University of KwaZulu-Natal
School of Mathematics, Statistics and Computer Science

Title

A Statistical Analysis of Dissolving Timber Pulp Properties Using Linear Mixed Models

By

Oliver Bodhlyera

2017

A Statistical Analysis of Chemical Pulp Properties Using Linear Mixed Models

By

Oliver Bodhlyera

Supervised by

Professor Temesgen Zewotir and Professor Shaun Ramroop

Submitted in fulfilment of the academic requirements for the degree of

DOCTOR OF PHILOSOPHY
in
Applied Statistics

in the

School of Mathematics, Statistics and Computer Science
University of KwaZulu-Natal
Pietermaritzburg
2017

Dedication

Dedicated to my wife and kids, for their patience, constant support and encouraging cheerfulness.

Declaration

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

I, Oliver Bodhlyera, declare that this thesis is my own, unaided work. This work 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, 2017.

_____ _____

Mr. Oliver Bodhlyera Date

_____ _____

Prof. Temesgen Zewotir Date

_____ _____

Prof. Shaun Ramroop Date

Acknowledgements

This work has been greatly encouraged and motivated by my supervisors, Professor Temesgen Zewotir and Professor Shaun Ramroop who kept on encouraging me to finish the study. Some of the ideas in this study were initiated by my supervisors who also meticulously checked how such ideas were put across in this thesis. I would also like to thank Dr Viren Chunilall for providing a thorough description of the experimentation process and how the data was collected in the chemistry laboratory. Viren provided all the data that was need for this study as well as a thorough description of how the chemical pulping process works and the general properties of dissolving pulp.

The Council for Scientific and Industrial Research (CSIR) and Sappi Saiccor of South Africa are acknowledged for, in-part, financially supporting the project and providing the data. The Forestry and Forest Products Research Centre, CSIR-Durban where the pulping, bleaching and data generation was conducted deserve special acknowledgement.

Abstract

The main focus of the study was to understand the behaviour of seven timber genotypes based on seven chemical properties observed during the chemical pulping process with the prime objective of developing methods of grouping different timber genotypes into compatible groups of timber that can be optimally processed together. Four related statistical methods were used in analysing the data and each had a specific objective. The random coefficients model was used to investigate how the genotypes evolve over the processing stages and it was discovered that the rates of change of the chemical properties studied depended on their initial readings at the beginning of processing. This trend applied for all seven genotypes of pulping trees studied.

The important results that came out of fitting the random coefficient model to the data is that the higher the raw stage readings (initial values) the higher the rates of change in the chemical properties over the processing stages. The changes were either increases or decreases in the chemical property studied. The random coefficient model was also used to suggest a rudimental mixing index for the different genotypes based on the average ranking of their slope parameters (rates of change) for the seven variables studied. It was found, for example, that the genotypes GUA and GUW are the least mixable ones.

Piecewise linear regression models were used to identify important variables when classifying genotypes and it was generally found that viscosity is not a very useful variable in the classification of genotypes. Using piecewise linear regression models together with kernel density estimation a mixing index (scale) was developed that can be used to determine which genotypes are the most mixable for chemical processing. A coparison of the random coefficient and the piecewise linear regression models shows that the two models yielded very similar conclusions on what genotypes are most mixable during processing.

Joint modelling was used to analysis the correlations between evolutions of different chemical properties studied. The various levels of correlations between these

variables were discussed. The main limitation of the joint modelling method was its computational challenges because of the many parameters that need to be estimated at the same time.

Table of Contents

List of Tables

List of Figures

Published Papers

1. Bodhlyera, O., Zewotir, T. and Ramroop (2014). Random coefficient model for changes in viscosity in dissolving pulp. *Wood research* 59(4):2014-571.

2. Bodhlyera, O., Zewotir, T., Ramroop, S. and Chunilall, V. (2015). Analysis of the changes in chemical properties of dissolving pulp during the bleaching process using piecewise linear regression models. *Cellulose Chemistry and Technology* 49(3):3-4.

Paper under Review

1. Bodhlyera, O., Zewotir, T. and Ramroop (2014). Classification of Timber Genotypes using Their Behaviour under Chemical Pulping Using Piecewise Regression and Kernel Density based clustering. Submitted to: *Wood Fibre Science* on 31 March 2017.

Conference presentations

1. Bodhlyera, O., Zewotir, T. and Bush, T. (2011). Fitting a Mixed Effects Model to Data from a Pulping Process. Annual Conference of the South African Statistical Association, December 2011, Potchefstroom.
2. Bodhlyera, O., Zewotir, T. and Ramroop (2014). Classification of Timber Genotypes using Their Behaviour under Chemical Pulping Using Piecewise Regression and Kernel Density based clustering. Annual Conference of the South African Statistical Association, November/December 2015, Pretoria.

Papers in preparation

1. Joint Modelling the Chemical Evolution of the Properties of Dissolving Pulp Under Chemical Processing.

Chapter 1

Introduction

This study is based on data recorded in laboratory experiments at The Forestry and Forest Products Research Centre of The Council for Scientific and Industrial Research (CSIR) in collaboration with Sappi Saiccor of South Africa. The laboratory experiments were carried out under similar conditions that the actual production process is carried out.

Wood pulp is commonly associated with the production of paper, which is considered an essential commodity. Paper has provided a means for people to keep written records, communicate ideas and information and create works of art. Paper has also been used for hygiene purposes. Apart from paper, wood pulp is also used to produce fabrics and other derivative chemicals with many industrial uses. This study looks at chemically processed wood pulp (dissolving wood pulp) which is used in the production of viscose fibre.

Dissolving wood pulp is bleached pulp which has more than 90% pure cellulose fibre with a high level of brightness and uniform molecular weight distribution (Patrick, 2011). It is used to make products such as rayon and acetate textile fibres, cellophane and other chemical products. Cellulose acetate, being important in textile and cigarette industries, is prepared from high quality celluloses such as wood pulps with α-cellulose content of more than 95% (He, Cui ad Wang, 2007).

.

The quality of dissolving wood pulp depends on the quality of the raw wood material and the pulp processing itself (Jahan et al, 2008) and several variables can be used to measure this dissolving pulp quality, of which some are lignin, viscosity, α-cellulose, γ-cellulose, copper numbers, xylose and glucose. The seven chemical properties listed above were analysed in this study with the main aim of better understanding how they change over the processing stages.

Chemical pulping and bleaching removes lignin, hemicellulose and other impurities through dissolution followed by washing. This process results in high purity *α*-cellulose pulp fibre which can be used in the manufacture of the products mentioned above. Cellulose can also be made into cellulose powder which has many industrial uses. During the process of extracting lignin through bleaching, other chemical properties are also altered, namely pulp viscosity, glucose level, degraded celluloses or hemicelluloses, sugars and other chemical properties. The process of chemical pulping has a very low solid matter yield of between 40% and 50% since lignin constitutes a large part of the raw wood pulp and in general most of the lignin and hemicelluloses are removed (Biermann, 1993).

The chemicals used in chemical pulp processing are costly, hence effort must not be spared in trying to optimize the usage of such chemicals. Depending on the type of raw material (tree species or genotype) used in the chemical pulp processing, different environmental and occupational exposures also result. Different wood species or genotypes used in chemical pulping require different types and quantities of chemicals, different in-plant processes, and result in different by-products with different product properties (Soskolne and Sieswerda, 2010). It is therefore imperative that if any different wood species are to be mixed during processing then the mixing should be done after careful consideration of their processing requirements. It would not be optimal to mix two genotypes which require completely different amounts and concentrations of chemicals for processing. This study suggests methods that can be used to optimally mix various tree genotypes during chemical processing according to their observed laboratory behaviour.

1.1. Significance of the study

Numerous studies have been made relative to wood properties, the causes of wood variation, and how best to develop wood for desired products (Zobel and Van Buijtenen, 2012). It was noted that, generally, hardwoods contain a larger proportion of cellulose and hemicellulose and less lignin as compared to softwoods, but hardwoods have a greater percentage of extractives (Soskolne and Sieswerda). Even within these two classes of wood, various genotypes still differ in their chemical requirements during processing. There is therefore, a need for an in depth study of the behaviours of different wood genotypes during chemical processing in order to

determine how to optimally mix them if the need arises, particularly when economic quantities are required. This study suggests some methods of determining wood genotypes that can be mixed for chemical processing according to their similarities in chemical behaviour. The methods suggested in this study are from a statistical point of view and further developments on such methods are to be expected, particularly to take into account biochemical considerations.

This study develops methods of optimally mixing different timber genotypes in chemical pulping. The study suggests that, to better understand chemical changes in pulping processes that involve several sub-processes, piecewise linear regression is useful and that coupling piecewise linear regression with kernel density estimation can help sort out genotypes, or in general, raw materials that behave the same under processing. The researcher is not aware of similar work having done elsewhere. Joint modelling was also used to better understand the evolution of various response variables over processing stages using pairwise fitting to circumvent computational limitation in fitting models with many parameters to be estimated.

1.2. Objectives of the study

The study will first look at statistical tests of the effects of different processing methods carried out during the experiments that produced the data. The prime objective is to develop methods of grouping different timber genotypes into compatible groups of timber that can be optimally processed together. To achieve this objective, a statistical analysis of seven timber genotypes is carried out. The study involves profiling the evolution of some important chemical properties (variables) of the genotypes under chemical pulping. Possible grouping criteria are suggested and such criteria are subject to scrutiny and further development.

1.3. Organisation of the study

This study comprises of four main statistical concepts used to analyse the data, arranged into four methodology chapters. Chapter 1 is the introduction and Chapter 2 looks at the general description of the data including how it was obtained from chemical laboratory experiments. Chapter 3 fits random coefficients models to the data in order to evaluate the effects of three pulping methods on the chemical

properties (variables) studied. In Chapter 4, piecewise linear regression models with dummy time variables are fit to the data in order to assess the effects of each of the three main phases of chemical pulp processing (dissolving pulp). Chapter 5 is a comparison of results from Chapters 3 and 4 with a view of assessing whether modelling using piecewise linear regression has any value addition over the random coefficient model. In Chapter 6, kernel density based clustering is used to develop a similarity matrix that can be used to decide which genotypes can be mixed optimally, that is, regarding their processing requirements and conditions. In Chapter 7 a joint modelling approach is used to better understand the evolution of six chemical properties (variables) together over the six processing stages. The correlations of the evolutions for the seven genotypes on six chemical properties were calculated. The seventh chemical property (xylose) could not be estimated jointly with the other variables as the estimating procedures failed to converge. Joint modelling helps in the understanding of the joint evolutionary behaviour of the chemical properties when considered together. Chapter 8 summarises the various findings of the study into a consolidated study outcome with possible extensions suggested.

Chapter 2

Description of the Data and Exploratory Data Analysis

2.1. The Pulping Process

In order to understand the nature of the data, it is necessary to understand the chemical pulping process in detail. In a pulping process, wood is converted into fibres either mechanically, thermally, chemically or through a combination of these techniques (Karlsson, 2006). Chemical delignification, an important process during pulping, includes all processes resulting in partial or total removal of lignin from wood by the action of suitable chemicals (Gierer, 1985). The lignin macromolecule is depolymerised through the cleavage of the ether linkages to become dissolved in the pulping liquor. The α-hydroxyl and α-ether groups are readily cleaved under simultaneous formation of benzilium ions (Funaoka et al, 1991). The cleavage of the open α-aryl ether linkages represents the fragmentation of lignin during acid sulphite pulping. The benzilium ions are sulphonated by attack of hydrated sulphur dioxide or bi-sulphite ions, resulting in the increased hydrophylic nature of the lignin molecule. The extent of delignification depends on the degree of sulphonation as well as the depolymerisation (Funaoka et al, 1991). The aim of chemical pulping is to break down the lignin bonds between the fibres using chemicals and heat, enabling easy removal by washing, whilst not destroying the cellulose and hemicellulose components. The removed lignin is a by-product that can be used in water treatment, dye manufacture, agricultural chemicals and in road construction (Sundstrom et al, 1983). Different wood species/genotypes have different levels of lignin content and those species/genotypes that contain more lignin would require more reagents to extract the lignin from the cellulose (Casey, 1983). This means that different wood species/genotypes have different lignin extraction behaviour as they go through the chemical processing stages and it is of interest to investigate this behaviour. The wood species/genotypes with similar physical and chemical characteristics would naturally be put into the same class and may be mixed during processing if larger processing quantities are needed and one genotype falls short of required quantitities.

2.2. The Bleaching Process

The laboratory bleaching sequence was a scaled down version of the commercial process. The results obtained for the viscosity and lignin content (K-number) at the oxygen delignification (O) stage were used to adjust the bleaching conditions. The chemical pulping and bleaching process considered here consists of six stages as indicated in Table 2.1 below. The first stage of the process will be called stage-1, i.e., the stage where wood is acid bi-sulphite pulped into the raw pulp for the bleaching stages.

Table 2.1. Numerical codes for processing stages

Stage	Process	Description
1	Wood to Raw Pulp	Delignification
2	O	Delignification
3	D_1	Brightness
4	E_0	Extract hemicelluloses and solubilise lignin degradation products
5	D_2	Brightness
6	P	Brightness and residual hemicellulose removal

The aim of adjusting the bleaching conditions was to produce dissolving pulp that conformed to the quality control parameters for α-cellulose, viscosity, copper number, glucose (%) and xylose (%) prescribed commercially for the 96α dissolving pulp grade. While it would be reasonable to consider the correlated chemical properties using multivariate techniques, this study looked at a single chemical property individually with the aim of modelling and comparing the behaviour of wood species/genotypes on the same chemical property.

2.3. The Data Collected from the Chemical Process

The wood species/genotypes analysed in this study are EDunnii, EGrandis, ENitens, ESmithii and the Eucalyptus clones GCG, GUA and GUW. The variable species/genotype is a fixed effect with seven levels, namely the seven genotypes which are known beforehand. The observation units are the pulp samples taken from pulped wood species/genotypes.

Trees were randomly selected from each of the seven species/genotypes, chipped and the raw pulp produced through acid bi-sulphite pulping. Independent samples were then taken from the raw pulp and processed. From each sample, measurements of various chemical properties were recorded at the six processing stages described in Table 2.1 above, and are shown in Figure 2.1 below. The samples were processed using three different bleaching conditions coded as A, B and C. Bleaching condition A is a set of the original bleaching conditions, whereas bleaching conditions B and C are revised sets of bleaching conditions specially set to 'fine tune' non-conforming final pulps. If the chemical properties of the final product do not fall within prescribed limits then the product will not be put on the market. This, in a way, produced a controlled response variable especially at the final stage of production. The bleaching conditions were found not to be significantly different hence they are not a prominent part of this study. The pulp samples are random effects as trees are chosen at random from a large number of possible trees.

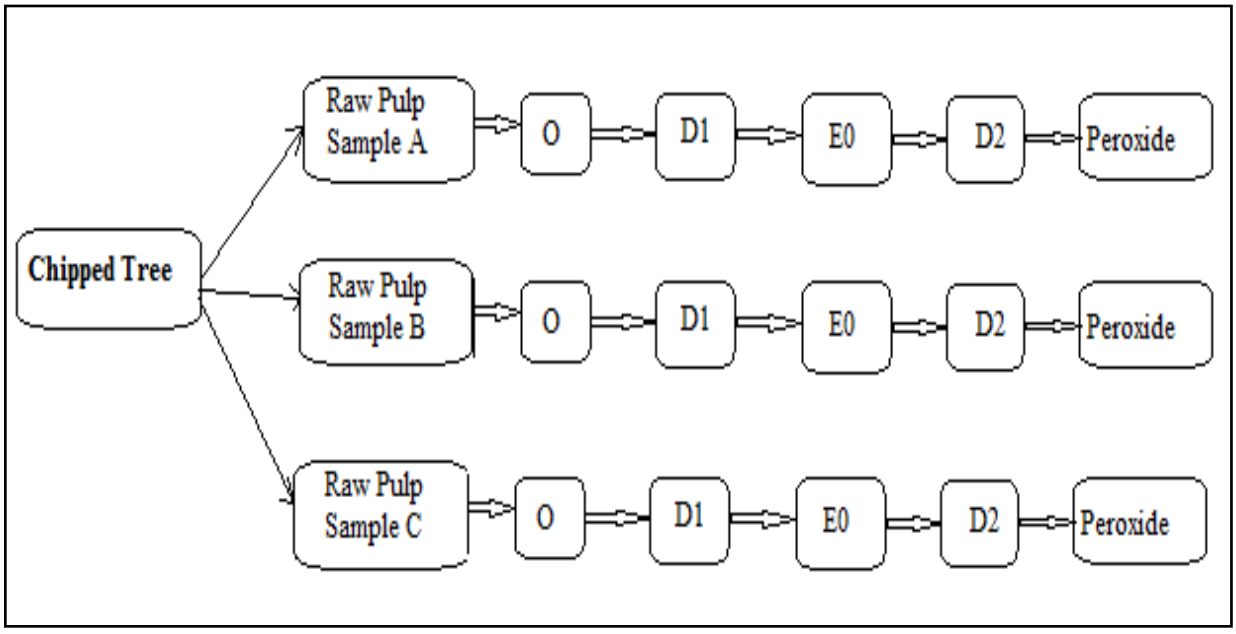


Figure 2.1. Processing stages and the pulp samples

The six stages in the chemical process fall under three sub-processes, namely, delignification, bleaching and finishing and these were carried out under laboratory conditions as described below

2.3.1. Delignification: Acid bi-sulphite pulping

The cooking liquor was prepared from acid bisulphite by bubbling SO_2 MgO slurry and circulated in the digester with wood chips. Temperature was ramped to 140°C and maintained for a period of time. The pressure in the digester was kept at 8.5 bars during the cooking process. At the end of the cooking period, the reaction mixture was allowed to cool down to room temperature. After pulping, an oxygen delignification step was included in a rotating digester. Pulp charge was 800 g (oven dry); consistency 11%; temperature 100°C; time at 100°C = 80 min (96α pulp).

2.3.2. Laboratory bleaching and finishing

The oxygen delignified pulp samples were bleached to target 96α pulp grade using the following four stage bleaching process: D_1 stage (ClO_2 treatment), E stage (NaOH treatment), D_2 stage (ClO_2 treatment), and a peroxide stage. From these processes, wet chemistry analysis variables were obtained as described in Section 2.3.3.

2.3.3. Wet Chemistry analysis – Chemical properties (Variables)

In this study, the following quality control parameters were measured during each step of chemical processing processing:

2.3.3.1. Cellulose content

Low molecular weight carbohydrates (hemicellulose and degraded cellulose) can be extracted from pulp samples using sodium hydroxide. Solubility of a pulp in alkali thus provides information on the degradation of cellulose and loss or retention of hemicellulose during the pulping and bleaching processes. Thus, it gives an indication of the amount of degraded cellulose/short chain glucan and hemicellulose present in the pulp. S_{10} (%) and S_{18} (%) indicate the proportions of low molecular weight carbohydrates that are soluble in 10% and 18% sodium hydroxide, respectively. The former alkali solubility gives an indication of the total extractable material, that is, degraded cellulose/short chain glucan and hemicellulose content in a pulp sample while the latter alkali solubility gives an indication of the total hemicellulose content of the pulp sample and is also known as the percentage gamma (γ %) cellulose content of pulp samples.

The quantity of degraded cellulose/short chain glucan, also known as percentage Beta (β %) cellulose, was determined by the difference in S_{10} (%) and S_{18} (%) alkali solubilities, that is,

Degraded cellulose/short chain glucan = S_{10} (%) – S_{18} (%).

The α-cellulose content is given by the following equation

$$\alpha\text{-cellulose} = 100 - \left(\frac{S_{10}\,\% \; + S_{18}\,\%}{2} \right)$$

S_{10} (%) and S_{18} (%) alkali solubilities were determined according to TAPPI method T235 OM-60 (Tappi T235 OM-60). The principle of the method is based on the extraction of carbohydrates with sodium hydroxide followed by oxidation with potassium dichromate. The procedure for S_{10} (%) alkali solubilities determination is as follows: 1.6g of the pulp sample is placed in 100 mL of 10 % sodium hydroxide (18 % sodium hydroxide for S_{18} (%) determination). The pulp and solution are stirred for a period of 3 minutes and thereafter left at 20ºC for a period of an hour. The pulp sample is filtered under vacuum using a sintered glass crucible (G3). Ten millilitres of 0.4N potassium dichromate and 30 mL of concentrated sulphuric acid are added to 10 mL of the filtrate. Thereafter 500 mL of deionised water is added and the solution is cooled. Approximately 20 mL of 10% potassium iodide is added to the cool solution and 5 minutes thereafter the solution is titrated with 0.1N sodium thiosulphate. A blank, without the pulp sample, is also titrated to give the blank titre. The alkali solubility is given by the following equation:

$$\text{Alkali solubility} = \frac{(\text{Blank titre} - \text{Sample titre}) \times 0.685\%}{\text{Weight of pulp sample}}$$

2.3.3.2. Viscosity

The viscosity of a pulp sample provides an estimate of the degree of polymerisation (DP) of the cellulose chain. Viscosity determination of pulp is one of the most informative procedures that is carried out to characterise a polymer, i.e., this test gives an indication of the degree of degradation (decrease in molecular weight of the polymer, i.e. cellulose) resulting from the pulping and bleaching processes. The viscosity measure involves dispersing 1g of dissolving pulp sample (cellulose I) in a mixture of (15 mL) sodium hydroxide and (80 mL) cuprammonium solution (concentration of ammonia 166 g/L and concentration of copper sulphate 94 g/L) for a period of 1 hour. The dispersed cellulose I is allowed to equilibrate at 20ºC for one

hour and is then siphoned into an Ostwald viscometer. The time taken for it to flow between two measured points is recorded and the viscosity is calculated using the specific viscometer coefficient at the corresponding temperature according to a TAPPI method (Tappi T230, Accessed 15 January 2013).

2.3.3.3. Lignin content (k-number)

The permanganate number (k-number method) was used to assess the lignin content after each stage of processing. The principle of the method is based on the direct oxidation of lignin in pulp by standard potassium permanganate and back-titrating the excess permanganate with ferrous ammonium sulphate (Mohr's salt) standard solution (Tappi UM251, Accessed 15 January 2013; Tasman and Berzins, 1957). The procedure for permanganate number determination is as follows: Approximately 20 mL of 10% sulphuric acid and 180 mL of water is added to 1 g of pulp sample in a conical flask. The mixture is then stirred using a magnetic stirrer. Twenty five millilitres of 0.1N potassium permanganate is added and after 3 minutes 25 mL of 0.1N ferrous ammonium sulphate is added followed by 10 drops of N-phenyl anthranilic acid indicator. The excess is back titrated with 0.1N potassium permanganate. A blank is also carried out with the exception of the pulp sample. The following calculation is used for permanganate number determination:

Permanganate number = (Sample titre – Blank titre) x 0.355.

2.3.3.4. Copper number (Cu number)

Pulping and bleaching is known to affect cellulose structure by the generation of oxidised positions and subsequent chain cleavage in pulp samples (Röhrling et al, 2002). The copper number gives an indication of the reducing end groups in a pulp sample. The copper number is a measure of the reducing properties of the pulp and is defined as the number of grams of metallic copper reduced from the cupric (Cu++) to cuprous (Cu+) state in alkaline solution by 100g cellulose under standard conditions. The copper number is inversely proportional to the viscosity of the pulp samples, that is, with a decrease in viscosity there is increased chain cleavage and hence more reducing end groups. The copper number also serves as an index of reducing impurities in pulp, such as oxycellulose, hydrocellulose, lignin and monosaccharides which possess reducing power. The procedure for determining copper number is as follows: 2.5 g of disintegrated pulp is mixed with a carbonate/bicarbonate (2.6/1, w/w)

and 0.4N copper sulphate solution (95/5, v/v) for exactly 3 hours. Thereafter the pulp is filtered and washed with 5% sodium carbonate followed by hot deionised water. Cuprous acid is dissolved by treating the cellulose on the filter with 45 mL of 0.2N ferric ammonium sulphate. This is left for 10 minutes then filtered off. The pulp is then washed with 250 mL of 2N sulphuric acid. The filtrate is then titrated with 0.04N $KMnO_4$. The blank is subtracted from the titre value to yield the number of grams of reduced copper in the pulp sample (Tappi T430 OM 94, Accessed 15 January 2013).

2.3.3.5. Glucose and xylose

The polysaccharides were measured after their conversion to monosaccharides (glucose and xylose) via a two-step hydrolysis procedure with 72% sulphuric acid. The first step in the hydrolysis process is the addition of 3 mL of sulphuric acid to 0.2g of oven dried pulp in a test tube with stirring. The contents of the test tube are then quantitatively transferred into a Schott bottle with 84 mL of water. The second step in the hydrolysis process involves placing the Schott bottle in an autoclave set at a temperature of 121ºC and pressure of 103 kPa for 1 hour. The contents are then allowed to cool and then filtered using a 0.45µm filter. The filtrate is then transferred to a 200mL volumetric flask and diluted to the mark. 50µl of the sample is placed in a vial and diluted with 500µl of water. Twenty microlitres of 1mg/ml fucose (internal standard) is added using the autosampler. The monosaccharide constituents (glucose, mannose, xylose, arabinose etc.) were analysed using high performance liquid chromatography coupled with pulsed amperometric detection (Davis, 1998). Reference standards of glucose and xylose was prepared. The standards were treated in the same way as the sample and analysed using high performance liquid chromatography coupled with pulsed amperometric detection. The concentrations of the monosaccharide constituents were obtained from the calibration curves of the standards.

The ideal levels of the measurements used to calculate the variables discussed in this study are shown in Table 2.2 below. Anything outside the ranges outlined would not meet the final product requirements.

Table 2.2. Ideal pulp characteristics for the 96α pulp

Final Pulp Characteristic	Ideal levels
Viscosity (cP)	28 - 35
Copper Number	0.43 - 0.54
S_{10}	6.4 - 7.0
S_{18}	2.7 - 3.3
S_{10}-S_{18}	3.7
α-cellulose	>95.3
K-number	0.25

2.4. Exploratory Data Analysis

Exploratory data analysis methods are useful tools that can be used on observed data to obtain an insight on what would be obtained when fitting an implicit or explicit statistical model (Gelman, 2004). In this section the basic properties and layout of the data are explored with a view on obtaining a pointer into the appropriate statistical tools to use for further analysis. Issues of correlation patterns, means and other data descriptors are explored in order to check which statistical models will be applicable to the data.

2.4.1. Theoretical aspects of Profile Plots and associated smoothing methods

To visualise the overall evolution of the pulp chemical properties (variables) over the processing stages (time), profile plots are used. It would have been easy to just plot the values of the chemical properties over time and join the points with straight lines but smoothed plots are much more appealing and outline the general movement of the response variables with the predictor(s) which is time or processing stage in this case. The profile plots are smoothed time plots of the chemical properties over time. The smoothing is done through the use of spline functions as discussed by several authors that include Craven and Wahba (1979) and Brumback and Rice (1998). Smoothed profile plots have also been referred to as LOESS curves (also called LOWESS curves).

2.4.1.1. Smoothing Spline based Profile Plots

Hastie and Tibshirani (1986) discussed in detail the use of splines to smooth data in the context of generalised additive models. In basic principles the smoothing function for the profile plots is a scatterplot smoother (Hastie and Tribshirani, 1986) such as the local average estimate defined as:

$$\hat{s}(x_i) = \text{Average}_{j \in N_i}\{y_j\}. \quad (2.1)$$

where N_i is a neighbourhood of x_i (i.e. a set of observations whose x values are in the neighbourhood of x_i where in our context x_i is time or processing stage. Another way to estimate $s(x_i)$ is to use a polynomial basis which is a space of functions from which $s(x_i)$ is derived. The smoothing function is then presented as weighted sums of the basis functions so that

$$s(x) = \sum_{k=1}^{q} \beta_k \, [b_k(X)] \quad (2.2)$$

where q is the maximum number of basis functions available and $b_k(x)$ is the k^{th} such function. Suppose that, as is the case with our smoothing problem, there is only one covariate x, then the smoothed value can be expressed as

$$s(x_i) = E(Y_i|x_i) = \mu_i \quad (2.3)$$

Or equivalently

$$y_i = s(x_i) + e_i \quad (2.4)$$

where $e_i \sim \text{i.i.d.} N(0, \sigma^2)$ and y_i is the observed response variable. Wood (2006) discussed in detail, possible forms of the smoothing function that include a polynomial basis and a cubic spline basis. Suppose that the smoothing function $s(x)$ is a polynomial of order q so that the space of polynomials of order q or below contains $s(x)$, the basis for this function space is $b_0(x) = 1, b_1(x) = x, b_2(x) = x^2, \ldots, b_q(x) = x^q$. A linear combination of these basis functions can then be used to get the smooth function

$$s(x) = \beta_0 + \sum_{k=1}^{q} \beta_k \, x^{k-1} = \beta_0 + \beta_1 \, x + \beta_2 \, x^2 + \cdots + \beta_q \, x^q. \quad (2.5)$$

If the terms x, x_2, …, x_q , are considered as covariates then model (2.5) is a normal regression model whose parameters can be estimated using the usual methods in generalised linear models. The profile plot will basically be the plot of $s(x_i)$ on x_i.

2.4.1.2. The cubic spline and penalised regression

A cubic spline consists of piecewise third degree polynomials joined at some n-2 knot points, where n is the number of data points. The polynomials must be continuous at the knot points up to the second derivative. A cubic spline representation of model (2.5) is generally of the form

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{i=1}^{r} \beta_{i+3} (x - k_i)_+^3 \qquad (2.6)$$

where $(x)_+ = x$ if $x>0$, 0 otherwise, k_i is the i^{th} knot and ($r=n-2$) is the total number of knots. Equation (2.6) can be written as a linear combination of $r+4$ basis functions to obtain a cubic polynomial with r+4 parameters which can be presented in matrix form as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. This allows the parameters to be estimated as the usual regression parameters.

The utilisation of cubic splines, or any other polynomial basis, is not truly nonparametric as the choices of the knots are basically parametric choices and the way they are chosen affects the fit of the model to the data.

The knot selection problem can be avoided by finding the smoothing function that minimises

$$\sum_{i=1}^{n} \{y_i - s(x_i)\}^2 + \lambda \int \{s''(x)\}^2 \, dx \qquad (2.7)$$

which can also be written as

$$\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \int \{s''(x)\}^2 dx \qquad (2.8)$$

The term $\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2$ captures the fit of the data to the smoothing function while the term $\lambda \int \{s''(x)\}^2 dx$ penalises the smooth function. The integral is over all values of x covered in the data points. The penalty function strikes a balance between smoothness and overfitting or 'wiggliness' (Wood, 2006). If the smoothing function is a straight line then $s''(x) = 0$ for all x. The parameter λ is the smoothing parameter which controls the trade-off between 'wiggliness' and smoothness. When $\lambda=0$ there will be no penalty on $s(x)$ resulting in an non penalised model since the second part of (2.8) disappears. Such an unpenalised model is very wiggly or it will be a mere interpolation of the data. When $\lambda \longrightarrow \infty$ the model will be heavily penalised rendering curvature impossible and this results in the fitting of a linear regression model.

