	Df	AIC	BIC	LogLik	Chisq	Chi Df	Pr(>Chisq)
Model 1	4	296.06	306.32	-144.03			
Model 2	5	292.86	305.68	-141.43	5.2012	1	0.02257 *

E. g x t

	Df	AIC	BIC	LogLik	Chisq	Chi Df	Pr(>Chisq)
Model 1	4	491.67	503.87	-241.84			
Model 2	5	492.43	507.68	-241.22	1.2409	1	0.2653

E. g x u

	Df	AIC	BIC	LogLik	Chisq	Chi Df	Pr(>Chisq)
Model 1	4	1215.5	1230.7	-603.77			
Model 2	5	1217.2	1236.1	-603.58	0.3843	1	0.5353

E. dunnii

	Df	AIC	BIC	LogLik	Chisq	Chi Df	Pr(>Chisq)
Model 1	4	361.79	373.50	-176.90			
Model 2	5	357.85	372.49	-173.92	5.9442	1	0.01477 *

Significance. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

When we compare the models excluding `TPH0` we find that the *E. g x c, E. g x u* and *E. g x t* models are not statistically different, indicating that `TPH0` does not have a significant influence on the development of dominant height in these cases. However, for both the *E. nitens* and *E. dunnii* models we find that they are statistically different (to 95 %), implying that there is a significant effect due to initial planted stems. The process followed to estimate a *p*-value is, however, not reliable when the degrees of freedom are small (Bates 2009), in this case the degrees of freedom can be assumed to be sufficient.

Since both the *E. nitens* and *E. dunnii* models are 'degenerate' this suggests they can therefore be analysed using traditional multiple regression analysis[19], this was also pursued (Table 7):

[19] See section 4.5.1 for a description of the methodology

Table 7: Results of multiple regression analysis on *E. dunnii* and *E. nitens.*

E. dunnii

lm(formula = Hdom ~ logAGE + TPH0, data = Dunn)				
Residuals:				
Min	**1Q**	**Median**	**3Q**	**Max**
-2.27858	-0.49648	-0.06738	0.54041	1.95757
Coefficients:				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	6.0612896	0.4556036	13.304	<2e-16 ***
logAGE	6.9789608	0.0779621	89.517	<2e-16 ***
TPH0	-0.0006865	0.0002816	-2.438	0.0161 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Residual standard error: 0.8628 on 135 degrees of freedom

Multiple R-squared: 0.9835, Adjusted R-squared: 0.9832

F-statistic: 4011 on 2 and 135 DF, p-value: < 2.2e-16

E. nitens

lm(formula = Hdom ~ logAGE + TPH0, data = Dunn)				
Residuals:				
Min	**1Q**	**Median**	**3Q**	**Max**
-1.7819	-1.0654	0.1104	1.0257	1.8495
Coefficients:				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.9429974	0.6488305	1.453	0.1495
logAGE	2.3115496	0.1499270	82.117	<2e-16 ***
TPH0	-0.0009778	0.0003899	-2.508	0.0139 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Residual standard error: 1.073 on 93 degrees of freedom

Multiple R-squared: 0.9864, Adjusted R-squared: 0.9861

F-statistic: 3375 on 2 and 93 DF, p-value: < 2.2e-16

In both cases the Akaike information criterion decreases when `TPHO` is removed (the *E. dunnii* model goes from 359.79 to 355.84, and *E. nitens* from 295.14 to 290.86) suggesting `TPH0` is not necessary in the model. However the *E. nitens* model fails the assumption test of normally distributed residuals (p = 0.0001190), and the *E. dunnii* model only barely passes (p = 0.005489) suggesting that these results should be viewed with caution.

2.6. Discussion

The majority of the species tested showed no initial planting density effect (653 observations out of 887 or 73.61%) it may, however, play a significant role in the development of dominant height for some species (notably in this study the cold tolerant *Eucalyptus nitens* and *Eucalyptus dunnii*).

However, the data set used was relatively small, and the two species constituted the smallest subsets within this data. It is also possible that the variability noticed between species, may in fact simply be due to the effect of site (since *E. nitens* and *E. dunnii* are represented by single trials), when the analysis was repeated without the species separation, the conclusion reached was that initial planted density did not play a significant role. Considering the other body of evidence, however, there is some indication that initial planted stems may be significant in some species, and that this may be a subject worthy of further research with a larger data set.

Chapter 3. OBJECTIVE TWO: THE INFLUENCE OF MEASUREMENT AGE ON ESTIMATIONS OF SITE INDEX

3.1. Introduction

Van Laar and Akça (1997) state that the prediction of dominant height in young stands is "uncertain", since the effect of both silviculture (e.g. fertiliser and weed control) and prevailing weather conditions on the young stand are more influential than the productivity of the site. Husch (1956) advocated the use of breast height age as a means of avoiding this early variability. As mentioned earlier, Coetzee (1990) in his analysis of the early results of a *Eucalyptus grandis* spacing trial in Zululand cautioned that Site Index calculations based on early observations (in this case at 3 years) should be treated with some care. Raley et al.(2003)[20] found that early (5 year) height measurements predicted future Site Index (at age 25) relatively poorly ($R^2 = 0.58$), whereas heights closer to the base age (10 years) predicted comparatively well ($R^2 = 0.83$). Johnson et al.(1997) found a similar trend for Douglas-fir *(Pseudotsuga menziesii)* in western Oregon, with lower correlations for early measurements (0.685 at age 7) with height at age 20, versus higher correlations at ages closer to the base age (0.833 at age 10, and 0.948 at age 15).

If early estimates of Site Index are unreliable this would have a direct impact on the inventory and tree breeding assessment policies of most commercial companies, and in particular those with short rotation crops. It would also mean that early measurements should be excluded in the data set used to model Site Index using site variables.

[20] Eleven *Pinus taeda* progeny trials, Western Gulf Forest Tree Improvement cooperative, Texas, USA

3.2. Objective

The objective of this analysis is to determine whether Site Index calculated from early measurements of dominant height are statistically different from those derived from later measurements.

3.3. Materials

The espacement trial data set as described in Section 2.3 has been used. The relationship between measurement age and Site Index reflects the Site Index model used to project dominant height to the relevant base age. Site Index was therefore recalculated using a Chapman-Richards 3-parameter difference form model, with the same coefficient set for all plots[21] and across all species. Thus reducing any effect the various models would have had on the analysis.

$$SI = HD_1 \left[\frac{1-\exp\left(\beta_1 AGE_{SI}\right)}{1-\exp\left(\beta_1 AGE_1\right)} \right]^{\beta_2}$$

Equation 2: Chapman-Richards 3-parameter difference form Site Index model (Fletcher 2010)

Where :

SI = Calculated Site Index

HD_1 = Dominant height measured at AGE_1

AGE_1 = Measurement age

AGE_{SI} = Site Index base age (In this case 8 years)

And coefficients : β_1 β_2

A grouping factor of age in 2 year intervals was added then to the plot data : i.e. A: 0 – 2; B: 2 – 4 ; C: 4 – 6 etc.[22]

The question which arises when viewing the *Eucalyptus* espacement trial data, is whether or not the age at which dominant height is measured has any influence on the accuracy of the estimate of Site

[21] Coefficients supplied by Mondi – these are covered by a confidentiality agreement and can therefore not be published here.

[22] More specifically A: 0 – 1.99 yrs ; B: 2 – 3.99 yrs ; C: 4 – 5.99 yrs etc.

Index. From the graphs below (Figures 17 & 18) it would appear that measurements taken prior to the age of two, underestimate the Site Index.

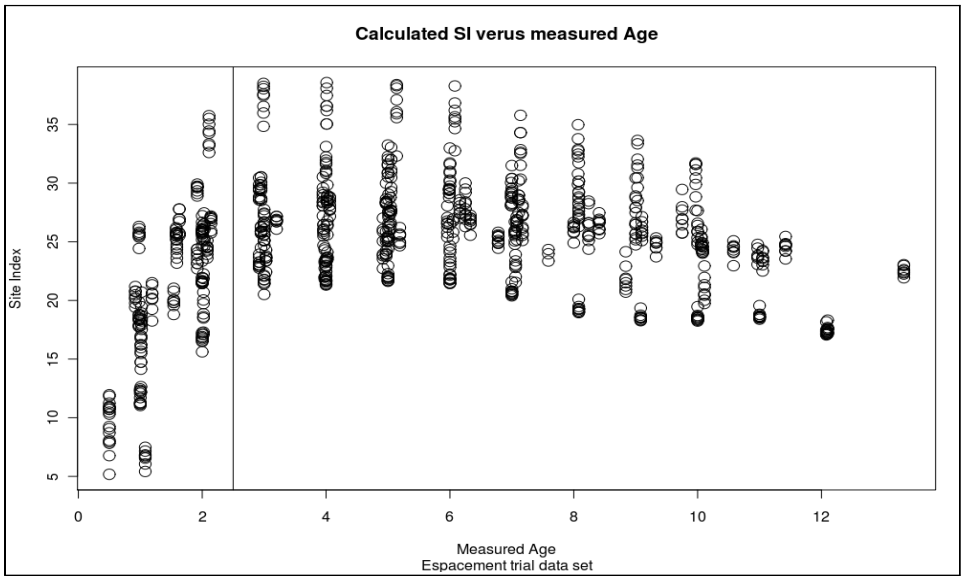


Figure 17: Calculated Site Index versus the age at which dominant height was measured (*Eucalyptus* espacement trial data).

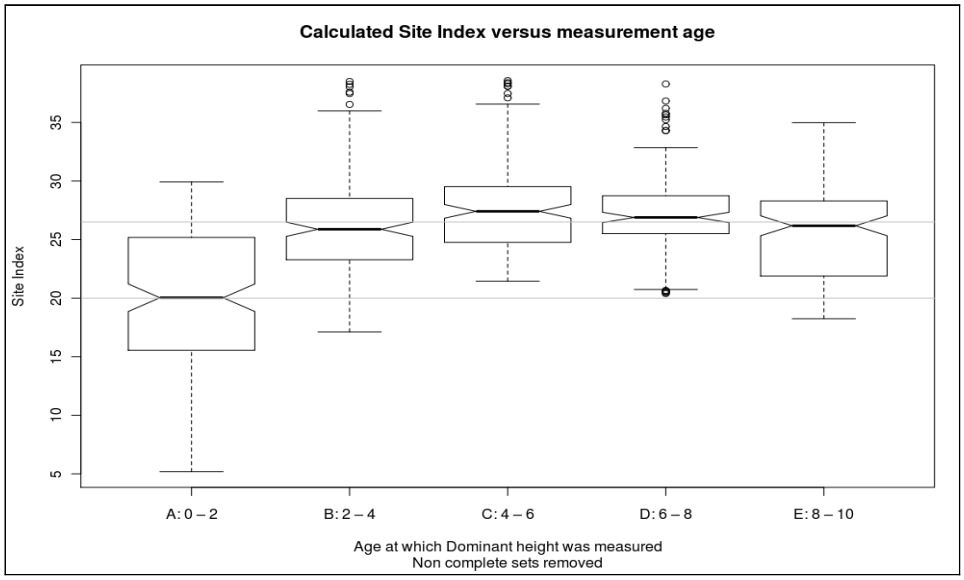


Figure 18: Box plot of calculated Site Index versus the age at which dominant height was measured, for complete sets.

Figure 17 shows that estimates of Site Index before the age of 2 years appear to be lower than those taken at older ages. Figure 18 shows the issue more clearly when the measurements where not all

trials are represented[23], are removed (i.e. above ten years). As can be seen the mean Site Index for the group of espacement trials is estimated at approximately 20 m when measured between 0 and 2 years, whereas the mean estimate for measurements taken at ages over 2 is over 25 m.

3.4. Method

The grouped data can be analysed using one-way analysis of variance for multiple samples with the following hypotheses (adapted from Dalgaard 2008):

Null hypothesis : $(H_0): \bar{x}_1 = \bar{x}_2 = \bar{x}_3 = etc.$

and an alternative hypothesis (H_a): that at least two groups differ.

Where :

$\bar{x}_n$ = mean Site Index of age grouping *n*.

The traditional analysis of variation requires an assumption of normality and equal variances for all groups. The QQ (Quantile versus Quantile) plots of each group (Figure 19) show that the groups are not normally distributed (parametric). QQ plots compare the sample quantities with theoretical quantities from a normal distribution. If the lines are straight, this indicates that the samples are likely to have come from a normal distribution (Dalgaard 2008).

[23] i.e. Incomplete sets.

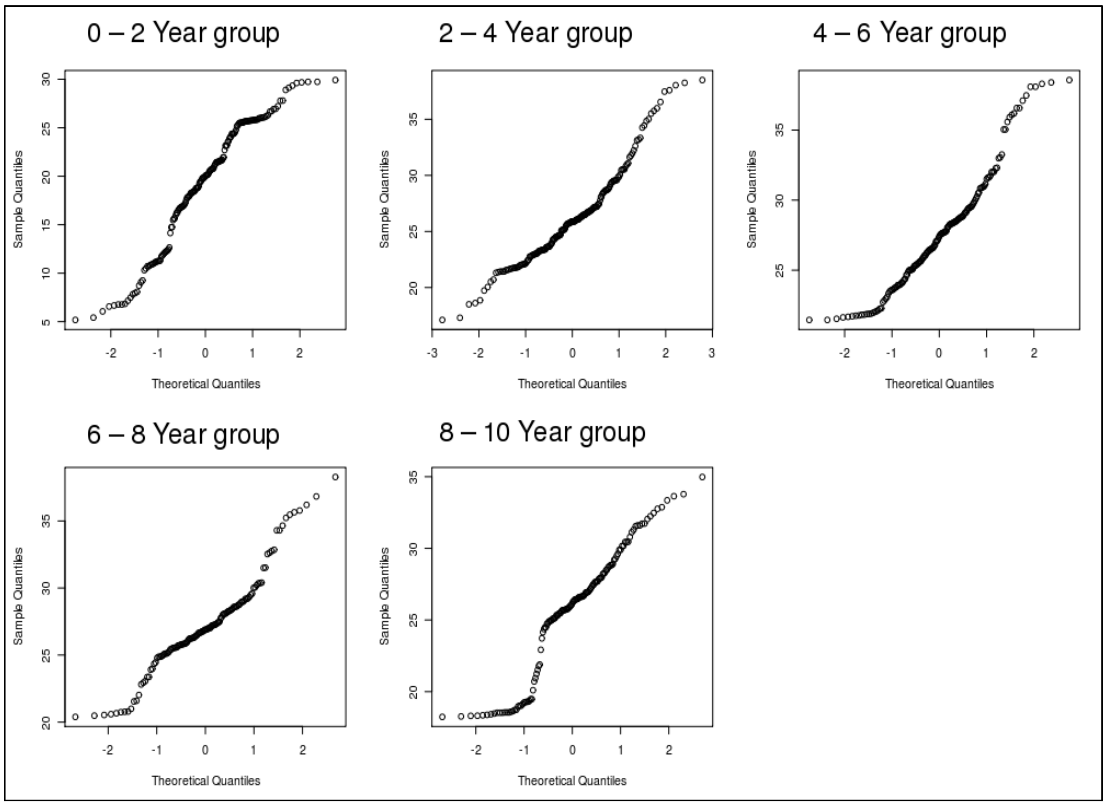


Figure 19: Quantile to Quantile (QQ) plots for the estimated Site Index by age grouping.

This is further confirmed in Table 8 by Shapiro-Wilk tests which show that all groups are non-parametric (to a 95 % level of significance):

Table 8: Shapiro-Wilk normality tests of estimated Site Index by measured age grouping, espacement trial data.

Shapiro-Wilk normality tests		
Age Grouping	**W**	***p*-value**
A: 0 – 2	0.9542	2.67e-005
B: 2 – 4	0.9529	7.63e-006
C: 4 – 6	0.9458	5.21e-006
D: 6 – 8	0.9467	4.87e-005
E: 8 – 10	0.9368	5.33e-006

A more normal transformation of the response can theoretically be obtained using a Box-Cox transformation. These transformations, as originally described by Box and Cox in 1964, take the

following form (Li 2005; Sakia 1992):

$$y(\lambda)=(y^{\lambda}-1)/\lambda$$

Equation 3: Box-Cox transformation (Li 2005; Sakia 1992)

where $\lambda \neq 0$

and

log y if $\lambda = 0$

Figure 20 shows the estimate of lambda (λ) at approximately 1.1. Since the 95 % confidence interval does not include 0 this would indicate that a natural log transformation would not improve the normality.

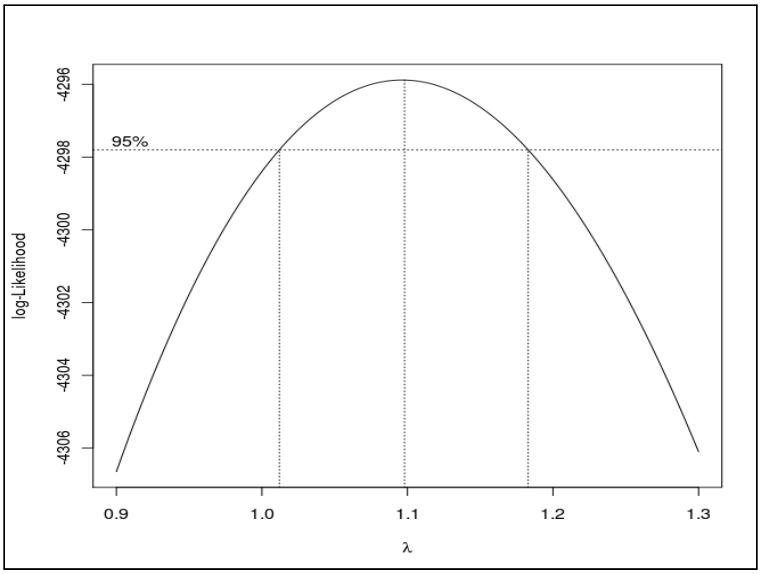


Figure 20: Box-Cox log likelihood lambda of dominant height on age.

As can be seen in Figure 21 the Box-Cox transformation of the response did have a noticeable effect on the between plot variation, however, as expected, it has little or no effect on the linearisation of the data. Since the data is only moderately non-normal, and a further transformation would only serve to complicate the analysis, the Box-Cox transformation was not used. A simple natural log transformation of the age predictor produced a relatively linear relationship, and is the least complex transformation, this was therefore applied.

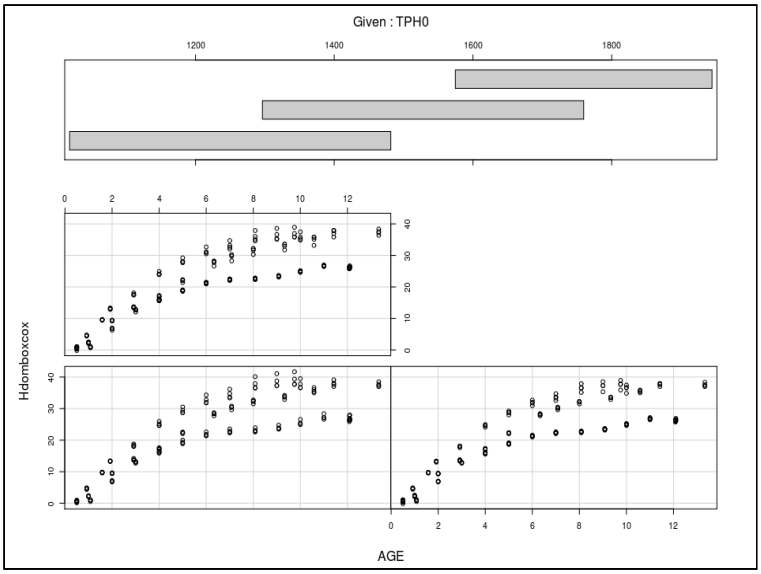


Figure 21: Co-plot of Box-Cox transformed dominant height on `TPH0` and age.

To test the second assumption of homoscedasticity a Bartlett test was performed (Table 9):

Table 9: Bartlett test of homogeneity of variances for the measured age groups, espacement trial data.

Bartlett test of homogeneity of variances		
Bartlett's K-squared	**df**	***p*-value**
65.49	4	2.03e-013

Since the p-value is < 0.05 we can confirm that the data is heteroscedastic.

Since the data is both heteroscedastic, and non-parametric, a traditional ANOVA approach would be inappropriate. The Dunnett modified Tukey-Kramer test, also known as Dunnett's T3 test, as described by Dunnett (1980) would be an appropriate test for this data, if the assumption of normality is eased[24]. This test conducts multiple pairwise comparisons, while adjusting for unequal variance and sample size. Although the test is robust there is a risk of incurring type I (and or type II) errors due to the failed assumption of normality. Using a non-parametric test such as the

[24] Other methods include : Dunn-Sidak; Dunnetts's C; Games-Howell;Scheffe; REGWF; SNK; Tukey's b; Bonferonii; Dunkan; Waller-Dunkan; Tamhane's T2; Welsch; Hochberg; Dayton; etc.

Kruskal-Wallis test runs a similar risk of type I errors if the assumption of homoscedasticity does not hold (as in this case). Cribbie and Keselman (2003) recommend the REGWQ[25] procedure where the data is both non-normal and when variances are unequal. Since the data is only moderately non-parametric both the Dunnett T3 and REGWQ procedures have been used.

3.5. Results

3.5.1. Espacement trial data

The results of Dunnett's T3 test (Table 10) shows that the (A) 0 – 2 age grouping is significantly different to all other age groups to a 99 % level of significance, there is also a significant difference between the (E) 8 to 10 year group and the (C; D) 4 – 8 year groups.

Table 10: Results of Dunnett's T3, on the espacement trial data.

Dunnett's Modified Tukey-Kramer Pairwise Multiple Comparison Test.						**Tukey HSD**
Pair wise comparisons	**Mean Difference**	**Lower Confidence Interval**	**Upper Conference Level**	**H_0 rejected (95 %)**	**H_0 rejected (99 %)**	**Illustrative Adjusted *p*-value**
A: 0 – 2 : B: 2 – 4	-6.996	-5.077	-8.915	*	*	0.00000
A: 0 – 2 : C: 4 – 6	-8.395	-6.454	-10.337	*	*	0.00000
A: 0 – 2 : D: 6 – 8	-7.995	-6.073	-9.918	*	*	0.00000
A: 0 – 2 : E: 8 – 10	-6.321	-4.296	-8.347	*	*	0.00000
B: 2 – 4 : C: 4 – 6	-1.400	0.071	-2.870	*		0.04023
B: 2 – 4 : D: 6 – 8	-1.000	0.446	-2.445			0.32366
B: 2 – 4 : E: 8 – 10	0.674	2.256	-0.907			0.69368
C: 4 – 6 : D: 6 – 8	0.400	1.873	-1.073			0.94730
C: 4 – 6 : E: 8 – 10	2.074	3.680	0.468	*	*	0.00102
D: 6 – 8 : E: 8 – 10	1.674	3.257	0.090	*	*	0.02491

The adjusted *p*-values are from a Tunkey Honestly Significant Difference test (HSD) - and are for illustration only[26]

Since the Site Index base age used is 8 years, one can assume that measurements taken around this age will be more accurate than measurements taken in the outer years. The (E) 8 to 10 year grouping is the least parametric group – this may have had an influence on the outcome of the pairwise comparisons. On the 95 % level of significance, there is also a difference between the (B) 2 to 4, and the (C) 4 to 6 year grouping.

[25] Ryan-Einot-Gabriel-Welsch Q multiple comparison test, based on a stepwise approach.

[26] The DTK.test function in R does not give *p*-values.

The results of the REGWQ test (Table 11) confirm the findings of the first test.

Table 11: Results the REGWQ test, on the espacement trial data.

Ryan – Einot – Gabriel – Welsch Q Pairwise Multiple Comparison Test.			
Pair wise comparisons	***t* statistic**	**Adjusted *p*-value**	**H_0 rejected (95 %)**
A: 0 – 2 : E: 8 – 10	16.80020	0.00000	*
A: 0 – 2 : B: 2 – 4	19.91230	0.00000	*
A: 0 – 2 : C: 4 – 6	23.27600	0.00000	*
A: 0 – 2 : D: 6 – 8	20.91300	0.00000	*
B: 2 – 4 : C: 4 – 6	3.97760	0.01390	*
B: 2 – 4 : D: 6 – 8	2.67250	0.05920	
B: 2 – 4 : E: 8 – 10	1.83310	0.19530	
C: 4 – 6 : D: 6 – 8	1.04510	0.46010	
C: 4 – 6 : E: 8 – 10	5.50410	6.00E-04	*
D: 6 – 8 : E: 8 – 10	4.21040	0.00840	*

3.5.2. PSP and inventory data

Since the permanent sample plot (PSP)[27] data supplied included in many cases multiple measurements of the same plot over time, and the spread of plots is over a wide geographical area it is safe to assume that any influence of site on the analysis can be discounted[28]. This data can be viewed as a random sampling (with replacement) of the population of Site Index estimates. For a complete description of the data see Chapter 4.

The same analytical procedure as outlined with the espacement trial data was then followed :

3.5.2.1. *Eucalyptus* Data

As can be seen from Figure 22, the *Eucalyptus* PSP/TSP data follows a similar, but less well defined pattern to that of the espacement trial data set - where it would appear that the initial estimates taken at young ages are different to those taken closer to the base age. In addition it would appear that the estimates taken in the later years are also different from those around the base age.

[27] And in some cases the TSP data (repeat measurements of the same compartment).

[28] In other words a site - age interaction of some sort.

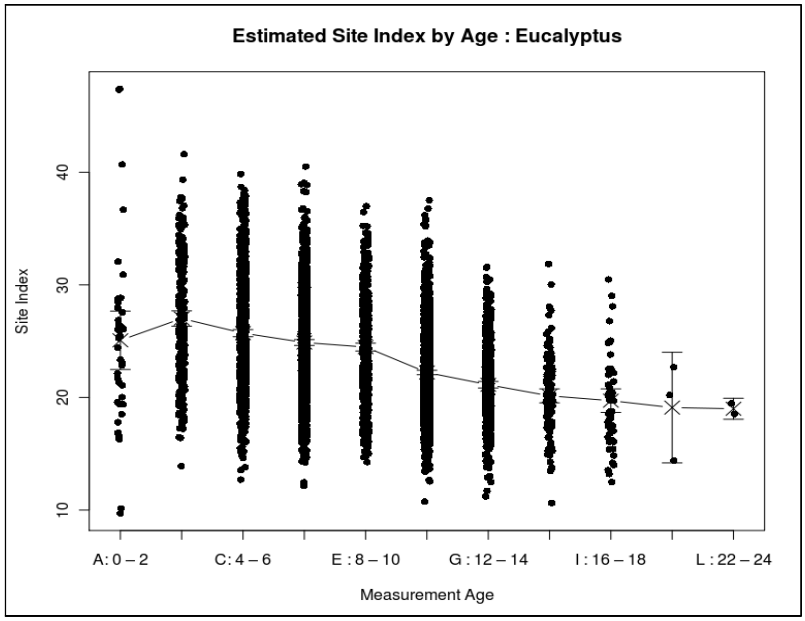


Figure 22: Site Index by grouped age classes for the *Eucalyptus* data.

Shapiro-Wilk tests on the *Eucalyptus* data (Table 12 below) show that groups B,C,D,E,F and G are non-parametric (to a 95 % level of significance).

Table 12: Results of the Shapiro-Wilk's test for normality, *Eucalyptus* data.

Shapiro-Wilk normality tests		
Age Grouping	**W**	***p* - value**
B : 2 - 4	0.98780	0.02927
C : 4 - 6	0.99020	2.289E-006
D : 6 - 8	0.99700	0.01992
E : 8 - 10	0.99340	0.00712
F : 10 - 12	0.98910	9.395E-010
G : 12 - 14	0.99440	0.04284
H : 14 - 16	0.98550	0.18650
I : 16 - 18	0.97220	0.22030
J: 18 - 20	0.94900	0.56480

The Bartlett test (Table 13) confirmed that the data is also heteroscedastic ($p < 0.05$).

Table 13: Bartlett test of homogeneity of variances for the measured age groups, *Eucalyptus* data.

Bartlett test of homogeneity of variances		
Bartlett's K-squared	**df**	***p*-value**
215.94	10	2.20E-016

Interestingly, the results of the Dunnett T3 (Table 14) test shows that almost all groups are significantly different from every other group.

Table 14: Dunnett's T3 test to 95 % level of significance, *Eucalyptus* data.

Dunnett's Modified Tukey-Kramer Pairwise Multiple Comparison Test.					**Tukey HSD**
Pair wise comparisons	**Mean Difference**	**Lower Confidence Interval**	**Upper Conference Level**	**H_0 rejected (95 %)**	**Illustrative Adjusted *p*-value**
B: 2 – 4 : C: 4 - 6	-1.29806	-2.44206	-0.15407	*	0.0006708
B: 2 – 4 : D: 6 - 8	-2.13155	-3.24308	-1.02001	*	0.0000000
B: 2 - 4 : E: 8 – 10	-2.53794	-3.70510	-1.37077	*	0.0000000
B: 2 – 4 : F: 10 - 12	-4.79531	-5.87554	-3.72	*	0.0000000
B: 2 – 4 : G: 12 - 14	-5.88203	-7.01528	-4.74878	*	0.0000000
B: 2 – 4 : H: 14 - 16	-6.87063	-8.30944	-5.43183	*	0.0000000
B: 2 – 4 : I: 16 - 18	-7.29672	-9.18981	-5.4	*	0.0000000
C: 4 – 6 : D: 6 – 8	-0.83349	-1.45719	-0.21	*	0.0001801
C: 4 – 6 : E: 8 – 10	-1.23988	-1.95743	-0.52	*	0.0000006
C: 4 – 6 : F: 10 – 12	-3.49725	-4.06275	-2.93	*	0.0000000
C: 4 – 6 : G: 12 – 14	-4.58397	-5.23945	-3.92849	*	0.0000000
C: 4 – 6 : H: 14 – 16	-5.57257	-6.66114	-4.48401	*	0.0000000
C: 4 – 6 : I: 16 – 18	-5.99866	-7.65594	-4.34137	*	0.0000000
D: 6 – 8 : E: 8 – 10	-0.40639	-1.07132	0.25853		0.5553723
D: 6 – 8 : F: 10 – 12	-2.66376	-3.16081	-2.16672	*	0.0000000
D: 6 – 8 : G: 12 – 14	-3.75048	-4.34737	-3.15	*	0.0000000
D: 6 – 8 : H: 14 – 16	-4.73909	-5.79216	-3.69	*	0.0000000
D: 6 – 8 : I: 16 – 18	5.16517	-6.80063	-3.52972	*	0.0000000
E: 8 – 10 : F: 10 – 12	-2.25737	-2.86769	-1.64705	*	0.0000000
E: 8 – 10 : G: 12 – 14	-3.34409	-4.03894	-2.65	*	0.0000000
E: 8 – 10 : H: 14 – 16	-4.33269	-5.44626	-3.22	*	0.0000000
E: 8 – 10 : I: 16 – 18	-4.75878	-6.43170	-3.08586	*	0.0000000
F: 10 – 12 : G: 12 – 14	-1.08672	-1.62166	-0.55178	*	0.0000134
F: 10 – 12 : H: 14 – 16	-2.07533	-3.09334	-1.05731	*	0.0000059
F: 10 – 12 : I: 16 – 18	-2.50141	-4.11573	-0.88709	*	0.0006713
G: 12 – 14 : H: 14 – 16	-0.98860	-2.05968	0.08247		0.2868638
G: 12 – 14 : I: 16 – 18	-1.41469	-3.06117	0.23		0.2890382
H: 14 – 16 : I: 16 – 18	-0.42609	-2.27289	1.42072		1.000000

The adjusted *p*-values are from a Tunkey Honestly Significant Difference test (HSD) - and are for illustration only

The exceptions being the pairwise comparison between the (D) 6 to 8 and (E) 8 to 10 groups (i.e. around the base age), and the comparisons between the oldest groupings (G, H and I). This result may point to the fact that the species variation within the age groupings of the *Eucalyptus* PSP/TSP data set is potentially too large. Since the espacement trial data represents re-measurements of the same site and species (and at the same time, fewer sites and species), it is probably a more appropriate data set. This is less likely to be the case with the *Pinus* and *Acacia* sets as they represent far fewer species. Again the results of the REGWQ test (Table 15) confirm the results of

Table 15: Results the REGWQ test, on the *Eucalyptus* data.

Ryan – Einot – Gabriel – Welsch Q Pairwise Multiple Comparison Test.			
Pair wise comparisons	***t* statistic**	**Adjusted *p*-value**	**H_0 rejected (95 %)**
B: 2 - 4 : E: 8 – 10	10.9850	0.00000	*
B: 2 – 4 : H: 14 - 16	20.5191	0.00000	*
B: 2 – 4 : I: 16 - 18	15.9789	0.00000	*
B: 2 – 4 : F: 10 - 12	22.8643	0.00000	*
B: 2 – 4 : G: 12 - 14	24.9447	0.00000	*
B: 2 – 4 : C: 4 - 6	5.9595	0.00000	*
B: 2 – 4 : D: 6 - 8	9.9411	0.00000	*
C: 4 – 6 : D: 6 – 8	6.3749	0.00000	*
C: 4 – 6 : E: 8 – 10	7.9218	0.00000	*
C: 4 – 6 : F: 10 – 12	28.4555	0.00000	*
C: 4 – 6 : G: 12 – 14	28.0419	0.00000	*
C: 4 – 6 : H: 14 – 16	19.3151	0.00000	*
C: 4 – 6 : I: 16 – 18	14.1529	0.00000	*
D: 6 – 8 : E: 8 – 10	2.6779	0.05830	
D: 6 – 8 : F: 10 – 12	22.8097	0.00000	*
D: 6 – 8 : G: 12 – 14	23.6001	0.00000	*
D: 6 – 8 : H: 14 – 16	16.5728	0.00000	*
D: 6 – 8 : I: 16 – 18	12.2365	0.00000	*
E: 8 – 10 : F: 10 – 12	15.5623	0.00000	*
E: 8 – 10 : G: 12 – 14	18.5049	0.00000	*
E: 8 – 10 : H: 14 – 16	14.5091	0.00000	*
E: 8 – 10 : I: 16 – 18	11.0466	0.00000	*
F: 10 – 12 : G: 12 – 14	7.1246	0.00000	*
F: 10 – 12 : H: 14 – 16	7.3474	0.00000	*
F: 10 – 12 : I: 16 – 18	5.9593	1.00E-004	*
G: 12 – 14 : H: 14 – 16	3.2700	0.02080	*
G: 12 – 14 : I: 16 – 18	3.2644	0.05470	
H: 14 – 16 : I: 16 – 18	0.8620	0.54220	

the Dunnett T3 test (with the exception of the (G) 12 – 14 ; (H) 14 – 16 comparison which is marginally significant.

3.5.2.2. *Pinus* Data

Figure 23 shows that the *Pinus* data is somewhat different from the *Eucalyptus* data in that there are overestimations of Site Index at younger ages, but that these are still generally within the same range as those estimated at ages closer to the base age. Since the first group (B: 2 - 4) in the *Pinus* data set only contains only one observation, this has been removed for the analysis.

