UNIVERSITY OF KWAZULU-NATAL

Application of Statistical Multivariate Techniques to Wood Quality Data

2010

ASNAKE WORKU NEGASH

Application of Statistical Multivariate Techniques to Wood Quality Data

By

ASNAKE WORKU NEGASH

Submitted in fulfilment of the academic

Requirements for the degree of

MASTER OF SCIENCE

in

BIOMETRY

in the

School of Statistics and Actuarial Science

University of KwaZulu – Natal

Pietermaritzburg

2010

Dedication

To my father, Worku Negash , my mother Enane Fantaye and my brother Mesele Worku.

Declaration

The research work described in this thesis was carried out in the School of Statistics and Actuarial Sciences, University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Prof. Henry Mwambi and Prof. Temesgen Zewotir, in collaboration with Dr Tammy Bush.

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

May, 2010

Mr Asnake Worku Negash	**Date**
Prof. Henry Mwambi	**Date**
Prof. Temesgen Zewotir	**Date**

Acknowledgment

I would like to thank my supervisors Prof. Henry Mwambi and Prof. Temesgen Zewotir for their tireless guidance, encouragement and supervision in the preparation of this thesis. In addition, I extend special thanks to Dr. Tammy Bush for all support during my study. I would like to thank the staff members and postgraduate students in the school of statistics and actuarial sciences for the hospitable environment they provided during my study.

I would like to thank Central Statistical Agency and Council for Scientific and Industrial Research (CSIR) for providing the data and financial support for my study.

I thank my wife Meserte Sahilu, my son Edomias and my family for all the social, moral and logistical support throughout my life.

There are too many people close to me who were supporting me during this study period. I will not mention you all by names, except to say, thank you and I really love you.

Abstract

Sappi is one of the leading producer and supplier of *Eucalyptus* pulp to the world market. It is also a great contributor to South Africa economy in terms of employment opportunity to the rural people through its large plantation and export earnings. Pulp mills production of quality wood pulp is mainly affected by the supply of non uniform raw material namely *Eucalyptus* tree supply from various plantations. Improvement in quality of the pulp depends directly on the improvement on the quality of the raw materials. Knowing factors which affect the pulp quality is important for tree breeders.

Thus, the main objective of this research is first to determine which of the anatomical, chemical and pulp properties of wood are significant factors that affect pulp properties namely viscosity, brightness and yield. Secondly the study will also investigate the effect of the difference in plantation location and site quality, trees age and species type difference on viscosity, brightness and yield of wood pulp.

In order to meet the above mentioned objectives, data for this research was obtained from Sappi's P186 trial and other two published reports from the Council for Scientific and Industrial Research (CSIR). Principal component analysis, cluster analysis, multiple regression analysis and multivariate linear regression analysis were used. These statistical analysis methods were used to carry out mean comparison of pulp quality measurements based on viscosity, brightness and yield of trees of different age, location, site quality and hybrid type and the results indicate that these four factors (age, location, site quality and hybrid type) and some anatomical and chemical measurements (fibre lumen diameter, kappa number, total hemicelluloses and total lignin) have significant effect on pulp quality measurements.

Contents

List of Tables

List of Figure

Chapter 1

Introduction

South Africa is ranked as the 16th largest producer of pulp in the world and is largely self-sufficient in supplying pulp for the manufacture of printing and writing, packaging and tissue grades of paper (PAMSA, 2002). The forest products sector plays an important role in the South African economy. In 2002 the forest products accounted for 1.3% of the GDP (FAO, 2004). The pulp and paper industry directly and indirectly employed 23,981 people in 2006 (Edwards, 2006). More than 1.5 million hectares of South Africa is covered with industrial tree plantations. The pulp and paper industry is the main driver of the expansion of plantations and consumes over two-thirds of all the timber from South Africa's plantations. More than half of the plantation area is planted with pine where about one-third of this area is *Eucalyptus* and about one-fifth is acacia.

Two main companies dominate the pulp and paper industry in South Africa. One of the two dominant companies is Sappi. Sappi was registered in 1936 and today owns 465,000 hectares of plantations in South Africa (it has a further 75,000 hectares in Swaziland). Sappi has a strong export focus and operates in an open global market worldwide. The primary product of Sappi Saiccor is dissolving pulp which is used in manufacturing of a wide range of products that touch people's lives on a daily basis. These include textiles, food, chemicals and plastics (Sappi, 2008). In South Africa, pulp production and export stands at 2.406 million, and 671 thousand tons respectively. The value of pulp exports in 2007 was valued at 3747 million Rand (PAMSA, 2007) and still Sappi is currently expanding its Saiccor dissolving pulp mill to a target of more than 200,000 tons a year. The company also plans to expand pulp production at its Ngodwana mill by 225,000 tons a year. The company is planning to convert the plantations feeding its mill from pine to *Eucalyptus* (Lung, 2007).

Eucalypts are an important source of fibre and pulp for the South African pulp and paper industry. However *Eucalypts* plantation resources are extremely variable in nature, both with respect to rate of growth and quality (Zbonak, Bush and Grzeskowiak, 2007). In order to

manage their quality and value it is imperative to define the important quality measures, understand the extent of variation in quality and what drives it (Turner, 2001). The extent of variation associated with mainly *Eucalypts* creates the biggest problem but also the greatest opportunities for maximizing the value of plantation resources. Variations of the plantation resourses affect the fast growing demand for Sappi pulp wood and therefore the aim to produce a consistent, high quality product is a key issue for global competitiveness of the forestry and forest product industry. The primary constraint in achieving this in the short to medium term lies in the variability of the quality of the timber resource.

An important characteristic of South African forestry holdings lies in the wide range of site types and species planted. This, results in a resource of varying quality and value. Pulp mills often receive a wide variety of fibre types in terms of basic density, wood age, and from a variety of geographic locations. The quality of the raw material delivered to the pulp mills has a major effect on the productivity and efficiency of a mill. Mill performance and the value of pulp produced are related to the uniformity of the raw material being processed. For example, a high pulp yield and fast rate of delignification is only a positive attribute if all the material in the digester is of a similar type. If fast and slow cooking fibres are mixed in the same digester, one will over cook the fibre leading to lower pulp yield. On the other hand undercooking leads to an unacceptably high lignin residual. This means in practice that maximum value gain must be achieved in the first instance, through minimizing and or effectively managing variation at the pulp mill. A uniform mill furnish allows improved control overcooking conditions, less waste and reduction of variation in finished product. To address these problems Sappi regularly conducts different scientific and technological research in collaboration with the Council for Scientific and Industrial Research (CSIR).

CSIR is one of the leading scientific and technology research, development and implementation organization in Africa. It under-takes direct research and development for socio-economic growth. With research in forestry and forest products focused to a large extent on tree improvement, forest assessment, wood science and fibre processing. Research work conducted by CSIR indicate that variation in site index, age, clone type also affect density, extractives, pulp yield, brightness and other anatomical and chemical properties of the *Eucalypts* wood and the pulp (Turner et al, 1999, Megown, 2000, Zbonak, 2006).

Knowing the effect of variation in location, site quality, hybrid type and age, wood anatomical and chemical properties helps tree breeders and pulp producers to improve raw material uniformity for pulp mills. This improvement in raw material will help to increases quality of the pulp industries product and decrease production cost of the industry.

Two major objectives of this research are

1. to determine which of the anatomical, chemical and pulp properties of wood are significant factors that affect pulp properties namely viscosity, brightness and yield; and
2. to see the effect of geographic characteristics (location and site quality), age and species type on viscosity, brightness and yield of wood pulp.

To address these issues different statistical techniques such as principal component analysis, cluster analysis, multiple regression analysis and multivariate analysis of variance will be used. For the analysis two statistical packages namely SAS and GENSTAT will be used.

This thesis is organized as follows. This chapter gives some background about Sappi and introduction to the research. Chapter 2 presents data sources and exploratory data analysis. In Chapter 3 the data is analysed using multiple regression analysis and multiple comparison methods. Chapter 4 contains application and result of multivariate regression analysis. Chapter 5 presents a cluster analysis of the problem and finally Chapter 6 provides a summary and conclusions of the study.

Chapter 2

Exploratory data analysis

2.1 The data

The data used in this research is mainly obtained from Sappi's P186 trial and consisted of various measures (average diameter at breast height, average height, average height up to diameter of 7cm) obtained from different sub-tropical Eucalypts grown at several locations in Zululand, KwaZulu-Natal province. There are also two other additional groups of measurements. First the anatomical measurements which consist of fibre diameter, cell wall thickness, fibre lumen diameter and vessel percentage (Bush and Naidoo, 2008). Secondly the chemical measurements which consist of glucose, cellulose, SG ratio, total extractives, total hemicelluloses, total lignin and density (Bush, Naidoo and Gounden, 2008). Chemical measurements values are mostly from laboratory derived results and the rest are predicted values.

The entire data set in this project can be classified into two different data sets. The first data set is called different hybrid data which contains measurements from eighteen different eucalypts hybrid types planted at different period of time in eight different locations of Zululand, KwaZulu-Natal province, which has two site qualities. The second data set is E-dunnii data which contains measurements of eucalyptus hybrid type E-dunnii planted at the same time in one location (hellelo) which has one site quality.

The pulp properties namely viscosity, brightness and yield are the dependent variables. While the chemical and anatomical measurements and pulp property kappa number are the explanatory variables. Other variables which are categorical in nature are location, site quality, age and hybrid combination which will be used to account for observable source of variation among the observations in the analysis. Table 2.1 gives the codes of hybrid

combinations, locations, site quality and age categories. Table 2.2 gives the names and the codes for the three different classes of the independent variables as well as the three dependent variables used in the thesis. Thus in the entire thesis these respective codes will be inter-changeably used together with the actual names.

Table 2.1 Classification and levels of categorical variables for different hybrid and E-dunnii data sets

Hybrid combination		Location	Site quality	Age(in years)
GC	GU × (ET + GP)	P/Ridge B1	I	5
GP	GU × (G × GU)	P/Ridge C10	II	7
GU	GU×((GP) +(G × GT))	P/Ridge C13	III	8
UG	GU × ((GP) × E.ter)	P/Ridge D13		9
G × GU	E.uro × E.ter	Salpine E05		
GU × GC	E.uro × Gra/Ter	Terra A01		
GU × GT	E. grandis	KT E10		
G × GT	E.urophylla	KT E09		
GU×U	E.dunnii	Hliello		
GU × GP				

Table 2.2 Variable codes and description

Anatomical measurements		**Chemical measurements**		**Pulp properties**	
description	**code**	**description**	**code**	**description**	**code**
Average diameter at breast height	**dbh**	Glucose	**glu**	Kappa number	**kno**
Average height	**aht**	Cellulose	**cel**	Viscosity	**vis**
Average height of tree up to diameter of 7cm	**htc**	SG ratio	**sgr**	Brightness	**bri**
Fibre diameter	**fd**	Total extractives		Yield	**yld**
Cell wall thickness	**cwt**	Total hemicelluloses	**ths**		
Fibre lumen diameter	**fld**	Total lignin	**tli**		
Vessel percentage	**vp**				
Density	**dey**				

2.2 Preliminary Data Analysis

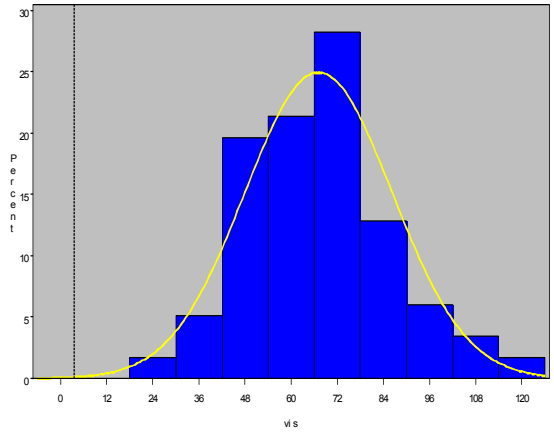
In this section some exploratory analysis of the data is presented. First we use the box plots as a graphical tool to check for possibility of outliers and differences between groups of the same variable. For example difference in site quality with respect to viscosity may give us insight about the effect of site on pulp viscosity. The scatter plots of the dependent variables (viscosity, brightness and yield) against the independent variables (chemical, anatomical and pulp property) are used to give an indication of the type of inherent relationship (linear or non linear). In addition normal probability plots for viscosity, brightness and yield are used to cheek if the response variables do conform to the normal distribution or not.

2.2.1 The different hybrid data set

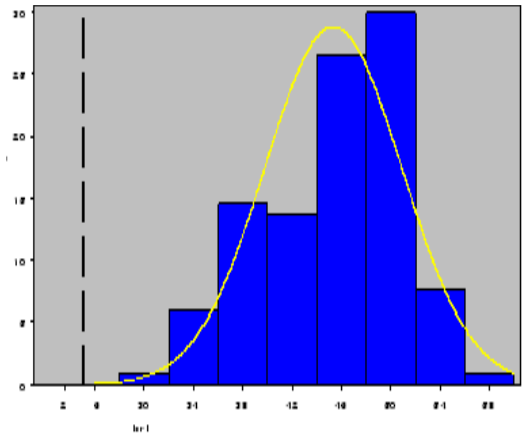
The summary statistics of viscosity, brightness and yield for the different hybrid data are presented in Table 2.3. Table 2.3 shows that there were 117 observations for the different hybrid data. For the variable yield the mean, median and the mode are not very far apart. But for viscosity and brightness the mode is far less than the median and the mean. In terms of variability, yield seems less variable than viscosity and brightness. The shape and peakdness measures show a slight positive skewness and peakedness for viscosity, a slight negative skewness and peakdness for brightness and yield. These characteristics are further supported by the graphical assessment presented in Figure 2.1. The graphical assessment shows that viscosity is slightly skewened to the right (Figure 2.1(a)). Brightness and yield are slightly skewened to the left as shown in Figures 2.1(b) and 2.1(c). Nevertheless the normal probability plots for viscosity, brightness and yield, in Figures 2.2- 2.4 show that all the three variables are approximately normally distributed except for some few outlier observations.

Table 2.3 Summary statistics of different hybrid data

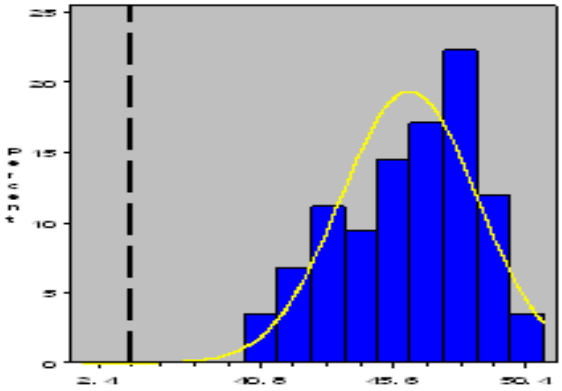
	Viscosity	Brightness	Yield
N	117	117	117
Mean	67.124217	45.325399	46.150282
Median	66.70000	46.45000	46.43000
Mode	47.60000	35.80000	44.19000
Variance	367.54357	30.751797	6.1259748
Std Deviation	19.171426	5.5454303	2.4750707
Skewness	0.4191059	-0.526332	-0.3769141
Kurtosis	0.3457136	-0.3330469	-0.7853556
Std Error Mean	1.7723989	0.5126752	0.2288204



(a) **Viscosity**



(b) Brightness



(c) Yield

Figure 2.1 The distribution of viscosity, brightness and yield for different hybrid data set

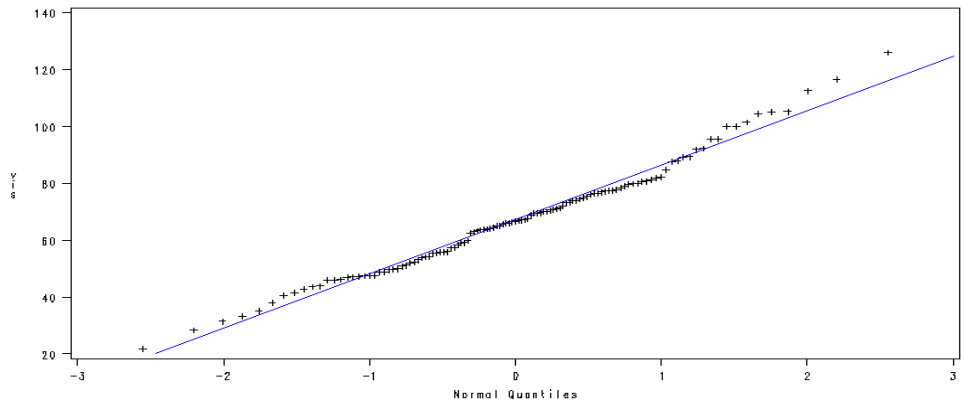


Figure 2.2 Normal Probability Plot of viscosity for different hybrid data

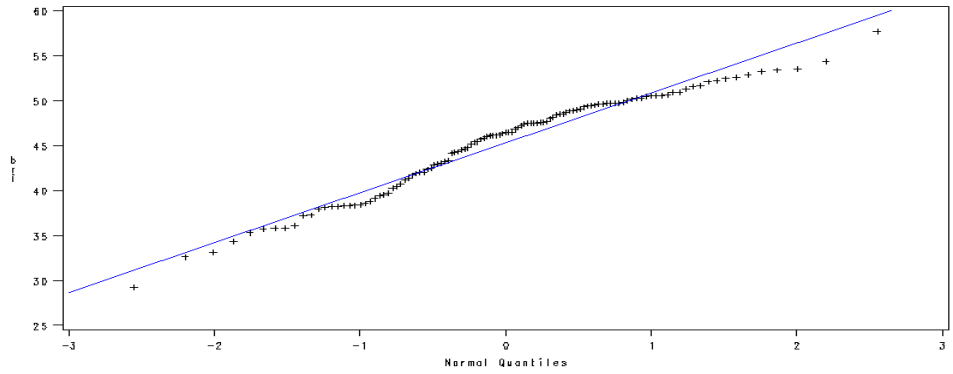


Figure 2.3 Normal Probability Plot of brightness for different hybrid data

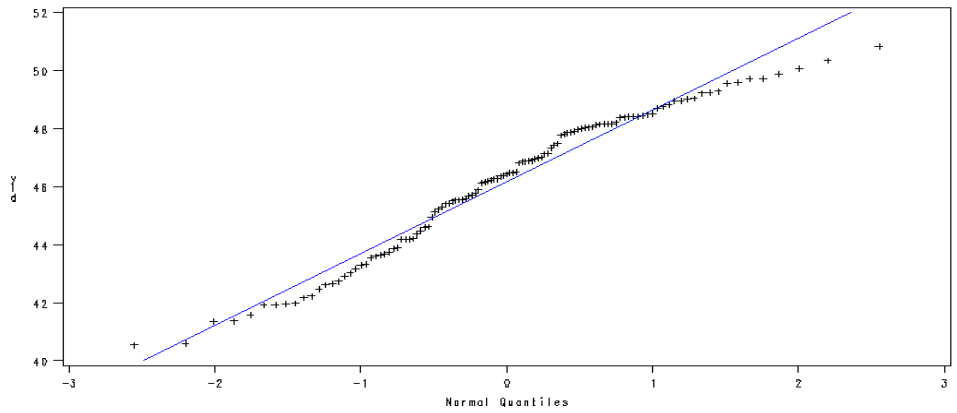


Figure 2.4 Normal Probability Plot of Yield for different hybrid data

A distributional assessment of viscosity, brightness and yield of the different hybrid data set against each categorical variable namely age, location and site quality are assessed using the box-plot in Figures 2.5 – 2.10. Viscosity versus age box-plot in Figure 2.5 indicate that age group 8 Eucalyptus trees have the highest mean and age group 7 have the lowest mean. Variation or dispersion of viscosity is lowest in Eucalyptus age category 9 compared to other age categories 5, 7 and 8 which have almost similar variation. The age category versus brightness box-plot, in Figure 2.6 shows that mean brightness increases with age, and also that the variation or dispersion is lowest for age category 9 while age category 7 brightness is the most variable. Similarly from the yield versus age box-plots in Figure 2.7 trees in age category 8 have the highest mean yield and those of age category 5 have the lowest mean yield. Yield distribution is most dispersed in age category 7 and least variable in age category 8.

The distribution of viscosity, brightness and yield by location is shown in Figures 2.8-2.10. These figures indicate that, Location KTE 10, P/ Ridge D13 and P/ Ridge C 13 have the highest variation in mean viscosity, brightness and yield respectively. Location KT G09 has the lowest mean variation for all of the three variables. Location P/ Ridge C 13 for viscosity, Salpine E05 for brightness and KT E10 for yield have the highest means compared to the other locations.

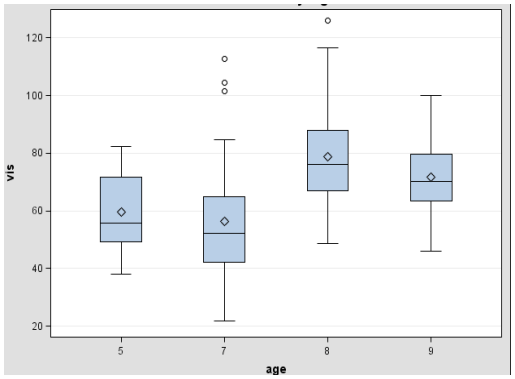


Figure 2.5 Distribution of viscosity by age for different hybrid data set

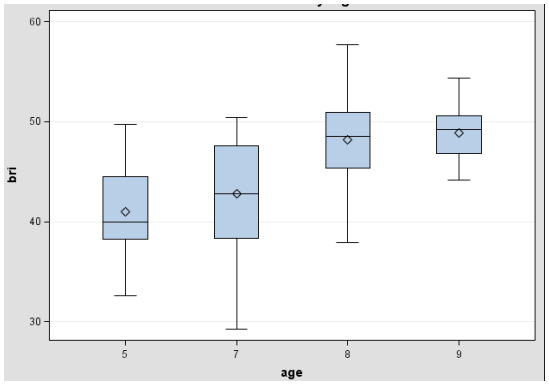


Figure 2.6 Distribution of brightness by age for different hybrid data set

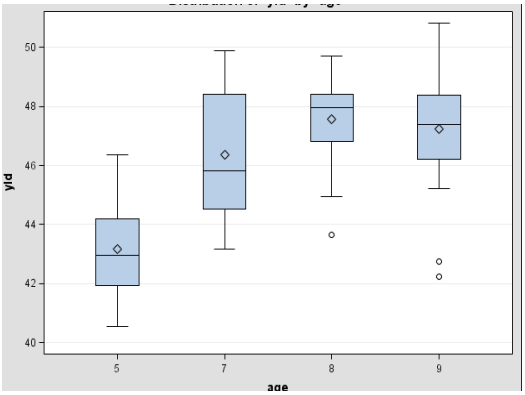


Figure 2.7 Distribution of yield by age for different hybrid data set

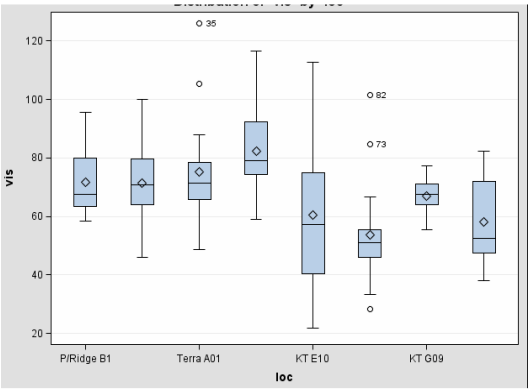


Figure 2.8 Distribution of viscosity by location for different hybrid data set

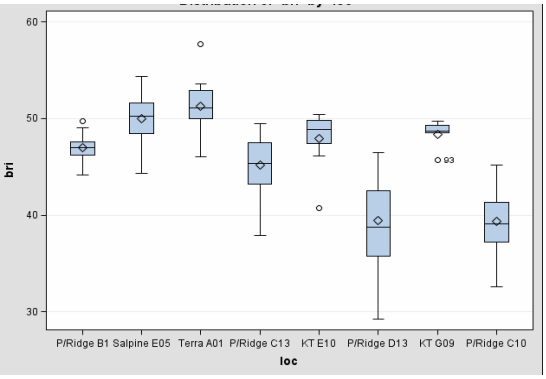


Figure 2.9 Distribution of brightness by location for different hybrid data set

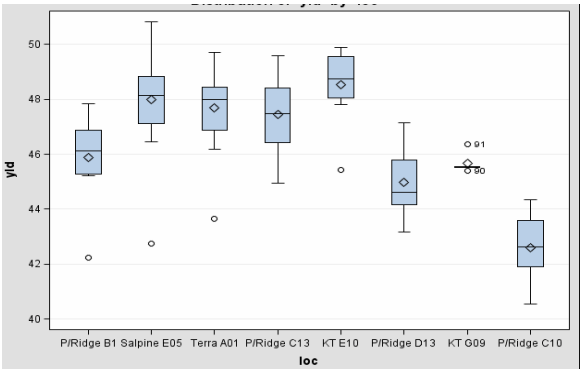


Figure 2.10 Distribution of yield by location for different hybrid data set

The distributional comparison of the two site quality groups in the different hybrid data set indicate that site quality '**I**' has a higher mean and lowest variation for all the three variables (viscosity, brightness and yield) compare to site quality "**II**" as presented in Figures 2.11-2.13 respectively.

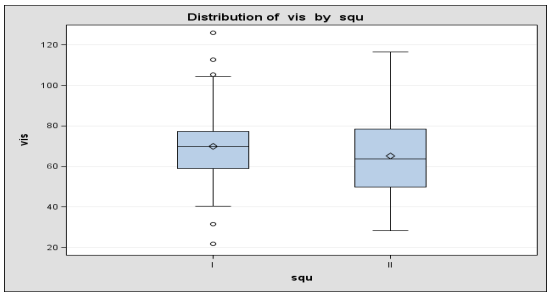


Figure 2.11 Distribution of viscosity by site quality for different hybrid data

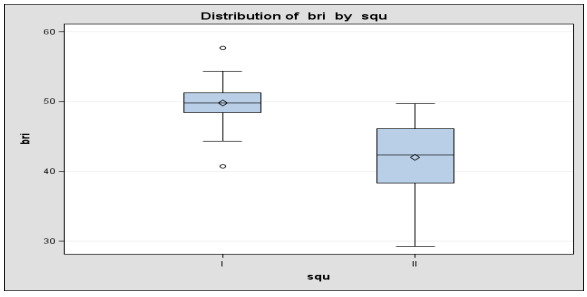


Figure 2.12 Distribution of brightness by site quality for different hybrid data

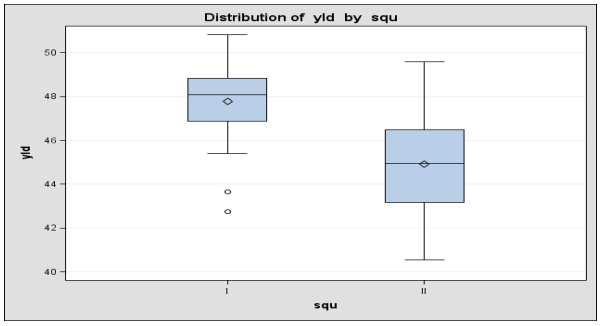


Figure 2.13 Distribution of yield by site quality for different hybrid data

2.2.2 The E-dunnii data set

The summary statistics of viscosity, brightness and yield using the E-dunnii data are presented in Table 2.4. Table 2.4 shows that there were 205 observations for viscosity and 297 observations for both brightness and yield. For brightness and yield the mean, mode and the median are almost equal. But for viscosity the mode is far less than the mean and the median. In terms of variability yield is less dispersed about its mean than brightness and viscosity. The shape and peakedness measures and the graphical assessment in Figure 2.14 show a slight positive skeweness for viscosity and yield, and a slight negative skeweness for brightness. The normal probability plots for viscosity, brightness and yield are shown in Figures 2.15-2.17. These figures show that all the three variables are approximately normally distributed except for the presence of some few outlier observations.