Equation (2.7) or (2.8) has a unique minimizer which is a natural cubic spline with knots at unique values of the covariate (Hastie and Tibshirani, 1990).

If the observed values of the covariate are $x_1 < x_2 < \cdots < x_n$ then the natural cubic spline is the smoothest function that minimizes $\int_{x_1}^{x_n} \{s''(x)\}^2 \, dx$ amongst all functions that are continuous on all data points and are twice differentiable (Wood, 2006). Natural cubic spline smoothing has the following properties:

(i) $s(x)$ exists and $s(x_k -) = s(x_k +)$ for *k*=2,3,…,*n*-1,
(ii) $s'(x)$ exists and $s'(x_k -) = s'(x_k +)$ for *k*=2,3,…,*n*-1,
(iii) $s''(x)$ exists and $s''(x_k -) = s''(x_k +)$ for *k*=2,3,…,*n*-1,
(iv) $s''(x_1) = s''(x_n) = 0$, that is, the function *s*(*x*) is linear at the end points and in regions outside the observed data.

The derivation of the natural cubic spline is well documented in literature, for example see Mathews and Fink (2004) and Rorres and Howard (1984).

Having decided on the spline basis to use, it is now left to estimate the parameter vector $\boldsymbol{\beta}$ in equation (2.8). Let $\{b_j(x)\}_{j=1}^{q}$ be the set of the natural cubic spline basis functions and $\boldsymbol{X}_{q\times q}$ denote the design matrix consisting of the basis functions evaluated at the observed values of the covariate:

(i) $X_{ij} = b_j(x_i)$, (i.e. the (i,j)th element of $\boldsymbol{X}$)
(ii) $s(x) = \sum_{j=1}^{q} \beta_j \left[b_j(x)\right]$
(iii) $s(x) = \boldsymbol{X\beta}$

Using (i)-(iii) above, equation (2.8) can now be expressed as

$$(\boldsymbol{y} - \boldsymbol{X\beta})^{\mathrm{T}}(y - \boldsymbol{X\beta}) + \lambda \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S\beta} \qquad (2.9)$$

Where the (*k,j*)th element of $\boldsymbol{S}$ is $S_{jk} = \int b_j''(x) b_k''(x) dx$. The solution for minimizing (2.9), by differentiating with respect to $\boldsymbol{\beta}$ and equating the differential to zero, can be found to be

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{S})^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{Y}. \qquad (2.10)$$

The fitted values of the response vector $\boldsymbol{y}$ are then presented as

$$\begin{aligned}\widehat{\boldsymbol{y}} &= \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{S})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \\ &= \boldsymbol{H}_{\boldsymbol{\lambda}}\boldsymbol{y}\end{aligned} \tag{2.11}$$

where $\boldsymbol{H}_{\boldsymbol{\lambda}} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\lambda}\boldsymbol{S})^{-1}\boldsymbol{X}^{\mathrm{T}}$.

2.4.1.3. Choice of the smoothing parameter

The smoothing parameter (*λ*) needs to be specified before the model parameters are estimated in (2.10). If *λ* is too large the data will be oversmoothed and if *λ* is too low then the data will be under smoothed. The best choice of *λ* is the one that minimises the difference between the true smoother *s*(*x*) and the estimated smoother $\hat{\boldsymbol{s}}(\boldsymbol{x})$, that is, the choice of *λ* must minimises

$$M = \frac{1}{n}\sum_{i=1}^{n}[\hat{s}(x_i) - s(x_i)]^2. \tag{2.12}$$

The true smoother in (2.12), that is $s(x_i)$, is not known but Wahba (1975) showed that the same result can be achieved by ordinary cross validation (OCV). Let $\hat{s}_{\lambda}^{(-i)}(x)$ be the smoothing function (with a particular choice of *λ*) fitted to the data when the *i*th observation is left out. The ordinary cross-validation (OCV) estimate of the prediction error is

$$OCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{s}_{\lambda}^{(-i)}(x_i)\right)^2.$$

The computation of OCV(*λ*) is quite tedious as for every choice of *λ* the parameters of the smoother $\hat{s}_{\lambda}^{(-i)}(x_i)$ have to be computed *n*-times. It can be shown that OCV(*λ*) can be approximated by

$$OCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{s}_{\lambda}(x_i)}{1 - h_{ii}}\right)^2.$$

where h_{ii} is a diagonal element of the hat matrix H. To further reduce computational time, a common value is used for all the h_{ii}'s, for *i*=1, 2,…*n.* Craven and Wahba (1979) suggested the use of the average of the all the diagonal elements of the hat (or influence) matrix H instead of the *n* h_{ii} values. Craven and Wahba's substitution results in the generalised cross validation (GCV) score which is given by

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{s}_{\lambda}(x_i)}{1 - \frac{\mathrm{tr}(\boldsymbol{H})}{n}}\right)^2 = \frac{n\sum_{i=1}^{n}\left(y_i - \hat{s}_{\lambda}(x_i)\right)^2}{[\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H})]^2}$$

It can also be shown that, apart from computational efficiency, the generalised cross validation score minimises E(*M*), where *M* is as defined in equation 2.12. The value of λ that has the minimum GCV score is the one that is chosen as the optimal smoothing parameter. There are many other methods that can be used to select the optimal smoothing parameter like the AIC or REML (Shao, 1997), but the GVC is well suited for complex models (Xiang and Wahba,1996; Zhang et al, 2002). In writing the programs to carry out exploratory data analysis in this study, insights and hints were obtained from the UCLA Statistical Consulting Group website (UCLA, accessed June 13, 2016).

2.4.2. Profile Plots for the seven chemical properties

A discussion of the profile plots of the seven chemical properties covered in this study is presented in this section. The profile plots are presented in Figures 2.2 to 2.8 and are discussed in the sections that follow.

2.4.2.1. Viscosity Plots

The profile plots for viscosity are presented in Figure 2.2. The plots indicate that the seven genotypes have varying numbers of subjects (pulp samples), exhibit different levels of variability across the processing stages and have less defined downward trends. The steepest decline in viscosity seem to occur for GCG and GUA genotypes. GUW seems to have more variation in viscosity readings at the earlier stages of processing than in the final stages. The opposite applies for EDunnii which has higher variability in the last three stages of processing than in the earlier stages. It is noted, for example, that viscosity for EGrandis has less variability across the six processing stages than EDunnii and ESmithii.

The general profiles of the seven genotypes would suggest that linear trends best describe the evolution of viscosity over the six processing stages.

2.4.2.2. Lignin Plots

In Figure 2.3 are the profile plots of lignin readings for the seven genotypes under study. The Lignin measurements exhibit more uniform variability per genotypes across the six processing stages.

The lignin trend seems to be similar for the seven genotypes. The trend could best be described as an exponential decline hence a negative exponential, or piecewise linear trend with negative slopes, could best be used to describe the change in lignin over the processing stages.

2.4.2.3. γ-cellulose Plots

Figure 2.4 shows that the variability of γ-cellulose values seems constant across the processing stages for most genotypes except for EDunnii (Figure 2.4(a)) which had one sample having noticeably higher value at stage 1 than the other samples. The genotype ESmithii, which had six samples, exhibits higher variability at each stage with the variability being constant across all stages.

Generally the γ-cellulose trend could best be described as three-part piecewise linear, although a linear trend could still be attempted if estimation problems are encountered and a more parsimonious model is required.

2.4.2.4. α-cellulose

The most important variable in the whole chemical processing scheme is α-cellulose. The profile plots of α-cellulose in Figure 2.5 show that the variability of α-cellulose values seem to be constant across the six processing stages for most genotypes except for ESmithii (Figure 2.5(c)) which had one sample having the stage 1 value looking more like an outlier. Special care should be taken when dealing with this sample as it starts off having the highest α-cellulose value but ends up with the least value from stage to the finishing stage. It might be necessary to treat this value as an unusual value and use imputation for the stage 1 value for this particular sample.

Generally the α-cellulose trends could best be described as three-part piecewise linear, with slopes that are reversals of those of γ-cellulose. In fact, the two variables (α-cellulose and γ-cellulose) evolve in inversely to each other. A rudimental linear trend could still be attempted if estimation problems are encountered and a more parsimonious model is required.

2.4.2.5. Copper Numbers Plots

Figure 2.6 shows that the sample values for copper numbers of each of the seven genotypes evolve very closely together. Within sample variability is constant across all time points or processing stages. However, there is slightly lower variation at stage 1 than all other stages for all genotypes.

Copper numbers trends also look like they are three-part piecewise linear, with negative slopes. A higher polynomial model could also be attempted although the possibility of non-convergence could be high. A simple linear trend could also be attempted if convergence is not attained with more complex models.

2.4.2.6. Glucose Plots

Glucose and α-cellulose have similar trends (compare Figures 2.5 and 2.7). Although between sample variation cannot be said to be exactly constant, there is no worrying deviation from an assumption of constant variance across all six processing stages or time points. While the evolution of glucose for some of the genotypes could be modelled with three-part piecewise linear models (EDunnii, EGrandis, ESmithiii, ENitens and GCG), two-part piecewise linear models could suffice for others (GUA and GUW). In all cases a parsimonious simple linear trend could still be attempted if convergence problems.

2.4.2.7. Xylose Plots

The profile plots for Xylose are presented in Figure 2.8. The plots show that variability is non-constant for some genotypes, for example, there is bigger variation in stage 5 for ENitens than in the other stages although not to worrying proportions. The genotype ESmithii, which happened to have the most number of samples, shows wider variation but there are no worrying indications that the variation is non-constant across the processing stages. Some of the profile lines criss-cross each other indicating that the samples react differently to the various stages of processing.

In general, all seven genotypes can be modelled by some form of linear trends (piecewise or otherwise). Nonlinear trends can also be seen especially with EGrandis, ENitens, GCG and to some extent GUW. With all the seven variables, whatever models are proposed in subsequent chapters, reference will be made to the profile

plots that have been described in this section. The profile plots will guide the selection of appropriate models in the different methodologies that are presented in the chapters that follow. To investigate the interdependence of the seven variables covered in this study, a correlation analysis is presented in Section 2.4.2 below.