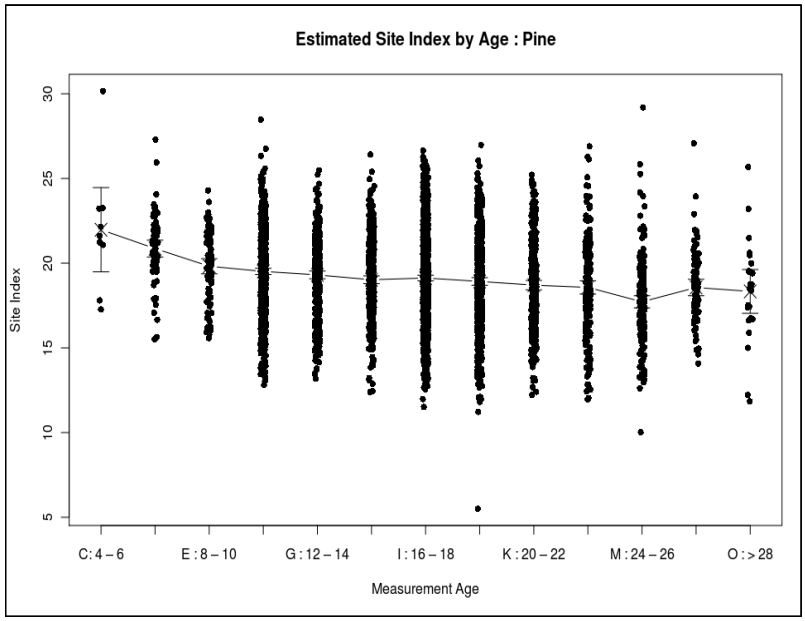


Figure 23: Site Index by grouped age classes for the *Pinus* data.

Again it appears that some of the groupings are non-parametric ($p < 0.05$). The test results (Table 16) show that groups F, G, I, J, M and N are non-parametric (to a 95 % level of significance). And as with the *Eucalyptus* data the *Pinus* data proves to be heteroscedastic (Table 17).

Table 16: Results of the Shapiro-Wilk's test for normality, *Pinus* data.

Shapiro-Wilk normality tests		
Age Grouping	**W**	***p* - value**
B : 2 - 4		
C : 4 - 6	0.87540	0.14040
D : 6 - 8	0.96840	0.07376
E : 8 - 10	0.97420	0.07603
F : 10 - 12	0.99550	0.02791
G : 12 - 14	0.99010	0.00633
H : 14 - 16	0.99710	0.68800
I : 16 - 18	0.99650	0.01603
J : 18 - 20	0.99380	0.01193
K : 20 - 22	0.99500	0.28810
L : 22 - 24	0.99070	0.19960
M : 24 - 26	0.95670	0.00001
N : 26 - 28	0.95940	0.01159
O : > 28	0.96750	0.62850

Table 17: Bartlett test of homogeneity of variances for the measured age groups, *Pinus* data.

Bartlett test of homogeneity of variances		
Bartlett's K-squared	**df**	***p*-value**
59.19	12	3.172E-008

From the results of the Dunnett T3 pairwise comparisons (Table 18) it would appear that the 6 to 8 (D) group is significantly different from all other groups except the 8 to 10 year (E) group. Surprisingly the younger 4 to 6 year grouping (C) is not identified as being significantly different, however, the illustrative *p*-values obtained from a Tunkey HSD test suggest that they may well be different. This result may be due to the size of the group - being the smallest. As well as group D, individual comparisons with the 24 to 26 (M) group are also significantly different - this may in part be due to the variation within this group (i.e. a matter of sample error).

Table 18: Dunnett's T3 test to 95 % level of significance, *Pinus* data.

Dunnett's Modified Tukey-Kramer Pairwise Multiple Comparison Test.					**Tukey HSD**
Pair wise comparisons	**Mean Difference**	**Lower Confidence Interval**	**Upper Conference Level**	**H_0 rejected (95 %)**	**Illustrative Adjusted *p*-value**
C: 4 – 6 D: 6 – 8	-1.11176	-6.70503	4.48151		0.992606
C: 4 – 6 E : 8 – 10	-2.15572	-7.76468	3.45324		0.462743
C: 4 – 6 F : 10 – 12	-2.46165	-7.99217	3.06887		0.189057
C: 4 – 6 G : 12 – 14	-2.66750	-8.20581	2.87082		0.107554
C: 4 – 6 H : 14 - 16	-2.95263	-8.49727	2.59200	*	0.041055
C: 4 – 6 I : 16 – 18	-2.84962	-8.37784	2.67861		0.055451
C: 4 – 6 J : 18 – 20	-3.05269	-8.58549	2.48012	*	0.027077
C: 4 – 6 K : 20 – 22	-3.27068	-8.79860	2.25724	*	0.011989
C: 4 – 6 L : 22 – 24	-3.40362	-8.97502	2.16778	*	0.007700
C: 4 – 6 M : 24 – 26	-4.27292	-9.80178	1.25595	*	0.000107
C: 4 – 6 N : 26 – 28	-3.41111	-8.83951	2.01729	*	0.011722
C: 4 – 6 O : > 28	-3.64203	-8.43482	1.15076	*	0.021625
D: 6 – 8 E : 8 – 10	-1.04396	-2.52661	0.43870		0.366388
D: 6 – 8 F : 10 – 12	-1.34989	-2.51898	-0.18080	*	0.002242
D: 6 – 8 G : 12 – 14	-1.55573	-2.76271	-0.34876	*	0.000262
D: 6 – 8 H : 14 - 16	-1.84087	-3.05238	-0.62936	*	0.000003
D: 6 – 8 I : 16 – 18	-1.73785	-2.89514	-0.58057	*	0.000005
D: 6 – 8 J : 18 – 20	-1.94092	-3.14436	-0.73748	*	0.000000
D: 6 – 8 K : 20 – 22	-2.15892	-3.40986	-0.90797	*	0.000000
D: 6 – 8 L : 22 – 24	-2.29186	-3.68791	-0.89581	*	0.000000
D: 6 – 8 M : 24 – 26	-3.16115	-4.47252	-1.84979	*	0.000000
D: 6 – 8 N : 26 – 28	-2.29935	-3.78091	-0.81779	*	0.000005
D: 6 – 8 O : > 28	-2.53027	-4.89686	-0.16367	*	0.003445
E : 8 – 10 F : 10 – 12	-0.30593	-1.38384	0.77197		0.998092
E : 8 – 10 G : 12 – 14	-0.51178	-1.63067	0.60712		0.902234
E : 8 – 10 H : 14 - 16	-0.79691	-1.92051	0.32669		0.300753
E : 8 – 10 I : 16 – 18	-0.69390	-1.75898	0.37119		0.435456
E : 8 – 10 J : 18 – 20	-0.89697	-2.01221	0.21828		0.112739
E : 8 – 10 K : 20 – 22	-1.11496	-2.28182	0.05190	*	0.018454
E : 8 – 10 L : 22 – 24	-1.24790	-2.56887	0.07307	*	0.009890
E : 8 – 10 M : 24 – 26	-2.11720	-3.34902	-0.88538	*	0.000000
E : 8 – 10 N : 26 – 28	-1.25539	-2.67094	0.16016		0.083569
E : 8 – 10 O : > 28	-1.48631	-3.82703	0.85441		0.411468
F : 10 – 12 G : 12 – 14	-0.20584	-0.85677	0.44508		0.986373
F : 10 – 12 H : 14 - 16	-0.49098	-1.14869	0.16673		0.098483
F : 10 – 12 I : 16 – 18	-0.38796	-0.94115	0.16523		0.085084
F : 10 – 12 J : 18 – 20	-0.59103	-1.23682	0.05475	*	0.002028

The adjusted *p*-values are from a Tunkey Honestly significant Difference test (HSD)- and are for illustration only

Table 18 continued.

Pair wise comparisons	**Mean Difference**	**Lower Confidence Interval**	**Upper Conference Level**	**H_0 rejected (95 %)**	**Illustrative Adjusted *p*-value**
F : 10 – 12 L : 22 – 24	-0.94197	-1.90128	0.01735	*	0.000258
F : 10 – 12 M : 24 – 26	-1.81126	-2.64767	-0.97486	*	0.000000
F : 10 – 12 N : 26 – 28	-0.94946	-2.05880	0.15989		0.087141
F : 10 – 12 O : > 28	-1.18038	-3.41344	1.05268		0.628966
G : 12 – 14 H : 14 - 16	-0.28514	-1.00829	0.43802		0.935118
G : 12 – 14 I : 16 – 18	-0.18212	-0.81159	0.44735		0.992167
G : 12 – 14 J : 18 – 20	-0.38519	-1.09728	0.32690		0.484473
G : 12 – 14 K : 20 – 22	-0.60319	-1.39639	0.19002		0.063186
G : 12 – 14 L : 22 – 24	-0.73612	-1.74115	0.26891	*	0.044181
G : 12 – 14 M : 24 – 26	-1.60542	-2.49347	-0.71737	*	0.000000
G : 12 – 14 N : 26 – 28	-0.74362	-1.89028	0.40305		0.472405
G : 12 – 14 O : > 28	-0.97453	-3.21957	1.27050		0.872147
H : 14 - 16 I : 16 – 18	0.10302	-0.53278	0.73881		0.999976
H : 14 - 16 J : 18 – 20	-0.10005	-0.81773	0.61762		0.999994
H : 14 - 16 K : 20 – 22	-0.31805	-1.11624	0.48014		0.894054
H : 14 - 16 L : 22 – 24	-0.45099	-1.45998	0.55800		0.698778
H : 14 - 16 M : 24 – 26	-1.32029	-2.21277	-0.42780	*	0.000001
H : 14 - 16 N : 26 – 28	-0.45848	-1.60840	0.69144		0.965620
H : 14 - 16 O : > 28	-0.68940	-2.93549	1.55670		0.990804
I : 16 – 18 J : 18 – 20	-0.20307	-0.82720	0.42106		0.944291
I : 16 – 18 K : 20 – 22	-0.42107	-1.13684	0.29470		0.268842
I : 16 – 18 L : 22 – 24	-0.55400	-1.49891	0.39090		0.188563
I : 16 – 18 M : 24 – 26	-1.42330	-2.24328	-0.60332	*	0.000000
I : 16 – 18 N : 26 – 28	-0.56150	-1.65919	0.53620		0.808352
I : 16 – 18 O : > 28	-0.79241	-3.02180	1.43698		0.966500
J : 18 – 20 K : 20 – 22	-0.21800	-1.00749	0.57150		0.988733
J : 18 – 20 L : 22 – 24	-0.35093	-1.35303	0.65116		0.897736
J : 18 – 20 M : 24 – 26	-1.22023	-2.10499	-0.33548	*	0.000001
J : 18 – 20 N : 26 – 28	-0.35843	-1.50268	0.78583		0.994581
J : 18 – 20 O : > 28	-0.58934	-2.83359	1.65491		0.997636
K : 20 – 22 L : 22 – 24	-0.13294	-1.19581	0.92993		0.999995
K : 20 – 22 M : 24 – 26	-1.00224	-1.95473	-0.04974	*	0.001105
K : 20 – 22 N : 26 – 28	-0.14043	-1.33499	1.05413		1.000000
K : 20 – 22 O : > 28	-0.37135	-2.63223	1.88954		0.999983
L : 22 – 24 M : 24 – 26	-0.86930	-2.00193	0.26333	*	0.044022
L : 22 – 24 N : 26 – 28	-0.00749	-1.34243	1.32745		1.000000
L : 22 – 24 O : > 28	-0.23841	-2.54877	2.07195		1.000000
M : 24 – 26 N : 26 – 28	0.86181	-0.39544	2.11906		0.369940
M : 24 – 26 O : > 28	0.63089	-1.65154	2.91332		0.996864
N : 26 – 28 O : > 28	-0.23092	-2.59144	2.12961		1.000000

The results of the REGWQ comparisons (see Appendix 7) are slightly different from the above results, but are also somewhat clearer : the majority of the (C) 4 – 6 year grouping, as well as the (D) 6 - 8 and (M) 24 – 26 groups prove to be significantly different from the other age groups.

3.5.2.3. *Acacia* Data

Figure 24 again shows that early estimates of Site Index appear to be substantially different from those estimates produced closer to the base age, however, in the *Acacia* example there does not appear to be much difference post base age.

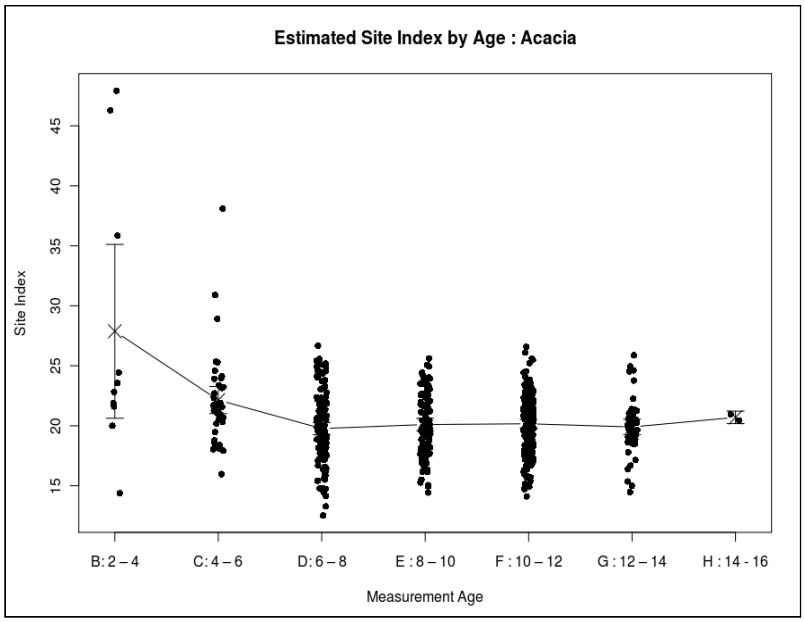


Figure 24: Site Index by grouped age classes for the *Acacia* data.

The results of the Shapiro-Wilk test (Table 19) show that groups B, C and G are non-parametric.

Table 19: Results of the Shapiro-Wilk's test for normality, *Acacia* data.

Shapiro-Wilk normality tests		
Age Grouping	W	***p* - value**
B : 2 - 4	0.82970	0.03315
C : 4 - 6	0.82800	1.260E-005
D : 6 - 8	0.99230	0.70060
E : 8 - 10	0.98500	0.30600
F : 10 - 12	0.98750	0.08733
G : 12 - 14	0.94680	0.02321

Again the data also proves to be heteroscedastic (Table 20).

Table 20: Bartlett test of homogeneity of variances for the measured age groups, *Acacia* data.

Bartlett test of homogeneity of variances		
Bartlett's K-squared	**df**	***p*-value**
59.19	12	3.172E-008

The results of the Dunnett T3 test (Table 21) show that the (C) 4 - 6 age group is significantly different from the 6 to 8, and 12 to 14 age groups (the illustrative *p*-values suggest also that the B, or 2 to 4 group is significantly different from the rest). This again confirms that the younger age estimates are different from those closer to the base age.

Table 21: Dunnett's T3 test to 95 % level of significance, *Acacia* data.

Dunnett's Modified Tukey-Kramer Pairwise Multiple Comparison Test.					**Tukey HSD**
Pair wise comparisons	**Mean Difference**	**Lower Confidence Interval**	**Upper Conference Level**	**H_0 rejected (95 %)**	**Illustrative Adjusted p value**
B: 2 – 4 : C: 4 - 6	-5.731636	-18.740061	7.276788	*	0.000002
B: 2 – 4 : D: 6 - 8	-8.104308	-20.988906	4.780290	*	0.000000
B: 2 – 4 : E: 8 – 10	-7.770745	-20.662953	5.121463	*	0.000000
B: 2 – 4 : F: 10 - 12	-7.706575	-20.570737	5.157587	*	0.000000
B: 2 – 4 : G: 12 - 14	-7.973196	-18.762662	2.816270	*	0.000000
C: 4 – 6 : D: 6 – 8	-2.372671	-4.560746	-0.184597	*	0.000173
C: 4 – 6 : E: 8 – 10	-2.039109	-4.237778	0.159561	*	0.003562
C: 4 – 6 : F: 10 – 12	-1.974939	-4.069264	0.119386	*	0.001893
C: 4 – 6 : G: 12 – 14	-2.241560	-4.166896	-0.316224	*	0.005774
D: 6 – 8 : E: 8 – 10	0.333563	-0.945080	1.612205		0.963882
D: 6 – 8 : F: 10 – 12	0.397733	-0.694246	1.489711		0.864749
D: 6 – 8 : G: 12 – 14	0.131112	-1.082377	1.344600		0.999845
E: 8 – 10 : F: 10 – 12	0.064170	-1.046855	1.175195		0.999980
E: 8 – 10 : G: 12 – 14	-0.202451	-1.427975	1.023073		0.998920
F: 10 – 12 : G: 12 – 14	-0.266621	-1.359450	0.826208		0.993979

The adjusted *p*-values are from a Tunkey Honestly significant Difference test (HSD) - and are for illustration only

The results of the REGWQ comparisons (Table 22) confirm the results of Dunnetts T3 and show that the (B) 2 - 4, and (C) 4 – 6 year age groupings differ from the other age groups.

Table 22: Results the REGWQ test, on the *Acacia* data.

Ryan – Einot – Gabriel – Welsch Q Pairwise Multiple Comparison Test.			
Pair wise comparisons	***t* statistic**	**Adjusted *p*-value**	**H_0 rejected (95 %)**
B: 2 - 4 : E: 8 – 10	10.7790	0.00000	*
B: 2 – 4 : F: 10 - 12	10.9224	0.00000	*
B: 2 – 4 : G: 12 - 14	10.5969	0.00000	*
B: 2 – 4 : C: 4 - 6	7.5203	0.00000	*
B: 2 – 4 : D: 6 - 8	11.3514	0.00000	*
C: 4 – 6 : D: 6 – 8	6.2530	1.00E-004	*
C: 4 – 6 : E: 8 – 10	5.1966	8.00E-004	*
C: 4 – 6 : F: 10 – 12	5.4339	1.00E-004	*
C: 4 – 6 : G: 12 – 14	5.0076	0.00240	*
D: 6 – 8 : E: 8 – 10	1.1591	0.69100	
D: 6 – 8 : F: 10 – 12	1.6113	0.66530	
D: 6 – 8 : G: 12 – 14	0.3647	0.79660	
E: 8 – 10 : F: 10 – 12	0.2409	0.86480	
E: 8 – 10 : G: 12 – 14	0.5426	0.70140	
F: 10 – 12 : G: 12 – 14	0.7784	0.84630	

3.6. Discussion

From the above analysis it is clear that measurements of dominant height taken at early ages should be used with caution when estimating Site Index. This may be a reflection of the projection model used to estimate Site Index – Seifert (2011) has suggested that it is possible that the measurements taken prior to the inflection point in the model do not extrapolate well. This result will have a direct impact on potential inventory policies and should be taken into account. It may be more appropriate to use modelled Site Index values for early ages than to project the measured dominant height.

For the purposes of the next objective (Site Index modelling), the following measurements have been be removed:

Eucalyptus : prior to age 2 (Group A).

Pinus: prior to age 8 (Groups A, B, C and D).

Acacia : prior to age 4 (Groups A and B).

Chapter 4. OBJECTIVE THREE: MODELLING SITE INDEX USING EDAPHIC AND CLIMATIC VARIABLES

4.1. Introduction

There have been numerous studies undertaken to link abiotic site factors to Site Index – almost all of which have used multiple linear regression (MLR) as the statistical model (Kimsey et al. 2008; Wang et al. 2005, McKenney et al. 2003).

Examples of the use of multiple linear regression as a model include:

- Ercanli et al. (2008) developed multiple regression models using edaphic, topographic, nutrient and climatic data, as well as integrated factors to estimate Site Index for oriental spruce *(Picea orientalis)* in Turkey. They found their best model explained 77 % of the variation, but that this model (which had 12 variables) required costly soil nutrient analysis and was not practical on a large scale.
- Louw and Scholes (2006) created two Site Index models using multiple regression for *Pinus patula* in the Mpumalanga escarpment area of South Africa, using site data from 31 plots. A third hybrid model was also produced which combined the second Site Index model with a conventional Chapman-Richards type height projection model. These were then compared to the output from the process based 3-PG model (using MAI_{20} as the comparator). The first model[29] had a R^2 of 0.69 and a mean square error of 1.5588, using effective soil depth, precipitation in the driest quarter of the year, and N mineralisation during the growth season as predictor variables. The second model with an R^2 of 0.74 and mean square error of 1.9767, used topographical position, profile parent material, effective rooting depth and driest quarter precipitation as predictors. The Site Index determined in the second model was then used to calibrated the Chapman-Richards height projection model. The first two models

[29] The first model predicted Site index at a base age of 10, the second model had a base age of 20.

predicted Site Index well, although the second model slightly over-predicted on poorer sites, and under-predicted on good sites. The authors point out that these models are "highly area-specific" and are applicable only to the area of the study.

- Sánchez-Rodriguez et al. (2002) found a regression model that showed Site Index for 47 stands of *Pinus radiata* in North-western Spain, to be positively correlated to foliar concentrations of P, soil pH and depth, and negatively correlated to total nitrogen in the soil. The resultant model explained 60 % of the observed variation.
- Wang and Klinka (1996) used synoptic variables (where various variables are used as explanatory measures of the main variables) of climate, soil moisture, aeration and soil nutrients in regression and limiting factor analysis to explain Site Index in white spruce *(Picea glauca)* in British Columbia Canada. The model explained 90 % of the variance, however, the variables used are not easily obtained or available.
- Louw (1997) also used multiple regression to model the Site Index of *Eucalyptus grandis* in Mpumalanga on attributes of topography, climate, physical and chemical soil properties[30]. He developed a model with soil depth, mean month precipitation for August, soil group He, and organic carbon % as the predictor variables. The model had an adjusted R^2 of 0.802 and root mean square error of 2.22, however, it tended to over predict on poor sites, and underestimate on good sites – this he ascribed to the fact that the Site Index model used to standardise to the reference age was developed for the Zululand coast (based on the Langepan CCT). Again some of the the variables used are not easily or cheaply obtainable.
- Corona et al. (1998) examined the relationship between environmental factors such as temperature, rainfall, soil pH, texture, clay content, altitude, aspect etc. on Site Index in Douglas fir *(Pseudotsuga menziesii)* in central Italy. Their multiple regression model explained 58 % of the variation in Site Index.
- Grey (1979a) used multiple regression on 120 plots of *Pinus patula* in the Glengarry area of the then Transkei. The author found that topographic variables such as altitude, accounted for between 42 and 48% of the variation, and that edaphic factors were poor predictors (Grey 1979b).

[30] Via 96 circular plots with a minimum of 30 trees each, each plot also had a soil pit dug. Chemical and physical soil data as well as climate, topography , parent material and foliar nutrient data was collected for each plot.

Although multiple linear regression is the most commonly used method for this type of modelling, it comes with a number of associated problems:

- The various studies have produced different relationships between the environmental factors and Site Index, and generally the correlations found have been on the low side when the geographic area is large. Both Kimsey et al. (2008) and Wang et al. (2005) suggest that the use of multiple linear regression (MLR) for this sort of study is problematic. MLR relies on the assumption that each independent variable (in this case the abiotic variables) effects the dependent variable (in this case the Site Index) uniformly across the geographic spread of the study area. This may not be the case. Kimsey et al. (2008) give an example of where two separate effects on Site Index may occur. In the theoretical example, MLR assumes that altitude will have the same effect at one point on Site Index as at any other point, however one point may be in a mountainous area which produces very different climatic conditions than another position. MLR is incapable of capturing the effects separately.
- Wang et al. (2005) also state that collinearity often exists between environmental variables, and consequently MLR has a tendency to inflate or deflate at least one of the regression coefficients and therefore the confidence interval of the predictions. Collinearity may even produce coefficients with the incorrect sign (Sheather 1999). Collinearity can also artificially increase the R^2 value, but without adding explanatory value (Seifert 2011).
- MLR also relies on having pre-defined assumptions about the relationship between the independent and dependent variable's (the mathematical form of the model needs to be specified before estimating the parameters). Since this is not known before hand there is a chance of introducing an unknown source of error (Wang et al. 2005).
- If MLR is used, a large amount of field work and analysis is needed to determine which set of site variables are important in explaining the growth/site relationship (van Laar & Akça 1997).
- Lastly, since the understanding of the interaction between the environment and tree growth is not well researched, the often quoted adage "*Correlation does not imply Causation*" may apply in MLR (Saigol 2009).

In order to avoid these issues newer alternative methodologies have been suggested:

- Kimsey et al. (2008) suggested the use of geographically weighted regression which relies

on developing local as opposed to global relationships within the MLR framework. They found that the use of this method improved on the MLR method by capturing an additional 29 % of the variation, they suggest, however, that non-parametric-non-linear models may produce more accurate predictions, but may be more difficult to interpret.

- Curt et al. (2001) compared the use of MLR with variance analysis and concluded that methods such as MLR were less explicative and robust than other methods. They found considerable unexplained variance in Site Index in areas that were classed as uniform by forest managers.
- Wang et al. (2005) suggested the use of non-parametric methods (i.e. no need to specify the mathematical form before estimating the parameters) such as generalised additive models (GAM)[31], neural networks (NNT) and regression trees (TREE). They found that GAM produced the best fit. Although the TREE model also produced an impressive fit they found the TREE model did not produce a smooth map of the Site Indexes since not enough Site Index classes were produced.

Regression trees have a number of advantages over multiple linear regression :

- Regression trees are visual by nature, which allows for simple interpretation of the output, this advantage increases as the number of independent variables and complexity increase (Gehrke et al. 2000).
- Trees require fewer assumptions on the sample data: Trees do not rely on assumptions about the relationship between independent and dependent variables and since few assumptions are made about the model or the data distribution, trees are able to model a much wider range of data distributions (Prasad et al. 2006, De'ath et al. 2000).
- Trees are exploratory as opposed to inferential (Gehrke et al. 2000) which means that a better fundamental understanding of the drivers of the dependent variable can be obtained, and are better at determining interactions between variables (Muller et al. 2008). They are good at revealing structure in data that has hierarchical or non-additive variables (Prasad et al. 2006). Thus avoiding the "*correlation is not causation*" problem.
- Trees are able to provide more detail on the effect of a specific variable. An example is give

[31] GAMs usually rely on spline smoothers, which in the case of Site Index modelling are used for spacial interpolation, data beyond the model region is required in order to avoid problematic over-swing of the spline at the edges in the model region (Seifert 2011).

by McKenney and Pedlar (2003) when they modelled Site Index for black spruce *(Picea mariana)* and jack pine *(Pinus banksiana)* in Ontario Canada using regression trees. The authors found that the thickness of the organic horizon appeared in the first split of one of their models, and again reappeared lower in the tree – this is difficult to produce using normal regression analysis.

- Trees are able to handle missing values in both the dependent and independent variables (Gehrke et al. 2000).
- Irrelevant independent variables are rarely selected (Elith et al. 2008).
- Trees are better at capturing non-additive effects, and interactions (McKenney et al. 2003).
- Trees can be constructed relatively quickly (Gehrke et al. 2000).
- Trees do not require data beyond the area of interest, such as GAM's (Seifert 2011).

There are a number of examples of the use of Classification and Regression Tree (CART) analysis in the forestry and general ecology fields. Many of these examples pertain to the analysis of remote sensing and imagery, particularly within the GIS environment. Recently in South Africa, van Aardt and Norris-Rogers (2008) compared CART analysis to stepwise discriminant analysis in the use of hyperspectral data to discriminate between *Eucalyptus* and *Acacia* species in typical South African plantation forest settings, and to define the age class to which these genera belonged. They found the method promising, but not as accurate as discriminant analysis (72 – 91 % accuracy versus 85 – 97 %). Comparisons by Moisen and Frescino (2002) of five modelling techniques for the automated mapping of forest inventory data in the western United States using satellite based information, also found CART to be less accurate than the other methods tested (Generalised additive models, Multivariate adaptive regression splines, Artificial neural networks and a simple linear model). Other examples of the use of CART in spatial analysis include the use of this technique to integrate forest soils data from multiple sources and differing scales (digital elevation models, remote sensing, digital climatic surfaces etc.). Ryan et al. (2000) used linear models and CART models to generate landscape level forest soil models based on point samples. They found that linear models often produced a simpler and more robust model when single soil properties were modelled, but that they lose their advantage when there are an increasing number of conditional relationships.

Another fairly common use of the tree technique in ecology is in the modelling of species

distributions under climate change scenarios. Guisan and Zimmermann (2000) included classification and regression trees as a modelling technique in their review of predictive habitat distribution modelling, they made no judgement as to which technique was more suitable or accurate, suggesting that the choice of technique depends more on the goal of the study than on the statistical method employed. Iverson and Prasad (2002) used DISTRIB (Deterministic regression tree analysis) to examine the relationship between current forest species distributions in the eastern United States and environmental drivers (climate, soil, land use and elevation variables), and to use this to model future distributions under various climate change scenarios. They found the method valuable in increasing the understanding of species-environment relationships, but limited in explaining many biological attributes under future species distributions accompanying climate change. Bourg et al. (2005) successfully combined classification tree analysis with digital data layers in GIS to predict the potential new habitats of a forest herb (turkeybeard: *Xerophyllum asphodeloides*), the model proved to be relatively accurate (predicting 74 % of the presence areas, and 90 % of the absence areas).

Other examples of the use of CART include work done on modelling tree mortality: Baker et al. (1993) used the technique to develop a model based on soil variables to predict mortality of *Pinus elliottii* and *Pinus taeda* caused by the root rot fungus *Heterobasidion annosum*. They found that the method was apparently accurate (80 %) and useful in improving the disease hazard rating. Dobbertin and Biging (1998) used the method to predict tree mortality for two species (*Pinus ponderosa* and *Abies concolor*). The accuracy of the models was not particularly impressive (11 – 36 %), however, they compared favourably with logistic models, and were able to identify important independent variables which logistic regression did not. Fan et al. (2006) used CART in combination with survival analysis (Kaplan-Meier product limit) which they called CARTBSA, to estimate tree survival rates in oak-dominated forests in Missouri USA, they recommend the method not as a replacement to traditional approaches such as logistic regression but as a complimentary methodology.

Trees have also been used in conjunction with other methods within the field of growth modelling: Raty and Kangas (2008) tested various methods including regression trees to try to localise a generic (national level) volume model by forming homogeneous sub-regions. Localisation was

necessary because the volume model was found to be regionally biased due to stem form. They found that regression trees did not perform as well as the other methods tested (only 50 – 58 %). Klemmt (2007) used a two stage approach applying classification trees to partition inventory data into classes based on observed age-height development. These classes were then used to determine the coefficients of Chapman – Richards growth equations within each leaf via regression, and subsequently used to adjust the SILVA[32] forest growth model. The main advantage of this approach is that actual observed data forms the basis of differentiation for site based parameters.

Regression trees have also been used previously to model Site Index. Mckenney and Pedlar (2003) modelled Site Index spatially against a range of environmental factors from 1140 plots of jack pine *(Pinus banksiana)* and black spruce *(Picea mariana)* in Ontario, Canada. The authors tested several modelling methodologies, but found that a regression tree produced the best results. The pruned tree for jack pine had four splits with five classes and a root mean squared prediction error (root-MSPR) of 2.55 m. The second tree for black spruce had six splits and seven classes with a root-MSPR of 2.84 m.

4.2. Objective

The objective of this section of the study is to explore various methodologies, including regression tree analysis, to model Site Index using readily available climatic and edaphic variables.

4.3. Materials

4.3.1. Summary of the data sources

The data used for this objective has been consolidated from a number of sources (see Figure 25). Firstly PSP and TSP data supplied by Sappi[33] and Mondi[34] (2, 3 & 4 in Figure 25; see also section 2.3.2) which contained the following data :

[32] See www.wwk.forst.wzw.tum.de/research/methods/**model**ling/**silva/**

[33] Kindly supplied by Nico Hattingh of Sappi.

[34] Kindly supplied by Yvonne Fletcher of Mondi.

- Plot number
- Genus and species
- Position (co-ordinates)
- Measured dominant height (m)
- Age at which dominant height was measured

The genus/species, dominant height and age data was then used to calculate a Site Index (see section 4.3.3).

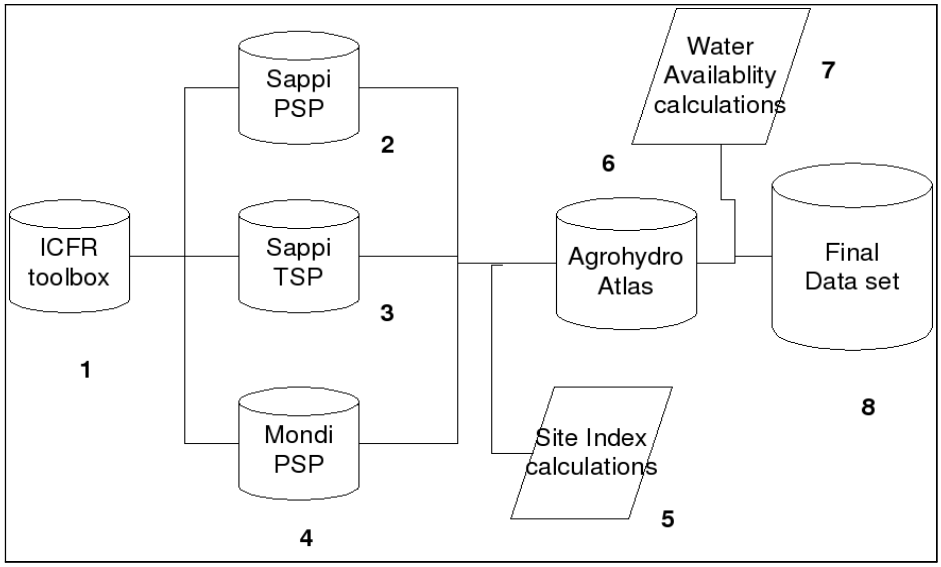


Figure 25: Showing the various steps followed to compile the data set.

The coordinates were then used to populate the following site data via the ICFR Forestry Productivity Toolbox[35] (1 in Figure 25; Table 23) :

[35] Kindly supplied by Trevor Morley of the ICFR. The Forestry Productivity Toolbox was developed by the ICFR to provide data on South African forestry sites and to help with various management options such as site species matching, potential productivity etc. (Kunz 2004).

Table 23: Site variables obtained from the ICFR Forest Productivity Toolbox (Kunz 2004).