Table 2.4 Summary statistics for E-dunnii data

	Viscosity	Brightness	Yield
N	205	207	207
Mean	49.8304878	48.077053	41.5112
Median	45.3	48.3	41.11
Mode	32.4	48.3	41.08
Variance	313.076039	14.076376	3.31767
Std Deviation	17.6939549	3.7518497	1.82145
Skewness	0.50333169	-0.739232	0.73076
Kurtosis	-0.555596	1.3126814	0.05478
Std Error Mean	1.23579941	0.2607716	0.1266

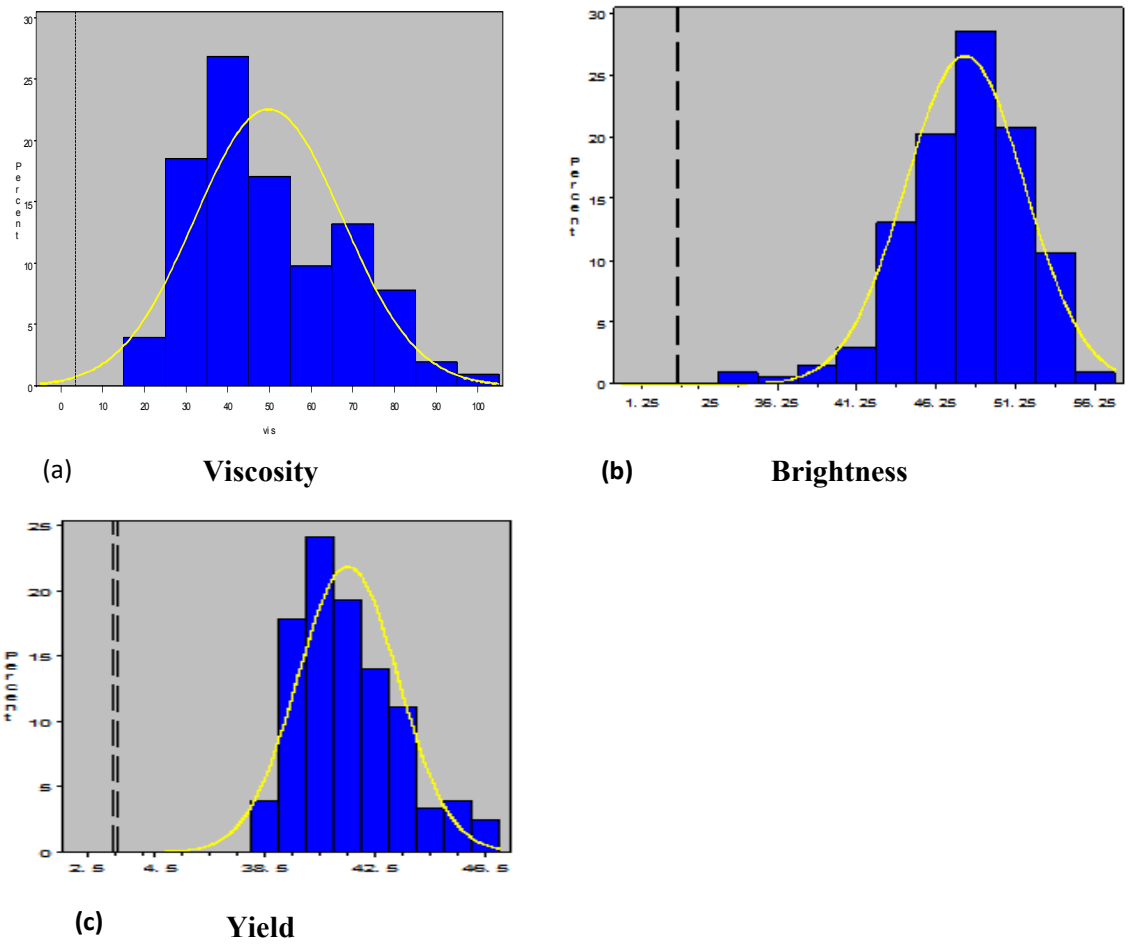


Figure 2.14 The distribution of viscosity, brightness and yield for E-dunnii data set

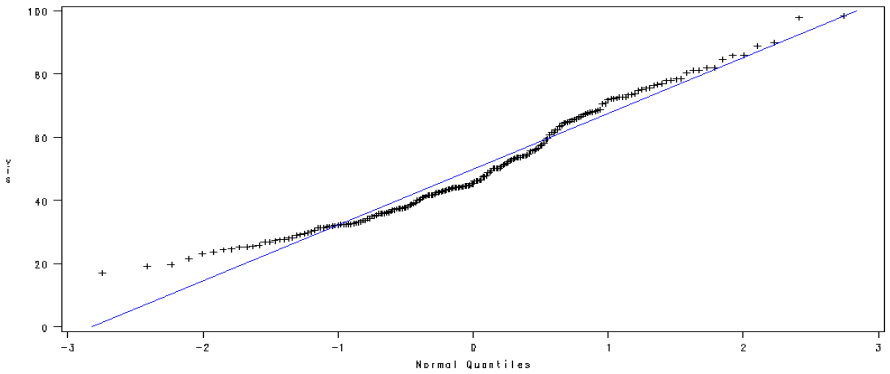


Figure 2.15 Normal Probability Plot of viscosity for E-dunnii data

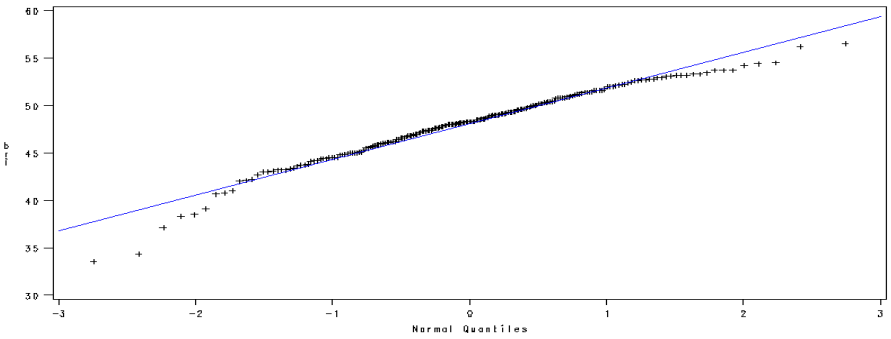


Figure 2.16 Normal Probability Plot of brightness for E-dunnii data

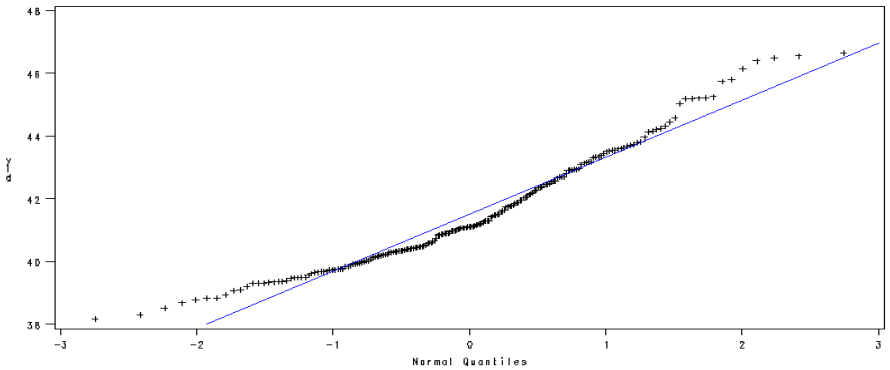


Figure 2.17 Normal Probability Plot of yield for E-dunnii data

2.3 Correlation and scatter plots

The correlation coefficient within or between any two anatomical, chemical and pulp property measurements indicate the strength and the direction of the linear relationship between the two measurements. Table 2.5 gives the correlation matrix for the different hybrid data set. We defined a strong correlation to be a value of magnitude 0.93-0.99. The values indicate a strong correlation within anatomical measurement namely average diameter at breast height, average height and average height of tree up to diameter of 7cm. There is also a high correlation (correlation coefficient values of 0.72-0.75) within measurement fibre lumen diameter, vessel percentage and cell wall thickens. Chemical measurement cellulose, total lignin and total extractives and pulp property measurement brightness, yield and kappa

number exhibit correlation ranges from mild ones up to high values that is correlation coefficient values from -0.55 between brightness and kappa number up to 0.79 between glucose and cellulose. There are also approximately similar correlation strengths between variables of E-dunnii data set as shown in Table 2.6.

Graphics such as the scatter plot matrix can also be very useful in choosing predictor variables in multiple regression models. The scatter plot matrix is just a two-dimensional array of two-dimensional plots, where each frame contains a scatter diagram. Thus, each plot is an attempt to shed light on the relationship between a pair of variables. This is often a better summary of the relationships than a numerical summary (such as displaying the correlation coefficients between each pair of variables) because it gives a sense of linearity or nonlinearity of the relationship and some awareness of how the individual data points are arranged over the range of values of any pair of variables.

The scatter plot matrix in Appendix A Figures A.1 and A.2 shows the presence of a possible linear relationship between viscosity or brightness or yield with anatomical or chemical measurements. For example in the scatter plot of the different hybrid data set for yield versus average height (Appendix A.1(a)) or glucose (Appendix A.1(d)) indicates the presence of positive linear relationship between these measurements. A scatter plot of brightness versus total extractives of the E-dunnii data set also shows a negative linear relationship between these two variables (Appendix A.2 (e)).

Table 2.5 Correlation matrix for variables using different hybrid data

vis																		
bri	0.2366																	
yld	0.1322	0.7021																
dbh	0.3081	0.7118	0.7627															
aht	0.3225	0.7653	0.8233	0.9292														
htc	0.3086	0.7498	0.8241	0.9462	0.9908													
fd	0.4028	0.5289	0.5678	0.5936	0.6311	0.6202												
cwt	0.2634	0.4669	0.4155	0.5341	0.5263	0.5301	0.3133											
fld	0.3207	0.2127	0.4271	0.3223	0.3879	0.371	0.7188	-0.166^{ns}										
vp	0.4617	0.4634	0.4199	0.5062	0.537	0.5321	0.7277	0.7557	0.2709									
dey	-0.125^{ns}	-0.347	-0.3161	-0.184^{ns}	-0.3217	-0.2715	-0.3259	0.174^{ns}	-0.4055	-0.074^{ns}								
kno	0.032^{ns}	-0.5573	-0.4569	-0.2549	-0.2657	-0.2449	-0.3821	-0.168^{ns}	-0.3165	-0.2559	0.3277							
glu	0.1973	0.649	0.7565	0.6973	0.7531	0.7519	0.542	0.6368	0.2776	0.5906	-0.22	-0.4003						
cel	0.2292	0.5291	0.5812	0.5407	0.5926	0.601	0.413	0.6125	0.191	0.5757	-0.183^{ns}	-0.3568	0.7895					
sgr	-0.2363	0.157_{ns}	0.2589	0.014^{ns}	0.019^{ns}	0.005^{ns}	0.011^{ns}	-0.2376	0.166^{ns}	-0.2669	-0.222	-0.4198	0.114^{ns}	-0.044^{ns}				
tex	-0.152^{ns}	-0.3729	-0.5152	-0.3927	-0.4675	-0.4436	-0.2773	-0.4918	-0.2065	-0.4191	0.2986	0.3775	-0.7029	-0.7169	-0.091^{ns}			
ths	-0.5958	-0.5968	-0.4804	-0.6139	-0.629	-0.6159	-0.6841	-0.4458	-0.4838	-0.6086	0.1947	0.3135	-0.4055	-0.3375	0.104^{ns}	0.2321		
tli	-0.101^{ns}	-0.3092	-0.3867	-0.1554	-0.2007	-0.1916	-0.0727	-0.3282	-0.024^{ns}	-0.264^{ns}	0.131^{ns}	0.3986	-0.5865	-0.6174	-0.171^{ns}	0.632^{ns}	-0.002	
	vis	bri	yld	dbh	aht	htc	fd	Cwt	fld	vp	dey	kno	glu	cel	sgr	tex	ths	tli

ns: Not significant at $\alpha = 0.05$

Table 2.6 Correlation matrix for variables using E-dunnii data

vis																			
bri	-0.2135																		
yld	-0.3731	0.4192																	
dbh	0.044^{ns}	-0.086^{ns}	0.2228																
aht	-0.040^{ns}	0.043^{ns}	0.3383	0.621															
htc	-0.048^{ns}	-0.044^{ns}	0.2587	0.7452	0.806														
fd	0.067^{ns}	-0.052^{ns}	0.109^{ns}	0.075^{ns}	0.112^{ns}	0.120^{ns}													
cwt	-0.05^{ns}	0.2987	0.3807	0.140^{ns}	0.3468	0.2513	-0.2929												
fld	0.3103	-0.2762	-0.2772	-0.1626	-0.1864	-0.1569	0.6914	-0.6811											
vp	-0.113^{ns}	0.2459	0.4923	0.098^{ns}	0.3112	0.2322	0.706	0.3348	0.143^{ns}										
dey	-0.137^{ns}	0.3099	0.122^{ns}	0.056^{ns}	0.010^{ns}	0.067^{ns}	-0.2435	0.060^{ns}	-0.2661	-0.1694									
kno	0.4167	-0.4527	-0.118^{ns}	0.2943	0.1655	0.2563	0.093^{ns}	-0.095^{ns}	0.095^{ns}	-0.130^{ns}	-0.050^{ns}								
glu	-0.2881	0.3386	0.4727	0.1897	0.2749	0.2057	-0.251	0.3628	-0.4067	0.100^{ns}	0.2302	-0.1794							
cel	-0.4083	0.4206	0.6623	0.019^{ns}	0.2189	0.113^{ns}	-0.0061	0.3979	-0.2424	0.4439	0.108^{ns}	-0.4752	0.457						
sgr	-0.133^{ns}	0.1681	0.094^{ns}	-0.008^{ns}	0.002^{ns}	-0.012^{ns}	-0.038^{ns}	-0.027^{ns}	-0.1311	-0.001^{ns}	0.132^{ns}	-0.0905	0.2516	0.079^{ns}					
tex	0.3307	-0.4082	-0.5226	-0.006^{ns}	-0.157	-0.052^{ns}	0.024^{ns}	-0.309	0.214^{ns}	-0.4061	-0.094^{ns}	0.481^{ns}	-0.3962	-0.7045	-0.108^{ns}				
ths	-0.5027	0.136^{ns}	0.2046	0.080^{ns}	-0.011^{ns}	-0.006^{ns}	-0.187^{ns}	-0.071^{ns}	-0.2987	-0.126^{ns}	0.2814	-0.2256	0.4465	0.1772	0.2518	-0.104^{ns}			
tli	0.148^{ns}	-0.2887	-0.2118	-0.06^{ns}	-0.109^{ns}	-0.058^{ns}	0.3635	-0.098^{ns}	0.334	0.2303	-0.41	0.1515	-0.5796	-0.129^{ns}	-0.321	0.2375	-0.4573		
	vis	bri	yld	dbh	aht	htc	fd	cwt	fld	vp	dey	kno	glu	cel	sgr	tex	ths	tli	

ns: Not significant at $\alpha = 0.05$

These observed linear relationships between variables were assessed using simple linear regression of each of the dependent variables (viscosity, brightness and yield) with each independent variables (chemical and anatomical measurements). Results from simple linear regression show the presence of some strong linear relationship between the dependent variables (viscosity, brightness and yield) and the independent variables (anatomical and chemical measurements) for each data set.

In summary the above exploratory analysis results indicate,

i. an approximate normal distribution for all the three dependent variables namely viscosity, brightness and yield
ii. a mean difference and distributional variability of viscosity, brightness and yield within different age group, location, site quality
iii. the presences of a significant linear correlation between and within anatomic, chemical and pulp property measurements and
iv. some significant simple linear regression of viscosity, brightness and yield with each of the anatomic and chemical measurements. All these results suggest the application of multiple linear regression and a multiple comparison of mean viscosity, brightness and yield which is under taken in the next chapter.

Chapter 3

Multiple Regression

Introduction

Regression analysis is a statistical methodology that utilizes the relation between two or more quantitative variables so that a response or outcome variable can be predicted from the other, or others (Kutner, Naershelm and Neter 2005). The relation between variables can be a functional or statistical relation. The functional relation is a perfect one and is expressed by a matimatical formula, Y= f(X). Given a particular value of x the function f yield the corresponding values of Y. The statistical relation unlike a functional relation is not a perfect one. It is expressed by a regression model which integrate some uncertainty in the determination of Y from a given value X= x.

A regression model is a formal means of expressing a tendency of the response variable to vary with the predictor variable in a systematic fashion and a scattering of points around the curve of statistical relationship. This model also postulates the presence of a probability distribution of the dependent variable (Y) for each levels of independent variable (X). The means of these probability distributions vary in some systematic fashion with X. This systematic relation of the mean of Y and X is the regression function of Y on X. In particular; we will deal with linear regression. First we present an overview of simple linear regression then we present a brief formulation of the multiple regression model.

3.1 Simple linear regression model

In simple linear regression model there is only one linear predictor variable and the regression function is linear. The model can be stated as follows

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (3.1)$$

where Y_i is the i^{th} value of the response variable.

x_i is the i^{th} predictor value which correspond to Y_i

β_o and β_1 are fixed parameters

ε_i is a random variable error term with mean $E(\varepsilon_i) = 0$ and variance σ^2

ε_i and ε_j are uncorrelated for all $i \neq j$ i = 1,2 3...n

The simple regression model (3.1) has also the following important characteristics

- Y_i is a random variable with the constant component $\beta_0 + \beta_1 x_i$ and random term ε_i
- $E(Y_i) = \beta_0 + \beta_1 x_i$ which is the regression function
- Two error term ε_i and ε_j are uncorrelated, so the responses Y_i and Y_j are uncorrelated
- The error term ε_i and the responses Y_i have the same constant variance(σ^2), regardless of the level of the predictor variable x.

The parameters β_o and β_1 in the simple regression model (3.1) are the regression coefficients. β_1 is the slope of the regression line. It indicates the change in Y per unit change in X. The parameter β_o is the Y intercept of the regression line which indicate the mean value of the Y at X = 0, if the scope of the model include X=0.

The regression model parameters estimation can be performed either by the method of least squares or by the method of maximum likelihood. The method of least squares requires that the estimators of $\beta_o$and β_1 are those values $\hat{\beta}_0$and $\hat{\beta}_1$ respectively, that minimize the sum of the n squared deviations for the given sample of observations $(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n)$.

Let (x_i , y_i) are observations, $i = 1,2,3, \cdots, n$. The error sum of squares is

$$Q = \sum(y_i - \beta_0 - \beta_1 x_i)^2 \quad (3.2)$$

The values of $\beta_o$and $\beta_1$that minimize Q can be derived by differentiating Q with respect to $\beta_o$and β_1. From which it follows that

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \frac{1}{n}(\sum y_i - \hat{\beta}_1 \sum x_i) = \bar{y} - \hat{\beta}_1 \bar{x}$$

where $\bar{x}$ and $\bar{y}$ are means of the X and Y observations respectively.

The least square estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ have several important properties first, $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combinations of the observations y_i and are also unbiased estimators of the model parameters $\beta_o$and β_1, respectively.

An important result concerning the quality of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ is the Gauss-Markov theorem which states that for the regression model (3.1) with assumptions E(ε_i) = 0 and var (ε_i) $= \sigma^2$ and uncorrelated errors the least square estimators are unbiased and have a minimum variance when compared with all other unbiased estimators that are linear combinations of y_i, that is least square estimates are best linear unbiased estimators (BLUE).

There are other useful properties of the least squares fit

- The sum of the residuals in any regression model that contains an intercept β_o is always zero, that is
$$\sum(y_i - \hat{y}) = \sum e_i = 0$$
- The sum of the observed values y_i equals the sum of the fitted value $\hat{y}$, or
$$\sum y_i = \sum \hat{y}_i$$
- The least square regression line always passes through the centroid [the point $(\bar{x}, \bar{y})$ of the data].
- The sum of the residuals weighted by the corresponding value of the regression variables always equals zero, that is,
$$\sum x_i e_i = 0.$$
- The sum of the residuals weighted by the corresponding fitted value always equals zero, that is
$$\sum \hat{y}\, e_i = 0.$$

The least squares method provides unbiased point estimators of β_o and β_1 that have minimum variance among all unbiased linear estimators whatever the distribution of the error terms ε_i and observation y_i is $i = 1,2, \cdots, n$. But to set up of interval estimates and to carry out hypothesise test the assumption of normally distributed error term ε_i is the standard assumption. It is also justifiable in many real world situations where regression analysis is applied.

Knowing the specified functional form of the probability distribution of the error terms is important to estimate the parameter β_o, β_1 and σ^2 by the method of maximum likelihood. The method of maximum likelihood chooses as estimates those values of the parameters that are most consistent with the sample data.

The regression model (3.1) implies that Y_i are independent normal random variables, with mean $E(Y_i) = \beta_0 + \beta_1 x_i$ and variance σ^2.

The density of an observation Y_i for normal error regression model (3.1) utilizing the fact that $E(Y_i) = \beta_0 + \beta_1 x_i$ and variance of Y_i σ^2 is

$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2}\left(\frac{(y_i - \beta_0 - \beta_1 x_i)}{\sigma}\right)^2\right]. \qquad (3.3)$$

The likelihood function for n observations $y_1, y_2, \ldots, y_n$ is

$$L(\beta_o, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{(y_i - \beta_0 - \beta_1 x_i)}{\sigma}\right)^2\right] \qquad (3.4)$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]. \qquad (3.5)$$

The log likelihood function for n observations $y_1, y_2, \cdots, y_n$ is

$$\ell = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \qquad (3.6)$$

The variance $\sigma^2$of the error terms is usually unknown and the values of β_o, β_1 and σ^2 that maximize the likelihood function or equivalently the log-likelihood functions are the maximum likelihood estimates. These estimates denoted by $\widehat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$, respectively can be estimated from the log likelihood function ℓ by taking partial derivatives of ℓ with respect to β_o, β_1 and σ^2, and equating each of the partials to zero. Simple algebraic manipulations gives estimating equations that are identical to the least square estimators for β_o and β_1. In other words the least squares estimates and maximum likelihood estimates of β_o and β_1 are identical under the normality assumption.

3.2 Multiple linear regression analysis

Multiple linear regression analysis is one of the most widely used statistical method. In multiple linear regression analysis one is often concerned with the nature and significance of the relationship between a dependent variable and several independent variables.

The main concerns in this method are:

- To identify the relative importance of the effect of given independent variables on the dependent variable.
- To determine the magnitude of the effect of given independent variables on the dependent variable.
- To identify which independent variable play a significant role on the dependent variable.
- Model adequacy where, one can consider the impact of including independent variables which were initially not included in a simpler model.

In general, multiple regression analysis is a statistical tool used to explain the relationship between the random dependent variable (Y) and the p fixed independent variables $[X_1, X_2, \cdots, X_P]$.

If we have p independent variables (X_1, $X_2, \cdots, X_p$), then a multiple regression model relating Y to $X_1, X_2, \cdots, X_P$ is given by

$$Y= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon_i , \quad i=1,2,\ldots,n \qquad (3.7)$$

where Y is the dependent variable

β_0 is the intercept

$\beta_1, \beta_2, \cdots, \beta_p$ are partial regression coefficients

$X_1, X_2, \cdots, X_p$ are independent variables

ε_i is an error term and we assume that the ε_i's are IID $N(0, \sigma^2)$, $i = 1,2, \cdots, n$.

From the model in (3.7) it follows that $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. The coefficient β_j measures the change in Y per unit change of X_j keeping other independent variables fixed.

In matrix model (3.7) can be written as

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad (3.8)$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \ldots & X_{p1} \\ 1 & X_{12} & X_{22} & \ldots & X_{P2} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{1n} & X_{2n} & \ldots & X_{pn} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

where $\mathbf{y}$ is an n- dimensional vector of observations, $\mathbf{X}$ is an n × (p+1) design matrix and $\boldsymbol{\varepsilon}$ is an n- dimensional vector of error terms and thus $\boldsymbol{\varepsilon} \sim \mathbf{N_n}(\mathbf{0}, \boldsymbol{\sigma^2 I})$ where $\mathbf{I}$ is n×n identity matrix.

The partial regression coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are unknown parameters, which we have to estimate in order to fully discuss the fitted regression model. The multiple regression model parameters estimation like the simple regression model can also be performed either by the method of least squares or by the method of maximum likelihood if the distributional assumption is imposed.

To estimate the parameters, we use least square equation which minimizes the error sum of squares, i.e. minimize $(\mathbf{y} - \mathbf{X\beta})'(\mathbf{y} - \mathbf{X\beta}) = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$.

To obtain the least square estimate we differentiate $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ with respect to $\boldsymbol{\beta}$, and solve the result by setting the derivative equal zero. This process results in the estimator

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} . \qquad (3.9)$$

The estimated variance-covariance matrix for $\widehat{\boldsymbol{\beta}}$ is given by;

$$\text{Var}\,(\hat{\beta}) = \sigma^2(\mathbf{X}'\,\mathbf{X})^{-1}$$

assuming that $\text{var}(\mathbf{y}) = \sigma^2\mathbf{I}$.

The method of maximum likelihood leads to the same estimator for $\boldsymbol{\beta}$ as those obtained by the method of least squares.

The analysis of variance approach may be used to test the adequacy of the model as follows. Let the total sum of squares be SST, the regression or explained sum of squares be SSR the error or residual sum of squares be SSE then

$$\text{SST} = \text{SSR} + \text{SSE}$$

where

$$\text{SST} = \mathbf{y}'\,\mathbf{y} - \left(\frac{1}{\text{n}}\right)\mathbf{y}'\mathbf{J}\mathbf{y}$$

$$\text{SSR} = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\,\boldsymbol{y} - \left(\frac{1}{\text{n}}\right)\mathbf{y}'\mathbf{J}\mathbf{y}$$

$$\text{SSE} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\,(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{y}'\,\mathbf{y} - \widehat{\boldsymbol{\beta}}'\mathbf{X}'\,\mathbf{y}\ .$$

Here **J** is an n × n matrix of ones. The degrees of freedom associated with SST, SSR and SSE are n-1, p-1 and n-p respectively. Thus the mean sum of squares due to regression (MSR) is given by

$$\text{MSR} = \frac{SSR}{p-1}$$

and the mean error sum of squares (MSE) is given by

$$\text{MSE} = \frac{SSE}{n-p}\ .$$

The mean error sum of squares (MSE) is an unbiased estimator of σ^2. The partitioned sum of squares is summarized in Table 3.1:

Table 3.1 Regression ANOVA table summary

Sources of variation	Degrees of freedom	Sum of squares	Mean of squares	F-ratio
Regression	p -1	$\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \left(\frac{1}{\mathrm{n}}\right)\mathbf{y}'\mathbf{J}\mathbf{Y}$	$\frac{SSR}{p-1}$	$\frac{SSR}{SSE}$
Error	n – p	$\mathbf{y}'\,\mathbf{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\,\boldsymbol{y}$	$\frac{SSE}{n-p}$	
Total	n -1	$\mathbf{y}'\,\mathbf{y} - \left(\frac{1}{\mathrm{n}}\right)\mathbf{y}'\mathbf{J}\mathbf{y}$		

To investigate whether the dependent variable (Y) and the independent variables X_1, X_2, X_3, $\cdots$, X_P have significant relation we have to test the hypothesis;

H_0: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$

H_1: not all β_j are zero [for j=1, 2 , $\cdots$, p at least one coefficient is different from zero]

To test the above hypothesis we use the test statistic $F_{cal} = \frac{\mathrm{MSR}}{\mathrm{MSE}}$ which is distributed as F with p and n-p degrees of freedom in the numerator and denominator respectively. If the $F_{cal} \leq F_{1-\alpha,p,n-p}$ then we fail to reject H_0 and conclude that, the independent variables do not have significant linear relationship to the dependent variable, where $F_{1-\alpha,p,n-p}$ is the (1-α/2)100 percentile from the F distribution based on α levels of significance. On the other hand, if $F_{cal} > F_{1-\alpha,p,n-p,}$this implies that we reject H_0 in favour of H_1 and we conclude that at least one independent variable has significant linear relation with the dependent variable.

If we reject the null hypotheses H_0: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_p = 0$ we have to test the coefficients individually to identify which one is significantly different from zero. To test each regression coefficient, the null hypothesis (H_0) and the alternative hypothesis (H_1) are stated as follows:

H_0: $\beta_j = 0$

H_1: $\beta_j \neq 0$

The test statistic is $\mathrm{t} = \frac{\hat{\beta}_j}{\mathrm{s.e}\left(\hat{\beta}_j\right)}$,

where

$\hat{\beta}_j$ is the estimated value of β_j and $s.e\left(\hat{\beta}_j\right)$ the standard error of $\hat{\beta}_j$. If $|t| \leq t_{1-\alpha,(n-p)}$ we fail to reject H_O and conclude that β_j is not significantly different from zero, otherwise if $|t|$

$> t_{1-\alpha,(n-p)}$ we reject H_O and conclude that β_j is significantly different from zero. Here $t_{1-\alpha,(n-p)}$ is the $(1-\alpha/2)100$ percentile from the t-distribution with n-p degrees of freedom.

It is useful to have some measure of how well the model fit the data. In multiple regression, coefficient of multiple determination (denoted by R^2) is the commonly used measure of model fit and it is given by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} .$$

It is a measure of the proportion of sum of squares explained by the model. However high R^2 does not necessarily imply model adequacy because it is an increasing function of the number of independent variables which is not a good caracterstic. Instead we prefer the adjusted coefficient of determination, which is given by

$$R^2_{adj} = 1 - \frac{(n-1)}{(n-p)} \frac{SSE}{SST} .$$

R^2_{adj} is more desirable because as p increases R^2_{adj}decreases but as $n \to \infty$ R^2_{adj} approaches the unadjusted R^2

3.3 Model diagnostics

The estimation and inference from the regression model depends on several assumptions. Thus one should always check the validity of these assumptions and conduct analysis to examine the adequacy of the model. Gross violation of the assumptions may yield an unstable model in the sense that a different sample could lead to a totally different model with opposite conclusions. We usually cannot detect departures from the underlying assumptions by examination of the standard summary statistics, such as the t or F statistics, or R^2. These are general model properties, and as such they do not ensure model adequacy.

Model assumptions need to be cheeked using regression diagnostics. Diagnostics methods are used to examine for instance the possibility that error variance are constant, or that there is any distributional deviation from normality. The data has to be cheeked for possible outliers that may exist. In general, model diagnostic methods are used to identify unusual behaviour of observations which is usually overlooked and can also be used to remedy these situations.

Diagnostics for the response variable are usually carried out indirectly through an examination of the residuals, so one should first define the residuals.

The residuals e_i is defined as the difference between the observed value y_i and the fitted value $\hat{y}_i$, it is also a measure of the variability in the response variable not explained by the regression model.

$$e_i = y_i - \hat{y}_i , \qquad i = 1,2, \dots , n.$$

Thus any unusual departures from the model assumption on the error should show up in the residuals. If the model is appropriate for the data, the observed residual e_i should reflect the properties assumed for the ε_i .

The relationship between $e \; and \; \varepsilon$ can be established by defining the 'Hat' matrix (**H**).

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} .$$

The hat matrix **H** symmetric and idempotent, that is $\mathbf{H}^2 = \mathbf{H}$; it is an n × n matrix that transforms the vector of observed values to a vector of fitted values. In geometric terms, the ith diagonal element of the hat matrix's is a standardized measure of the distance between the X value for the ith observation and the means of X values of all n observations. A large value h_{ii} indicates that ith observation is distant far from the center of all X observations.

The element h_{ii} of the **H** matrix also may be interpreted as the amount of leverage exerted by the ith observation y_i on the fitted value $\hat{y}_i$.The diagonal elements are often called the leverages.

Traditionally (Montgomery et al. , 2001, Siverpersad, 2007), it is assumed that an observation with a leverage which exceeds $\frac{2P}{n}$ is considered a severe leverage point, however, this cut off only applies to situations where $\frac{2P}{n} \leq 1$.

Deviations from assumption are often best detected by working with residuals that have the same precision. The variance of the residuals or errors can be estimated by the mean square of error (MSE) of a regression. The standardized residuals would be

$$d_i = \frac{e_i}{\sqrt{MSE}} \quad i = 1, 2, \cdots , n.$$

The standardized residuals have mean zero and approximately unit variance. Consequently, a large absolute value standardized residual ($|d_i| > 3$), potentially indicates an outlier (Montgomery et al. , 2001).

Studentized residuals, also known as internally sudentized residuals, are defined as

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \quad i = 1, 2, \cdots, n.$$

where $s = \frac{\sqrt{\sum e_i{}^2}}{n-p}$.

Studentized residuals are scale-free, and its structure makes it t-like (however it does not exactly follow the t –distribution). It is very useful to the residual diagnostic analysis. The appearance of a large studentized residuals (r_i) values indicate a possible regression model assumption violation.