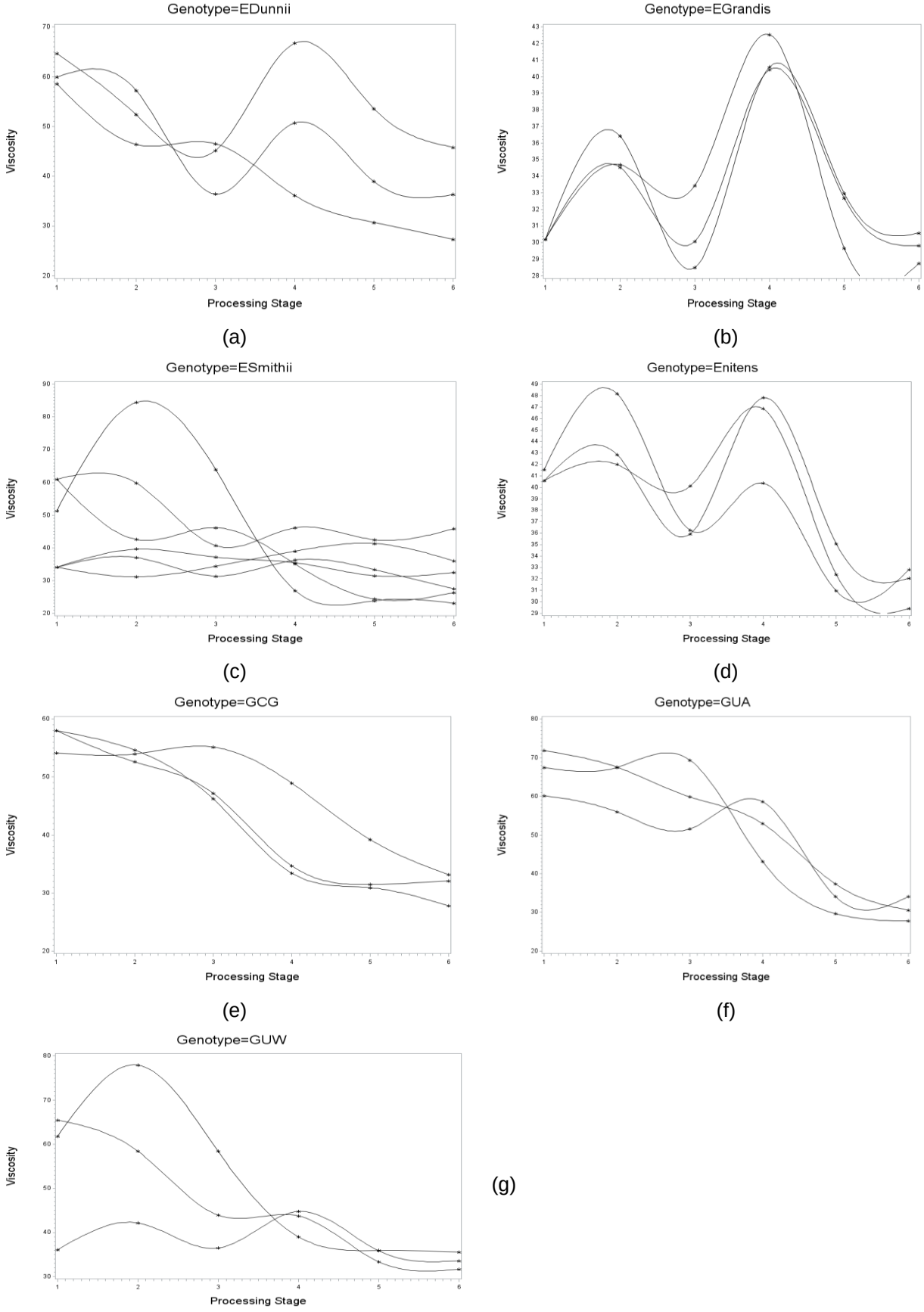


Figure 2.2. Profile plots of Viscosity for all seven genotypes

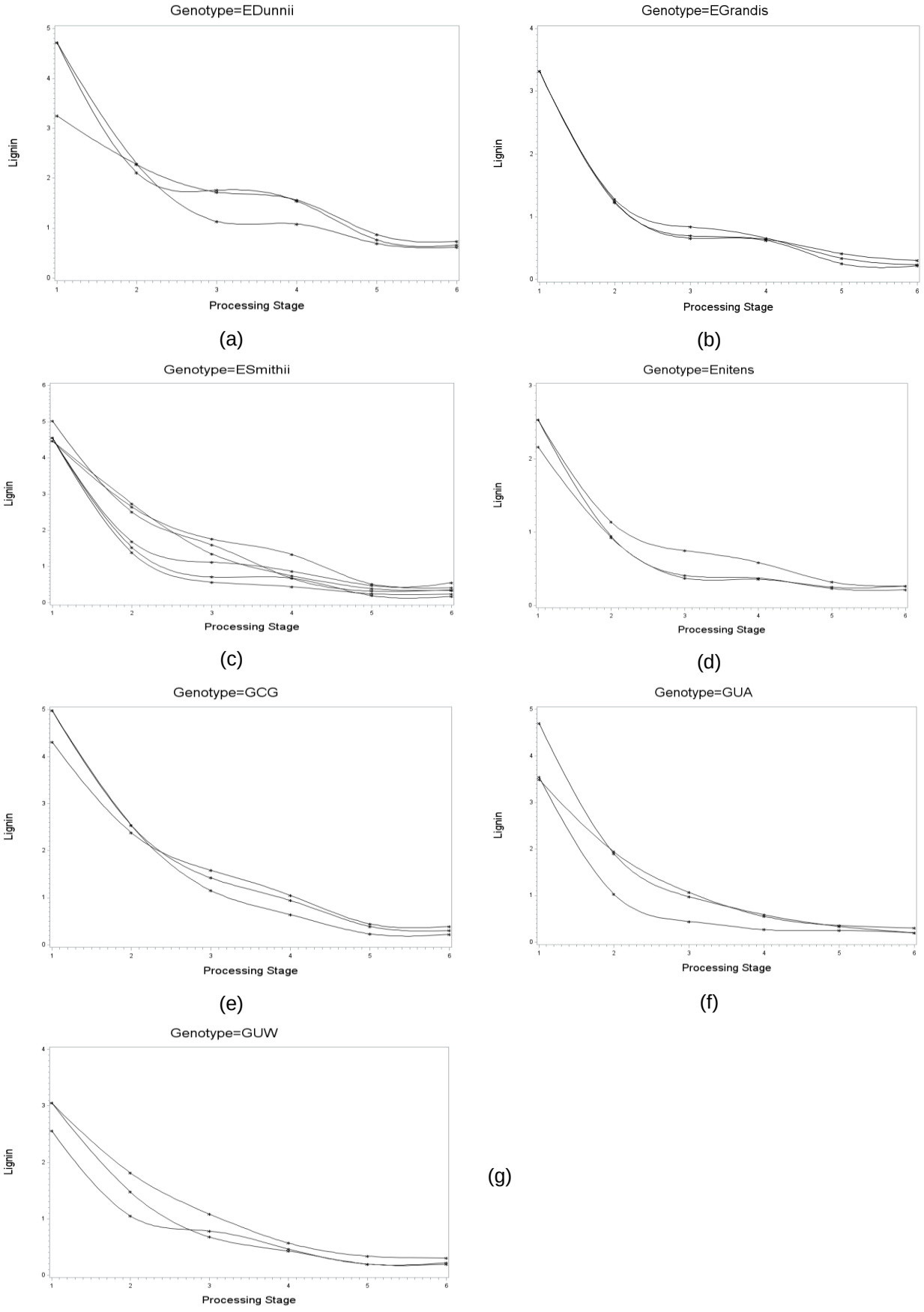


Figure 2.3. Profile plots of Lignin for all seven genotypes

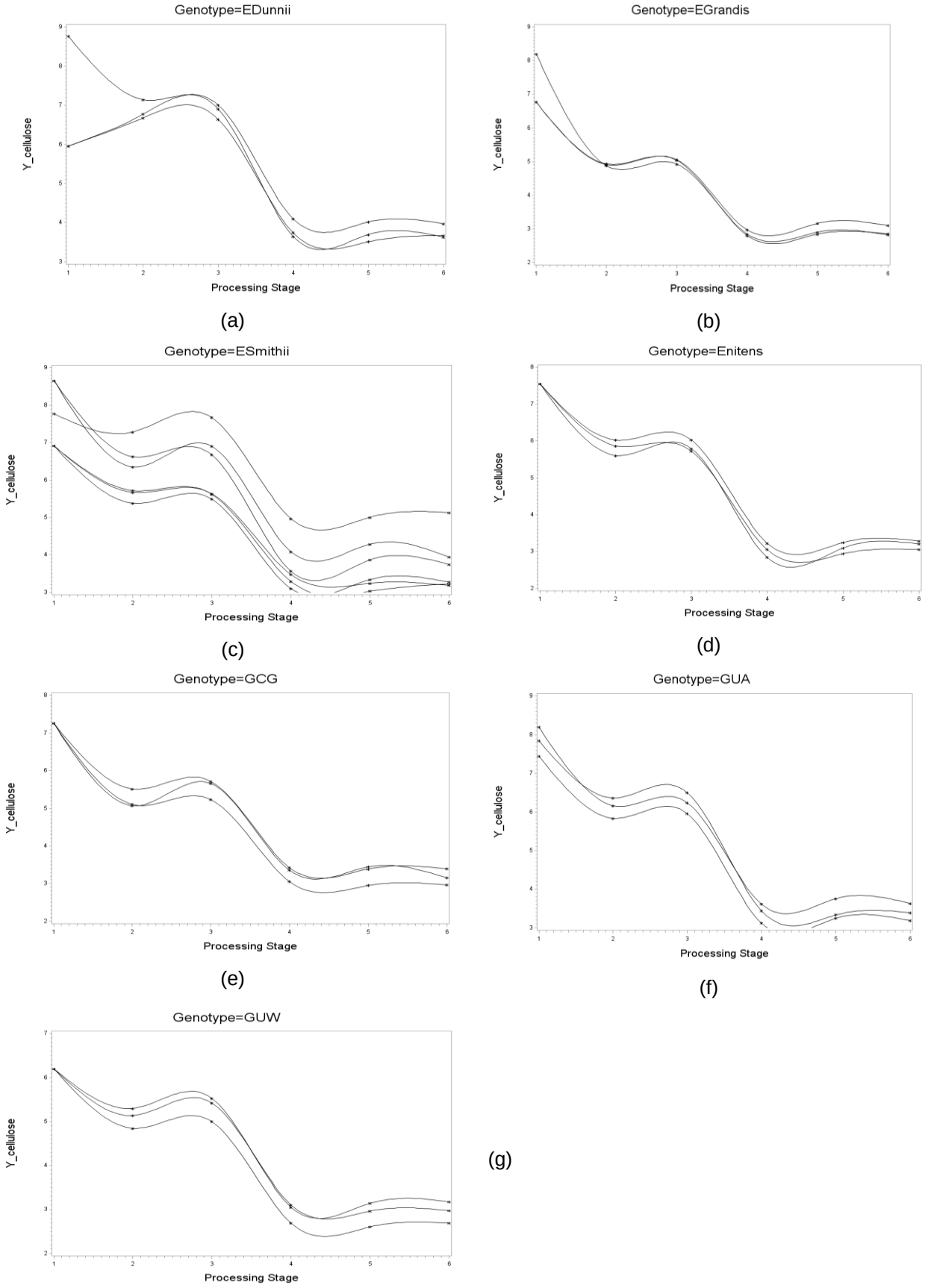


Figure 2.4. Profile plots of γ-cellulose for all seven genotypes

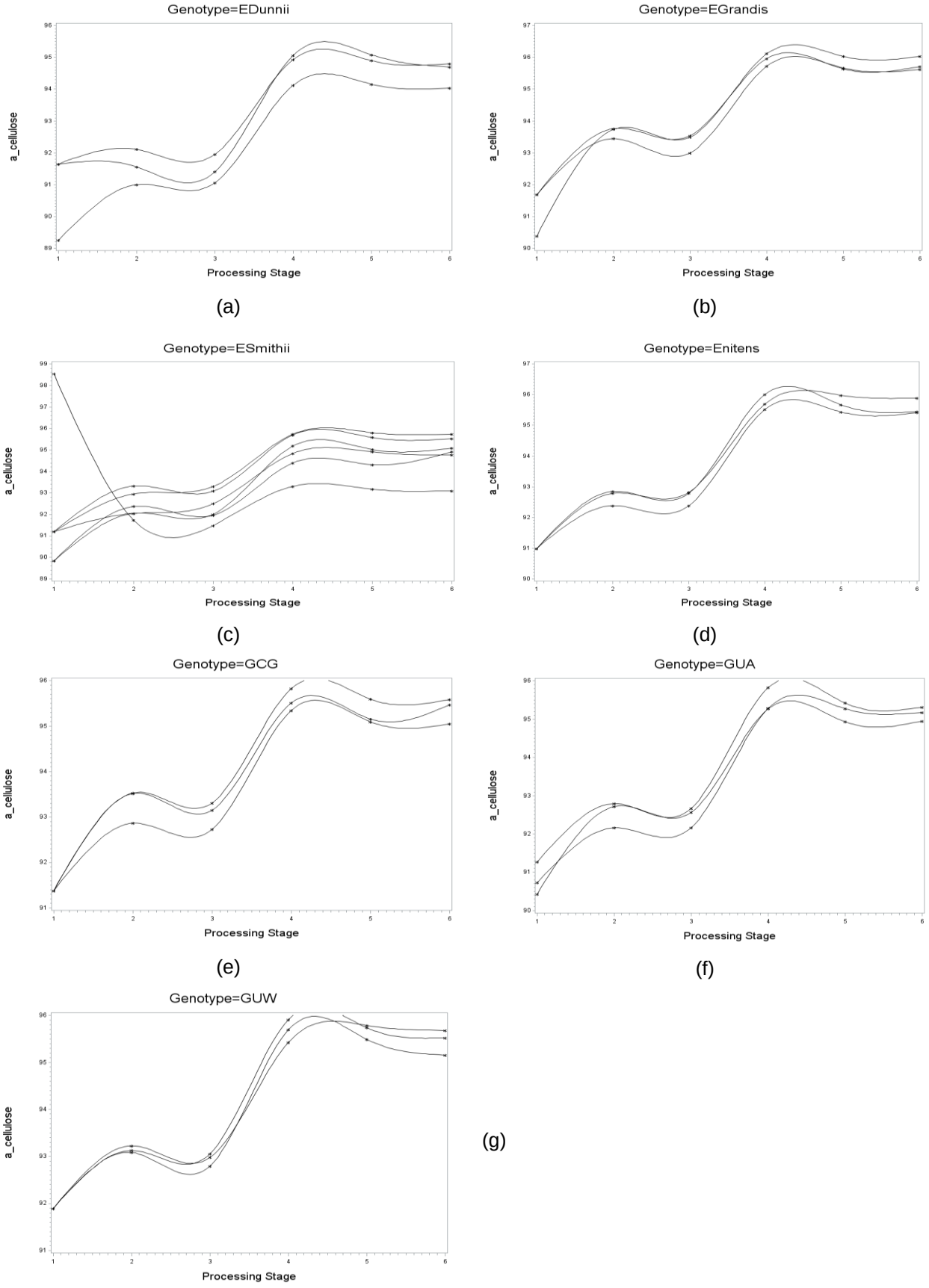


Figure 2.5. Profile plots of α-cellulose for all seven genotypes

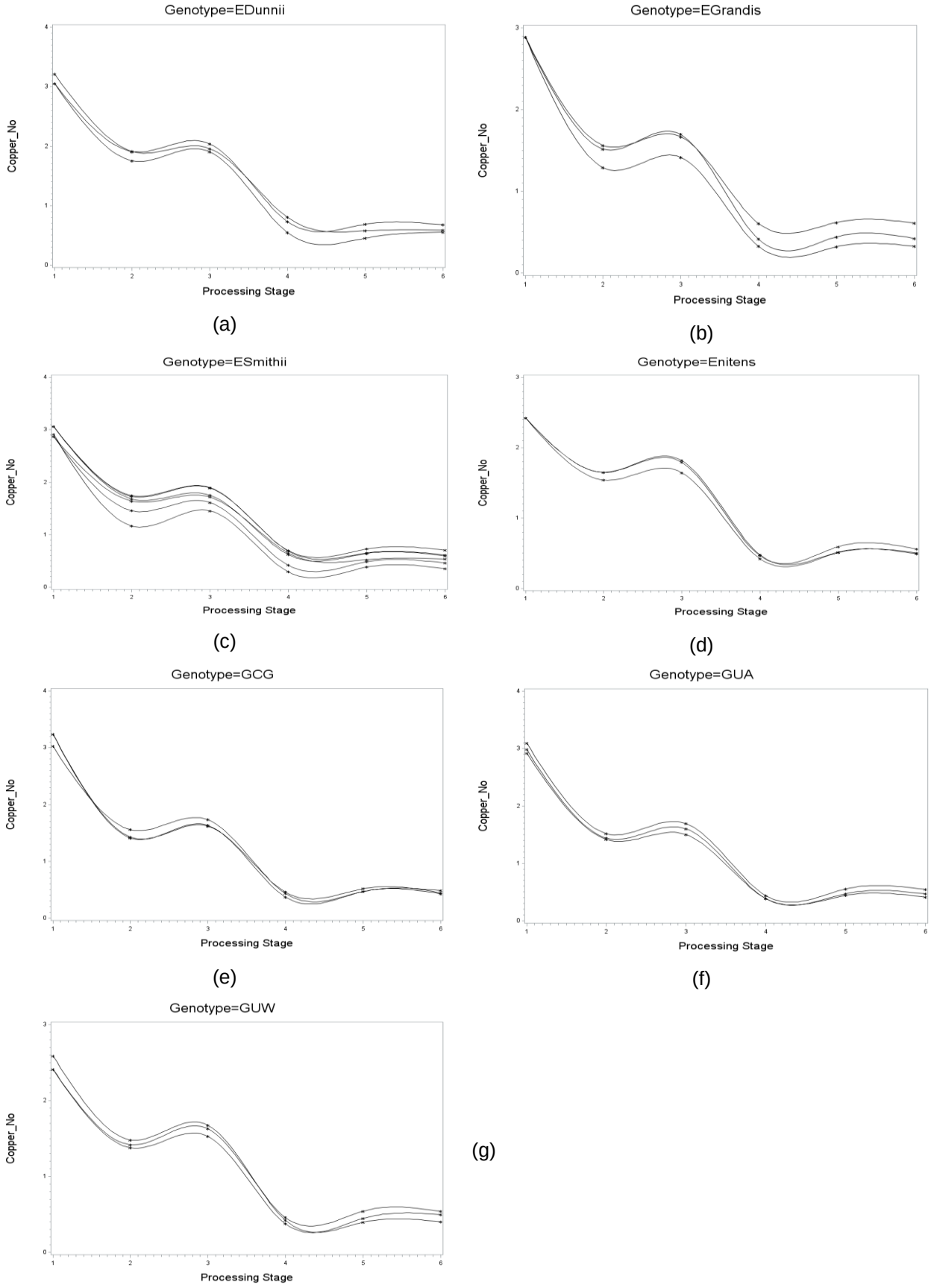


Figure 2.6. Profile plots of Copper numbers for all seven genotypes

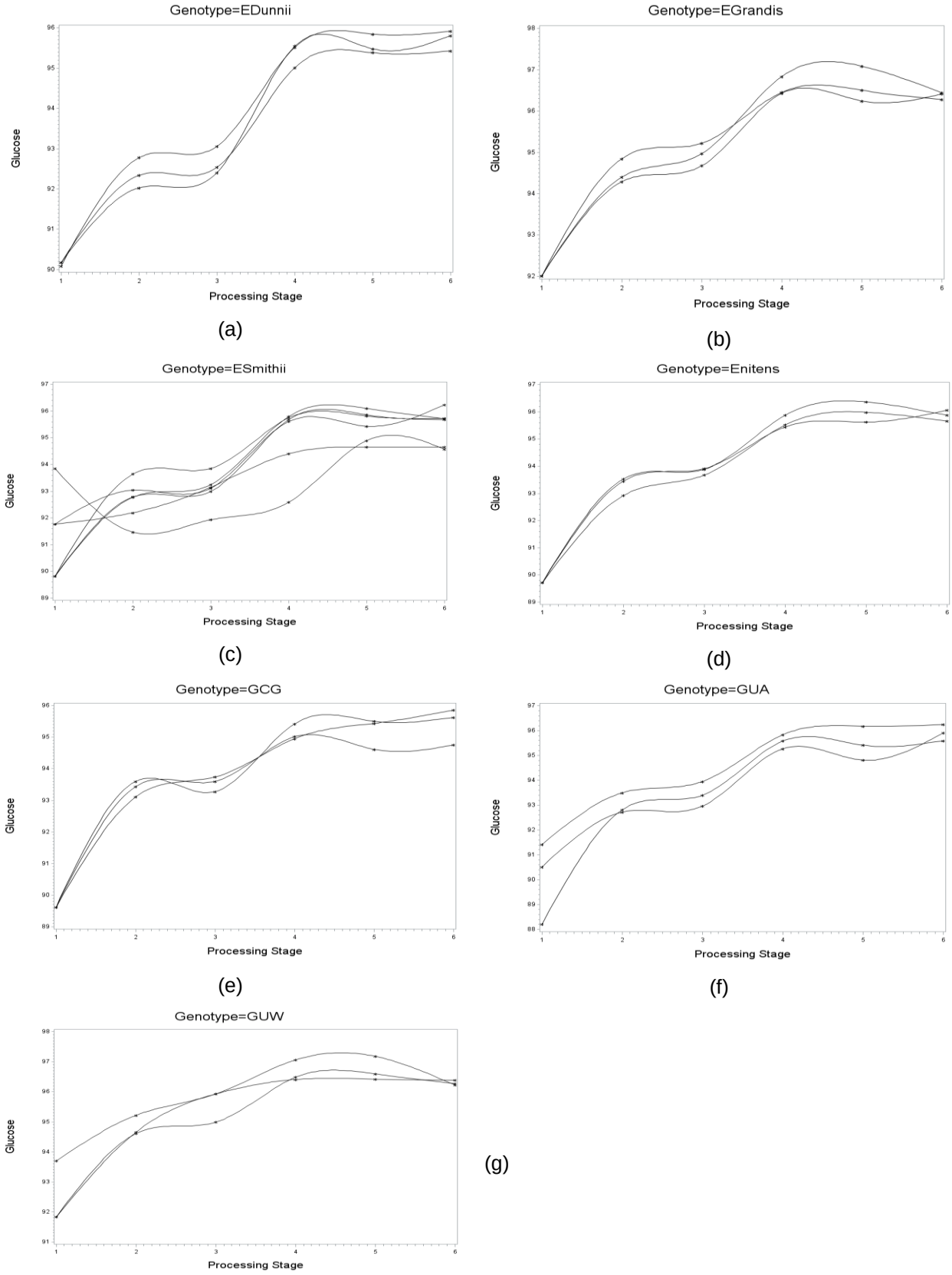


Figure 2.7. Profile plots of Glucose for all seven genotypes

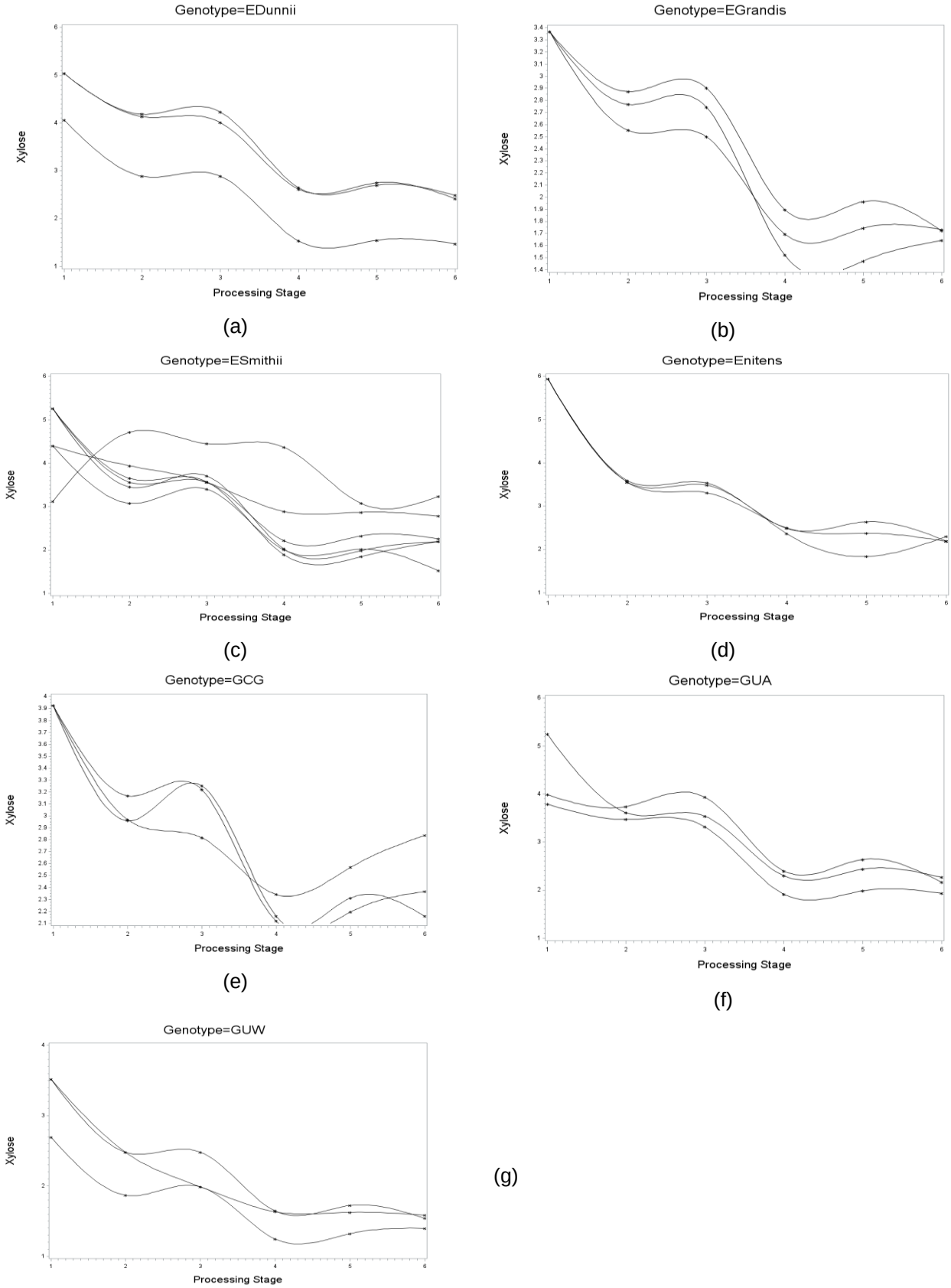


Figure 2.8. Profile plots of Xylose for all seven genotypes

2.4.3. Scatter Plots and Correlation Analysis

Exploration of complex relationships between variables not only requires examination of correlation tables but also visual inspection of scatter plots which provide meaningful insights into associations under investigation (Kulesz et al, accessed 17 June, 2016). Matange and Heath (2011) provide detailed SAS programming procedures to obtain various kinds of exploratory graphs. Most of the results discussed in this section were obtained using adaptations of such procedures. The SGSCATTER procedure in SAS was used to produce scatter plots with confidence ellipses and histograms with normal density curves for the variables covered in the study. An inspection of the histograms indicated that the data did not severely deviate from normality hence procedures that are based on the normal distribution can be applied.

Multi-panel scatter plots were used as they are useful data visualizations which allow for a graphical display of relations between multiple variables in a condensed manner. In multivariate repeated measures the correlations can be realised in two fronts, that is, serial correlation within each variable and pairwise correlations between the variables measured. Serial correlation will measure how values taken on the same individual affect each other sequentially. There have been discussions on the issue of correlation amongst repeated measurements for univariate responses (Hearne et al, 1983; Hamlett et al, 2003) and multivariate cases (Roy, 2006). Hearne et al state that, the robustness of univariate procedures for repeated measures depends on careful consideration of the serial correlation between values observed on the same subject over time. Roy on the other hand estimated the correlation coefficient between two variables with repeated observations on each variable using a linear mixed effects model. In all these discussions, the role of both forms of correlation play a pivotal role in the determination of a valid model for the data.

2.4.3.1. Correlations between chemical property variables

The graphical displays of the between variable correlations are presented in Figure 2.9 below which shows that viscosity has the least correlation to the other six variables as indicated by the scatter plots together with the confidence ellipsoids which are more circular than elongated. The other variables are correlated to varying degrees and directions with some positively and others negatively correlated. The exact degrees of

correlation and significance tests are presented in Table 2.3 below. The results in Table 2.3 show that the viscosity column has the lowest correlation values (maximum of 0.5079) which means that viscosity has the weakest relationship with the other variables. On the other hand, the other variables are highly correlated with each other hence are expected to affect each other's evolutions.

Table 2.3. Correlations between chemical property variables

	Viscosity	Lignin	a_cellulose	Y_cellulose	Copper No	Glucose	Xylose
Viscosity	1						
Lignin	0.5079	1					
a_cellulose	-0.3734	-0.581	1				
Y_cellulose	0.4510	0.724	-0.908	1			
Copper_No	0.4027	0.862	-0.816	0.8951	1		
Glucose	-0.4378	-0.797	0.855	-0.8837	-0.890	1	
Xylose	0.2864	0.610	-0.757	0.7957	0.735	-0.903	1

2.4.3.2. Correlations between processing stages

In data involving repeated measurements on the same subject, it is assumed that serial correlations will exist among sequential observations. Significant correlations among values taken on the same subject at different time points (stages) are an indication that the observations are not independent hence a model that assumes independence will not be appropriate for the data.

Results in Table 2.4 show that viscosity is the least serially correlated variable than the other six, particularly at longer stage lags. With the exception of stages 1 and 2 (correlation=0.735), 1 and 3 (correlation = 0.748), stages 2 and 3 (correlation = 0.844), 4 and 5 (correlation=0.746), 4 and 6 (correlation=0.645) and stages 5 and 6 (correlation=0.883), the rest of stages have very small correlations.

With the other six variables, it can be seen that stage 1 is the least correlated to the other 5 stages of chemical processing. Serial correlations are more vividly illustrated by the panel plots in Figures 2.10 (a), (b), (c) and (d) below. As a general assessment of the nature of correlation, the more elongated the confidence ellipsoid of the plot, the more correlated the stages involved. If on the other hand, the confidence ellipsoid is more circular than elongated, then the correlation between the two stages involved is closer to zero. As far as lignin, γ-cellulose, α-cellulose, copper number, glucose and

xylose are concerned, the stages that follow after stage 1 show strong serial correlations (elongated ellipsoids) which attest to the fact that there is indeed correlation between values measured on the same subject for these variables. These correlations have a bearing on the mixed model that will best fit the data, particularly the choice of the covariance structure.

Table 2.4. Serial correlations for the seven genotypes

Viscosity

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.735	1				
Stage3	0.748	0.844	1			
Stage4	0.300	0.016	0.023	1		
Stage5	0.174	-0.041	-0.140	0.746	1	
Stage6	0.199	0.020	-0.17	0.645	0.883	1

Lignin

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.747	1				
Stage3	0.582	0.882	1			
Stage4	0.430	0.648	0.851	1		
Stage5	0.277	0.504	0.709	0.903	1	
Stage6	0.270	0.542	0.7456	0.908	0.963	1

γ-cellulose

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.381	1				
Stage3	0.386	0.969	1			
Stage4	0.374	0.845	0.902	1		
Stage5	0.472	0.808	0.894	0.953	1	
Stage6	0.392	0.856	0.912	0.961	0.971	1

α-cellulose

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.558	1				
Stage3	0.586	0.958	1			
Stage4	0.725	0.776	0.825	1		
Stage5	0.738	0.709	0.783	0.957	1	
Stage6	0.656	0.783	0.8311	0.926	0.954	1

Copper Number

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.221	1				
Stage3	0.235	0.959	1			
Stage4	0.334	0.900	0.841	1		
Stage5	0.184	0.731	0.725	0.801	1	
Stage6	0.140	0.820	0.805	0.852	0.937	1

Glucose

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.309	1				
Stage3	0.368	0.956	1			
Stage4	0.038	0.829	0.785	1		
Stage5	0.326	0.771	0.773	0.819	1	
Stage6	0.138	0.698	0.6558	0.838	0.770	1

Xylose

	Stage1	Stage2	Stage3	Stage4	Stage5	Stage6
Stage1	1					
Stage2	0.512	1				
Stage3	0.518	0.953	1			
Stage4	0.221	0.854	0.775	1		
Stage5	0.333	0.828	0.779	0.872	1	
Stage6	0.307	0.793	0.715	0.875	0.856	1

2.4.4. The assumption of normality and normality tests

The assumption of normality is important for the application of general linear and allied models. Since the seven chemical variables are expected to have different means at each stage of processing, normality tests are conducted at everey stage. The histograms presented in Figures 2.10 (a) to 2.10(d) show that, generally, the chemical variables exhibit normality and there is no serious departure from normality. Considering the limited number of subjects for which repeated measurements were taken under each genotype, hence stage, it is not possible to have very smooth histograms with a perfect normal outline. The histograms produced under these circumstances suggest that it is safe to assume normality even in the absence of formal normality tests. However, to ascertain the presence of normality in the data, the Kolmogorov–Smirnov test (K-S-t) was conducted. It is a general feature of the data that the means differ by stage hence the Kolmogorov–Smirnov test was conducted to test for normality at each of the six stages as stated earlier, with the results presented in Table 2.5 below.

Table 2.5. Kolmogorov–Smirnov tests of normality for the seven chemical variables at the six stages.

		Chemical Property (Variable)						
Stage		Viscosity	Lignin	Y_cellulose	a_cellulose	Copper_No	Glucose	Xylose
Stage 1	(K-S-t) value	0.213	0.212	0.108	0.160	0.244	0.209	0.155
	p-value	<0.010	<0.010	>0.150	0.113	<0.010	<0.010	0.137
	comment	-	-	Normal	Normal	-	-	Normal
Stage 2	(K-S-t) value	0.135	0.128	0.106	0.128	0.100	0.127	0.133
	p-value	>0.150	>0.150	>0.150	>0.150	>0.150	>0.150	>0.150
	comment	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Stage 3	(K-S-t) value	0.147	0.126	0.156	0.128	0.101	0.172	0.138
	p-value	>0.150	>0.150	0.131	>0.150	>0.150	0.051	>0.150
	comment	Normal	Normal	Normal	Normal	Normal	Normal	Normal
Stage 4	(K-S-t) value	0.106	0.235	0.136	0.539	0.216	0.170	0.145
	p-value	>0.150	<0.010	>0.150	<0.010	<0.010	0.071	>0.150
	comment	Normal	-	Normal	-	-	Normal	Normal
Stage 5	(K-S-t) value	0.144	0.177	0.156	0.539	0.102	0.102	0.121
	p-value	>0.150	0.051	0.132	<0.010	>0.150	>0.150	>0.150
	comment	Normal	Normal	Normal	-	Normal	Normal	Normal
Stage 6	(K-S-t) value	0.147	0.232	0.163	0.539	0.065	0.169	0.169
	p-value	>0.150	<0.010	0.085	<0.010	>0.150	0.082	0.076
	comment	Normal	-	Normal	-	Normal	Normal	Normal

In most of the cases, the data was found to be normally distributed (p-values>0.05), and of the few cases in which normality was not conclusive it is mainly due to the low number of subjects, otherwise the data is generally normally distributed, hence standard general linear models can be applied to the data.

2.4.5. Relevance of the exploratory data analysis

The results discussed in this chapter are of great significance on the techniques that need to be employed to model the data in order to achieve the objectives outlined in Chapter 1. Chapter 3 is based on techniques that cannot assume complete independence amongst all observations. This follows from the fact that results in Section 2.4.2 suggest that there is serial correlation amongst observations taken on the same subject.

The profile plots suggest that the evolutions, over the processing stages (time), of the seven chemical property variables might best be modelled by nonlinear models such as piecewise linear regression models presented in Chapter 4.

The profile plots also showed that the seven genotypes had different trajectories on the seven chemical property variables analysed. Some seem to evolve closer to each other than others. Trying to identify which genotypes evolve with similar patterns is very important as this will lead to genotype classification or clustering. Genotypes that fall within the same clusters are deemed similar hence can be processed together as their processing requirements in terms of chemicals and processing conditions will also be similar.

The strong correlations among the seven chemical property variables also suggest the need for multivariate analysis of the data and this is done using the joint modelling approach in Chapter 6.

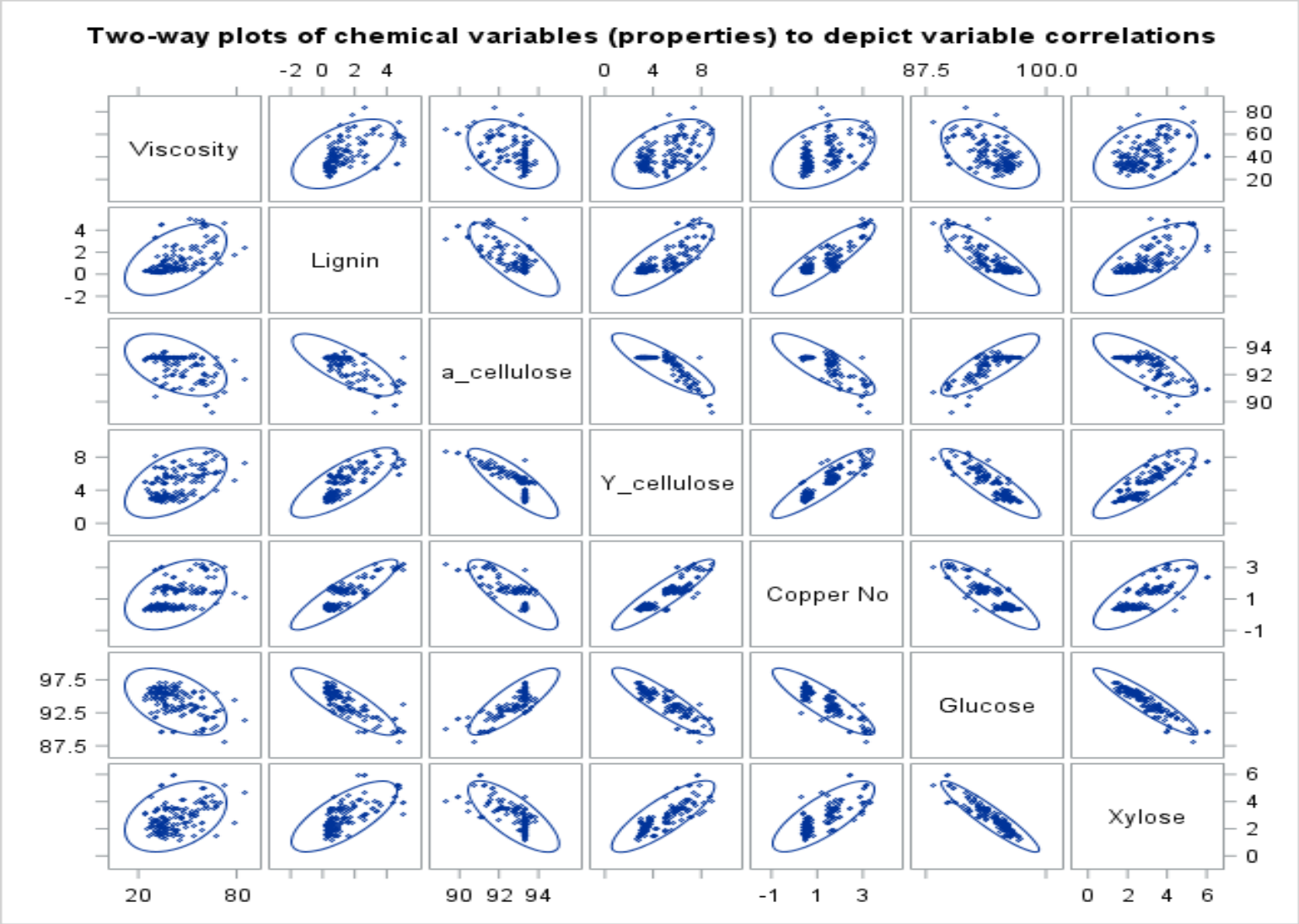
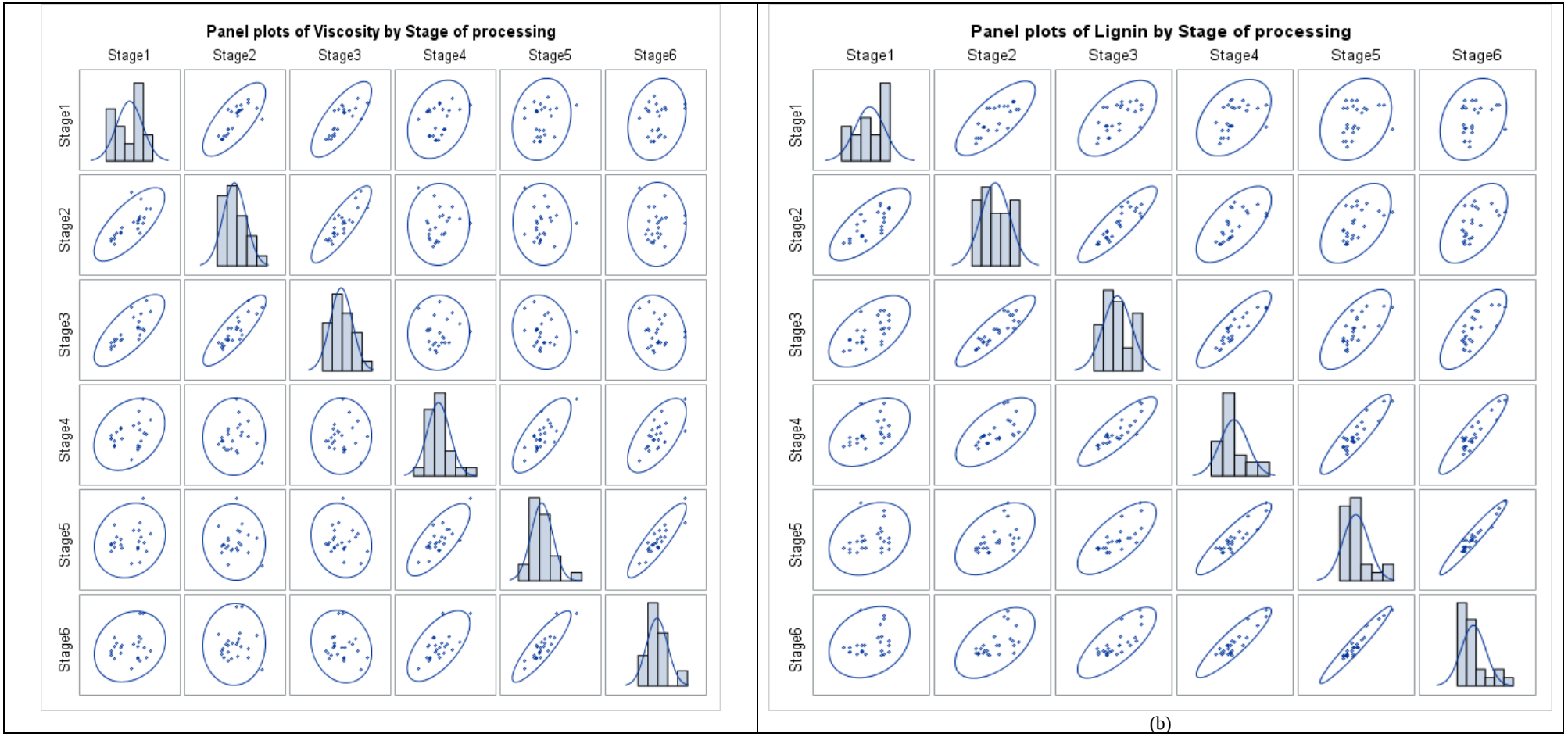
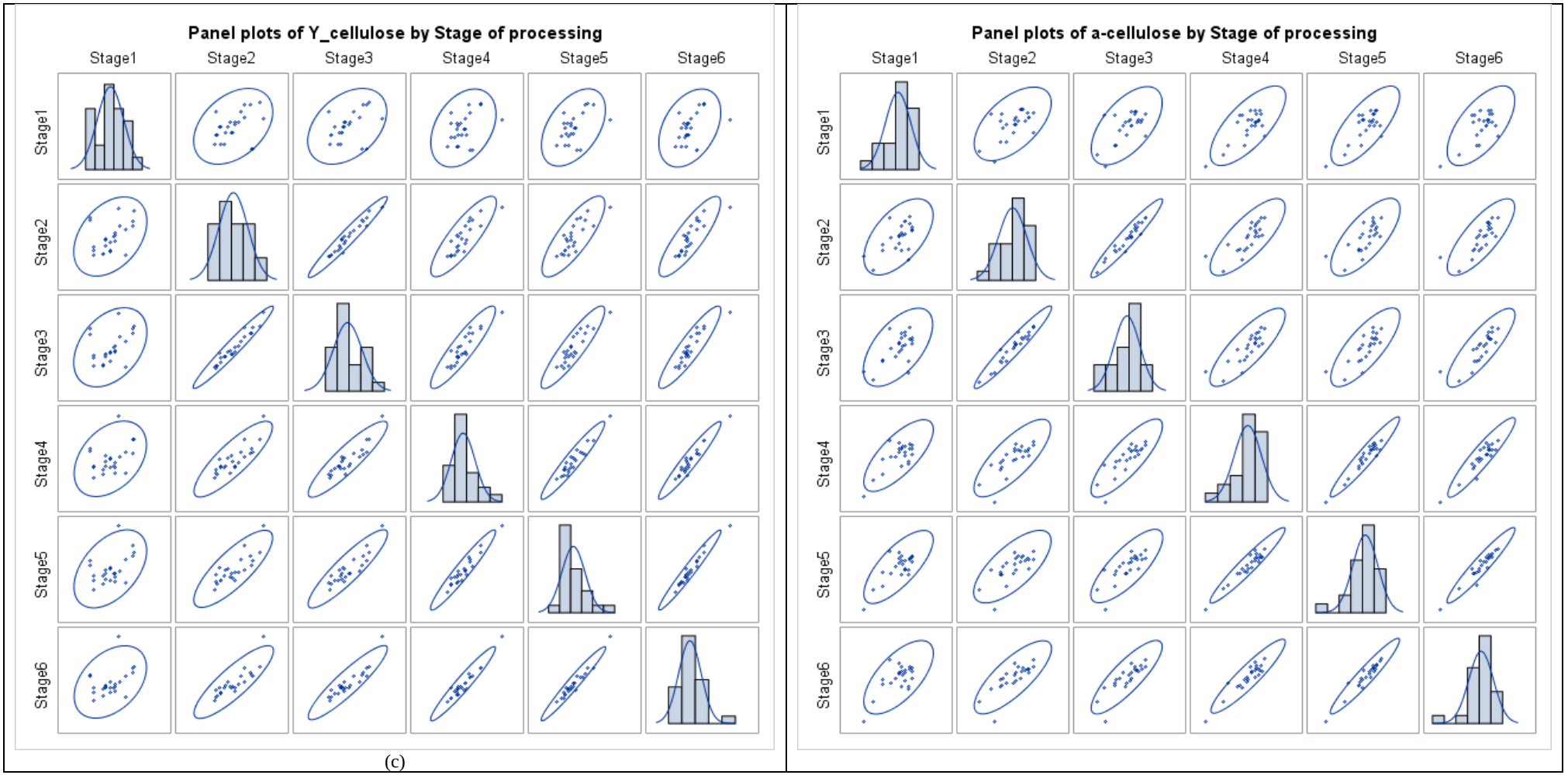


Figure 2.9. Scatter plots of chemical properties to depict their correlations.