Site Variable	Unit
Mean Annual Precipitation or MAP	mm
Probability of obtaining < 650 mm of annual rainfall in any given year	%
Probability of obtaining > 850 mm of annual rainfall in any given year	%
Mean Monthly Precipitation	mm (by month)[36]
Mean Annual Temperature or MAT	^{o}C
Site classification based on climate	CT=Cool temperate; WT=Warm temperate; ST=Sub-tropical
Monthly means of Minimum daily Temperature	^{o}C (by month)
Monthly means of Maximum daily Temperature	^{o}C (by month)
Total Annual Potential: A-pan equivalent Evaporation	mm
Mean Monthly A-pan Evaporation	mm (by month)
Total Annual Solar radiation	MJ/m^2/day
Mean Monthly Solar radiation	MJ/m^2/day(by month)
Topsoil texture-from the 1:250 000 scale land types	~
Total soil depth-from the 1:250 000 scale land types	mm
Permanent Wilting Point of topsoil horizon	mm/m
Field Capacity (Drained Upper Limit) of topsoil horizon	mm/m
Total Porosity of topsoil horizon	mm/m
Geology-from the 1:1 000 000 scale geology map	~
Lithology-from the 1:1 000 000 scale geology map	~
Physiographic region	refer to Kunz & Pallet (2000)
Soil texture derived from parent material	~
Soil depth derived from parent material	mm
Wilting point derived from parent material	mm/m
Field capacity derived from parent material	mm/m
Total porosity derived from parent material	mm/m
Altitude from the 1`x1` of a degree grid	m
Slope derived from the 1`x1` altitude grid	Deg
Aspect derived from the 1`x 1` altitude grid	Deg
Altitude derived from the 1:200/400m altitude grid	m
Slope derived from the 1:200/400m altitude grid	Deg
Aspect derived from the 1:200/400m altitude grid	Deg

A single variable was chosen where variables could be considered synonymous (such as Median monthly precipitation, or Mean monthly rainfall). Given that water availability (or lack there of) is one of the most important drivers of tree growth in South Africa, it is important to ensure that the data included in the model is meaningful. The climatic attributes supplied (Temperature, rainfall etc.) in and of their own may not be good indicators of water availability, or of the supply – demand balance. A monthly water balance (using a generic dry-land crop model) per plot site was therefore

[36] i.e. for each month of the year.

calculated (7 in Figure 25). The water balance was calculated in the following manner (Kunneke 2011):

$$WB = (PAW + EffRAIN) - PET$$

Equation 4: Water balance calculation (Kunneke 2011)

where :

WB = Water balance (monthly)

PAW = Plant available water

EffRain = Effective rainfall (monthly rainfall converted using crop factor)

PET = Potential evapotranspiration

The same plot co-ordinates were also used to populate further site data from the South African Atlas of Agrohydrology and Climatology [37] (6 in Figure 25; Table 24):

Table 24: Site variables obtained from the South African Atlas of Agrohydrology and Climatology (Schulze 1997).

Site variable	**Units**
Altitude 200m	m
Solar Radiation	$MJ.m^{-2}.day^{-1}$ (by month)
Mean Annual Precipitation (2003)	mm
Rainfall Concentration	%
Rainfall Seasonality	Seasons
Means Of Daily Maximum Temperature	^{o}C (by month)
Means Of Daily Minimum Temperature	^{o}C (by month)
Daily Mean Temperature	^{o}C (by month)
Temperature Range (T - T)	^{o}C (by month)
Mean Annual Temperature	^{o}C
Heat Units	^{o}days (by month)
Average First Date of Heavy Frost	Day of year
Average Last Date of Heavy Frost	Day of year
Average Duration of Frost Period	Days
Average Number of Days with Frost	Days
Standard Deviation of Number of Days with Frost	Days
Daily Mean Relative Humidity	% (by month)
Daily Minimum Relative Humidity	% (by month)
Potential Evaporation	mm (by month)
Potential Evaporation Mean Annual	mm

[37] Kindly supplied by Anton Kunneke of the University of Stellenbosch.

Table 24 continued.

Site variable	Units
Potential Evapotranspiration	mm (by month)
Wilting Point top soil - 84 soil zones	mm
Grid Wilting Point top soil - 84 soil zones	mm
Wilting Point sub soil - 84 soil zones	mm
Grid Top soil to sub soil daily drainage fraction	fraction
Grid sub soil daily drainage fraction	fraction
Grid Initial Crop Numbers (Acocks)	ACRU Crop Number
Moisture Growing Season Mean Start of Season	month
Moisture Growing Season Mean End of Season	month
Moisture Growing Season Duration of Season	day
Gross Irrigation Requirements Median Annual	mm

4.3.2. PSP and TSP data

The following is a summary of the PSP and TSP plot data supplied (pre removal of outliers, errors and missing sets):

Table 25: Breakdown of PSP's and TSP's by genus.

Genus	PSP	TSP	Total
Acacia	154	379	533
Eucalyptus	1033	4507	5540
Pinus	445	3862	4307
Total	**1632**	**8748**	**10380**

Table 26: Breakdown of PSP's and TSP's by company and genus.

	Mondi		Sappi		
Genus	**PSP**	**TSP**	**PSP**	**TSP**	**Total**
Acacia	118	0	36	379	533
Eucalyptus	568	0	465	4507	5540
Pinus	112	0	333	3862	4307
Total	**798**	**0**	**834**	**8748**	**10380**

4.3.3. Conversion of dominant height data to Site Index

Dominant height data was provided for the full data set (Mondi and Sappi, PSP and TSP data) and

Site Index was only provided for a small portion of the data set[38], as well as with varying base ages. The dominant height data was therefore converted to Site Index using the following equations (Fletcher 2006 & Fletcher 2010)[39]. This was applied to the complete data set in order to ensure consistency.

- **SICR4 - Chapman Richards 4-parameter – difference form**

$$SI = HD_1 \left(\frac{1-\exp\left(\beta_1\left(AGE_1+\beta_2\right)\right)}{1-\exp\left(\beta_1\left(AGE_{SI}+\beta_2\right)\right)} \right)^{\beta_3}$$

Equation 5: Chapman - Richards 4 parameter difference form Site Index model (Fletcher 2010)

- **SICR2 - Chapman-Richards 2-parameter - difference form**

$$SI = HD_1 \left[\frac{1-\exp\left(\beta_1 AGE_{SI}\right)}{1-\exp\left(\beta_1 AGE_1\right)} \right]$$

Equation 6: Chapman - Richards 2 parameter difference form Site Index model (Fletcher 2010)

- **SICR3 - Chapman-Richards 3-parameter - difference form**

$$SI = HD_1 \left[\frac{1-\exp\left(\beta_1 AGE_{SI}\right)}{1-\exp\left(\beta_1 AGE_1\right)} \right]^{\beta_2}$$

Equation 7: Chapman - Richards 3 parameter difference form Site Index model (Fletcher 2010)

- **SICLJ - Clutter and Jones - difference form**

$$SI = \exp\left[\frac{\ln\left(HD_1\right) - \frac{\beta_2}{AGE_1} + \beta_3}{\exp\left[\beta_1\left(\frac{1}{AGE_1} - \frac{1}{AGE_{SI}}\right)\right]} + \frac{\beta_2}{AGE_{SI}} - \beta_3 \right]$$

Equation 8: Clutter and Jones Difference form Site Index model (Fletcher 2010)

Each plot was assigned a region based on it's geographical position (this was done using Quantum GIS 1.4.0.), see Figure 26 below .

[38] 7.65 % of the data set. A further motivation for conversion was the discovery of some errors in the Site Indexes supplied.

[39] Since the coefficients are covered by a confidentiality agreement these are not published here.

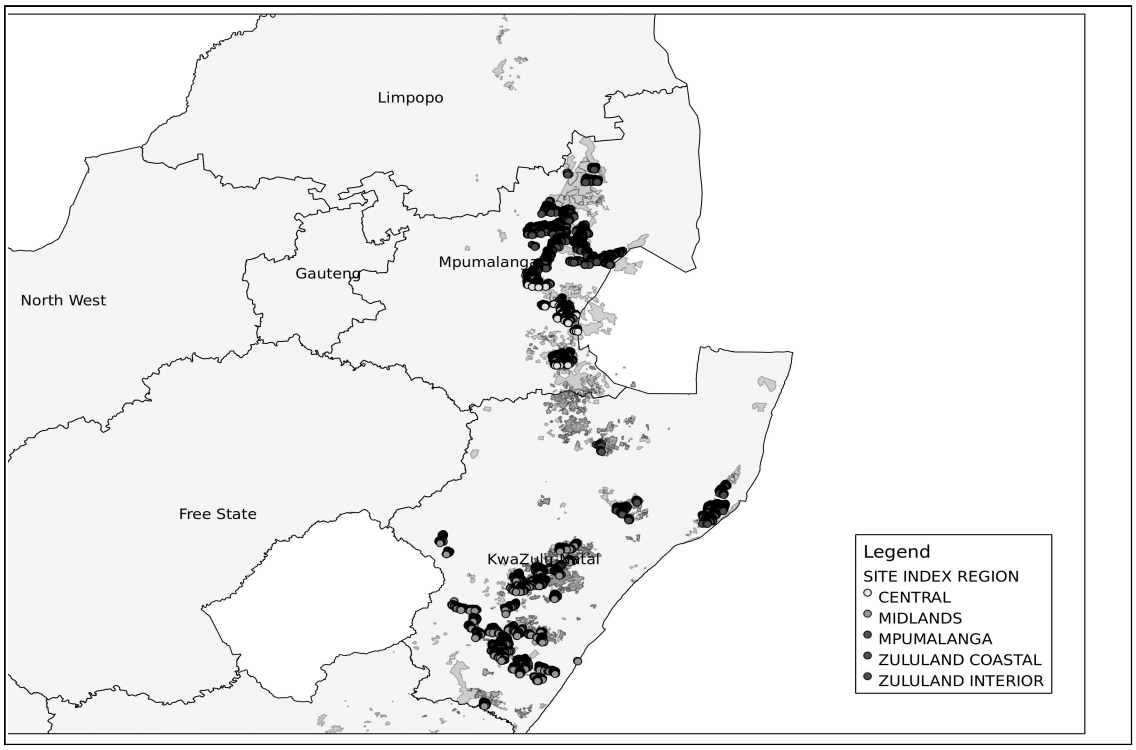


Figure 26: Map showing the assigned regions used to convert dominant height to Site Index.

Each plot was then assigned a "Site Index Species" based on it's current species (choosing the closest similar species based on the species growth characteristics[40]), and the coefficients by species and geographic location were then applied to the relevant model (Table 27).

Table 27: Showing the "Site Index Species" and equations used for conversion. (See Appendix 2 for full species names).

Site Index Species.	Grouped Species					Equation used
AMEA						5
EDUN	ENIT					6
EFAS	EFRA	ECLO	ECAM	EURO		7
EGRA	ESAL	EGXN				7
EGXC						7
EGXU						7
EMAC	EG+M	EMIX	EREG	ERUB		7
ESMI	EELA	EEMA				6
PELL	PCAR	PE+R	PE+T	PECH	PMIX	7
PPAT	PGRE					8
PTAE	PROX	PPSE				6

[40] Based on personal experience.

4.3.4. Site Index base age

This conversion ensured that all Site Indexes are on the same base age, and are comparable. Sappi currently uses the following base ages : **5** for both *Eucalyptus* and *Acacia*, and **15** for *Pinus* (Hattingh 2010), Mondi uses **10** for *Acacia*, **8** for *Eucalyptus* and **15** for *Pinus* (Fletcher 2010). The Mondi base ages were used since they are closer to the normal felling ages for each of the genera. Thus the data obtained was recalculated to fit these base ages.

4.3.5. Comparison between supplied Site Index and calculated Site Index

From the following graphs (Figures 27 and 28) it can be seen that, barring a few outliers[41] the calculated Site Indexes match well with those supplied.

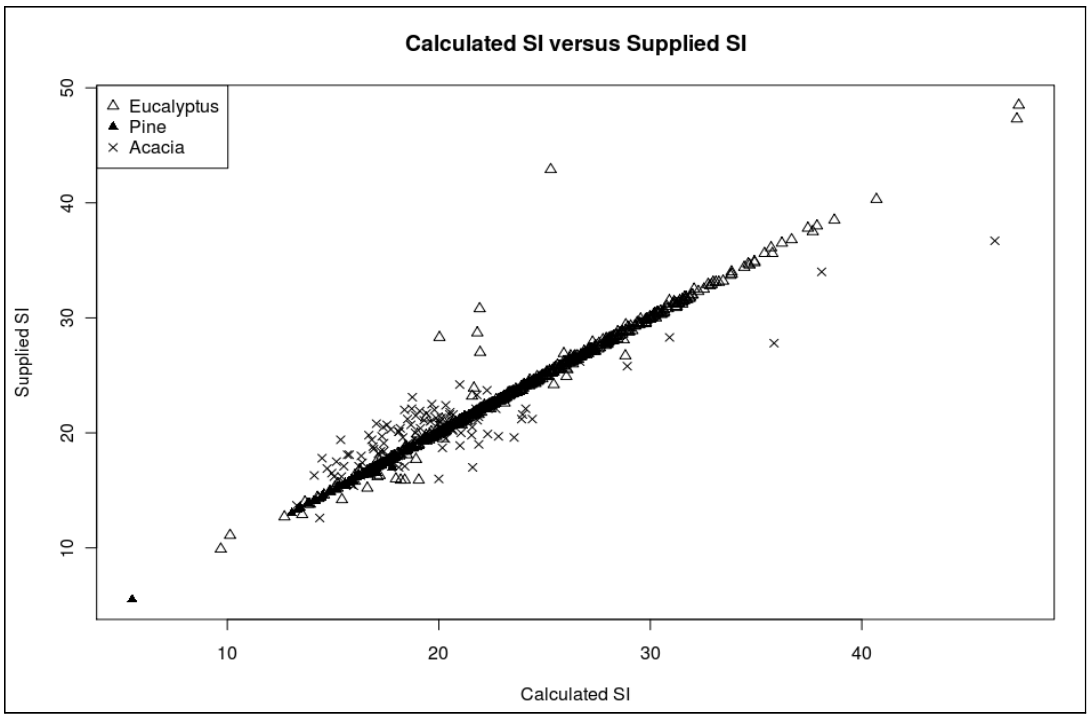


Figure 27: Showing the calculated Site Index versus the supplied Site Index.

[41] These can be ascribed in some cases to errors in the supplied Site Indexes .

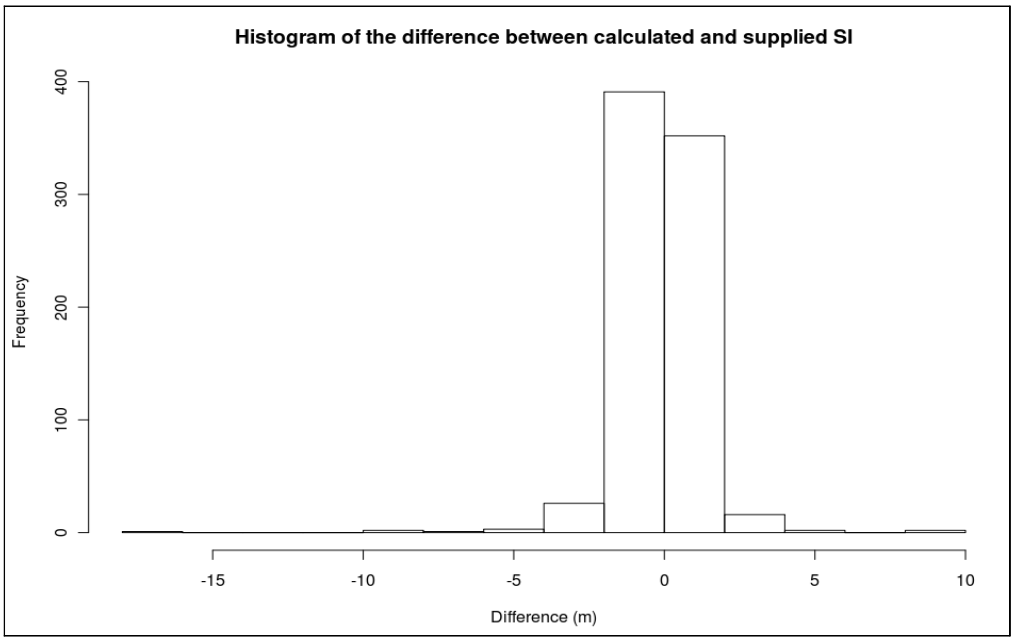


Figure 28: Histogram of the differences between calculated and supplied Site Indexes.

Two sided paired t-tests were used to confirm that there was no statistical difference between the supplied and calculated data sets (Table 28).

Table 28: Results of the paired two sided t tests between supplied and calculated Site Index.

	99 % confidence interval:					
Genus	**Mean**	**Lower**	**Upper**	**t**	**df**	**p-value**
Eucalyptus	-0.07035	-0.18166	0.04096	-1.6335	567	0.1029
Pinus	0.01714	-0.01283	0.04711	1.4991	111	0.1367
Acacia	-0.35534	-0.90175	0.19106	-1.7034	115	0.0912

In all cases the *p*-value was greater than 0.05, we can therefore not reject the null hypothesis that the difference in population means is equal to zero. The assumption of normality of the differences was confirmed (via a Shapiro-Wilk's test for normality).

4.3.6. Removal of observations from the data set

Obvious errors (for example dominant height of 237 meters), as well as plots with missing height, site, or age data were removed from the data set. Any other missing data values were assigned a value of NA. As a result of the analysis in Chapter 3 all plots measured below the ages specified

were removed.

Unfortunately, both stepwise model selection for multiple regression, and the random forest scripts in R are unable to handle categorical predictor variables with more than 32 categories. One such variable exists in the data set: Geological Type. The two categories with the lowest number of observations were therefore removed so that the methodologies could be compared on the same data set: Geological type Vg, with 3 observations, and Zt with one observation.

4.3.7. Data summaries

A summary of the full data set used can be found in Appendix 8. In total 232 predictor variables were considered for modelling[42]. There were a total of 5457 *Eucalyptus* observations, 4226 *Pinus* and 520 *Acacia* observations.

4.4. Method

For reasons of clarity only the results of the *Eucalyptus* modelling are given in the main text – the results of the *Pinus* and *Acacia* regression tree models can be found in Appendix 9.

4.4.1. Classification and Regression trees

Trees are graphical representations of the data – with the root node representing the complete, un-partitioned data set, and the branches and leaf (or terminal) nodes below (De'ath et al. 2000). They do not necessarily have to be binary (nodes do not have to split into two sub-nodes, single nodes are also possible), but most are (Wilkinson 1992). They are constructed by repeatedly splitting subsets of the data using all the independent variables to create child nodes, starting with the full data set or root node (Ture et al. 2009).

Figure 29 shows a stylised diagram of a regression tree – at each node of the tree a split criterion is located (this is where the data can be separated into two significantly different sub-sets): if this split

[42] See Appendix 4.

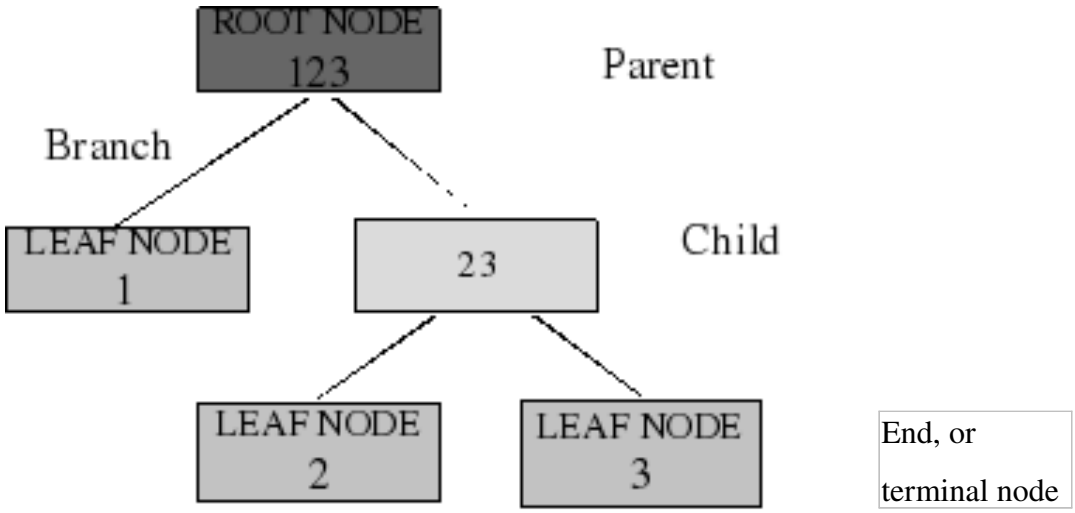


Figure 29: Showing the root and leaf nodes – the numbers 1, 2, 3 represent significant data subsets. (Adapted from various sources – van Diepen et al. 2006; Gehrke et al. 2000; Wilkinson 1992)

criterion is met then the observations are sent to the left branch, if not, then the right. This can then be easily translated into the **IF THEN ELSE**[43] statement form.

Trees explain variation of a dependent (response) variable by one or more independent (explanatory) variables. Where the dependent variable is numeric or continuous a **regression tree** is used, and where the dependent variable is categorical (named, nominal or ordinal) a **classification tree** is used (De'ath et al. 2000). In other words a regression tree is used if the dependent variable is quantitative, and a classification tree if it is qualitative (Moisen et al. 2002). The independent variables in both cases can be either type. The difference between the two trees is essentially in the method of splitting and classification of the nodes (De'ath et al. 2000). In this thesis only regression tree analysis has been used.

There are three basic steps in the tree methodology (Muller and Mockel 2008):

- The data set is split into two subsets using the most effective predictor (splitting rules).

[43] **IF** the criterion is met **THEN** follow the left tree branch **ELSE** follow the right. A simplistic example: where Site Index is split into two sets : 10m or 20m with the split on soil depth > or < 1.5 meters. so : **IF** soil depth >1.5m **THEN** SI=20 **ELSE** SI=10.

- This splitting is repeated within the subgroups until the sub-groups become too small, or there are no further splits (stopping rules).
- Results are displayed in a binary tree structure, and pruning takes place if necessary.

Everitt and Hothorn (2010) give the following mathematical description of the method:
In the initial step a covariate x_j is selected from the available covariates $x_1, \ldots, x_q$ and a split point is estimated which splits the response values y_i into two groups. For a nominal covariate x_j the two groups are defined by a set of levels A where either $x_i \in A$ or $x_i \notin A$. And for an ordered covariate x_j the split point is a number ξ which divides the two groups so that the first group contains the observations $x_j \geq \xi$ and the second $x_j < \xi$. Once the splits have been estimated this is then repeated recursively on each sub-group until a stopping criterion is reached. The various Tree building available algorithms differ in terms of covariate selection, the estimation of the split point and on the stopping criteria.

The algorithm used in this thesis is that implemented in the **rpart** package of R (Therneau et al. 2010) using the "anova" method, which looks at all possible splits for all of the covariates, and chooses the split based on the following splitting criteria (Therneau et al. 2011):

$$SS_T - (SS_L + SS_R)$$

Equation 9: Splitting criteria for anova based regression tree (Therneau et al. 2011)

where

$SS_T = (y_{1i} - \bar{y})^2$ is the sum of squares for the node, and

SS_L, SS_R the sum of squares of the left and right child node, in essence maximising the between-group sum of squares.

4.4.1.1. Stopping rules

Stopping rules control when (or whether) the tree growing process should be stopped or not. Tree quality depends more on the quality of the stopping rules than on the splitting rules (Murthy 1998). Some of the rules used include (Breiman et al. 1984):

- If all cases in a node have the same value for the dependent variable.
- If all cases in a node have the same independent variable value.

- If the user has specified a maximum tree depth (i.e. number of splits from the root) and the tree has reached this value.
- If the user has specified a minimum node size and the next node is less than this value.
- If the best split for the node contributes less than a user specified improvement to deviation.

4.4.1.2. Pruning and cross-validation

Due to the nature of the method, trees will continue to split the data until there are no further valid splits, or until one of the stopping rules becomes valid. However, the final number of splits may be unnecessary (i.e. over-fitted). Where the variation can be explained with fewer leaf (terminal) nodes, and the additional nodes do not add to the improvement of prediction, it is important to have some method of reducing the number of nodes to an optimal size (Murthy 1998). It is also important to allow the tree to grow fully before this is done (i.e. not to overly restrict the splits in the initial tree) since it is possible that significant variables could be missed if this is done. Once the tree building has stopped pruning methods can be employed to ideally trim back the tree to the optimal size (Wilkinson 1992; Breiman et al. 1984). The size of the tree is controlled by the best trade-off between explained deviance or variation and the tree size (Thuiller et al. 2003).

'V-fold' cross-validation is used to calculate the error rates used in the pruning (see Section 4.4.3). A large tree is grown from the full data set, this is the tree that will be pruned back. The full data set is then divided into roughly-equal parts, each containing a similar distribution for the dependent variable. The next step is to construct the largest possible tree from the set, less one of the parts, and use the remaining data set to obtain initial estimates of the error rate of selected sub-trees. The same process is then repeated each time using a different data subset as a test sample. The process continues until each part of the data has been used as a test sample. The results of the test samples are then combined to form error rates for trees of each possible size, these are applied to the tree based on the entire sample (Breiman et al. 1984). The rpart package does this automatically based on 10 randomly selected subsets of the data each of size 10/n (Therneau et al. 2011).

Trees can be described by their size (i.e. the number of leaf nodes) and their fit, or how well the tree will predict. Fit is defined by either relative error (RE) which in classification trees is a measure of the level of homogeneity of the leaf nodes over the level in the root node, and in regression trees the

amount of variance not explained by the tree, or by the cross-validation relative error (CVRE) (De'ath 2002). Classification tree nodes are typically characterised by the distribution of the dependent variable (as well as group size and the value of the independent variables that define the node). Regression tree nodes are characterised by the mean value, group size and the values of the independent variables that define the node (De'ath et al. 2000).

4.4.2. Regression Tree Model

The initial tree built is a (relatively) large one (see Figure 30), with a minimum number of splits set at 20 specified as follows , the number of cross-validations set at 20 and the complexity parameter[44] at 0.06:

```
Regression.tree.model <- rpart(SI ~. ,method = "anova", control = rpart.control(minsplit = 20, xval=20, cp = 0.06), data)
```

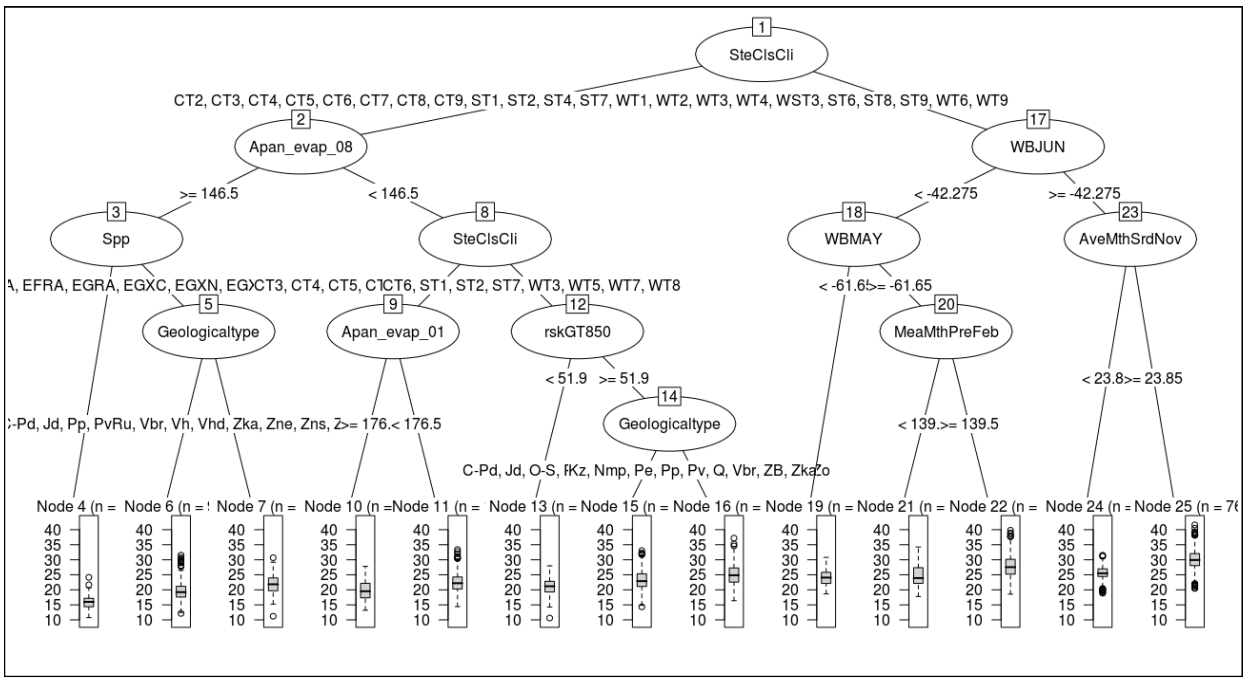


Figure 30: Initial large 12 split *Eucalyptus* regression tree

Even though a minimum number of 20 splits was specified, only 12 splits could be obtained with the complexity parameter specified, and only 10 of the 232 available variables were actually used to

[44] The complexity parameter (CP) is essentially a cost on each additional split, the increasing "cost" of a larger tree is traded off against a reduction in lack of fit (Maindonald et al.2007). Any split which does not improve the fit by the value specified will be ignored. The CP is set low initially to grow a bigger tree.

construct the tree[45]:

SteClsCli - Site classification based on climate[46];

Apan_evap_08, 01 – Mean monthly Apan evapotranspiration for August and January;

MeaMthPreFeb - Mean Monthly Precipitation for Febuary;

AveMthSrdNov – Average monthly solar radiation for November;

Spp - Species;

rskGT850 - Probability of obtaining > 850 mm of annual rainfall in any given year;

WBJUN , MAY - Water balance for June and May;

Geologicaltype – Geological type.

The parent node is first split by Site classification (based on climate), then by mean monthly Apan evapotranspiration in August (node 2) and water balance for June (node 17). Interestingly, these are both restrictive (i.e. they represent low rainfall months) and may in fact represent the same driver. The left branch is then further divided by species and again by site classification. The right branch by the water balance in May, and the average Solar radiation in August. It is only further in the splitting that factors such as geological type and the % potential of obtaining greater than 850 mm in any given year start to appear. This gives us a strong impression that climate – and in particular rainfall are the strongest drivers of height growth in this model.

Looking in more detail at the predictive ability of this large model, it is clear from Figure 31 that the model can be simplified further – the apparent R^2 value increases above 50 % by the 6th to 8th split then does not increase much as the splits increase, and the relative error flattens out at about the same point. The point at which complexity no longer adds value to the model is where the tree needs to be pruned back to. Figure 32 shows the observed Site Indexes versus those predicted by the initial model.

[45] The full model output can be found in Appendix 9.

[46] See Appendix 5.

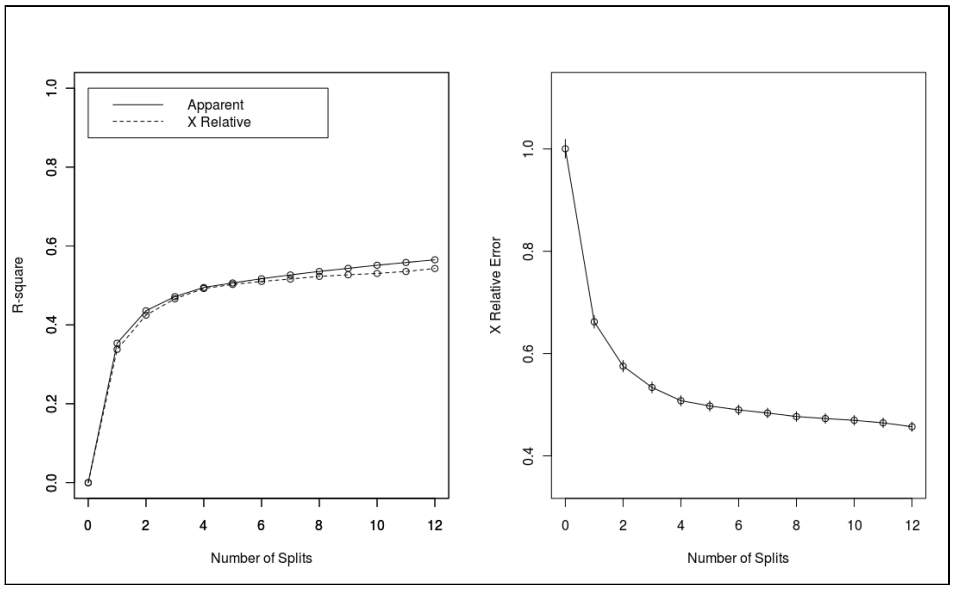


Figure 31: The apparent and cross-validated relative R^2 by number of splits, and the cross-validated relative error by number of splits for the first large *Eucalyptus* regression tree.

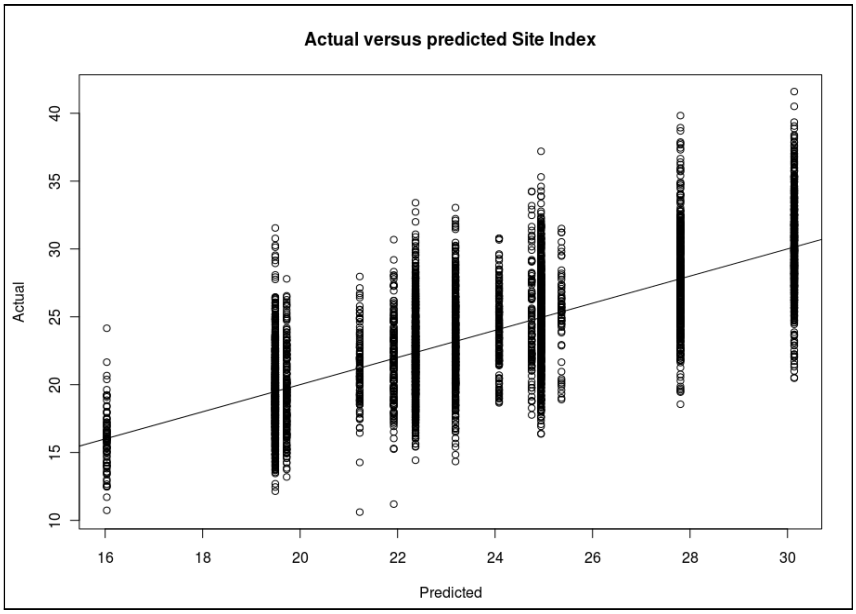


Figure 32: Observed versus predicted Site Index for the first large *Eucalyptus* regression tree.

4.4.2.1. Pruning

In order to select the most parsimonious model, the initial large model is now pruned back. Looking at the table below the 'xerror' column gives estimates of cross-validation prediction error for the different number of splits, the lowest prediction error obviously donates the best number of splits, however, in this case it does not reach a minimum (each successive split reduces the error). However, the complexity parameter (CP), which is a measure of the value added (improved fit) versus the complexity, steadily decreases to the CP number specified in the model call (see Table 29).

Table 29: Error and complexity (by cross-validation) for the number of splits.

CP	nsplit	rel error	xerror	xstd
0.353471293	0	1.0000000	1.0004317	0.018324305
0.082309356	1	0.6465287	0.6621257	0.012381169
0.035738010	2	0.5642194	0.5754456	0.011289866
0.023304024	3	0.5284813	0.5338072	0.010691860
0.011452004	4	0.5051773	0.5078477	0.010084055
0.010650137	5	0.4937253	0.4974974	0.009890438
0.009611191	6	0.4830752	0.4898202	0.009672843
0.008866080	7	0.4734640	0.4837786	0.009684178
0.007939401	8	0.4645979	0.4769056	0.009579612
0.007843509	9	0.4566585	0.4730089	0.009484857
0.006994593	10	0.4488150	0.4695541	0.009477397
0.006800850	11	0.4418204	0.4646040	0.009450548
0.006000000	12	0.4350196	0.4570242	0.009308809

Where: CP – complexity Parameter, nsplit – number of splits, rel error – relative error, xerror – cross-validated error, xstd – cross-validated standard error

Viewed graphically (Figure 33), one can see that although the cross-validated relative error continues to improve with increasing tree size, the amount of improvement flattens off after a tree size of approximately 5 - 8 splits. Thus the ideal model would have a complexity parameter somewhere between 0.011 and 0.0069.