The other residual that is computed from $y_i - \hat{y}_{(i)}$, where $\hat{y}_{(i)}$ is the fitted value of the ith response based on all observations except the ith one is called PRESS residuals (because of their use in computing the prediction error sum of squares) sometimes it is also called deleted residual.

If the ith observation y_i is really unusual, the regression model based on all observation may be influenced by this observation. This could produce a fitted value $\hat{y}_i$ that is very similar to the observed value y_i, and consequentially, the ordinary residual e_i will be small. Therefore, it will be hard to detect the outlier. However, if the ith observation is deleted, then $\hat{y}_{(i)}$ cannot be influenced by that observation, so the resulting residual should be likely to indicate the presence of the outlier. The ith PRESS residual $e_{(i)}$ is

$$e_{(i)} = y_i - \hat{y}_{(i)} .$$

The variance of the ith PRESS residual is

$$var(e_{(i)}) = \frac{\sigma^2}{1-h_{ii}} .$$

So that a standardized PRESS residual is $\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}}$.

which, if we use MSE to estimate σ^2, it is just the studentized residual. Generally, a large difference between the ordinary residual and the PRESS residual will indicate a point where the model fit the data well, but a model built without that point predicts poorly.

Let an estimate of σ^2 based on a data set with the ith observation removed be $s_{(i)}{}^2$

$$s_{(i)}{}^2 = \frac{(n-p)MSR - e_i{}^2/(1-h_{ii})}{n-p-1}.$$

The estimate of σ^2 is used instead of MSR to produce an externally studentized residual, usually called R-Student, given by

$$t_i = \frac{e_i}{\sqrt{s_{(i)}{}^2(1-h_{ii})}}, i = 1,2,\dots,n.$$

In many situations t_i will differ little from the studentized residual r_i. However, if the ith observation is influential, then $s_{(i)}{}^2$can differ significantly from MSE, and thus the R-student statistic will be more sensitive to this point.

One of a model diagnostics measure is Cooks's distance D_i. Cooks's distance D_i is an aggregate influence measure, which is used to measure the effect of the ith observation on all n fitted values (Kutner and Neter, 2002, Moeti, 2007). Cook's distance measures the effect of deleting a given observation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a Cook's distance of 1 or more are considered to merit closer examination in the analysis.

Cook's distance is defined as

$$D_i = \frac{e_i}{\mathrm{P}\times(MSE)}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

where h_{ii} is the i^{-th} diagonal element of the hat matrix ; e_i is the crude residual (i.e. the difference between the observed value and the value fitted by the proposed model); MSE is the mean square error of the regression model and *p* is the number of fitted parameters in the model.

Residuals can produce information regarding the validity of the normality assumption on the model errors. Normality of ε_i is required for the validity of hypothesis testing and confidence interval estimation.

A simple method of checking the normality assumption, apart from the histogram and the stem and leaf plot, is the construction of normal probability plots of the residuals. The normal probability plot is a plot of each residual versus its expected values under the normality assumption. The residual plot should be approximately a straight line if the

normality assumption holds. A departure from a straight line indicates that the distribution is not normal.

3.4 Application of Multiple Linear Regression to the different hybrid data

In wood science pulp properties namely viscosity, brightness, and yield are expected to be mainly influenced by anatomical and chemical properties of wood. Hence, to investigate what effect these anatomical and chemical properties have on the pulp properties simultaneously, it is appropriate to use multiple regression techniques as a method to deal with multiple predictor variables.

Multiple regressions with viscosity, brightness and yield as dependent variables for the different hybrid data is presented in this section.

3.4.1 Viscosity

Results from multiple regression analysis with viscosity as the dependent variable for the different hybrid data are presented in Tables 3.2, 3.3 and Fig 3.1. The P- value for the F-statistics is less than 0.0001 (see Table3.2). This shows that there is an overall highly significant linear relation between the dependent variable viscosity and at least with one of the independent variables among the anatomical, chemical measurement and pulp properties kappa number. The R-square value indicates that 52.73% of the total variation is explained by the fit.

Individual t-tests in Table 3.3 indicate that kappa number (kno), total hemicelluloses (ths) and total lignin (tli) significantly affect viscosity at 5% significance level. However the variance inflation factor (VIF) value for some variables are also greater than 10 which is an indication of the problem of multicollinearity that may be present in the analysis. The residual plots of Figure 3.1 aid in some model diagnostics. The raw residuals versus predicted values (Figure 3.1(a)) indicate that the variance of the errors is not constant. The outward-opening funnel pattern of the residuals implies that variance is an increasing function of viscosity. The Hat

diagonal versus R-student residual plot (Figure 3.1(c)) also indicates the presence of some leverage and outlier observations. Even if some observations are leverage and outliers, the cook's distance versus observation number plot (Figure 3.1(b)) shows that all the cook's distance values are below one which indicates that neither the outlier and leverage observations are influential. The normality plot (Figure 3.1(d)) also show that the assumption of normal distribution of residuals is not seriously violated.

Table 3.2 ANOVA with viscosity as the dependent variable (Different hybrid data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	20682	1378.82897	7.14	<.0001
Error	96	18540	193.12754		
Corrected Total	111	39223			
R-Square = 0.5273			Adj R-Square = 0.4534		

Table 3.3 Parameter estimates of independent variables for the viscosity model (Different hybrid data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	369.71656	130.07918	2.84	0.0055	0
dbh	1	1.50601	1.24213	1.21	0.2283	11.46528
aht	1	1.61402	2.22225	0.73	0.4694	77.24169
htc	1	-2.43717	2.27631	-1.07	0.2870	93.44716
fd	1	-10.30190	8.89429	-1.16	0.2496	9.71431
cwt	1	-13.55714	14.75988	-0.92	0.3607	9.35549
fld	1	3.70440	3.31790	1.12	0.2670	4.27375
vp	1	5.78612	2.93938	1.97	0.0519	9.73220
dey	1	-40.48337	57.91027	-0.70	0.4862	2.03038
kno	1	9.85184	2.72017	3.62	0.0005	1.93946
glu	1	-1.05619	1.37410	-0.77	0.4440	7.07488
cel	1	0.45645	1.21617	0.38	0.7083	4.01700
sgr	1	-2.07626	7.24148	-0.29	0.7749	1.70555
tex	1	0.64141	1.48223	0.43	0.6662	3.22848
ths	1	-3.81009	0.70878	-5.38	<.0001	3.00036
tli	1	-3.92831	1.58187	-2.48	0.0148	2.87055

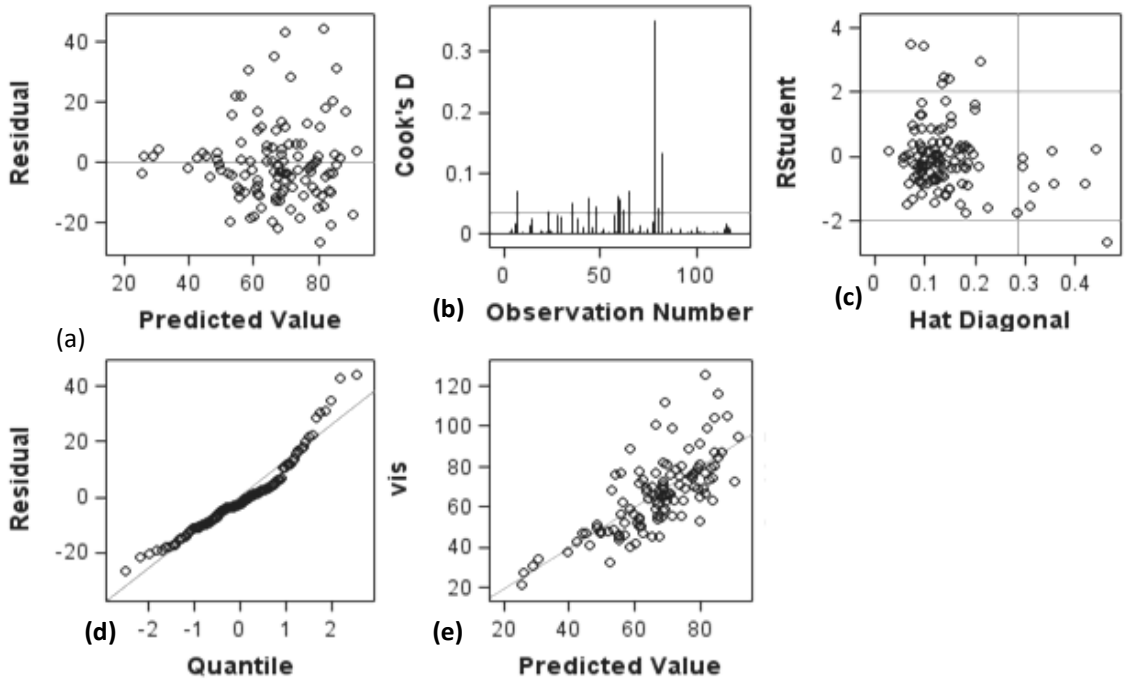


Figure 3.1 Model diagnostics with viscosity as the dependent variable (different hybrid data) (a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed viscosity versus predicted values

3.4.2 Box-Cox transformation of viscosity

The generalized least squares method can be used to solve the problem of non constant variance of residuals directly by specifying model for both the mean and variance, and then estimating the relevant parameters simultaneously. The alternative approach commonly used and applied in statistics is to transform the data in such a way as to obtain a new regression model with the desired properties namely constant variance and / or normally or symmetrically distributed errors (Carroll and Ruppert, 1980). The multiple regression diagnosis results showed that the transformation of viscosity to solve the problem of non constant variance of residuals is important. One convenient way of transformation is the Box-Cox transformation.

Box-Cox transformation is a transformation of the response variable of the form

$$y(\lambda) = \frac{(y^{\lambda}-1)}{\lambda} \quad \text{for } \lambda \neq 0$$

$$y(0) = \log(y) \text{ for } \lambda = 0$$

The definition of $\lambda = 0$ is the limit of the first expression, i.e $y(\lambda)$, as $\lambda \rightarrow 0$.

Note that the inclusion of this special case makes the transformation a continuous function of λ.

This family of transformations of the positive dependent variable y is controlled by the parameter λ. Transformations such as square root, inverse, quadratic and cubic, are special cases of Box-Cox transformation.

Using Proc TRANSREG in SAS a Box-Cox transformation was performed for the dependent variable viscosity as a remedial measure for the non constant error variance of the regression. Preliminary results are shown in Table 3.4. The best value of lambda was determined as zero and thus natural log transformations of viscosity is the most suitable to solve the problem of non-constant variance of the residuals. Next multiple regression using log of viscosity as dependent variable was performed and the ANOVA Table 3.5 now show 62.61% of the total variation is explained by the fit, which is an improvement over the earlier fitted model of untransformed viscosity which explained 52.73% of the total variation. The regression coefficients based on log viscosity as dependent variable and individual t-tests (Table 3.6) indicate that vessel percentage (vp), kappa number (kno), total hemicelluloses (ths) and total lignin (tli) significantly affect log of viscosity at 5% significance level. Model diagnostics in Figures 3.2 (a) – (f) also supports a log transformation of viscosity as a means to resolve the problem of a non constant variance and leads to an improvement on the fitted model.

Table 3.4 Box-Cox transformation of viscosity (Different hybrid data)

TRANSREG Univariate History for BoxCox (vis)				
Model Statement Specification Details				
Type	DF	Variable	Description	Value
Dep	1	BoxCox(vis)	Lambda Used	0
			Lambda	-0.06
			R-square	0.62161
			Adj R-Sq	0.5625

Table 3.5 ANOVA with log viscosity as the dependent variable (Different hybrid data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	6.13462	0.40897	10.51	<.0001
Error	96	3.73428	0.03890		
Corrected Total	111	9.86890			
R-Square = 0.6216			Adj R-Sq = 0.5625		

Table 3.6 Parameter estimates of independent variables for the log viscosity model (Different hybrid data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	8.89798	1.84609	4.82	<.0001	0
dbh	1	0.02126	0.01763	1.21	0.2308	11.46528
aht	1	0.01486	0.03154	0.47	0.6385	77.24169
htc	1	-0.03073	0.03231	-0.95	0.3439	93.44716
fd	1	-0.18075	0.12623	-1.43	0.1554	9.71431
cwt	1	-0.13666	0.20947	-0.65	0.5157	9.35549
fld	1	0.06870	0.04709	1.46	0.1478	4.27375
vp	1	0.09964	0.04172	2.39	0.0189	9.73220
kno	1	0.16141	0.03860	4.18	<.0001	1.93946
dey	1	-0.83127	0.82187	-1.01	0.3143	2.03038
glu	1	-0.02289	0.01950	-1.17	0.2434	7.07488
cel	1	0.01502	0.01726	0.87	0.3862	4.01700
sgr	1	-0.02527	0.10277	-0.25	0.8063	1.70555
tex	1	0.01524	0.02104	0.72	0.4706	3.22848
ths	1	-0.06584	0.01006	-6.55	<.0001	3.00036
tli	1	-0.05744	0.02245	-2.56	0.0121	2.87055

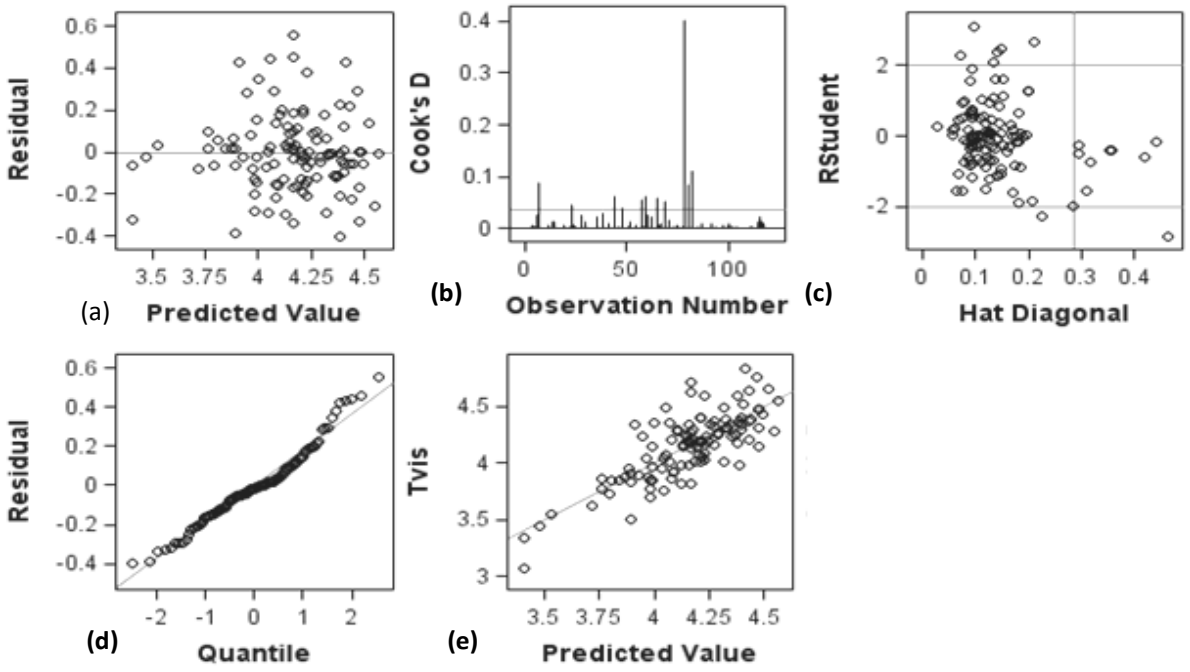


Figure 3.2 Diagnostic test with log viscosity as dependent variable (different hybrid data) (a) raw residuals versus predicted values (b Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed log viscosity versus predicted values

3.4.3 Brightness

Multiple regression analysis ANOVA results with brightness as the dependent variable and anatomical, chemical and pulp property measurement kappa number as explanatory variables are presented in Table 3.7 below. The P- value of F statistics is <.0001 indicating the presence of highly significant linear relationship between brightness and at least with one of the explanatory variables (anatomical and chemical measurement and pulp property kappa number).

The individual t-test values for each independent variable presented in Table 3.8 indicate that fibre diameter (fd), kappa number (kno), total hemicelluloses (ths) and total lignin (tli) significantly affect brightness at 5% significance level. The regression model also explains 81.20 % of the total variation based on the unadjusted R- square but adjusted R- square is 78.27%. Some VIF values are greater than 10 which indicate the presence of a multicollinearity problem.

Residual plots for model diagnosis are presented in Fig 3.3. A plot of raw residuals versus predicted values (Figure 3.3 a) does not show any systematic pattern, meaning that the assumption of constant error term variance is not violated. The plot R-studentized residuals versus Hat diagonal elements shows the presence of some outlier and leverage observations (Figure 3.3c). But the Cook's distance versus observation number plot (Figure 3.3b) indicate that these leverage and outlier observations are not influential because the cook's distance of each observation is below one. The residual versus normal quantile plot (Figures 3.3d) also indicate that the error term is approximately normally distributed.

Table 3.7 ANOVA with brightness as the dependent variable (Different hybrid data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	2767.95225	184.53015	27.65	<.0001
Error	96	640.69869	6.67394		
Corrected Total	111	3408.65095			
R-Square = 0.8120			Adj R-Square = 0.7827		

Table 3.8 Parameter estimates of independent variables for the Brightness model (Different hybrid data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	61.29732	24.18115	2.53	0.0129	0
dbh	1	-0.17701	0.23091	-0.77	0.4452	11.46528
aht	1	0.76408	0.41311	1.85	0.0674	77.24169
htc	1	-0.03030	0.42316	-0.07	0.9431	93.44716
fd	1	2.41367	1.65341	1.46	0.1476	9.71431
cwt	1	-1.50851	2.74380	-0.55	0.5837	9.35549
fld	1	-2.83416	0.61678	-4.60	<.0001	4.27375
vp	1	-0.44382	0.54642	-0.81	0.4187	9.73220
kno	1	-2.87815	0.50567	-5.69	<.0001	1.93946
dey	1	-14.61864	10.76527	-1.36	0.1777	2.03038
glu	1	0.15302	0.25544	0.60	0.5506	7.07488
cel	1	0.14785	0.22608	0.65	0.5147	4.01700
sgr	1	0.41887	1.34616	0.31	0.7564	1.70555
tex	1	0.62520	0.27554	2.27	0.0255	3.22848
ths	1	-0.46418	0.13176	-3.52	0.0007	3.00036
tli	1	-0.47669	0.29406	-1.62	0.1083	2.87055

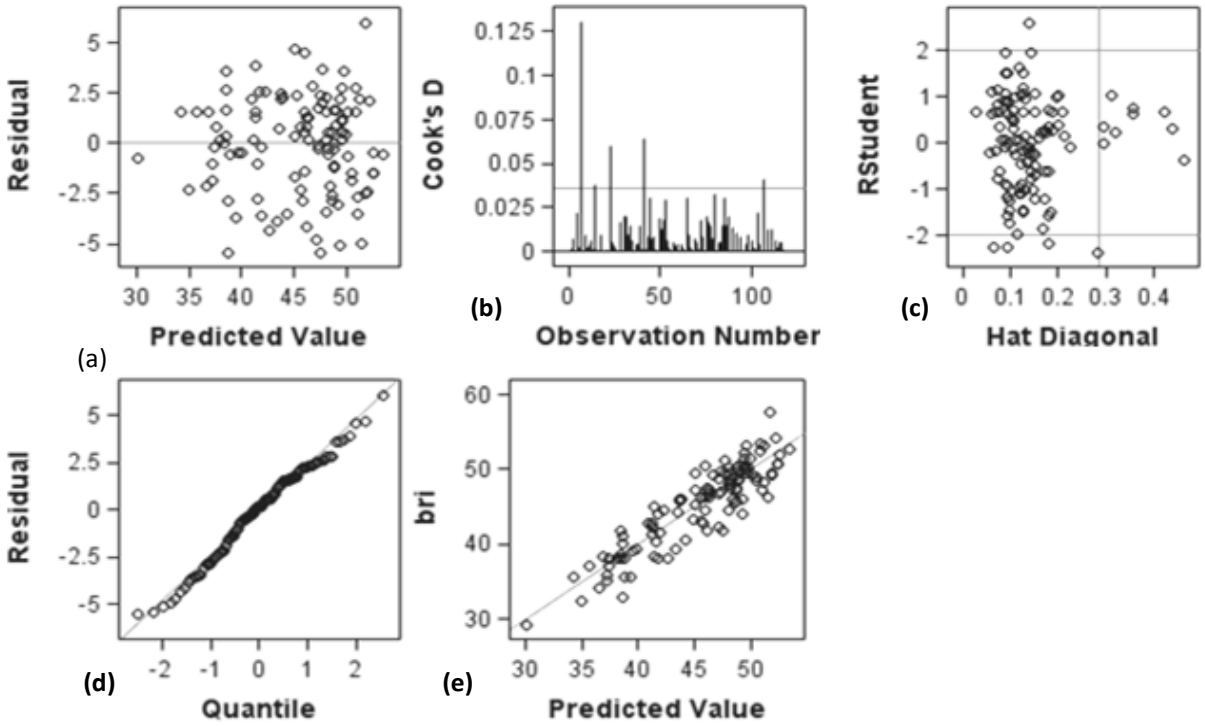


Figure 3.3 Model diagnostics with brightness as the dependent variable (different hybrid data) **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed brightness versus predicted values**

3.4.4 Yield

The ANOVA table of the multiple regression model where yield as dependent variable is presented in Table 3.9. The P-value of the F-statistics is < 0.0001 and it indicates the presence of highly significant linear relationship between yield and the independent variables at 5% significance level. A high proportion of the total variation is 81.27% explained by the fitted model. From the individual t-test values of each independent variables (Table 3.10) the average height of a tree up to diameter of 7cm (htc), kappa number (Kno), cellulose (cel) and total lignin (tli) significantly affect yield at 5% significance level. The multicolinarity measure namely variance inflation factor (VIF) values in Table 3.10 for some variables are greater than 10 indicating the presence of multicolinarity problem.

Model diagnosis plots of residuals in Figure 3.4 show no systematic patterns indicating that the assumption of constant error term variance is not violated (Figure 3.4 a). To test for

outlier or leverage and influential observations, a plot of R-student residuals versus Hat diagonal elements and cook's distance versus observation number, are used respectively (Figures 3.4 c, b). The first one shows the presence of some outlier and leverage observations but from the second plot namely the cook's distances of each observation are all below one. This indicates that those leverage and outlier observations are not influential. The residual versus normal quantile plot (Figures 3.4d) also indicate that the assumption of normally distributed errors is not seriously violated.

Table 3.9 ANOVA with yield as the dependent variable (Different hybrid data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	560.30359	37.35357	27.76	<.0001
Error	96	129.16803	1.34550		
Corrected Total	111	689.47162			
R-Square = 0.8127			Adj R-Square = 0.7834		

Table 3.10 Parameter estimates of independent variables for the yield model (Different hybrid data)

Parameter Estimates						
Variable	DF	Parameter Estimation	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	37.46268	10.85744	3.45	0.0008	0
dbh	1	-0.13912	0.10368	-1.34	0.1828	11.46528
aht	1	-0.11510	0.18549	-0.62	0.5364	77.24169
htc	1	0.49637	0.19000	2.61	0.0104	93.44716
fd	1	0.71399	0.74239	0.96	0.3386	9.71431
cwt	1	1.78132	1.23198	1.45	0.1515	9.35549
fld	1	0.39733	0.27694	1.43	0.1546	4.27375
vp	1	-0.38166	0.24534	-1.56	0.1231	9.73220
kno	1	-0.50032	0.22705	-2.20	0.0299	1.93946
dey	1	-0.12164	4.83365	-0.03	0.9800	2.03038
glu	1	0.05340	0.11469	0.47	0.6426	7.07488
cel	1	-0.11211	0.10151	-1.10	0.2722	4.01700
sgr	1	1.25220	0.60443	2.07	0.0410	1.70555
tex	1	-0.02186	0.12372	-0.18	0.8601	3.22848
ths	1	0.05014	0.05916	0.85	0.3988	3.00036
tli	1	-0.31192	0.13204	-2.36	0.0202	2.87055

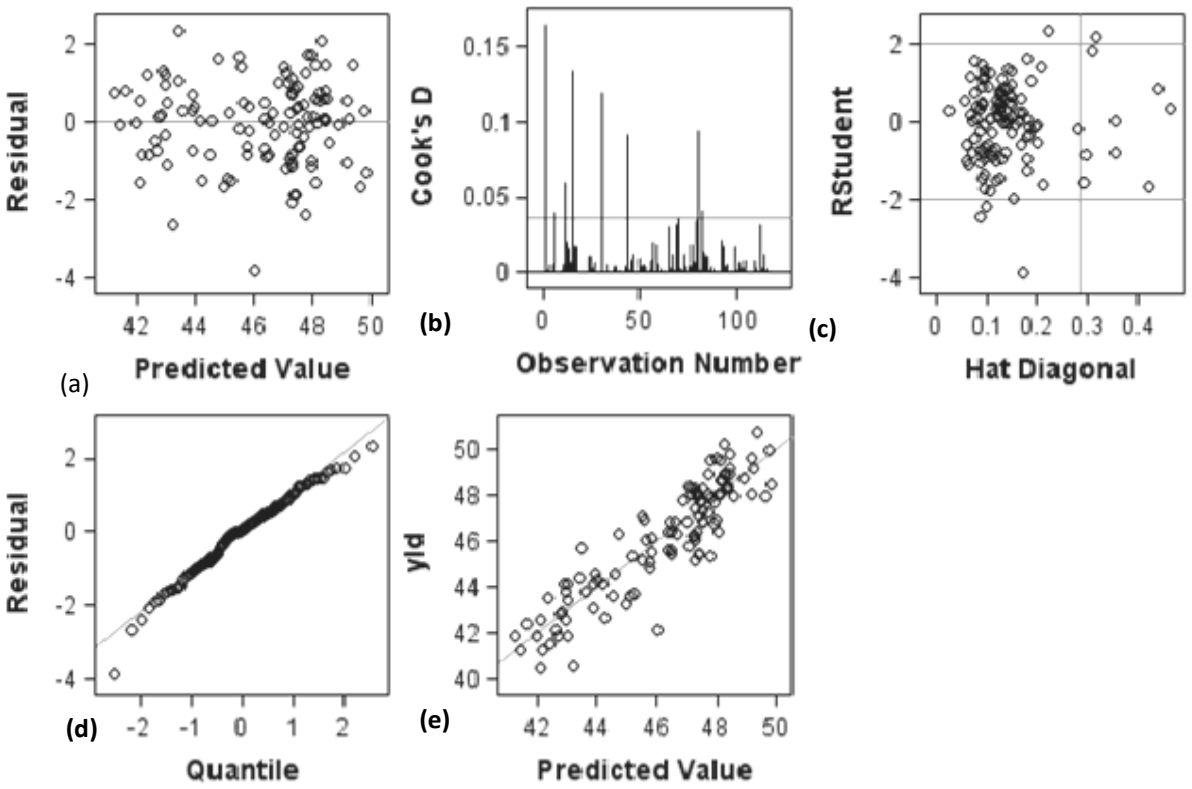


Figure 3.4 Model diagnostics with yield as the dependent variable (different hybrid data) **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residual versus Hat diagonals (d) normal probability plot of residual (e) observed yield versus predicted values**

3.5 Application of Multiple Linear Regressions to the E-Dunnii data

3.5.1 Viscosity.

The multiple regression analysis results for the dependent variable viscosity of the E-Dunnii data are presented in Tables 3.11 and 3.12. The P- value of the F- statistics is < 0.0001 (see Table 3.11). This shows that there is highly significant linear relation between the dependent variable viscosity and at least with one of the independent variables (anatomical, chemical measurements and pulp properties kappa number). The R-square value indicates that 53.84% of the total variation is explained by the fit.

The Parameter estimates and individual t-tests in Table 3.12 for estimation of the regression coefficients with viscosity as the dependent variable, indicate that average diameter at breast height (dbh), average height of a tree up to diameter of 7 cm (htc), cell wall thickness (cwt), fibre lumen diameter (fld), kappa number (Kno), cellulose (cel), total hemicelluloses (ths) and total lignin (tli) significantly affect viscosity at 5% significance level. However the variance inflation factor (VIF) values for some variables are greater than 10 which are an indication that the problem of multicollinearity may be present in the analysis.

The residual plots in Figure 3.5 shows some model diagnostics. The plots of raw residuals versus predicted values (Figure 3.5 a) indicate that the variance of the errors component may not constant. The outward-opening funnel pattern of the residual plots implies that variance is an increasing function of viscosity. The Hat diagonal elements versus R-student residual plot (Figure 3.5 c) also indicate the possible presence of some leverage and outlier observations. Even if some observations are leverage and outliers the cook's distance versus observation number plots (figure 3.5 b) show that all the cook's distance values are below one which implies that the outlier or leverage observations are not influential. The normality plots (Figures 3.5 d) further show that the assumption of normal distribution of residuals is not seriously violated.