(b)

Figure 2.10 (a). Scatter plots for stages of processing for Viscosity and Lignin



(c)

Figure 2.1 (b). Scatter plots for stages of processing for γ-cellulose and α-cellulose

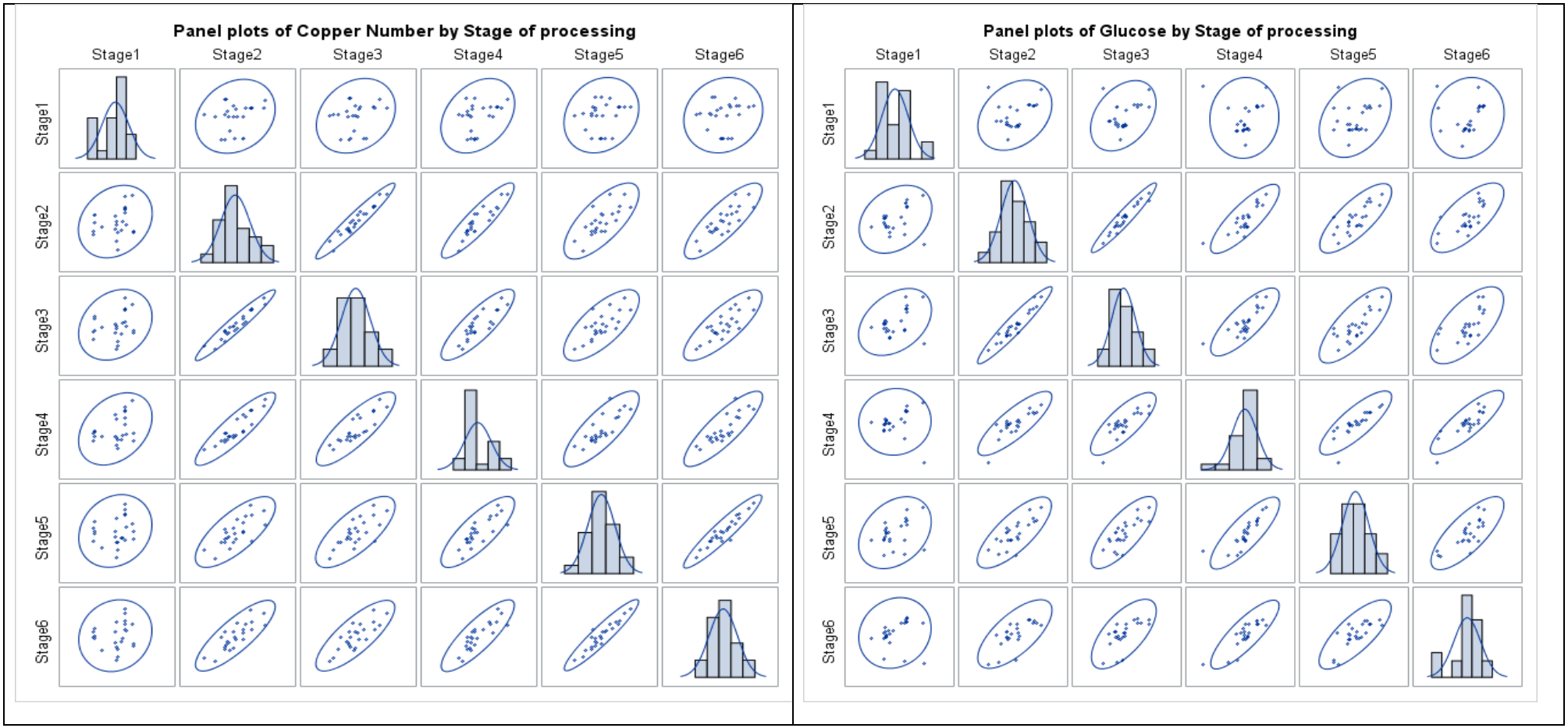


Figure 2.10 (c) Scatter plots for stages of processing for Copper number and Glucose

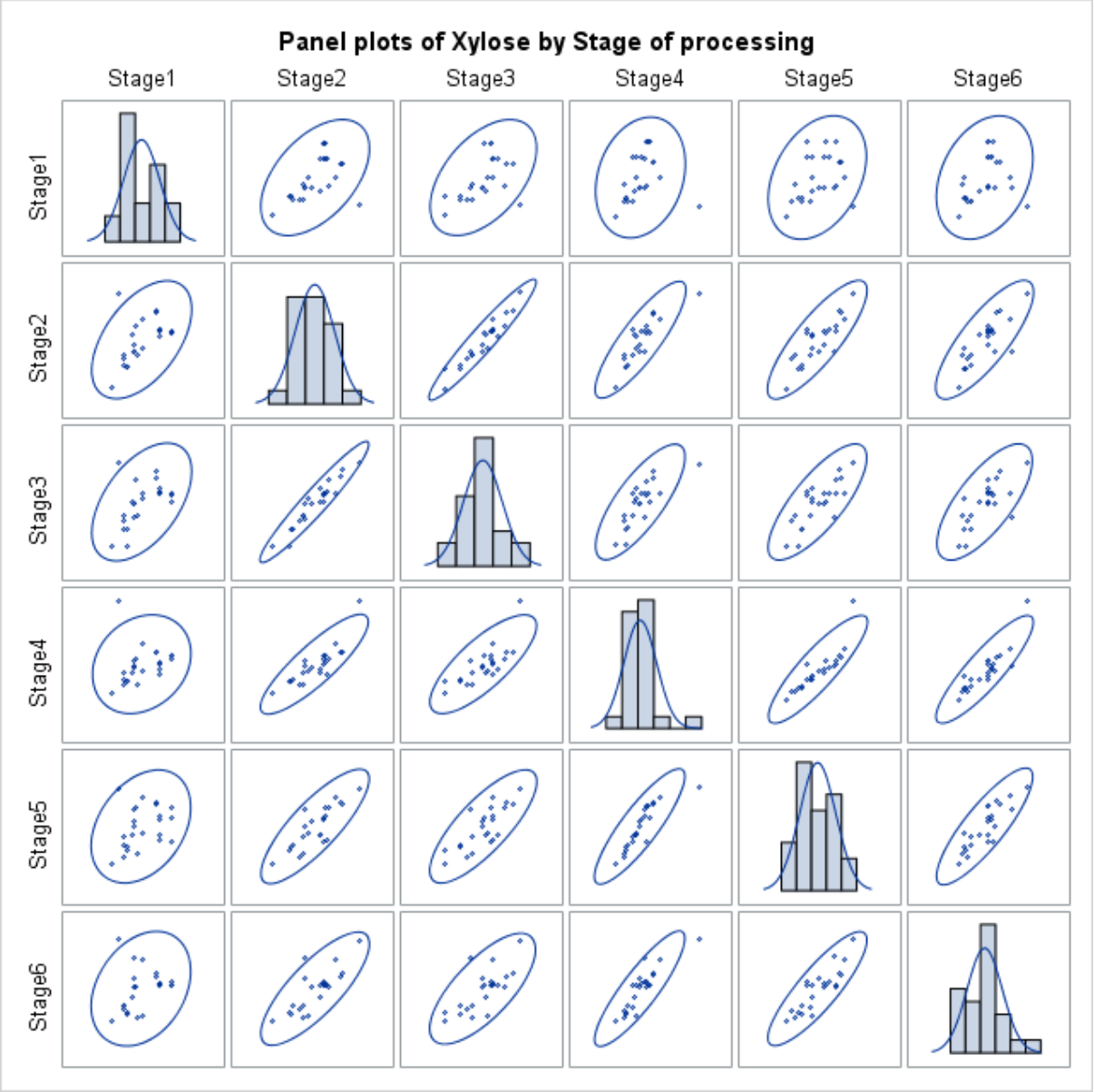


Figure 2.10 (d) Scatter plots for stages of processing for Xylose

Chapter 3

Fitting Random Coefficient Models to Timber Pulp Chemical Properties

3.1. Introduction

Results from the profile plots and the correlation analysis of Chapter 2, have indicated that different genotypes have different trajectories on the seven chemical property variables. This suggests that the chemical property variables can be modelled using mixed models in the context of random coefficient models. The random effects in this case are the different model coefficients which vary by genotype and individual samples on which repeated measurements were taken. Random coefficient models will allow us to distinguish which genotype have similar evolutional profiles through the processing stages and this allows for the determination of which genotypes are similar. Coull (2011) carried out an analysis of a model incorportating random intercepts and functional slopes (splines) in the assessment of susceptibility in longitudinal designs. While coull estimated precise intercepts in his model, the slope parameter estimates were not single values but rather spline functions. In this study the slope parameters, which are random, will be estimated (or predicted) as single entinties, just like the intercepts, rather than spline functions.

3.2. The Linear Mixed Model for Repeated Measures and the Random Coefficient Model

In order to understand the use of linear mixed models (LMM) to model random coefficients, it is necessary to outline the basic features of the LMM and its extension to more complex forms. The data used in this study have correlated variables as well repeated measurements of the same variables that are correlated in time. Verbeke and Molenbrghs (2000) and Molenbrghs and Verbeke (2005) make reference to correlated data as data falling under various data structures that include, clustered data, repeated measurements, longitudinal data and spatially correlated data. The serial correlations discussed in Chapter 2 above indicate that our data fall in this broad family of data structures and in particular, the repeated measurements category. Rizopoulos (2012) gives an outline of LMM's and how they can be used to model such data with illustrations in the R-software. The linear mixed model falls under a family of

models called the generalised linear mixed models (GLMM) which is an extension of generalised linear models (GLM). To understand all these models, it would be convenient to start with a discussion of GLMs.

In this study LMMs were used to analysis the chemical properties of dissolving timber pulp with a view to see how different genotypes behave. It is expected that, genotypes with similar behaviour under chemical processing will have similar random coefficient models for the various responses variables discussed in Chapter 2. Random coefficient models were thus explored and their usefulness in determining genotypes with similar evolutions over the processing stages investigated.

3.2.1. Generalised Linear Models (GLM)

Consider a univariate response variable Y, and *p*-predictor variables (or covariates) $X_1, \ldots, X_p$. A likelihood function is assumed for Y and the mean of Y (that is $E[Y] = \mu$) or a function of μ, say $g(\mu)$, is modelled as a linear function of the covariates, that is

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j. \qquad (3.1)$$

The function $g(\mu)$ is a link function that makes it possible to obtain a linear combination of the covariates that is related to μ.

Suppose that the response variable Y has a density function that is a member of the exponential family of distributions. The density function of Y can then be written as

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right) \qquad (3.2)$$

where θ is the natural parameter, which is the main focus of the estimation process (usually $\theta = E[Y] = \mu$), ϕ is a scale parameter which is usually assumed to be known, a is a function of ϕ, b is a function of θ, and c is a joint function of y and ϕ (Hastie and Tibshirani, 1986, 1990). The mean is related to the natural parameter by $\mu = b'(\theta)$ and if $g(\mu) = \theta$ the link function is said to be in canonical form, that is, $g(\mu)$ has an identity link function and is expressed as

$$g(\mu) = \mu = E(Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

For some given data, the maximum likelihood estimates of $\beta_0, \beta_1, \ldots, \beta_p$ are obtained using the Fisher scoring procedure as outlined by Rustagi's (1994), Smyth (2002) and Wang (2007).

3.2.2. Parameter Estimation in Generalised Linear Models - The Fisher scoring procedure

This section describes the Fisher scoring procedure as outlined by Rustagi (1994) and Wang (2007). Let $Y_1,\ldots, Y_n$ be a random sample of independent random variables each with density function $f(y;\boldsymbol{\theta})$, which is twice differentiable, and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^p$, where $\boldsymbol{\Theta}$ is a parameter space. The parameter vector $\boldsymbol{\theta}$ is p-dimensional and can be written as $\boldsymbol{\theta}^{\mathrm{T}} = [\theta_1, \ldots, \theta_p]$. The objective is to obtain the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$, of the parameter vector $\boldsymbol{\theta}$. Under the assumption of independence, the likelihood function of the random sample $\boldsymbol{Y} = [Y_1, \ldots, Y_n]$, is given by $L = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta})$ and the log-likelihood function is given by

$$\ell(\boldsymbol{\theta}) = \ln(L) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \log\{f(y_i;\boldsymbol{\theta})\}. \tag{3.3}$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ can be obtained by solving the simultaneous equations derived from

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_2} = \cdots = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} = 0. \tag{3.4}$$

These simultaneous equations do not always have direct solutions and numerical methods such as the Newton Raphson procedure are often used to find the estimate of the parameter vector $\boldsymbol{\theta}$. The Newton Raphson procedure starts with a carefully chosen initial estimate $\boldsymbol{\theta}_0$, which is then iteratively updated according to the recursive formula

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m - [\nabla^2 \ell(\boldsymbol{\theta}_m)]^{-1} \nabla \ell(\boldsymbol{\theta}_m) \tag{3.5}$$

where

$$\nabla \ell(\boldsymbol{\theta}_m) = \begin{pmatrix} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_1} \\ \vdots \\ \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix} \tag{3.6}$$

and

$$\boldsymbol{\nabla}^2 \ell(\boldsymbol{\theta}_m) = \begin{pmatrix} \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_1^2} & \cdots & \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_1 \, \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_p \, \partial \theta_1} & \cdots & \dfrac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \theta_p^2} \end{pmatrix}. \tag{3.7}$$

The estimate of the parameter vector, that is $\widehat{\boldsymbol{\theta}}$, is obtained when a convergence criteria is met. The Newton Raphson procedure requires the computation of the matrix $\boldsymbol{\nabla^2}\ell(\boldsymbol{\theta}_m)$, at each iteration. The Fisher scoring procedure, on the other hand, makes use of $\boldsymbol{I}(\boldsymbol{\theta}) = -E[\boldsymbol{\nabla}^2\ell(\boldsymbol{\theta})]$ in place of $\boldsymbol{\nabla}^2\ell(\boldsymbol{\theta}_{\mathrm{m}})$. The matrix $\boldsymbol{I}(\boldsymbol{\theta}) = -\mathrm{E}[\boldsymbol{\nabla}^2\ell(\boldsymbol{\theta})]$ is the Fisher information matrix. The Fisher information may not depend on the current value of $\boldsymbol{\theta}$ hence the same information matrix can be used for all iterations. This is a major improvement on the Newton Raphson procedure as far as computation time is concerned. The Fisher scoring recursive formula is given by

$$\boldsymbol{\theta}_{m+1} = \boldsymbol{\theta}_m - \boldsymbol{I}(\boldsymbol{\theta}_m)^{-1}\boldsymbol{\nabla}\ell(\boldsymbol{\theta}_m). \tag{3.8}$$

The performance of the procedure depends on the careful selection of the initial value and the stopping criteria should also be specified. Most statistical software consider the procedure to have converged if there is no significant change in the mean square error.

3.1.3. Estimation of LMM parameters by Maximum Likelihood Estimation (MLE)

The linear mixed model is a special case of the GLMMs where the assumption of normality is made on the response variable. The model can be expressed as

$$\begin{cases} \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon} \\ \boldsymbol{b} \sim N(\boldsymbol{0}, \boldsymbol{G}) \\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{R}) \end{cases} \tag{3.9}$$

Where $\boldsymbol{X}$ is a design matrix corresponding to the fixed effects and $\boldsymbol{Z}$ is the design matrix corresponding to the random effects. The marginal distribution of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V}) \quad \text{where } \boldsymbol{V} = \boldsymbol{R} + \boldsymbol{ZGZ}^{\mathrm{T}}.$$

Let $\boldsymbol{\varphi}$ be the vector of the q-variance components in $\boldsymbol{V}$, the log-likelihood function for model (3.9) is then given by

$$l\{\boldsymbol{y};\boldsymbol{\beta},\boldsymbol{\varphi}\} = k - \tfrac{1}{2}\log(|\boldsymbol{V}|) - \tfrac{1}{2}(\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta}) \qquad (3.10)$$

where k is a constant. If the matrix $\boldsymbol{V}$ is of full rank then the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ can be estimated by solving the following system of partial differential equations:

$$\frac{\partial l}{\partial \beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X\beta} = 0 \qquad (3.11)$$

and

$$\frac{\partial l}{\partial \varphi_i} = \tfrac{1}{2}\left\{(\boldsymbol{y}-\boldsymbol{X\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \varphi_i}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X\beta})\right\} - tr\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \varphi_i}\right) = 0, \qquad (3.12)$$

where φ_i is the i^{th} component of the p-dimensional vector $\boldsymbol{\varphi}$. Solving (3.11) and (3.12) gives

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{y} \qquad (3.13)$$

Equation (3.13) still requires the estimation of the matrix $\boldsymbol{V}$ whose components are found by solving

$$\boldsymbol{y}^{\mathrm{T}}\frac{\partial \boldsymbol{V}}{\partial \varphi_i}\boldsymbol{P}\boldsymbol{y} = tr\left(\boldsymbol{V}^{-1}\frac{\partial \boldsymbol{V}}{\partial \varphi_i}\right) \qquad (3.14)$$

where the matrix $\boldsymbol{P}$ is given by

$$\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{V}^{-1} \qquad (3.15)$$

The estimate of $\boldsymbol{V}$, that is $\widehat{\boldsymbol{V}}$, is then substituted into (3.13) to obtain the MLEs of $\boldsymbol{\beta}$.

3.2.3. Estimation of LMM parameters by Restricted Maximum Likelihood Estimation (REML)

Estimates of variance components obtained using MLE are biased and this calls for the use of REML which works by transforming the data to eliminate the fixed effects then work with the transformed data to estimate the variance components. Let matrix $\boldsymbol{A}$ be an $n\times(n\text{-}p)$ matrix such that rank(**A**)=n-p and $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}$=0. The matrix $\boldsymbol{A}$ (the restriction matrix) is the transformation matrix that is applied on the original response data vector $\boldsymbol{y}$ to obtain $\boldsymbol{u} = \boldsymbol{A}^{\mathbf{T}}\boldsymbol{y}$ and this implies that $\boldsymbol{u}\sim\boldsymbol{N}(\mathbf{0},\boldsymbol{A}^{\mathbf{T}}\boldsymbol{V}\boldsymbol{A})$. Based on the new restricted response variable $\boldsymbol{u}$ the restricted log-likelihood is given by

$$l_R\{\boldsymbol{u}\colon\boldsymbol{\varphi}\} = k - \tfrac{1}{2}\log(|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}|) - \tfrac{1}{2}\boldsymbol{u}^{\mathrm{T}}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A})^{-1}\boldsymbol{u}. \qquad (3.16)$$

Estimates of the variance components are obtained by partially differentiating $\boldsymbol{l_R}\{\boldsymbol{u}: \boldsymbol{\varphi}\}$ with respect to the variance components to obtain a system of partial differential equations of the form

$$\frac{\partial l_R}{\partial \varphi_i} = \frac{1}{2}\left\{\boldsymbol{y}^{\mathrm{T}}\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \varphi_i}\boldsymbol{P}\boldsymbol{y} - tr\left(\boldsymbol{P}\frac{\partial \boldsymbol{V}}{\partial \varphi_i}\right)\right\} = 0, \qquad \text{for } i = 1, \dots q, \tag{3.17}$$

where $\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^{\mathbf{T}}\boldsymbol{V}\boldsymbol{A})^{-\mathbf{1}}\boldsymbol{A}^{\mathbf{T}}$. The REML estimates are obtained through the transformation matrix ***A*** but they do not depend on the same matrix. This means that the matrix ***A*** is not unique to any set of estimates. Once the variance components vector $\widehat{\boldsymbol{\varphi}}$ is estimated the fixed effects parameter vector $\widehat{\boldsymbol{\beta}}$ can then be estimated using MLE. Both MLE and REML are based on the assumption of normality which is not always the case. Other methods that deal with situations where normality is violated have been put forward and these include quasi-likelihood methods as described by Wedderburn (1974) and quasi-likelihood methods for count data by Wooldridge (1997).

3.2.4. Estimation of random effects by Best Linear Unbiased Predictors (BLUP)

In this study, the random components of the random coefficients model are closely analysed as they outline the differences in evolution between genotypes. Random effects are predicted rather than estimated hence they are termed best linear unbiased predictors rather than best linear unbiased estimators. The theory of Best Linear Unbiased Predictors (BLUPs) as compared to Best Linear Unbiased Estimators (BLUEs) is well known and appears in many statistical publications, for example, Henderson (1953), Searle (1995), Robinson (1991), Coull (2011) and many others. A brief discussion of such predictors and how they apply to random effects of mixed models, with particular regards to random coefficients and piecewise linear regression models, is hereby presented.

The linear mixed effects model (equation 3.9) can be stated as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{e} \tag{3.18}$$

where

$$\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}\right). \tag{3.19}$$

Given a vector of observed values $\boldsymbol{y}$, we wish to find the best guess or prediction of the vector $\boldsymbol{b}$. It is important to note that, while we treat the vector of fixed effects, that is $\boldsymbol{\beta}$, as parameters to be estimated by $\widehat{\boldsymbol{\beta}}$, we cannot treat $\boldsymbol{b}$ in the same way. The vector $\boldsymbol{b}$ is a random variable rather than a fixed parameter vector hence we can talk of predicting $\boldsymbol{b}$ rather than estimating it. We thus talk of a Best Linear Unbiased Predictor (BLUP) for $\boldsymbol{b}$, rather than a Best Linear Unbiased Estimator (BLUE). The BLUP for $\boldsymbol{b}$, which is denoted by $\widehat{\boldsymbol{b}}$ has the following properties:

(i) $\widehat{\boldsymbol{b}}$ must be a linear function of $\boldsymbol{y}$;

(ii) $\widehat{\boldsymbol{b}}$ must be an unbiased predictor for $\boldsymbol{b}$, that is, $E\left(\widehat{\boldsymbol{b}} - \boldsymbol{b}\right) = \boldsymbol{0}$ and

(iii) $\mathrm{Var}\left(\widehat{\boldsymbol{b}} - \boldsymbol{b}\right) \leq \mathrm{Var}\left(\widetilde{\boldsymbol{b}} - \boldsymbol{b}\right)$, where $\widetilde{\boldsymbol{b}}$ is any other linear unbiased predictor for $\boldsymbol{b}$.

The BLUP for $\boldsymbol{b}$ is a conditional expectation which is the BLUE for $E(\boldsymbol{b}|\boldsymbol{y})$. Consider any two multivariate normal vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with the multivariate normal distribution

$$\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right) \tag{3.20}$$

The conditional distribution of any two such multivariate vectors is given by

$$\boldsymbol{X}_2|\boldsymbol{X}_1 \sim N(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{X}_1 - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \tag{3.21}$$

Now consider equations 3.18 and 3.19 above. It can be shown that

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{b}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X\beta}\\ \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}\boldsymbol{Z} & \boldsymbol{I}\\ \mathbf{I} & \boldsymbol{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{e}\end{bmatrix} \tag{3.22}$$

with

$$\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{b}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{X\beta}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\boldsymbol{ZGZ}^{\mathrm{T}} + \boldsymbol{R} & \boldsymbol{ZG}\\ \boldsymbol{GZ}^{\mathrm{T}} & \boldsymbol{G}\end{bmatrix}\right) \tag{3.23}$$

and assuming that the covariance matrix of $\boldsymbol{y}$, that is $\boldsymbol{V} = \boldsymbol{ZGZ}^{\mathrm{T}} + \boldsymbol{R}$, is positive definite then

$$\boldsymbol{b}|\boldsymbol{y} \sim N(\boldsymbol{GZ}^{\mathrm{T}}(\boldsymbol{ZGZ}^{\mathrm{T}} + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}), \qquad \boldsymbol{G} - \boldsymbol{GZ}^{\mathrm{T}}(\boldsymbol{ZGZ}^{\mathrm{T}} + \boldsymbol{R})^{-1}\boldsymbol{ZG}) \tag{3.24}$$

The BLUP for $\boldsymbol{u}$ is thus given by

$$E(\boldsymbol{b}|\boldsymbol{y}) = \boldsymbol{GZ}^{\mathrm{T}}(\boldsymbol{ZGZ}^{\mathrm{T}} + \boldsymbol{R})^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{3.25}$$

or

$$E(\boldsymbol{b}|\boldsymbol{y}) = \boldsymbol{GZ}^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{3.26}$$

Since the matrices $\boldsymbol{G}$ and $\boldsymbol{V}$ are estimated from observed data by $\widehat{\boldsymbol{G}}$ and $\widehat{\boldsymbol{V}}$ respectively, the BLUP for $\boldsymbol{b}$ is approximated by

$$\widehat{\boldsymbol{b}} = E(\boldsymbol{b}|\boldsymbol{y}) = \widehat{\boldsymbol{G}}\boldsymbol{Z}^{\mathrm{T}}\widehat{\boldsymbol{V}}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right). \tag{3.27}$$

3.2.5. Use of LMM for the longitudinal pulp data

The data at hand is longitudinal in nature as measurements are taken sequentially in time. The linear mixed model for repeated measures (longitudinal data) for the pulp data has genotype and processing stage as fixed effects and the pulp samples as random effects on which repeated measurements are taken. The model can be expressed as

$$Y_{ijt} = f_{ij} + \tau_t + I_{ijt} + e_{ijt} \tag{3.28}$$

where f_{ij} is the part of the model that is due to the fixed effects and this can be expressed as

$$f_{ij} = \mu + \alpha_i,$$

where μ is the overall mean and α_i is the genotype effect. The effect of stage (or time) t is denoted by τ_t. The term I_{ijt} of model (3.28) is the interaction between processing stage and genotype and e_{ijt} is the random effect part of the model which is the random error associated with subject i, under the j^{th} treatment at stage t. Model (3.28) can also be written as

$$Y_{ijt} = \mu + \alpha_i + \tau_t + I_{ijt} + e_{ijt} \tag{3.29}$$

The subjects (pulp samples) are assumed to be independent while the observations of each pulp sample over the processing stages are correlated according to some suitable covariance structure. If the complete set of observations is put into a single vector $\boldsymbol{Y}$, noting that there are L subjects in total and T processing stages, the covariance matrix of $\boldsymbol{Y}$ can be written as

$$Var(\boldsymbol{Y}) = \boldsymbol{I}_L \otimes \boldsymbol{\Sigma_T} \tag{3.30}$$

where the covariance matrix $\boldsymbol{\Sigma_T}$ shows how the values of a single subject at different stages are related to each other. The matrix $\boldsymbol{I}_L$ is an $L \times L$ identity matrix while $\boldsymbol{\Sigma}_T$ has one of the many possible covariance structures. The best fitting covariance structure

is determined by considering known covariance structures and choosing one with the best fit according to the Akaike information criteria or *AIC* (Burnhan and Anderson, 2004). A correct choice of a covariance structure for $\boldsymbol{\Sigma}_T$ will greatly affect the quality of the model parameters obtained (Littell et al, 2006). The covariance matrix of the observations on each subject over all the time periods can also be decomposed into

$$\boldsymbol{\Sigma_T} = \sigma_T^2 \boldsymbol{J} + \boldsymbol{R} \tag{3.31}$$

where $\sigma_T^2 \boldsymbol{J}$ is the part of variation due to the subject, $\boldsymbol{R}$ is the covariance matrix of observations within the same subject due to the different stages. Having identified a suitable covariance structure, the model parameters can be estimated by, either using Maximum Likelihood (*ML*), or Restricted Maximum Likelihood (*REML*) methods.

3.2.6. Fitting the Random Coefficient Model to the Pulp Data

According to Swamy (1970), the random coefficient regression model is similar to the linear mixed model for repeated measures. Such a model has also been described by Bollen and Curran (2006) as a latent curve model. Under this model, each genotype (or treatment) has its parameters estimated separately to form a family of parameters which have overall mean parameters for all treatments and parameters specific to each treatment. The parameters for all treatments can then be used to compare the performances of the different genotypes.

The effect of time on the response variables can be linear or may take any form suggested by the profile plot of the response variable on time. As an example, a quadratic random coefficient regression model would of the form:

$$Y_{gt} = \alpha_{0g} + \alpha_{1g} t + \alpha_{2g} t^2 + \varepsilon_{gt} \tag{3.32}$$

Where Y_{gt} is a response variable observed for genotype g at time t, α_{0g} is the overall slope parameter for all subjects under genotype g, that is, the initial value of the response variable *Y* at time *t*=0 (or raw pulp stage) for genotype g, α_{1g} is the overall linear slope of samples under genotype g, α_{2gb} is the overall curvature (quadratic term) of samples under genotype g and ε_{gt} is the random error term associated with genotype g at time t.

The parameters are treated as variables which vary by genotype and can be further factorized into $\alpha_{0g} = (\beta_0 + b_{0g})$, $\alpha_{1g} = (\beta_1 + b_{1g})$ and $\alpha_{2g} = (\beta_2 + b_{2g})$ where the quantities β_0, β_1 and β_2 are the overall intercept, slope and curvature parameters respectively and $b_{0g} \sim N(0, \sigma_0^2)$, $b_{1g} \sim N(0, \sigma_1^2)$ and $b_{2g} \sim N(0, \sigma_2^2)$ are genotype specific variables. Substituting the factorized forms of α_{0g}, α_{1g} and α_{2g} in (3.32) gives

$$Y_{gt} = (\beta_0 + b_{0g}) + (\beta_1 + b_{1g})t + (\beta_2 + b_{2g})t^2 + \varepsilon_{gt} \qquad (3.33)$$

where $E(Y_{bgi}) = \beta_0 + \beta_1 t + \beta_2 t^2$ is the population growth model for all genotypes combined. The quantities b_{0g}, b_{1g} and b_{2g} are the variable parts (random effects) of the model parameters that depend on genotype and have zero means with a covariance structure which can be written as:

$$Cov(b_{0g}, b_{1g}, b_{2g}) = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{10} & \sigma_1^2 & \sigma_{12} \\ \sigma_{20} & \sigma_{21} & \sigma_2^2 \end{bmatrix} \qquad (3.34)$$

where $\sigma_i^2 = \text{variance}(b_{ig})$, $\sigma_{ij} = \sigma_{ji} = \text{covariance}(b_{ig}, b_{jg})$ for i, j = 0, 1, 2. If $\sigma_0^2 = 0$ then all the genotypes have identical intercepts (or stage 0 values) equal to β_0. Likewise, if $\sigma_1^2 = 0$ then the linear slopes of all genotype/bleaching condition combinations are identical. The covariance σ_{01} shows the association between the raw stage value (intercept) and the linear slope, σ_{02} shows the association between the intercept and the quadratic term of the model and σ_{12} shows the association between the slope parameter and the quadratic term of the model. Higher or lower order random coefficient regression models can also be considered depending on the relationship between the response variable and time *t* (or processing stage in this case) as determined using the profile plots of the variables.

3.3. Model fitting and results discussion

The SAS procedure Proc MIXED was used to fit the random coefficients model to the data using restricted maximum likelihood estimates (REML) (Liu et al, 2007). The SAS code is presented in Appendix A1.2. The results of the data analysis are presented below.

3.3.1. Choice of covariance structures

Table 3.1 below shows the covariance structures that fitted the data best out of many covariance structures attempted. The results, as presented in Table 3.1, show that, the covariance structure that fits the viscosity, lignin and α-cellulose data best is the unstructured one (AIC=957.7, 306.2 and 438.5 respectively), for copper number the compound symmetry covariance structure is of best fit (AIC=187.4), for γ-cellulose the Toeplitz covariance structure is of best fit (AIC=368.3) and for xylose the AR(1) covariance structure is of best fit. The covariance structure that was found to be of best fit to any variable, will be used in the fitting of a random coefficient model that best describes the trajectory (linear or quadratic) of the concerned variable.

Table 3.1. Fit Statistics for Covariance Structures for random coefficient regression models for the seven chemical pulping variables.

Covariance Structure	Number of parameters	AIC by Variable						
		Viscosity	Lignin	γ-cellulose	α-cellulose	Copper Number	Glucose	Xylose
Unstructured	**4**	**957.7**	**306.2**	368.9	**438.5**	188.8	419.0	259.0
ANTE(1)	4	972.6	315.7	369.5	441.0	197.1	**417.6**	259.1
AR(1)	3	976.1	311.7	369.6	441.0	197.1	417.9	**257.1**
ARMA(1,1)	4	978.1	317.7	371.6	443.0	199.1	419.9	259.1
CS	3	976.1	306.7	369.6	440.3	**187.4**	419.0	257.3
Toeplitz	4	975.7	308.8	**368.3**	440.3	188.8	419.0	259.0

3.3.2. Random coefficient models for viscosity

Profile plots for viscosity (Figure 2.2) suggest that viscosity is mainly linearly related to processing stage (time) although a slight curvature might also need to be investigated. After fitting a quadratic random coefficients model (3.33), the quadratic terms for all genotypes turned up not to be significant hence a linear regression model was fitted to the data. Slope parameters for some of the genotypes were found to be significant with each genotype having its own set of intercept and slope parameters. Such coefficients are considered random (Swamy, 1970). In this section the parameters of model (3.33), without the quadratic term, are estimated for each genotype.

The results for the random coefficient regression models for the various genotypes are presented in Table 3.2 below. The slope parameters of the models for the seven genotypes indicated that Egrandis and Enitens had the lowest and non-significant rates of change of viscosity over the six processing stages (Slope=-1.995 with *p*-value=0.2253 and Slope=-2.1222 with *p*-value=0.3159 respectively). In general, the genotypes with the lowest viscosities before processing also had the lowest rate of change of viscosity over the processing stages (Intercept for Egrandis=38.34463 and Intercept for Enitens=43.9554).