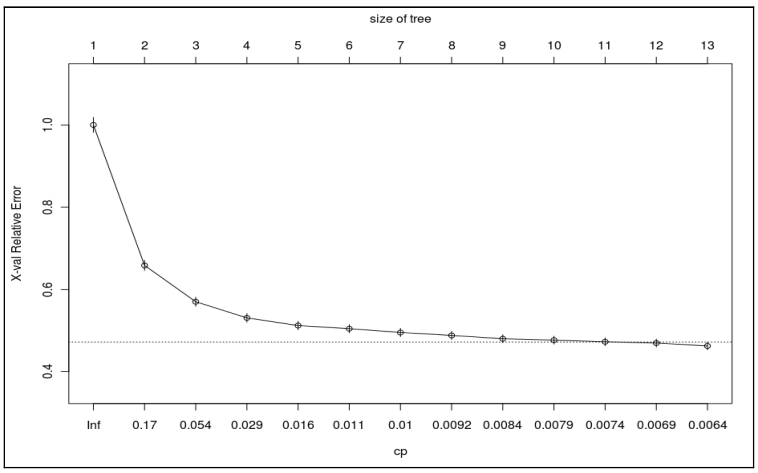


Figure 33: Relative cross-validated error and complexity parameter by tree size for the first large regression tree.

De'ath et al. (2000) give two methods of choosing the "best tree": (1). Where a plateau has been reached in the error estimate, the tree with the minimum error is chosen, and (2). A method suggested by Breiman et al. (1984), where the smallest tree is chosen which falls within one standard error of the minimum, the 1 SE rule. Therefore if the 12 split tree is assumed to have the smallest cross-validated error : 0.4570 plus the standard error of 0.0093, giving us a cross-validated error of 0.4663 – equating to a tree with 11 splits, still a large tree.

Simpler is often better, however, so a complexity parameter of 0.015 was chosen heuristically, which produced a much more parsimonious model with only three explanatory variables (see Table 30 and Figure 34): Site classification based on climate, potential A-pan equivalent evapotranspiration for August and the water balance for June. The model is easily interpreted as follows: Sites falling into each of the split criteria receive the predicted Site Index: so for example a Cool Temperate 5 site, with a potential A-pan evapotranspiration value below 146.5 mm in August will receive a predicted Site Index value of 21.93.

The final model has a relative coefficient of determination (R^2) of 0.4944, meaning that these three explanatory variables explain approximately 49.44 % of the variation in Site Index! This is remarkable given the complex nature of the problem, and the fact that it covers the full geographic

spread of the data supplied[47]. The model may, however, not be immediately useful, given that in essence there are only 5 site classes – it would need to be broken down to a lower level. This could be done by site classification or some other geographic breakdown.

Table 30: Summary of the *Eucalyptus* Regression tree model.

Node	Split Variable	Split criteria	Number of observations	Mean SI
1	Root		5457	23.72845
2	SteClsCli	CT2,CT3, CT2,CT3,CT4,CT5,CT6,CT7,CT8,CT9,ST1,ST2,ST4,ST7,WT1,WT2,WT3,WT4,WT5,WT7,WT8	3895	21.94915
4	Apan_evap_08	>=146.5	1245	19.60063 *
5	Apan_evap_08	<146.5	2651	23.05085
10	SteClsCli	CT3,CT4,CT5,CT7,CT8,CT9,ST4,WT1,WT2,WT4	1226	21.92878*
11	SteClsCli	CT6,ST1,ST2,ST7,WT3,WT5,WT7,WT8	1428	24.01419 *
3	SteClsCli	ST3,ST6,ST8,ST9,WT6,WT9	1558	28.18128
6	WBJUN	< -42.275	728	26.39460 *
7	WBJUN	>=-42.275	830	29.74839*

- indicates a terminal node.

[47] Almost the entire commercial forestry area of South Africa (excluding the cape provinces).

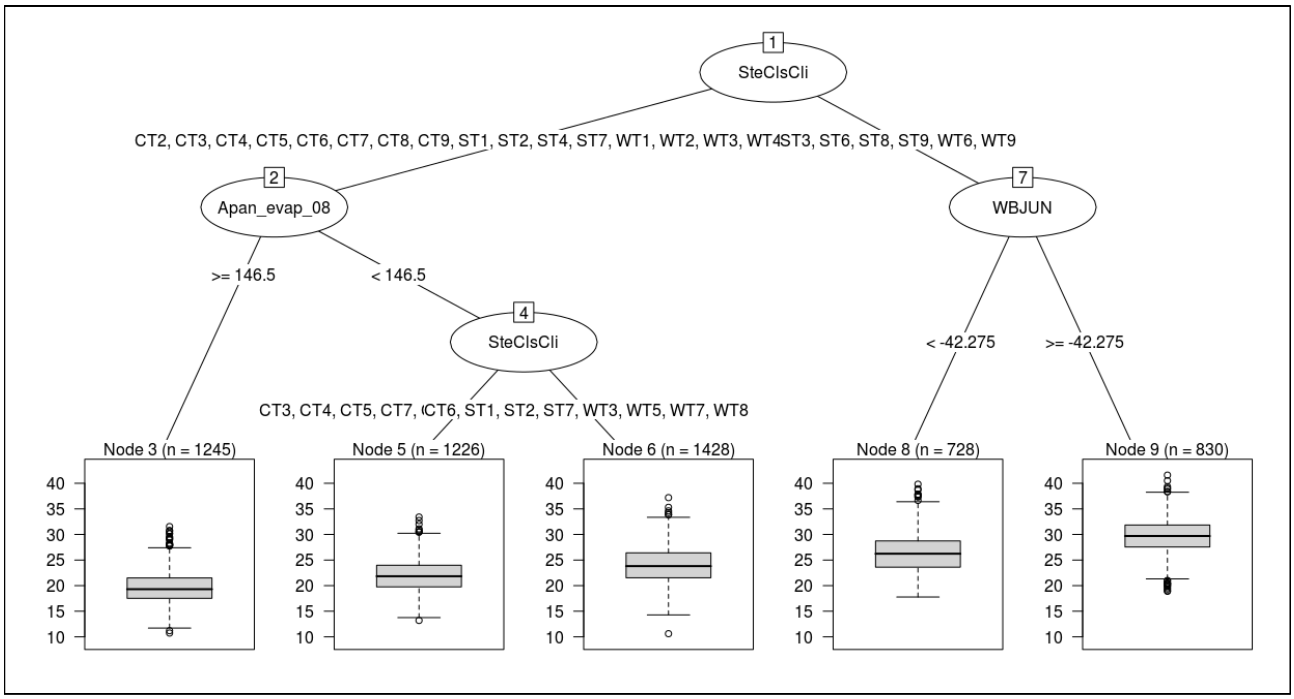


Figure 34: 5 split pruned *Eucalyptus* regression tree model.

Figure 35 clearly shows the "blocky" nature of regression trees – since there are only 5 terminal nodes. As suggested this can be reduced by building separate tree models on a lower level within the data – ideally by site class.

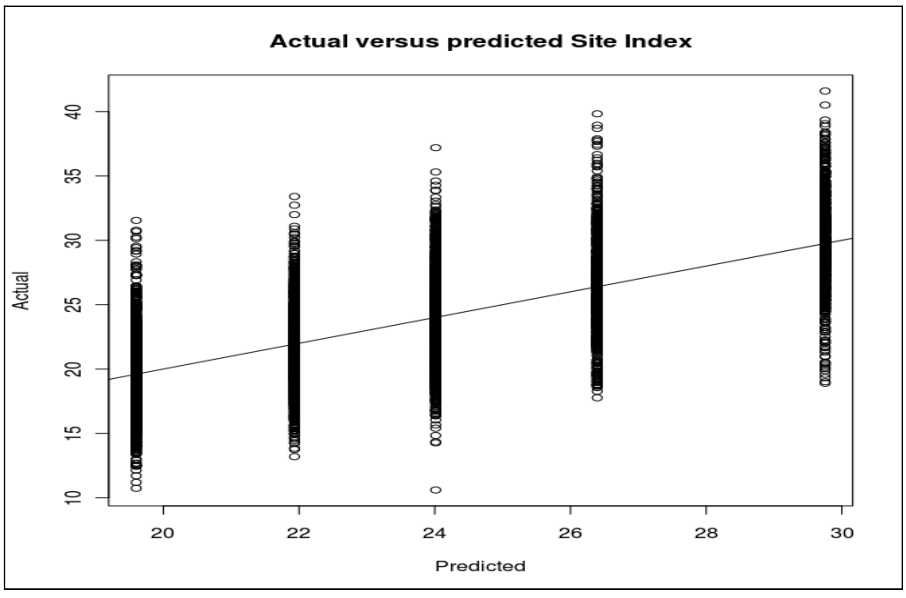


Figure 35: Observed versus predicted Site Index from the 5 split pruned *Eucalyptus* regression tree model

4.4.2.2. Post – Hoc tests

Stability

One of the problems with regression trees is that they can be unstable, since the point at which the data is split as well as the choice of the splitting variable depends on the distribution of the observations in the data set used. A change in this distribution can cause a very different tree to be built if the first splitting variable or cut point is chosen differently (Strobl et al. 2009). In order to test the stability of this model the same analysis was repeated 100 times by means of bootstrapping and the optimal number of splits recorded. The result showed that in all 100 models the same number of splits were indicated meaning the model is a stable one. Where this is not the case, alternative aggregate methods can be used such as random forest or bagging (Everitt et al. 2010; Strobl et al. 2009).

Residuals

As can be seen in Figure 36, the residuals of the pruned model are normally distributed.

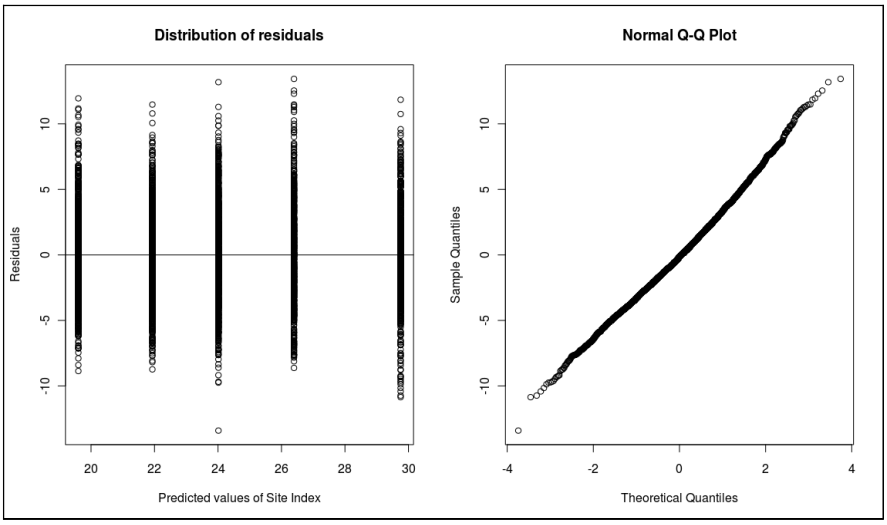


Figure 36: Distribution of the residuals of the 5 split pruned *Eucalyptus* regression tree model.

4.4.3. Multiple linear regression

As stated previously the majority of previous attempts to link abiotic data to Site Index have used the multiple linear regression method.

Dalgaard (2008) gives the basic model for multiple regression analysis as follows :

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$

Equation 10: Multiple Linear regression form (Dalgaard 2008)

In which ε is assumed to be independent, is normally distributed and has a variance of zero

where

$x_1 \ldots x_k$ are the explanatory or predictor variable's (from 1 to *k*)

and

$\beta_1 \ldots \beta_k$ the parameter coefficients estimated using least squares, where the values of β are found that minimize the sum of squared residuals. For each unit increase of the explanatory variable (x) the independent variable (y) will on average increase (or decrease) by the β coefficient.

In R this is specified using the lm() function as follows:

$$\text{MLR.model} < - \text{lm}(\text{SI} \sim \text{Var}_1 + \text{Var}_2 + \ldots \text{Var}_x , \text{data})$$

where $\text{Var}_{(1, \ldots x)}$ represent the individual physiological factors.

The number of independent variables was then reduced to find the most parsimonious model using the stepwise model search function step() based on the Akaike information criterion (AIC) with backwards and forwards elimination.

The model is only supplied as a traditional modelling approach as a comparison to more modern methods. There are a number of serious statistical problems with this model (see also section 4.1), and a substantial amount of work would be needed to check and or correct for these – the problems include :

- No second order (or higher) powers of the variables have been included, and in order to properly model Site Index using multiple linear regression more work would need to be done to determine whether there are any polynomial relationships.

- No interaction terms were specified.
- The likelihood of multicollinearity, where two or more predictor variables are highly correlated and carry similar information about the response (this is highly likely in this model as the majority of variables had a variable inflation factor above 5 [48]). Multicollinearity boosts the R^2 value without adding explanatory value. The least squares method struggles to distinguish the separate effects, and may even produce coefficients with an incorrect sign (Sheather 2009).
- The potential of "spurious correlations" where two (or more) variables may produce an association because they are related to another variable which has not been included in the model (Sheather 2009).
- The plethora of variables does not allow for a clear interpretation of their influence.

Bearing the above problems in mind the final model produced should be viewed as a simplistic representation and has only been included as an illustration of the methodology for comparison, and since the model is of no predictive value the coefficients have not been published here. Model validation on an independent data set was also not performed.

176 variables (this includes the dummy variables created to handle categorical variables, 98 excluding dummies) were ultimately included in the model with a residual standard error of 2.884 (5284 df) and an adjusted R^2 of 0.6296. The model itself was highly significant with an F-statistic of 55.23 and a p-value of 2.2e-16. The model also had normal residuals. Figure 37 below shows the observed versus predicted values given by the model.

[48] Correlation between the predictor variables increases the variance of the estimated coefficients, the variance inflation factor is calculated as a measure of multicollinearity as follows : $1/(1-R^2_j)$ for variable estimated coefficient β_j. A factor of 5 is often used as a cut-off (Sheather 2009).

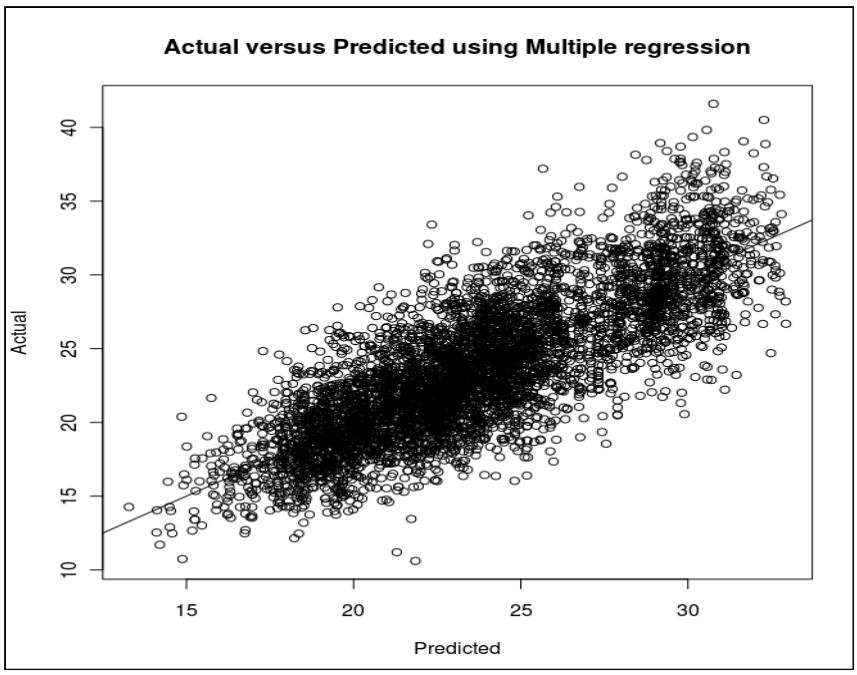


Figure 37: Observed versus predicted Site Index using the *Eucalyptus* linear multiple regression model.

4.4.4. Multiple linear regression using variables identified by the regression tree

One potential method to avoid the problems associated with the previous model, is to use the variables identified by the regression tree as the explanatory variables. A model was constructed using only the ten variables identified in the 12 split regression tree model, vis:

SteClsCli - Site classification based on climate[49];
Apan_evap_08, 01 – Mean monthly Apan evapotranspiration for August and January;
MeaMthPreFeb - Mean Monthly Precipitation for Febuary;
AveMthSrdNov – Average monthly solar radiation for November
Spp - Species;
rskGT850 - Probability of obtaining > 850 mm of annual rainfall in any given year;
WBJUN , MAY - Water balance for June and May;
Geologicaltype – Geological type.

The variable inflation factors were then calculated for each of the above explanatory variables (Table 31):

[49] See Appendix 5.

Table 31: Variance inflation factors for the 10 explanatory variables used in the alternative *Eucalyptus* multiple regression model.

Variable	GVIF	Df	GVIF^(1/(2*Df))
Spp	4.166684e+01	20	1.097728
rskGT850	5.727031e+00	1	2.393122
MeaMthPreFeb	1.009096e+01	1	3.176628
SteClsCli	1.906706e+05	24	1.288267
AveMthSrdNov	1.424741e+01	1	3.774574
Geologicaltype	1.101451e+05	31	1.205930
Apan_evap_01	2.947988e+01	1	5.429538
Apan_evap_08	2.487156e+01	1	4.987140
WBJUN	1.838488e+01	1	4.287759
WBMAY	2.037958e+01	1	4.514375

Where : GVIF = Generalised variance inflation factor

Df = Degrees of freedom

GVIF^(1/(2*Df)) = adjusted (for degrees of freedom) generalised variance inflation factor

The variance inflation factor gives an indication of the amount by which the variance of the regression coefficient is increased due to collinearity, if these exceed 5 the resultant regression coefficients will be poorly estimated (Sheather 2009). We can see that there is still some level of collinearity, however, it is greatly reduced from the previous multiple regression model and only the potential evapotranspiration is above 5 on an adjusted basis.

As there are fewer explanatory variables it is possible to look closer at the data. From the pairwise scatterplot shown below in Figure 38, the only relationship which appears to be clearly discernible is between the mean monthly precipitation for February (MeMthPreFeb), and the risk of obtaining greater than 850 mm of rainfall in any given year (rskGT850). RskGT850 also has a variance inflation factor (unadjusted) of 5.7270 this variable was therefore dropped from the model.

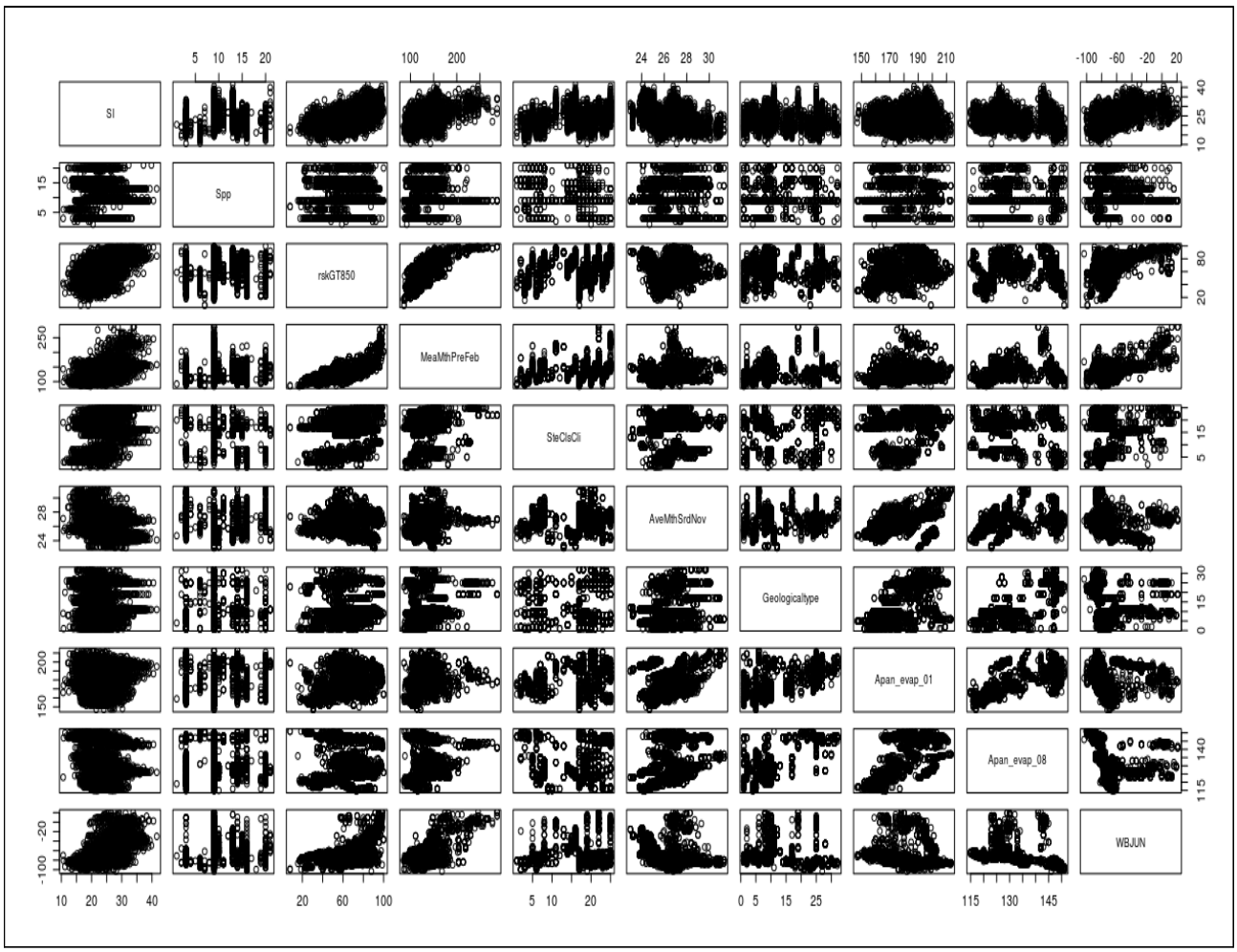


Figure 38: Pairwise plots of the data used in the alternative *Eucalyptus* multiple regression model using the variables identified in the regression tree.

The model was specified in R as before using the lm() function and then reduced using the stepwise model search function step() based on the Akaike information criterion (AIC) with backwards and forwards elimination.

The resultant model used 7 variables (79 if dummy variables used to cater for categorical variables are included), is highly significant (p-value of <2.2e-16) and has an adjusted R^2 value of 0.5833, and a root mean square error of 3.0359.

The model residuals proved to be normally distributed (see Figure 39).

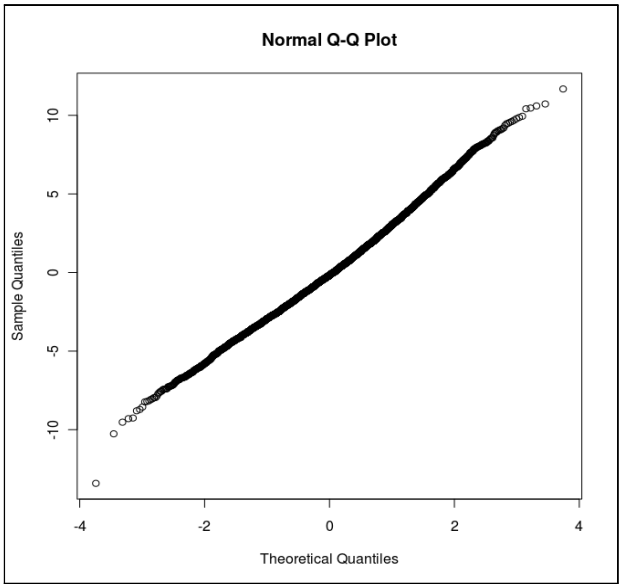


Figure 39: QQ plot of the residuals of the alternative *Eucalyptus* multiple regression model using the variables identified in the regression tree.

4.4.5. Hybrid or model trees

A number of authors (De Ville 2006, Muňoz et al. 2004, Neville 1998) suggest using decision trees to stratify the data to be used within some other model (normally multiple linear regression). A model is generally built in each leaf of the tree using the data stratified by that branch. These models are called "hybrid" or "model trees". A well known method for building a model tree with linear regression models in the leaves is the M5 algorithm (Quinlan 1992)[50]. This produces a regression tree that has a linear model in the terminal node rather than a mean value. Model trees are an ideal method of combining the strengths of linear modelling, with the data structuring strengths of trees. A simple model tree was built using the **RWeka** package in R (Hornik et al. 2006) specified as follows (restricted to 1000 observations per node in order to produce a small tree similar to the pruned regression tree):

```
Model.tree <- M5P(SI~.,control=Weka_control(R=F,M=1000),data)
```

As with the regression tree, 10 way cross-validation is employed. The tree had 6 terminal nodes each

[50]The M5 algorithm builds the regression tree in a similar way to the regression tree (i.e. by continually trying to reducing the variance in the target variable, in this case the splitting criterion is to reduce the standard deviation (Wang and Witten 1996)), other methods exist which try to maximise the quality of the terminal node model.

with a linear model associated with the node (see Figure 40, and Appendix 11 for the full model), with an R^2 of 0.5209 and a mean absolute error of 2.5906.

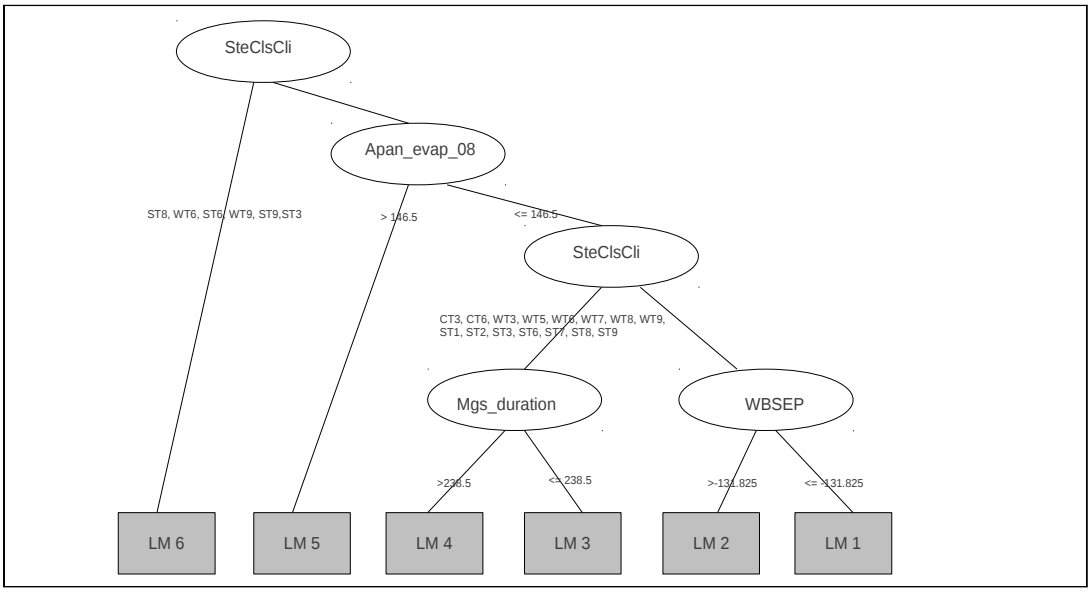


Figure 40: *Eucalyptus* Hybrid / model tree, each terminal node contains a linear model.

As well as the 4 variables used in the tree, a further 2 were included in the linear models - Species, and Water balance for June. The terminal node linear models are restricted to the variables used by tests or linear models in the sub-tree, and theses are then simplified by removal of variables which do not contribute much to the model – in extreme cases this can leave a constant only (Quinlan 1992). Each terminal node linear model had the following form :

$$SI = \beta(\text{Species[list]}) + \beta(\text{Site Classification[list]}) + \beta(\text{Mean growth season duration}) + \beta(\text{Potential Apan evapotranspiration for August}) + \beta(\text{Water Balance September}) + \beta(\text{Water Balance June}) + \epsilon$$

The resultant hybrid regression tree model is likely to have higher predictive accuracy and does not have the "blocky" nature of a regression tree. However, it comes with a disadvantage with regards to interpretation, since, it has more than one model contributing to the prediction (Torgo 1997). One other advantage which model trees have over regression trees is that they are able to return values lying outside the observed range of the data (i.e. they can extrapolate), something regression trees are unable to do (Quinlan 1992).

4.4.6. Random Forest

Another method that appears to be useful to overcome the "blocky" nature of tree models is the random forest[51], however, the method does not have the advantages of the simple tree model in that it does not allow for a better or simpler understanding of the drivers, because the model cannot be viewed like a tree model. The random tree method would be a better choice for modelling interactions which are already fairly well understood, whereas the regression tree method is a better choice of method where the subject is not well known. The main disadvantage of the random forest model is that it is a "black box" - it is not possible to see inside the model, or to view the individual trees used to create the model (Prasad et al. 2006).

The random tree method was developed by Breiman (2001) to reduce over-fitting of the data – it is similar to other tree methods in that a sample is drawn to grow multiple trees – the difference being that *each tree is grown with a randomised set of independent variables*. A large number of unpruned trees are grown (a forest between 500 and 2000 trees) and aggregated by averaging (Strobl et al. 2009, Prasad et al. 2006, Breiman 2001).

Liaw and Wiener (2002) give the random forest algorithm as follows :

- n_{tree} samples are bootstrapped from the full data set.
- For each of the samples above, an un-prunned regression tree is grown. Instead of choosing the best split from all of the variables, a random sample m_{try} of the predictors is chosen (the default for regression is $p/3$ where p is the number of predictors), and the best split is chosen from this sub-set of predictors.
- The n_{tree} trees are then aggregated (by averaging in the case of regression, and by majority in the case of classification.
- At each iteration an estimate of the error rate is calculated using the tree created, by predicting the data using the "out-of-bag" sample (i.e. the data not used to create the tree). These are then aggregated for all the trees grown.

The random forest model was specified using the **randomForest** package in R (Breiman et al.

[51] A similar method known as 'Bagging Trees' involves replicating sample data to make up for the test sample drawn. (Prasad et al. 2006).

2010). 1000 trees were specified to create the "forest". Due to the nature of the method, pruning is unnecessary, but the size of tree grown can be restricted by specifying the node size, and the maximum number of terminal nodes. In order to keep the tree sizes reasonably small the following was specified: minimum node size of 40 (default is 5), and maximum number of terminal nodes at 15. At each node 78 variables were tested. The model was specified in R as follows:

```
Random.forest <- randomForest(SI ~., importance = TRUE, ntree=1000, proximity = TRUE, nodesize = 40,maxnodes = 15, na.action=na.omit, data)
```

The model produced had a mean of squared residuals of 9.4152 and an R^2 of 58.05 (an unrestricted random forest model produced an R^2 of 67.19 but obviously with very large tree sizes). Every one of the 1000 trees produced had 15 end nodes – showing that the tree sizes could all have been far larger. Potentially as many as 30 variables could have been used to build a tree with 15 terminal nodes. Figure 41 below shows the observed versus predicted values given by the model.

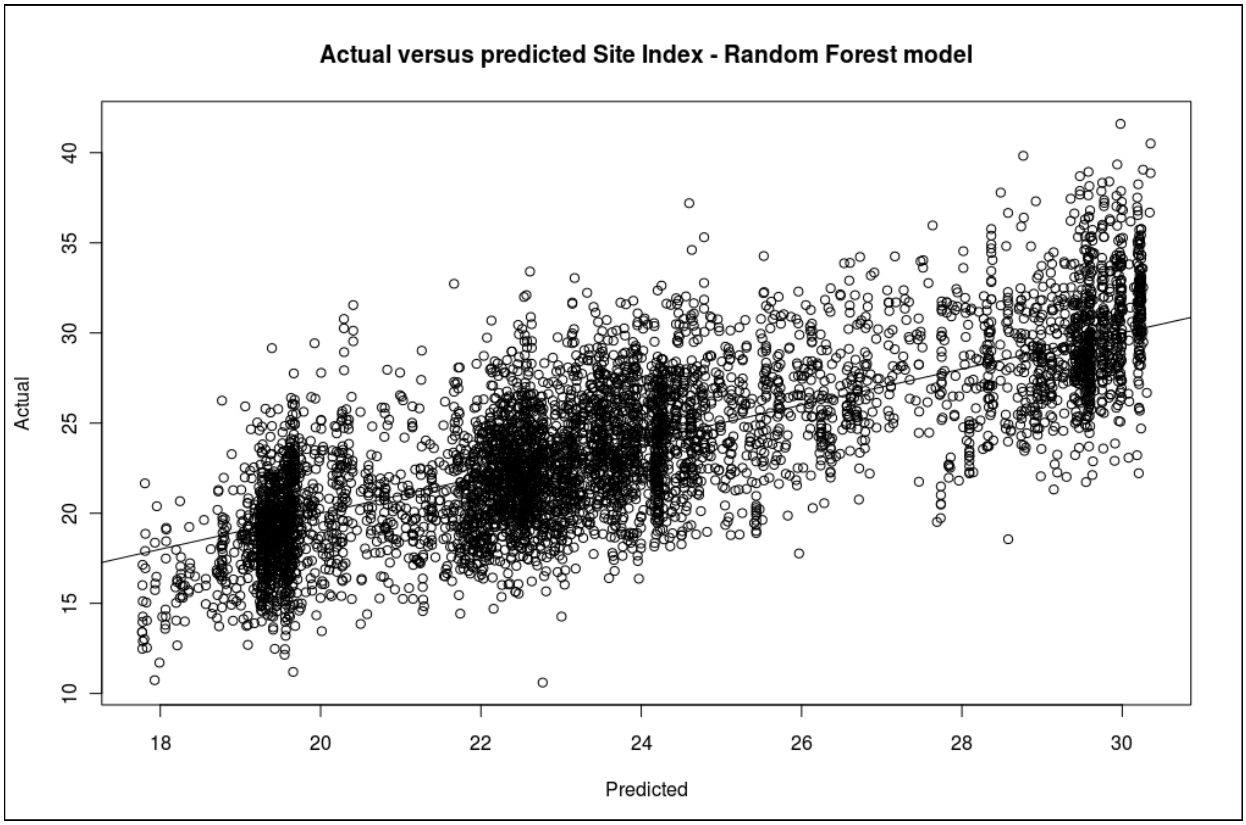


Figure 41: Actual versus predicted values of Site Index for the *Eucalyptus* random forest model.