Table 3.11 ANOVA with viscosity as the dependent variable (E-dunnii data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	30425	2028.35301	12.44	<.0001
Error	160	26082	163.01026		
Corrected Total	175	56507			
R-Square = 0.5384			Adj R-Square = 0.4952		

Table 3.12 Parameter estimates of independent variables for the viscosity model (E-dunnii data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	201.90888	85.58628	2.36	0.0195	0
dbh	1	2.90849	0.88109	3.30	0.0012	2.56954
aht	1	0.32983	1.44473	0.23	0.8197	3.29393
htc	1	-2.84573	1.15373	-2.47	0.0147	4.08711
fd	1	-14.38909	4.55712	-3.16	0.0019	13.21663
cwt	1	22.30387	7.50280	2.97	0.0034	4.63420
fld	1	14.55535	2.44984	5.94	<.0001	5.99532
vp	1	2.96440	1.71097	1.73	0.0851	11.20939
kno	1	7.50801	2.15827	3.48	0.0006	1.70694
dey	1	15.62947	34.37036	0.45	0.6499	1.29937
glu	1	-0.82641	0.88702	-0.93	0.3529	2.41556
cel	1	-1.89528	0.77140	-2.46	0.0151	2.80618
sgr	1	2.51144	3.46833	0.72	0.4701	1.19430
tex	1	0.91744	0.71436	1.28	0.2009	2.72304
ths	1	-1.38609	0.34811	-3.98	0.0001	1.97814
tli	1	-2.14395	1.01160	-2.12	0.0356	2.24588

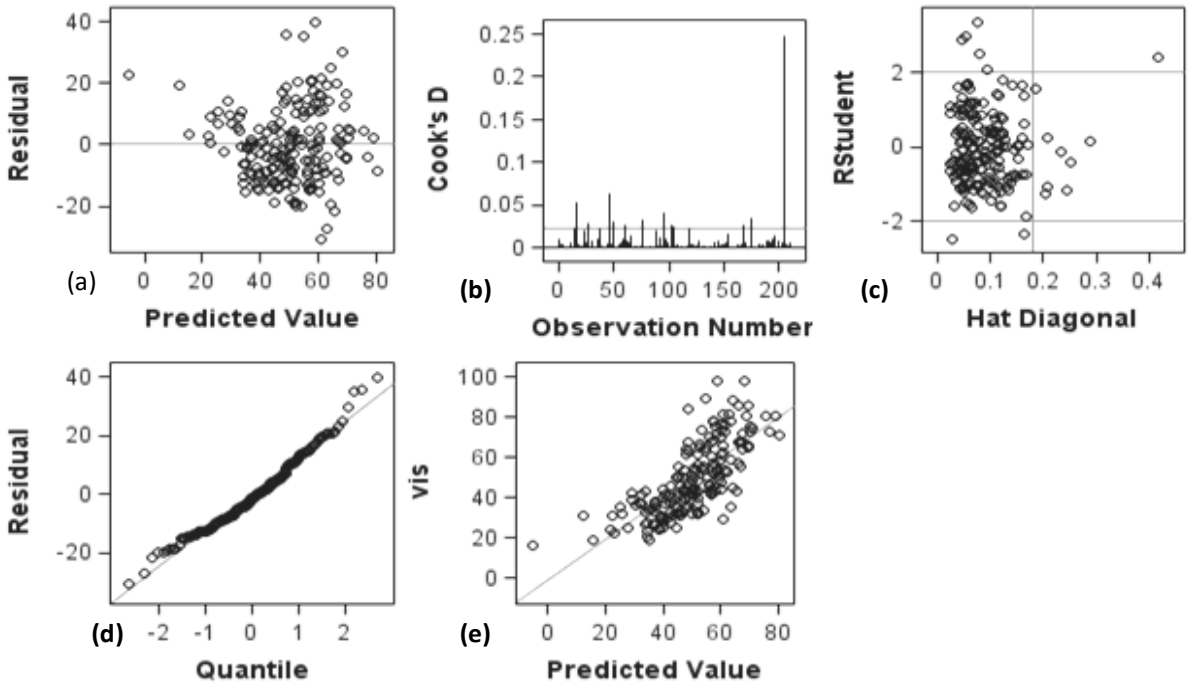


Figure 3.5 Model diagnostics with viscosity as the dependent variable **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residual versus Hat diagonals (d) normal probability plot of residual (e) observed viscosity versus predicted values**

3.5.2 Box-Cox transformation of viscosity

Using Proc TRANSREG in SAS a Box-Cox transformation was performed on the dependent variable viscosity as a remedial measure to make the data achieve a constant variance of the residuals. The results of this transformation are shown in Table 3.13. The best value of lambda is zero implying that the natural log transformations of viscosity is the most suitable to solve the problem of non-constant variance. The multiple regression ANOVA (Table 3.14) results show that a total of 59.54% of the total variation is explained by such a fit. The regression coefficients with log viscosity as the dependent variable and individual t-tests indicate that average diameter at breast height (dbh), average height of a tree up to diameter of 7 cm (htc), fibre diameter (fd), cell wall thickness (cwt), fibre lumen diameter (fld), kappa number (Kno), cellulose (cel), total hemicelluloses (ths) and total lignin (tli) are significantly associated with log viscosity at 5% significance level as tabulated in Table 3.15

Table 3.13 Box-Cox transformation of viscosity (E-dunnii data)

TRANSREG Univariate Algorithm History for BoxCox(vis)				
Model Statement Specification Details				
Type	DF	Variable	Description	Value
Dep	3	BoxCox(vis)	Lambda Used	0
			Lambda	-0.11
			R- square	0.59734
			Adj R- Square	0.5574

Table 3.14 Analysis of variance with log viscosity as the dependent variable (E-dunnii data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	14.21766	0.94784	15.69	<.0001
Error	160	9.66272	0.06039		
Corrected Total	175	23.88037			
R-Square = 0.5954			Adj R-Square = 0.5574		

Table 3.15 Parameter estimates of independent variables for the log viscosity model (E-dunnii data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6.90847	1.64735	4.19	<.0001	0
dbh	1	0.05017	0.01696	2.96	0.0036	2.56954
aht	1	0.00121	0.02781	0.04	0.9655	3.29393
htc	1	-0.04680	0.02221	-2.11	0.0366	4.08711
fd	1	-0.30593	0.08771	-3.49	0.0006	13.21663
cwt	1	0.45019	0.14441	3.12	0.0022	4.63420
fld	1	0.29726	0.04715	6.30	<.0001	5.99532
vp	1	0.06216	0.03293	1.89	0.0609	11.20939
kno	1	0.15860	0.04154	3.82	0.0002	1.70694
dey	1	0.31383	0.66156	0.47	0.6359	1.29937
glu	1	-0.00283	0.01707	-0.17	0.8688	2.41556
cel	1	-0.04605	0.01485	-3.10	0.0023	2.80618
sgr	1	0.06157	0.06676	0.92	0.3577	1.19430
tex	1	0.02210	0.01375	1.61	0.1099	2.72304
ths	1	-0.03315	0.00670	-4.95	<.0001	1.97814
tli	1	-0.04077	0.01947	-2.09	0.0379	2.245

Model diagnostics depicted in Figures 3.6 (a) – (e) also supports that a log transformation of viscosity resolves the problem of non-constant error variance.

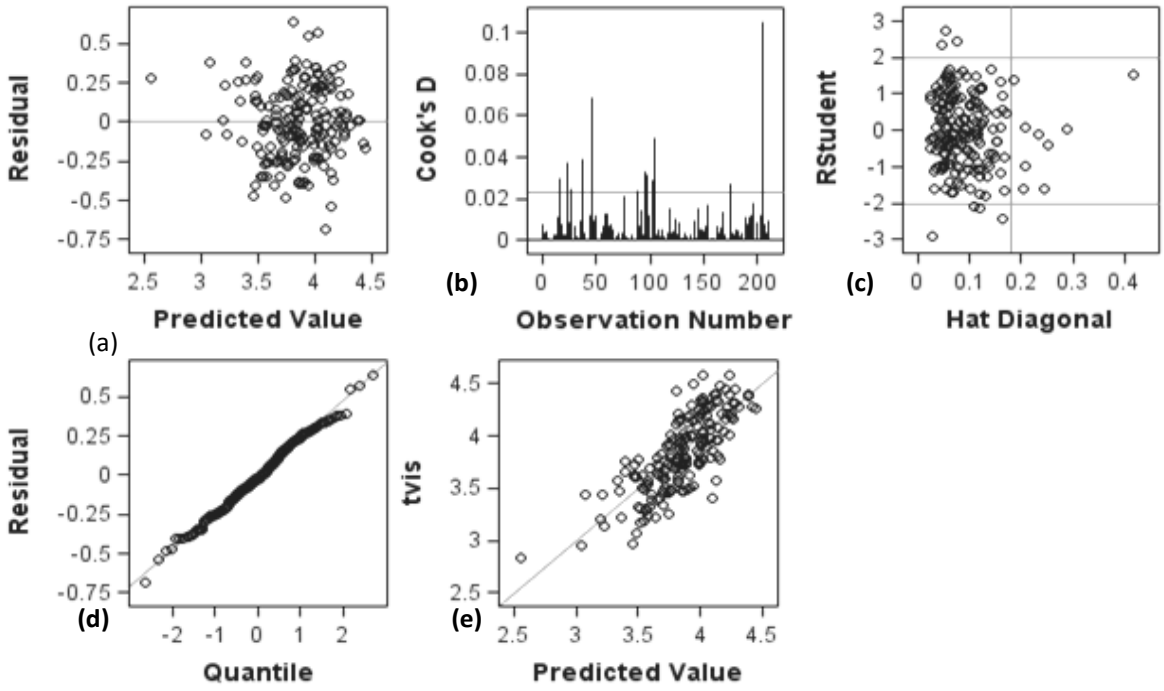


Figure 3.6 Model diagnostics with log viscosity as dependent variable **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed log viscosity versus predicted values**

3.5.3 Brightness

Multiple regression analysis results with brightness as the dependent variable against explanatory variables falling under anatomical, chemical and pulp property measurement kappa number are presented in Table 3.16. The P- value of the F- statistics is < 0.0001 it indicate the presence of a highly significant linear relationship between brightness and at least with one of the explanatory variables (anatomical and chemical measurements including the pulp property kappa number). The individual t-test values for each independent variable in Table 3.17 indicate that fibre lumen diameter (fld), kappa number (Kno), density (dey) and total hemicelluloses (ths) significantly affect brightness at 5% significance level. The regression model overall explains 43.52 % of the total variation. VIF values of fibre diameter is greater than 10 which shows the presence of a multicollinearity problem involving fibre

diameter with one or more other predictors. Residuals plots for model diagnosis are presented in Figures 3.7 (a) – (e). Raw residuals versus predicted values plot do not shown any systematic pattern, meaning the assumption of constant error variance, is not violated. The R-studentized residuals versus Hat diagonal elements plot none the less show the presence of some outlier and leverage observations, but however the cook's distance versus observation number plot indicates that these leverage and outlier observations are not influential since the cook's distance of each observation is below one. The residual versus normal quantile plot and the histogram of residual further indicate that the normality assumption of error component terms is not seriously violated.

Table 3.16 ANOVA with brightness as the dependent variable (E-dunnii data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	928.40754	61.89384	8.32	<.0001
Error	162	1204.65376	7.43613		
Corrected Total	177	2133.06131			
R-Square = 0.4352		Adj R-Square = 0.3830			

Table 3.17 Parameter estimates of independent variables for the brightness model (E-dunnii data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	34.05298	17.76302	1.92	0.0570	0
dbh	1	-0.06311	0.18692	-0.34	0.7361	2.65259
aht	1	0.08451	0.30821	0.27	0.7843	3.40325
htc	1	-0.27038	0.24463	-1.11	0.2707	4.31566
fd	1	1.53757	0.92490	1.66	0.0984	12.08371
cwt	1	0.16406	1.59472	0.10	0.9182	4.60374
fld	1	-1.31197	0.51191	-2.56	0.0113	5.74745
vp	1	0.04253	0.33964	0.13	0.9005	9.92957
kno	1	-2.08311	0.45095	-4.62	<.0001	1.63938
dey	1	27.10937	7.31109	3.71	0.0003	1.29420
glu	1	0.26805	0.18928	1.42	0.1587	2.45733
cel	1	0.07605	0.14411	0.53	0.5984	2.38867
sgr	1	0.33051	0.73689	0.45	0.6544	1.19419
tex	1	-0.07574	0.13663	-0.55	0.5801	2.29359
ths	1	-0.16993	0.07364	-2.31	0.0223	1.97100
tli	1	-0.31145	0.21397	-1.46	0.1474	2.21891

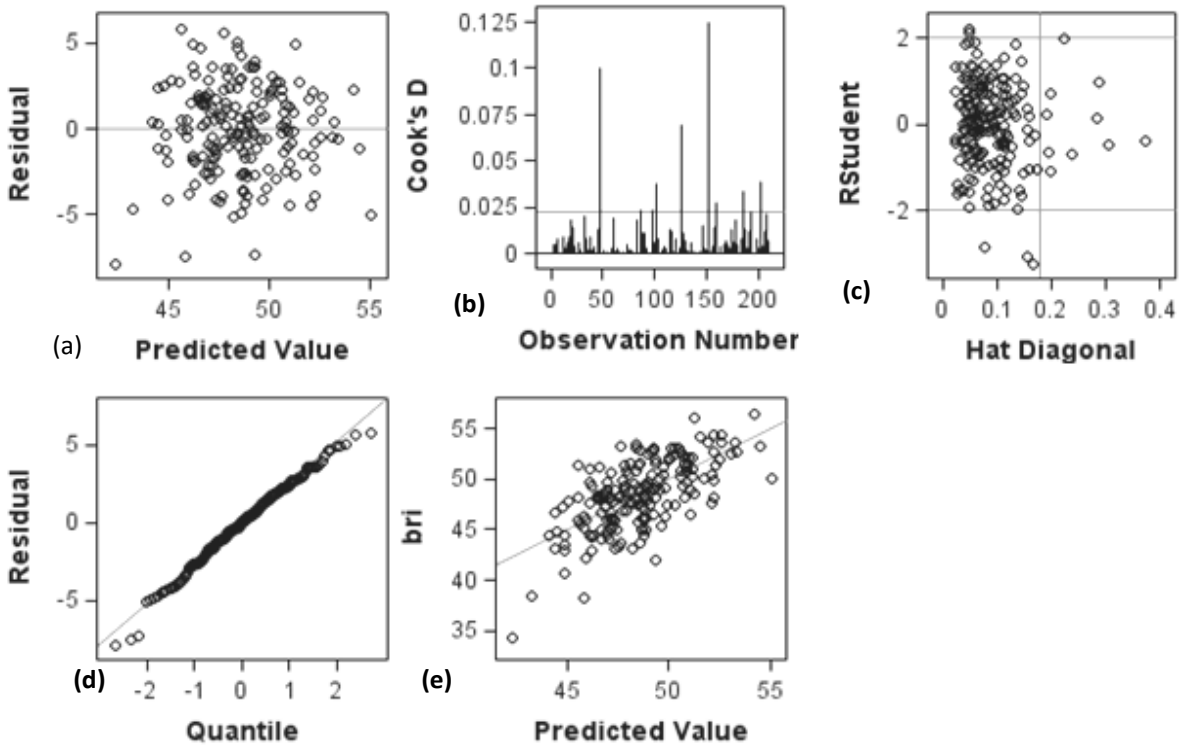


Figure 3.7 Model diagnostics with brightness as dependent variable **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed brightness versus predicted value**

3.5.4 Yield

Similar to the regression of viscosity and brightness a multiple regression model with yield as the dependent variable was performed. The P- value of F- statistics is < 0.0001 (see Table 3.18), this indicates highly significant linear relationship between yield and at least with one of the independent variables. It is noted that 59.06% of the total variation explained by the fitted model. The individual t-test values in Table 3.19 show that average height of a tree up to vessel percentage (vp), kappa number (Kno) and cellulose (cel) significantly affect yield at 5% significance level. The multicolinarity detecting measure namely variance inflation factor (VIF) value for vessel percentage (vp) is greater than 10 which indicates the presence of a multicolinarity problem.

Model diagnostic plots in Figure 3.8 (a) plots of raw residuals versus predicted values, do not show any systematic pattern, and hence the assumption of constant error variance is not violated. The test for outlier or leverage and influential observation is detected from a plot of R-student residuals versus Hat diagonal element and cook's distance versus observation number respectively (Figures 3.8 b, c). The first one shows the presence of some outlier and leverage observations but from the second plot the cook's distance of each observation is below one. Thus those leverage and outlier observations are not influential.

Table 3.18 ANOVA with yield as the dependent variable (E-dunnii data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	354.75172	23.65011	15.58	<.0001
Error	162	245.89222	1.51785		
Corrected Total	177	600.64394			
R-Square = 0.5906			Adj R-Square = 0.5527		

Table 3.19 Parameter estimates of independent variables for the yield model (E-dunnii data)

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	29.22635	8.02524	3.64	0.0004	0
dbh	1	0.11611	0.08445	1.37	0.1711	2.65259
aht	1	0.03906	0.13925	0.28	0.7794	3.40325
htc	1	-0.05821	0.11052	-0.53	0.5991	4.31566
fd	1	0.06764	0.41787	0.16	0.8716	12.08371
cwt	1	-0.86506	0.72048	-1.20	0.2316	4.60374
fld	1	-0.44914	0.23128	-1.94	0.0539	5.74745
vp	1	0.39956	0.15345	2.60	0.0101	9.92957
kno	1	0.56862	0.20374	2.79	0.0059	1.63938
dey	1	1.42385	3.30311	0.43	0.6670	1.29420
glu	1	0.09575	0.08551	1.12	0.2645	2.45733
cel	1	0.33118	0.06511	5.09	<.0001	2.38867
sgr	1	-0.36421	0.33292	-1.09	0.2756	1.19419
tex	1	-0.07798	0.06173	-1.26	0.2083	2.29359
ths	1	0.02726	0.03327	0.82	0.4139	1.97100
tli	1	-0.17428	0.09667	-1.80	0.0733	2.21891

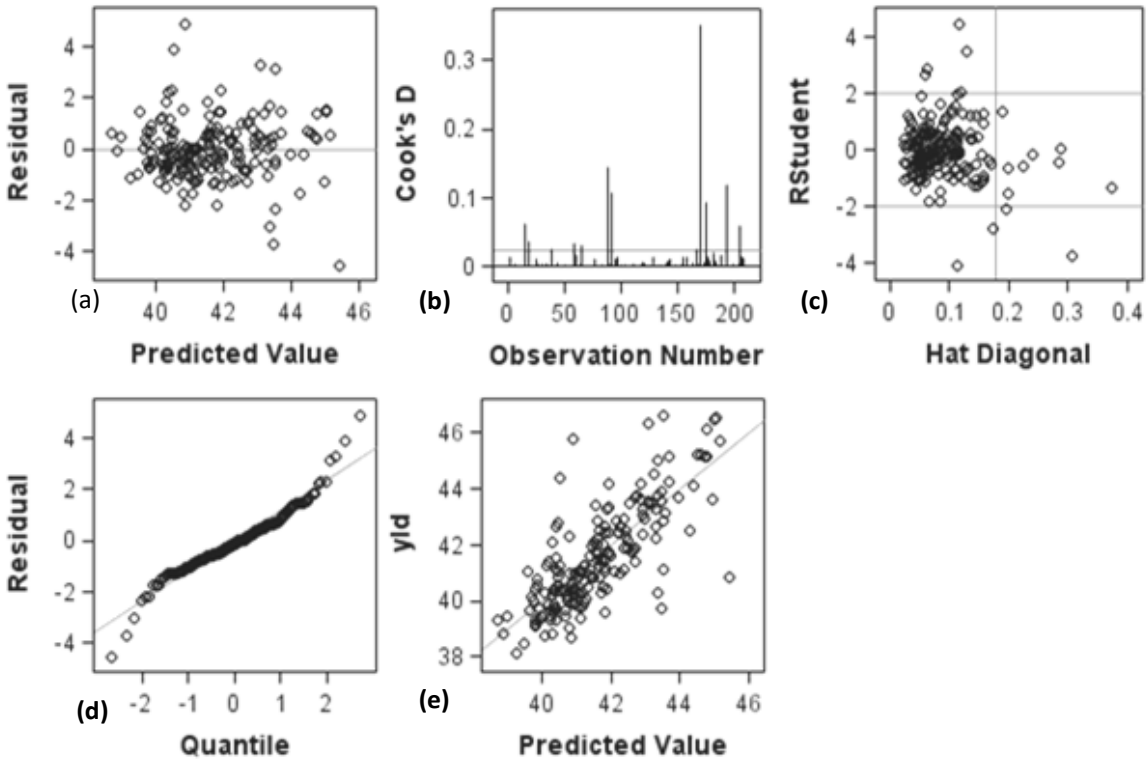


Figure 3.8 Model diagnostics with yield as dependent variable **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed yield versus predicted value**

3.6 Multicollinearity

Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. This phenomenon is expected wherever we are dealing with a regression with several or multiple independent variables. In this situation the regression coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole rather it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors may still indicate how well the entire bundle of predictors predict the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant in relation to others.

Multicollinearity therefore is problematic when one's purpose is explanation rather than mere prediction (Vaughan and Berry, 2005). One consequence to this is that the individual P-values can be misleading (a P- value can be high, even though the variable is unimportant).

The second problem is that the confidence intervals on individual regression coefficients may be very wide. The confidence intervals may even include zero, which means one cannot even be confident whether an increase in that independent value is associated with an increase, or a decrease, in the dependent variable. Because the confidence intervals are so wide, excluding a subject or observation (or adding a new one) can change the coefficients dramatically and may even change their signs.

3.6.1 Multicollinearity Diagnostics

We now consider the correlation matrix presented in Tables 2.5 and 2.6 in relation with the above discussion of multicollinearity and its diagnostics in the regression analysis. Multicollinearity is a matter of degree, not a matter of presence or absence. The higher the degree of multicollinearity, the greater the likelihood of the effects or consequences of multicollinearity. Several techniques have been proposed for detecting multicollinearity. The first one is through an examination of the off-diagonal elements of the correlation matrix and it is a very simple measure of multicollinearity. If the independent variables are nearly linearly dependent, then the off diagonal element $|r_{ij}|$ will be near unity implying a high correlation between X_i and X_j .

The second method for detecting multicollinearity is through the variance inflation factor (VIF). The diagonal elements of the inverse of the **X′X** matrix are very useful for detecting multicollinearity. The jth diagonal element of $C = (X' X)^{-1}$ matrix can be written as $C_{jj} = (1-R_j^2)^{-1}$, where R_j^2 is the coefficient of determination obtained when X_j is regressed on the remaining p-1 regressors. If X_j is nearly orthogonal to the remaining p-1 regressors, R_j^2 is small and C_{jj} is close to unity, while if X_j is nearly linearly dependent on some subset or all of the remaining regressors, R_j^2 is near unity and C_{jj} will be large. Since the variance of the jth regression coefficient is $C_{jj}\sigma^2$, we can view C_{jj} as the factor by which the variance of $\hat{\beta}_j$ is increased due to near linear dependences among the regressors. We call this the variance inflation factor or VIF given by

$$VIF_j = C_{jj} = (1-R_j^2)^{-1}$$

Marquardt (1970). The VIF for each term in the model measures the combined effect of the dependences among the regressors on the variance of that term. One or more large VIF indicate multicollinearity. Practical experience indicates that if any of the VIFs exceeds 10, it

is an indication that the associated regression coefficients are poorly estimated because of multicollinearity. This therefore justifies the reason for under taking multicolinearity remedial measures in any regression application problem.

The third method is that the determinant of $\mathbf{X^TX}$ can also be used as an index of multicollinearity. Since the $\mathbf{X^TX}$ matrix is in correlation form, the possible range of values of the determinant is $0 < |\mathbf{X^TX}| \leq 1$. If $|\mathbf{X^TX}| = 1$ the regressors are orthogonal, while if $|\mathbf{X^TX}| = 0$, there is an exact linear dependence among the regressors. The degree of the multicollinearity becomes more severe as $|\mathbf{X^TX}|$ approaches zero. While this measure of multicollinearity is easy to apply, it does not provide any information on the source of the multicollinearity.

The F statistic for significance of regression and the individual t statistics can sometimes indicate the presence of multicollinearity. Specifically, if the overall F statistic is significant but most or all individual t statistics are non significant, then this indicates the presence of multicollinearity. Another indicator of multicollinearity is when the signs or magnitude of the estimated regression coefficients are contrary to what was a prior expected.

First we note the presence of high correlation of up to 0.9908 (see Table 2.5 and 2.6) between wood anatomical measurements (average height, average height of tree up to diameter 7cm, average diameter at breast height, cell wall, fibre diameter and fibre lumen diameter) and chemical measurements (cellulose, total lignin and glucose). Secondly the regression models of the three dependent variables viscosity, brightness and yield shown in Tables 3.3, 3.8 and 3.10 respectively, indicate that some of the VIF values are greater than 10 and finally the third indicator about the change of the signs of the regression coefficients is also evidence. For example kappa number (kno) and density in the regression model of viscosity are contrary to prior expectation which points to the presences of multicolinearity problem in the three multiple linear regression models or analyses.

3.6.2 Remedies of Multicollinearity

The best solution to the multicollinearity problem is to try to avoid it by not including redundant independent variables in the regression model. If we can identify redundancy

among independent variables already in the model, several remedies can be used to attempt to lessen the influence of the multicollinearity.

One possible solution is to remove one or more of the highly correlated independent variables using stepwise or any other model selection method. This remedial measure has two important limitations. First, no direct information is obtained about the dropped predictor variables. Secondly, the magnitudes of the regression coefficients for the predictor variables remaining in the model may be affected by the correlated predictor variables not included in the model. The second remedial measure is to add more observations to the data used in building the regression model, so that the multicollinearity is possibly lessened. That is, sometimes the data we have collected makes it appear as though two or more independent variables are related when, in fact, no strong relationship exists. Collecting additional data may then lessen the (apparent) multicollinearity. The third remedial measure for multicollinerity that can be used with ordinary least squares is to form one or several composite indexes based on the highly correlated predictor variables, an index being a linear combination of the correlated predictor variables. The methodology of principal components provides composite indexes that are uncorrelated. Finally, there are estimation procedures that are modifications of the least squares estimation procedure. When multicollinerity exists, these procedures are capable of producing point estimates that are better than the least squares point estimates in the sense that they are closer to the true values of parameters. One such procedure is called ridge regression.

The presence of high correlation within the anatomical and chemical measurements as shown in section 2.3 indicates that these subset of variables carry the same information about the response variable. Because of this using a principal component analysis for variable reduction may be necessary.

3.6.3 Principal Component Analysis

Principal component analysis is basically or fundamentally a variable reduction procedure. It is useful when we obtain data for a number of variables and believe that there is a correlation among these variables (O'Rorke, Hatcher and Stepanski, 2005).

Principal component analysis of different hybrid data for strongly correlated set of variables namely (*i*) average diameter at breast height (dbh), average height (aht) and average height of a tree up to a diameter of 7cm (htc), (*ii*) highly correlated variables (cell wall thickens(cwt), fibre lumen diameter (fld) and vessel percentage(vp)) and chemical variables (cellulose(cel), total extractives(tex) and total lignin(tli)) was carried out with results tabulated in Tables 3.20, 3.21 and 3.22 respectively. An approximately equal value of rotated factors indicates that variable reduction using principal component analysis is not supportive. So the utility of another or alternative means of variables selection for example stepwise regression may be necessary. This method is discussed in the next section.

Table 3.20 Summary of Principal components analysis for average diameter at breast height, average height, and average height of a tree up to a diameter of 7cm (Different hybrid data)

Principal components analysis							
			Latent vectors (loadings)			Communalities	Rotated factors
	Latent roots	Variable	1	2	3	1	1
1	2.909	dbh	-0.56995	0.8138	0.11353	0.3248	-0.57
2	0.083	aht	-0.5793	-0.496	0.64687	0.3356	-0.5793
3	0.008	htc	-0.58272	-0.3029	-0.7541	0.3396	-0.5827

Table 3.21 Summary of Principal components analysis for cell wall thickens, fibre lumen diameter and vessel percentage (Different hybrid data)

Principal components analysis							
			Latent vectors (loadings)			Communalities	Rotated factors
	Latent roots	Variable	1	2	3	1	1
1	1.761	cwt	0.68755	0.31514	0.65419	0.4727	0.6875
2	1.104	fld	0.10628	-0.93489	0.33866	0.0113	0.1063
3	0.135	vp	0.71832	-0.16332	-0.67627	0.5160	0.7183

Table 3.22 Summary of Principal components analysis for cellulose, total extractives and total lignin (Different hybrid data)

Principal components analysis							
			Latent vectors (loadings)			Communalities	Rotated factors
	Latent roots	Variable	1	2	3	1	1
1	2.324	cel	0.58355	-0.4589	0.66999	0.3405	0.5836
2	0.398	tex	-0.58866	0.3293	0.73827	0.3465	-0.5887
3	0.278	tli	-0.55942	-0.82521	-0.07797	0.3129	-0.5594

3.7 Model selection with viscosity, brightness and yield as dependent variables using stepwise regression

From any set of p-1 predictors, 2^p-1 alternative models can be constructed. This calculation is based on the fact that each predictor can be either included or excluded from the model. The number of models grows rapidly with the number of predictors. Evaluating all of the possible alternatives can be very difficult. To simplify this task, a variety of automatic computer selection procedures have been developed and one of these is stepwise regression. Selection of variables that enter into the models is done through stepwise selection.

Model selection summary result for different hybrid data set using SAS for the log of viscosity model is shown in Table 3.25. The variables ultimately included in the multiple regression model are total hemicelluloses, kapp number, total lignin, glucose, vessel percentage and fibre lumen diameter. For the brightness model independent variables that were retained include average height, kapp number, fibre lumen diameter, total hemicelluloses, density, total lignin and total extractives and lastly for the yield model, average height of a tree up to a diameter of 7cm, kappa number, Sg ratio, total lignin and fibre lumen diameter were included as shown in Tables 3.28 and 3.31, respectively. Tables 3.24, 3.27 and 3.30 respectively show that except for fibre lumen diameter all selected variables have a significant effect. Model diagnostic plots in Figures 3.9 - 3.14 shows an improvement in the fitted model with none of the model assumption is violated. The same model selection method namely a stepwise regression was applied to the E-dunnii data set with log viscosity, brightness and yield as dependent variables. A summary model selection

details for each of these three variables are presented in Appendix B Tables B.1, B.2 and B.3. Model diagnostic information for selected and fitted models show that there is no linear model assumption violation (see Appendix B Figures B.1- B.3).