Table 3.2 Parameter estimates for the random coefficient regression model for viscosity.

Model parameter estimates, Standard deviations and p-values for t-tests						
	Intercept			Slope		
Genotype	Parameter	(Std Dev)	p-value	Parameter	(Std Dev)	p-value
E.dunnii	63.5289	2.9355	<0.0001*	-5.8961	0.9630	<0.0001*
E.grandis	38.4463	4.9086	<0.0001*	-1.9950	1.6115	0.2253
E.smithii	48.6429	4.4809	<0.0001*	-3.5687	1.4711	0.0215*
E.nitens	43.9554	6.3370	<0.0001*	-2.1222	2.0804	0.3159
GCG	58.1603	6.3370	<0.0001*	-5.6770	2.0804	0.0105*
GUA	70.8950	6.3370	<0.0001*	-7.9262	2.0804	0.0006*
GUW	58.1603	6.3370	<0.0001*	-5.1765	2.0804	0.0186*

*significant parameter at 5% significance level

A diagrammatic presentation of the random coefficients regression models for the viscosity data is shown in Figure 3.1 below. The genotypes with the steepest slopes also had the highest raw stage viscosities. In order of highest intercepts and hence in terms of the steepest slopes, the genotypes, as indicated in Figure 3.1, can be ordered as:

1.GUA, 2. Edunnii, 3. GCG, 4. GUW, 5. Esmithii, 6. E nitens and 7. Egrandis.

The covariance matrix for the slope and intercept parameters for all genotypes is given as

$$Cov(b_{0g}, b_{1g}) = \begin{bmatrix} 84.829 & -23.833 \\ -23.833 & 9.096 \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}) = \begin{bmatrix} 1.000 & -0.858 \\ -0.858 & 1.000 \end{bmatrix}.$$

The correlation between the intercept and the slope parameters is $r = -0.858$ which is a strong negative. This shows the dependence of the rate of change of viscosity to

initial viscosity levels. This implies that genotypes which start off with high viscosity levels have higher rates of change of viscosity.

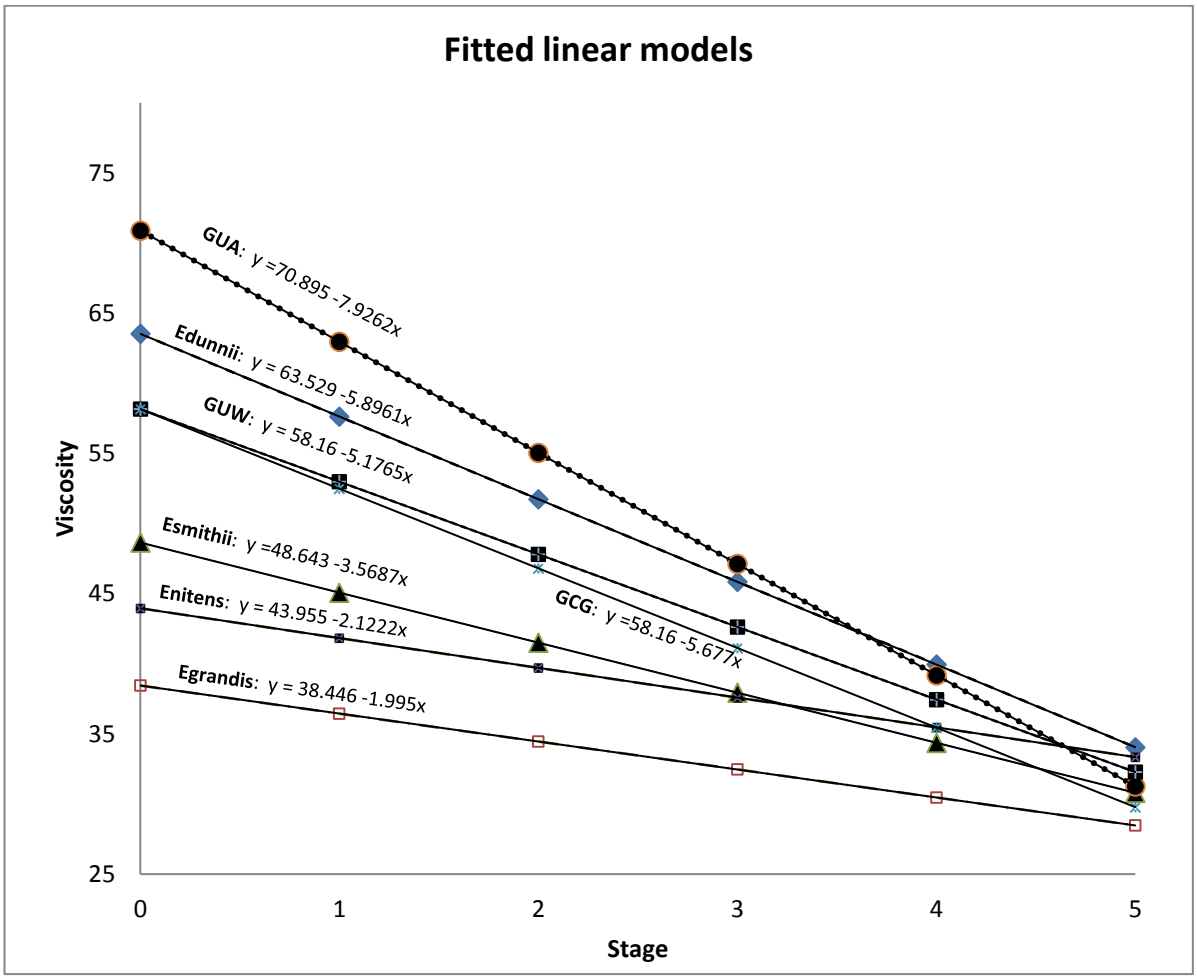


Figure 3.1 Random coefficients regression models for the seven genotypes

Results in Table 3.3 below shows that the low slope parameter of Egrandis is significantly different to the slope parameters of Edunnii (Difference in slope= -3.9011, p-value=0.0464) and GUA (Difference in slope= -5.312, p-value=0.0317). The other genotypes do not have significantly different slope parameters but this is mainly because the parameter estimates have high standard deviations as shown in Table 3.2 (they range from 0.9630 to 2.0804).

Table 3.3: Intercept and slope parameter estimated differences for the random coefficient regression model for viscosity.

Genotype			Differences in intercepts and slopes viscosities (*p*-values in brackets)					
			E.grandis	E.nitens	Esmithii	GUW	GCG	Edunnii
E.grandis	Intercept	38.4463	-					
	Slope	-1.9950	-					
E.nitens	Intercept	43.9554	5.5091 (0.4972)	-				
	Slope	-2.1222	0.1272 (0.9618)	-				
Esmithii	Intercept	48.6429	10.1966 (0.1355)	4.6875 (0.5504)	-			
	Slope	-3.5687	1.5737 (0.4763)	-1.4465 (0.5745)	-			
GUW	Intercept	58.1603	19.7140 (0.0199)*	14.2049 (0.1234)	9.5174 (0.2296)	-		
	Slope	-5.1765	3.1816 (0.2361)	3.0544 (0.3075)	1.6079 (0.5328)	-		
GCG	Intercept	58.1764	19.7301 (0.0198)*	14.2209 (0.1230)	9.5334 (0.2289)	0.01606 (0.9986)	-	
	Slope	-5.6770	3.6821 (0.1720)	3.5548 (0.2364)	2.1083 (0.4145)	-0.5005 (0.8661)	-	
Edunnii	Intercept	63.5289	25.0826 (0.0001)*	19.5734 (0.0088)*	14.886 (0.0093)*	5.3686 (0.4480)	5.3525 (0.4494)	-
	Slope	-5.8961	-3.9011 (0.0464)$^{+}$	-3.7739 (0.1102)	-2.3274 (0.1956)	-0.7195 (0.7558)	-0.2191 (0.9245)	-
GUA	Intercept	70.8950	32.4487 (0.0003)*	26.9396 (0.0053)*	22.2521 (0.0075)*	12.7347 (0.1656)	12.7186 (0.1661)	7.3661 (0.2999)
	Slope	-7.9262	5.9312 (0.0317)$^{+}$	5.8040 (0.0578)	4.3575 (0.0976)	-2.7496 (0.3575)	2.2492 (0.4506)	2.0301 (0.3829)

*Genotypes with significantly different intercept parameters

$^{+}$ Genotypes with significantly different slope parameters

3.3.3. Random coefficient models for Lignin

Profile plots for lignin (Figure 2.3) suggest that lignin could be quadratically related to processing stage (time) since lignin plots appear to have distinct curvatures over time. A quadratic random coefficients model (3.33) was fitted to the lignin data with the results presented in Table 3.4 below.

Table 3.4 Parameter estimates for the random coefficient regression model for Lignin

Lignin model parameter estimates, Standard deviations and p-values for t-tests									
Genotype	Intercept			Slope			Curvature		
	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value
E.dunnii	5.662	0.618	<0.0001*	-1.857	0.267	<0.0001*	0.174	0.034	<0.0001*
E.grandis	4.704	0.618	<0.0001*	-1.869	0.267	<0.0001*	0.193	0.034	<0.0001*
E.smithii	6.654	0.536	<0.0001*	-2.572	0.189	<0.0001*	0.258	0.024	<0.0001*
E.nitens	3.499	0.618	<0.0001*	-1.407	0.267	<0.0001*	0.148	0.034	<0.0001*
GCG	6.855	0.618	<0.0001*	-2.515	0.267	<0.0001*	0.240	0.034	<0.0001*
GUA	5.771	0.618	<0.0001*	-2.352	0.267	<0.0001*	0.243	0.034	<0.0001*
GUW	4.180	0.437	<0.0001*	-1.568	0.267	<0.0001*	0.154	0.034	<0.0001*

*significant parameter at 5% significance level

The covariances and correlations between the intercept, slope and quadratic (curvature) parameters for lignin for all genotypes are given as

$$Cov(b_{0g}, b_{1g}, b_{2g}) = \begin{bmatrix} 0.164 & -0.071 & 0.006 \\ -0.071 & 0.038 & -0.003 \\ 0.006 & -0.003 & < 0.0001 \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}, b_{2g}) = \begin{bmatrix} 1.000 & -0.895 & - \\ -0.895 & 1.000 & - \\ - & - & 1.000 \end{bmatrix}.$$

The correlation between the intercept and the slope parameters is $r = -0.895$, which is strong and negative. This shows the dependence of the rate of change of lignin to initial lignin levels. The relationship of the quadratic term to the intercept and slope parameters could not me computed as the variance of the quadratic term could not be computed due to non-convergence.

Table 3.5: Intercept, slope and curvature parameter estimated differences for the random coefficient regression model for Lignin.

Genotype	Parameter Estimates		Differences in intercepts, slopes and curvatures for lignin (*p*-values in brackets)											
			E.grandis		E.nitens		Esmithii		GUW		GCG		Edunnii	
			Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value
E.grandis	Intercept	4.704	-											
	Slope	-1.869	-											
	Curvature	0.193	-											
E.nitens	Intercept	3.499	1.205	0.055	-									
	Slope	-1.407	-0.462	0.225	-									
	Curvature	0.148	-0.462	0.225	-									
Esmithii	Intercept	6.654	-1.951	0.001*	3.155	<0.001*	-							
	Slope	-2.572	0.703	0.035*	-1.165	<0.001*	-							
	Curvature	0.258	0.703	0.035*	-1.165	0.001*	-							
GUW	Intercept	4.180	0.524	0.400	-0.681	0.275	2.474	<0.001*	-					
	Slope	-1.568	-0.301	0.428	0.161	0.670	-1.004	0.003*	-					
	Curvature	0.154	-0.301	0.428	0.161	0.670	-1.004	0.003*	-					
GCG	Intercept	6.855	-2.151	0.001*	-3.356	<0.001*	-0.201	0.709	2.675	<.0001*	-			
	Slope	-2.515	0.646	0.091	1.108	0.004*	-0.057	0.861	-0.947	0.014*	-			
	Curvature	0.240	0.646	0.091	1.108	0.004*	-0.057	0.861	-0.947	0.014*	-			
Edunnii	Intercept	5.662	0.958	0.126	2.163	0.001*	-0.993	0.068	1.482	0.019*	-1.193	0.058	-	
	Slope	-1.857	0.012	0.975	-0.450	0.237	0.715	0.032*	-0.289	0.447	0.658	0.085	-	
	Curvature	0.174	0.012	0.975	-0.450	0.237	0.715	0.032*	-0.289	0.447	0.658	0.085	-	
GUA	Intercept	5.771	-1.067	0.089	-2.272	0.001*	0.883	0.103	1.591	0.012*	1.084	0.084	-0.110	0.860
	Slope	-2.352	0.483	0.205	0.945	0.015*	-0.221	0.502	-0.784	0.041*	-0.163	0.666	0.495	0.194
	Curvature	0.243	0.483	0.205	0.945	0.015*	-0.221	0.502	-0.784	0.041*	-0.163	0.666	0.495	0.194

3.3.4. Random coefficient models for γ-cellulose

Profile plots in (Figure 2.4) do not suggest a distinct trajectory for the evolution of γ-cellulose over the processing stages. An attempt at fitting a cubic polynomial to model the evolution of γ-cellulose showed that only the linear trend component of the model was significant. A linear trend random coefficient model was fitted to the data with the results presented in Table 3.6 below.

Table 3.6 Parameter estimates for the random coefficient regression model for γ-cellulose

Y-Cellulose Model parameter estimates, Standard deviations and p-values for t-tests						
Genotype	Intercept			Slope		
	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value
E.dunnii	8.131	0.605	<0.0001*	-0.803	0.113	<0.0001*
E.grandis	7.274	0.605	<0.0001*	-0.845	0.113	<0.0001*
E.smithii	8.150	0.524	<0.0001*	-0.832	0.080	<0.0001*
E.nitens	8.046	0.605	<0.0001*	-0.940	0.113	<0.0001*
GCG	7.480	0.605	<0.0001*	-0.817	0.113	<0.0001*
GUA	8.367	0.605	<0.0001*	-0.943	0.113	<0.0001*
GUW	6.754	0.428	<0.0001*	-0.720	0.113	<0.0001*

*significant parameter at 5% significance level

The covariance matrix for the slope and intercept parameters for all genotypes is given as

$$Cov(b_{0g}, b_{1g}) = \begin{bmatrix} 0.2750 & -0.0247 \\ -0.0247 & 0.0032 \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}) = \begin{bmatrix} 1 & -0.8326 \\ -0.8326 & 1 \end{bmatrix}.$$

The correlation between the intercept and the slope parameters is $r = -0.8326$ which is strong and negative. This shows the dependence of the rate of change of viscosity to initial viscosity levels. This implies that genotypes which start off with high viscosity levels have higher rates of change of viscosity. Results in Table 3.7 show that no genotypes differ in their rates of change of γ-cellulose (p-values>0.05). Gamma cellulose cannot be used to distinguish any difference in evolutional behavior between genotypes as they do not have significantly different slopes.

Table 3.7: Intercept, slope and curvature parameter estimated differences for the random coefficient regression model for γ-cellulose.

Genotype	Parameter Estimates		Differences in intercepts and slopes for γ-cellulose (t-tests *p*-values in brackets)											
			E.grandis		E.nitens		Esmithii		GUW		GCG		Edunnii	
			Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value
E.grandis	Intercept	7.274	-											
	Slope	-0.845	-											
E.nitens	Intercept	8.046	-0.772	0.205	-									
	Slope	-0.940	0.095	0.555	-									
Esmithii	Intercept	8.150	-0.876	0.098	0.104	0.843	-							
	Slope	-0.832	-0.012	0.929	0.107	0.441	-							
GUW	Intercept	6.754	0.521	0.391	1.293	0.035	1.397	0.009	-					
	Slope	-0.720	-0.125	0.438	-0.220	0.173	-0.112	0.420	-					
GCG	Intercept	7.480	-0.205	0.735	0.567	0.351	0.671	0.203	0.726	0.233	-			
	Slope	-0.817	-0.028	0.861	-0.123	0.444	-0.016	0.910	0.126	0.434	-			
Edunnii	Intercept	8.131	0.857	0.160	0.085	0.889	-0.019	0.971	1.377	0.025	0.652	0.284	-	
	Slope	-0.803	0.042	0.793	0.137	0.394	0.030	0.831	-0.083	0.607	0.014	0.931	-	
GUA	Intercept	8.367	-1.093	0.074	-0.321	0.597	-0.217	0.680	1.613	0.009	-0.887	0.146	-0.236	0.697
	Slope	-0.943	0.098	0.544	0.003	0.986	0.110	0.429	-0.223	0.168	-0.097	0.548	0.140	0.385

3.3.5. Random coefficient models for α-cellulose

The polynomial of the trajectory that best models the evolution of α-cellulose as suggested by the profile plots in Figure 2.5 could be linear but an attempt was also made on the quadratic polynomial which turned out not significant. The linear random coefficient model was thus fitted to the data with the results presented in Table 3.8. below.

Table 3.8 Parameter estimates for the random coefficient regression model for α-cellulose

α-Cellulose Model parameter estimates, Standard deviations and p-values for t-tests						
Genotype	Intercept			Slope		
	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value
E.dunnii	89.865	1.086	<0.0001*	0.887	0.205	<0.0001*
E.grandis	91.128	1.086	<0.0001*	0.904	0.205	<0.0001*
E.smithii	91.136	0.941	<0.0001*	0.689	0.145	<0.0001*
E.nitens	90.368	1.086	<0.0001*	1.006	0.205	<0.0001*
GCG	91.153	1.086	<0.0001*	0.811	0.205	<0.0001*
GUA	90.344	1.086	<0.0001*	0.933	0.205	<0.0001*
GUW	91.317	0.768	<0.0001*	0.804	0.205	0.0002*

*significant parameter at 5% significance level

The covariance matrix for the slope and intercept parameters for all genotypes is given as

$$Cov(b_{0g}, b_{1g}) = \begin{bmatrix} 0.8685 & -0.2476 \\ -0.2476 & 0.0767 \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}) = \begin{bmatrix} 1.0000 & -0.9593 \\ -0.9593 & 1.0000 \end{bmatrix}.$$

Pairwise comparisons of intercept and slope parameters for the seven genotypes are presented in Table 3.9 below. None of the variables have significantly different intercept or slope parameters (all p-values>0.05).

Table 3.9: Intercept and slope parameter estimated differences for the random coefficient regression model for α-cellulose.

Genotype	Parameter Estimates		Differences in intercepts and slopes for a-cellulose (t-tests *p*-values in brackets)											
			E.grandis		E.nitens		Esmithii		GUW		GCG		Edunnii	
			Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value
E.grandis	Intercept	91.128	-											
	Slope	0.904	-											
E.nitens	Intercept	90.368	0.759	0.486	-									
	Slope	1.006	-0.101	0.728	-									
Esmithii	Intercept	91.136	-0.009	0.993	0.768	0.416	-							
	Slope	0.689	0.216	0.392	-0.317	0.210	-							
GUW	Intercept	91.317	-0.189	0.862	-0.948	0.385	-0.181	0.848	-					
	Slope	0.804	0.101	0.730	0.202	0.489	-0.115	0.647	-					
GCG	Intercept	91.153	-0.026	0.981	-0.785	0.472	-0.017	0.986	-0.164	0.881	-			
	Slope	0.811	0.093	0.748	0.195	0.504	-0.122	0.627	0.007	0.981	-			
Edunnii	Intercept	89.865	-1.263	0.248	-0.503	0.644	-1.271	0.180	-1.452	0.185	-1.288	0.239	-	
	Slope	0.887	-0.018	0.951	-0.119	0.683	0.198	0.432	0.083	0.776	0.076	0.795	-	
GUA	Intercept	90.344	0.784	0.472	0.025	0.982	0.793	0.402	-0.973	0.372	0.810	0.458	-0.479	0.660
	Slope	0.933	-0.028	0.923	0.073	0.802	-0.244	0.334	0.129	0.659	-0.122	0.676	-0.046	0.875

3.3.6. Random coefficient models for Copper number

The polynomial of the trajectory that best models the evolution of copper number as suggested by the profile plots in Figure 2.6 could be linear. The linear random coefficient model was thus fitted to the data with the results presented in Table 3.10 below. All the parameters are significantly different from zero (p-values<0.0001) hence there are significant changes in copper number for all genotypes over the processing stages.

Table 3.10 Parameter estimates for the random coefficient regression model for α-cellulose

Copper number Model parameter estimates, Standard deviations and p-values for t-tests						
Genotype	Intercept			Slope		
	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value
E.dunnii	3.231	0.336	<0.0001*	-0.504	0.061	<0.0001*
E.grandis	2.847	0.336	<0.0001*	-0.466	0.061	<0.0001*
E.smithii	2.954	0.291	<0.0001*	-0.467	0.043	<0.0001*
E.nitens	2.621	0.336	<0.0001*	-0.402	0.061	<0.0001*
GCG	3.050	0.336	<0.0001*	-0.507	0.061	<0.0001*
GUA	2.910	0.336	<0.0001*	-0.478	0.061	<0.0001*
GUW	2.549	0.237	<0.0001*	-0.401	0.061	<0.0001*

*significant parameter at 5% significance level

The covariance matrix for the slope and intercept parameters for all genotypes is given as

$$Cov(b_{0g}, b_{1g}) = \begin{bmatrix} 0.00367 & -0.00317 \\ -0.00317 & 0.00460 \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}) = \begin{bmatrix} 1.0000 & -0.7715 \\ -0.7715 & 1.0000 \end{bmatrix}.$$

There is a strong negative correlation between the slope and the intercept parameters for the genotypes which, as with the other variables, suggests that the initial readings of copper numbers affect the evolution of the variable over the processing stages. Pairwise comparisons of intercept and slope parameters for the seven genotypes are presented in Table 3.11 below. Only EDunnii and GUW have significantly different intercepts (p-value=0.045) and none of the variables have significantly different slope parameters (all p-values>0.05).

Table 3.11: Intercept and slope parameter estimated differences for the random coefficient regression model for copper number.

Genotype	Parameter Estimates		Differences in intercepts and slopes for copper number (t-tests *p*-values in brackets)											
			E.grandis		E.nitens		Esmithii		GUW		GCG		Edunnii	
			Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value
E.grandis	Intercept	2.847	-											
	Slope	-0.466	-											
E.nitens	Intercept	2.621	0.226	0.503	-									
	Slope	-0.402	-0.064	0.456	-									
Esmithii	Intercept	2.954	-0.108	0.712	0.333	0.255	-							
	Slope	-0.467	0.000	0.995	-0.065	0.386	-							
GUW	Intercept	2.549	0.298	0.377	0.072	0.830	0.406	0.166	-					
	Slope	-0.401	-0.065	0.455	0.000	0.998	-0.065	0.385	-					
GCG	Intercept	3.050	-0.204	0.546	-0.429	0.204	-0.096	0.742	0.502	0.138	-			
	Slope	-0.507	0.041	0.634	0.106	0.223	0.041	0.587	-0.106	0.222	-			
Edunnii	Intercept	3.231	0.384	0.256	0.610	0.073	0.276	0.344	0.682	0.045	0.180	0.593	-	
	Slope	-0.504	-0.038	0.661	-0.102	0.238	-0.037	0.617	-0.103	0.237	0.003	0.970	-	
GUA	Intercept	2.910	-0.063	0.851	-0.289	0.391	0.044	0.880	0.362	0.284	0.140	0.677	0.320	0.342
	Slope	-0.478	0.012	0.888	0.077	0.376	0.012	0.876	-0.077	0.375	-0.029	0.737	-0.026	0.765

3.3.7. Random coefficient models for Glucose

The model to describe the trajectory that describes the evolution of glucose can be decided based on the profile plots in Figure 2.7. As the plots do not exhibit any serous departure from linearity the linear random coefficient model was fitted to the data with the results presented in Table 3.12 below. All the slope parameters are positive and significantly different from zero indicating that there are significant changes in glucose over the processing stages.

Table 3.12 Parameter estimates for the random coefficient regression model for glucose.

Glucose model parameter estimates, Standard deviations and p-values for t-tests						
Genotype	Intercept			Slope		
	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value
E.dunnii	89.629	0.733	<0.0001*	1.146	0.133	<0.0001*
E.grandis	92.197	0.733	<0.0001*	0.851	0.133	<0.0001*
E.smithii	90.317	0.645	<0.0001*	0.970	0.097	<0.0001*
E.nitens	89.989	0.733	<0.0001*	1.161	0.133	<0.0001*
GCG	90.113	0.733	<0.0001*	1.028	0.133	<0.0001*
GUA	90.020	0.733	<0.0001*	1.111	0.133	<0.0001*
GUW	92.834	0.518	<0.0001*	0.742	0.133	<0.0001*

**significant parameter at 5% significance level*

The covariance matrix for the slope and intercept parameters for all genotypes is given as

$$Cov(b_{0g}, b_{1g}) = \begin{bmatrix} 0.01769 & -0.00916 \\ -0.00916 & \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}) = \begin{bmatrix} 1.0000 & -0.9433 \\ -0.9433 & 1.0000 \end{bmatrix}.$$

There is a strong negative correlation between the slope and the intercept parameters for the genotypes (r=-0.9433) which suggests that the higher initial readings of glucose the slower the rate of change over the processing stages. This could be because glucose cannot be expected to increase forever hence if it is high already then there is less room for more increase. Results in Table 3.13 show that there is significant difference in rate of change of glues between GUW and ENitens (d=0.419, p-value=0.028) and GUW and EDunnii (d=0.404, p-value=0.034). It would be wise not to mix these differing genotypes during processing if this difference is due to different chemical requirements.

Table 3.13: Intercept and slope parameter estimated differences for the random coefficient regression model for glucose.

Genotype	Parameter Estimates		Differences in intercepts and slopes for glucose (t-tests *p*-values in brackets)											
			E.grandis		E.nitens		Esmithii		GUW		GCG		Edunnii	
			Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value
E.grandis	Intercept	92.197	-											
	Slope	0.851	-											
E.nitens	Intercept	89.989	2.208	0.003	-									
	Slope	1.161	-0.311	0.102	-									
Esmithii	Intercept	90.317	1.880	0.004	0.328	0.612	-							
	Slope	0.970	-0.120	0.470	-0.191	0.249	-							
GUW	Intercept	92.834	-0.637	0.387	-2.845	0.000	-2.517	0.000	-					
	Slope	0.742	0.109	0.565	0.419	0.028	0.228	0.169	-					
GCG	Intercept	90.113	2.083	0.006	-0.124	0.866	0.203	0.753	-2.721	0.000	-			
	Slope	1.028	-0.177	0.348	0.133	0.481	-0.058	0.726	0.286	0.132	-			
Edunnii	Intercept	89.629	-2.567	0.001	-0.360	0.625	-0.687	0.289	-3.205	<.0001	-0.484	0.511	-	
	Slope	1.146	0.295	0.120	-0.015	0.936	0.176	0.289	0.404	0.034	0.118	0.533	-	
GUA	Intercept	90.020	2.176	0.004	-0.031	0.966	0.296	0.647	-2.814	0.000	0.093	0.899	-0.391	0.595
	Slope	1.111	-0.261	0.170	0.050	0.791	-0.141	0.395	0.369	0.053	-0.083	0.660	0.035	0.854

3.3.8. Random coefficient models for Xylose

The model to describe the trajectory that best describes the evolution of xylose can be decided based on the profile plots in Figure 2.8. The plots do not exhibit any serous departure from linearity hence the linear random coefficient model was fitted to the data with the results presented in Table 3.14 below. All the slope parameters are negative and significantly different from zero indicating that there are significant changes in xylose over the processing stages.

Table 3.14 Parameter estimates for the random coefficient regression model for Xylose.

Xylose model parameter estimates, Standard deviations and p-values for t-tests						
Genotype	Intercept			Slope		
	Parameter	Std Dev	p-value	Parameter	Std Dev	p-value
E.dunnii	5.005	0.361	<0.0001*	-0.531	0.064	<0.0001*
E.grandis	3.560	0.361	<0.0001*	-0.353	0.064	<0.0001*
E.smithii	5.085	0.317	<0.0001*	-0.517	0.047	<0.0001*
E.nitens	5.657	0.361	<0.0001*	-0.669	0.064	<0.0001*
GCG	3.873	0.361	<0.0001*	-0.294	0.064	<0.0001*
GUA	4.662	0.361	<0.0001*	-0.464	0.064	<0.0001*
GUW	3.189	0.255	<0.0001*	-0.328	0.064	<0.0001*

**significant parameter at 5% significance level*

The covariance matrix for the slope and intercept parameters for all genotypes is given as

$$Cov(b_{0g}, b_{1g}) = \begin{bmatrix} 0.0084 & 0.0140 \\ 0.0140 & 0.0256 \end{bmatrix},$$

with the correlation matrix,

$$Corr(b_{0g}, b_{1g}) = \begin{bmatrix} 1.0000 & 0.9547 \\ 0.9547 & 1.0000 \end{bmatrix}.$$

There is a strong positive correlation between the slope and the intercept parameters for the genotypes (r=0.9547) which suggests that the higher the initial readings the higher the rate of change over the processing stages. Results in Table 3.15 show that there is significant difference in rate of change of xylose between EGrandis and ESmithii, (d=0.164, p-value=0.042), EGrandis and ENitens (d=0.315, p-value=<0.001), ENitens and GUW (d=-0.340, p-value=<0.0001), ESmithii and GUW (d=-0.189, p-value=0.019), GUW and EDunii (d=-0.203, p-value=0.028), WSmithii and GCG (d=-0.224, p-value=0.006), ENitens and GUA (d=-0.204, p-value=0.027) and lastly GCG and EDunnii (d=-0.238, p-value=0.010). It would be advisable not to mix these differing genotypes during processing.

Table 3.15: Intercept and slope parameter estimated differences for the random coefficient regression model for xylose.