As stated previously, random forest models are "black boxes" - so it is not possible to visualise or

export the model, however, there are some metrics available to help in interpretation (Prasad et al. 2006). Figure 42 above shows the importance of the predictor variables used to construct the random forest trees, both in terms of the percentage increase in mean square error (%IncMSE) if the variable is excluded, and in the purity of the node. Site Classification by climate, and the water balance for June, and the potential A-pan evapotranspiration are again highlighted as important predictor variables under both measures.

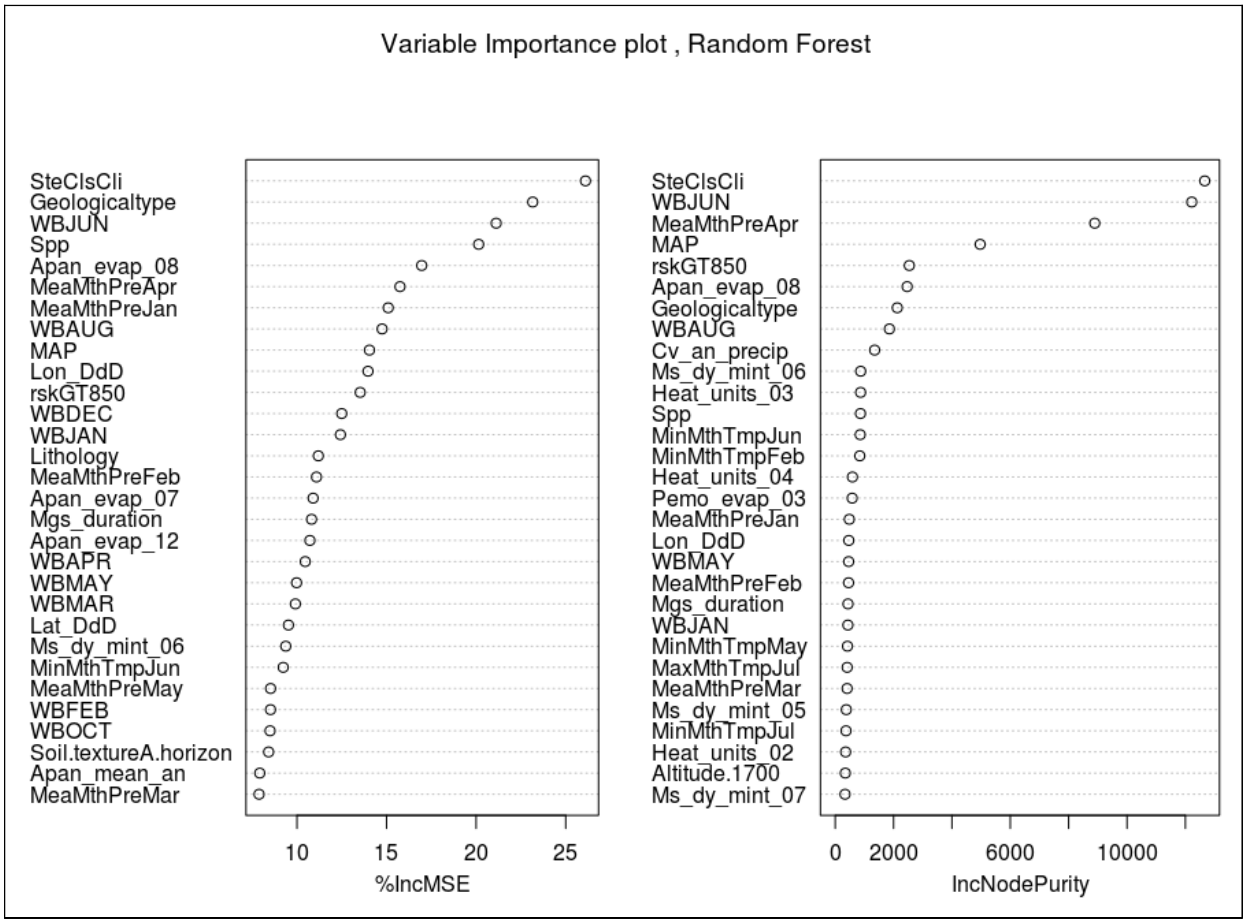


Figure 42: Variable importance for the *Eucalyptus* random forest model.

4.5. Results

It must be remembered that the intention of this section of the study was not to produce a valid, viable Site Index model, rather to compare alternative modelling methodologies. As such, focus is on the performance of each model methodology, rather than the variables included.

From Table 32 below it can be seen that the alternative multiple regression model (using the variables identified by regression tree analysis) was able to explain more variation and had a lower root mean square error than the alternative models. However, this was at the expense of a large number of variables. The alternative multiple regression model used only seven explanatory variables, yet this is extended to 76 due to the dummy variables required to handle categorical predictors and this thereby increases the complexity.

Table 32: Model comparison – fit versus number of variables used for the *Eucalyptus* default Site Index models.

Model	Fit (R^2)	RMSE[52]	Number of variables used
Regression Tree (5 splits)	49.44	3.3685	3
Regression Tree (12 splits)	55.79	3.1256	10
Multiple regression	62.96**	2.8376	98 (176*)
Alternative MLR – using tree variables	58.33	3.0359	7 (76*)
Hybrid model	52.10	3.3694	6
Random Forest	58.05	3.0687	30

*Including dummy variables. ** Model has serious multicollinearity problems and should not be seen as the best fit.

The initial multiple regression model has serious problems associated with it and would be entirely inappropriate for this type of data without a large amount of additional work. The random forest model performed better than the regression trees and the hybrid, again at the expense of complexity. The simple regression tree produces the best fit relative to the number of variables used. The main advantages and disadvantages of each of the modelling methodologies is discussed in Table 33.

[52] Root Square Mean Error is in meters and was calculated as the square root of the mean squared difference between predicted and actual Site Index.

Table 33: The main advantages and disadvantages of the various modelling approaches.

Approach	Main Advantages	Main Disadvantages
Regression Tree	Simple, and able to capture large variation with few variables. Few assumptions needed on the sample data.	"Blocky" in nature – but can be overcome by breaking the model down further. Unable to extrapolate beyond the limits of the data used to construct the model.
Multiple Linear Regression	Able to capture relatively large amounts of variation. Produces a continual "smooth" response.	Need to have some understanding of the relationship between variables. Problems with multicollinearity. Complex with many variables used.
Alternative MLR	Reduces the multicollinearity issues associated with the first MLR model.	Still complex due to the number of variables used.
Hybrid or Model trees	Simple, and able to capture large variation with few variables.	Mix of models can complicate interpretation and verification.
Random Forest	Can easily show the importance of individual variables. Does not over fit. Not as "blocky" as regression Trees .	"Black box" - unable to export. Uses many variables

4.6. Discussion

As can be seen the Regression tree model has the huge advantage of simplicity, it was able to capture almost 50 % of the variation in Site Index using only three input variables, and although the alternative model strategies produce better coefficients of determination, this is at the expense of complexity. Considering that this model is a generic model for the entire forest growth area covered by the data (essentially the entire commercial forest area of South Africa)[53], the performance is remarkable. Further work would be needed to localise the model to allow it to be more readily usable in a commercial environment, however, it is clear that the regression tree method is ideally suited to this type of data.

Multiple Linear regression comes with a level of complexity due to the nature of the data and statistical problems which would be difficult to overcome for this type of data set. And the black box nature of the random forest model, coupled with the fact that it is not possible to implement in an external environment such as GIS makes it and multiple regression less suitable as an alternative than a hybrid or model tree.

The main advantage to using regression trees is simplicity, and above all the ability to easily explain or uncover the main independent variables and their interaction. This advantage is also seemingly its biggest disadvantage – the resultant model is not smooth and continuous. Whether this is actually a

[53] And covers all species.

problem is debatable, since all models derived from a sample set are only explanations of the data set used, and conversely any information extracted from a model is really only information about the model itself – it is not, and cannot be a true reflection of the whole. Continuous "smooth" model outputs may simply be giving a false sense of certainty and accuracy – the fact that output can be displayed to three decimal points does not mean that it is accurate to this degree!

The move in forestry from traditional yield tables to continuous empirical models has given some managers the impression that prediction and projection have become more accurate – this may not always be the case. Having the model in the classed format of a table at least allowed the user to intuitively realise that the output was not definitive. If, however, a continuous model is a necessity the hybrid or model tree seems to be a potential alternative.

Chapter 5. CONCLUSIONS AND RECOMMENDATIONS

As with many studies the questions asked and answered often lead to opportunities for further research, this study is no exception. Since this study is divided into three main sections the conclusions and recommendations have been divided on the same basis.

Initial planted density.

Although the majority of the observations used in this portion of the study proved the assumption that dominant height is unaffected by the initial planted density, two exceptions were found in *E. dunnii* and *E. nitens*. Since these constituted the smallest subset of the data, and represented only single trials it is possible that the effect found was due to some variable not included in the analysis. However there is evidence from other studies that there may well be a role played by stocking for some species, and that this may be a subject worth pursuing with a more dedicated and larger data series. If an effect is found, this will have knock on consequences within the growth model configurations where an adjustment factor will need to be introduced to the Site Index models. Any potential effect will also have consequences on other research since Site Index is often used to differentiate between treatments, it is important that researchers are aware of potential interactions between Site Index and stand density – it is possible that if they do not recognise the additional co-variable that the conclusions reached may not be valid.

Measurement age of dominant height.

It is clear from this study that early estimates of Site Index via dominant height measurement should be used with caution, or not at all. The early measurements were shown to be either over or under estimates of the true value. Measurements taken below 2 years in the case of *Eucalyptus*, 8 years in the case of *Pinus*, and 4 years for *Acacia* were shown to be significantly different from measurements taken at other ages. The *Eucalyptus* result is, however, based on the espacement trial

data set, which used a single Site Index model and may need to be repeated on separate Site Index models since the results from the PSP/TSP data set using numerous models produced less definitive results. One potential source of the over/under estimates may be due to the inability of the projection model to extrapolate if the measurement is taken before the point of inflection (Seifert 2011), this may also explain the results obtained using the PSP/TSP data. This may be a fruitful avenue of further study.

The results of this portion of the study will have direct consequences for inventory policies since Site Index estimation is a critical output.

Site Index modelling

The purpose of this section was to compare various alternative modelling methods, rather, than to present a final model ready for application. As such, focus was not on finding the right variables but on comparison of various methodologies. The non-parametric methods tested proved to be comparable to, if not better than, traditional regression, without the statistical flaws of the latter. Regression trees in particular are of enormous benefit, as they are able to provide a better visualisation of the underlying drivers, without the complexity and potential problems associated with multiple linear regression. Further localisation of the models either using regression trees, or if a more continuous output is required, a hybrid tree model would be of commercial value. Hybrid tree models have the further advantage of being able to extrapolate beyond the limits of the data used to build the model – something regression trees are unable to do. Additional site data, such as the growth day classes and temperature classes recently introduced by Louw et al. (2011) should also be tested as predictor variables. Since the majority of the important variables identified are related to water, or the lack thereof, it may be worth including variables related to water stress periods. Most commercial forestry companies in South Africa have detailed soil databases, this as well as any other site related data, such as crop specific evapotranspiration or nutrient data could also be included.

By way of illustration a localised regression tree for the ST9 site class (Sub tropical class 9, from Smith et al. (2005)[54] with a mean annual temperature between 21 and 22 ºC, and mean annual

[54] See Appendix 5.

precipitation greater than 1075 mm/yr) was constructed for the *Eucalyptus* data (See Figure 43 below), this tree has a root mean square error of 2.4187 m[55] – a substantial 28 % improvement over the generic 5 split regression tree which had a 3.3685 m RMSE. Even further localisation could prove beneficial.

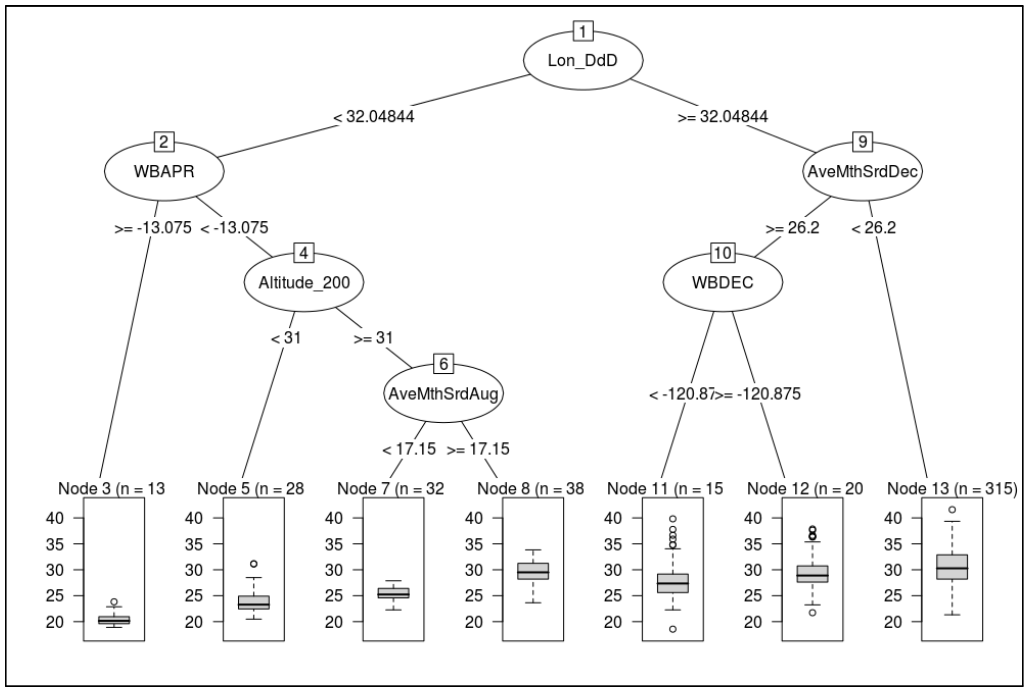


Figure 43: Localised regression tree for *Eucalyptus* in the ST9 climate class.

[55] A hybrid model tree on the same site class produced a root mean square error of 2.9606 m.

Chapter 6. INTEGRATING THE SITE INDEX MODEL INTO THE PLANNING PROCESS

It is important to consider how the introduction of a default Site Index model will affect the current planning process. Equally important is to consider if the costs associated in both developing and integrating the new model are justified. A brief investigation into the changes necessary to implement the model, as well as methods of calculating the value of information are given here.

6.1. The new process

Figure 44 below shows the possible future Site Index process, the most obvious difference between this new process and the one currently employed[56] is the reversal of the roles played by the enumerated Site Index and the actual production – in the current process the enumerated Site Index

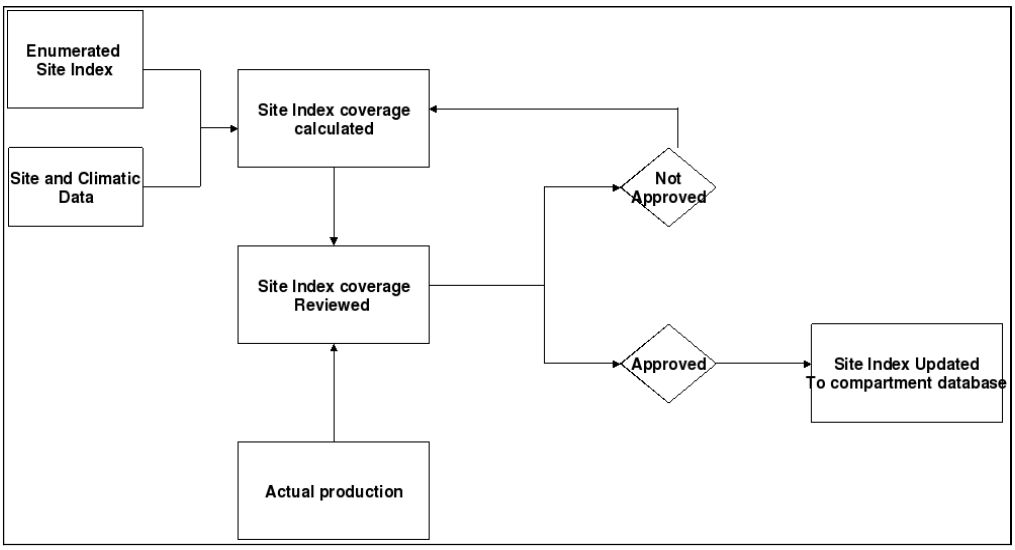


Figure 44: The envisaged future default Site Index process.

is used simply as part of the review/checking process, and the actual or expected production is used to directly calculate the Site Index. In the new process, however, the measured Site Index is used

[56] Based on the process followed by Mondi Limited.

directly to calculate a Site Index Coverage, and the actual production is used to check and review the coverage.

In the new process, the enumerated Site Index together with Site and Climatic data are used to calculate a predictive coverage of Site Index for those sites which do not have measured Site Indexes. This predictive coverage is then reviewed using the actual production, if approved the coverage can be used to update the compartment database. Since the coverage is generated within a Geographic Information System (GIS) this data can be easily mapped, tested and updated and will eliminate the need for the current large manual systems.

This new process has a number of key advantages over the current process.

- The key advantage is that the process is based on scientific methodology, and is therefore a more defensible and audit-able input into the planning system.
- Since the process is automated, there is less scope for error.
- The process is easily repeatable – which would permit additional use-age of the process such as allowing for climate change scenario planning. All that this would require would be an alternative climate coverage in the GIS.
- Since the Site Index is based on measured data it would allow for a more thorough review of how well the growth models are performing. The current method uses the growth models to generate the default Site Index, which means that it is not possible to test the performance of the growth models directly.
- The new coverage will be on compartment level (see Appendix 6 for a description of the geographic hierarchy) – the current process is generally complied on an Area, Working plan or Farm level depending on the data available. The new process will therefore bring a higher level of focus or intensity. The level of accuracy will be dependant on the lowest level of accuracy of the data used.
- The default Site Index coverage can be used to check specific enumerations – the measured Site Index can be compared to the predicted default. It can also be used to identify under performing stands (expected versus actual productivity) for further investigation, and as a means of improving corporate governance by identifying sites for further investigation and or audit.

- The default Site Index coverage could be used to evaluate future potential forest sites, and for the ranking of sites into areas of high or low productivity and therefore input and or investment costs.
- The coverage can also be used to identify sites for research (e.g. for climate change effects, stem form and volume function, biomass, wood and fibre quality, silviculture etc.). Currently this can only be done on compartments that have been enumerated, or on the level which the default Site Index has been calculated (e.g. on a Farm level.) – a distinct disadvantage if the research site is needed on an un-planted site, or if sites with varying Site Indexes are needed on the same farm.

6.2. Additional processes

As already mentioned , the proposed default Site Index process will allow for other processes such as climate change scenario planning, and the selection of research sites. It will also mean that current growth and yield processes could be changed to take advantage of the new data. Examples of this are :

- the placement of PSP (permanent sample plots), and espacement trials
- the grouping of growth and yield data for modelling purposes
- the linking of growth models to compartments.

6.3. Discussion

A long term (strategic) plan is a prediction about the future. It can be influenced by the probability of random (i.e. cannot be modelled) events or elements such as fire or land reform[57], as well as non-random events or elements such as felling age or planted area. Models of the non-random events are used to make the predictions. A model of Site Index is the last of these “non-random” elements which is not generally used by South African forest planners. The potential for these new models has only recently become obtainable due to the availability of the base data (specifically the climatic

[57] Risk adjustments can be made for these random events, but by there nature they cannot be predicted or modelled. The term land reform refers to the land reform process as carried out by the South African government.

and edaphic data).

The addition, these new models will add significantly to the planning "tool box", however, they will also affect the current planning process, and introduce new processes which will naturally have an effect on the roles and responsibilities of the staff involved. A certain amount of change management will therefore be necessary.

One aspect of this change is the altering of the focus from an entirely empirically centred growth and yield environment, to one that incorporates physiological processes. This will require a different approach and way of thinking. If the new model is used for other purposes (such as climate change scenario planning) some effort will be required in order to incorporate these models into the new process. A model designed for one purpose cannot be easily converted for use in another way without careful consideration.

An issue which has not been discussed is when to update the Default Site Index model. It is possible that with an increase in input data, any new model produced will be significantly different from the previous model. This is quite likely over a longer period due to climatic changes (e.g. over a drought cycle or cyclonic event), or due to changes in the genetic material planted. These changes should not, however, be material over shorter periods (up to 3 years). It would therefore be recommended that updating the model should not be done at too short an interval, unless there are specific reasons to do so, because constant changes could be detrimental to the integrity of the plan.

In a typical long term plan it is not uncommon for the construction of the plan to be based on less than 3 – 6 % measured (enumerated via inventory) compartments[58]. Figure 45 below shows a real world example (Ntonjaneni 2005). As you move closer to the present, the percentage of compartments in the plan that are enumerated increases, the relevance of inventory therefore also increases as you move from strategic (1 - 30 yrs), to tactical (1- 5 yrs) to operational (1-2 yrs) level plans.

[58] Generally only the older age classes are enumerated.

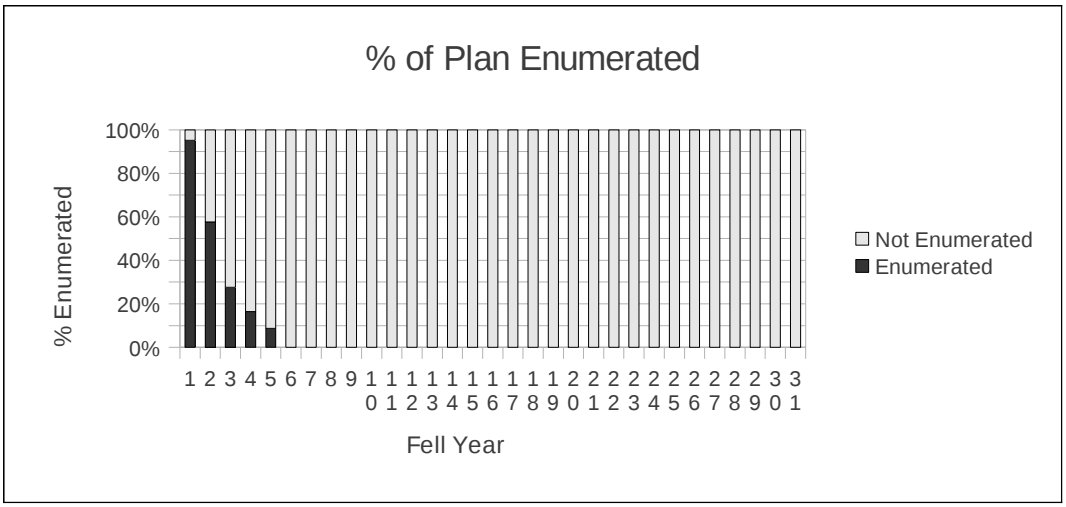


Figure 45: An example of a particularly well enumerated plan (6.18 % of the total plan is enumerated). 93.82 % of this plan is therefore based on default data.

A great deal of effort and attention is normally placed on collecting, analysing and using inventory data, while comparatively little effort is placed on the default data. As one can see this is misplaced from a strategic planning perspective - inventory data is simply one piece of the holistic data "puzzle".

6.4. Weighing the cost of data acquisition

The quantity of data or information can be defined by the characteristics the data has in terms of accuracy, dependability, reliability, relevance, timeliness, completeness and presentation. Its value can be said to be the difference between the value of the plan with, and without the data. (Kangas 2009; Duvemo 2009). Although there have not been many studies on the subject of valuing forest planning data, there are a few methods available to allow forest planners to weigh the cost of improved data. Traditionally this has been done simply based on the cost of acquiring the data and the level of accuracy of the new data, or by minimising the cost of data collection based on some level of accuracy (for example setting a minimum level of accuracy for enumerations – and choosing the least costly method to acquire this data). The problem with these approaches is that they do not

allow one to judge the value of the data on the level of decision making. (Kangas 2009).

6.5. Cost-plus-loss analysis

One method to include the value of the data for decision making is via *Cost-plus-loss analysis*, whereby the expected losses caused by poor decision making (due to inaccurate data) are added to the cost of acquiring the data (see Figure 46). The point where the sum of the two costs is at its minimum is then the most optimal.

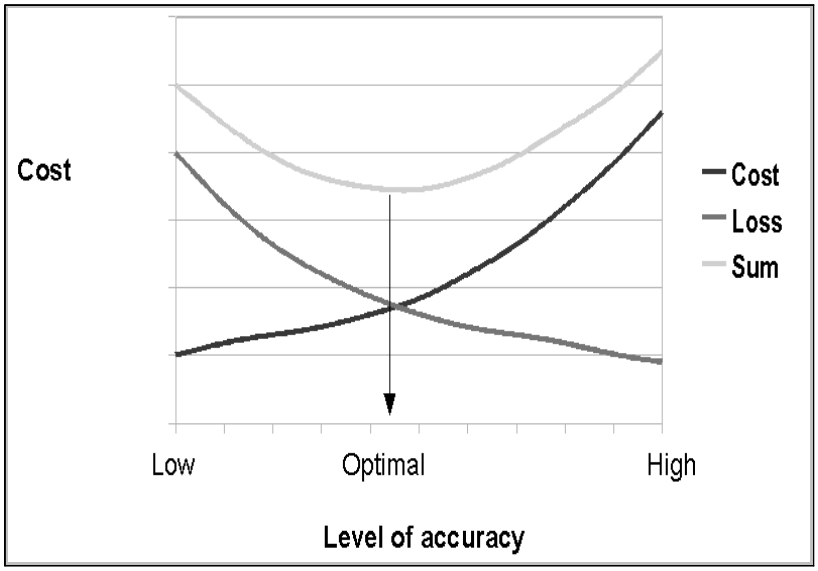


Figure 46: The loss due to poor decision making based on poor data , plus the cost of improving the accuracy is the total cost. (After Holström 2001; Magnusson 2006).

The most difficult aspect of cost-plus-loss analysis is to calculate the value of the potential loss – in most studies it is assumed that the plan is aimed at maximising the net present value (NPV) of the forest, so the losses are defined by the effect on net present value (Kangas 2009; Holström 2001; Eid 2000). The difference between two plans (with / without the data, or data accuracy) in respect to NPV can be calculated either directly as a function of accuracy, or more commonly by means of simulation. In simulation there are also two possible methods : by either using real data and real errors, or by simulating the errors (Kangas 2009).

Figure 47 shows how NPV losses can occur due to incorrect or sub-optimal data (in most studies this data is inventory data – but the principle can be extended to any type of data which would have an effect on the value or timing of the plan). The top line (light grey) shows the NPV from the harvesting plan using correct or optimal data. When the plan is based on incorrect or sub-optimal data the harvest is carried out at time T_{err} rather than at $T_{opt.}$

The NPV loss can be calculated as follows (Eid 2000):

If:

$NPVerr_{xy}$ is the NPV of stand number **x** (x = 1,2,....m) in data set **y** (y = 1,2,...n) and

$NPVopt_x$ is the NPV of stand number **x** in the corresponding data set

Then the Net present value due to an error in a certain data variable of stand number **x** and data set **y** is calculated as follows :

$$NPVloss_{xy} = NPVopt_x - NPVerr_{xy}$$

Equation 11 Net present value due to an error (Eid 2000)

Using the above calculation allows for a comparison between each individual variable in a data set – and that particular variable's effect on net present value. Where potential NPV losses can be expected to be large, more effort/cost can be spent on the collection of that particular piece of data.

The expected losses for a stand are then calculated as the sum of the NPV errors due to each data element error for that compartment, and the NPV losses for all stands is the sum of all the stand losses.

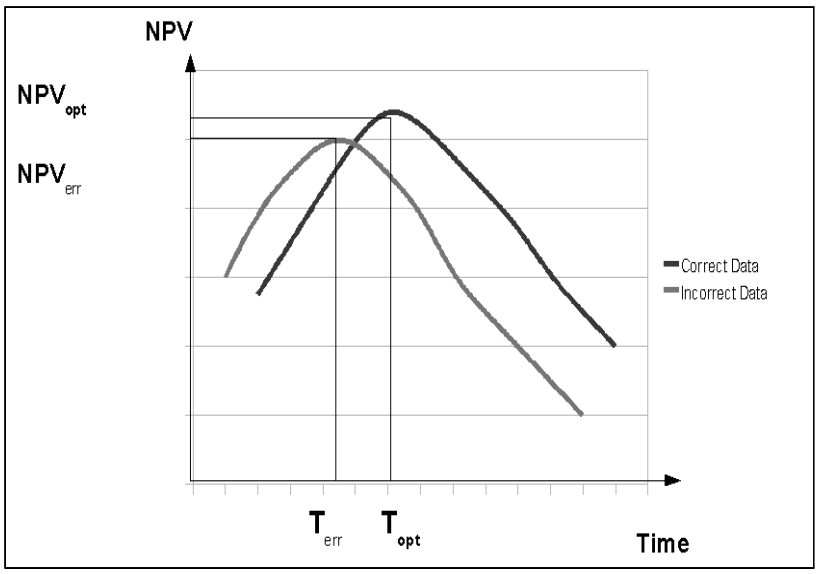


Figure 47: How Net Present Value losses can occur over time due to erroneous data (After Eid 2000; Kangas 2009)

6.6. Value of Information

It is possible to go even further than cost-plus-loss analysis by calculating the value of any individual piece of information or information concerning a particular variable in decision making. The *Value of information* (VOI) is the difference between the value of a particular project with particular information or data, and the value without that data (less the cost of acquiring the information). The value of an particular piece of information will depend on a number of factors including (Duvemo 2009):

- How uncertain the decision maker/s are and what their attitude is towards risk.
- What is at stake.
- What the cost is of using and assessing the information, and
- What the costs are of the alternatives or next-best substitute is.

Kangas (2009) also includes what the quality of prior information is, on the list affecting VOI.

In the *value of information* concept acquiring new information is only rational if the expected value of information is positive, in other words the expected cost is less than the expected gain. It can be

summarised as follows (Duvemo 2009):

$$EVOI = (EV_{after} - EV_{before}) - EC$$

Equation 12 Expected value of information (Duvemo 2009)

where

EVOI - Expected value of information

EV_{after} - Expected value after the new information or data is added

EV_{before} - Expected value without the information and

EC - Expected cost of acquiring the information

If forest data or information is not appropriately valued, it is possible that forestry companies experience large losses (in net present value) which they are unaware of due to the fact that they have developed plans based on data that is too imprecise or of poor quality (Borders et al. 2008; Kätsch 2006).

6.7. Discussion

Duvemo and Lämås (2006) summarised into four major components the possible sources of error in forest growth prediction and planning :

- The inherent randomness of nature itself.
- Incomplete models due to a lack of data.
- Errors in the description of the present (or initial) state.
- Errors in the parametrisation of the growth models due to the data used in model development.

Obviously not much can be done about the randomness of nature – other than to study causal relationships (i.e. how and why trees are influenced by nature) and try to gain a better understanding of these influences and events. Risk adjustments can then be introduced into the plan to help correct for these sources of errors[59]. It is nevertheless possible to fully correct for the other three major causes of error - given an unlimited budget. However, forest companies do not have unlimited funds

[59] Such as adjusting for the risk of fire or drought.

for planning, it is therefore up to the forest planner to ensure that the best possible value is obtained from the funding available. It is possible to re-use data, for example, in Norway where old compartment boundary data was used in conjunction with photogrammetry to estimate stand volume, they found that it was possible to use this old data to reduce the time taken to do the analysis (Aasland 2002). The longer data is used (for example soil data, which should not change rapidly over time) the more one can spend on acquiring this data, and at a higher level of accuracy. Using existing data for other uses (e.g. Site Index modelling using existing inventory, edaphic and climatic data) is another method of obtaining the most value out of the funding available.

Developments in computer science will enable forest planners to use and present plans and data in new and exciting ways in future examples include virtual reality, or 3D visualisation which will not only improve our understanding of the alternatives but potentially help us spot overlooked possibilities, or unforeseen consequences (Wang et al. 2006).

Finally it must be remembered that there are many other issues other than the quality of the input data which will affect the integrity of the plan, and that complexity does not equal accuracy!

REFERENCES

- **Aasland T.** (2002). *Use of old inventory data for forest inventory and management planning*. Doctoral Theses Agricultural University of Norway
- **Assmann E.** (1971). *The principles of forest yield study*. English edition, pp 159 – 205. Pergamon Press.
- **Avery T.E., Burkhart H.E**. (2002). *Forest Measurements*. McGraw-Hill series in Forest resources.
- **Baker F.A., Verbyla D.L., Hodges C.S., Ross E.W.** (1993). *Classification and regression tree analysis for assessing hazard of pine mortality caused by Herterobasidion-annosum*. Plant Disease 77(2), pp 136 – 139.
- **Bates D.M.** (2005) *Fitting linear mixed models in R*. R News , the Newsletter of the R Project. Volume 5/1, pp 27 – 30.
- **Bates D.M.** (2009) On line discussion : http://markmail.org/message/56c4ck4mmjyouqfo .
- **Bates D.M.** (2010) Lme4: Mixed-effects modeling with R. Draft book. Springer .
- **Bates D.M., Maecher M., Bolker B.** (2011). *Linear mixed-effects models using S4 classes*. Lme4 R package version 0.999375-42.
- **Batho A., García O.** (2006). *De Perthuis and the origins of Site Index: A Historical note*. FBMIS 1, pp 1- 10.
- **Bernardo A.L., Reis M.G.F., Reis G.G., Harrison R.B., Firme D.J**. (1998). *Effect of spacing on growth and biomass distribution in Eucalyptus camaldulensis, E.pellita and E. urophylla plantations in south-eastern Brazil.* Forest Ecology and Management 104, pp 1 – 13.
- **Borders B.E., Harrison W.M., Clutter M.L., Shiver B.D., Souter R.A**. (2008). *The value of timber inventory information for management planning*. Canadian Journal of Forest Research 38, pp 2287 – 2294.
- **Bourg N.A., McShea W.J., Gill D.E.** (2005) *Putting a CART before the search: Successful habitat prediction for a rare forest herb*. Ecology 86(10), pp 2793-2804.
- **Bredenkamp B.** (1987). *Effects of spacing and age on growth of Eucalyptus grandis on a*

dry Zululand site. South African Forestry Journal 140, pp 24 – 28.