Table 3.23 Analysis of variance with log-viscosity as the dependent variable (Different hybrid reduced data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	5.92040	0.98673	26.24	<.0001
Error	105	3.94851	0.03760		
Corrected Total	111	9.86890			

Table 3.24 Parameter estimates with log- viscosity as the dependent variables (Different hybrid reduced data)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	8.18533	1.04151	2.32267	61.77	<.0001
fld	0.04483	0.02636	0.10876	2.89	0.0920
vp	0.05928	0.01897	0.36710	9.76	0.0023
kno	0.16055	0.03245	0.92049	24.48	<.0001
glu	-0.04015	0.01115	0.48758	12.97	0.0005
ths	-0.06236	0.00838	2.08414	55.42	<.0001
tli	-0.07473	0.01797	0.65042	17.30	<.0001

Table 3.25 Summary of stepwise selection for the log viscosity model (Different hybrid data)

Summary of Stepwise Selection								
Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ths	ths	1	0.4213	0.4213	38.8267	80.07	<.0001
2	kno	kno	2	0.0649	0.4862	24.3572	13.77	0.0003
3	tli	tli	3	0.0444	0.5306	15.0873	10.22	0.0018
4	glu	glu	4	0.0239	0.5545	11.0344	5.73	0.0184
5	vp	vp	5	0.0344	0.5889	4.3031	8.87	0.0036
6	fld	fld	6	0.0110	0.5999	3.5072	2.89	0.0920

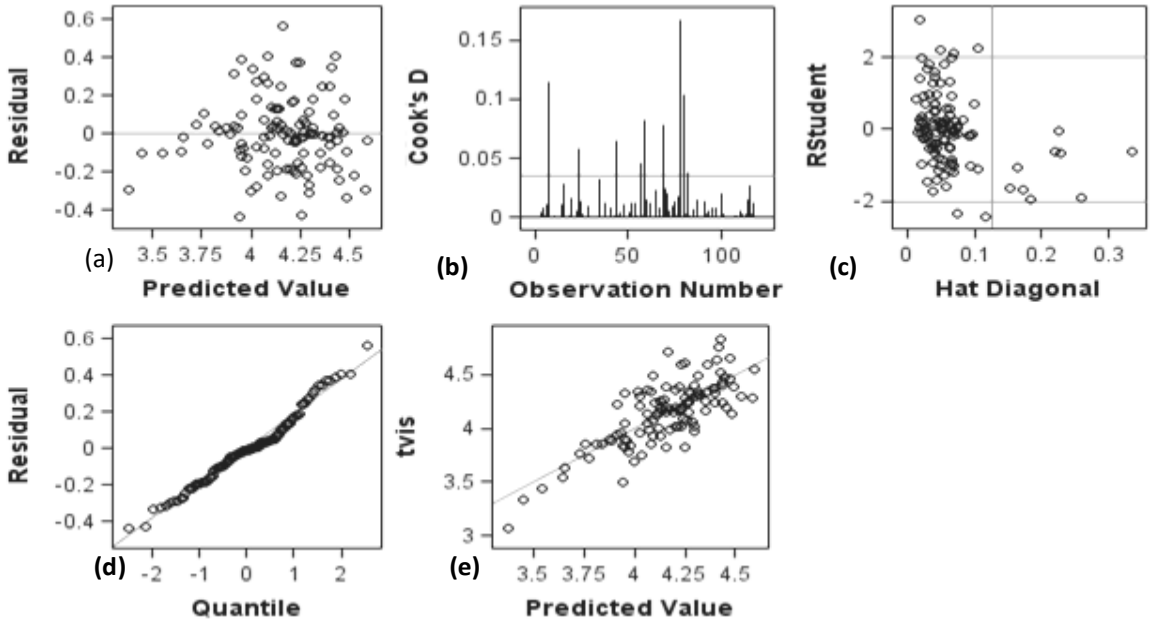


Figure 3.9 Model diagnostics with log viscosity as the dependent variable (different hybrid reduced data) **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed log viscosity versus predicted values**

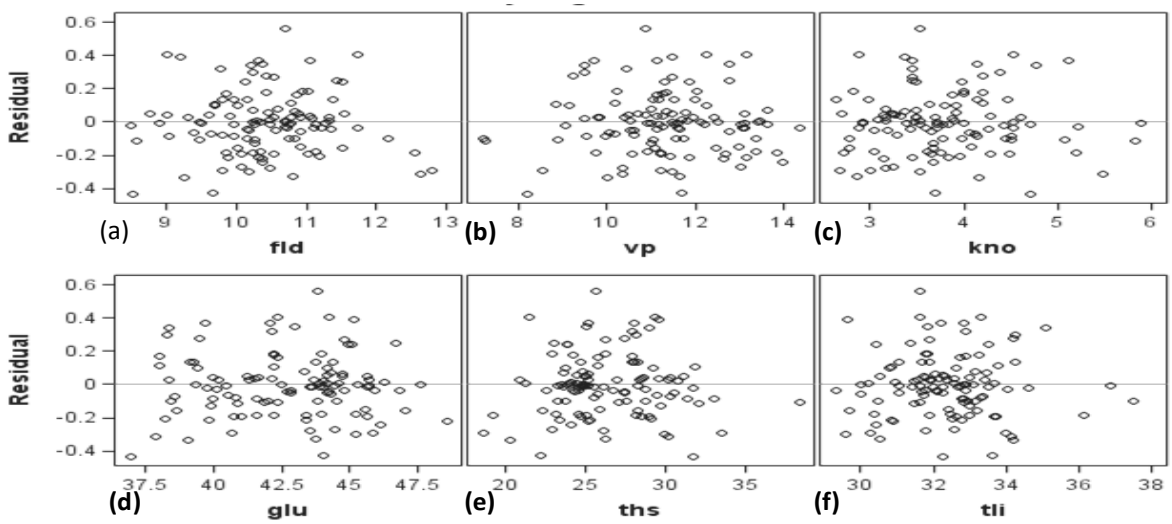


Figure 3.10 Model diagnostics of selected variables of log of viscosity model (different hybrid reduced data) **(a) fibre lumen diameter (b) vessel percentage (c) Kappa number (d) glucose (e) total hemicelluloses (f) total lignin**

Table 3.26 ANOVA with brightness as the dependent variable (Different hybrid reduced data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	2730.93909	390.13416	59.87	<.0001
Error	104	677.71186	6.51646		
Corrected Total	111	3408.65095			

Table 3.27 Parameter estimates with brightness as the dependent variables (Different hybrid reduced data)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	93.70327	10.78933	491.51032	75.43	<.0001
aht	0.69978	0.06736	703.37089	107.94	<.0001
fld	-1.90811	0.36494	178.14450	27.34	<.0001
kno	-3.11826	0.43644	332.65116	51.05	<.0001
dey	-20.91759	8.70472	37.62922	5.77	0.0180
tex	0.68680	0.22249	62.09784	9.53	0.0026
ths	-0.38283	0.10764	82.42372	12.65	0.0006
tli	-0.55722	0.23836	35.61225	5.46	0.0213

Table 3.28 Summary of stepwise selection for the brightness model (Different hybrid data)

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	aht		1	0.5857	0.5857	103.594	155.51	<.0001
2	kno		2	0.1348	0.7205	36.7659	52.55	<.0001
3	fld		3	0.0342	0.7547	21.3094	15.05	0.0002
4	ths		4	0.0198	0.7745	13.1900	9.40	0.0027
5	cwt		5	0.0091	0.7836	10.5192	4.48	0.0366
6	fd		6	0.0065	0.7901	9.2231	3.23	0.0753
7	tli		7	0.0068	0.7968	7.7579	3.47	0.0652
8	tex		8	0.0051	0.8019	7.1603	2.64	0.1069
9	dey		9	0.0046	0.8066	6.7896	2.45	0.1208
10		cwt	8	0.0034	0.8031	6.5394	1.81	0.1819
11		fd	7	0.0020	0.8012	5.5459	1.03	0.3123

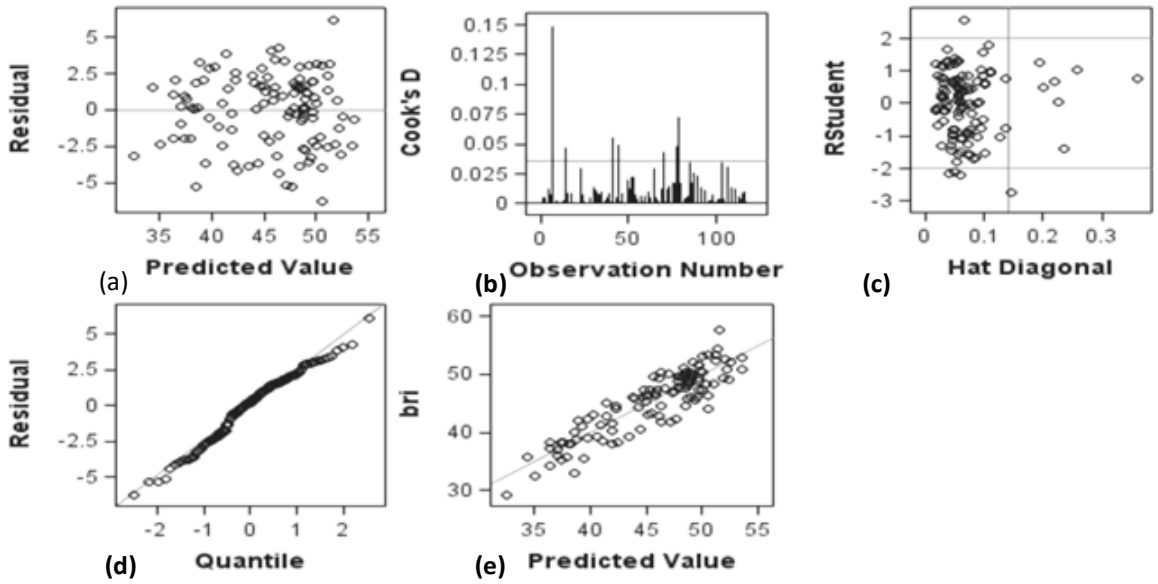


Figure 3.11 Model diagnostics with brightness as the dependent variable (different hybrid reduced data) **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed brightness versus predicted values**

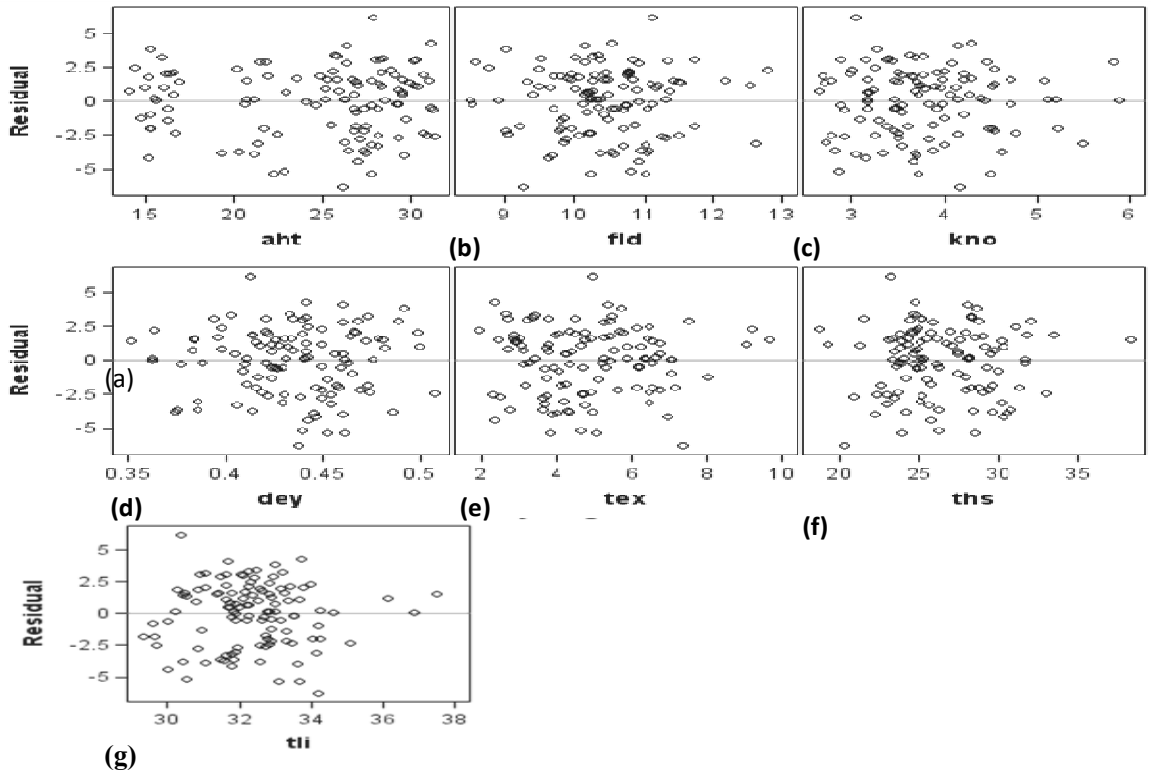


Figure 3.12 Model diagnostics of selected variables of Brightness model (different hybrid reduced data) **(a) average height (b) fibre lumen diameter (c) Kappa number (d) density (e) total extractives (f) total hemicelluloses (g) total lignin**

Table 3.29 ANOVA with yield as the dependent variable (Different hybrid reduced data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	551.24699	110.24940	84.55	<.0001
Error	106	138.22463	1.30401		
Corrected Total	111	689.47162			

Table 3.30 Parameter estimates with yield as the dependent variables (Different hybrid reduced data)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	45.43916	3.10794	278.73681	213.75	<.0001
htc	0.32450	0.02150	297.15174	227.88	<.0001
fld	0.27108	0.14914	4.30838	3.30	0.0719
kno	-0.41812	0.20029	5.68270	4.36	0.0392
sgr	1.70492	0.50690	14.75155	11.31	0.0011
tli	-0.30284	0.08527	16.44914	12.61	0.0006

Table 3.31 Summary of stepwise selection for the yield model (Different hybrid data)

Summary of Stepwise Selection of yield								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	htc		1	0.6791	0.6791	58.1307	232.83	<.0001
2	kno		2	0.0692	0.7484	24.2804	29.99	<.0001
3	sgr		3	0.0243	0.7727	13.7025	11.54	0.0010
4	tli		4	0.0206	0.7933	5.0359	10.66	0.0015
5	fld		5	0.0062	0.7995	3.8005	3.30	0.0719

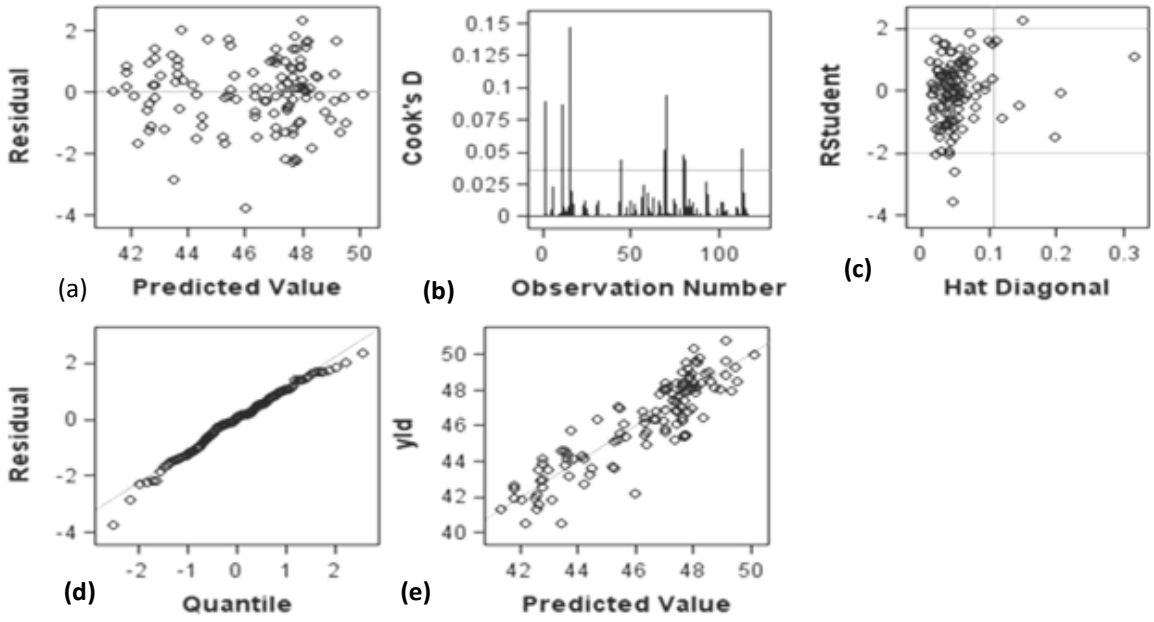


Figure 3.13 Model diagnostics with yield as the dependent variable (different hybrid reduced data) **(a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed yield versus predicted values**

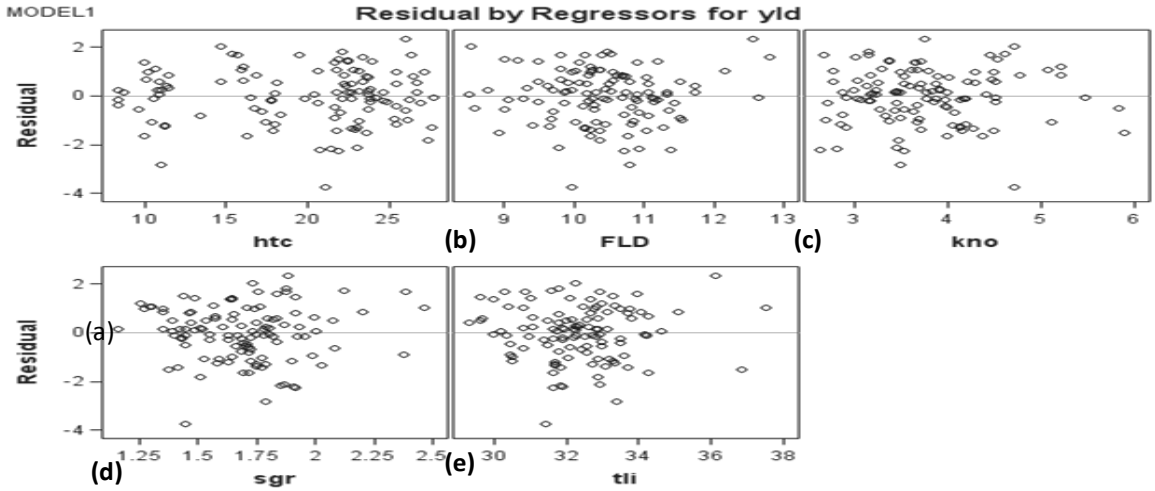


Figure 3.14 Model diagnostics of selected variables of yield model (different hybrid reduced data) **(a) average diameter at breast height (b) fibre lumen diameter (c) Kappa number (d) SG ratio (e) total lignin**

3.8 Multiple Comparisons

In Chapter 2 the presence of a mean difference and distributional variability of viscosity, brightness and yield of trees with different age group, location, site quality was evident. To now further, understanding of differences in age, site quality, location and hybrid type on viscosity, brightness and yield an application of a mean comparison under these grouping factors is important.

When comparing more than two means, an ANOVA F-test only tells us whether the mean are significantly different from each other, but it does not tell us which means differ from which other means. Multiple comparison procedures also called mean separation tests, give us more detailed information about the differences among the group means. The main goal in multiple comparisons is to compare the average effects of three or more treatments or groups of subjects to decide which treatments or groups are better/different, which ones are worse, and by how much, while controlling the probability of making an incorrect decision.

There are a number of multiple comparison procedures that are available. These include least significance difference (LSD) method, Duncan's multiple range test, Tukey method, Dunnett method and many more. The selection of the appropriate multiple comparison method depends on the desired inference. For more details on methods of multiple comparison, one may refer to Gomez and Gomez(1984), Montogomery (1990), Hsu(1991), Dean and Voss(1999), Milliken and Johnson (2002).

3.8.1 A mean comparison on different hybrid data

A mean comparison of viscosity, brightness and yield measurements for different age group of trees for the different hybrid data were performed using Duncan's multiple range test where performed. ANOVA sub-tables in Table 3.32 give P- values of the F- statistics and they indicate the presence of significant difference (P-value <.0001) in mean viscosity, brightness and yield of trees over different age groups. Duncan grouping summary for each variable in Table 3.33 indicate that, mean viscosity and brightness for a tree of age category 8 and 9, 7 and 5 are not significantly different. The mean yield of trees in age groups 8 and 9, 9

and 7 are not significantly different but trees of age group 5 mean yield is different from all the other age groups.

Table 3.32 Mean comparison ANOVA of viscosity, brightness and yield at different age categories (Different hybrid data)

Variable	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Viscosity	Model	3	9930.857	3310.286	11.44	<.0001
	Error	113	32704.2	289.4177		
	Corrected Total	116	42635.05			
Brightness	Model	3	1338.52	446.1732	22.62	<.0001
	Error	113	2228.689	19.72291		
	Corrected Total	116	3567.208			
Yield	Model	3	352.8535	117.6178	37.15	<.0001
	Error	113	357.7596	3.166014		
	Corrected Total	116	710.6131			

Table 3.33 Duncan grouping of means for viscosity, brightness and yield over different age categories (Different hybrid data)

Viscosity				Brightness				Yield				
Duncan Grouping	Mean	N	age	Duncan Grouping	Mean	N	age	Duncan Grouping		Mean	N	age
A	78.856	33	8	A	48.895	28	9		A	47.5736	33	8
A	71.566	28	9	A	48.164	33	8	B	A	47.2464	28	9
								B		46.3707	28	7
B	59.701	28	5	B	42.779	28	7					
B	56.279	28	7	B	40.957	28	5		C	43.1562	28	5

NB : Means with the same letter are not significantly different

Similarly Duncan's multiple range test of mean viscosity, brightness and yield of trees over different locations were performed and results tabulated in Table 3.34. The P- value of the F-statistics is < 0.0001 for all the three variables as it seen in the mean comparison ANOVA table. This is evidence of a statistically significant difference in mean viscosity, brightness and yield over some location. A Duncan grouping of trees over different locations given in

Table 3.35 shows all the possible grouping of means by location based for mean viscosity, brightness and yield. The tables show that locations can be grouped into four, five and three classes based on mean viscosity, brightness and yield respectively. Unlike groupings based on mean viscosity and brightness grouping based on yield show clear distinct groups.

Table 3.34 Mean comparison ANOVA of viscosity, brightness and yield at different locations (Different hybrid data)

Variable	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Viscosity	Model	7	10994.46	1570.637	5.41	<.0001
	Error	109	31640.59	290.2807		
	Corrected Total	116	42635.05			
Brightness	Model	7	2517.754	359.6791	37.36	<.0001
	Error	109	1049.455	9.628026		
	Corrected Total	116	3567.208			
Yield	Model	7	506.5476	72.36394	38.65	<.0001
	Error	109	204.0655	1.872161		
	Corrected Total	116	710.6131			

Table 3.35 Duncan grouping of means for viscosity, brightness and yield over different locations (Different hybrid data)

Viscosity							Brightness					Yield			
Duncan Grouping				Mean	N	location	Duncan Grouping		Mean	N	location	Duncan Grouping	Mean	N	location
		A		82.28	17	P/Ridge C13		A	51.3	16	Terra A01	A	48.5441	11	KT E10
B		A		75.22	16	Terra A01	B	A	49.953	18	Salpine E05	A	48.0036	18	Salpine E05
B		A	C	71.78	10	P/Ridge B1	B	C	48.38	5	KT G09	A	47.6972	16	Terra A01
B		A	C	71.44	18	Salpine E05	B	C	47.936	11	KT E10	A	47.4574	17	P/Ridge C13
B	D	A	C	67.07	5	KT G09	D	C	46.992	10	P/Ridge B1	B	45.8833	10	P/Ridge B1
B	D		C	60.495	11	KT E10	D		45.212	17	P/Ridge C13	B	45.679	5	KT G09
	D		C	58.099	23	P/Ridge C10		E	39.441	17	P/Ridge D13	B	44.9644	17	P/Ridge D13
	D			53.55	17	P/Ridge D13		E	39.344	23	P/Ridge C10	C	42.6078	23	P/Ridge C10

NB : Means with the same letter are not significantly different

A multiple range test for mean comparison of viscosity, brightness and yield over the two site quality was performed and the ANOVA results presented in ANOVA Table 3.36. The results indicate that mean viscosity between the two site quality are not significantly different but mean brightness and yield over site quality I is statistically significantly different from site quality II. This significant difference between the two site quality is also supported by a Duncan grouping as shown in Table 3.37.

Table 3.36 Mean comparison ANOVA of viscosity, brightness and yield for different site quality (Different hybrid data)

Variable	Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Viscosity	Model	1	628.8919	628.8919	1.72	0.1921
	Error	115	42006.16	365.271		
	Corrected Total	116	42635.05			
Brightness	Model	1	1734.936	1734.936	108.89	<.0001
	Error	115	1832.272	15.9328		
	Corrected Total	116	3567.208			
Yield	Model	1	235.3305	235.3305	56.94	<.0001
	Error	115	475.2826	4.132892		
	Corrected Total	116	710.6131			

Table 3.37 Duncan grouping of means for viscosity, brightness and yield of the two site quality (Different hybrid data)

Viscosity				Brightness				Yield			
Duncan Grouping	Mean	N	Site quality	Duncan Grouping	Mean	N	Site quality	Duncan Grouping	Mean	N	Site quality
A	69.8	50	I	A	49.8	50	I	A	47.8	50	I
A	65.1	67	II	B	42.0	67	II	B	44.9	67	II

NB : Means with the same letter are not significantly different

Lastly a Duncan's multiple range test of mean viscosity, brightness and yield of trees over different hybrid types are presented in Tables 3.38 -3.43. The P-values of the ANOVA F-statistics are less than < 0.0001 as shown in Tables 3.38, 3.40 and 3.42. This shows an overall significant difference in mean of viscosity, brightness and yield between some hybrid types. Tables 3.39, 3.41 and 3.43 indicate that means of the same letter are not significantly different. But these comparisons present a problem because of the number of observations for most hybrid types is below five which may affect the reliability of the comparisons result.

Table 3.38 Mean comparison ANOVA of viscosity for different hybrid type (Different hybrid data)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	22229.63226	1111.48161	5.23	<.0001
Error	96	20405.42135	212.55647		
Corrected Total	116	42635.05361			

Table 3.39 Duncan grouping of means for viscosity over different hybrid type (Different hybrid data)

Means with the same letter are not significantly different.						
Duncan Grouping				Mean	N	Hybrid type
		A		92.95	8	ZGU CN
B		A		85.60	3	GU x U
B		A	C	78.40	1	E. uro x Gra/Ter
B		A	C	77.79	36	GU
B		A	C	77.50	1	GU x GC
B	D	A	C	71.25	1	G x GT
B	D	A	C	69.08	6	E. grandis
B	D	A	C	66.17	12	GP
B	D	A	C	64.48	2	G x GU
B	D	A	C	64.10	1	E. urophylla
B	D	A	C	57.78	2	E. uro x E. ter
B	D	A	C	57.73	22	UG
B	D	A	C	57.25	1	GU x ((GP) + (GXGT))
B	D		C	55.95	1	GU x GT
B	D		C	52.14	4	GU x ((GP)x E. ter))
B	D		C	50.53	2	GU x ((GP) + (GxGT))
B	D		C	49.80	1	GC
	D		C	47.28	5	GU x GP
	D		C	41.72	5	GU x (ET+GP)
	D		C	41.55	1	GU x((GP) + (GxGT))
	D			39.53	2	GU x (G x GU)

Table 3.40 Mean comparison ANOVA of brightness for different hybrid types (Different hybrid data)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	2012.667171	100.633359	6.21	<.0001
Error	96	1554.541321	16.193139		
Corrected Total	116	3567.208492			

Table 3.41 Duncan grouping of means for brightness over different hybrid type (Different hybrid data)

Means with the same letter are not significantly different.							
Duncan Grouping					Mean	N	Hybrid type
			A		52.100	1	G x GT
	B		A		50.550	1	GU x GT
	B		A		50.325	2	G x GU
	B		A	C	48.900	1	E. urophylla
	B		A	C	48.703	36	GU
	B		A	C	48.183	12	GP
	B		A	C	47.900	2	GU x (G x GU)
	B		A	C	47.600	1	GU x ((GP) + (GXGT))
	B		A	C	47.212	8	ZGU CN
	B	D	A	C	45.192	6	E. grandis
	B	D	A	C	45.133	3	GU x U
E	B	D	A	C	44.390	5	GU x (ET+GP)
E	B	D	A	C	44.300	1	GU x GC
E	B	D	A	C	43.360	5	GU x GP
E	B	D	A	C	42.850	1	GC
E	B	D		C	41.266	22	UG
E		D		C	39.500	1	GU x((GP) + (GxGT))
E		D			37.200	1	E. uro x Gra/Ter
E		D			35.400	2	E. uro x E. ter
E					35.112	4	GU x ((GP)x E. ter))
E					35.000	2	GU x ((GP) + (GxGT))

Table 3.42 Mean comparison ANOVA of yield for different hybrid type (Different hybrid data)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	430.2298428	21.5114921	7.37	<.0001
Error	96	280.3832391	2.9206587		
Corrected Total	116	710.6130819			

Table 3.43 Duncan grouping of means for yield over different hybrid type (Different hybrid data)

Means with the same letter are not significantly different.							
Duncan Grouping					Mean	N	Hybrid type
			A		48.690	1	GU x ((GP) + (GXGT))
			A		48.405	1	GU x GT
	B		A		48.080	12	GP
	B		A		47.738	36	GU
	B		A	C	47.373	2	G x GU
	B		A	C	47.330	2	GU x (G x GU)
	B		A	C	47.179	5	GU x (ET+GP)
	B	D	A	C	46.870	1	GU x((GP) + (GxGT))
	B	D	A	C	46.505	1	G x GT
	B	D	A	C	46.430	1	E. urophylla
E	B	D	A	C	46.103	3	GU x U
E	B	D	A	C	45.988	2	GU x ((GP) + (GxGT))
E	B	D	A	C	45.871	6	E. grandis
E	B	D	A	C	45.652	5	GU x GP
E	B	D	A	C	45.548	8	ZGU CN
E	B	D	A	C	44.989	4	GU x ((GP)x E. ter))
E	B	D	F	C	43.870	1	GC
E		D	F	C	43.300	22	UG
E		D	F		42.735	1	GU x GC
E			F		42.009	2	E. uro x E. ter
			F		40.540	1	E. uro x Gra/Ter

3.9 Summary

Application of multiple regressions to each pulp property measurement viscosity, brightness and yield shows the presence of a highly significant linear relation with wood anatomic and chemical measurements. But the total variation explained by the fitted models varies from one data set to the other and also within pulp properties measurement of each data set. The explained variation ranges from a lowest total variation explained of 43.52% for brightness of E-dunnii data set to a highest total variation explained of 81.27% for yield for the different hybrid data set. This proportion variation explained difference may be because of the presence of additional variability of age, location, site quality and hybrid type in the different hybrid data set that is not accounted for by the models. This is also further supported by a

significant effect of these categorical variables on the pulp viscosity, brightness and yield as shown in the multiple comparisons tests.