Genotype	Parameter Estimates		**Differences in intercepts and slopes for xylose** (t-tests *p*-values in brackets)											
			E.grandis		E.nitens		Esmithii		GUW		GCG		Edunnii	
			Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value	Difference	p-value
E.grandis	Intercept	3.560	-											
	Slope	-0.353	-											
E.nitens	Intercept	5.657	-2.097	<.0001	-									
	Slope	-0.669	0.315	<.0001	-									
Esmithii	Intercept	5.085	-1.526	<.0001	-0.572	0.075	-							
	Slope	-0.517	0.164	0.042	0.151	0.059	-							
GUW	Intercept	3.189	0.371	0.307	2.468	<.0001	1.896	<.0001	-					
	Slope	-0.328	-0.025	0.783	-0.340	<.0001	-0.189	0.019	-					
GCG	Intercept	3.873	-0.314	0.387	1.784	<.0001	1.212	0.000	0.684	0.061	-			
	Slope	-0.294	-0.060	0.512	-0.375	<.0001	-0.224	0.006	0.035	0.704	-			
Edunnii	Intercept	5.005	1.446	0.000	-0.652	0.074	-0.080	0.802	1.816	<.0001	1.132	0.002	-	
	Slope	-0.531	-0.178	0.053	0.137	0.133	-0.014	0.860	-0.203	0.028	-0.238	0.010	-	
GUA	Intercept	4.662	-1.102	0.003	0.995	0.007	0.423	0.185	1.473	<.0001	-0.789	0.031	0.343	0.344
	Slope	-0.464	0.111	0.224	-0.204	0.027	-0.053	0.508	-0.136	0.136	0.171	0.063	-0.067	0.463

3.4. Conclusion

The random coefficient model sought to look at the family of parameters of the models that were fitted to the seven genotypes to describe the behaviour of seven chemical properties of dissolving pulp going through six processing stages. The random coefficient model explored the variations in the parameter estimates across the seven genotypes and compared them as well as comparing how model parameters of the same genotype relate to each other.

An important result coming from fitting this model to data for the seven chemical properties of dissolving pulp is that the higher the raw stage readings the higher the rate of change in the chemical properties over the processing stages. This result means that the system makes more efficient use of the bleaching chemicals in dealing with samples that start off with high readings at the raw stage. This implies that genotypes which start off with similar readings respond in a similar manner to the chemical pulping process. Such genotypes can be mixed during processing.

The high correlations between the intercepts and the rates of change of all variables over the processing stages (slopes) indicate strong relationships between raw pulp readings and rates of change over the processing stages. This means that the system is more efficient when processing genotypes that start off with high readings of the chemical properties, which might be pointing to the fact that genotypes with low values at the raw pulp stage require lesser chemical concentrations as excess chemicals are not utilised to the extent they are utilised by genotypes with higher raw stage readings.

Since slope parameters measure the rates of changes of the chemical properties over the processing stages, it would be of interest to order the genotypes according to their slope parameters so as to have an overall idea the rate at which they evolve over the processing stages. A suggested method would be to rank all the genotypes according to the magnitude of their rates of change for each variable and then find the average ranks across all seven chemical property variables. Table 3.16 presents these rankings, of which the average rank would be used as a crude similarity index.

Table 3.16: Chemical properties' slope rankings for all genotypes.

Genotype	Viscosity		Lignin		a-cellulose		y-cellulose		copper no		Glucose		Xylose		Average Rank
	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	
E.dunnii	-5.896	6	-1.857	3	0.887	4	-0.803	2	-0.504	6	1.146	6	-0.531	6	**4.71**
E.grandis	-1.995	1	-1.869	4	0.904	5	-0.845	5	-0.466	3	0.851	2	-0.353	3	**3.29**
E.smithii	-3.569	3	-2.572	7	0.689	1	-0.832	4	-0.467	4	0.970	3	-0.517	5	**3.86**
E.nitens	-2.122	2	-1.407	1	1.006	7	-0.940	6	-0.402	2	1.161	7	-0.669	7	**4.57**
GCG	-5.677	5	-2.515	6	0.811	3	-0.817	3	-0.507	7	1.028	4	-0.294	1	**4.14**
GUA	-7.926	7	-2.352	5	0.933	6	-0.943	7	-0.478	5	1.111	5	-0.464	4	**5.57**
GUW	-5.177	4	-1.568	2	0.804	2	-0.720	1	-0.401	1	0.742	1	-0.328	2	**1.86**

The rankings are based on the absolute values of the slope parameters which measure the rate of change in the direction the variables are expected to change. The results show that GUW (mean rank=1.86) is the least mixable genotype as it tends to have the lowest rate of change which is not comparable to any of the other genotypes. The genotype GUA (mean rank=5.57), on the other hand, stands at the extreme of being the most process responsive genotype. Its closest mixable partner is E.dunii (mean rank=4.71). Using this simple logic, decisions could be made on how mixable the different genotypes are. This can be very helpful with more complex systems that involve higher numbers of varieties in raw materials that feed into their processes.

This chapter attempted to fit an overall linear model across all stages without regards to what each stage is meant to achieve in the chemical processing scheme. This might be the reason why some subtle differences in the behaviours of the chemicals properties might not have been captured. Chapter 4 that follows, attempts to attribute each stage to the three distinct sub-processes inherent in the chemical processing scheme. The benefits of Chapter 4 would be to identify differences in behaviour of the genotypes within the process rather than offer an outside view of the overall process. Piecewise linear regression would model each sub-process as a linear component of a much bigger nonlinear process.

The limitations of the study were mainly the consideration of processing stages as time points as there was no controlled time lapse between stages. The stages are therefore points of measurements which are not necessarily on an interval scale.

The residual plots for the models fitted in sections 3.3.2 to 3.3.8 show that the residuals exhibit normality hence the assumption of normality an holds. The models can also be deemed adequate based on the residual plots (See Figures A1.1 to A1.7).

Chapter 4

Piecewise Linear Regression Models with Dummy Time Variables

4.1. Introduction

Chemicals used in pulp processing are costly, it is necessary to optimise the usage of such chemicals by identifying and combining wood species/genotypes with similar chemical properties under the chemical pulping process for maximum utilization of processing resources and uniformity of the final product. This chapter identifies tree genotypes that exhibit similar processing behavior by modeling the chemical properties of dissolving pulp at all the processing stages using piecewise linear regression models. Piecewise linear regression models are deemed appropriate for this data since there are three known sub-processes, in series, in the chemical processing of dissolving pulp, namely, delignification, bleaching and finishing. Species/genotypes with similar rates of change (or slopes) in the chemical properties under consideration during the three sub-processes will be mixed together during processing in the future if economic quantities cannot be achieved with just one species/genotype. Species/genotype which differ significantly in the way they respond to the processing stages would better be processed separately. It is expected that each of the three sub-processes will have a different effect on the response variables hence the resultant models are three segment, piecewise linear regression models. In this study the response variables are seven important chemical properties of dissolving pulp and the independent variable is the stage of processing.

The main objective in the processing of dissolving pulp is to remove lignin, retaining *α*-cellulose while at the same time maintaining other properties like viscosity within certain product specified limits. The modeling of the evolutions of lignin, viscosity, γ-cellulose, α-cellulose copper number, glucose and xylose over the processing stages, for each of the seven genotypes, will highlight the differences in species/genotype responses to chemical processing.

4.2. Graphical presentation of chemical properties over processing stages

Graphs of genotype means by stage of processing, for each chemical property are presented to show genotypes that are close together or behave in a similar manner during processing. Figure 4.1 below shows the percentage content of α-cellulose as the processing stages unfold. It is apparently clear from Figure 4.1 that α-cellulose percentages increase from the first to the last stage for all genotypes. Generally stage D_1 has the effect of slightly reducing the α-cellulose level for all species/genotypes.

The relationship between α-cellulose content and processing stage is not easy to generalize for all species/genotypes hence the piecewise linear regression method discussed in this chapter seeks to better describe the patterns in the data. Different species/genotypes are expected to have varying model parameters for the piecewise linear regression model. Species/Genotypes with parameters that do not differ significantly can be classified as having similar evolutions across the processing stages of the chemical pulping process and such species/genotypes will require similar amounts of chemicals in each of the six processing stages.

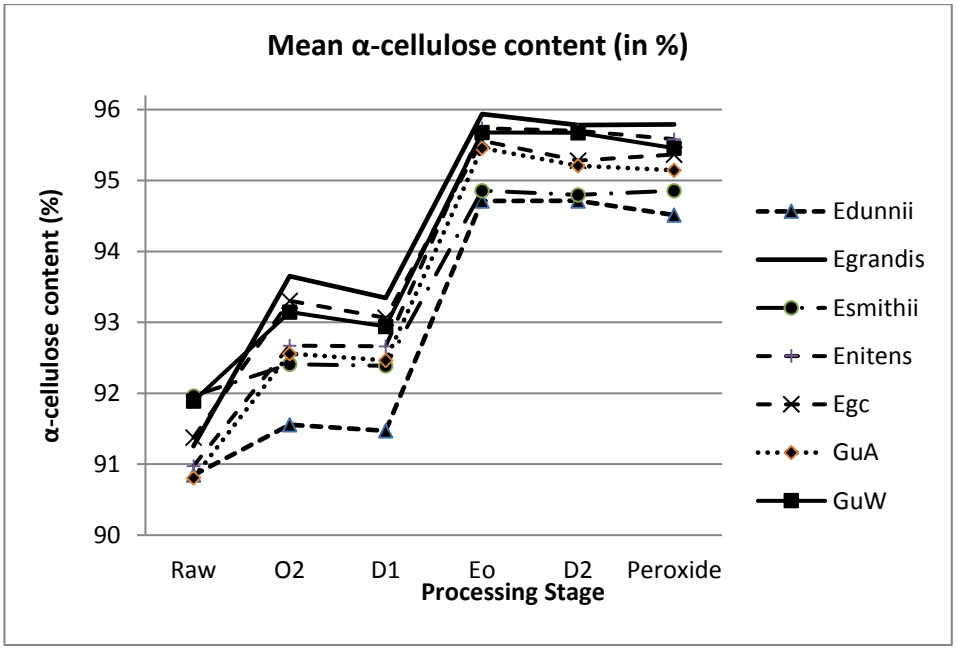


Figure 4.1. Mean α-cellulose content (in %) by stage for different Genotypes

Figure 4.2 shows how γ-cellulose levels for different species/genotypes evolve over the processing stages. The graph for γ-cellulose (Figure 4.2) is more of an inverted

version of the graph of α-cellulose (Figure 4.1). This is due to the fact that α-cellulose is closely associated with γ-cellulose and degraded cellulose.

The viscosity profiles for the various species/genotypes as shown in Figure 4.3 indicate a general declining trend over the processing stages. The process is designed to reduce the viscosity of the product until ideal pulp characteristics are achieved. A final product with pulp characteristics outside the product specific margins is discarded.

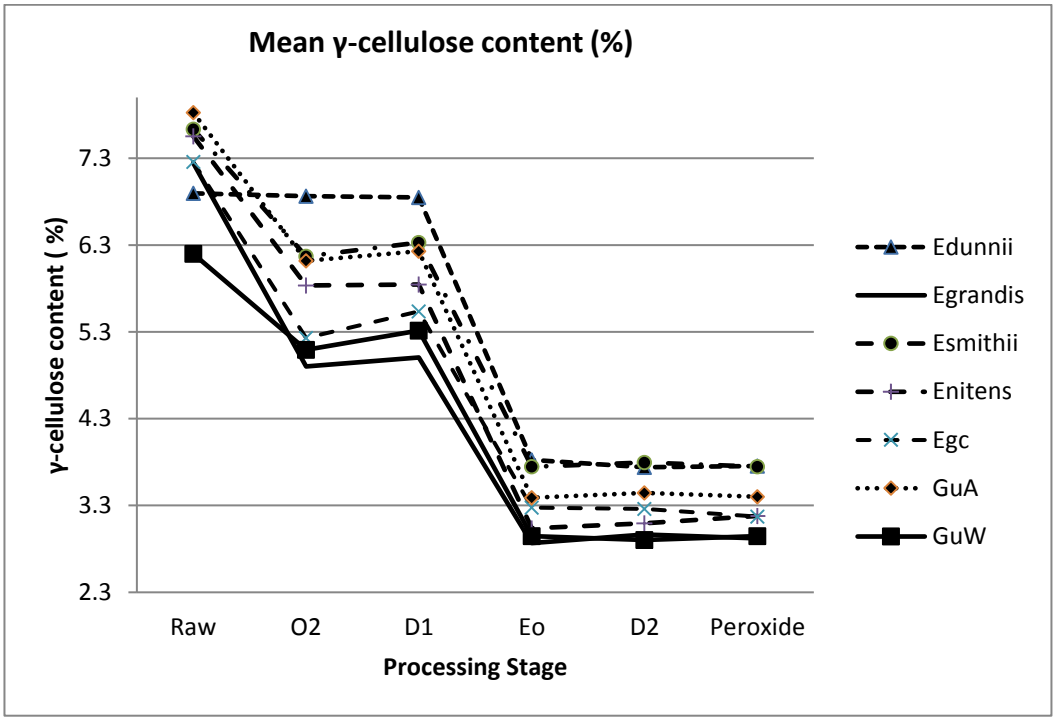


Figure 4.2. Mean γ-cellulose content (in %) by stage for different Genotypes

Lignin content for all species/genotypes under study decreases over the processing stages as shown in Figure 4.4.

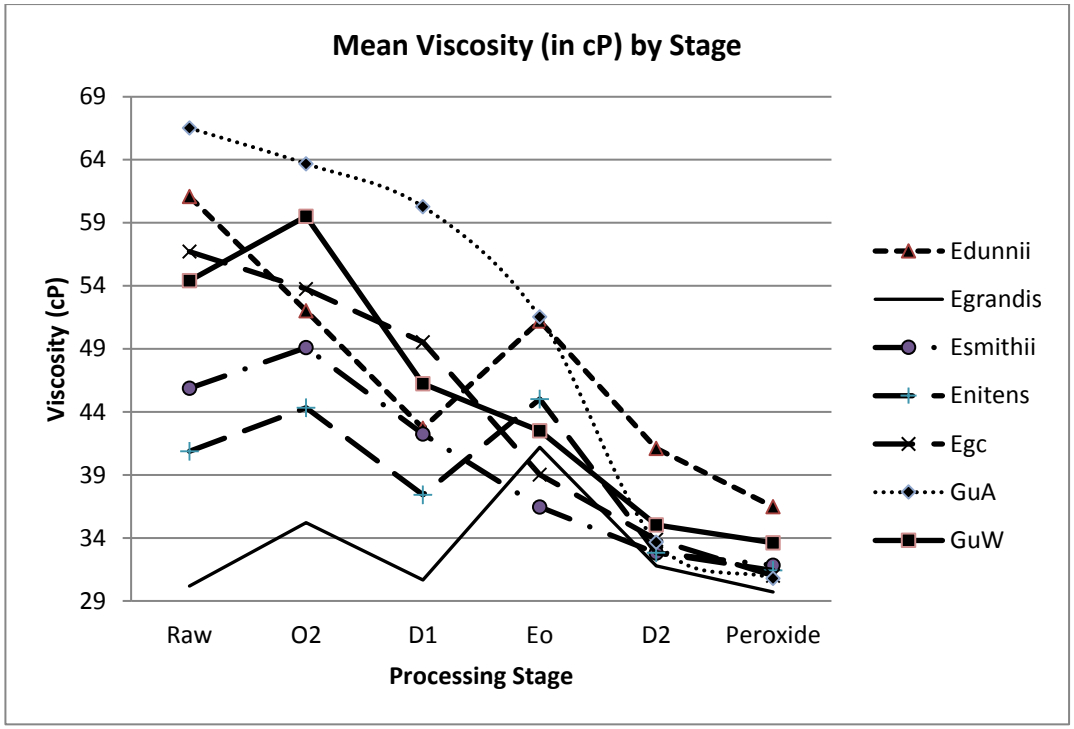


Figure 4.3. Mean viscosities by stage for different genotypes

The decrease in lignin is not linear over the six stages but can be piecewise linear if the stages are grouped into sub-processes, namely delignification, bleaching and finishing.

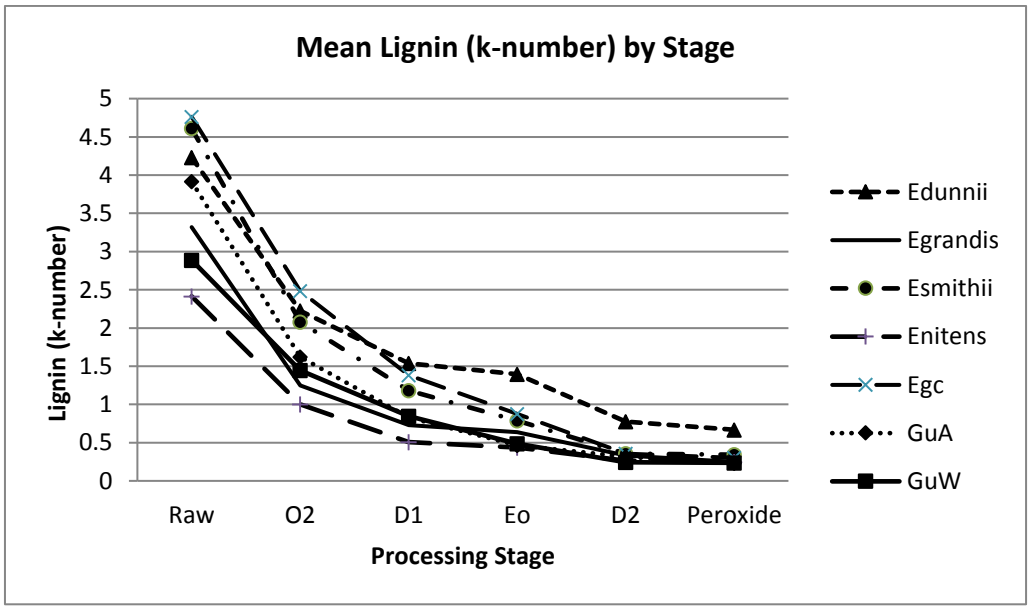


Figure 4.4. Mean Lignin content by stage for different Genotypes.

Figure 4.5 shows that copper numbers decrease with each processing stage with the O_2 and Eo stages accounting, to a larger extent, for the decrease in copper numbers. It is also clear from this graph that the species/genotypes do not vary much in their copper numbers as the lines are very close together.

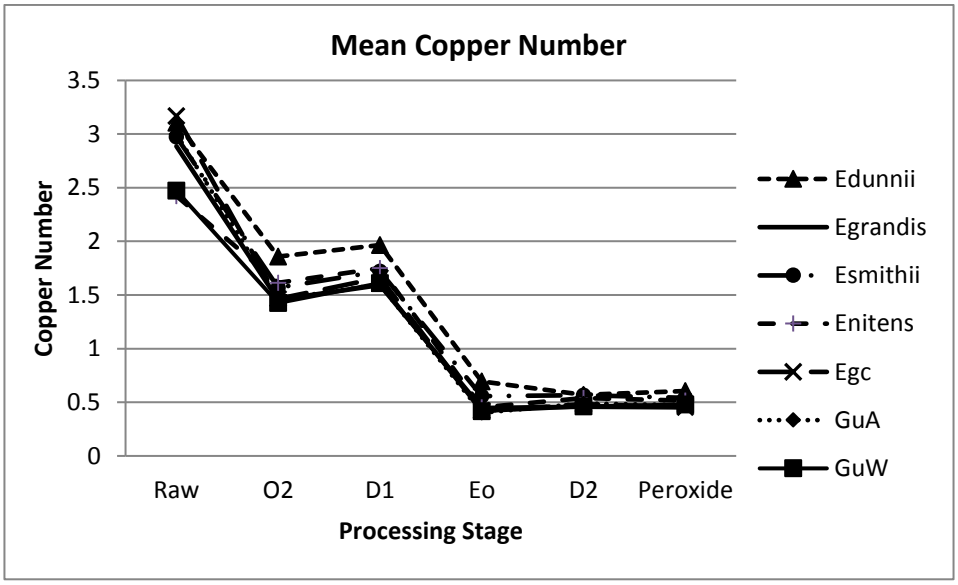


Figure 4.5. Mean Copper numbers by stage for different genotypes

Mean glucose levels, as indicated in Figure 4.6, increase as the processing stages unfold with *E.grandis* and *GuW* having higher glucose levels across the stages than the other five species/genotypes.

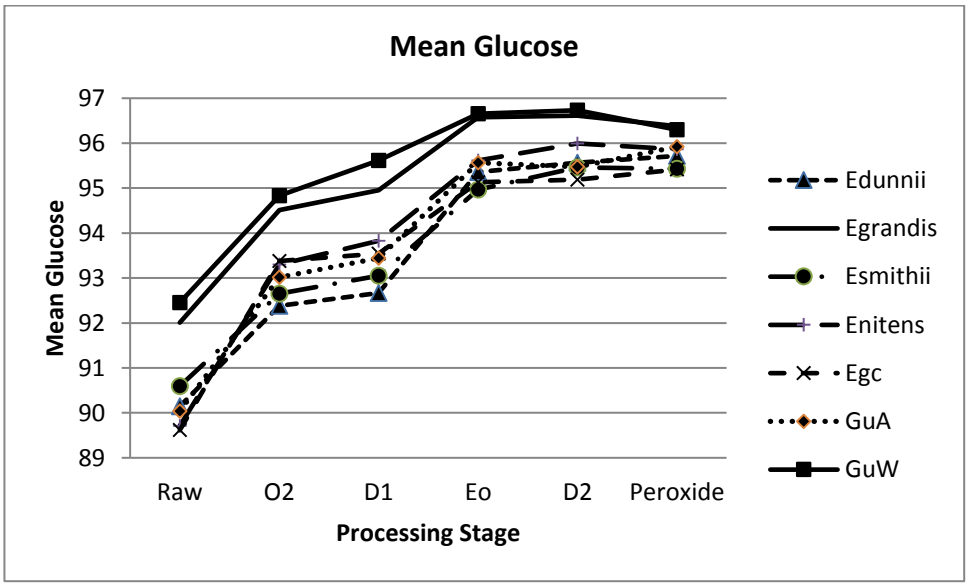


Figure 4.6. Mean Glucose by stage for different Genotypes

Figure 4.7 shows the changes in xylose over the processing stages. It was observed that mean xylose levels decrease as the processing stages unfold with *E.grandis* and *GuW* having closer and lower means by stage. These two species/genotypes also had very similar α-cellulose, γ-cellulose, lignin and copper numbers levels. Based on this similarity the two genotypes can be deemed mixable during processing.

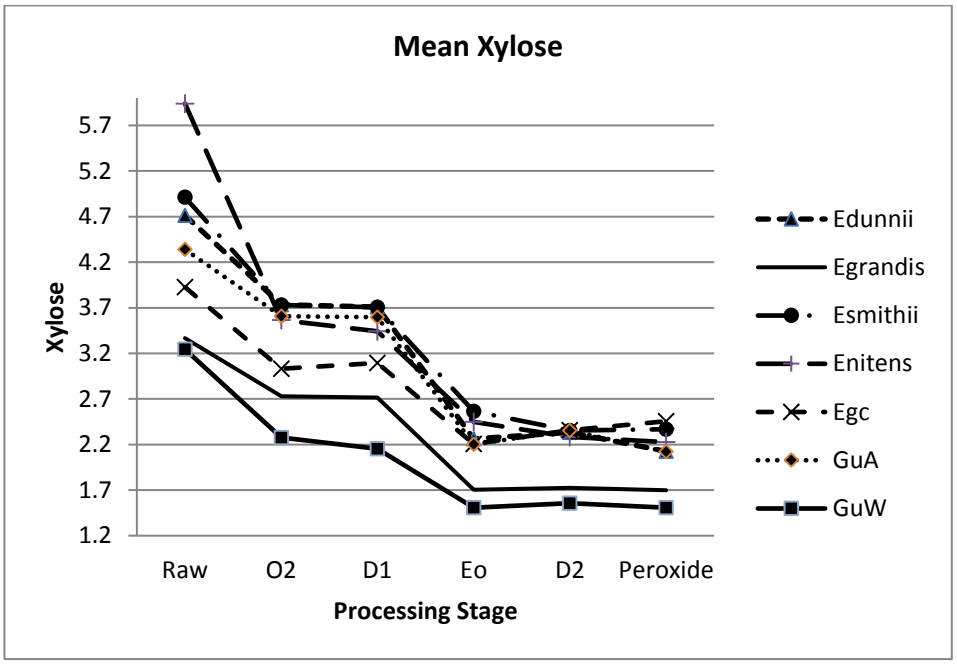


Figure 4.7. Mean Xylose by stage for different genotypes

Figures 4.1 to 4.7 suggest that in general, the trajectories of the seven chemical properties profiled in these graphs can best be modelled using nonlinear models of some kind as all graphs exhibit nonlinear relationships with stage (time). This, together with the knowledge of the inherent three processing stages in chemical pulping leads to the suggestion that a three part piecewise linear regression model be attempted for the seven response variable with stage as the independent variable.

4.3. The Piecewise Linear Regression Model

The piecewise linear model is part of the mixed modelling framework (Bryk and Raudenbush, 1992; Snijders and Bosker, 1999). The basic principle of fitting a piecewise linear model is to identify transition points in the data and fit linear functions between such points. The transition points maybe determined by the theoretical background of the problem (Bollen, 2006) and in this study the stages at which a

transition is made from one sub-process to the next in the chemical pulping process are such transition points.

The piecewise linear regression model is an additive function of an intercept, a linear component leading to the next transition point, following by more linear components separated by the transition points that have been decided upon. Bollen (2006) used the manipulation of the independent variable to come up with a piecewise linear model without having to mention the transition points at all times. The idea is to say that if we have three linear parts in the model, that is, if we have two transition points, then the piecewise linear model can be specified as

$$Y_i = \beta_{0i} + \beta_{1i}t_1 + \beta_{2i}t_2 + \beta_{3i}t_3 + \varepsilon_i. \tag{4.1}$$

The model parameters in 4.1 vary from subject to subject or from genotype to genotype hence they are considered as random effects which can be treated in a similar manner as in the random coefficient model, that is,

$$\begin{aligned} \beta_{0i} &= \beta_0 + b_{0i} \\ \beta_{1i} &= \beta_1 + b_{1i} \\ \beta_{2i} &= \beta_2 + b_{2i} \\ \beta_{3i} &= \beta_3 + b_{3i} \\ \text{with } b_{ki} &\sim N\left(0, \sigma_{b_k}^2\right) \end{aligned} \tag{4.2}$$

The data representations in Figures 4.1 to 4.7 show that there is a degree of non-linearity in the data thus the piecewise linear regression approach is a useful method to model such data using two or more piecewise linear splines (Bollen and Curran, 2006). It is more appropriate to use this method for the pulp processing data since there are three known basic sub-processes in the whole chemical pulping process and there will be a spline for each of these sub-processes in the model. The transition points or knots are points where the parameters of the model change from one spline to the other giving the model a broken stick appearance (Fitzmaurice et al., 2004). The three sub-processes are:

(i) delignification,
(ii) bleaching and
(iii) finishing (peroxide stage)

The delignification sub-process is activated at the O stage followed by the bleaching sub-process spanning stages D_1, E_o and D_2, and the finishing sub-process which is activated at the last stage (where peroxide is used). The variable t_1 is used to represent the delignification sub-process, t_2 for the bleaching and t_3 for the finishing sub-processes. The values of t_1, t_2 and t_3 for each sub-process are as defined in Table 4.1 below. The whole process can be described in terms of t_1, t_2 and t_3 by equation 4.3 below:

$$Y_i = \begin{cases} \beta_{0i} + \beta_{1i} t_1 + \varepsilon_i & \text{Delignification} \\ (\beta_{0i} + \beta_{1i}) + \beta_{2i} t_2 + \varepsilon_i & \text{Bleaching} \\ (\beta_{0i} + \beta_{1i} + 3\beta_{2i}) + \beta_{3i} t_3 + \varepsilon_i & \text{Finishing,} \end{cases} \tag{4.3}$$

where the term $(3\beta_{2i})$ in the expression for “finishing ” accounts for the fact that there are three bleaching stages. The response variable Y_i is the pulp characteristic of interest of which seven are modelled independently in this study, viz., viscosity, lignin, γ-cellulose, α-cellulose, copper numbers, glucose and xylose. It is assumed that the error terms (ε_i ’s) within each pulp sample are correlated at different processing stages according to a suitable covariance structure which will be determined by choosing one with the lowest AIC value (Bozdogan, 1987).

Chemical pulp processing is a continuous process and measurements, on the seven chemical properties under study, were taken at six time points or stages. The knots of the piecewise regression model are set as the stages at which a different sub-process starts. This is so because each sub-process has a different effect on the chemical properties. There are two knots that separate the three sub-process and these are stages 2 and 5 as indicated in Table 4.1. Instead of fitting the piecewise linear regression model with one covariate (stage as indicated in Table 4.1) together with two indicator variables for the two knots, time or stage is recoded into three time variates. Each of the new time variates is set to zero for the starting point of each process and increases by a unit for each progression of the process.

Table 4.1. Values of t for the three main chemical sub-processes in dissolving pulp

Stage	t_1 (Delignification)	t_2 (Bleaching)	t_3 (Finishing)
Raw	0	0	0
O	1	0	0
D_1	1	1	0
E_0	1	2	0
D_2	1	3	0
Finishing(P)	1	3	1

In model (4.3) above β_{1i}, β_{2i} and β_{3i} are rates of change of the response variable due to delignification, bleaching and the finishing stage respectively. Since delignification occurs during the two initial stages, that is, at the raw pulp and the O stages, to be followed by bleaching thereafter, we let $t_1 = 0$ for the raw stage and $t_1 = 1$ from the O stage up to the finishing stage as the delignification sub-process ends at the O stage, with $t_1 = 1$. The bleaching sub-process begins at stage D_1 ($t_2 = 1$) and continues in stages E_0 ($t_2 = 2$) and D_2 ($t_2 = 3$). At the finishing stage there is no bleaching occurring so t_2 remains unchanged at $t_2 = 3$ and on the last stage of the process. A value of $t_i = 0$ for i=1, 2 or 3, indicates that chemical sub-process t_i has not been activated and if t_i remains constant for subsequent stages then the chemical sub-process ascribed to t_i has stopped. For example t_1 remains at $t_1 = 1$ for stages O, D_1, E_o, D_2 and the finishing stage because it is activated only at stage O, ends at stage D_1 and does not occur in subsequent stages.

The intercept of the delignification sub-process is β_{0i} with slope parameter β_1 and the intercept of the bleaching sub-process is $(\beta_{0i} + \beta_{1i})$ since these are the predicted values of the response variable when delignification and bleaching start respectively. In the same way, the intercept of the finishing sub-process is $\beta_{0i} + \beta_{1i} + 3\beta_{2i}$. Model (4.3) together with the values of t_1, t_2 and t_3, as outlined in Table 4.1, can be generalised as

$$E[Y_i] = \beta_{0i} + \beta_{1i}t_1 + \beta_{2i}t_2 + \beta_{3i}t_3. \qquad (4.4)$$

where β_{0i} (the delignification intercept) is the initial value of the response variable at the raw stage. The parameters β_{1i}, β_{2i} and β_{3i} can be compared for different

species/genotypes to see which species/genotypes have the same response rates to the three sub-processes in chemical pulp processing.

Figure 4.8 below, illustrates how the time points have been defined to indicate the three sub-processes in the chemical pulping process. The knots (1 and 2) indicate when one sub process ends and the next one starts.