- **Bredenkamp B.** (1993). *Top Height: a definition for use in South Africa.* South African Forestry Journal 167, pp 55.
- **Breiman L.** (2001). *Random Forests*. Machine learning 45(1), pp 5 – 32.
- **Breiman L., Cutler A., Laiw A., Wiener M.** (2010). *Breiman and Cutler's random forests for classification and regression.* RandomForest package in R. Version 4.5-36.
- **Breiman L., Friedman J.H., Olshen R., Stone C.J.** (1984). *Classification and Regression Trees*. Wadsworth and Brooks/Cloe Advanced Books and Software, Pacific California.
- **Coetzee J.** (1990). E*arly growth trends observed with a Eucalyptus grandis spacing trial at Kwambonambi in the Zululand area*. Annual Research Report - Institute for Commercial Forestry Research: University of Natal, 1990, pp 203 -214.
- **Coetzee J.** (1994) *The development of Top Height with age for application to short rotation non-thinning crops. E.grandis.* Institute for Commercial Forestry Research Bulletin Series (09/94), Pietermaritzburg.
- **Coetzee J., Chiswell K., Storey P., Arbuthnot A.L.** (1996). *The final results of the E. grandis spacing trial, Kwambonambi, for age two to ten years*. Institute for Commercial Forestry Research Bulletin series (10/96), Pietermaritzburg.
- **Coetzee J., Naicker S.** (1998a). *The final results for age two to eleven years in the case of the Kia-Ora E.grandis spacing trial*. Institute for Commercial Forestry Research Bulletin Series (03/98), Pietermaritzburg.
- **Coetzee J., Naicker S.** (1998b). *The final results for age two to ten years of the Tanhurst (M5) E.grandis spacing trial*. Institute for Commercial Forestry Research Bulletin Series (11/98), Pietermaritzburg.
- **Corona P., Scotti R., Tarchiani N.** (1998). *Relationship between environmental factors and site index in Douglas-fir plantations in central Italy.* Forest Ecology and Management 110 (1-3), pp 195 – 207.
- **Cribbie R.A., Keselman H.J.** (2003). *The effects of Nonnormality on Parametric, nonparametric, and model comparison approaches to pairwise comparisons*. Educational and Psychological Measurement, Vol. 63 No. 4, pp 615-635.
- **Curt T., Bouchand M., Agrech G.** (2001). *Predicting site index of Douglas-Fir plantations*

from ecological variables in the Massif Central area of France. Forest Ecology and Management 149 (1-3), pp 61 – 74.

- **Dalgaard P.** (2008). *Introductory Statistics with R. Second Edition*. Springer Texts in Statistics.
- **De Ville B.** (2006). *Decision trees for business intelligence and data mining: using SAS © Enterprise Miner™*. Cary. NC: SAS Institute Inc.
- **De'ath G.** (2002). Multivariate regression Trees: A *new technique for modelling Species-Environment relationships*. Ecological Society of America 83(4), pp 1105 – 1117.
- **De'ath G., Fabricius K.E.** (2000). *Classification and Regression Trees : A powerful yet simple technique for ecological data analysis*. Ecological Society of America. 81(11), pp 3178 - 3192 .
- **Dobbertin M., Biging G.S.** (1998). *Using the non-parametric classifier CART to model forest tree mortality*. Forest Science 44(4), pp 507-516.
- **Dunnet C.W** (1980). *Pairwise Multiple Comparisons in the Unequal Variance Case.* Journal of the American Statistical Association 75, No 372, pp. 796 – 800.
- **Duvemo K.** (2009). *The Influence of Data Uncertainty on Planning and Decision Processes in Forest Management.* Doctoral Thesis. Swedish University of Agricultural Sciences. Umeå.
- **Duvemo K., Lämås T.** (2006)**.** *The influence of forest data quality on planning processes in forestry*. Scandinavian Journal of Forest Research 21, pp 327-339.
- **Eid T.** (2000). *Use of Uncertain Inventory Data in Forestry Scenario Models and Consequential Incorrect Harvest Decisions.* Silva Fennica 34(2), pp 89 – 100.
- **Elith J., Leathwick J.R., Hastie T.** (2008). *A working guide to boosted regression trees.* Animal Ecology 77, pp 802-813.
- **Ercanli I., Gunlu A., Altun L., Baskent E.Z.** (2008). *Relationship between site index of oriental spruce (Picea orientalis) and ecological variables in Macka, Turkey*. Scandinavian journal of forest research. 23(4), pp 319 – 329.
- **Everitt B.S., Hothorn T**. (2010). *A handbook of statistical analysis using R. Second Edition*. Chapman and Hall, Boca Raton, FL.
- **Fan Z., Kabrick J.M., Shifley S.R.** (2006). *Classification and regression tree based survival analysis in oak-dominated forest of Missouri's Ozark highlands*. Canadian Journal

of Forest Research 36, pp 1740-1748.

- **Faraway J.** (2006). *Extending the linear Model with R*. Chapman and Hall, London.
- **Fletcher Y.** (2006). *Growth and yield model configurations and coefficients currently in use in Mondi Business Paper and Mondi Shanduka Newsprint*. Third edition. Mondi internal document.
- **Fletcher Y.** (2010). Data supplied via personal correspondence.
- **Fox J.** (2002). *Linear Mixed Models.* Appendix to an R and S-PLUS Companion to Applied Regression.
- **García O.** (1983). *A stochastic Differential Equation Model for Height Growth of Forest Stands*. International Biometric Society. Biometrics, Vol 39 (4), pp 1059 – 1072.
- **García O.** (2004). *Site Index: Concepts and Methods*. Second International Conference on Forest Measurements and Quantitative Methods and Management, Arkansas USA.
- **García O.** (2005). *Comparing and Combining Stem Analysis and Permanent Sample Plot Data in Site Index Models*. Forest Science Vol 51 (4), pp 277 – 283.
- **García O.** (2010). *Dynamical implications of the variability representation in site-index modelling.* European Journal of Forest Research. Springer 2010
- **Gehrke J., Ramakrishnan R., Ganti V.** (2000). *Rainforest – A Framework for Fast Decision Tree Construction of Large Datasets*. Data Mining and Knowledge Discovery 4, pp 127 – 162.
- **Grey D.C.** (1979a). *Site Quality prediction for Pinus patula in the Glengarry area, Transkei*. South African Forestry Journal 111, pp 44 – 48.
- **Grey D.C.** (1979b). *Soil classification and site index of Pinus patula.* South African Forestry Journal 111, pp 64 – 65.
- **Grey D.C.** (1989). *Site Index - A Review*. South African Journal of Forestry Volume 148. Issue 1. March 1989, pp 28 - 32.
- **Guisan A., Zimmermann N.E.** (2000). *Predictive habitat distribution models in ecology*. Ecological Modelling 135, pp 147 – 186.
- **Hattingh N.** (2010). Data supplied via personal correspondence.
- **Holström H.** (2001). *Data acquisition for Forestry Planning by Remote Sensing Based*

Sample Plot Imputation. Doctoral Thesis. Swedish University of Agricultural Sciences. Umeå.

- **Hornik K., Buchta C., Hothorn T., Karatzoglou A., Meyer D., Zeileis A.** (2006). *The RWeka Package*. Version 0.2-4. An R interface to Weka.
- **Husch B**. (1956). *Use of age at DBH as a variable in the Site Index concept*. Journal of Forestry 54:340.
- **Husch B., Beers T.W., Kershaw J. A. Jr.** (2003). *Forest Mensuration. Fourth Edition*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- **Iverson L.R., Prasad A.M.** (2002). *Potential distribution of tree species habitat under five climate change scenarios in the eastern US.* Forest ecology and Management 155, pp 205 – 222.
- **Johnson G.R., Sniezko R.A., Mandel N.L.** (1997). A*ge Trends in Douglas-fir Genetic Parameters and implications of Optimum Selection Age.* Silvae Genetica Vol. 46, 6.
- **Johnston D.R., Grayson A.J., Bradley R.T**. (1967). *Forest Planning*. Faber and Faber London.
- **Kangas A.S.** (2009). *Value of forest information.* European Journal of Forest Research. Springer-Verlag.
- **Kätsch C.** (2006). P*recision Forestry and Information – Information Management a forgotten task* ? Proceedings of the international precision forestry symposium. Stellenbosch University, pp 175 – 186.
- **Kaufmann M.R., Ryan M.G**. (1986). *Physiographic, stand, and environmental effects on individual tree growth and growth efficiency in sub alpine forests*. Tree Physiology Vol 2, pp 47 – 59.
- **Kimsey M.J., Moore J., McDaniel P.** (2008). *A Geographically Weighted Regression Analysis of Douglas-Fir Site Index in North Central Idaho*. Forest Science 54(3), pp 356 – 366.
- **Klemmt H-J.** (2007). *Standortabhängige Ableitung der Höhenwuchsleistung aus Forstinventurdaten mit Hilfe von Data-Mining-Methoden. Grundlage für die regionale, standortbezogene Feinjustierung des forstlichen Wuchsmodells SILVA*. Unpublished doctoral dissertation. Technical University of Munich

- **Kunneke A.** (2011). Water balance data as supplied by A. Kunneke. University of Stellenbosch.
- **Kunz R.P.** (2004). *Forestry Productivity Toolbox (FPT)*. Institute for Commercial Forestry Research, Pietermaritzburg, RSA.
- Kunz R.P., Pallett R.N. (2000). A stratification system based on climate and lithology for locating commercial forestry permanent sample plots. Institute for Commercial Forestry Research Bulletin Series 01/00, Pietermaritzburg.
- **Li P.** (2005). *Box-Cox Transformations: An Overview*. Department of Statistics, University of Connecticut.
- **Liaw A., Wiener M.** (2002). *Classification and Regression by randomForest.* R news Vol. 2/3, pp 18 – 22.
- **Loetsch F., Zöhrer F., Haller K.E.** (1973). *Forest Inventory Volume 2.* BLV Verlagsgesellschaft, München.
- **Louw J.H.** (1997). *A site-growth study of Eucalyptus grandis in the Mpumalanga escarpment area*. South African Forestry Journal 180, pp 1 – 13.
- **Louw J.H., Germishuizen I., Smith C.W.** (2011). *A stratification of the South African forestry landscape based on climatic parameters.* Southern Forests: a Journal of Forest Science 73:1, pp 51-62.
- **Louw J.H., Scholes M.C.** (2002). *Forest site classification and evaluation: a South African perspective*. Forest Ecology and Management 171, pp 153 – 168.
- **Louw J.H., Scholes M.C.** (2006). *Site index functions using site descriptors of Pinus patula plantations in South Africa*. Forest Ecology and Management 225, pp 94 – 103.
- **Magnusson M.** (2006). *Evaluation of Remote Sensing Techniques for Estimation of Forest Variables at Stand Level.* Doctoral Thesis. Swedish University of Agricultural Sciences. Umeå.
- **Maindonald J.H., Braun W.J.** (2007). *Data Analysis and Graphics Using R – an Example-Based Approach*. Cambridge University Press.
- **Martín-Benito D., Gea-Izquierdo G., del Río M., Cañellas I.** (2008). *Long-term trends in dominant-height growth of black pine using dynamic models.* Forest Ecology and Management 256, pp 1230-1238.

- **McFarlane D.W., Green E.J., Burkhart H.E.** (2000). *Population density influences assessment and application of site index*. Canadian Journal of Forestry Research 30, pp 1472 – 1475.
- **McKenney D.W., Pedlar J.H.** (2003). *Spatial models of site index based on climate and soil properties for two boreal tree species in Ontario, Canada*. Forest Ecology and Management 175, pp 497 – 507.
- **Meredieu C., Perret S., Dreyfus P.** (2003). *Modelling Dominant Height Growth: Effect of Stand Density*. Chapter 10, Modelling forest systems, (eds. Amaro A., Reed D., Soares P) . CAB International 2003.
- **Moisen G.G., Frescino T.S.** (2002). *Comparing five modelling techniques for predicting forest characteristics*. Ecological Modelling 157, pp 209 – 225.
- **Morkel R.** (2005). *Internal Mondi Business Paper presentation on Planning time line and Process.*
- **Muller R., Mockel M.** (2008). *Logistic regression and CART in the analysis of multimarker studies*. Clinica Chimica Acta 394, pp 1 – 6.
- **Muňoz J., Ángel M.** (2004). *Comparison of statistical methods commonly used in predictive modelling*. Journal of Vegetation Science 15, pp 285-292.
- **Murthy S.** (1998). *Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey*. Data Mining and Knowledge Discovery 2, pp 245 – 389. Kluwer Academic Publishers, Boston.
- **Nakai M., Ke W.** (2009). *Statistical Models for Longitudinal Data Analysis*. Applied Mathematical Sciences, Vol. 3, 2009 no 40, pp 1979 – 1989.
- **Neville P.G.** (1998). *Growing Trees for Stratified Modeling*. SAS Institute, Inc. Cary, NC 27513.
- **Nigh G.D., Love B.A**. (1999). *How well can we select undamaged site trees for estimating site index?* Canadian Journal of Forest Research Vol 29, pp 1989 – 1999.
- **Philip M.S.** (1994). *Measuring trees and Forests*. Second edition. CAB international. University press Cambridge.
- **Pienaar L.V.** (1965). *Quantitative theory of forest growth*. (pp 32 – 34, 101 - 107) Doctoral thesis, University of Washington.

- **Prasad A.M., Iverson L.R., Liaw A.** (2006). *Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction*. Ecosystems 9, pp 181 – 199.
- **Quinlan J.R.** (1992). *Learning with Continuous classes*. In Proceedings AI'92 (Adams and Sterling, Eds), Singapore: World Scientific 1992, pp 343 – 348.
- **R Development Core Team** (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.
- **Raley E.M., Gwaze D.P., Byram T.D**. (2003). *An Evaluation of Height as an Early Selection Criterion for Volume and Predictor of Site Index Gain in the Western Gulf*. Proceedings of the 27th Southern Forest Tree Improvement Conference, Oklahoma, pp 45 - 55.
- **Raty M., Kangas A.** (2008). *Localizing general models with classification and regression trees*. Scandinavian Journal of Forest Research 23:5, pp 419 – 430.
- **Rose C.E., Cieszewski C.J., Carmean W.H.** (2003). *Three methods for avoiding the impacts of incompatible site index and height prediction models demonstrated on jack pine curves for Ontario.* The Forestry Chronicle Vol 79 no 5, pp. 928 – 935.
- **Ryan P.J., McKenzie N.J., O'Connell D.O., Loughhead A.N., Leppert P.M., Jacquier D., Ashton L.** (2000). *Integrating forest soils information across scales: spatial prediction of soil properties under Australian forests*. Forest Ecology and Management 138, pp 139 -157.
- **Saigol Z.** (2009). *Pearl: Causation, Action, and Conterfactuals. IR Lab, School of computer Science. University of Birmingham.*
- **Sakar D**. (2008). *Fitting Mixed-Effects Models Using the lme4 Package in R*. Fred Hutchinson Cancer research Center.
- **Sakia R.M.** (1992). *The Box-Cox transformation technique: a review*. The statistician 41, pp 169 – 178.
- **Salford Systems.** (no date). *Critical Features of High performance Decision Trees*. Salford Systems 8880 Rio San Diego Drive, Ste. 1045, San Diego, CA. www.salford-systems.com .
- **Sánchez-Rodriguez F., Rodriguez-Soalleiro R., Español E., López C.A., Merino A.**

(2002). *Influence of edaphic factors and tree nutritive status on the productivity of Pinus radiata D.Don plantations in north-western Spain*. Forest Ecology and Management 171 (1-2), pp 181-189.

- **Schafer G.N.** (1988a). *A site growth model for Pinus elliottii in the Southern Cape.* South African Forestry Journal 146, pp 12 – 17.
- **Schafer G.N.** (1988b). *A site growth model for Pinus pinaster in the Southern Cape*. South African Forestry Journal 146, pp 18 – 22.
- **Schönau A.P.G., Coetzee J.** (1989). *Initial spacing, stand density and thinning in Eucalyptus plantations*. Forest Ecology and Management 29, pp 245 – 266.
- **Schulze R.E.** (1997). *South African Atlas of Agrohydrology and – Climatology*. Water Research Commission, Pretoria, Report TT82/96.
- **Seifert T.** (2011). Personal correspondence.
- **Sharma M., Amateis R.L., Burkhart H.E.** (2002). *Top height definition and its effect on site index determination in thinned and unthinned loblolly pine plantations*. Forest Ecology and Management 168, pp 163 - 175.
- **Sheather S.J.** (2009). *A Modern Approach to Regression with R*. Springer Texts in Statistics.
- **Skovsgaard J.P., Vanclay J.K.** (2008). *Forest site productivity: a review of the evolution of dendrometic concepts for even-aged stands*. Journal of Forestry, Vol 81, pp 13 – 31.
- **Smith C.W., Kassier H.W., Cunningham L**. (2005). *The effect of stand density and climatic conditions on the growth and yield of Eucalyptus grandis*. Institute for Commercial Forestry Research Bulletin Series 09/05, Pietermaritzburg.
- **Smith, C.W., Pallett, R.N., Kunz, R.P., Gardner R.A.W., du Plessis, M.** (2005). *A strategic forestry site classification for the summer rainfall region of Southern Africa based on climate, geology and soils.* Institute for Commercial Forestry Research Bulletin Series 03/05, Pietermaritzburg.
- **Strobl C., Malley J., Tutz G.** (2009). *An introduction to Recursive Partitioning: Rational, Application and Characteristics of Classification and Regression Trees, Bagging and Random Forests.* Technical Report Number 55, 2009. Department of Statistics, University of Munich.

- **Subasinghe S.M.C.U.P.** (2008). *Growth models and their use in plantation forestry*. Silver Jubilee proceedings of the department of Forestry and Environmental Science, University of Sri Jayewardenepura, December 2008.
- **Therneau T.M., Atkinson B., Ripley B.** (2010). *Recursive Partitioning.* rpart, R package version 3.1-46
- **Therneau T.M., Atkinson E.J.** (2011). *An Introduction to Recursive Partitioning Using the RPART Routines.* Mayo Foundation
- **Thuiller W., Araujo M.B., Lovorel S.** (2003). G*eneralized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales*. Journal of Vegetation Science 14, pp 669 – 680.
- **Torgo L.** (1997). *Functional models for Regression Tree Leaves.* Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997). Fisher, D. H (Ed.), pp 385 – 393.
- **Ture M., Tokatli F., Kurt I.** (2009). *Using Kaplan-Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4.5 and ID3) in determining recurrence-free survival of breast cancer patients*. Expert Systems with Applications 36, pp 2017 – 2026.
- **van Aardt J.A.N., Norris-Rogers M.** (2008). *Spectral-age interactions in managed, even-aged Eucalyptus plantations: application of discriminant analysis and classification and regression trees approaches to hyperspectral data*. International journal of Remote Sensing, 29:6, pp 1841 – 1845.
- **van Diepen M., Franses P.H.** (2006). *Evaluating chi-squared automatic interaction detection*. ScienceDirect. Information Systems 31, pp 814 – 831.
- **van Laar A.** (1978). *The growth of Unthinned Pinus Patula in Relation to Spacing*. South African Forestry Journal (107), pp 3 – 11.
- **van Laar A., Akça A.** (1997). *Forest Mensuration.* Cuvillier Verlag, Göttingen.
- **van Laar A., Bredenkamp B.V.** (1979). T*he effect of initial spacing on some growth parameters of Eucalyptus grandis*. South African Forestry Journal 111, pp 58 – 63.
- **Vanclay J.K.** (1994). *Modelling Forest Growth and Yield. Applications to Mixed Tropical Forests* CAB International, Walingford UK.
- **Wang G.G., Klinka K.** (1996). *Use of synoptic variables in predicting white spruce site*

index. Forest Ecology and Management 80 (1-3), pp 95 – 105.

- **Wang X., Song B., Chen J., Crow T.R., LaCroix J.J.** (2006). *Challenges in Visualizing Forests and Landscapes*. Forest Ecology. Journal of Forestry. Sept 2006, pp 316 – 319.
- **Wang Y., Raulier F., Ung C.H.** (2005). *Evaluation of spatial predictions of site index obtained by parametric and nonparametric methods – A case study of lodgepole pine productivity.* Forest Ecology and Management 214, pp 201 – 211.
- **Wang Y., Witten I.H.** (1996). *Induction of Model Trees for Predicting Continuous Classes*. Working paper 96/23. Department of Computer Science. The University of Waikato Hamilton, New Zealand.
- **West P.W.** (2004). *Tree and Forest Measurement.* Springer-Verlag Berlin Heidelberg.
- **Wilkinson L.** (1992). *Tree Structured Data Analysis: AID, CHAID and CART.* Paper presented at the 1992 Sun Valley, ID, Sawtooth/SYSTAT Joint software Conference.
- **Zumrawi A.A.M.A** (1986). *Effects of stand density on Site Index in thinned stands of Douglas-fir in the Pacific Northwest*. Masters thesis. Oregon State University.

APPENDIX 1 Random effects specification

Random effects specification.

A model with correlated random effects:

lmer(Hdom ~ 1 + logAGE + (1+ logAGE|Plotid), data = mydata1)

A model with uncorrelated random effects: here we have a simple scalar random effect for `Plotid`, and a random effect for the slope with respect to `logAGE`, indexed by `Plotid`:

lmer(Hdom ~ 1 + logAGE + (1|Plotid) + (0*+logAGE|Plotid), data = mydata1)

Models with both a random intercept and random slope with respect to `TPH0` for each site :

lmer (Hdom ~ logAGE + (TPH0 | Plotid), data = mydata1)

Next we can examine a model with an intercept and slope for each site, but assuming independence of these random effects with :

lmer (Hdom ~ logAGE + TPH0 + (1|Plotid) + (TPH0-1 | Plotid), data = mydata1)

This model has a random intercept and a random slope for each site (as does the previous model), however, in this model these random effects are assumed to be independent within site.

* The intercept is implicit in linear models – to suppress it we can use 0 + term or term – 1, it needs to be explicit if it is the only term in the expression.

APPENDIX 2 Abbreviated Species names

Abbreviated species names

ECAM	*Eucalyptus*	*camaldulensis*
ECLO		*cloeziana*
EDUN		*dunnii*
EELA		*elata*
EEMA		*maculata*
EFAS		*fastigata*
EFRA		*fraxinoides*
EG+M		*grandis and macarthurii*
EGRA		*grandis*
EGXC		*grandis cross camaldulensis*
EGXU		*grandis cross urophilla*
EGXN		*grandis cross nitens*
EGXT		*grandis cross tereticornis*
EMAC		*macarhturii*
EMIX		*mixed species*
ENIT		*nitens*
EREG		*regnans*
ERUB		*rubida*
ESAL		*saligna*
ESMI		*smithii*
EURO		*urophilla*
PCAR	*Pinus*	*caribaea*
PE+R		*elliotti and radiata*
PE+T		*elliottii and taeda*
PECH		*elliottii cross caribaea var hondurensis*
PELL		*elliotti*
PGRE		*greggii*
PKES		*kesiya*
PMIX		*mixed species*
PP+E		*patula and elliottii*
PP+T		*patula and taeda*
PPAT		*patula*
PPSE		*pseudostrobus*
PROX		*roxburghii*
PTAE		*taeda*
PTEC		*tecunumannii*

+ donates the two separate species in the same compartment

x donates a hybrid cross (generally cloned)

APPENDIX 3 List of Acronyms

Acronym	Meaning
Apan_evap_08	Potential Apan evapotranspiration for August (09 = September etc.) mm
APO	Annual plan of operations
BA	Basal Area
CART	Classification and Regression Trees
Cp /CP	Complexity parameter
df	Degrees of freedom
E. g x c	*E. grandis x E. camaldulensis*
E. g x t	*E. grandis x E. tereticornis*
E. g x u	*E. grandis x E. urophylla*
GIS	Geographic information system
ha	Hectare
Hdom / HD	Dominant height
HSS	Harvest Scheduling System
ICFR	Institute for Commercial Forestry Research
logAge	Natural logarithm of age
MAI_{max}	Maximum Mean Annual Increment (m^3/ha/yr)
MAI_n	Mean Annual Increment (m^3/ha/yr), to base age n
Mgs_duration	Mean growth season duration (days)
MLR	Multiple linear regression
Mondi	Mondi South Africa – a pulp, wood chip and liner board company with approximately 327 000 ha of commercial plantations under management. http://www.mondigroup.com/desktopdefault.aspx/tabid-349/
NPV	Net Present Value
PSP	Permanent Sample Plot
REGWQ	Ryan, Einot, Gabriel, Welsch Q test
RSME	Root square mean error
SAPPI	Sappi Forests South Africa – a pulp and saw timber forestry company with approximately 489 000 ha of commercial plantations under management (excluding Swaziland). http://www.sappi.com/regions/sa/SappiSouthernAfrica/Sappi%20Forests/Pages/default.aspx
SI	Site Index
SPHA	Stems Per Hectare
Spp	Species
SteClsCli	Site Classification by Climate
TPH	Current stems per Hectare
TPH0	Initial planted stems per Hectare
TSP	Temporary Sample Plot
VOI	Value of Information
WBJUN	Water Balance for June (SEP = September etc.) mm

APPENDIX 4 Variables considered in modelling

Site variable	Unit of measure.
Altitude 200m	m
Solar Radiation	$MJ.m^{-2}.day^{-1}$ (by month)
Mean Annual Precipitation (2003)	mm
Rainfall Concentration	%
Rainfall Seasonality	Seasons
Means Of Daily Maximum Temperature	^{o}C (by month)
Means Of Daily Minimum Temperature	^{o}C (by month)
Daily Mean Temperature	^{o}C (by month)
Temperature Range (T - T)	^{o}C (by month)
Mean Annual Temperature	^{o}C
Heat Units	^{o}days (by month)
Average First Date of Heavy Frost	Day of year
Average Last Date of Heavy Frost	Day of year
Average Duration of Frost Period	Days
Average Number of Days with Frost	Days
Standard Deviation of Number of Days with Frost	Days
Daily Mean Relative Humidity	% (by month)
Daily Minimum Relative Humidity	% (by month)
Potential Evaporation	mm (by month)
Potential Evaporation Mean Annual	mm
Potential Evapotranspiration	mm (by month)
Wilting Point top soil - 84 soil zones	mm
Grid Wilting Point top soil - 84 soil zones	mm
Wilting Point sub soil - 84 soil zones	mm
Grid Top soil to sub soil daily drainage fraction	fraction
Grid sub soil daily drainage fraction	fraction
Grid Initial Crop Numbers (Acocks)	ACRU Crop Number
Moisture Growing Season Mean Start of Season	month
Moisture Growing Season Mean End of Season	month
Moisture Growing Season Duration of Season	day
Gross Irrigation Requirements Median Annual	mm
Mean Annual Precipitation or MAP	mm
Probability of obtaining < 650 mm of annual rainfall in any given year	%
Probability of obtaining > 850 mm of annual rainfall in any given year	%
Mean Monthly Precipitation	mm (by month)
Mean Annual Temperature or MAT	^{o}C
Site classification based on climate	CT=Cool temperate; WT=Warm temperate; ST=Sub-tropical
Monthly means of Minimum daily Temperature	^{o}C (by month)
Monthly means of Maximum daily Temperature	^{o}C (by month)
Total Annual Potential: A-pan equivalent Evaporation	mm
Mean Monthly A-pan Evaporation	mm (by month)
Total Annual Solar radiation	$MJ/m^2/day$

Appendix 4 continued.

Site variable	**Unit of measure.**
Mean Monthly Solar radiation	$MJ/m^2/day$(by month)
Topsoil texture-from the 1:250 000 scale land types	~
Total soil depth-from the 1:250 000 scale land types	mm
Permanent Wilting Point of topsoil horizon	mm/m
Field Capacity (Drained Upper Limit) of topsoil horizon	mm/m
Total Porosity of topsoil horizon	mm/m
Geology-from the 1:1 000 000 scale geology map	~
Lithology-from the 1:1 000 000 scale geology map	~
Physiographic region	refer to Kunz & Pallet (2000)
Soil texture derived from parent material	~
Soil depth derived from parent material	mm
Wilting point derived from parent material	mm/m
Field capacity derived from parent material	mm/m
Total porosity derived from parent material	mm/m
Altitude from the 1`x1` of a degree grid	m
Slope derived from the 1`x1` altitude grid	Deg
Aspect derived from the 1`x 1` altitude grid	Deg
Altitude derived from the 1:200/400m altitude grid	m
Slope derived from the 1:200/400m altitude grid	Deg
Aspect derived from the 1:200/400m altitude grid	Deg
Water balance	mm (by month)

APPENDIX 5 Site Classification based on Climate

	Cold		
MAT °C	Below 10		
Class	Cold	Cold	Cold
	Dry	Moist	Wet
MAP mm	<700	700-800	>800

	Cool Temperate								
MAT °C		10-14			14-15			15-16	
Class	CT1	CT2	CT3	CT4	CT5	CT6	CT7	CT8	CT9
	Dry	Moist	Wet	Dry	Moist	Wet	Dry	Moist	Wet
MAP mm	<700	700-800	>800	<800	800-900	>900	<825	825-925	>925

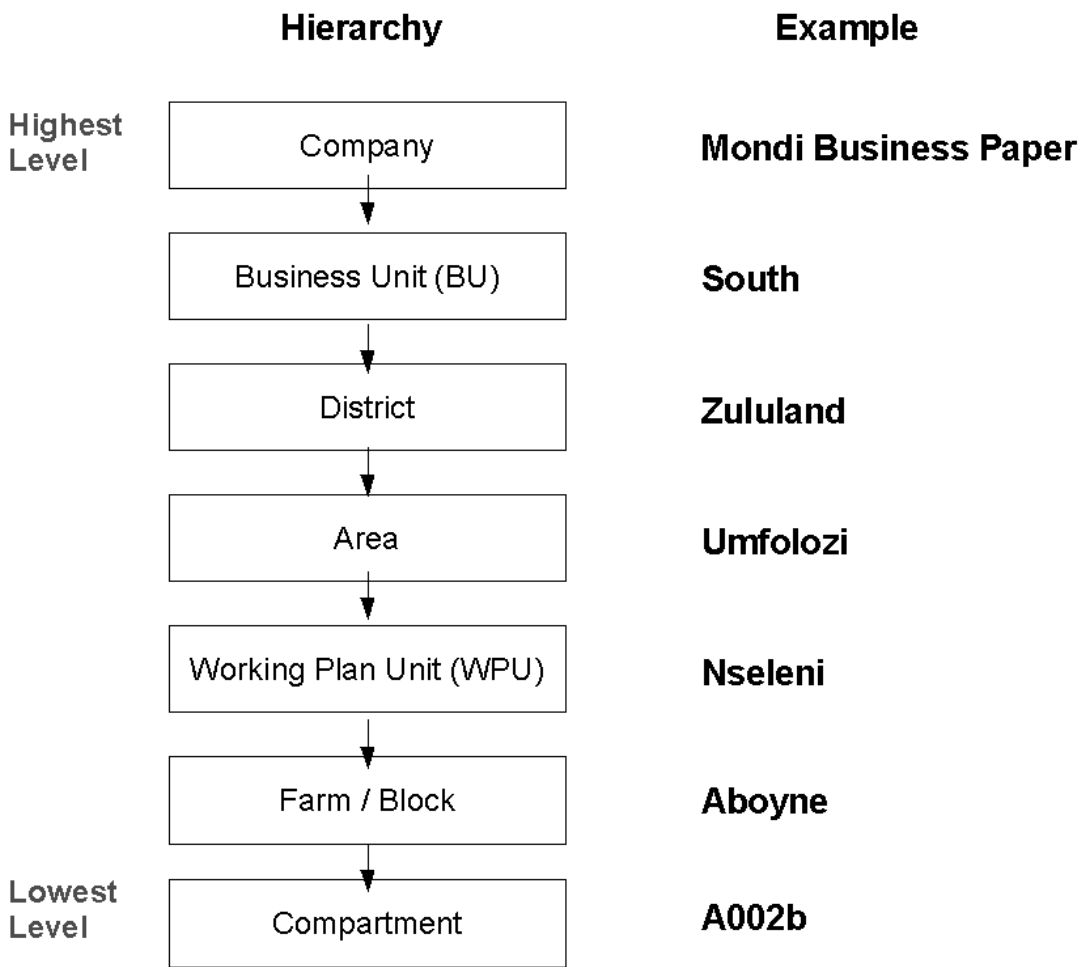
	Warm Temperate								
MAT °C		16-17			17-18			18-19	
Class	WT1	WT2	WT3	WT4	WT5	WT6	WT7	WT8	WT9
	Dry	Moist	Wet	Dry	Moist	Wet	Dry	Moist	Wet
MAP mm	<850	850-950	>950	<875	875-975	>975	<900	900-1000	>1000

	Sub Tropical								
MAT °C		19-20			20-21			21-22	
Class	ST1	ST2	ST3	ST4	ST5	ST6	ST7	ST8	ST9
	Dry	Moist	Wet	Dry	Moist	Wet	Dry	Moist	Wet
MAP mm	<925	925-1025	>1025	<950	950-1050	>1050	<975	975-1075	>1075

	Tropical		
MAT °C	Above 22		
Class	Tropical	Tropical	Tropical
	Dry	Moist	Wet
MAP mm	<975	975-1075	>1075

Smith C.W., Pallett R.N., Kunz R.P., Gardner R.A.W., du Plessis M. (2005). *A strategic forestry site classification for the summer rainfall region of Southern Africa based on climate, geology and soils.* ICFR Bulletin Series 03/05, Pietermaritzburg.

APPENDIX 6 Geographic hierarchy



All databases have some form of hierarchy to enable collation analysis and reporting of data on various levels , all forestry companies have a geographic hierarchy for the same reasons. This geographic hierarchy is normally a refection in some way of the management hierarchy.

APPENDIX 7 Results of the REGWQ test on the *Pinus* data

Ryan – Einot – Gabriel – Welsch Q Pairwise Multiple Comparison Test.			
Pair wise comparisons	***t* statistic**	**Adjusted *p*-value**	**H_0 rejected (95 %)**
C: 4 - 6-D: 6 - 8	1.7100	0.22630	
C: 4 - 6-E : 8 - 10	3.3600	0.04640	
C: 4 - 6-F : 10 - 12	4.0000	0.02420	
C: 4 - 6-G : 12 - 14	4.3200	0.01940	*
C: 4 - 6-H : 14 - 16	4.7800	0.01310	*
C: 4 - 6-I : 16 - 18	4.6400	0.01330	*
C: 4 - 6-J : 18 - 20	4.9600	0.01090	*
C: 4 - 6-K : 20 - 22	5.2800	0.00590	*
C: 4 - 6-L : 22 - 24	5.4500	0.00460	*
C: 4 - 6-M : 24 - 26	6.8300	0.00010	*
C: 4 - 6-N : 26 - 28	5.2900	0.00850	*
C: 4 - 6-O : > 28	5.0500	0.01870	*
D: 6 - 8-E : 8 - 10	3.5500	0.01200	
D: 6 - 8-F : 10 - 12	5.8900	0.01180	
D: 6 - 8-G : 12 - 14	6.5700	0.00000	*
D: 6 - 8-H : 14 - 16	7.7600	0.00000	*
D: 6 - 8-I : 16 - 18	7.6800	0.00000	*
D: 6 - 8-J : 18 - 20	8.3900	0.00000	*
D: 6 - 8-K : 20 - 22	9.0100	0.00000	*
D: 6 - 8-L : 22 - 24	9.0500	0.00000	*
D: 6 - 8-M : 24 - 26	12.3400	0.00000	*
D: 6 - 8-N : 26 - 28	7.6800	0.00000	*
D: 6 - 8-O : > 28	5.7400	0.00250	*
E : 8 - 10-F : 10 - 12	1.4800	0.29560	
E : 8 - 10-G : 12 - 14	2.3800	0.21230	
E : 8 - 10-H : 14 - 16	3.7000	0.06770	
E : 8 - 10-I : 16 - 18	3.4100	0.07500	
E : 8 - 10-J : 18 - 20	4.2900	0.02930	
E : 8 - 10-K : 20 - 22	5.1100	0.02720	
E : 8 - 10-L : 22 - 24	5.3600	0.03110	
E : 8 - 10-M : 24 - 26	8.9600	0.00000	*
E : 8 - 10-N : 26 - 28	4.4400	0.04450	
E : 8 - 10-O : > 28	3.4600	0.29870	
F : 10 - 12-G : 12 - 14	1.8400	0.19440	
F : 10 - 12-H : 14 - 16	4.3600	0.01570	
F : 10 - 12-I : 16 - 18	4.4300	0.01180	
F : 10 - 12-J : 18 - 20	5.9200	0.01950	
F : 10 - 12-K : 20 - 22	6.8800	0.02340	
F : 10 - 12-L : 22 - 24	6.5700	0.02720	
F : 10 - 12-M : 24 - 26	12.2000	0.00000	*
F : 10 - 12-N : 26 - 28	4.4200	0.03760	
F : 10 - 12-O : > 28	3.0400	0.44000	
G : 12 - 14-H : 14 - 16	2.2400	0.25270	
G : 12 - 14-I : 16 - 18	1.7200	0.22330	
G : 12 - 14-J : 18 - 20	3.3200	0.08840	

Appendix 7 continued.