Model diagnosis results further showed that there were no serious violation of model assumptions, except a non constant error variance for viscosity in the fitted model of both data sets and a multicollinearity problem to all the fitted regression models. A natural log transformation of viscosity and a stepwise regression was applied to address the problem of non constant error variance and multicollinearity respectively. Model selection in terms of variables included in the viscosity, brightness and yield models results differ in terms of the number and type of variables included in the fitted model. Kappa number and total lignin are the only variables included in all the final fitted models.

In this chapter multiple regressions of viscosity, brightness and yield were under taken individually without consideration of the presence of a significant correlation between these three variables. An attempt to model the three dependent variables jointly together is addressed in the next chapter using multivariate analysis methods.

Chapter 4

Multivariate linear regression

Multivariate regression procedures take into account the correlation among the dependent variables which is ignored by univarate analysis and this allows construction of simultaneous confidence intervals (Kim and Timm 2007). In general the multivariate regression model is used to explain the relationship between q dependent variables $y_1, y_2, y_3, \cdots, y_q$ and p independent variables $x_1, x_2, x_3, \cdots, x_p$ (Johnson, 2002). The p-dimensional multivariate linear regression model is

$$\mathrm{E}(\mathbf{Y}) = \mathbf{X\beta} \text{ or } \mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

where

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1q} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2q} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & y_{n3} & \cdots & y_{nq} \end{bmatrix} = \left[\boldsymbol{Y_1} \vdots \boldsymbol{Y_2} \vdots \boldsymbol{Y_3} \vdots \boldsymbol{Y_j} \vdots \ \ldots \ \vdots \boldsymbol{Y_q} \right]$$

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1p} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0q} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1q} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq} \end{bmatrix} = \left[\boldsymbol{\beta_1} \vdots \boldsymbol{\beta_2} \vdots \cdots \vdots \boldsymbol{\beta_q} \right]$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} & \cdots & \varepsilon_{1q} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} & \cdots & \varepsilon_{2q} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \varepsilon_{n3} & \cdots & \varepsilon_{nq} \end{bmatrix} = \left[\varepsilon_1 \vdots \varepsilon_2 \vdots \cdots \vdots \varepsilon_{pq}\right] = \begin{bmatrix} \varepsilon'_1 \\ \cdots \\ \varepsilon'_2 \\ \cdots \\ \vdots \\ \cdots \\ \varepsilon'_q \end{bmatrix}$$

where E(ε) = 0 and Var(ε) = Σ.

The p observations on the i[th] trial have covariance matrix $\sum = [(\sigma_{ij})]$ but observations from different observation unit (tree) are uncorrelated and also in this model the rows of **Y** and **ε** are assumed to be distributed as multivariate normal.

4.1 Parameter estimation

The matrix of residual sum of squares and cross products is $\Omega = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$; differentiating Ω with respect to β, yields $2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = 0$ and thus

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,,$$

provided that n > p or, more exactly provided that $\mathbf{X}'\mathbf{X}$ is non-singular. Differentiation of Ω also minimizes both the determinant and the trace of Ω means. This means that the p individual $\widehat{\boldsymbol{\beta}}_\mathbf{i}$ in $\widehat{\boldsymbol{\beta}}$ are identical to those which would be obtained by separate multiple regression of each $\mathbf{y}_i$ on the dependent variables $X_1, X_2, X_3 \ldots X_p$. Collecting those univariate least squares estimates, $\hat{\beta}_i$

$$\widehat{\boldsymbol{\beta}}_\mathbf{i} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_\mathbf{i} \qquad (4.1)$$

We can obtain $\hat{\beta} = \left[\hat{\beta}_1 \vdots \hat{\beta}_2 \vdots \cdots \vdots \hat{\beta}_q\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left(\mathbf{Y_1} \vdots \mathbf{Y_2} \vdots \mathbf{Y_3} \vdots \mathbf{Y_4} \vdots \ \ldots \ \vdots \mathbf{Y_q}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Under normality assumption of **ε** these least squares estimators are also maximum likelihood estimators of $\boldsymbol{\beta}$.

So we have matrices of predicted values $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ and we have a resulting matrices of residuals

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$$

and the variance $\sum$ is estimated by $\widehat{\sum}$.

$$\widehat{\sum} = \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{n - p - 1}$$

Multivariate linear regression coefficients β from experimental data is used to evaluate the marginal or partial effect of a predictor to the dependent variable given the other predictor variables in the model. The least squares estimation $\hat{\beta} = [\hat{\beta}_1 \vdots \hat{\beta}_2 \vdots \cdots \vdots \hat{\beta}_q]$ determined under the multivariate regression model with full rank **X** have the following properties.

i. $\mathrm{E}(\widehat{\boldsymbol{\beta}}_i) = \boldsymbol{\beta}_i$ i. e, $\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

ii. Cov $(\widehat{\boldsymbol{\beta}}_i, \widehat{\boldsymbol{\beta}}) = \delta_{ij}(\mathbf{X}'\mathbf{X})^{-1}$ *i, j* = 1,2,…,q

iii. $\mathrm{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ and $\mathrm{E}\left(\frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-p-1}\right) = \sum$

iv. $\mathrm{Cov}\left(\hat{\boldsymbol{\varepsilon}}, \widehat{\boldsymbol{\beta}}\right) = \mathbf{0}$.

v. $\mathrm{Cov}(\widehat{\sum}, \widehat{\boldsymbol{\beta}}) = 0$

Further, with full rank **X** and normally distributed errors $\boldsymbol{\varepsilon}$, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \sum^*)$. where the elements of $\sum^*$ are as given in *(ii)*.

4.2 Multivariate test statistics

Other than the likelihood ratio test other tests have been proposed for testing the multivariate regression parameter **β**. Most computer package programs routinely calculate four multivariate test statistics, namely Wilks's Lambda, Pilai's trace, Hoteling-Lawley trace and Roy's greatest root (Verbeke, 2004; Johnson and Wichern, 2002,).

Let **H** denote the sums of squares and cross products matrix, and let **E** denote the error sums of squares and cross products matrix. The above four statistics can be defined in terms of **E** and **H** directly, or in terms of the nonzero eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_s$ of $\mathbf{HE^{-1}}$, where s = min (p, r-q) where

$$\mathbf{E} = \left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \qquad (4.2)$$

$$\mathbf{H} = \left(\mathbf{L}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'\right)^{-1}\left(\mathbf{L}\widehat{\boldsymbol{\beta}}\right) \qquad (4.3)$$

where **L** is a (p+1) × q matrix with rank(L) = q. The four multivariate test statistics are defined as follows:

The first statistic called the Wilk's Λ was the first MANOVA test statistic to be developed and one of the most important for several multivariate procedures in addition to MANOVA.

$$\text{Wilk's lambda } (\Lambda) = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \prod_{i=1}^{s}\frac{1}{1+\lambda_i}. \qquad (4.4)$$

The second statistic is the Pillai's trace. Some statisticians consider it to be the most powerful and most robust of the four statistics. Its formula is given by

$$\text{Pillai's trace } = \text{tr}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_{i=1}^{s}\frac{\lambda_i}{1+\lambda_i}. \qquad (4.5)$$

The third test statistic is the Hotelling-Lawley's trace.

$$\text{Hotelling} - \text{Lawley trace} = \text{tr}[\mathbf{HE}^{-1}] = \sum_{i=1}^{s}\lambda_i. \qquad (4.6)$$

The fourth and last statistic is the Roy's largest root. This gives an upper bound for the F statistic.

$$\text{Roy's greatest root} = \frac{\lambda_1}{1+\lambda_1} \qquad (4.7)$$

where λ_1 is the largest or dominant eigenvalue of $\mathbf{HE^{-1}}$.

4.3 Application of multivariate regression analysis

The application of multivariate regression analysis for the different hybrid data set using SAS PROC GLM is presented in Tables 4.2.1-4.2.4. The four multivariate test results for each independent variables indicate that fibre lumen diameter (fld), kappa number (kno), total hemicelluloses (ths) and total lignin simultaneously have significant effect on pulp properties viscosity, brightness and yield at a time at 5% significance level. Other anatomic and chemical measurements have no concurrent significant effect.

Table 4.2.1 Multivariate Analysis of Variance of fibre lumen diameter (Different hybrid data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall fibre lumen diameter Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.77873974	8.90	3	94	<.0001
Pillai's Trace	0.22126026	8.90	3	94	<.0001
Hotelling-Lawley Trace	0.28412607	8.90	3	94	<.0001
Roy's Greatest Root	0.28412607	8.90	3	94	<.0001

Table 4.2.2 Multivariate Analysis of Variance of kappa number (Different hybrid data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall kappa number Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.64990194	16.88	3	94	<.0001
Pillai's Trace	0.35009806	16.88	3	94	<.0001
Hotelling-Lawley Trace	0.53869367	16.88	3	94	<.0001
Roy's Greatest Root	0.53869367	16.88	3	94	<.0001

Table 4.2.3 Multivariate Analysis of Variance of total hemicelluloses (Different hybrid data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall total hemicelluloses Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.63505085	18.01	3	94	<.0001
Pillai's Trace	0.36494915	18.01	3	94	<.0001
Hotelling-Lawley Trace	0.57467704	18.01	3	94	<.0001
Roy's Greatest Root	0.57467704	18.01	3	94	<.0001

Table 4.2.4 Multivariate Analysis of Variance of total lignin(Different hybrid data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall total lignin Effect					
	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.83850441	6.03	3	94	0.0008
Pillai's Trace	0.16149559	6.03	3	94	0.0008
Hotelling-Lawley Trace	0.19259957	6.03	3	94	0.0008
Roy's Greatest Root	0.19259957	6.03	3	94	0.0008

Similar application of multivariate regression to the E-dunnii data set was undertaken and the accompanying results are shown in Tables 4.3.1-4.3.10. The results show that average diameter at breast height (dbh), fibre diameter (fd), cell wall thickness (cwt), fibre lumen diameter (fld), vessel percentage (vp), kappa number (kno), density (dey), cellulose (cel), total hemicelluloses (ths) and total lignin (tli) have concurrent significant effect at 5% significance level on viscosity, brightness and yield. But from anatomic measurements, average height and average height of a tree up diameter of 7cm and from chemical measurements glucose, Sg ratio and total extractives are no joint significant effect on viscosity, brightness and yield.

Table 4.3.1 Multivariate Analysis of Variance of average diameter at breast height (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall average diameter at breast height Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.92041464	4.55	3	158	0.0043
Pillai's Trace	0.07958536	4.55	3	158	0.0043
Hotelling-Lawley Trace	0.08646685	4.55	3	158	0.0043
Roy's Greatest Root	0.08646685	4.55	3	158	0.0043

Table 4.3.2 Multivariate Analysis of Variance of fibre diameter (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall fibre diameter Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.90797047	5.34	3	158	0.0016
Pillai's Trace	0.09202953	5.34	3	158	0.0016
Hotelling-Lawley Trace	0.10135740	5.34	3	158	0.0016
Roy's Greatest Root	0.10135740	5.34	3	158	0.0016

Table 4.3.3 Multivariate Analysis of Variance of cell wall thickness (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall cell wall thickness Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94119375	3.29	3	158	0.0222
Pillai's Trace	0.05880625	3.29	3	158	0.0222
Hotelling-Lawley Trace	0.06248049	3.29	3	158	0.0222
Roy's Greatest Root	0.06248049	3.29	3	158	0.0222

Table 4.3.4 Multivariate Analysis of Variance of fibre lumen diameter (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall fibre lumen diameter Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.76008703	16.62	3	158	<.0001
Pillai's Trace	0.23991297	16.62	3	158	<.0001
Hotelling-Lawley Trace	0.31563882	16.62	3	158	<.0001
Roy's Greatest Root	0.31563882	16.62	3	158	<.0001

Table 4.3.5 Multivariate Analysis of Variance of vessel percentage (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall vessel percentage Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.94664370	2.97	3	158	0.0337
Pillai's Trace	0.05335630	2.97	3	158	0.0337
Hotelling-Lawley Trace	0.05636365	2.97	3	158	0.0337
Roy's Greatest Root	0.05636365	2.97	3	158	0.0337

Table 4.3.6 Multivariate Analysis of Variance of kappa number (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall kappa number Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.66810783	26.16	3	158	<.0001
Pillai's Trace	0.33189217	26.16	3	158	<.0001
Hotelling-Lawley Trace	0.49676438	26.16	3	158	<.0001
Roy's Greatest Root	0.49676438	26.16	3	158	<.0001

Table 4.3.7 Multivariate Analysis of Variance of density (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall density Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.91932931	4.62	3	158	0.0040
Pillai's Trace	0.08067069	4.62	3	158	0.0040
Hotelling-Lawley Trace	0.08774950	4.62	3	158	0.0040
Roy's Greatest Root	0.08774950	4.62	3	158	0.0040

Table 4.3.8 Multivariate Analysis of Variance of cellulose (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall cellulose Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.78744470	14.22	3	158	<.0001
Pillai's Trace	0.21255530	14.22	3	158	<.0001
Hotelling-Lawley Trace	0.26993045	14.22	3	158	<.0001
Roy's Greatest Root	0.26993045	14.22	3	158	<.0001

Table 4.3.9 Multivariate Analysis of Variance of total hemicelluloses (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall total hemicelluloses Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.85455146	8.96	3	158	<.0001
Pillai's Trace	0.14544854	8.96	3	158	<.0001
Hotelling-Lawley Trace	0.17020454	8.96	3	158	<.0001
Roy's Greatest Root	0.17020454	8.96	3	158	<.0001

Table 4.3.10 Multivariate Analysis of Variance of total lignin (E-dunnii data)

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall total lignin Effect					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.93899854	3.42	3	158	0.0188
Pillai's Trace	0.06100146	3.42	3	158	0.0188
Hotelling-Lawley Trace	0.06496438	3.42	3	158	0.0188
Roy's Greatest Root	0.06496438	3.42	3	158	0.0188

In summary fibre lumen diameter, kappa number and total hemicelluloses are the only variables that concurrently significantly affect viscosity, brightness and yield in both data sets. One way of wood quality improvement is good management of the raw material supply for processing which in effect means determine groups of *Eucalypts* wood hybrid type that have similar characteristics. One way of making a homogeneous group is by classification of trees in terms of viscosity, brightness and yield an issue that will be addressed in the next chapter, by means of cluster analysis.

Chapter 5

Cluster analysis for the combined data

Mill performance and value of pulp production are related to the uniformity of raw material supply to a mill. Management of *Eucalypts* raw material in terms of differences in viscosity, brightness and yield is important during production process. Relating these properties to some discernible sources of variation like location and hybrid type will help to improve the uniformity of *Eucalypts* raw supply. This chapter deals with the possibility of classifying *Eucalypts* based on viscosity, brightness and yield and the relation to location and hybrid type using cluster analysis. First we will deal with some brief theory of cluster analysis.

The term cluster analysis does not imply a particular statistical method or model as do a regression analysis. We often do not have to make any assumptions about the underlying distribution of the data. Ideally we have data that may have come from several populations, but it is not known which population they came from. Cluster analysis or clustering is the process of grouping of similar objects using data about the objects. It is part of the general scientific process of searching for patterns in data than trying to construct laws that explain the pattern. The idea of grouping together observations that are 'similar', raises questions of how one defines similarity and how similar do they have allocate them into the same group. Using cluster analysis, we can also form groups of related variables. In cluster analysis there is no a prior assumption made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities).

The input required in order to undertake cluster analysis is a similarity measure(s) or data from which similarities can be computed. There are three main types of data structures used in cluster analysis (Johnson and Wichern, 2002). The first is the d-dimensional vector data $X_1, X_2, \cdots, X_n$ arising from measuring or observing d characteristics on each of the n objects or individuals. The characteristics or variables may be quantitative or qualitative and the data can be expressed in a matrix form $X = [(x_{ij})]$ where X is of dimension $d \times n$.

The aim of the cluster analysis is to devise a classification scheme for grouping the X_i into g clusters. In this analysis the characteristics of the cluster and, in most cases, the number of clusters have to be determined from the data itself.

A second type of data encountered in cluster analysis consist of an n × n proximity matrix denoted by $[(C_{ij})]$ or $[(D_{ij})]$, where (C_{ij}) or (D_{ij}), is a measure of similarity (dissimilarity) between the i^{th} and j^{ih} subjects. An element (C_{ij}) or (D_{ij}) is called a proximity and the data is referred to as proximity data.

A third type of data that is already in a cluster format is what might be called sorted data. All the three types of data can be converted into proximity data. Once we have a proximity matrix, we can then proceed to form clusters of objects that are similar or close to one another based on the proximities.

In order to achieve the basic objective of cluster analysis, that is to discover natural groupings of the items (or variables), we must first develop a quantitative scale on which to measure the association (similarity) between objects.

5.1 Similarity Measures

Most efforts to produce a rather simple group structure from a complex data set require a measure of closeness or similarity. Important considerations include the nature of the variables, scales of measurements and subject matter knowledge.

When items (units or cases) are clustered, proximity is usually indicated by some sort of distance. On the other hand, variables are usually grouped on the basis of correlation coefficient or like measures of association.

The statistical distance between the same two observations is

$$d(x,y) = \sqrt{(x-y)'\boldsymbol{A}'(x-y)}\,. \qquad (5.1)$$

Ordinarily, $\mathbf{A} = \mathbf{S}^{-1}$, where **S** contains the sample variance and covariance. However, without prior knowledge of the distinct groups, these sample quantities cannot be computed. For this reason, Euclidean distance is often preferred for clustering.

Another distance measures are the Minkowski metric, Canberrra metric and the Czekanowski coefficients, where the last two measures are defined for nonnegative variables only.

Table 5.1 Summary table of distance measure

1	Minkowski metric	$d(x,y) = \left[\sum_{i=1}^{p} \lvert x_i - y_i \rvert^2\right]^{1/2}$
2	Canberrra metric	$d(x,y) = \left[\sum_{i=1}^{P} \frac{\lvert x_i - y_i \rvert}{(x_i + y_i)}\right]$
3	Czekanowski coefficients	$d(x,y) = 1 - 2\frac{\sum_{i=1}^{P} min(x_i, y_i)}{\sum_{i=1}^{P}(x_i + y_i)}$

5.2 Clustering methods

There are basically three types of clustering available, the first one is hierarchical clustering. In this clustering method the clusters are themselves grouped into bigger clusters, the process being repeated at different levels to form what is technically known as a tree of clusters (Everitt and Dunn, 2001, Everitt, 1974) . Such a tree can be constructed from the bottom up using an agglomerative method that proceeds by a series of successive fusions of n objects into clusters, or from the top down using a divisive method which partitions the total set of n objects in to finer and finer partitions. The former method begins with a single cluster of n objects, where the latter method consists of the reversus process. The graphical presentation of hierarchical clustering is called a dendogram. The second method is partitioning, here the objects are partitioned into non overlapping clusters. The last one is overlapping clusters.

5.2.1 Hierarchical clustering method

Hierarchical clustering agglomerative techniques all begin with clusters each containing just one object, a proximity matrix for the n objects and a measure of distance between two clusters, where each cluster contains one or more objects. The first step is to fuse the two nearest objects into a single cluster so that we now have n-2 clusters containing one object each and a single cluster of two objects. The second step is to fuse the two nearest of the n-1

clusters to form n-2 cluster. We continue in this manner until at the $(n-1)^{th}$ step we fuse the two clusters left into a single cluster of n objects. A number of different distance measure for cluster have been proposed and the one that are or have been widely used are single linkage (nearest neighbor) method, complete linkage (farthest neighbor) method, centroid method and median method.

To demonstrate the use of single linkage method let C_1 and C_2 be two clusters. Next the distance between them is defined to be the smallest dissimilarity between a member of C_1 and C_2 that is

$$d(C_1)(C_2) = \min\{d_{ij}: i \in C_1; j \in C_2\}. \quad (5.2)$$

Complete linkage method is the opposite of single linkage method. In this method distance between clusters is defined in terms of the largest dissimilarity between a member of C_1 and C_2 given by

$$d(C_1)(C_2) = \max\{d_{ij}: i \in C_1; j \in C_2\} \quad (5.3)$$

At each step we fuse the two clusters that are closest, that is, those with minimum dissimilarity $d(C_1)(C_2)$.

Under the centroid method the distance between two clusters is defined to be the distance between the cluster centroids. If $\bar{X}_1 = \sum_{i=c} \frac{x_i}{n_i}$ is the centroid of n_1 member cluster C_1 and $\bar{X}_2$ similarly defined for C_2 then $d(C_1)(C_2) = P(\bar{X}_1, \bar{X}_2)$; where P is the proximity measure such as the squared Euclidean distance. In general the procedure starts with a proximity matrix with elements P (X_i , X_j) and at each stage the two nearest clusters are fused and replaced by the centroid of the new cluster.

Median method is the same as the centroid method, except that a new cluster is replaced by the unweighted average $\bar{X} = \frac{1}{2}(\bar{X}_1 + \bar{X}_2)$. This method was introduced to overcome a shortcoming of the centroid method, namely, that if a small group fuses with a large one, it loses its identity and the new centroid may lie in the large group.

5.3 Non- hierarchical methods

Nonhirarcical clustering techniques are designed to group items, rather than variables, into a collection of K clusters. The number of clusters, K, will either be specified in advance or determined as part of the clustering procedure. This method starts from either an initial partition of items into groups or an initial set of seed points, which will form the nuclei of clusters. One way to start is to randomly select seed point from among the items or to randomly partition the items into initial groups. One of the more popular non-hierarchical procedures is the K-means method.

K-Means Method

The K-means method in its simplest version is a process composed of these three key steps:

Step 1. Partition the data into K initial clusters. This may be done at random.

Step 2. Determine the centroid (that is the mean) for each cluster. For each observation, reassign it to the cluster that is closest. That is, compute the distance to each centroid and assign it to the one that is smallest. It is best to use the standardized data and normally use the Euclidean distance. If an observation is reassigned, recompute the centroid for the cluster receiving the new observation and for the one losing that observation. This is like an updating step.

Step 3. Repeat Step 2 until no more reassignments have been made. Alternatively, we could begin the procedure by specifying K initial centroids and proceed as in Step 2.

Since the procedure depends on the initial choice of clusters and the number of clusters, it is suggested that the process be repeated for different choices. In particular, the specification of K could lead to unusual clustering and outlying observations can produce unusual clusters. There are numerous ways we can sort cases into groups. The choice of a method depends on, among other things, the size of the data file. Methods commonly used for small data sets are impractical for data files with thousands of cases. Identifying groups of individuals or objects that are similar to each other but different from individuals in other groups can be intellectually satisfying, profitable, or sometimes both.

5.4 Application of Cluster Analysis on Combined Data

The main source of variation of Eucalypts wood supply destined to a Sappi mill is trees hybrid type and plantation area difference. To include all location and hybrid types in the cluster analysis the usage of combined data is important. The aim of the current application of cluster analysis is to group the trees based on viscosity, brightness and yield into homogenous groups and associate the grouping with location and hybrid type and then characterize the group.

A SAS PROC CLUSTER method was applied to perform cluster analysis of K- means method with median distance as a measure of similarity. Sorted data by location and hybrid type result from a cluster analysis was applied to the combined hybrid data which indicated some possible grouping of trees based on location and hybrid type. A PROC CLUSTER method output summary results for the combined data are presented in Tables 5.2, 5.3 and 5.5. and dendogram Figuer 5.1

Table 5.2 indicates that most of the trees planted in location KTG09, P\RidgeB1, P\RidgeC10, P\Ridge C13, Salpine E05 and Terra A1 are found in the first group and trees on location P\Ridge C10 and P\Ridge D13 form the second group. Trees planted in P\Ridge C13 form the third group but trees planted on location Hilelo and KTE 10 are evenly distributed in different groups.

Table 5.2 Summary of clusters according to location (Combined data)

CLUSTER	LOCATION (Grouping of trees based on viscosity yield and brightens)
1	KT GO9, P/RidgeB1, P/RidgeC10 , P/Ridge C 13,Salpine E05,Terra A1
2	P/RidgeC10, P/Ridge d 10
3	
4	P/RidgeC13
5	
	HILELO and KT E10 evenly distributed

Association of cluster result with trees hybrid type (Table 5.3) also indicate that hybrids E. grandis, E.uro×ETera, E.Uro×GRA/Ter, E.urophlla, G×GT, G×GU, GP, GU, GU((GP)×E.ter)), GU×GC, UG form the first group, GP, GU×((GP)+G×GT)), G×((GP)+(G×GT)), GU×GP, GU×GT, UG second group,E. Grandis, E. Smithii, GU×(G×GU) third group and GU the fourth group but , GU×U and E. dunnii evenly distributed into different group.

Table 5.3 Summary of clusters according to hybrid type (Combined data)

CLUSTER	HYBRIDE TYPE (Grouping of trees based on viscosity yield and brightens)
1	E. grandis, E .uro × E Tera, E. Uro×Gra/ Ter, E.urophlla, G× GT, G×GU, GP, GU,GU((GP)×E .ter)),GU×GC,UG
2	G×GU, GC, GP GU×((GP)+G×GT)),GU ×((GP)+(G×GT)), GU×GP, GU×GT, UG
3	E. grandis,E.smithii,GU×(G×GU)
4	GU
5	

Note :- GU×U and E. dunnii evenly distributed

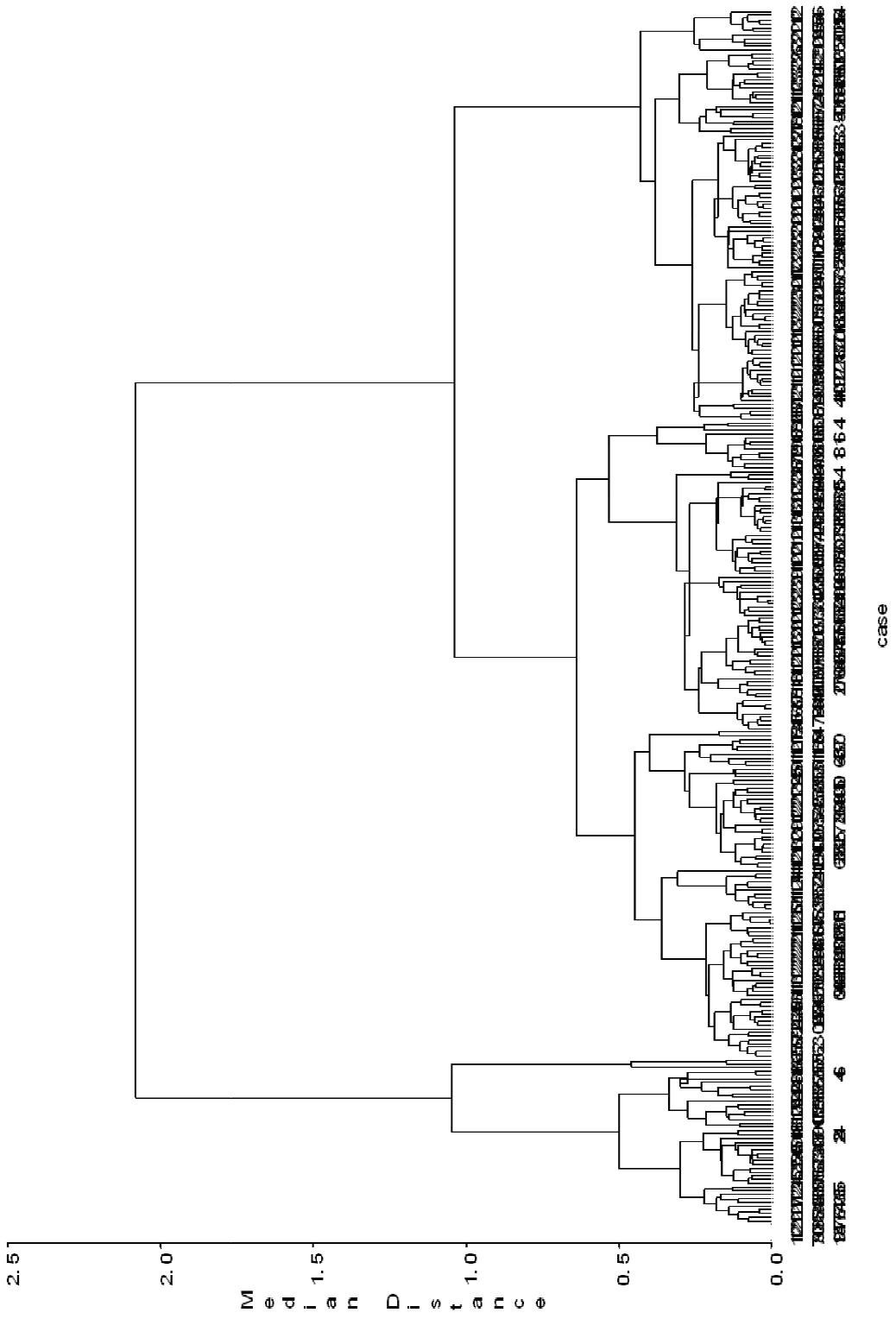


Figure 5.1 Dendogram of combined data

From the overall resolute of cluster analysis we conclude that it could not able characterization of the group. However there is some homogeneity interims of viscosity, brightness and yield. In the next chapter we will summarize all the analysis applied in this thesis.

Chapter 6

Summary and conclusions

The main objective of this thesis was to determine which of the anatomical, chemical and pulp properties of wood that are significant factors affecting pulp properties namely viscosity, brightness and yield and also to assess the effect of geographic characteristics (location and site quality), age and species type on viscosity, brightness and yield of wood pulp. The data used in this thesis came from three different sources. Exploratory data analysis was first performed on the different hybrid data set and E-dunnii data set. The results of this initial preliminary assessment indicate the presence of a linear relationship between some of the independent anatomic, chemical and pulp property measurements and the dependent variables namely viscosity, brightness and yield on both data sets but the relationship differed in terms of strength and number of variables.

Some summary statistics in addition to normal probability plots and histograms each of dependent variables viscosity, brightness and yield for each data sets indicated the presence of some outlier observations and approximately normally distributed. A box-plot analysis of viscosity, brightness and yield versus age, location and site quality supported the presence of some outliers and a mean variation between different age categoris of trees, locations of where trees grown and site quality.

A correlation matrix assessment between each variable also indicated the presence of redundancies in some of the information. Values of correlation coefficient up to 0.9904 between some variables were found which necessitated the need for some data reduction methods such as principal component analysis. The rotated factors from the use of principal component analysis as was shown in Tables 3.20, 3.21 and 3.22 did not support the possibility of data reduction using principal component analysis. Other method of data reduction was used in particular stepwise regression in this case.