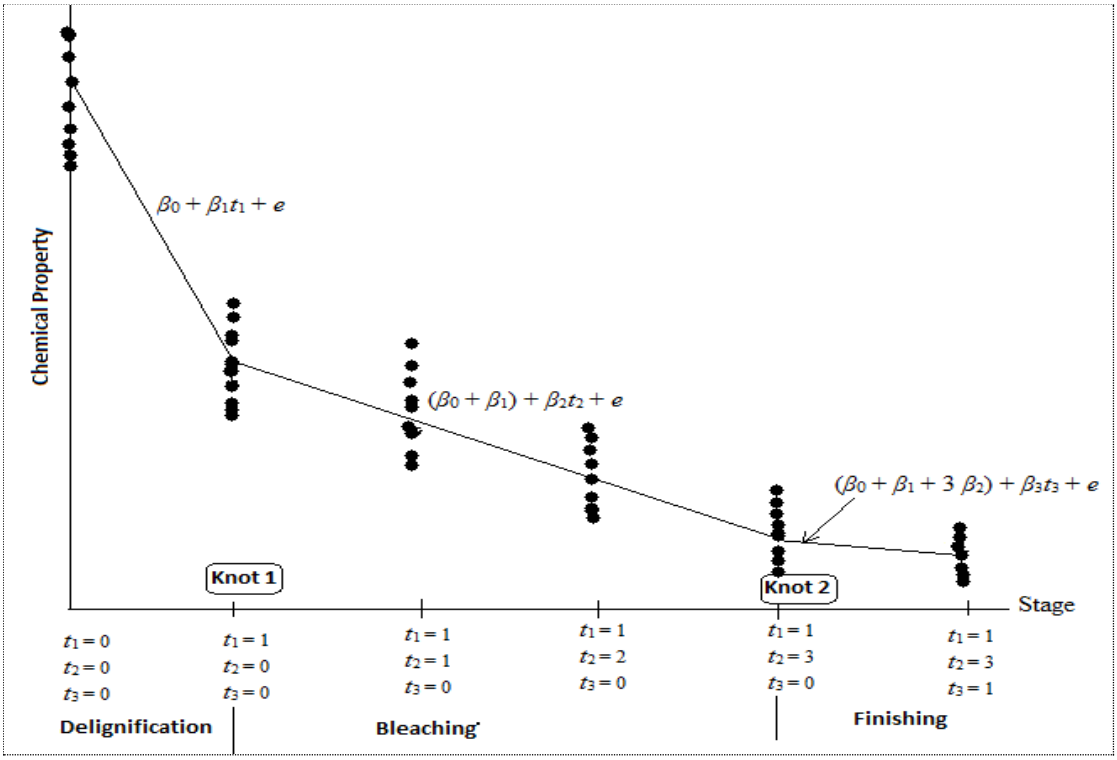


Figure 4.8. Piecewise regression lines for the chemical pulping process

4.4. Fitting the Piecewise Linear Regression Model to the chemical pulp properties data

The SAS procedure Proc Mixed was used to analyse the data and the results are presented in the sections that follow. The procedure Proc Mixed in SAS has the versatility to be used for the computation of parameter estimates for various models that include repeated measures models such as random coefficient models and piecewise regression models (Dechateau et al., 1998).

Piecewise regression models were fitted to the data for viscosity, lignin, α-cellulose, γ-cellulose, copper number, glucose and xylose in order to compare the response patterns of the seven species/genotypes. The purpose of fitting these models is to analyse the effects of each of the three main sub-processes namely delignification,

bleaching and finishing (peroxide stage) on the dissolving pulp chemical properties mentioned above. The choice of the covariance structure for the six processing stages was done after considering a few commonly used covariance structures, the results of which are presented in Table 4.2. Table 4.3 is a summary of Analysis of Variance (ANOVA) tests carried out to evaluate if the various species/genotypes have significantly different initial values and response characteristics to delignification, bleaching and finishing.

The rate of change in a chemical property due to any stage of a sub-process is represented by the slope parameter estimate of the sub-process. In this study β_0 is the intercept or raw stage value, β_1 is the rate of change of a chemical property due to delignification, β_2 is the rate of change due to bleaching and β_3 is the rate of change due to the finishing sub-process. Rates of change of the seven chemical properties discussed in this study are presented in Tables 4.2 to 4.10.

4.4.1. Viscosity data

For viscosity, the unstructured covariance structure had the lowest AIC value (Table 4.2: AIC=863.7). In fact the unstructured covariance structure was fitted for all the chemical properties studied as it had the lowest AIC values for all chemical properties.

Table 4.2. AIC values for different covariance structures for the piecewise regression models

Covariance Structure	Number of parameters	Wet chemistry property (AIC values)						
		Viscosity	Lignin	α-cellulose	γ-cellulose	Copper number	Glucose	Xylose
Unstructured	7	**863.7***	**67.3***	**372.4***	**284.9***	**65.2***	**276.5***	**173.5***
ANTE(1)	6	869.6	-	398.0	314.8	99.3	281.5	179.5
AR(1)	3	885.2	107.7	394.2	313.5	97.3	283.3	190.3
ARMA(1,1)	4	887.2	109.7	396.2	313.5	99.3	283.0	189.1
CS	3	886.0	107.7	394.2	313.5	93.8	283.3	190.3
Toeplitz	4	887.2	108.8	394.9	314.4	96.5	282.1	186.4
SP(Pow)	3	888.7	107.7	394.2	313.5	97.3	283.3	190.3
SP(Gau)	3	888.7	107.7	394.2	316.4	97.3	287.1	196.7

The seven species/genotypes had significantly different raw pulp viscosities (Table 4.3: F=205.55, df1=7, df2=65, p-value<0.000) but had no significantly different delignification slopes for viscosity (Table 4.3: F=0.22, df1=7, df2=17, p-value=0.976). The seven species/genotypes did not have significantly different bleaching slopes for

viscosity (Table 4.3: F=1.53, df1=7, df2=17, p-value<0.224) and they also did not have significantly different finishing stage viscosity slopes (Table 4.3: F=0.80, df1=7, df2=17, p-value=0.595). This means that viscosity cannot be used as a classifying variable for the species/genotypes. The mean viscosity values are shown in Figure 4.3 above.

Table 4.3. Tests for the effects of delignification, bleaching and finishing on genotype.

Effect		Viscosity	Lignin	γ-cellulose	α-cellulose	Copper number	Glucose	Xylose
Intercept by Genotype	F	205.55	808.21	329.67	23411.40	279.41	72851.0	480.62
	p-value	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
Delignification by Genotype	F	0.22	70.14	6.78	2.52	28.04	38.01	14.01
	p-value	0.976	0.000*	0.001*	0.056	0.000*	0.000*	0.000*
Bleaching by Genotype	F	1.53	15.32	29.05	15.29	31.350	41.01	26.57
	p-value	0.224	0.000*	0.000*	0.000*	0.000*	0.000*	0.000*
Finishing by Genotype	F	0.80	0.190	0.21	0.21	0.160	0.57	0.13
	p-value	0.595	0.983	0.980	0.980	0.990	0.768	0.995

Degrees of freedom for numerator=7 for all cases
Degrees of freedom for denominator=65 for intercept and 17 for delignification, bleaching and finishing

The parameter estimates for the piecewise linear regression models for the viscosity data for the seven species/genotypes are obtained from Table 4.4 as:

E.dunnii: $\hat{Y}=61.083-10.681t_1-2.427t_2-6.647t_3$

E.grandis: $\hat{Y}=30.183+4.501t_1-0.019t_2-5.028t_3$

E.smithii: $\hat{Y}=45.883-2.473t_1-5.471t_2-0.100t_3$

E.nitens: $\hat{Y}=40.882+3.062t_1-2.696t_2-4.440t_3$

E.gc: $\hat{Y}=56.713-2.143t_1-7.016t_2-2.516t_3$

E.guA: $\hat{Y}=66.517+0.592t_1-9.878t_2-6.687t_3$

E.guW: $\hat{Y}=54.413+2.986t_1-7.718t_2-0.630t_3$

From these model estimates the viscosity levels can be estimated at each processing stage by substituting the values of t_1, t_2 and t_3 as defined in Table 4.1. The *t*-tests, as indicated by the *p*-values which are all greater than 5% for delignification, bleaching and finishing, are not significant (Table 4.4: *p*-values>0.050) which indicates that no specific sub-process reduces viscosity significantly for all seven species/genotypes.

This means that viscosity is reduced steadily across the three sub-processes without any particular sub-process reducing viscosity significantly.

Table 4.4. Piecewise linear regression model parameter estimates and t-tests for viscosity

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	*t*-test (df=65) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value
Edunnii	61.083	15.94	-10.681	-1.02	-2.427	-0.47	-6.647	-1.33
	(3.832)	0.000*	(10.516)	0.320	(5.114)	0.641	(4.996)	0.201
Egrandis	30.183	7.88	4.501	0.43	0.019	0.00	-5.028	-1.01
	(3.832)	0.000*	(10.516)	0.674	(5.114)	0.997	(4.996)	0.328
Esmithii	45.883	16.93	2.473	0.33	-5.471	-1.51	-0.100	-0.03
	(2.710)	0.000*	(7.436)	0.744	(3.616)	0.149	(3.533)	0.978
Enitens	40.882	10.67	3.062	0.29	-2.696	-0.53	-4.440	-0.89
	(3.832)	0.000*	(10.516)	0.774	(5.114)	0.605	(4.996)	0.387
E gc	56.713	14.80	-2.143	-0.20	-7.016	-1.37	-2.516	-0.50
	(3.832)	0.000*	(10.516)	0.841	(5.114)	0.188	(4.996)	0.621
EguA	66.517	17.36	0.592	0.06	-9.878	-1.93	-6.687	-1.34
	(3.832)	0.000*	(10.516)	0.956	(5.114)	0.070	(4.996)	0.198
EguW	54.413	14.20	2.986	0.28	-7.718	-1.51	-0.630	-0.13
	(3.832)	0.000*	(10.516)	0.780	(5.114)	0.150	(4.996)	0.901

*significant parameters at the 5% significant level

4.4.2. Lignin data

For the lignin data, the unstructured covariance structure had the lowest AIC value (Table 4.2: AIC=67.3) hence it was fitted to the data. The rate of lignin decrease by species/genotype over the sub-processes can be used to highlight the differences in the response patterns of the seven species/genotypes to the three sub-processes. Ideally most of the lignin must be removed in the delignification stage but this does not remove all the lignin to product specified levels. The species/genotypes have significantly different raw stage lignin levels (Table 4.3: F=808.21, df1=7, df2=17, *p*-value=0.000). The results in Table 4 also show that the seven genotypes have significantly different slopes for lignin at delignification (Table 4.3: F=70.14. df1=7, df2=17, *p*-value=0.000) and bleaching (Table 4.3: F=15.32, df1=7, df2=17, *p*-value=0.000). There are no significant differences among species/genotypes in lignin content due to the finishing sub-process (Table 4.3: F=0.190. df1=7, df2=17, *p*-value=0.983). The results above mean that lignin levels at the raw, delignification and bleaching stages can be used to classify species/genotypes according to their slope parameters.

The piecewise linear regression parameters estimates of lignin are summarized in Table 4.5 which shows the estimates, their standard deviations and the corresponding *t*-tests.

Table 4.5: Piecewise linear regression model parameter estimates and t-tests for Lignin

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	*t*-test (df=65) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value
Edunnii	4.230	28.62	-2.073	-7.26	-0.449	-3.50	-0.141	-0.73
	(0.148)	0.000*	(0.286)	0.000*	(0.128)	0.003*	(0.193)	0.473
Egrandis	3.319	22.46	-2.157	-7.55	-0.284	-2.21	-0.062	-0.32
	(0.148)	0.000*	(0.286)	0.000*	(0.128)	0.041*	(0.193)	0.751
Esmithii	4.609	44.10	-2.673	-13.24	-0.556	-6.12	0.074	0.54
	(0.105)	0.000*	(0.202)	0.000*	(0.091)	0.000*	(0.136)	0.593
Enitens	2.414	16.33	-1.520	-5.32	-0.227	-1.77	0.036	0.19
	(0.148)	0.000*	(0.286)	0.000*	(0.128)	0.095	(0.193)	0.854
E gc	4.763	32.22	-2.453	-8.59	-0.690	-5.37	0.062	0.32
	(0.148)	0.000*	(0.286)	0.000*	(0.128)	0.000*	(0.193)	0.753
EguA	3.917	26.50	-2.467	-8.64	-0.428	-3.34	0.066	0.34
	(0.148)	0.000*	(0.286)	0.000*	(0.128)	0.004*	(0.192)	0.737
EguW	2.887	19.54	-1.538	-5.39	-0.396	-3.09	0.077	0.40
	(0.148)	0.000*	(0.286)	0.000*	(0.128)	0.007*	(0.193)	0.695

*significant parameters at the 5% significant level

The piecewise linear regression models built from Table 7 are given as follows:

E.dunnii: $\hat{Y}=4.230-2.073t_1-0.449t_2-0.141t_3$

E.grandis: $\hat{Y}=3.319-2.157t_1-0.284t_2-0.062t_3$

E.smithii: $\hat{Y}=4.609-2.673t_1-0.556t_2+0.074t_3$

E.nitens: $\hat{Y}=2.414-1.520t_1-0.227t_2+0.036t_3$

E.gc: $\hat{Y}=4.763-2.453t_1-0.6909t_2+0.062t_3$

E.guA: $\hat{Y}=3.917-2.467t_1-0.428t_2+0.066t_3$

E.guW: $\hat{Y}=2.887-1.538t_1-0.396t_2+0.077t_3$

Lignin levels can thus be estimated by substituting the appropriate values of t_1, t_2 and t_3 for any stage of the process for each species/genotype, where t_1, t_2 and t_3 are as defined in Table 4.1. The small but positive slopes for all genotypes at the finishing stage indicate that the finishing sub-processes slightly increases lignin levels. However this lignin increase at the finishing stage is not significant as shown by the *p*-values of β_3 which are not significant for all species/genotypes (Table 4.5).

4.4.3. γ-cellulose data

For the *γ*-cellulose data, the unstructured covariance structure had the lowest AIC value (Table 4.2: AIC=284.9) hence it was used in the analysis. As with viscosity and lignin the finishing stage had no significant effect on *γ*-cellulose (Table 4.3: F=0.21, df1=7, df2=17, *p*-value=0.980). However there were significant differences in the changes in *γ*-cellulose levels among the seven species/genotypes due to delignification (Table 4.3: F=6.78, df1=7, df2=17, *p*-value=0.001) and bleaching (Table 4.3: F=29.05, df1=7, df2=17, *p*-value=0.000) sub-processes. This means that *γ*-cellulose is an important classifying variable for the seven species/genotypes.

The piecewise linear regression model parameters estimates for *γ*-cellulose are summarized in Table 4.6. Results in Table 4.6 show that *E.dunnii* does not have a significant reduction of *γ*-cellulose due to delignification (Table 4.6: β_1=0.283, *t*=0.51, df=17, *p*-value=0.619). It is the only genotype that has this behaviour out of the seven genotypes studied. The other species/genotypes had significant reductions in *γ*-cellulose levels during both delignification and bleaching.

Table 4.6. Piecewise linear regression model parameter estimates and t-tests for γ-cellulose.

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	*t*-test (df=65) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value
Edunnii	6.896	16.05	0.283	0.51	-1.240	-6.28	0.296	0.53
	(0.430)	0.000*	(0.560)	0.619	(0.197)	0.000*	(0.560)	0.604
Egrandis	7.244	16.86	-2.117	-3.78	-0.795	-4.03	0.179	0.32
	(0.430)	0.000*	(0.560)	0.002*	(0.197)	0.001*	(0.560)	0.754
Esmithii	7.635	25.13	-1.170	-2.95	-0.970	-6.96	0.195	0.49
	(0.304)	0.000*	(0.396)	0.009*	(0.140)	0.000*	(0.396)	0.628
Enitens	7.553	17.58	-1.446	-2.58	-1.103	-5.59	0.382	0.68
	(0.430)	0.000*	(0.560)	0.019*	(0.197)	0.000*	(0.560)	0.504
E gc	7.256	16.89	-1.707	-3.05	-0.816	-4.14	0.072	0.13
	(0.433)	0.000*	(0.560)	0.007*	(0.197)	0.001*	(0.560)	0.899
EguA	7.826	18.22	-1.401	-2.50	-1.086	-5.51	0.235	0.42
	(0.430)	0.000*	(0.560)	0.023*	(0.197)	0.000*	(0.560)	0.680
EguW	6.198	14.43	-0.794	-1.42	-0.894	-4.53	0.222	0.40
	(0.430)	0.000*	(0.560)	0.175	(0.197)	0.000*	(0.560)	0.697

*significant parameters at the 5% significant level

The corresponding piecewise linear regression models derived from Table 8 for the seven species/genotypes are as follows:

E.dunnii: $\hat{Y}$=6.896+0.283t_1-1.240t_2+0.296t_3

E.grandis: $\hat{Y}$=7.244-2.117t_1-0.795t_2+0.179t_3

E.smithii: $\hat{Y}$=7.635-1.170t_1-0.970t_2+0.195t_3

E.nitens: $\hat{Y}$=7.553-1.466t_1-1.103t_2+0.382t_3

E.gc: $\hat{Y}$=7.256-1.707t_1-0.816t_2+0.072t_3

E.guA: $\hat{Y}$=7.826-1.401t_1-1.086t_2+0.235t_3

E.guW: $\hat{Y}$=6.198-0.794t_1-0.894t_2+0.222t_3

The levels of *γ*-cellulose can be estimated in a similar way described above for viscosity and lignin.

4.4.4. α-cellulose data

The covariance structure with the smallest AIC value for the α-cellulose data is the unstructured one (Table 4.2: AIC=372.4) and this was fitted to the data. The seven species/genotypes start with significantly different α-cellulose levels at the raw stage (Table 4.3: F=23411.40, df1=7, df2=65, p-value=0.000) and the sub-process of delignification does not produce significantly different rates of change in α-cellulose across the seven species/genotypes (Table 4.3: F=2.52, df1=7, df2=17, p-value=0.056). The sub-process of bleaching affects the rates of change of α-cellulose levels of the different species/genotypes in a significantly different manner (Table 4.3: F=15.29, df1=7, df2=17, p-value=0.000). As with the other chemical properties discussed above, the effects of the finishing sub-process do not differ significantly across the seven species/genotypes (Table 4.3: F=0.21, df1=7, df2=17, p-value=0.980). Since the rates of change in α-cellulose levels differ among the seven species/genotypes during the bleaching sub-process, α-cellulose can be used as a classifying variable. The piecewise linear regression model parameter estimates are presented in Table 4.7 below.

Table 4.7. Piecewise linear regression model parameter estimates and t-tests for α-cellulose

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	*t*-test (df=65) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value
Edunnii	90.846	142.28	0.361	0.43	1.271	4.45	-0.508	-0.61
	(0.639)	0.000*	(0.833)	0.670	(0.286)	0.000*	(0.833)	0.550
Egrandis	91.256	142.92	2.074	2.49	0.899	3.15	-0.238	-0.29
	(0.639)	0.000*	(0.833)	0.023*	(0.286)	0.006*	(0.833)	0.778
Esmithii	91.965	203.69	0.202	0.34	0.964	4.77	-0.202	-0.34
	(0.452)	0.000*	(0.589)	0.735	(0.202)	0.000*	(0.589)	0.735
Enitens	90.976	142.48	1.393	1.67	1.216	4.26	-0.432	-0.52
	(0.639)	0.000*	(0.833)	0.113	(0.286)	0.001*	(0.833)	0.611
E gc	91.375	143.11	1.663	2.00	0.843	2.95	-0.200	-0.24
	(0.639)	0.000*	(0.833)	0.062	(0.286)	0.009*	(0.833)	0.813
EguA	90.808	142.22	1.474	1.77	1.094	3.83	-0.416	-0.50
	(0.639)	0.000*	(0.833)	0.095	(0.286)	0.001*	(0.833)	0.624
EguW	91.890	143.91	0.923	1.11	1.031	3.61	-0.451	-0.54
	(0.639)	0.000*	(0.833)	0.283	(0.286)	0.002*	(0.833)	0.595

*significant parameters at the 5% significant level

The piecewise linear regression models which can be used to predict the α-cellulose levels of each genotype at each processing stage are derived from Table 9 and presented below:

E.dunnii: $\hat{Y}=90.846+0.361t_1+1.271t_2-0.508t_3$

E.grandis: $\hat{Y}=91.256+2.074t_1+0.899t_2-0.238t_3$

E.smithii: $\hat{Y}=91.965+0.202t_1+0.964t_2-0.202t_3$

E.nitens: $\hat{Y}=90.976+1.393t_1+1.216t_2-0.432t_3$

E.gc: $\hat{Y}=91.375+1.663t_1+0.843t_2-0.200t_3$

E.guA: $\hat{Y}=90.808+1.474t_1+1.094t_2-0.416t_3$

E.guW: $\hat{Y}=91.890+0.923t_1+1.031t_2-0.451t_3$

4.4.5. Copper Numbers data

The unstructured covariance structure had the best fit to the copper numbers data (Table 4.2: AIC=65.2). The delignification and bleaching rates of change in copper numbers were found to be significantly different among the seven species/genotypes (Table 4.3: F=28.04, df1=7, df2=17, p-value=0.000) and (Table 4.3: F=31.35, df1=7, df2=17, p-value=0.000) respectively. The finishing sub-process as with the other chemical properties did not produce significantly different rates of change in copper

numbers (Table 4.3: F=0.16, df1=7, df2=17, p-value=0.980). In addition the seven species/genotypes start off with significantly different copper numbers (Table 4.3: F=279.41, df1=7, df2=65, p-value=0.000). This means that copper numbers is an important chemical property that can be used in classifying the seven species/genotypes.

The copper numbers' piecewise linear regression model parameter estimates for the seven species/genotypes are presented in Table 4.8 below.

Table 4.8. Piecewise linear regression model parameter estimates and t-tests for Copper Number.

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	*t*-test (df=65) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value
Edunnii	3.107	16.83	-1.064	-4.42	-0.514	-6.22	0.104	0.43
	(0.185)	0.000*	(0.241)	0.000*	(0.083)	0.000*	(0.241)	0.669
Egrandis	2.886	15.63	-1.277	-5.30	-0.414	-5.01	0.083	0.34
	(0.185)	0.000*	(0.241)	0.000*	(0.083)	0.001*	(0.241)	0.736
Esmithii	2.974	22.78	-1.245	-7.32	-0.417	-7.14	0.063	0.37
	(0.131)	0.000*	(0.170)	0.000*	(0.058)	0.000*	(0.170)	0.714
Enitens	2.423	13.12	-0.657	-2.73	-0.452	-5.47	0.105	0.44
	(0.185)	0.000*	(0.241)	0.014*	(0.083)	0.000*	(0.241)	0.669
E gc	3.168	17.16	-1.534	-6.37	-0.418	-5.07	0.075	0.31
	(0.185)	0.000*	(0.241)	0.000*	(0.083)	0.001*	(0.241)	0.759
EguA	2.999	16.24	-1.397	-5.80	-0.410	-4.97	0.102	0.42
	(0.185)	0.000*	(0.241)	0.000*	(0.083)	0.000*	(0.241)	0.678
EguW	2.472	13.39	-0.881	-3.66	-0.408	-4.94	0.112	0.47
	(0.430)	0.000*	(0.241)	0.002*	(0.083)	0.000*	(0.241)	0.647

*significant parameters at the 5% significant level

All rates of change of copper numbers due to delignification and bleaching are significant for all species/genotypes (Table 4.8: all *p*-values for t-test<0.05). From Table 4.8 the piecewise linear regression models for copper numbers can be constructed as:

E.dunnii: $\hat{Y}=3.107-1.064t_1-0.514t_2+0.104t_3$

E.grandis: $\hat{Y}=2.886-1.277t_1-0.414t_2+0.083t_3$

E.smithii: $\hat{Y}=2.974-1.245t_1-0.417t_2+0.063t_3$

E.nitens: $\hat{Y}=2.423-0.657t_1-0.452t_2+0.105t_3$

E.gc: $\hat{Y}=3.168-1.534t_1-0.418t_2+0.075t_3$

E.guA: $\hat{Y}=2.999-1.397t_1-0.410t_2+0.102t_3$

E.guW: $\hat{Y}=2.472-0.881t_1-0.408t_2+0.112t_3$

The piecewise regression models can be used to estimate copper numbers for the seven species/genotypes at each stage by substituting the values of t_1, t_2 and t_3 that were described in Table 4.1.

The correlation between the percentage of γ-cellulose at the beginning (raw pulp stage) and at the end of processing was found to be r=0.766. This means that there is a strong relationship between the initial and final percentage levels of γ-cellulose.

4.4.6. Glucose Data

Having the lowest AIC value, the unstructured covariance structure was fitted to the glucose data (Table 4.2: AIC=276.5). The effects of delignification and bleaching were significantly different on the rates of change of glucose for the seven species/genotypes (Table 4.3: F=38.01, df1=7, df2=17, p-value=0.000) and (Table 4.3: F=41.01, df1=7, df2=17, p-value=0.000) respectively. In general glucose had significant rates of change during delignification and bleaching for all genotypes (Table 4.9: $\beta_{1'}$s>0 and $\beta_{2'}$s>0 with p-values for t-tests<0.05 for all genotypes). The changes in glucose due to the finishing stage were not significant for all species/genotypes. Table 4.9: Piecewise linear regression model parameter estimates and t-tests for Glucose (96α)

Table 4.9. Piecewise linear regression model parameter estimates and t-tests for Glucose

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	t-test (df=65) p-value	Parameter (Std Dev)	t-test (df=17) p-value	Parameter (Std Dev)	t-test (df=17) p-value	Parameter (Std Dev)	t-test (df=17) p-value
Edunnii	90.146 (0.352)	256.46 0.000*	2.010 (0.461)	4.36 0.000*	1.226 (0.157)	7.80 0.000*	-0.116 (0.458)	-0.25 0.803
Egrandis	92.009 (0.352)	261.76 0.000*	2.467 (0.461)	5.35 0.000*	0.792 (0.157)	5.04 0.001*	-0.474 (0.458)	-1.03 0.315
Esmithii	90.595 (0.272)	332.74 0.000*	1.884 (0.345)	5.47 0.000*	1.035 (0.111)	9.31 0.000*	-0.152 (0.324)	-0.47 0.645
Enitens	89.712 (0.352)	255.23 0.000*	3.493 (0.461)	7.57 0.014*	0.987 (0.157)	6.28 0.000*	-0.298 (0.458)	-0.65 0.524
E gc	89.619 (0.352)	254.96 0.000*	3.640 (0.461)	7.89 0.000*	0.701 (0.157)	4.46 0.001*	0.054 (0.458)	0.12 0.908
EguA	90.042 (0.352)	256.17 0.000*	2.908 (0.461)	6.30 0.000*	0.949 (0.157)	6.04 0.000*	0.124 (0.458)	0.27 0.791
EguW	92.454 (0.352)	263.03 0.000*	2.493 (0.461)	5.41 0.000*	0.675 (0.157)	4.29 0.001*	-0.672 (0.458)	-1.47 0.161

*significant parameters at the 5% significant level

The piecewise linear regression model parameter estimates derived from Table 11 are shown below and these models can be used to estimate glucose levels at each stage of chemical processing using the values of t_1, t_2 and t_3 defined in Table 2 above.

E.dunnii: $\hat{Y}=90.146+2.010t_1+1.226t_2-0.116t_3$

E.grandis: $\hat{Y}=92.009+2.467t_1+0.792t_2-0.474t_3$

E.smithii: $\hat{Y}=90.595+1.884t_1+1.035t_2-0.153t_3$

E.nitens: $\hat{Y}=89.712+3.493t_1+0.987t_2-0.298t_3$

E.gc: $\hat{Y}=89.619+3.640t_1+0.701t_2-0.054t_3$

E.guA: $\hat{Y}=90.042+2.908t_1+0.949t_2-0.124t_3$

E.guW: $\hat{Y}=92.454+2.493t_1+0.675t_2-0.672t_3$

4.4.7. Xylose data

With the lowest AIC, the unstructured covariance structure was of best fit to the xylose data (Table 4.2: AIC=173.5). The rates of change in xylose due to delignification and bleaching differed significantly across the seven species/genotype (Table 4.3: F=14.01, df1=7, df2=17, p-value=0.000) and (Table 4.3: F=26.57, df1=7, df2=17, p-value=0.000). This renders xylose an important classification variable for the seven species/genotypes. The finishing sub-process as with the other chemical properties did not have a significant effect on the final xylose readings.

There were significant rates of decrease in xylose during the delignification and bleaching processes for most species/genotypes (Table 4.10 β_1's<0, β_2's<0 with p-values<0.05 for t-tests) except for EguA which did not have a significant decrease in xylose during delignification (Table 4.10: β_1=-0.626, t=-1.95, df=17, p-value=0.068). The finishing stage did not have a significant effect on the xylose values just like with the other chemical properties.

Table 4.10. Piecewise linear regression model parameter estimates and t-tests for Xylose.

Genotype	β_{0i}		β_{1i}		β_{2i}		β_{3i}	
	Parameter (Std Dev)	*t*-test (df=65) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value	Parameter (Std Dev)	*t*-test (df=17) p-value
Edunnii	4.714	22.04	-0.857	-2.66	-0.565	-5.91	-0.037	-0.13
	(0.214)	0.000*	(0.322)	0.016*	(0.096)	0.000*	(0.279)	0.895
Egrandis	3.367	15.75	-0.545	-1.69	-0.402	-4.21	0.084	0.30
	(0.214)	0.000*	(0.322)	0.109*	(0.096)	0.001*	(0.279)	0.766
Esmithii	4.912	29.66	-1.032	-4.35	-0.528	-7.80	0.069	0.35
	(0.166)	0.000*	(0.237)	0.000*	(0.068)	0.000*	(0.197)	0.729
Enitens	5.939	27.77	-2.279	-7.09	-0.484	-5.06	0.017	0.06
	(0.214)	0.000*	(0.322)	0.000*	(0.096)	0.000*	(0.279)	0.952
E gc	3.927	18.37	-0.817	-2.54	-0.291	-3.04	0.218	0.78
	(0.214)	0.000*	(0.322)	0.021*	(0.096)	0.007*	(0.279)	0.445
EguA	4.340	20.29	-0.626	-1.95	-0.516	-5.39	-0.046	-0.16
	(0.214)	0.000*	(0.322)	0.068	(0.096)	0.000*	(0.279)	0.871
EguW	3.244	15.17	-0.951	-2.96	-0.280	-2.93	0.055	0.20
	(0.214)	0.000*	(0.322)	0.009*	(0.096)	0.009*	(0.279)	0.846

*significant parameters at the 5% significant level

The parameter estimates for the piecewise linear regression models for xylose derived from Table 4.10 are presented below:

E.dunnii: $\hat{Y}=4.714-0.857t_1-0.565t_2-0.037t_3$

E.grandis: $\hat{Y}=3.367-0.545t_1-0.402t_2+0.084t_3$

E.smithii: $\hat{Y}=4.912-1.032t_1-0.528t_2+0.069t_3$

E.nitens: $\hat{Y}=5.939-2.279t_1-0.484t_2+0.017t_3$

E.gc: $\hat{Y}=3.927-0.817t_1-0.291t_2+0.218t_3$

E.guA: $\hat{Y}=4.340-0.626t_1-0.516t_2-0.046t_3$

E.guW: $\hat{Y}=3.244-0.951t_1-0.280t_2+0.055t_3$

Although some parameter estimates for the finishing sub-process are negative most of them are generally positive. It was observed that for all chemical properties, the finishing stage has the general effect of reversing the trend in bleaching but such reversal is not significant. Glucose is also an important classifying variable for the seven species/genotypes.