Ryan – Einot – Gabriel – Welsch Q Pairwise Multiple Comparison Test.			
G : 12 - 14-K : 20 - 22	4.5800	0.01950	
G : 12 - 14-L : 22 - 24	4.7400	0.02340	
G : 12 - 14-M : 24 - 26	10.0400	0.00000	*
G : 12 - 14-N : 26 - 28	3.3400	0.21580	
G : 12 - 14-O : > 28	2.4800	0.65170	
H : 14 - 16 -J : 18 - 20	0.8600	0.54410	
H : 14 - 16 -K : 20 - 22	2.4100	0.20430	
H : 14 - 16 -L : 22 - 24	2.9000	0.16970	
H : 14 - 16 -M : 24 - 26	8.2400	0.00000	*
H : 14 - 16 -N : 26 - 28	2.0600	0.59250	
H : 14 - 16 -O : > 28	1.7500	0.81710	
I : 16 - 18-H : 14 - 16	0.9700	0.49280	
I : 16 - 18-J : 18 - 20	2.1900	0.26750	
I : 16 - 18-K : 20 - 22	3.7800	0.03810	
I : 16 - 18-L : 22 - 24	4.0000	0.03760	
I : 16 - 18-M : 24 - 26	9.9000	0.00000	*
I : 16 - 18-N : 26 - 28	2.6600	0.41580	
I : 16 - 18-O : > 28	2.0500	0.77480	
J : 18 - 20-K : 20 - 22	1.8000	0.20440	
J : 18 - 20-L : 22 - 24	2.3900	0.20790	
J : 18 - 20-M : 24 - 26	8.0500	0.00000	*
J : 18 - 20-N : 26 - 28	1.6500	0.64650	
J : 18 - 20-O : > 28	1.5100	0.82220	
K : 20 - 22-L : 22 - 24	0.8400	0.55480	
K : 20 - 22-M : 24 - 26	6.1200	0.00020	*
K : 20 - 22-N : 26 - 28	0.6200	0.89870	
K : 20 - 22-O : > 28	0.9400	0.91010	
L : 22 - 24-M : 24 - 26	4.7500	0.00440	*
L : 22 - 24-N : 26 - 28	0.0300	0.98240	
L : 22 - 24-O : > 28	0.5900	0.90810	
N : 26 - 28-M : 24 - 26	3.5500	0.03270	
N : 26 - 28-O : > 28	0.5300	0.70640	
O : > 28-M : 24 - 26	1.5600	0.27050	

APPENDIX 8 Summary of the Site data

```
    Lon_DdD         Lat_DdD            SI              Spp            MAP          rskLE650         rskGT850       MeaMthPreJan
Min.   :29.39   Min.   :24.85   Min.   :10.02   EGRA   :2434   Min.   : 618   Min.   : 0.000   Min.   :  3.60   Min.   : 78.0
1st Qu.:30.33   1st Qu.:25.88   1st Qu.:18.42   PPAT   :1944   1st Qu.: 840   1st Qu.: 4.300   1st Qu.: 51.30   1st Qu.:132.0
Median :30.58   Median :28.54   Median :20.85   PELL   :1699   Median : 909   Median : 8.300   Median : 61.20   Median :144.0
Mean   :30.70   Mean   :27.84   Mean   :21.61   EGXU   : 591   Mean   : 943   Mean   : 9.943   Mean   : 62.49   Mean   :146.5
3rd Qu.:30.87   3rd Qu.:29.38   3rd Qu.:24.03   ENIT   : 570   3rd Qu.:1018   3rd Qu.:13.100   3rd Qu.: 75.50   3rd Qu.:159.0
Max.   :32.30   Max.   :30.71   Max.   :41.60   EDUN   : 531   Max.   :1635   Max.   :67.900   Max.   :100.00   Max.   :285.0
                                                (Other):2434

 MeaMthPreFeb    MeaMthPreMar    MeaMthPreApr    MeaMthPreMay    MeaMthPreJun    MeaMthPreJul    MeaMthPreAug    MeaMthPreSep
Min.   : 78.0   Min.   : 67.0   Min.   : 34.00   Min.   :  8.00   Min.   : 4.00   Min.   : 4.00   Min.   : 5.00   Min.   : 19.00
1st Qu.:113.0   1st Qu.: 94.0   1st Qu.: 47.00   1st Qu.: 19.00   1st Qu.: 9.00   1st Qu.:10.00   1st Qu.:13.00   1st Qu.: 34.00
Median :130.0   Median :109.0   Median : 54.00   Median : 24.00   Median :12.00   Median :13.00   Median :24.00   Median : 47.00
Mean   :132.5   Mean   :112.5   Mean   : 57.05   Mean   : 29.07   Mean   :16.54   Mean   :17.06   Mean   :24.16   Mean   : 48.99
3rd Qu.:147.0   3rd Qu.:129.0   3rd Qu.: 61.00   3rd Qu.: 29.00   3rd Qu.:16.00   3rd Qu.:17.00   3rd Qu.:31.00   3rd Qu.: 59.00
Max.   :287.0   Max.   :211.0   Max.   :110.00   Max.   :121.00   Max.   :67.00   Max.   :68.00   Max.   :75.00   Max.   :118.00

 MeaMthPreOct    MeaMthPreNov    MeaMthPreDec     MeaAnnTmp        SteClsCli      MinMthTmpJan    MinMthTmpFeb    MinMthTmpMar
Min.   : 53.0   Min.   : 77.0   Min.   : 78.0   Min.   :13.10   WT1    :1054   Min.   :11.40   Min.   :11.30   Min.   :10.10
1st Qu.: 80.0   1st Qu.:110.0   1st Qu.:119.0   1st Qu.:15.70   WT2    : 935   1st Qu.:13.90   1st Qu.:13.70   1st Qu.:12.60
Median : 88.0   Median :124.0   Median :135.0   Median :16.70   CT8    : 867   Median :14.90   Median :14.80   Median :13.70
Mean   : 89.9   Mean   :123.3   Mean   :135.4   Mean   :17.10   WT3    : 792   Mean   :15.47   Mean   :15.39   Mean   :14.34
3rd Qu.: 98.0   3rd Qu.:133.0   3rd Qu.:149.0   3rd Qu.:18.10   ST9    : 787   3rd Qu.:16.40   3rd Qu.:16.30   3rd Qu.:15.30
Max.   :164.0   Max.   :197.0   Max.   :260.0   Max.   :22.00   CT5    : 748   Max.   :20.70   Max.   :20.80   Max.   :20.00
                                                                (Other):5020

 MinMthTmpApr    MinMthTmpMay    MinMthTmpJun    MinMthTmpJul    MinMthTmpAug    MinMthTmpSep    MinMthTmpOct
Min.   : 7.40   Min.   : 4.100   Min.   : 0.800   Min.   : 1.100   Min.   : 3.300   Min.   : 5.80   Min.   : 7.50
1st Qu.: 9.80   1st Qu.: 6.400   1st Qu.: 3.300   1st Qu.: 3.300   1st Qu.: 5.500   1st Qu.: 8.30   1st Qu.:10.20
Median :11.00   Median : 8.000   Median : 5.100   Median : 5.100   Median : 6.900   Median : 9.30   Median :11.10
Mean   :11.66   Mean   : 8.436   Mean   : 5.522   Mean   : 5.475   Mean   : 7.368   Mean   : 9.94   Mean   :11.67
3rd Qu.:12.50   3rd Qu.: 9.200   3rd Qu.: 6.500   3rd Qu.: 6.400   3rd Qu.: 8.100   3rd Qu.:10.60   3rd Qu.:12.50
Max.   :17.80   Max.   :15.500   Max.   :13.000   Max.   :12.700   Max.   :13.900   Max.   :15.70   Max.   :16.80

 MinMthTmpNov    MinMthTmpDec    MaxMthTmpJan    MaxMthTmpFeb    MaxMthTmpMar    MaxMthTmpApr    MaxMthTmpMay    MaxMthTmpJun
Min.   : 9.10   Min.   :10.60   Min.   :21.00   Min.   :20.60   Min.   :20.20   Min.   :18.70   Min.   :16.60   Min.   :14.40
1st Qu.:11.80   1st Qu.:13.10   1st Qu.:24.30   1st Qu.:24.20   1st Qu.:23.50   1st Qu.:21.80   1st Qu.:19.90   1st Qu.:17.80
Median :12.60   Median :14.10   Median :25.60   Median :25.60   Median :24.90   Median :23.10   Median :21.20   Median :19.10
Mean   :13.21   Mean   :14.68   Mean   :25.87   Mean   :25.78   Mean   :25.05   Mean   :23.31   Mean   :21.46   Mean   :19.30
3rd Qu.:14.10   3rd Qu.:15.60   3rd Qu.:27.10   3rd Qu.:27.00   3rd Qu.:26.30   3rd Qu.:24.60   3rd Qu.:22.80   3rd Qu.:20.50
Max.   :18.20   Max.   :19.80   Max.   :30.50   Max.   :30.10   Max.   :29.60   Max.   :27.80   Max.   :26.10   Max.   :24.30

 MaxMthTmpJul    MaxMthTmpAug    MaxMthTmpSep    MaxMthTmpOct    MaxMthTmpNov    MaxMthTmpDec      TotAnnPev       TotAnnSrd
Min.   :14.70   Min.   :16.70   Min.   :19.00   Min.   :19.10   Min.   :19.60   Min.   :20.50   Min.   :1465   Min.   :235.1
1st Qu.:17.90   1st Qu.:19.80   1st Qu.:21.60   1st Qu.:22.20   1st Qu.:22.70   1st Qu.:24.00   1st Qu.:1674   1st Qu.:253.4
Median :19.30   Median :20.90   Median :22.40   Median :23.10   Median :23.70   Median :25.30   Median :1756   Median :263.3
Mean   :19.48   Mean   :21.15   Mean   :22.85   Mean   :23.49   Mean   :24.03   Mean   :25.53   Mean   :1745   Mean   :263.6
3rd Qu.:20.80   3rd Qu.:22.50   3rd Qu.:24.30   3rd Qu.:25.00   3rd Qu.:25.50   3rd Qu.:26.80   3rd Qu.:1821   3rd Qu.:272.7
Max.   :24.20   Max.   :25.10   Max.   :26.40   Max.   :27.30   Max.   :27.80   Max.   :29.80   Max.   :1972   Max.   :298.5

 AveMthSrdJan    AveMthSrdFeb    AveMthSrdMar    AveMthSrdApr    AveMthSrdMay    AveMthSrdJun    AveMthSrdJul    AveMthSrdAug
Min.   :23.90   Min.   :22.50   Min.   :20.40   Min.   :17.80   Min.   :14.50   Min.   :12.80   Min.   :13.40   Min.   :16.70
1st Qu.:25.80   1st Qu.:24.40   1st Qu.:22.00   1st Qu.:18.90   1st Qu.:15.90   1st Qu.:14.30   1st Qu.:15.00   1st Qu.:18.10
Median :26.50   Median :25.10   Median :22.60   Median :19.70   Median :16.50   Median :14.80   Median :15.70   Median :19.10
Mean   :26.67   Mean   :25.22   Mean   :22.73   Mean   :19.69   Mean   :16.77   Mean   :15.12   Mean   :15.96   Mean   :19.15
3rd Qu.:27.50   3rd Qu.:26.10   3rd Qu.:23.40   3rd Qu.:20.40   3rd Qu.:17.70   3rd Qu.:16.10   3rd Qu.:17.00   3rd Qu.:20.30
Max.   :31.20   Max.   :28.80   Max.   :25.70   Max.   :22.00   Max.   :18.90   Max.   :17.20   Max.   :18.10   Max.   :21.50
```

Appendix 8 continued.

```
 AveMthSrdSep    AveMthSrdOct    AveMthSrdNov    AveMthSrdDec   Soil.textureA.horizon   Soil.depth        Soil.PWP
 Min.   :19.30   Min.   :21.70   Min.   :23.00   Min.   :24.30   SaClLm :4882          Min.   : 225.2   Min.   : 50.0
 1st Qu.:21.70   1st Qu.:24.60   1st Qu.:25.40   1st Qu.:26.50   SaCl   :1783          1st Qu.: 536.5   1st Qu.:159.0
 Median :23.10   Median :25.60   Median :26.40   Median :27.40   ClLm   :1362          Median : 708.6   Median :159.0
 Mean   :22.80   Mean   :25.47   Mean   :26.47   Mean   :27.55   Sa     : 750          Mean   : 720.3   Mean   :168.8
 3rd Qu.:23.90   3rd Qu.:26.50   3rd Qu.:27.50   3rd Qu.:28.50   Cl     : 514          3rd Qu.: 849.7   3rd Qu.:195.0
 Max.   :25.80   Max.   :29.10   Max.   :31.30   Max.   :32.50   LmSa   : 378          Max.   :1200.0   Max.   :298.0
                                                                 (Other): 534

    Soil.FCP        Soil.TPO     Geologicaltype          Lithology    Physiographicregion Geol.textureOption.1  Geol.depth
 Min.   :112.0   Min.   :402.0   Q      :1201   SHALE         :2715   Min.   : 1.00       ClLm  : 739          1000  :1210
 1st Qu.:254.0   1st Qu.:402.0   Pp     : 939   GRANITE       :1332   1st Qu.: 4.00       Lm    : 913          1200  :7012
 Median :254.0   Median :423.0   Vt     : 827   SEDIMENTARY   :1201   Median : 8.00       NoData:   3          1500  : 739
 Mean   :265.3   Mean   :424.4   ZB     : 807   ECCA SANDSTONE: 715   Mean   :10.98       Sa    :1239          2500  :1239
 3rd Qu.:312.0   3rd Qu.:432.0   Pvo    : 717   TM SANDSTONE  : 627   3rd Qu.:16.00       SaCl  :3301          NoData:   3
 Max.   :416.0   Max.   :482.0   Pv     : 715   MUDSTONE      : 610   Max.   :41.00       SaLm  : 297
                                 (Other):4997   (Other)       :3003                       SiCl  :3711

  Geol.PWP       Geol.FCP       Geol.TPO     Altitude.1700    Slope.1700       Aspect.1700      Altitude.200     Slope.200
 150   :4214   120   :1239   450   :1239   Min.   :   8   Min.   : 0.050   Min.   :  0.00   Min.   :   7   Min.   : 0.050
 250   : 739   190   : 297   480   : 297   1st Qu.: 920   1st Qu.: 0.830   1st Qu.: 81.03   1st Qu.: 913   1st Qu.: 2.140
 270   :3711   240   :3301   520   :3301   Median :1207   Median : 1.800   Median :136.63   Median :1199   Median : 4.240
 50    :1239   300   : 913   570   : 913   Mean   :1102   Mean   : 2.308   Mean   :148.04   Mean   :1095   Mean   : 5.677
 80    : 297   360   : 739   580   : 739   3rd Qu.:1433   3rd Qu.: 3.270   3rd Qu.:193.69   3rd Qu.:1416   3rd Qu.: 7.957
 NoData:   3   400   :3711   630   :3711   Max.   :2033   Max.   :10.750   Max.   :359.74   Max.   :2030   Max.   :32.200

  Aspect.200      G_ex_abresp     G_ex_bfresp     G_ex_cropno     G_ex_depaho     G_ex_depbho       G_ex_fc1        G_ex_fc2
 Min.   :  0.00   Min.   :31.00   Min.   :31.00   Min.   :10000   Min.   :20.00   Min.   :23.00   Min.   :18.00   Min.   :20.0
 1st Qu.: 80.38   1st Qu.:40.00   1st Qu.:40.00   1st Qu.:10018   1st Qu.:20.00   1st Qu.:34.00   1st Qu.:20.00   1st Qu.:20.0
 Median :157.43   Median :40.00   Median :40.00   Median :10205   Median :20.00   Median :40.00   Median :23.00   Median :30.0
 Mean   :165.38   Mean   :44.33   Mean   :44.33   Mean   :19171   Mean   :23.84   Mean   :46.33   Mean   :22.64   Mean   :27.1
 3rd Qu.:257.00   3rd Qu.:46.00   3rd Qu.:46.00   3rd Qu.:10219   3rd Qu.:25.00   3rd Qu.:50.00   3rd Qu.:26.00   3rd Qu.:33.0
 Max.   :358.96   Max.   :64.00   Max.   :64.00   Max.   :70000   Max.   :34.00   Max.   :69.00   Max.   :28.00   Max.   :33.0
 NA's   :  1.00

    G_ex_po1        G_ex_po2        G_ex_wp1        G_ex_wp2      Mgs_duration    Altitude_200   Apan_evap_01    Apan_evap_02
 Min.   :39.00   Min.   :40.00   Min.   : 8.00   Min.   : 9.00   Min.   :167.0   Min.   :  15   Min.   :147.0   Min.   :128.0
 1st Qu.:40.00   1st Qu.:40.00   1st Qu.:10.00   1st Qu.: 9.00   1st Qu.:210.0   1st Qu.: 915   1st Qu.:170.0   1st Qu.:146.0
 Median :42.00   Median :42.00   Median :13.00   Median :18.00   Median :224.0   Median :1208   Median :177.0   Median :153.0
 Mean   :43.49   Mean   :41.94   Mean   :12.32   Mean   :15.13   Mean   :235.7   Mean   :1099   Mean   :180.1   Mean   :155.7
 3rd Qu.:47.00   3rd Qu.:43.00   3rd Qu.:15.00   3rd Qu.:20.00   3rd Qu.:241.0   3rd Qu.:1417   3rd Qu.:194.0   3rd Qu.:167.0
 Max.   :49.00   Max.   :44.00   Max.   :17.00   Max.   :20.00   Max.   :365.0   Max.   :2029   Max.   :213.0   Max.   :184.0

 Apan_evap_03    Apan_evap_04    Apan_evap_05    Apan_evap_06    Apan_evap_07    Apan_evap_08    Apan_evap_09    Apan_evap_10
 Min.   :127.0   Min.   :106.0   Min.   : 90.0   Min.   : 78.00   Min.   : 91.0   Min.   :115.0   Min.   :129.0   Min.   :133.0
 1st Qu.:145.0   1st Qu.:123.0   1st Qu.:104.0   1st Qu.: 89.00   1st Qu.:101.0   1st Qu.:127.0   1st Qu.:143.0   1st Qu.:158.0
 Median :152.0   Median :129.0   Median :111.0   Median : 93.00   Median :107.0   Median :135.0   Median :151.0   Median :168.0
 Mean   :152.9   Mean   :128.1   Mean   :111.6   Mean   : 93.74   Mean   :105.5   Mean   :136.4   Mean   :153.9   Mean   :169.0
 3rd Qu.:164.0   3rd Qu.:135.0   3rd Qu.:121.0   3rd Qu.: 98.00   3rd Qu.:110.0   3rd Qu.:147.0   3rd Qu.:166.0   3rd Qu.:181.0
 Max.   :174.0   Max.   :149.0   Max.   :139.0   Max.   :114.00   Max.   :122.0   Max.   :153.0   Max.   :178.0   Max.   :202.0

 Apan_evap_11    Apan_evap_12    Apan_mean_an   Cv_an_precip   Dly_mean_t_01   Dly_mean_t_02   Dly_mean_t_03   Dly_mean_t_04
 Min.   :135.0   Min.   :158.0   Min.   :1465   Min.   :15.00   Min.   :16.00   Min.   :15.00   Min.   :15.00   Min.   :13.00
 1st Qu.:156.0   1st Qu.:176.0   1st Qu.:1674   1st Qu.:18.00   1st Qu.:19.00   1st Qu.:19.00   1st Qu.:18.00   1st Qu.:15.00
 Median :167.0   Median :185.0   Median :1756   Median :21.00   Median :20.00   Median :20.00   Median :19.00   Median :17.00
 Mean   :166.0   Mean   :187.1   Mean   :1745   Mean   :20.05   Mean   :20.23   Mean   :20.11   Mean   :19.24   Mean   :17.04
 3rd Qu.:175.0   3rd Qu.:199.0   3rd Qu.:1821   3rd Qu.:22.00   3rd Qu.:21.00   3rd Qu.:21.00   3rd Qu.:20.00   3rd Qu.:18.00
 Max.   :194.0   Max.   :219.0   Max.   :1972   Max.   :28.00   Max.   :25.00   Max.   :25.00   Max.   :24.00   Max.   :22.00

 Dly_mean_t_05   Dly_mean_t_06   Dly_mean_t_07   Dly_mean_t_08   Dly_mean_t_09   Dly_mean_t_10   Dly_mean_t_11   Dly_mean_t_12
 Min.   :10.00   Min.   : 7.00   Min.   : 8.00   Min.   :10.00   Min.   :12.00   Min.   :13.00   Min.   :14.00   Min.   :15.00
 1st Qu.:13.00   1st Qu.:10.00   1st Qu.:10.00   1st Qu.:12.00   1st Qu.:15.00   1st Qu.:16.00   1st Qu.:17.00   1st Qu.:18.00
 Median :14.00   Median :11.00   Median :12.00   Median :13.00   Median :15.00   Median :17.00   Median :18.00   Median :19.00
 Mean   :14.51   Mean   :11.95   Mean   :11.96   Mean   :13.76   Mean   :15.93   Mean   :17.11   Mean   :18.15   Mean   :19.62
 3rd Qu.:15.00   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:15.00   3rd Qu.:17.00   3rd Qu.:18.00   3rd Qu.:19.00   3rd Qu.:21.00
 Max.   :20.00   Max.   :18.00   Max.   :18.00   Max.   :19.00   Max.   :20.00   Max.   :21.00   Max.   :22.00   Max.   :24.00
```

Appendix 8 continued.

```
   Frost_days       Frost_durtn       Frost_end      Frost_start     Frost_stdev     Heat_units_01   Heat_units_02   Heat_units_03
Min.   : 0.000   Min.   :  0.00   Min.   :  0.0   Min.   :  0.0   Min.   : 0.000   Min.   :192.0   Min.   :166.0   Min.   :173.0
1st Qu.: 2.000   1st Qu.:  4.00   1st Qu.:186.0   1st Qu.:147.0   1st Qu.: 3.000   1st Qu.:285.0   1st Qu.:254.0   1st Qu.:252.0
Median : 4.000   Median : 40.00   Median :210.0   Median :161.0   Median : 4.000   Median :316.0   Median :285.0   Median :288.0
Mean   : 8.767   Mean   : 42.55   Mean   :185.0   Mean   :143.3   Mean   : 4.688   Mean   :330.3   Mean   :295.9   Mean   :300.2
3rd Qu.:15.000   3rd Qu.: 73.00   3rd Qu.:227.0   3rd Qu.:176.0   3rd Qu.: 7.000   3rd Qu.:361.0   3rd Qu.:327.0   3rd Qu.:333.0
Max.   :43.000   Max.   :122.00   Max.   :257.0   Max.   :183.0   Max.   :10.000   Max.   :480.0   Max.   :431.0   Max.   :455.0

Heat_units_04   Heat_units_05 Heat_units_06   Heat_units_07    Heat_units_08   Heat_units_09   Heat_units_10   Heat_units_11
Min.   :100.0   Min.   : 12   Min.   :  0.0   Min.   :  0.00   Min.   :  0.0   Min.   : 84.0   Min.   :111.0   Min.   :133.0
1st Qu.:178.0   1st Qu.:105   1st Qu.: 22.0   1st Qu.: 26.00   1st Qu.: 86.0   1st Qu.:151.0   1st Qu.:192.0   1st Qu.:216.0
Median :211.0   Median :137   Median : 57.0   Median : 62.00   Median :117.0   Median :178.0   Median :221.0   Median :246.0
Mean   :224.3   Mean   :153   Mean   : 75.5   Mean   : 79.22   Mean   :131.5   Mean   :191.6   Mean   :234.4   Mean   :258.4
3rd Qu.:258.0   3rd Qu.:184   3rd Qu.: 99.0   3rd Qu.:105.00   3rd Qu.:161.0   3rd Qu.:219.0   3rd Qu.:268.0   3rd Qu.:291.0
Max.   :379.0   Max.   :325   Max.   :247.0   Max.   :251.00   Max.   :285.0   Max.   :321.0   Max.   :359.0   Max.   :387.0

Heat_units_12    Mean_an_prec     Mean_humid_01   Mean_humid_02   Mean_humid_03   Mean_humid_04   Mean_humid_05   Mean_humid_06
Min.   :178.0   Min.   :  13.00   Min.   :65.00   Min.   :66.00   Min.   :64.0   Min.   :61.00   Min.   :56.00   Min.   :53.00
1st Qu.:269.0   1st Qu.:  15.00   1st Qu.:70.00   1st Qu.:70.00   1st Qu.:69.0   1st Qu.:66.00   1st Qu.:62.00   1st Qu.:59.00
Median :300.0   Median :  16.00   Median :72.00   Median :72.00   Median :71.0   Median :68.00   Median :64.00   Median :62.00
Mean   :312.8   Mean   :  61.03   Mean   :71.69   Mean   :71.69   Mean   :70.8   Mean   :68.25   Mean   :64.62   Mean   :62.39
3rd Qu.:345.0   3rd Qu.:  18.00   3rd Qu.:73.00   3rd Qu.:73.00   3rd Qu.:72.0   3rd Qu.:70.00   3rd Qu.:67.00   3rd Qu.:65.00
Max.   :455.0   Max.   :1553.00   Max.   :78.00   Max.   :78.00   Max.   :77.0   Max.   :75.00   Max.   :73.00   Max.   :71.00

Mean_humid_07   Mean_humid_08   Mean_humid_09   Mean_humid_10   Mean_humid_11   Mean_humid_12    Min_humid_01    Min_humid_02
Min.   :52.00   Min.   :54.00   Min.   :58.00   Min.   :61.00   Min.   :63.00   Min.   :64.00   Min.   :43.00   Min.   :44.00
1st Qu.:59.00   1st Qu.:60.00   1st Qu.:63.00   1st Qu.:66.00   1st Qu.:69.00   1st Qu.:69.00   1st Qu.:51.00   1st Qu.:50.00
Median :61.00   Median :62.00   Median :64.00   Median :67.00   Median :70.00   Median :71.00   Median :52.00   Median :52.00
Mean   :61.91   Mean   :62.83   Mean   :65.19   Mean   :67.88   Mean   :70.34   Mean   :70.54   Mean   :52.29   Mean   :52.33
3rd Qu.:65.00   3rd Qu.:65.00   3rd Qu.:67.00   3rd Qu.:69.00   3rd Qu.:72.00   3rd Qu.:72.00   3rd Qu.:54.00   3rd Qu.:54.00
Max.   :71.00   Max.   :72.00   Max.   :75.00   Max.   :76.00   Max.   :77.00   Max.   :77.00   Max.   :61.00   Max.   :61.00

 Min_humid_03   Min_humid_04   Min_humid_05   Min_humid_06   Min_humid_07   Min_humid_08   Min_humid_09   Min_humid_10
Min.   :43.0   Min.   :38.00   Min.   :32.00   Min.   :30.00   Min.   :29.00   Min.   :31.0   Min.   :35.00   Min.   :39.00
1st Qu.:49.0   1st Qu.:45.00   1st Qu.:39.00   1st Qu.:36.00   1st Qu.:36.00   1st Qu.:37.0   1st Qu.:41.00   1st Qu.:45.00
Median :51.0   Median :47.00   Median :42.00   Median :39.00   Median :39.00   Median :40.0   Median :42.00   Median :46.00
Mean   :51.1   Mean   :47.61   Mean   :42.93   Mean   :40.19   Mean   :39.60   Mean   :40.7   Mean   :43.68   Mean   :47.12
3rd Qu.:53.0   3rd Qu.:50.00   3rd Qu.:46.00   3rd Qu.:44.00   3rd Qu.:43.00   3rd Qu.:44.0   3rd Qu.:46.00   3rd Qu.:49.00
Max.   :60.0   Max.   :58.00   Max.   :55.00   Max.   :52.00   Max.   :51.00   Max.   :53.0   Max.   :57.00   Max.   :58.00

 Min_humid_11    Min_humid_12   Ms_dy_maxt_01   Ms_dy_maxt_02   Ms_dy_maxt_03   Ms_dy_maxt_04   Ms_dy_maxt_05   Ms_dy_maxt_06
Min.   :41.00   Min.   :41.00   Min.   :21.00   Min.   :20.00   Min.   :20.00   Min.   :18.00   Min.   :16.00   Min.   :14.00
1st Qu.:48.00   1st Qu.:49.00   1st Qu.:24.00   1st Qu.:24.00   1st Qu.:23.00   1st Qu.:21.00   1st Qu.:19.00   1st Qu.:17.00
Median :50.00   Median :51.00   Median :25.00   Median :25.00   Median :24.00   Median :23.00   Median :21.00   Median :19.00
Mean   :50.44   Mean   :50.74   Mean   :25.42   Mean   :25.34   Mean   :24.62   Mean   :22.85   Mean   :21.02   Mean   :18.83
3rd Qu.:52.00   3rd Qu.:53.00   3rd Qu.:27.00   3rd Qu.:27.00   3rd Qu.:26.00   3rd Qu.:24.00   3rd Qu.:22.00   3rd Qu.:20.00
Max.   :61.00   Max.   :60.00   Max.   :30.00   Max.   :30.00   Max.   :29.00   Max.   :27.00   Max.   :26.00   Max.   :24.00

Ms_dy_maxt_07   Ms_dy_maxt_08   Ms_dy_maxt_09   Ms_dy_maxt_10   Ms_dy_maxt_11   Ms_dy_maxt_12   Ms_dy_mint_01   Ms_dy_mint_02
Min.   :14.00   Min.   :16.00   Min.   :19.0   Min.   :19.00   Min.   :19.00   Min.   :20.00   Min.   :11.00   Min.   :11.00
1st Qu.:17.00   1st Qu.:19.00   1st Qu.:21.0   1st Qu.:22.00   1st Qu.:22.00   1st Qu.:24.00   1st Qu.:13.00   1st Qu.:13.00
Median :19.00   Median :20.00   Median :22.0   Median :23.00   Median :23.00   Median :25.00   Median :14.00   Median :14.00
Mean   :19.02   Mean   :20.71   Mean   :22.4   Mean   :23.03   Mean   :23.58   Mean   :25.07   Mean   :15.04   Mean   :14.95
3rd Qu.:20.00   3rd Qu.:22.00   3rd Qu.:24.0   3rd Qu.:25.00   3rd Qu.:25.00   3rd Qu.:26.00   3rd Qu.:16.00   3rd Qu.:16.00
Max.   :24.00   Max.   :25.00   Max.   :26.0   Max.   :27.00   Max.   :27.00   Max.   :29.00   Max.   :20.00   Max.   :20.00

Ms_dy_mint_03   Ms_dy_mint_04   Ms_dy_mint_05    Ms_dy_mint_06    Ms_dy_mint_07    Ms_dy_mint_08    Ms_dy_mint_09
Min.   :10.00   Min.   : 7.00   Min.   : 4.000   Min.   : 0.000   Min.   : 1.000   Min.   : 3.000   Min.   : 5.000
1st Qu.:12.00   1st Qu.: 9.00   1st Qu.: 6.000   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 5.000   1st Qu.: 8.000
Median :13.00   Median :11.00   Median : 8.000   Median : 5.000   Median : 5.000   Median : 6.000   Median : 9.000
Mean   :13.87   Mean   :11.20   Mean   : 7.982   Mean   : 5.106   Mean   : 5.048   Mean   : 6.922   Mean   : 9.486
3rd Qu.:15.00   3rd Qu.:12.00   3rd Qu.: 9.000   3rd Qu.: 6.000   3rd Qu.: 6.000   3rd Qu.: 8.000   3rd Qu.:10.000
Max.   :20.00   Max.   :17.00   Max.   :15.000   Max.   :13.000   Max.   :12.000   Max.   :13.000   Max.   :15.000
```