All anatomic, chemical and pulp property measurements are continuous variables and categorical variables namely age, location, site quality and hybrid types were among the fixed effects. The presence of a linear relationship between the dependent and the independent variables was supportive of the application of a multiple linear regression analysis for both

data sets and for each dependent variables (viscosity, brightness and yield). The results for both data sets showed the existence of a highly significant linear relationship with F-test p-values <.0001 at 5% significance level. The coefficient of determination R-square values of 52.73%, 81.20% and 81.27% for viscosity, brightness and yield respectively were obtained for the different hybrid data and 53.84%, 43.52% and 59.06% for the E-dunnii data set. But the presence of high correlation between some of the independent variables suggested the need for multicolinearity tests using a variance inflation factor (VIF) threshod of >10 for some independent variables. This led to the application of stepwise regression analysis for variable selection. The result of this application for each of the dependent variables (viscosity, brightness and yield) for the two data sets were different in terms of the type and number of variables selected. Furthermore, a test of multiple regression model assumptions or model diagnosis using residuals like the R-studentized residuals versus predicted values plots for viscosity indicated that the variance of the errors was not constant implying that the assumption of constant variance of error term was violated.

An alternative approach commonly used in applied statistics as a remedial measure for non-normality is the use of transformation of variable which in this case was applied to the viscosity data in such a way as to obtain a new regression model which clearly possessed properties of constant variance and / or normally distributed errors. One of the appropriate or convenient and robust way of transforming data is Box-Cox transformation. Using Proc TRANSREG in SAS a Box-Cox transformation was performed and the best value of lambda was conveniently chosen as zero for all the data sets implying the natural log transformations of viscosity was the most suitable way of solving the problem of non-constant variance of the error residuals. An application of multiple regression analysis using the transformed viscosity (log viscosity) and model diagnostics using residual plot showed that the problem of non constant variance of the error term was solved and an improvement on the fitted models followed. R-square values also increased from 52.73% to 62.161% for the different hybrid data set and from 53.84% to 59.54% for E-dunnii data set.

Another statistical application tool applied on the different hybrid data set for categorical explanatory variables was a multiple comparison procedures also called mean separation tests. This gave us more detailed information about the differences among the means of viscosity, brightness and yield over different locations and site quality where eucalyptus trees

were planted and age group of trees using Duncan's multiple range test. The results also showed a significant mean difference (P-value <0.0001) of mean viscosity, brightness and yield between some age groups, locations and the two site quality and hybrid combinations. Even though the mean of viscosity, brightness and yield comparison between different hybrid combinations were significant for some hybrid type the number of observations was too small namely below five which may have affected the reliability of this mean comparison result over hybrid type.

The presence of a moderate correlation between the dependent variables viscosity, brightness and yield also required a statistical analysis test which considered this correlation and their linear relationship with tree anatomic, chemical and pulp property measurements. To deal with this aspect of the data multivariate linear regression for continuous explanatory variables namely MANOVA was used.

An application of multivariate linear regression using SAS PROC GLM supported the presence of significant linear relationship between the dependent variables included into the model at once and independent variables for all the three data sets but the number and type of variables that affect the dependent variables vary from one dependent variable to the other and also from one data set to the other.

One of the problems of Sappi pulp mill is the non-uniform supply of eucalypts pulp wood in terms of viscosity, brightness and yield. Classification of different hybrid type of trees based on those three variables was important. A cluster analysis was applied to the combined data set and its results indicated some sort of grouping of trees of different hybrid type as presented in Table 5.2. This classification may help in terms of management by ensuring a uniform pulp wood supply of raw material destined to pulp mills and controlling for production process variation.

In conclusion, significant effect of age, hybrid type, location and site quality difference on pulp viscosity, brightness and yield was established. This mean that plantations site quality improvement by using suitable agricultural practice and selection of beast hybrid types and improvement of trees through breeding will help to ensure uniformity of pulp wood supplied to the pulp mills. The multiple regression models were used for prediction of viscosity, brightness and yield and also to identify which anatomic and chemical measurements are

significant contributors to the production of quality pulp. Further research including finding a functional relationship between some selected anatomic and chemical measurements and pulp viscosity, brightness and yield will be important. Further statistical techniques to deal with non-homogeneous wood quality due to factors such as location, type of data, site quality and others not considered in this work can be explored.

References

BERK, R. A. (2004). *Regression Analysis A constractive Critique,* California, Sage Publications, Inc.

BUSH, T. & NAIDOO, W. (2008). The development of Near Infra-red(NIR) caliberation models to predict selected anatomical properties of a range of subtropical *Eucalypts*. Forestery and Forest Products Research Center CSIR, UKZN.

BUSH, T., NAIDOO, W. & GOUNDEN, N. (2008). The development of Near Infra-red (NIR) calibration models to predict the pulp yield, vescosity, density and selected chemical properties of a range of subtropical Eucalypts and Eucalyptus Dunnii sampl. Forest and Forest Products Research Center CSIR, UKZN.

CARROLL, R. J. & RUPPERT, D. (1980). *Transformation and Weighting in Regression,* New York, Chapman and Hall.

CHATFIELD, C. & J.COOLINS, A. (1980). *Introduction to Multivariate Analysis* London, Chapman and Hall Ltd.

DEAN, A. & VOSS, D. (1990). *Design and Analysis of Experiments. Springer Texts in Statistics,* New York, Springer.

EDWARDS, M. (2006). The forestry sector in South Africa "*Its contribution to employment income generation and livelihoods*". Johannesburg.

EVERITT, B. (1977). *Cluster Analysis,* London, Heinemann Educational Books Ltd.

EVERITT, B. S. & DUNN, G. (2001). *Applied Multivariate Data Analysis,* London, Arnold.

EVERITT, B. S. & DUNN, G. (2001). *Applied Multivariate Data Analysis,* London, Arnold.

FAO (2004). Country profiles. www.fao.org accessed in September 2009.

GOMEZ, K. A. & GOMEZ, A. A. (1984). *Statistical Procedures for Agricultural Research,* New York, John Wiley and Sons, Inc.

HSU, J. C. (1996). *Multiple Comparison: Theory and Methods* London, Chapman and Hall.

JOHNSON, R. A. & WICHERN, D. W. (2002). *Applied Multivariate Statistical Analysis,* New Jersey, Prentice-Hall, Inc.

KIM, K. & TIMM, N. (2007). *Univariate and Multivariate General Linear Models,* New York, Chapman and Hall/CRC.

LANG, C. (June 2007). Pulp companies and their expansions plans, www.pulpmillwatch.org accessed in September 2009.

MARQUARDT, D. W. (1970). Generalized inverse, Ridge Regression, Biased Linear Estimate, and Nonlinear Estimation. *Technometrics,* **12(3)** 605 -607.

MILLIKEN, G. A. & E.JOHNSON, D. (2002). *Analysis of Messy Data* Florida, Chapman and Hall/CRC.

MOETI, A. (2007). Factors Affecting The Health Status of The People of Lesotho. *School of Statistics and Actuarial Science.* Pietermaritzburg, University of KwaZulu-Natal.

MONTOGMERY, D. C. (1991). *Design and Analysis of Experiment,* New York, John Wiley and Sons, Inc.

MONTOGMERY, D. C., PECK, E. A. & VINNING, G. G. (2001). *Introduction to Linear Regrssion Analysis,* New York, John Wiley and Sons, Inc.

MORRISON, D. F. (1983). *Applied Linear Statistical Methods,* New Jersey, Prentice -Hall, Inc.

PAPER MANUFACTURERS ASSOCIATION OF SOUTH AFRICA (2002). South African Pulp and Paper Industry Statistical Data. Johannesburg.

PAPER MANUFACTURERS ASSOCIATION OF SOUTH AFRICA (Junuary to December 2007). South African Pulp and Paper industry Statistical Data. Johannesburg.

ROURKE, N. O., HATCHER, L. & STEPANSKI, E. J. (2005). *A Step-by Step Approach to Using SAS for Univariate and Multivariate Statistics,* North Carolina John Wiley and Sons, Inc.

Sappi (2008). Our company www. Sappi.com accessed in September 2009.

SEBER, G. A. F. (1984). *Multivariate Observations,* New York, John Wiley and Sons, Inc.

SEBER, G. A. F. (2004). *Multivariate Observations,* New York, John Wiley and Sons, Inc.

SIVERPERSAD, I. (2007). Linear Model Diagnostics and Measurement Error. *School of Statistics and Actuarial Sciences.* Pitermaritzburg, KwaZulu-Natal.

TABACHNICK, B. G. & FIDELL, L. S. (2001). *Using Multivariate Statistics,* Boston, Allyn and Bacon.

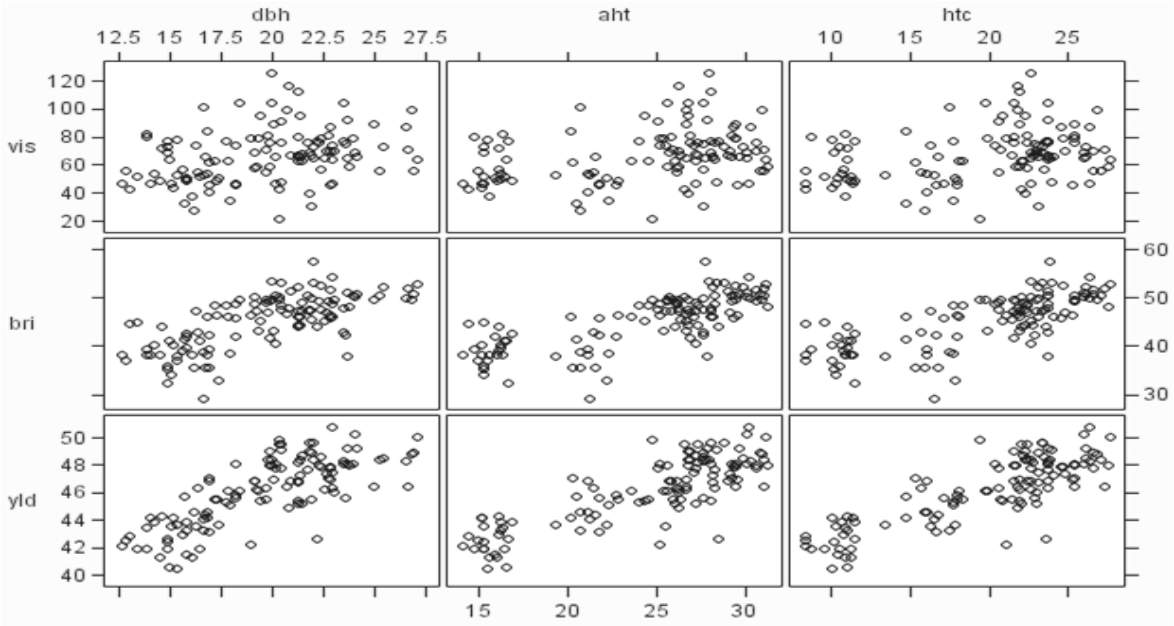
TURNER, P. (2001). Strategic and tactical options for managing the quality and value of Eucalypt plantation resources. South Africa, Forest and Forest Products Research Centre CSIR.

VAUGHAN, T. S. & BERRY, K. E. (2005). Using Monte Carlo Techniques to detemonstrate the meaning and implications of multicolinearity. *Journal of Statistical Education [online] retrived April 20, 2008,* **13(1).**

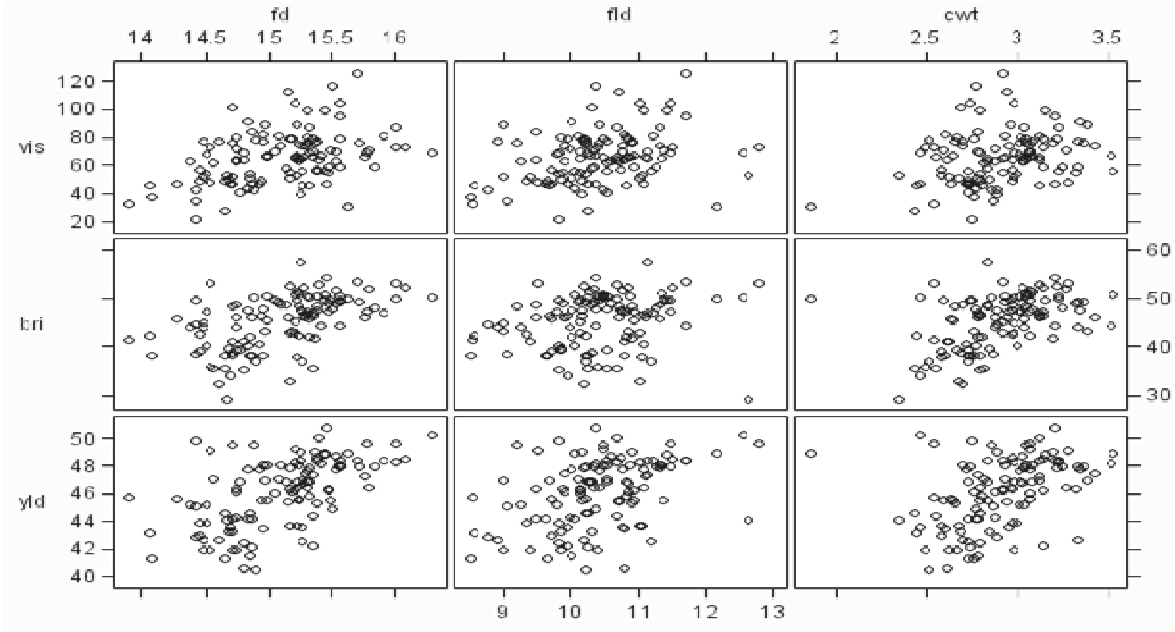
VERBEKE, G., GEYS, H. & MOLENBERGHS, G. (2004). *Correlated and Multivariate Data,* Limburgs, Limburgs Universitar Centrum.

Appendix A: Scatter Plot for Different Hybrid and E-Dunnii Data Set

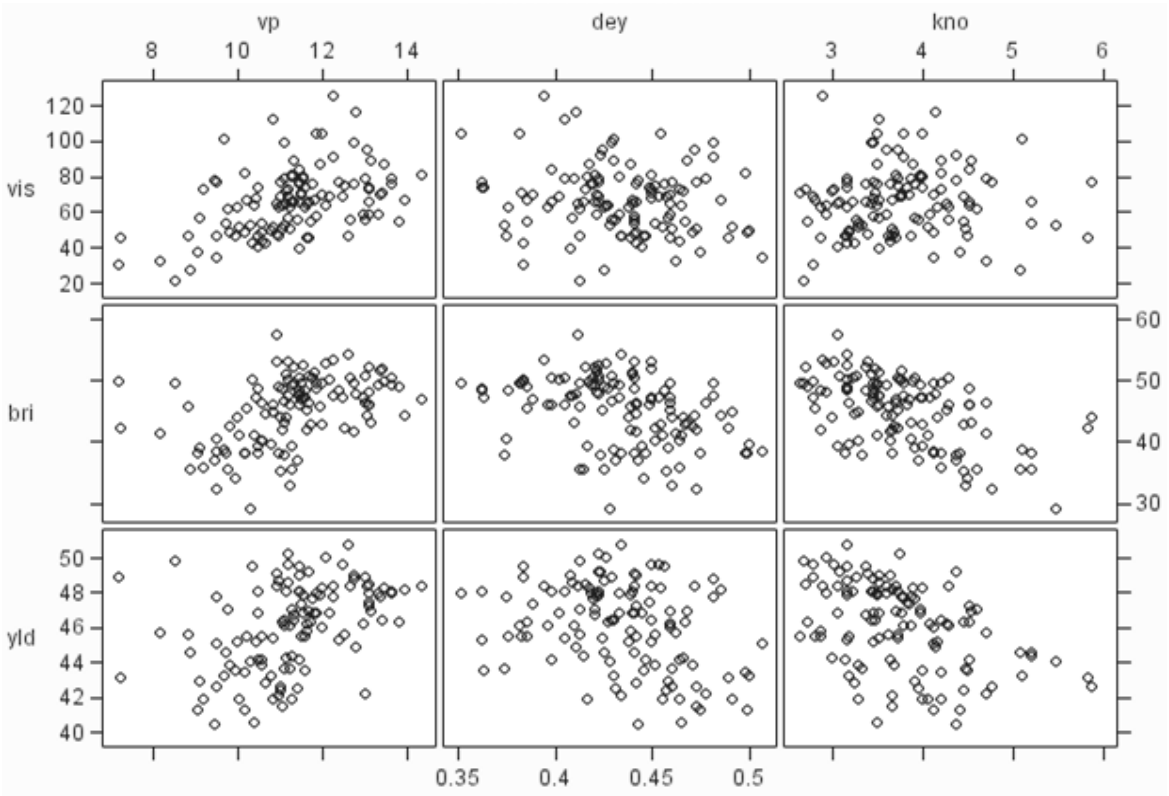
Figure A.1: Different Hybrid Data: Scatter plots of the three dependent variables versus independent variables.



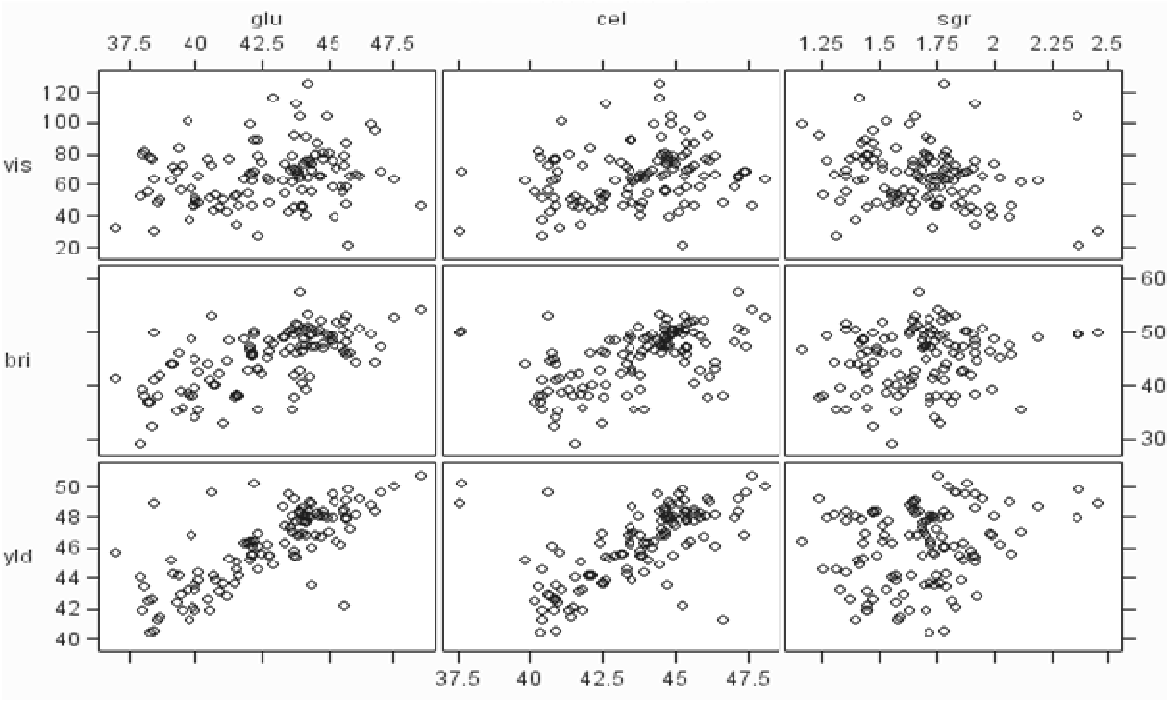
(a) Scatter plot matrix for viscosity, brightness and yield by Average diameter at breast height, Average height and Average height of tree up to diameter of 7cm



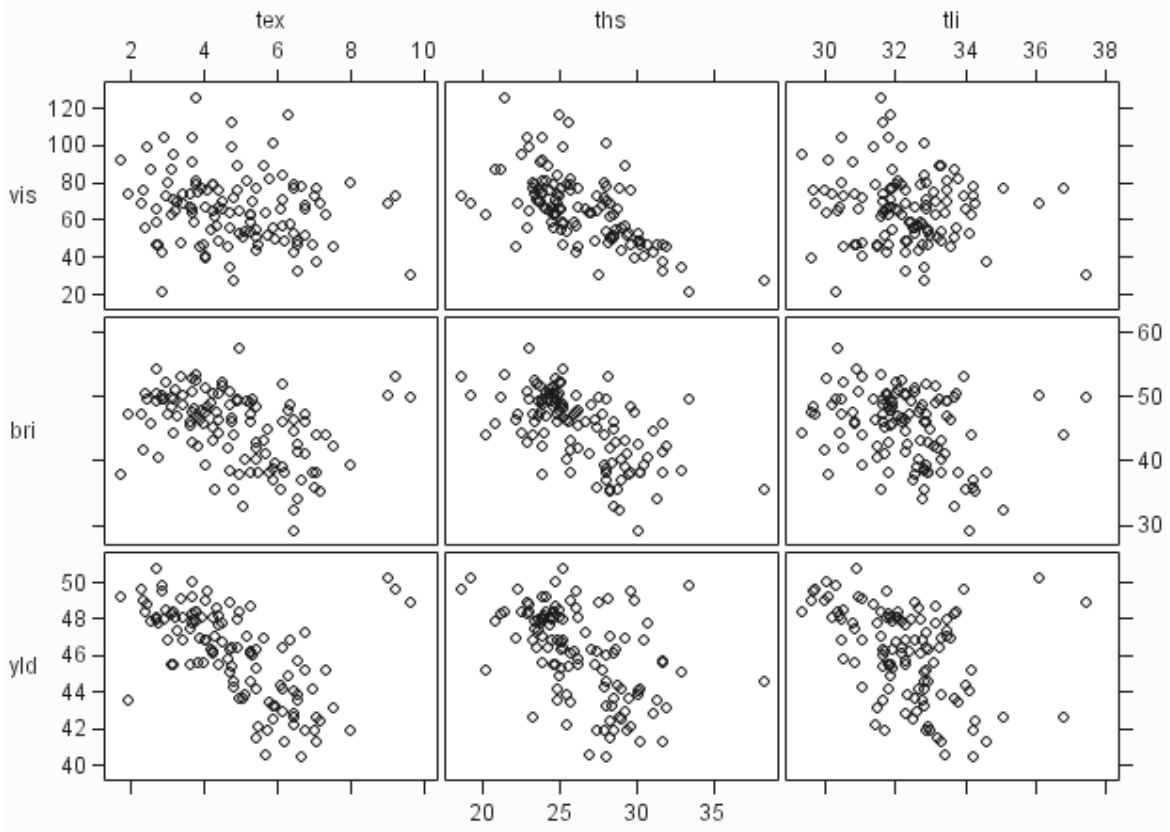
(b) Scatter plot matrix for viscosity, brightness and yield by Fibre diameter, Fibre lumen diameter and Cell wall thickness



(c) Scatter plot matrix for viscosity, brightness and yield by Vessel percentage, Density, Kappa number

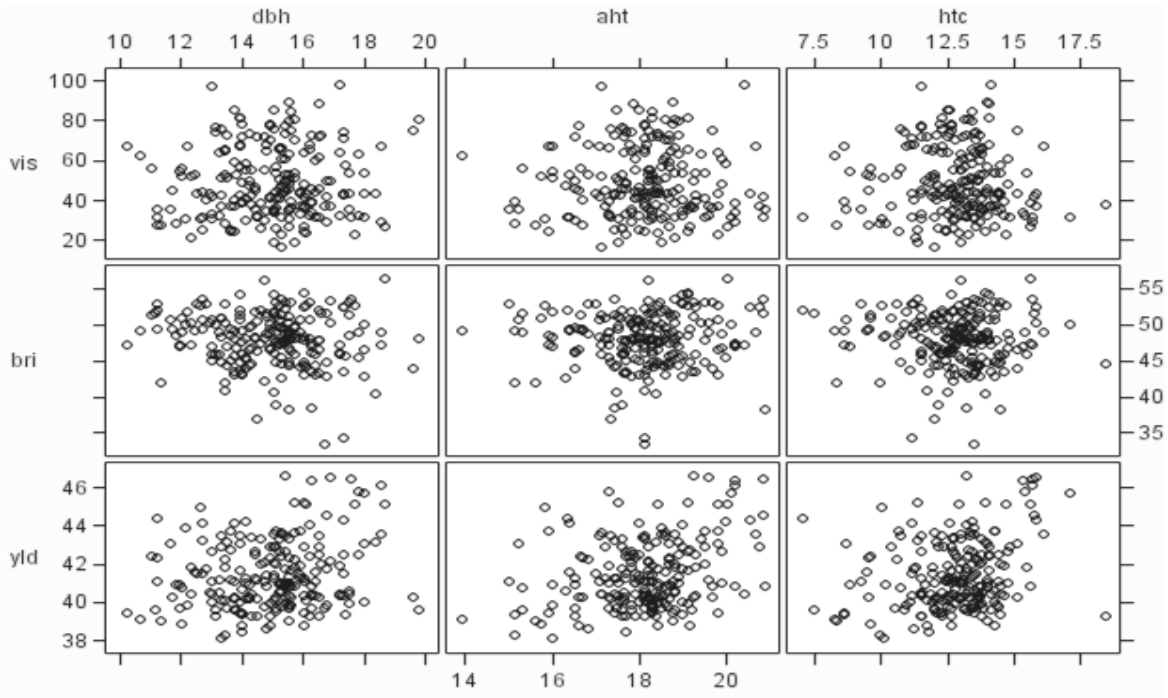


(d) Scatter plot matrix for viscosity, brightness and yield by Glucose, Cellulose and SG ratio

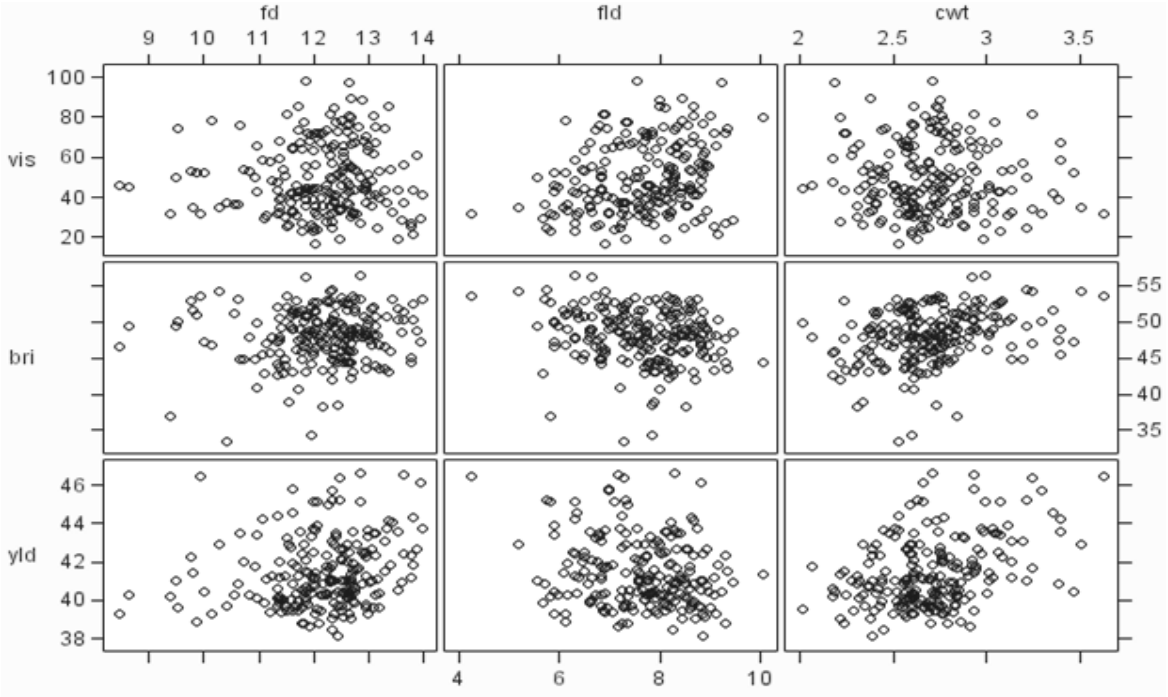


(e) Scatter plot matrix for viscosity, brightness and yield by Total extractives, Total hemicelluloses and Total lignin

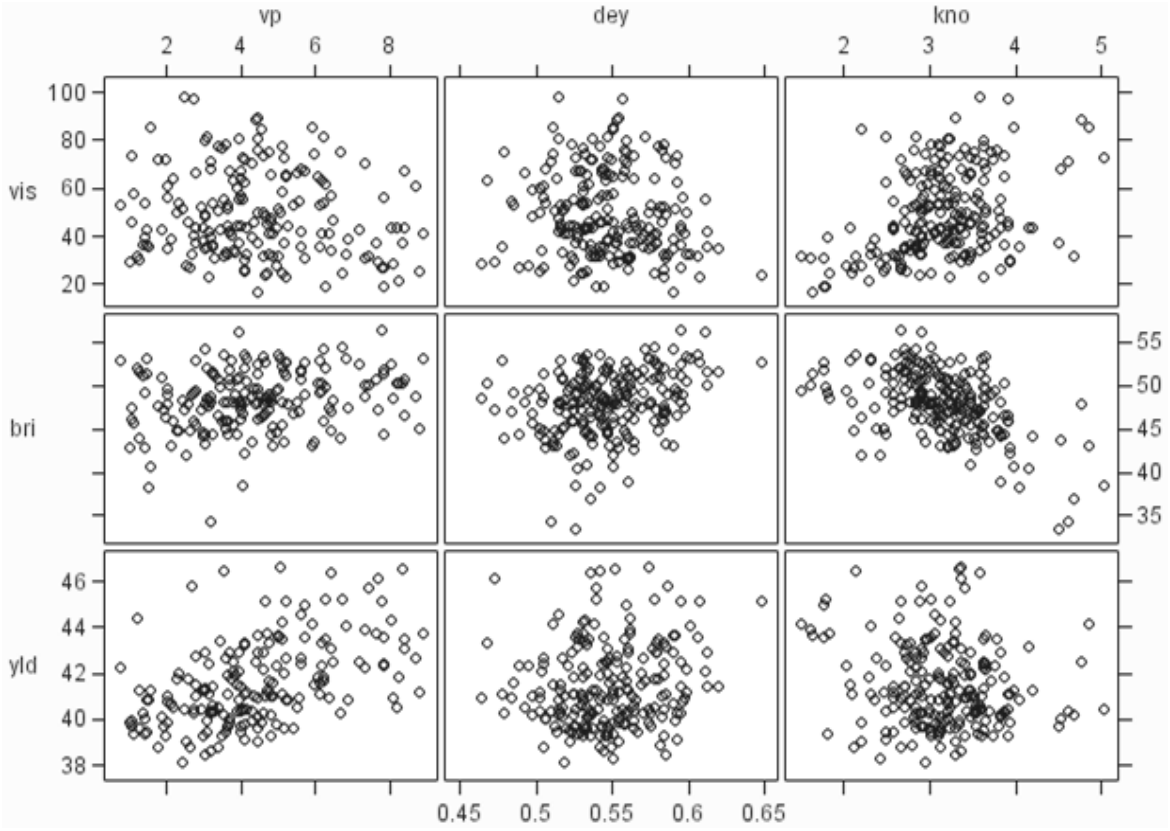
Figure A.2: E-Dunnii Data: Scatter plots of the three dependent variables versus independent variables.