Appendix 8 continued.

```
Ms_dy_mint_10   Ms_dy_mint_11   Ms_dy_mint_12    Pemo_evap_01    Pemo_evap_02    Pemo_evap_03    Pemo_evap_04     Pemo_evap_05
Min.   : 7.00   Min.   : 9.00   Min.   :10.00   Min.   :108.0   Min.   : 95.0   Min.   : 96.0   Min.   : 75.00   Min.   :59
1st Qu.:10.00   1st Qu.:11.00   1st Qu.:13.00   1st Qu.:128.0   1st Qu.:110.0   1st Qu.:106.0   1st Qu.: 85.00   1st Qu.:67
Median :11.00   Median :12.00   Median :14.00   Median :137.0   Median :118.0   Median :111.0   Median : 88.00   Median :70
Mean   :11.23   Mean   :12.72   Mean   :14.23   Mean   :137.2   Mean   :118.2   Mean   :112.7   Mean   : 89.52   Mean   :72
3rd Qu.:12.00   3rd Qu.:14.00   3rd Qu.:15.00   3rd Qu.:147.0   3rd Qu.:126.0   3rd Qu.:120.0   3rd Qu.: 94.00   3rd Qu.:77
Max.   :16.00   Max.   :18.00   Max.   :19.00   Max.   :161.0   Max.   :146.0   Max.   :134.0   Max.   :108.00   Max.   :94

 Pemo_evap_06    Pemo_evap_07    Pemo_evap_08    Pemo_evap_09    Pemo_evap_10    Pemo_evap_11    Pemo_evap_12   Rain_concentn
Min.   :50.00   Min.   :56.00   Min.   : 74.00   Min.   : 87.0   Min.   : 98.0   Min.   :103.0   Min.   :117.0   Min.   :22.00
1st Qu.:55.00   1st Qu.:61.00   1st Qu.: 82.00   1st Qu.:102.0   1st Qu.:117.0   1st Qu.:119.0   1st Qu.:134.0   1st Qu.:49.00
Median :56.00   Median :64.00   Median : 91.00   Median :108.0   Median :125.0   Median :127.0   Median :141.0   Median :52.00
Mean   :57.67   Mean   :65.32   Mean   : 88.44   Mean   :107.6   Mean   :125.5   Mean   :126.8   Mean   :142.9   Mean   :49.99
3rd Qu.:59.00   3rd Qu.:68.00   3rd Qu.: 92.50   3rd Qu.:113.0   3rd Qu.:134.0   3rd Qu.:135.0   3rd Qu.:150.0   3rd Qu.:57.00
Max.   :74.00   Max.   :80.00   Max.   :105.00   Max.   :125.0   Max.   :150.0   Max.   :147.0   Max.   :167.0   Max.   :60.00

 Rain_seasons   Solar_radn_01   Solar_radn_02   Solar_radn_03   Solar_radn_04   Solar_radn_05   Solar_radn_06   Solar_radn_07
Min.   :3.000   Min.   :23.90   Min.   :22.50   Min.   :20.40   Min.   :17.80   Min.   :14.50   Min.   :12.80   Min.   :13.40
1st Qu.:3.000   1st Qu.:25.80   1st Qu.:24.40   1st Qu.:22.00   1st Qu.:18.90   1st Qu.:15.90   1st Qu.:14.30   1st Qu.:15.00
Median :4.000   Median :26.50   Median :25.10   Median :22.60   Median :19.70   Median :16.50   Median :14.80   Median :15.70
Mean   :3.663   Mean   :26.67   Mean   :25.22   Mean   :22.73   Mean   :19.69   Mean   :16.77   Mean   :15.12   Mean   :15.96
3rd Qu.:4.000   3rd Qu.:27.50   3rd Qu.:26.10   3rd Qu.:23.40   3rd Qu.:20.40   3rd Qu.:17.70   3rd Qu.:16.10   3rd Qu.:17.00
Max.   :6.000   Max.   :31.20   Max.   :28.80   Max.   :25.70   Max.   :22.00   Max.   :18.90   Max.   :17.20   Max.   :18.10

Solar_radn_08   Solar_radn_09   Solar_radn_10   Solar_radn_11   Solar_radn_12   Temp_range_01    Temp_range_02    Temp_range_03
Min.   :16.70   Min.   :19.30   Min.   :21.70   Min.   :23.00   Min.   :24.30   Min.   : 8.000   Min.   : 8.000   Min.   : 8.00
1st Qu.:18.10   1st Qu.:21.70   1st Qu.:24.60   1st Qu.:25.40   1st Qu.:26.50   1st Qu.: 9.000   1st Qu.: 9.000   1st Qu.:10.00
Median :19.10   Median :23.10   Median :25.60   Median :26.40   Median :27.40   Median :10.000   Median :10.000   Median :10.00
Mean   :19.15   Mean   :22.80   Mean   :25.47   Mean   :26.47   Mean   :27.55   Mean   : 9.943   Mean   : 9.957   Mean   :10.23
3rd Qu.:20.30   3rd Qu.:23.90   3rd Qu.:26.50   3rd Qu.:27.50   3rd Qu.:28.50   3rd Qu.:10.000   3rd Qu.:11.000   3rd Qu.:11.00
Max.   :21.50   Max.   :25.80   Max.   :29.10   Max.   :31.30   Max.   :32.50   Max.   :13.000   Max.   :13.000   Max.   :13.00

Temp_range_04   Temp_range_05   Temp_range_06   Temp_range_07   Temp_range_08   Temp_range_09   Temp_range_10   Temp_range_11
Min.   : 8.00   Min.   : 9.00   Min.   :10.00   Min.   :10.00   Min.   :10.00   Min.   : 9.00   Min.   : 8.00   Min.   : 8.00
1st Qu.:10.00   1st Qu.:11.00   1st Qu.:12.00   1st Qu.:12.00   1st Qu.:12.00   1st Qu.:12.00   1st Qu.:11.00   1st Qu.:10.00
Median :11.00   Median :13.00   Median :13.00   Median :14.00   Median :14.00   Median :13.00   Median :12.00   Median :10.00
Mean   :11.15   Mean   :12.53   Mean   :13.30   Mean   :13.52   Mean   :13.32   Mean   :12.43   Mean   :11.34   Mean   :10.37
3rd Qu.:12.00   3rd Qu.:14.00   3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:15.00   3rd Qu.:13.00   3rd Qu.:12.00   3rd Qu.:11.00
Max.   :15.00   Max.   :17.00   Max.   :18.00   Max.   :18.00   Max.   :17.00   Max.   :16.00   Max.   :14.00   Max.   :14.00

Temp_range_12       WBDEC            WBNOV            WBOCT            WBSEP            WBAUG            WBJUL
Min.   : 8.00   Min.   :-141.40   Min.   :-144.60   Min.   :-187.8   Min.   :-179.05   Min.   :-155.00   Min.   :-120.10
1st Qu.:10.00   1st Qu.: -88.10   1st Qu.:-103.65   1st Qu.:-156.5   1st Qu.:-160.20   1st Qu.:-145.10   1st Qu.:-108.10
Median :10.00   Median : -73.05   Median : -87.40   Median :-128.4   Median :-133.75   Median :-129.45   Median :-100.10
Mean   :10.41   Mean   : -76.33   Mean   : -89.26   Mean   :-133.8   Mean   :-139.49   Mean   :-128.85   Mean   : -97.89
3rd Qu.:11.00   3rd Qu.: -63.30   3rd Qu.: -76.90   3rd Qu.:-112.5   3rd Qu.:-123.40   3rd Qu.:-116.55   3rd Qu.: -91.25
Max.   :14.00   Max.   :  -4.60   Max.   : -22.35   Max.   : -81.3   Max.   : -78.15   Max.   : -76.55   Max.   : -46.10

    WBJUN            WBMAY            WBAPR            WBMAR            WBFEB            WBJAN
Min.   :-103.55   Min.   :-100.90   Min.   :-82.25   Min.   :-99.50   Min.   :-103.50   Min.   :-91.63
1st Qu.: -84.60   1st Qu.: -77.85   1st Qu.:-46.70   1st Qu.:-58.35   1st Qu.: -53.85   1st Qu.: -4.64
Median : -78.85   Median : -65.30   Median :-34.15   Median :-45.25   Median : -31.50   Median : 16.12
Mean   : -71.74   Mean   : -55.43   Mean   :-19.97   Mean   :-28.05   Mean   : -18.28   Mean   : 20.11
3rd Qu.: -69.40   3rd Qu.: -51.30   3rd Qu.:-15.40   3rd Qu.:-20.11   3rd Qu.:   7.07   3rd Qu.: 41.52
Max.   :  20.95   Max.   : 104.10   Max.   :184.15   Max.   :199.95   Max.   : 190.50   Max.   :168.99
```

APPENDIX 9 Regression tree models

1.1. *Eucalyptus* regression tree

FULL MODEL : * donates a terminal node.

1) root 5461 122681.2000 23.73320

2)SteClsCli=CT2,CT3,CT4,CT5,CT6,CT7,CT8,CT9,ST1,ST2,ST4,ST7,WT1,WT2,WT3,WT4,WT5,WT7,WT8 3900 53823.7500 21.95042

4) Apan_evap_08>=146.5 1245 11761.1000 19.60063

8) Spp=ECLO,EFAS,EG+M,ERUB 80 421.1911 16.03375 *

9) Spp=EDUN,EELA,EEMA,EFRA,EGRA,EGXC,EGXN,EGXU,EMAC,EMIX,ENIT,ESMI 1165 10252.2100 19.84556

18) Geologicaltype=C-Pd,Jd,Pp,Pv,Q,RB,Rmp,Vm,Vsi,Vt,ZB,ZC,Z-R 995 7750.9900 19.49081 *

19) Geologicaltype=Ru,Vbr,Vh,Vhd,Zka,Zne,Zns,Zo 170 1643.1110 21.92188 *

5) Apan_evap_08< 146.5 2655 31964.8400 23.05229

10) SteClsCli=CT3,CT4,CT5,CT7,CT8,CT9,ST4,WT1,WT2,WT4 1227 11954.7200 21.93282

20) Apan_evap_01>=176.5 202 1929.0970 19.72460 *

21) Apan_evap_01< 176.5 1025 8846.5110 22.36800 *

11) SteClsCli=CT6,ST1,ST2,ST7,WT3,WT5,WT7,WT8 1428 17151.1600 24.01419

22) rskGT850< 51.9 115 850.4866 21.22357 *

23) rskGT850>=51.9 1313 15326.6600 24.25861

46) Geologicaltype=C-Pd,Jd,O-S,Pa,Pvo,Qm,TRt,Vm,Zn,Zne,Zns,Zo 512 5066.4670 23.18785 *

47) Geologicaltype=Kz,Nmp,Pe,Pp,Pv,Q,Vbr,ZB,Zka 801 9297.9390 24.94305 *

3) SteClsCli=ST3,ST6,ST8,ST9,WT6,WT9 1561 25493.1500 28.18731

6) WBJUN< -42.275 728 10925.0900 26.39460

12) WBMAY< -61.65 183 1404.9500 24.08268 *

13) WBMAY>=-61.65 545 8213.5670 27.17090

26) MeaMthPreFeb< 139.5 113 1526.0950 24.75168 *

27) MeaMthPreFeb>=139.5 432 5853.1360 27.80370 *

7) WBJUN>=-42.275 833 10183.6800 29.75405

14) AveMthSrdNov< 23.85 67 616.1522 25.36284 *

15) AveMthSrdNov>=23.85 766 8162.5840 30.13813 *

Appendix 9 continued.

1.2. *Acacia* regression tree

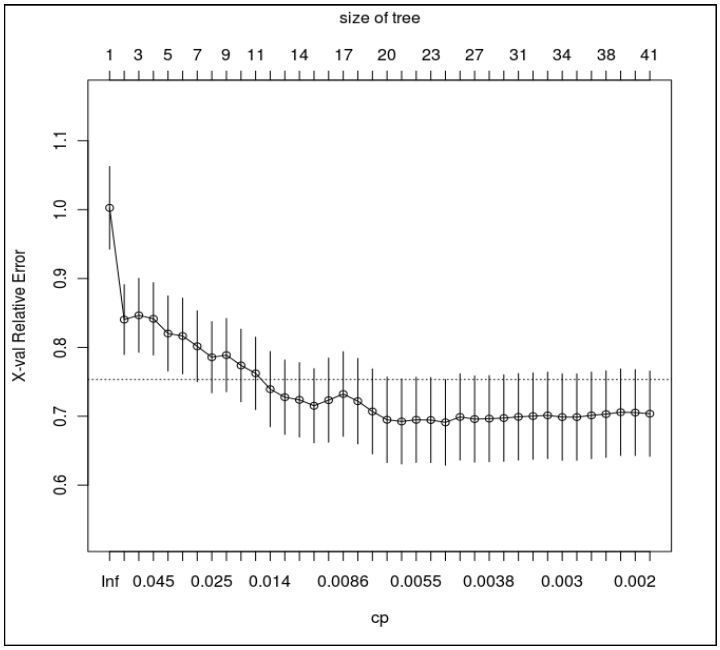


Figure 48: Cross-validated relative error and CP by tree size for the *Acacia* regression tree

Minimum relative error is for a tree of 24 splits. Pruned tree to 7 splits:

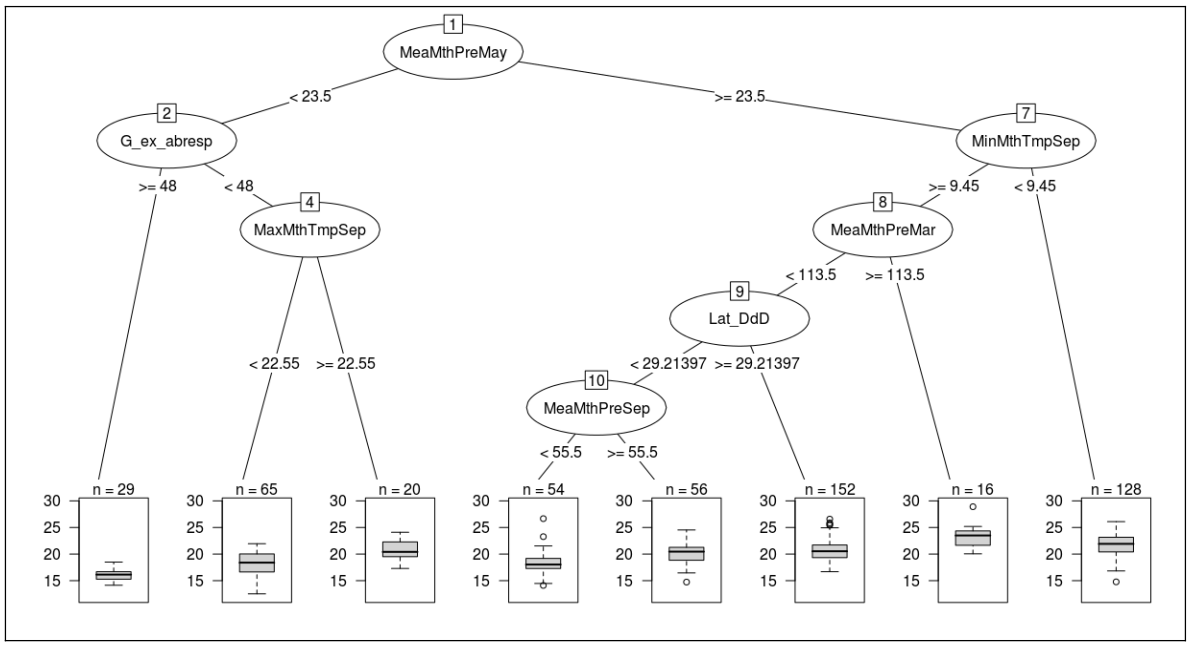


Figure 49: Pruned *Acacia* regression Tree (CP = 0.026)

Appendix 9 continued.

Where :

MeaMthPreMay = Mean monthly precipitation for May

G_ex_abresp = Grid top soil to sub soil daily drainage fraction

Max/Min MthTmpSep = Maximum / Minimum monthly temperature for September

MeaMthPre Sep/Mar = Mean monthly precipitation September / March

Lat_DdD = Latitude

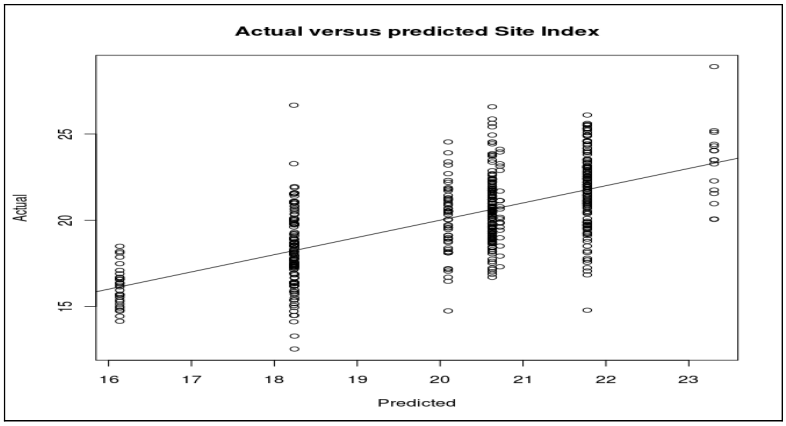


Figure 50: Actual versus predicted Site Index for the *Acacia* regression tree.

Root squared mean error (RMSE) = 2.008928, R^2 = 40.705

Full model :

node), split, n, deviance, yval * denotes terminal node

1) root 520 3539.27700 20.14227
2) MeaMthPreMay< 23.5 114 666.56360 18.14167
4) G_ex_abresp>=48 29 37.83612 16.13552 *
5) G_ex_abresp< 48 85 472.19280 18.82612
10) MaxMthTmpSep< 22.55 65 305.22090 18.24292 *
11) MaxMthTmpSep>=22.55 20 73.01486 20.72150 *
3) MeaMthPreMay>=23.5 406 2288.32200 20.70401
6) MinMthTmpSep>=9.45 278 1448.61800 20.21076
12) MeaMthPreMar< 113.5 262 1211.38000 20.02176
24) Lat_DdD< 29.21397 110 555.52710 19.18309
48) MeaMthPreSep< 55.5 54 250.58810 18.23537 *
49) MeaMthPreSep>=55.5 56 209.66840 20.09696 *
25) Lat_DdD>=29.21397 152 522.49230 20.62868 *
13) MeaMthPreMar>=113.5 16 74.62759 23.30562 *
7) MinMthTmpSep< 9.45 128 625.16240 21.77531 *

Appendix 9 continued.

1.3. *Pinus* regression tree

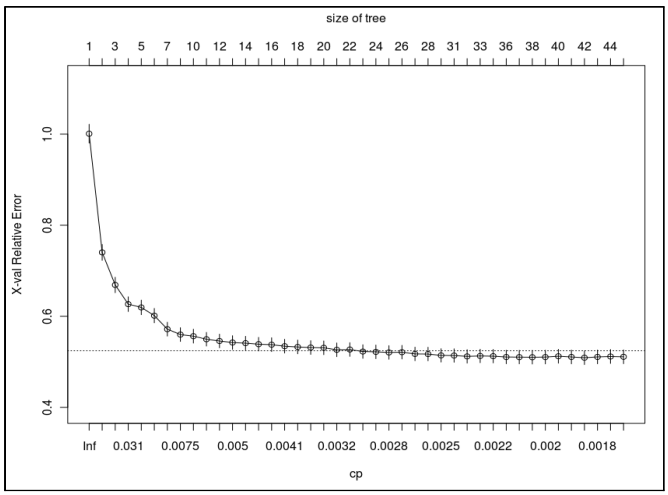


Figure 51: Cross-validated relative error and CP by tree size for the *Pinus* regression tree.

Minimum relative error is for a tree of 39 splits. Pruned tree to 7 splits:

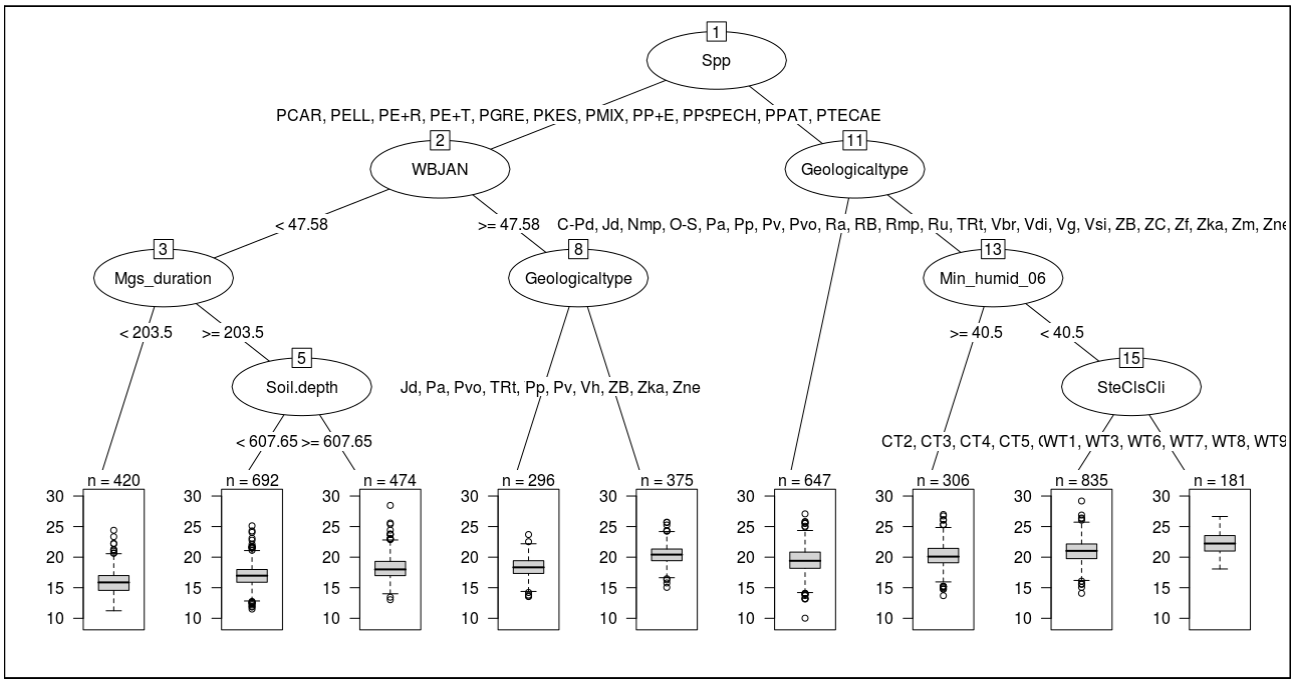


Figure 52: Pruned *Pinus* regression tree (CP = 0.0075)

Where

Spp = Species

Appendix 9 continued.

WBJAN = Water balance for January

Geologicaltype = Geological type

Mgs_duration = Mean growth season duration

Min_humid_06 = minimum humidity for June

SteClsCli = Site Classification by Climate

Soil.depth = Soil depth

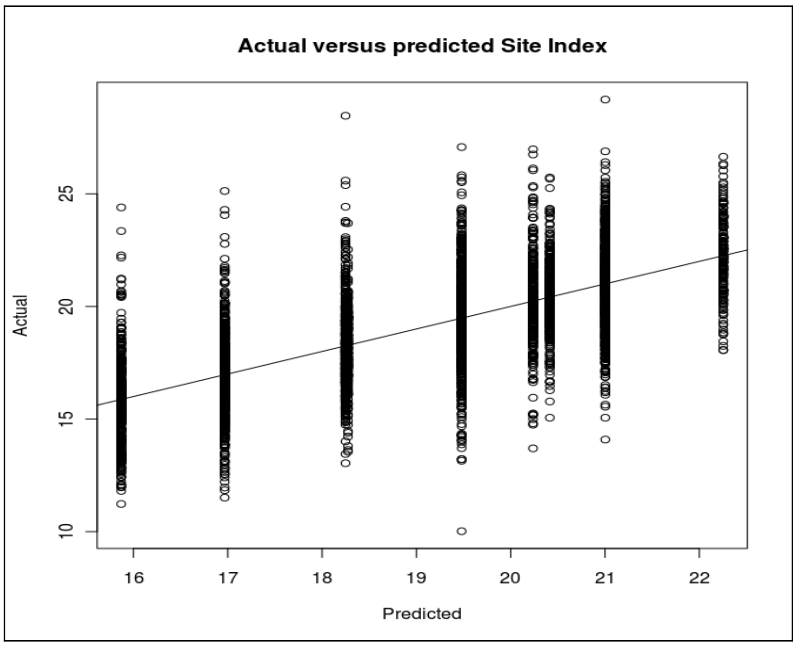


Figure 53: Actual versus predicted Site Index for the *Pinus* regression tree.

RMSE = 1.89585, R^2 = 42.00211

Full model :

1) root 4226 29211.3700 19.04483

2) Spp=PCAR,PELL,PE+R,PE+T,PGRE,PKES,PMIX,PP+E,PPSE,PP+T,PROX,PTAE 2257 11949.2900 17.77667

4) WBJAN< 47.58 1586 6829.5130 17.05988

8) Mgs_duration< 203.5 420 1609.9940 15.87031 *

9) Mgs_duration>=203.5 1166 4411.1040 17.48837

18) Soil.depth< 607.65 692 2216.3460 16.96757 *

19) Soil.depth>=607.65 474 1733.0520 18.24869 *

5) WBJAN>=47.58 671 2378.8750 19.47089

10) Geologicaltype=Jd,Pa,Pvo,TRt,Vm,Vt,Zf,Zm,Zo,Z-R 296 691.5144 18.27483 *

11) Geologicaltype=Pp,Pv,Vh,ZB,Zka,Zne 375 929.6712 20.41499 *

3) Spp=PECH,PPAT,PTEC 1969 9471.5670 20.49849

6) Geologicaltype=Vh,Vhd,Vm,Vt,Vw,Water,Zns,Zo,Z-R 647 3166.1950 19.47980 *

7) Geologicaltype=C-Pd,Jd,Nmp,O-S,Pa,Pp,Pv,Pvo,Ra,RB,Rmp,Ru,TRt,Vbr,Vdi,Vg,Vsi,ZB,ZC,Zf,Zka,Zm,Zne 1322 5305.3710 20.99704

14) Min_humid_06>=40.5 306 1315.1590 20.23712 *

15) Min_humid_06< 40.5 1016 3760.2840 21.22592

30) SteClsCli=CT2,CT3,CT4,CT5,CT6,CT7,CT8,CT9,WT2,WT4,WT5 835 2973.2400 21.00299 *

31) SteClsCli=WT1,WT3,WT6,WT7,WT8,WT9 181 554.1246 22.25431 *

APPENDIX 10 Alternative *Eucalyptus* multiple regression model using the explanatory variables identified in the regression tree

Call:

lm(formula = SI ~ Spp + MeaMthPreFeb + SteClsCli + AveMthSrdNov +
Geologicaltype + Apan_evap_01 + WBJUN, data = Euc10)

Residuals:

Min	1Q	Median	3Q	Max
-13.4127	-2.0761	-0.1691	1.9112	11.6781

Coefficients:

Variable	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	26.849274	3.717069	7.223	5.78e-13	***
Spp[T.ECLO]	-5.697923	3.440112	-1.656	0.097716	.
Spp[T.EDUN]	1.661898	3.077623	0.540	0.589224	
Spp[T.EELA]	-0.061184	3.192958	-0.019	0.984712	
Spp[T.EEMA]	3.904362	4.345558	0.898	0.368974	
Spp[T.EFAS]	-1.723114	3.101602	-0.556	0.578537	
Spp[T.EFRA]	2.336738	3.257089	0.717	0.473139	
Spp[T.EG+M]	-1.538632	3.558831	-0.432	0.665510	
Spp[T.EGRA]	2.054852	3.074246	0.668	0.503902	
Spp[T.EGXC]	0.093015	3.082704	0.030	0.975930	
Spp[T.EGXN]	5.780823	3.116577	1.855	0.063671	.
Spp[T.EGXT]	2.517743	3.768331	0.668	0.504078	
Spp[T.EGXU]	1.839125	3.080237	0.597	0.550484	
Spp[T.EMAC]	0.295496	3.079784	0.096	0.923566	
Spp[T.EMIX]	0.051268	3.098127	0.017	0.986798	
Spp[T.ENIT]	1.156339	3.082461	0.375	0.707575	
Spp[T.EREG]	4.039438	4.345397	0.930	0.352625	
Spp[T.ERUB]	-3.674069	4.440965	-0.827	0.408096	
Spp[T.ESAL]	1.263792	3.174542	0.398	0.690571	
Spp[T.ESMI]	1.731747	3.082148	0.562	0.574232	
Spp[T.EURO]	2.834405	3.213509	0.882	0.377801	
MeaMthPreFeb	0.035560	0.003897	9.125	< 2e-16	***
SteClsCli[T.CT3]	0.870543	1.411242	0.617	0.537351	
SteClsCli[T.CT4]	-0.848082	1.297265	-0.654	0.513303	
SteClsCli[T.CT5]	-1.244947	1.281114	-0.972	0.331209	
SteClsCli[T.CT6]	0.706540	1.318223	0.536	0.591995	
SteClsCli[T.CT7]	-0.600276	1.321213	-0.454	0.649605	
SteClsCli[T.CT8]	-0.431379	1.290498	-0.334	0.738186	
SteClsCli[T.CT9]	-0.065894	1.305243	-0.050	0.959738	
SteClsCli[T.ST1]	2.437050	1.411212	1.727	0.084240	.
SteClsCli[T.ST2]	2.524536	1.473964	1.713	0.086816	.
SteClsCli[T.ST3]	4.662008	1.389656	3.355	0.000800	***
SteClsCli[T.ST4]	1.675204	2.526231	0.663	0.507280	
SteClsCli[T.ST6]	4.919229	1.494238	3.292	0.001001	**
SteClsCli[T.ST7]	1.640384	1.408778	1.164	0.244313	

Appendix 10 Continued.

Variable	Estimate	Std. Error	t value	Pr(>|t|)	
SteClsCli[T.ST8]	1.654936	1.399511	1.183	0.237056	
SteClsCli[T.ST9]	3.567424	1.402997	2.543	0.011027	*
SteClsCli[T.WT1]	-0.082907	1.292872	-0.064	0.948872	
SteClsCli[T.WT2]	-0.009913	1.297707	-0.008	0.993905	
SteClsCli[T.WT3]	0.364927	1.298987	0.281	0.778774	
SteClsCli[T.WT4]	0.857862	1.308873	0.655	0.512225	
SteClsCli[T.WT5]	1.732455	1.312176	1.320	0.186794	
SteClsCli[T.WT6]	2.286185	1.327255	1.722	0.085038	.
SteClsCli[T.WT7]	1.763377	1.322533	1.333	0.182479	
SteClsCli[T.WT8]	3.339266	1.323067	2.524	0.011635	*
SteClsCli[T.WT9]	3.533651	1.331462	2.654	0.007979	**
AveMthSrdNov	0.227708	0.092899	2.451	0.014272	*
Geologicaltype[T.Jd]	0.508688	0.415987	1.223	0.221441	
Geologicaltype[T.Kz]	4.514187	0.921179	4.900	9.84e-07	***
Geologicaltype[T.Nmp]	0.179810	0.395347	0.455	0.649260	
Geologicaltype[T.O-S]	0.284042	0.303425	0.936	0.349253	
Geologicaltype[T.Pa]	-0.345460	0.468208	-0.738	0.460648	
Geologicaltype[T.Pe]	-0.452648	0.516122	-0.877	0.380516	
Geologicaltype[T.Pp]	1.013916	0.297569	3.407	0.000661	***
Geologicaltype[T.Pv]	0.818928	0.316673	2.586	0.009735	**
Geologicaltype[T.Pvo]	-0.131548	0.351164	-0.375	0.707968	
Geologicaltype[T.Q]	5.125765	0.540695	9.480	< 2e-16	***
Geologicaltype[T.Qb]	0.854647	0.821097	1.041	0.297988	
Geologicaltype[T.Qm]	3.661870	0.954465	3.837	0.000126	***
Geologicaltype[T.RB]	1.879957	0.771216	2.438	0.014815	*
Geologicaltype[T.Rmp]	1.290813	0.366418	3.523	0.000431	***
Geologicaltype[T.Ru]	3.902222	0.822648	4.743	2.16e-06	***
Geologicaltype[T.TRt]	-1.384723	0.502306	-2.757	0.005858	**
Geologicaltype[T.Vbr]	1.768307	0.826452	2.140	0.032429	*
Geologicaltype[T.Vdi]	0.952964	0.578339	1.648	0.099460	.
Geologicaltype[T.Vh]	2.245967	0.650548	3.452	0.000560	***
Geologicaltype[T.Vhd]	3.416696	0.996721	3.428	0.000613	***
Geologicaltype[T.Vm]	-1.020562	0.476198	-2.143	0.032146	*
Geologicaltype[T.Vsi]	1.444058	0.650514	2.220	0.026469	*
Geologicaltype[T.Vt]	-0.474286	0.363816	-1.304	0.192411	
Geologicaltype[T.ZB]	1.320333	0.325824	4.052	5.14e-05	***
Geologicaltype[T.ZC]	0.015330	0.542714	0.028	0.977466	
Geologicaltype[T.Zka]	1.122810	0.351446	3.195	0.001407	**
Geologicaltype[T.Zn]	0.803299	0.673609	1.193	0.233106	
Geologicaltype[T.Zne]	-0.440602	0.696138	-0.633	0.526811	
Geologicaltype[T.Zns]	-1.717184	0.876446	-1.959	0.050134	.
Geologicaltype[T.Zo]	-0.573434	0.542421	-1.057	0.290480	
Geologicaltype[T.Z-R]	0.277815	0.650577	0.427	0.669375	
Apan_evap_01	-0.091338	0.009033	-10.111	< 2e-16	***
WBJUN	0.025247	0.004077	6.192	6.37e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.059 on 5377 degrees of freedom

Multiple R-squared: 0.5893, Adjusted R-squared: 0.5833

F-statistic: 97.67 on 79 and 5377 DF, p-value: < 2.2e-16

APPENDIX 11 M5 pruned *Eucalyptus* Model tree

(using smoothed linear models)

Tree:

SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3 <= 0.5 :

| Apan_evap_08 <= 146.5 :

| | SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3 <= 0.5 :

| | | WBSEP <= -131.825 : LM1 (244/66.458%)

| | | WBSEP > -131.825 : LM2 (973/61.743%)

| | SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3 > 0.5 :

| | | Mgs_duration <= 238.5 : LM3 (721/69.947%)

| | | Mgs_duration > 238.5 : LM4 (715/72.207%)

| Apan_evap_08 > 146.5 : LM5 (1245/62.257%)

SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3 > 0.5 : LM6 (1558/77.198%)

LM num: 1

SI = 0.0202 * Spp=ENIT,ECAM,EFRA,EDUN,ESMI,ESAL,EEMA,EMIX,EGXN,EGRA,EGXC,EREG,EGXT,EGXU,EURO

+ 0.0398 * SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3

+ 0.0087 * SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3

+ 0.0005 * Mgs_duration

- 0.0004 * Apan_evap_08

+ 0.0038 * WBSEP

+ 0.0002 * WBJUN

+ 20.5488

LM num: 2

SI = 0.0202 * Spp=ENIT,ECAM,EFRA,EDUN,ESMI,ESAL,EEMA,EMIX,EGXN,EGRA,EGXC,EREG,EGXT,EGXU,EURO

+ 0.0398 * SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3

+ 0.0087 * SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3

+ 0.0005 * Mgs_duration

- 0.0004 * Apan_evap_08

+ 0.0009 * WBSEP

+ 0.0002 * WBJUN

+ 22.449

Appendix 11 continued.

LM num: 3

SI = 0.0202 * Spp=ENIT,ECAM,EFRA,EDUN,ESMI,ESAL,EEMA,EMIX,EGXN,EGRA,EGXC,EREG,EGXT,EGXU,EURO
+ 0.0366 * SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0087 * SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0009 * Mgs_duration
- 0.0004 * Apan_evap_08
- 0.0001 * WBSEP
+ 0.0002 * WBJUN
+ 23.04

LM num: 4

SI = 0.0202 * Spp=ENIT,ECAM,EFRA,EDUN,ESMI,ESAL,EEMA,EMIX,EGXN,EGRA,EGXC,EREG,EGXT,EGXU,EURO
+ 0.0366 * SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0087 * SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0009 * Mgs_duration
- 0.0004 * Apan_evap_08
- 0.0001 * WBSEP
+ 0.0002 * WBJUN
+ 24.5622

LM num: 5

SI = 1.8389 * Spp=ENIT,ECAM,EFRA,EDUN,ESMI,ESAL,EEMA,EMIX,EGXN,EGRA,EGXC,EREG,EGXT,EGXU,EURO
+ 0.0313 * SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0087 * SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0004 * Mgs_duration
- 0.0007 * Apan_evap_08
- 0.0001 * WBSEP
+ 0.0002 * WBJUN
+ 18.4046

LM num: 6

SI = 0.0217 * Spp=ENIT,ECAM,EFRA,EDUN,ESMI,ESAL,EEMA,EMIX,EGXN,EGRA,EGXC,EREG,EGXT,EGXU,EURO
+ 0.0193 * SteClsCli=WT7,CT3,WT5,CT6,WT3,ST1,WT8,ST2,ST7,ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0217 * SteClsCli=ST8,WT6,ST6,WT9,ST9,ST3
+ 0.0001 * Mgs_duration
- 0.0005 * Apan_evap_08
- 0.0002 * WBSEP
+ 0.0717 * WBJUN
+ 31.0906

Number of Rules : 6

Appendix 11 continued.

=== Summary ===

Correlation coefficient	0.7218
Mean absolute error	2.5906
Root mean squared error	3.2791
Relative absolute error	67.4107 %
Root relative squared error	69.2135 %
Total Number of Instances	5456