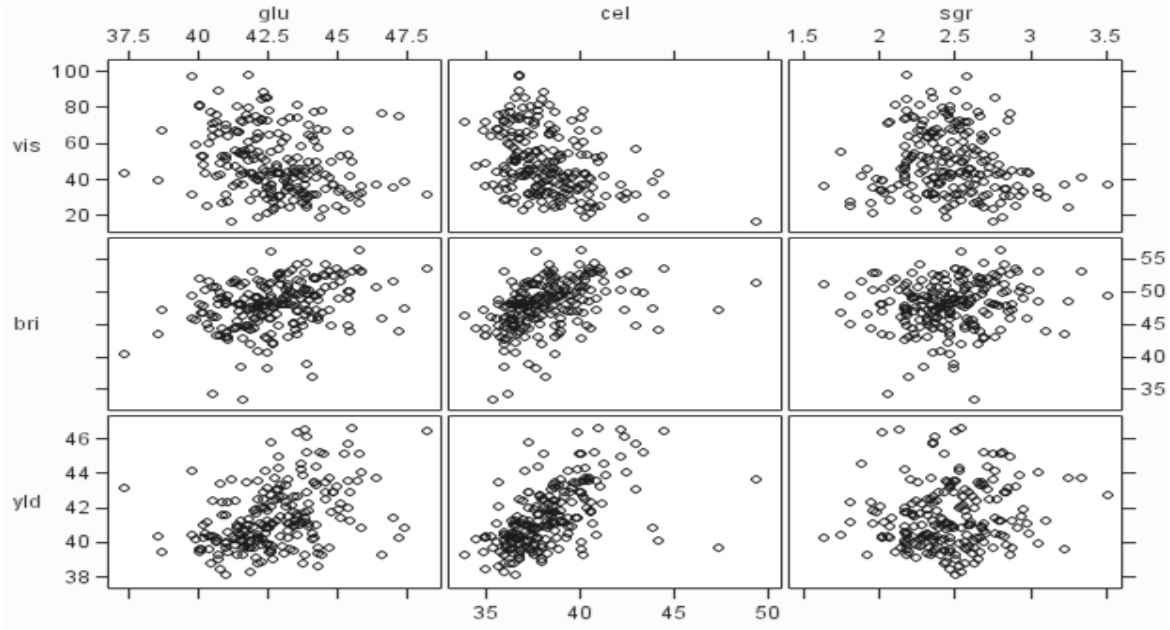
(a) Scatter plot matrix for viscosity, brightness and yield by Average diameter at breast height, Average height and Average height of tree up to diameter of 7cm



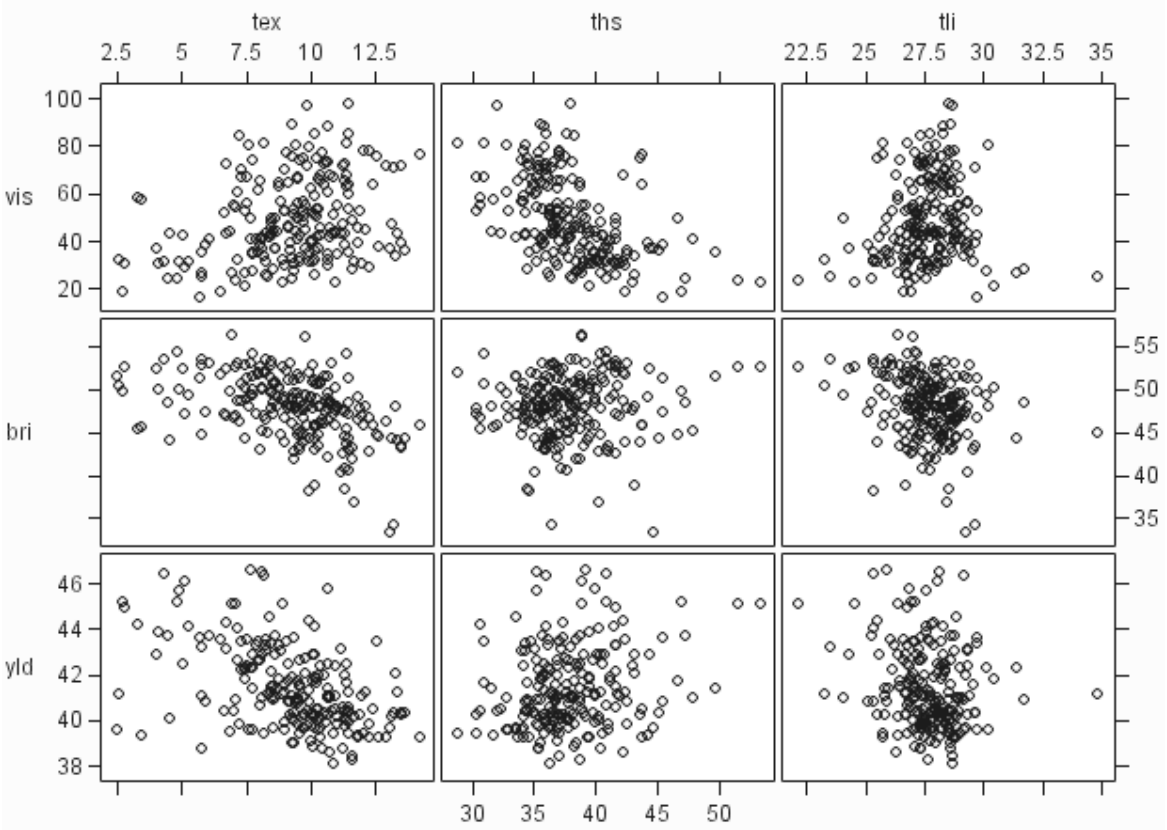
(b) Scatter plot matrix for viscosity, brightness and yield by Fibre diameter, Fibre lumen diameter and Cell wall thickness



(c) Scatter plot matrix for viscosity, brightness and yield by Vessel percentage, Density, Kappa number



(d) Scatter plot matrix for viscosity, brightness and yield by Glucose, Cellulose and SG ratio



(e) Scatter plot matrix for viscosity, brightness and yield by Total extractives, Total hemicelluloses and Total lignin

Appendix B Model selection for viscosity, brightness and yield for E-Dunnii data

Table B.1 Analysis of variance with log viscosity as the dependent variable (E-dunnii reduced data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	13.43961	1.91994	30.89	<.0001
Error	168	10.44077	0.06215		
Corrected Total	175	23.88037			

Table B.2 Parameter estimates of independent variables in a model log of viscosity as the dependent variable (E-dunnii reduced data)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	5.84708	0.83154	3.07279	49.44	<.0001
fd	-0.15870	0.03824	1.07046	17.22	<.0001
cwt	0.54245	0.11695	1.33699	21.51	<.0001
fld	0.25532	0.04272	2.21980	35.72	<.0001
kno	0.17036	0.03783	1.26055	20.28	<.0001
cel	-0.04894	0.01154	1.11713	17.98	<.0001
ths	-0.03061	0.00632	1.45654	23.44	<.0001
tli	-0.03568	0.01568	0.32187	5.18	0.0241

Table B.3 Summary of stepwise selection for the log viscosity model (E-dunnii data)

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	ths		ths	1	0.2876	0.2876	109.708	70.24	<.0001
2	cel		cel	2	0.1249	0.4125	62.3201	36.78	<.0001
3	kno		kno	3	0.0383	0.4508	49.1567	12.01	0.0007
4	tli		tli	4	0.0182	0.4690	43.9760	5.85	0.0166
5	fld		fld	5	0.0169	0.4859	39.2979	5.58	0.0193
6	cwt		cwt	6	0.0321	0.5180	28.6086	11.25	0.0010
7	fd		fd	7	0.0448	0.5628	12.8833	17.22	<.0001
8	dbh		dbh	8	0.0036	0.5664	13.4534	1.39	0.2396
9		dbh	dbh	7	0.0036	0.5628	12.8833	1.39	0.23

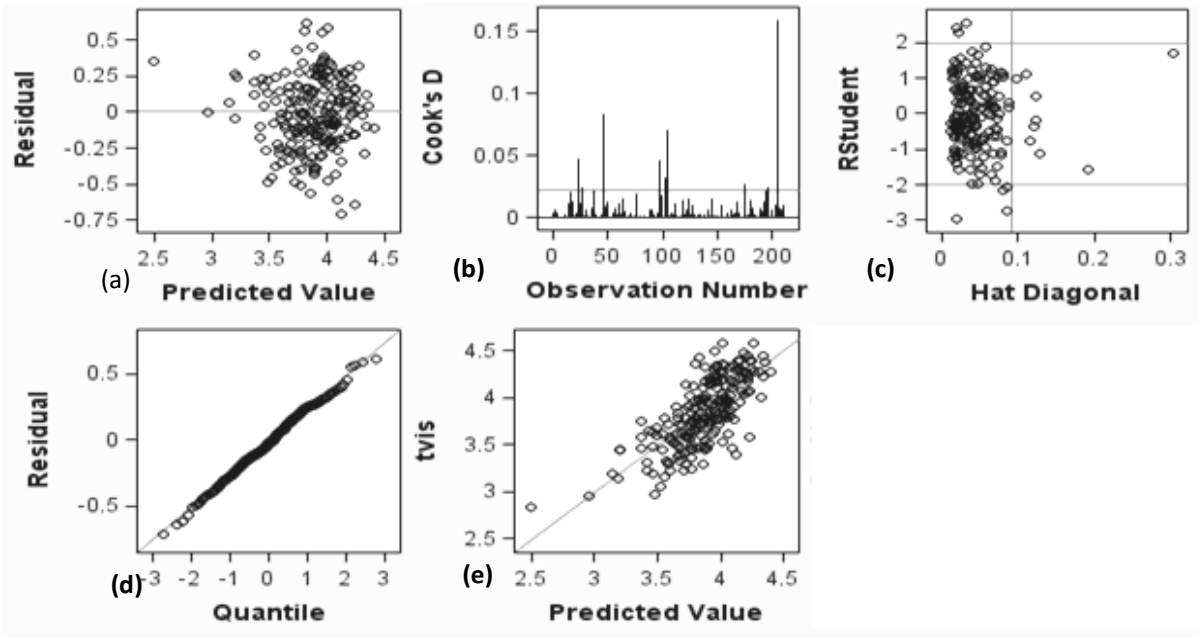


Figure B.1 Model diagnostics with log viscosity as the dependent variable (E-dunnii reduced data) (a) raw residuals versus predicted values (b) Cook's distance (c) standardized residual versus Hat diagonals (d) normal probability plot of residual (e) observed log viscosity versus predicted value

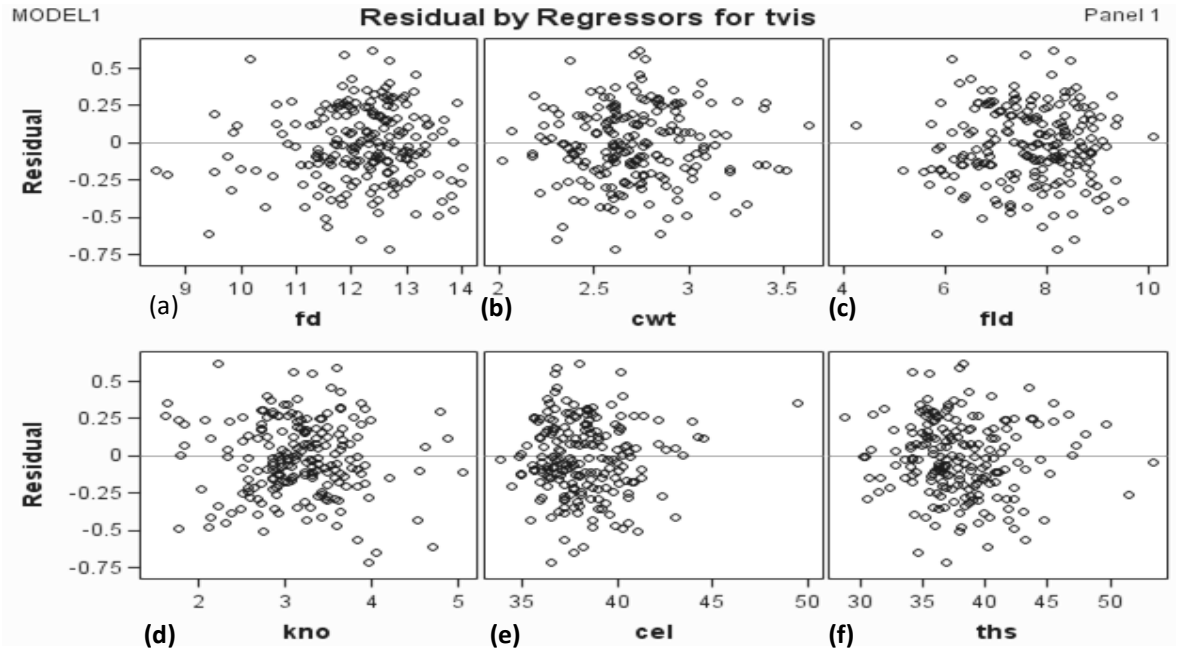


Figure B.2 Model diagnostics of selected variables of log viscosity model (E-dunnii reduced data) (a) fibre diameter (b) cell wall thickness (c) fibre lumen diameter (d) Kappa number (e) cellulose (f) total hemicelluloses (g) total lignin

Table B.4 ANOVA with brightness as the dependent variable (E-dunnii reduced data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	884.91793	126.41685	17.22	<.0001
Error	170	1248.14337	7.34202		
Corrected Total	177	2133.06131			

Table B.5 Parameter estimates of independent variables in a model with brightness as the dependent variable (E-dunnii reduced data)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	64.31767	8.22021	449.47946	61.22	<.0001
htc	-0.23248	0.13211	22.73581	3.10	0.0803
fld	-0.78673	0.23414	82.89456	11.29	0.0010
vp	0.58264	0.12045	171.78361	23.40	<.0001
kno	-2.08901	0.38897	211.77260	28.84	<.0001
dey	27.20555	7.14906	106.32445	14.48	0.0002
ths	-0.12130	0.06085	29.17792	3.97	0.0478
tli	-0.47444	0.17649	53.05509	7.23	0.0079

Table B.6 Summary of stepwise selection for the brightness model (E-dunnii data)

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	kno		kno	1	0.2054	0.2054	53.9255	45.50	<.0001
2	dey		dey	2	0.0806	0.2860	32.8030	19.76	<.0001
3	cwt		cwt	3	0.0563	0.3423	18.6636	14.89	0.0002
4	vp		vp	4	0.0246	0.3669	13.6116	6.72	0.0104
5	tli		tli	5	0.0174	0.3843	10.6064	4.87	0.0286
6	htc		htc	6	0.0102	0.3945	9.6836	2.88	0.0916
7	fld		fld	7	0.0080	0.4025	9.4005	2.26	0.1342
8		cwt	cwt	6	0.0013	0.4012	7.7722	0.37	0.5446
9	ths		ths	7	0.0137	0.4149	5.8484	3.97	0.0478
10	glu		glu	8	0.0071	0.4220	5.7977	2.09	0.1501
11		glu	glu	7	0.0071	0.4149	5.8484	2.09	0.1501

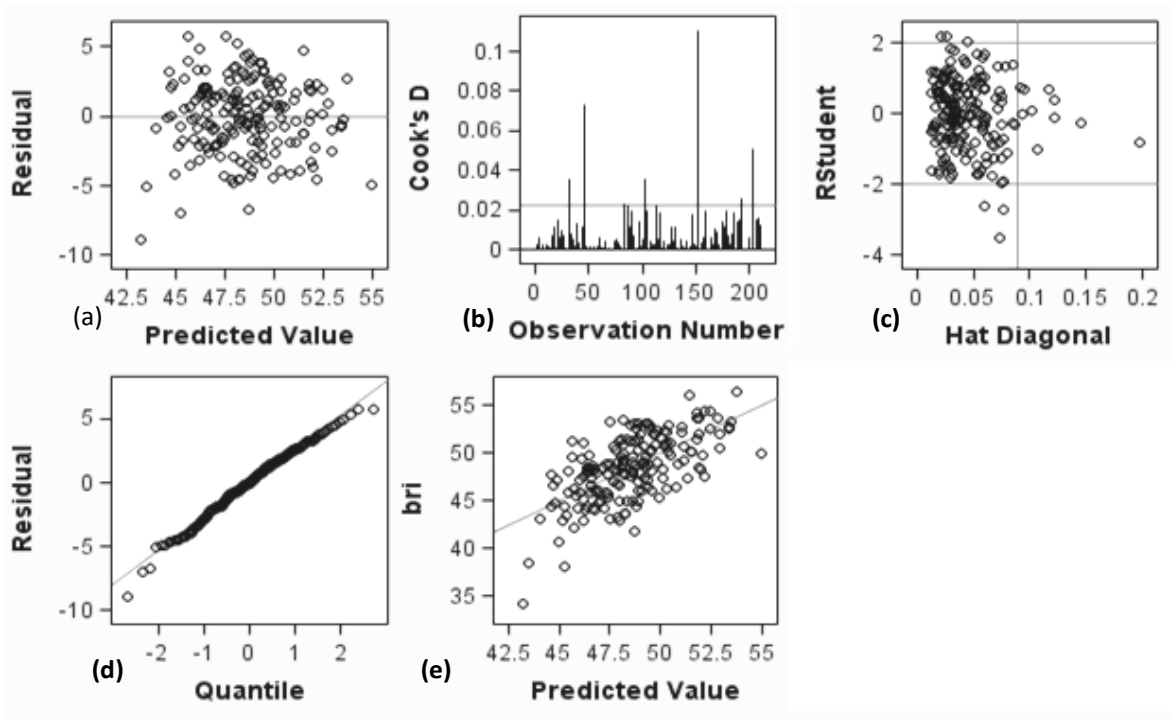


Figure B.3 Model diagnostics with brightness as dependent variable (E-dunnii reduced data) (a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed brightness versus predicted value

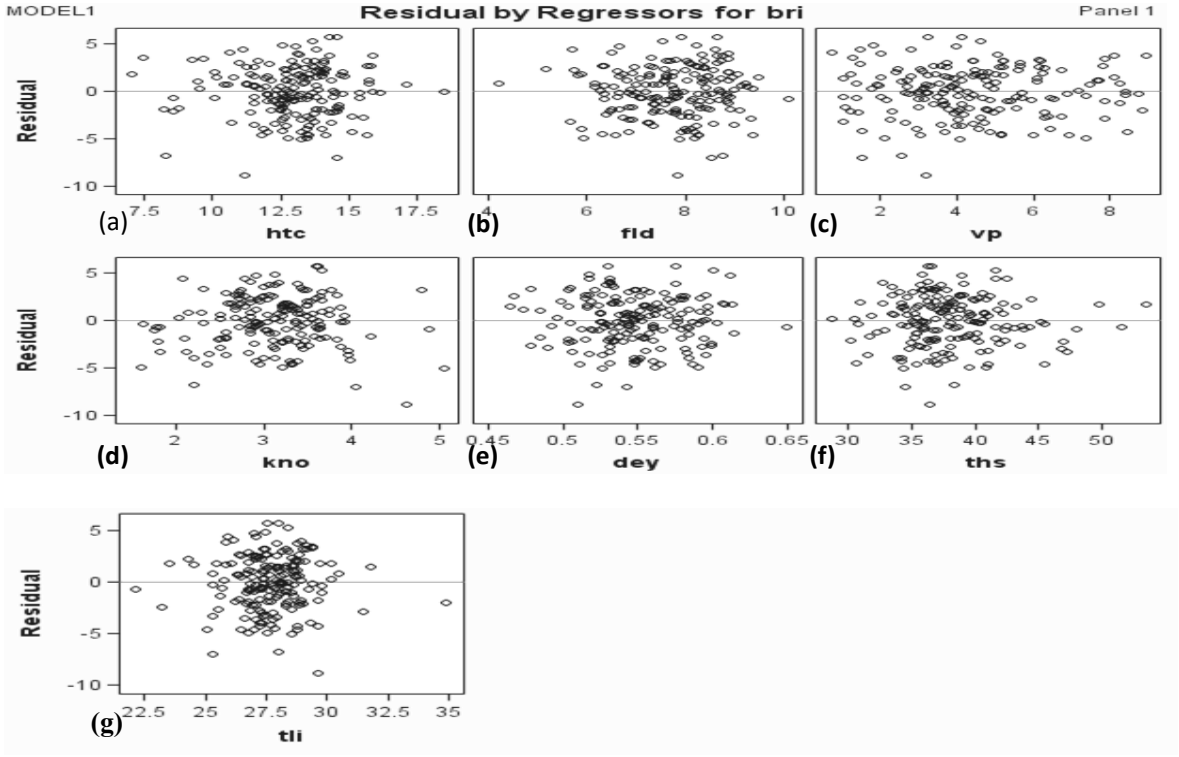


Figure B.4 Model diagnostics of selected variables of Brightness model (E-dunnii reduced data) **(a) average height of a tree up to diameter of 7 cm (b) fibre lumen diameter (c) vessel percentage (d) Kappa number (e) density (f) total hemicelluloses (g) total lignin**

Table B.7 ANOVA with yield as the dependent variable (E-dunnii reduced data)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	346.39956	49.48565	33.09	<.0001
Error	170	254.24439	1.49556		
Corrected Total	177	600.64394			

Table B.8 Parameter estimates of independent variables in model with yield as the dependent variable (E-dunnii reduced data)

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	35.67596	3.79461	132.19642	88.39	<.0001
dbh	0.11173	0.05644	5.86194	3.92	0.0493
cwt	-1.06608	0.57396	5.15965	3.45	0.0650
fld	-0.51763	0.16381	14.93341	9.99	0.0019
vp	0.44559	0.06681	66.52370	44.48	<.0001
kno	0.48037	0.18722	9.84539	6.58	0.0112
cel	0.39044	0.05304	81.03543	54.18	<.0001
tli	-0.26122	0.07129	20.08077	13.43	0.0003

Table B.9 Summary of stepwise selection for the yield model (E-dunnii data)

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	cel		cel	1	0.3664	0.3664	76.7211	101.79	<.0001
2	vp		vp	2	0.0816	0.4480	46.4472	25.85	<.0001
3	glu		glu	3	0.0577	0.5057	25.6054	20.32	<.0001
4	dbh		dbh	4	0.0240	0.5297	18.0954	8.84	0.0034
5	fld		fld	5	0.0161	0.5458	13.7325	6.09	0.0146
6	tli		tli	6	0.0123	0.5581	10.8661	4.76	0.0305
7		glu	glu	5	0.0041	0.5540	10.4752	1.57	0.2114
8	kno		kno	6	0.0141	0.5681	6.9019	5.58	0.0193
9	cwt		cwt	7	0.0086	0.5767	5.5026	3.45	0.0650
10	glu		glu	8	0.0041	0.5808	5.8983	1.63	0.2029
11		glu	glu	7	0.0041	0.5767	5.5026	1.63	0.2029

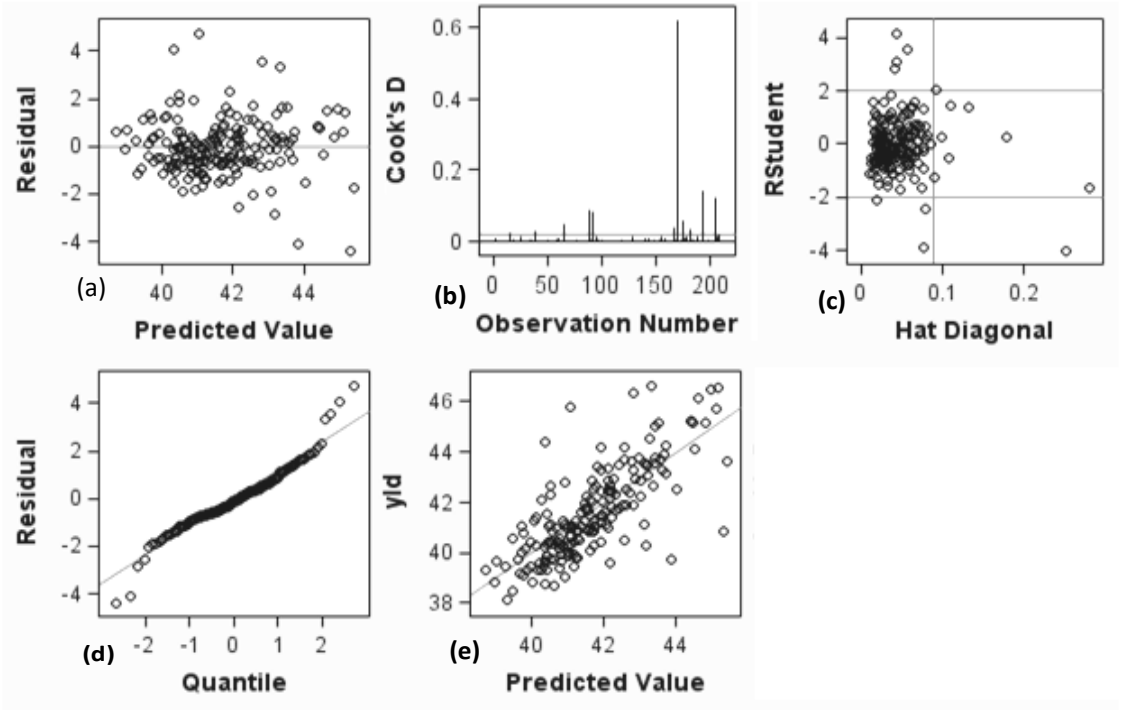


Figure B.5 Model diagnostics with yield as dependent variable (E-dunnii reduced data) (a) raw residuals versus predicted values (b) Cook's distance (c) standardized residuals versus Hat diagonals (d) normal probability plot of residual (e) observed yield versus predicted values

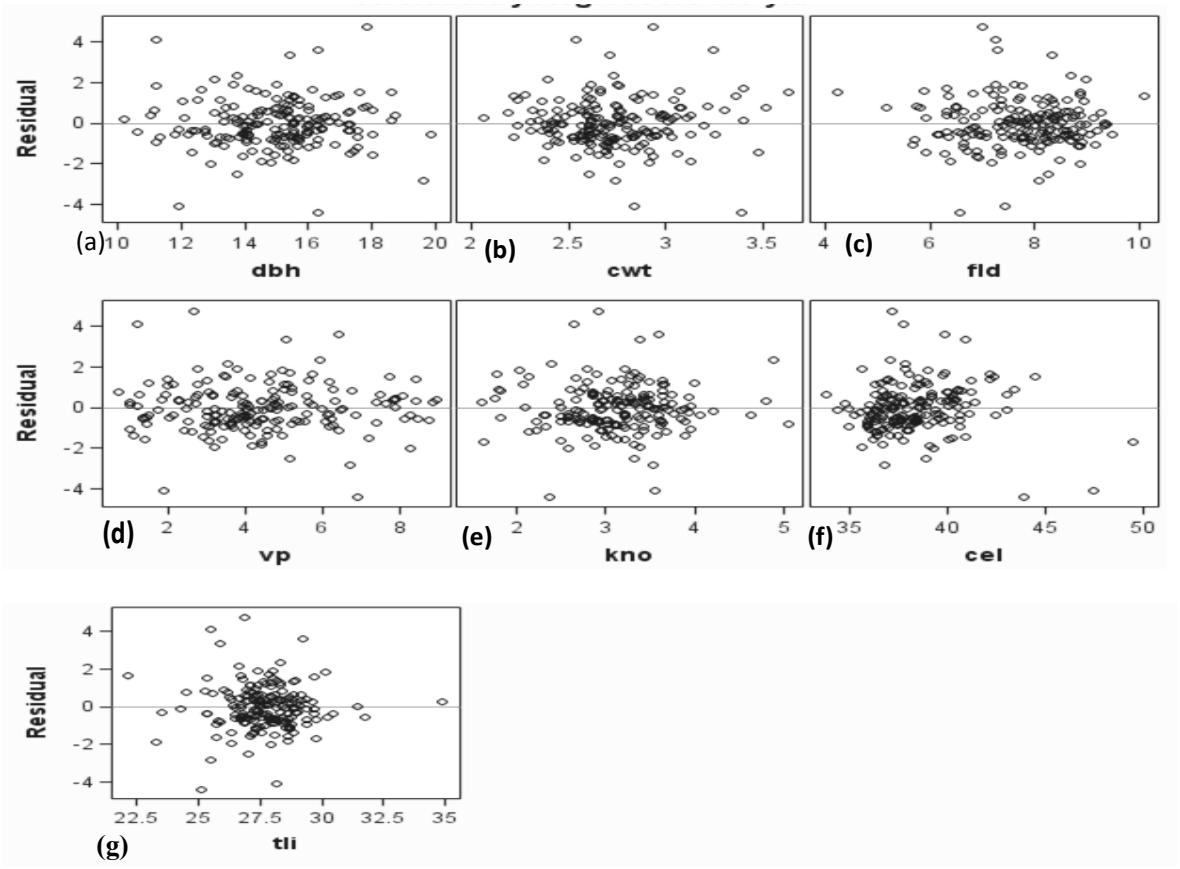


Figure B.6 Model diagnostics of selected variables of yield model (E-dunnii reduced data) **(a) average diameter at breast height (b) cell wall thickness (c) fibre lumen diameter (d) vessel percentage (e) Kappa number (f) cellulose (g) total lignin**