4.5. Conclusion

The piecewise linear regression models had the capability of outlining the effect of each sub-process of chemical pulping on the seven reactivity variables studied. The ability of the model to state, by the model parameters, the effect of each sub-process on the chemical properties is a value addition to the study of chemical pulping processes. This can be extended to other types of pulp processing with known sub-processes i.e. kraft pulping, neutral sulphite pulping.

Based on the results from the piecewise linear regression models it was established that the six chemical properties lignin, γ-cellulose, α-cellulose, copper numbers, glucose and xylose were important classification variables for species/genotypes while viscosity, based on the results obtained, was not. This means that when one wants to compare or group wood species/genotypes using their chemical properties for the purpose of deciding which ones are mixable during processing, they do not need to consider viscosity.

Using the coding of the stages as shown in Table 4.1, the levels of the chemical properties studied can be estimated at each stage using the piecewise linear regression models developed in this study. This is could be useful to businesses involved in the manufacture of dissolving pulp as the model can be used as a predictive tool to assess species/genotype properties without having to carry out the actual bleaching especially if such models have already been developed for the concerned timber species/genotype. This will reduce the use of costly chemicals as well as limit the generation of harmful waste. Another advantage of the developed models is that the parameter estimates for the various species/genotypes can be grouped according to their sizes in order to classify the species/genotypes into groups of mixable species or genotypes during chemical processing. This reduces the trial and error involved in selecting specific clones and species for specific grades of dissolving pulp. The methodology can thus be used for other pulps earmarked for other products in the timber industry.

For further studies, it would of interest to develop a classifying method based on multivariate statistical techniques such as cluster analysis. Chapter 5 compares results

in Chapters 3 and 4 and comes up with comparative clustering results while Chapter 6 makes use of results in Chapter 4 to come up with an alternative grouping mechanist to identify genotypes that can be optimally mixed during processing.

The residual plots for the models fitted in sections 4.4.1 to 4.4.7 show that the residuals exhibit normality hence the assumption of normality on the data holds. The models can also be deemed adequate based on the residual plots (See Figures A1.8 to A1.14).

Chapter 5

Comparison of Random Coefficient and Piecewise Linear Regression Models using Best Linear Unbiased Predictors (BLUP)

5.1. Introduction

This brief chapter presents a side by side comparison of results in Chapters 3 and 4. Chapters 3 and 4 sought to describe the evolution of the chemical process by means of two types of longitudinal models, namely, random coefficient models and piecewise linear regression models. The random coefficient model was used to describe the overall evolution of the seven chemical properties without due regard to the sub-processes in the system while the piecewise linear regression model paid particular attention to well the known sub-processes in the system. Although the two models have different number of parameters, hence not directly comparable, it is still of interest to try to present them side by side. The model parameters were calculated as random effects using Best Linear Unbiased Predictors (BLUP). The theory around the estiamation (prediction) of random effects as BLUPS, which are basically conditional expectations, is described in Chapter 3 Section 3.2.4.

5.2. Comparisons of the Random Coefficient and Piecewise Linear Regression Models

The main focus of this section is to look at the results obtained for random coefficient models and present them side by side with the piecewise linear regression results. The random coefficient model has only one overall slope value while the piecewise linear regression model has three slope values, each corresponding to a particular sub-process in the system. In order to obtain an overall slope value for the piecewise linear regression model, which will then be comparable to the random coefficient model slope, it is necessary to obtain a weighted mean slope for the piecewise linear regression model. The overall slope value for the piecewise linear regression model is calculated as

$$Average\ Slope = \frac{1}{6}(2\beta_{1i} + 3\beta_{2i} + \beta_{3i}) \tag{5.1}$$

Since the process of delignification, with slope β_{1i}, spans two of the six stages, it has a weighting of 2/6. The bleaching process, with slope β_{2i} spans three stages hence it gets a weighting of 3/6 and the finishing stage (β_{3i}) has a weighting of 1/6. The averages slopes for the piecewise linear regression models are calculated and presented in the tables that will follow.

5.2.1. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for viscosity

AS far as viscosity is concerned, the results in Table 5.1 indicate that for viscosity the genotype with the lowest rate of change is Egrandis (RC slope=-1.995 and PLR slope=0.672). For all other genotypes the rankings are almost the same for both the RC and PLR slopes although they differ in magnitude. Only the last two genotypes, that is Edunnii and GUA, have their ranks swapped around but if we were to decide which genotypes are mixable based on their ranks, the two would still be deemed close together and mixable. The rankings according to the two models are:

RC: 1. Egrandis, 2. ENitens, 3. Esmithii, 4. GUW, 5. GCG, 6. Edunnii, 7. GUA

PLR: 1. Egrandis, 2. ENitens, 3. Esmithii, 4. GUW, 5. GCG, 6. GUA, 7. Edunnii

Althogh the two models have different values for slopes, they yield similar reults with regard to the determining of which genotypes are mixable during processing.

Table 5.1. Comparison of the Random coefficient and the Piecewise linear regression model for Viscosity.

Random Effects Predictions for **Viscosity** Models									
Genotype	Random Coefficients			Piecewise linear regression					
	β_{0i}	β_{1i}	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	63.529	-5.896	6	61.083	-10.681	-2.427	-6.647	-5.882	7
E.grandis	38.446	-1.995	1	30.183	4.501	0.019	-5.028	0.672	1
E.smithii	48.643	-3.569	3	45.883	2.473	-5.471	-0.100	-1.928	3
E.nitens	43.955	-2.122	2	40.882	3.062	-2.696	-4.440	-1.067	2
GCG	58.160	-5.677	5	56.713	-2.143	-7.016	-2.516	-4.642	5
GUA	70.895	-7.926	7	66.517	0.592	-9.878	-6.687	-5.856	6
GUW	58.160	-5.177	4	54.413	2.986	-7.718	-0.630	-2.969	4

5.2.2. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for viscosity

The RC model statistics for lignin, presented in Table 5.2 below, contain quadradic terms which make the RC slope values not easily comparable with the PLR average slopes. In the RC model all slope values are negative indicating that lignin decreases with each stage. The RC quadratic terms are positive which means that the negative slopes are progressively reduced with each stage. Fow example, the slope value for Edunnii is -1.857 and the quadratic term is 0.174. This means that, with every stage, the negative slope keeps becoming less steep at the rate of 0.174 times the square of the stage level or value. The genotype Esmithii starts off with the highest negative slope (β_{1i}=-2.572) but it also has the most rapid decline in the absolute value of the negative slope (β_{2i}=0.258). Under the RC model, the average slope between the beginning of the process and the last stage can be calculated as:

$$Average\ Slope = \frac{\hat{y}_{stage\ 6} - \hat{y}_0}{6}$$

where $\hat{y}_{\text{stage } 6}$ is the estimated lignin value at stage 6 and $\hat{y}_0$ is the intercept of the RC model.

Table 5.2. Comparison of the Random coefficient and the Piecewise linear regression model for Lignin.

Random Effects Predictions for **Lignin** Models											
Genotype	Random Coefficients					Piecewise linear regression					
	β_{0i}	β_{1i}	β_{2i}	Average slope	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	5.662	-1.857	0.174	-0.976	4	4.23	-2.073	-0.449	-0.141	-0.939	4
E.grandis	4.704	-1.869	0.193	-0.853	3	3.319	-2.157	-0.284	-0.062	-0.871	3
E.smithii	6.654	-2.572	0.258	-1.229	6	4.609	-2.673	-0.556	0.074	-1.157	7
E.nitens	3.499	-1.407	0.148	-0.623	1	2.414	-1.52	-0.227	0.036	-0.614	1
GCG	6.855	-2.515	0.240	-1.290	7	4.763	-2.453	-0.69	0.062	-1.152	6
GUA	5.771	-2.352	0.243	-1.073	5	3.917	-2.467	-0.428	0.066	-1.025	5
GUW	4.18	-1.568	0.154	-0.773	2	2.887	-1.538	-0.396	0.077	-0.698	2

RC: 1. ENitens, 2. GUW, 3. Egrandis, 4. Edunnii, 5. GUA, 6. Esmithii, 7. GCG

PLR: 1. ENitens, 2. GUW, 3. Egrandis, 4. Edunnii, 5. GUA, 6. GCG, 7. Esmithii.

Again, the two methods are producing consistent ranking results with the slight difference that the last two genotypes are swapped around but still having adjacent ranks.

5.2.3. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for γ-Cellulose

The γ-cellulose results in Table 5.3 indicate that the seven genotypes can be ranked as follows:

RC: 1. GUW, 2. Edunnii, 3. GCG, 4. Esmithii, 5. Egrandis, 6. ENitens, 7. GUA

PLR: 1. GUW, 2. Edunnii, 3. Esmithii, 4. GCG, 5. Egrandis, 6. ENitens, 7. GUA

Althogh the two models yield almost the same results with only Esmithii and GCG swapped around on ranks 3 and 4.

Table 5.3. Comparison of the Random coefficient and the Piecewise linear regression model for Lignin.

Random Effects Predictions for **γ-Cellulose** Models									
Genotype	Random Coefficients			Piecewise linear regression					
	β_{0i}	β_{1i}	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	8.131	-0.803	2	6.896	0.283	-1.24	0.296	-0.730	2
E.grandis	7.274	-0.845	5	7.244	-2.117	-0.795	0.179	-0.853	5
E.smithii	8.150	-0.832	4	7.635	-1.17	-0.97	0.195	-0.809	3
E.nitens	8.046	-0.940	6	7.553	-1.446	-1.103	0.382	-0.913	6
GCG	7.480	-0.817	3	7.256	-1.707	-0.816	0.072	-0.817	4
GUA	8.367	-0.943	7	7.826	-1.401	-1.086	0.235	-0.918	7
GUW	6.754	-0.720	1	6.198	-0.794	-0.894	0.222	-0.691	1

5.2.4. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for α-Cellulose

As far as *α*-cellulose is concerned, the two model produced exactly the same genotype rankings. Results in Table 5.4 indicate that the seven genotypes can be ranked as follows:

RC: 1. Esmithii, 2. GUW, 3. GCG, 4. Edunnii, 5. Egrandis, 6. GUA, 7. ENitens

PLR: 1. Esmithii, 2. GUW, 3. GCG, 4. Edunnii, 5. Egrandis, 6. GUA, 7. ENitens

Table 5.4. Comparison of the Random coefficient and the Piecewise linear regression model for α-Cellulose.

Random Effects Predictions for **α-Cellulose** Models									
Genotype	Random Coefficients			Piecewise linear regression					
	β_{0i}	β_{1i}	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	89.865	0.887	4	90.846	0.361	1.271	-0.508	0.823	4
E.grandis	91.128	0.904	5	91.256	2.074	0.899	-0.238	0.905	5
E.smithii	91.136	0.689	1	91.965	0.202	0.964	-0.202	0.643	1
E.nitens	90.368	1.006	7	90.976	1.393	1.216	-0.432	0.971	7
GCG	91.153	0.811	3	91.375	1.663	0.843	-0.2	0.806	3
GUA	90.344	0.933	6	90.808	1.474	1.094	-0.416	0.906	6
GUW	91.317	0.804	2	91.89	0.923	1.031	-0.451	0.766	2

5.2.5. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for Copper Number

The comparative results for copper number are presented in Table 5.5 below. The results show consistency in the slope rankings for the two models with very slight variations. In general, it can be said that the two models achieve similar genotype rankings in terms of rates of changes in copper number readings. Rates of changes in copper number readings can be ranked as follows:

RC: 1. GUW, 2. ENitens, 3. Egrandis, 4. Esmithii, 5. GUA, 6. Edunnii, 7. GCG.
PLR: 1. ENitens, 2. GUW, 3.5 Egrandis, 3.5. Esmithii, 5. GUA, 6. Edunnii, 7. GCG.

Apart from the swap of GUW and Enitens in ranks 1 and 2 the two models achieve the similar ranking bearing in mind that genotype with adjacent ranks have a high possibility of being mixable.

Table 5.5. Comparison of the Random coefficient and the Piecewise linear regression model for Copper Number.

Random Effects Predictions for **Copper Number** Models									
Genotype	Random Coefficients			Piecewise linear regression					
	β_{0i}	β_{1i}	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	3.231	-0.504	6	3.107	-1.064	-0.514	0.104	-0.503	6
E.grandis	2.847	-0.466	3	2.886	-1.277	-0.414	0.083	-0.475	3.5
E.smithii	2.954	-0.467	4	2.974	-1.245	-0.417	0.063	-0.475	3.5
E.nitens	2.621	-0.402	2	2.423	-0.657	-0.452	0.105	-0.393	1
GCG	3.050	-0.507	7	3.168	-1.534	-0.418	0.075	-0.522	7
GUA	2.910	-0.478	5	2.999	-1.397	-0.41	0.102	-0.489	5
GUW	2.549	-0.401	1	2.472	-0.881	-0.408	0.112	-0.400	2

5.2.6. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for Glucose

The comparative results for glucose are presented in Table 5.6 below. The results show the same consistency in the slope rankings for the two models with very slight variations. In general, it can be said that the two models achieve similar genotype rankings in terms of rates of changes in glucose readings. Rates of changes in glucose readings can be ranked as follows:

RC: 1. GUW, 2. Egrandis, 3. Esmithii, 4. GCG, 5. GUA, 6. Edunnii, 7. ENitens.

PLR: 1. GUW, 2. Egrandis, 3. Esmithii, 4. GCG, 5. Edunnii, 6. GUA, 7. ENitens.

The two models produced the similar ranking results with the slight variation in the swap of ranks 5 and 6 between genotypes GUA and Edunnii.

Table 5.6. Comparison of the Random coefficient and the Piecewise linear regression model for Glucose.

Random Effects Predictions for **Glucose** Models									
Genotype	Random Coefficients			Piecewise linear regression					
	β_{0i}	β_{1i}	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	89.629	1.146	6	90.146	2.01	1.226	-0.116	1.133	5
E.grandis	92.197	0.851	2	92.009	2.467	0.792	-0.474	0.860	2
E.smithii	90.317	0.970	3	90.595	1.884	1.035	-0.152	0.979	3
E.nitens	89.989	1.161	7	89.712	3.493	0.987	-0.298	1.191	7
GCG	90.113	1.028	4	89.619	3.64	0.701	0.054	1.083	4
GUA	90.020	1.111	5	90.042	2.908	0.949	0.124	1.138	6
GUW	92.834	0.742	1	92.454	2.493	0.675	-0.672	0.754	1

5.2.7. Comparison of the Random Coefficient (RC) and the Piecewise Linear Regression (PLR) models for Xylose

The comparative results for xylose are presented in Table 5.7 below. The results show that the two methods yielded the same ranking order on the seven genotypes with rates of changes ranked as follows:

RC: 1. GCG, 2. GUW, 3. Egrandis, 4. GUA, 5. Esmithii, 6. Edunnii, 7. ENitens.

PLR: 1. GCG, 2. GUW, 3. Egrandis, 4. GUA, 5. Esmithii, 6. Edunnii, 7. ENitens.

Table 5.7. Comparison of the Random coefficient and the Piecewise linear regression model for Xylose.

Random Effects Predictions for **Xylose** Models									
Genotype	Random Coefficients			Piecewise linear regression					
	β_{0i}	β_{1i}	|Slope| Rank	β_{0i}	β_{1i}	β_{2i}	β_{3i}	Average Slope	|Slope| Rank
E.dunnii	5.005	-0.531	6	4.714	-0.857	-0.565	-0.037	-0.526	6
E.grandis	3.560	-0.353	3	3.367	-0.545	-0.402	0.084	-0.345	3
E.smithii	5.085	-0.517	5	4.912	-1.032	-0.528	0.069	-0.513	5
E.nitens	5.657	-0.669	7	5.939	-2.279	-0.484	0.017	-0.700	7
GCG	3.873	-0.294	1	3.927	-0.817	-0.291	0.218	-0.294	1
GUA	4.662	-0.464	4	4.34	-0.626	-0.516	-0.046	-0.456	4
GUW	3.189	-0.328	2	3.244	-0.951	-0.28	0.055	-0.336	2

5.3. Genotype comparisons and clustering based on average slopes

Comparisons of the seven genotypes, based on their average slope predictions, provide us with a measurement that can be used to formally decide on how mixable any pair of genotypes are. Such comparisons can be done through non-parametric tests such as the Friedman's test, followed by appropriate non-parametric post-hoc tests like Nemenyi's post-hoc tests (Pohlert, 2016). The genotypes can also be clustered using any standard clustering prodecdure such as the nearest neighbour hierarchical clustering method (Johnson and Wichern, 1998). Friedman's tests and nearest neighbour hierarchical clustering based on the slope estimates from the two models, that is the RC and the PLR, are presented in this section.

According to Garcia et al (2010), the most well-known nonparametric procedure for testing for the differences between more than two related samples is the Friedman test. In this case each sample (subject) has seven variables (chemical properties) measured at each of the six stages. Therefore, the seven chemical property readings are related by source, that is, the sampling unit which is regarded as a blocking variable in this study.

The Friedman's test is used to test for the randomised block design model given by:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij} \qquad (5.2)$$

with β_j being the j^{th} block (chemical property j) and α_i being the i^{th} treatment or genotype and e_{ij} is the error component of the model. The null hypothesis is that all genotypes have similar average slope distributions within all chemical properties against the alternative that some genotypes do not have identical slope distributions. The Friedman's test statistic, which has the χ^2 distribution with k-1 degrees of freedom, where k is the number of genotypes, is calculated as:

$$\hat{\chi}^2 = \left[\frac{12}{nk(k+1)}\sum_{i=1}^{k} R_i^2\right] - 3n(k+1) \qquad (5.3)$$

where R_i is the rank sum of the i^{th} genotype across all blocks (chemical properties), n is the number of slope estimates in each block, that is 7 in this case as seven genotypes are consided in each block and k is the number of genotypes compared. In the event that the Friedman's test turns out significant, the significance of the difference between any two genotypes, i and j, is determined using the test statistic:

$$z = \left(\bar{R}_i - \bar{R}_j\right) \Big/ \sqrt{\frac{k(k+1)}{6n}} \qquad (5.4)$$

where $\bar{R}_i$ and $\bar{R}_j$ are mean ranks of the two genotypes concerned and $z \sim N(0,1)$.

5.3.1. Genotype comparisons and clusterig based on average RC slopes

The summary of all the random coefficient model based mean slope ranks is presented in Table 5.8. Comarisons of the slopes based on their within block (chemical property) ranks were carried out using Friedman's test and Nemenyi's post-hoc tests.

The Friedman's test results presented at the bottom of Table 5.8 show that there are significant differences among the slope rank sums of the seven genotypes (*Fr*=13.531, *df* = 6, *p*-value=0.035). Nemenyi's post-hoc tests were used to determine which genotypes have significantly different mean ranks using the test statistic described in equation (5.4).

Table 5.8. Summary of random coefficient slope ranks

RC Slope Ranks									
Genotype	**Viscosity**	**Lignin**	**γ-cellulose**	**α-cellulose**	**Copper number**	**Glucose**	**Xylose**	**Rank sum**	**Mean Rank**
E.dunnii	6	4	2	4	6	6	6	34	4.857
E.grandis	1	3	5	5	3	2	3	22	3.143
E.smithii	3	6	4	1	4	3	5	26	3.714
E.nitens	2	1	6	7	2	7	7	32	4.571
GCG	5	7	3	3	7	4	1	30	4.286
GUA	7	5	7	6	5	5	4	39	5.571
GUW	4	2	1	2	1	1	2	13	1.857
Friedman's Test Statistic (Fr) = 13.531, degrees of freedom = 6, p-value = 0.035									

Table 5.9. Post-hoc tests based on the Friedman's test for the RC slopes.

Genotype	**Genotype**	E.dunnii	E.grandis	E.smithii	E.nitens	GCG	GUA	GUW
	Means Ranks	4.857	3.143	3.714	4.571	4.286	5.571	1.857
E.grandis	$\bar{R}_i - \bar{R}_j$	1.714						
	z	1.485						
	p-value	0.138						
E.smithii	$\bar{R}_i - \bar{R}_j$	1.143	-0.571					
	z	0.990	-0.495					
	p-value	0.322	0.621					
E.nitens	$\bar{R}_i - \bar{R}_j$	0.286	-1.429	-0.857				
	z	0.247	-1.237	-0.742				
	p-value	0.805	0.216	0.458				
GCG	$\bar{R}_i - \bar{R}_j$	0.571	-1.143	-0.571	0.286			
	z	0.495	-0.990	-0.495	0.247			
	p-value	0.621	0.322	0.621	0.805			
GUA	$\bar{R}_i - \bar{R}_j$	-0.714	-2.429	-1.857	-1.000	-1.286		
	z	-0.619	-2.103	-1.608	-0.866	-1.113		
	p-value	0.536	**0.035**	0.108	0.386	0.266		
GUW	$\bar{R}_i - \bar{R}_j$	3.000	1.286	1.857	2.714	2.429	3.714	
	z	2.598	1.113	1.608	2.351	2.103	3.217	
	p-value	**0.009**	0.266	0.108	**0.019**	**0.035**	**0.001**	

The results in Table 5.9 show that the genotype GUW has significantly different mean ranks to E.dunii ($[\bar{R}_i - \bar{R}_j]$=3.000, $z = 2.598$, p-value=0.009), E.nitens ($[\bar{R}_i - \bar{R}_j]$=2.714, $z = 2.351$, p-value=0.019), GCG ($[\bar{R}_i - \bar{R}_j]$=2.429, $z = 2.103$, p-

value=0.035) and GUA ($[\bar{R}_i - \bar{R}_j]$=3.714, $z = 3.217$, p-value=0.001). This means that it might be wise not to mix GUW with these other genotypes during processing. The other significant difference in mean rank is between GUA and E.grandis ($[\bar{R}_i - \bar{R}_j]$=-2.429, $z = -2.103$, p-value=0.035), which means that it is best not to mix these two genotypes during processing.

The Eucledean distances between the seven genotypes that is constructed from the ranks in Table 5.8, is presented in Table 5.10 with a hierarchical nearest neighbour dendogram based on the Eucledean distance matrix presented in Figure 5.1. It is clear from Figure 5.1 that the most mixable genotypes are Esmithii, GCG and Egrandis as these have the smallest distances between them. The next most mixable genotypes are Edunnii and GUA. The nearest neigbour clustering method yields almost similar results to the Friedman's test post-hoc analysis as far as the isolation of the genotype GUW from the other genotypes is concerned except that Eucledian distances place E.nitens furthest from the other genotypes (see Figure 5.1). How large the Eucledian distances should be to ascertain non-mixability might need further investigated.

Table 5.10. Eucledean distances based on genotype RC slope ranks

Genotype	**RC Euclidean Distances based on ranks**						
	E.dunnii	E.grandi	E.smithi	E.nitens	GCG	GUA	GUW
E.dunnii	-						
E.grandi	8.367	-					
E.smithi	6.325	6.000	-				
E.nitens	8.246	7.211	9.487	-			
GCG	6.481	8.000	6.000	11.832	-		
GUA	6.083	7.681	7.550	8.062	6.856	-	
GUW	8.888	6.403	7.000	10.817	8.775	10.296	-

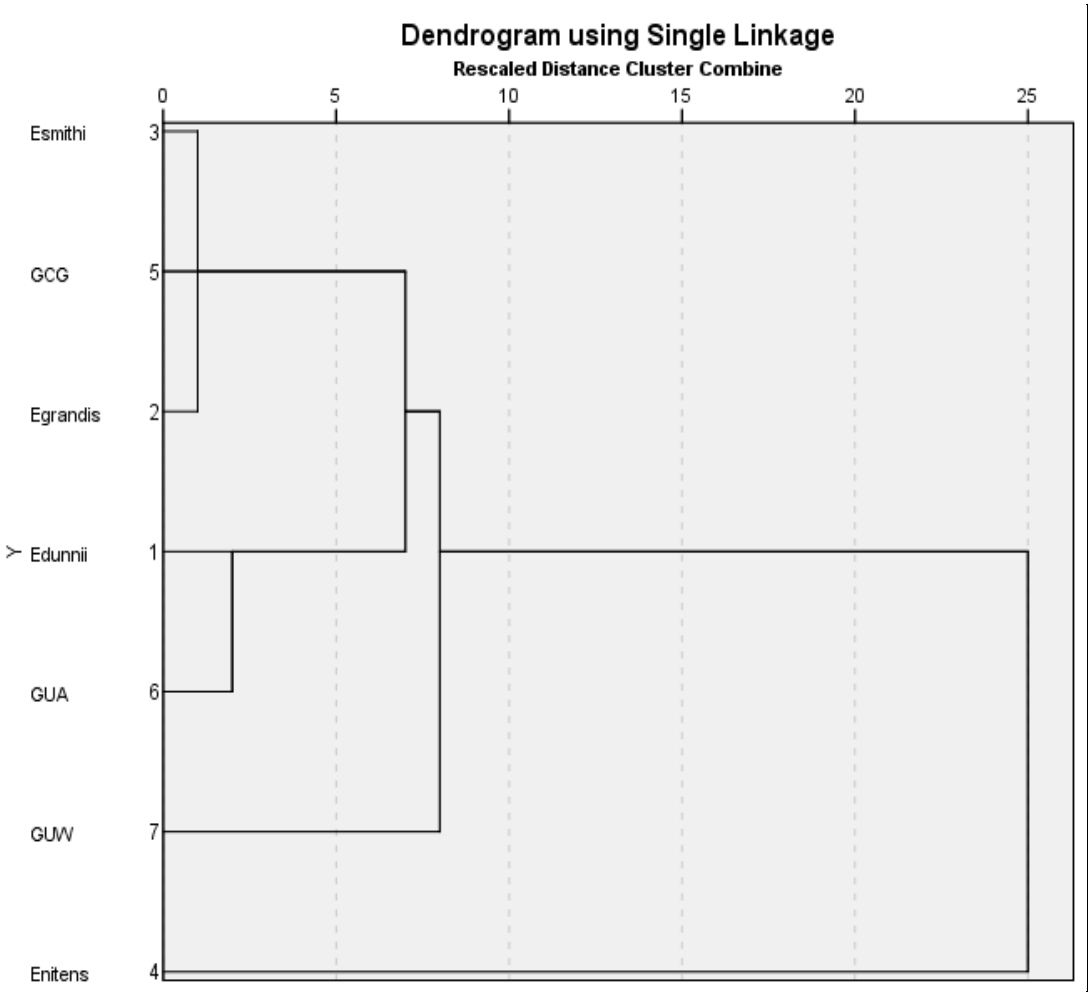


Figure 5.1. Nearest neigbhour dendogram based on the RC slope ranks.

5.3.2. Genotype comparisons and clusterig based on PLR slopes

The summary of all the piecewise linear regression slope ranks is presented in Table 5.11. The Friedman's test results presented at the bottom of Table 5.11 show that there are no significant differences in the slope rank sums of the seven genotypes at the 5% level of significance (Fr=12.321, df = 6, p-value=0.055). This result is very marginal as the p-value is very close to 0.05. At the 10% level of significance we may conclude that some of the genotypes have significantly different slope rank sums. Considering the marginal nature of the results, post-hoc tests were carried out to assess if any pair of genotyes have significantly different slope ranks with the results as presented in Table 5.12.

Table 5.11. Summary of piecewise linear regression slope ranks

PLR Slope Ranks									
Genotype	**Viscosity**	**Lignin**	**γ-cellulose**	**α-cellulose**	**Copper number**	**Glucose**	**Xylose**	Rank Sum	Mean Rank
E.dunnii	7	4	2	4	6	5	6	34	1156
E.grandis	1	3	5	5	3.5	2	3	22.5	506.25
E.smithii	3	7	3	1	3.5	3	5	25.5	650.25
E.nitens	2	1	6	7	1	7	7	31	961
GCG	5	6	4	3	7	4	1	30	900
GUA	6	5	7	6	5	6	4	39	1521
GUW	4	2	1	2	2	1	2	14	196
Friedman's Test Statistic (Fr) = 12.321, degrees of freedom = 6, p-value = 0.055									

The results in Table 5.12 show similar outcomes as with Table 5.9 in that the genotype GUW has significantly different mean rank to E.dunii ($\left[\overline{\mathrm{R}}_{\mathrm{i}} - \overline{\mathrm{R}}_{\mathrm{j}}\right]$=3.333, $z = 2.887$, p-value=0.004), E.nitens ($\left[\bar{R}_i - \bar{R}_j\right]$=2.833, $z = 2.454$, p-value=0.014), GCG ($\left[\bar{R}_i - \bar{R}_j\right] = 2.667$, $z = 2.309$, p-value=0.000) and GUA ($\left[\bar{R}_i - \bar{R}_j\right]$=4.167, $z = 3.608$, p-value=0.000). This leaves GUW as different from most of the other genotypes. The genotypes GUA and E.grandis are also significantly different ($\left[\overline{\mathrm{R}}_{\mathrm{i}} - \overline{\mathrm{R}}_{\mathrm{j}}\right]$=-2.429, $\mathrm{z} = -2.103$, p-value=0.035).

Table 5.12. Post-hoc tests based on the Friedman's test for the PLR slopes.

Genotype	**Genotype**	E.dunnii	E.grandis	E.smithii	E.nitens	GCG	GUA	GUW
	Means Ranks	5.667	3.750	4.250	5.167	5.000	6.500	2.333
E.grandis	$\bar{R}_i - \bar{R}_j$	1.917						
	z	1.660						
	p-value	0.097						
E.smithii	$\bar{R}_i - \bar{R}_j$	1.417	-0.500					
	z	1.227	-0.433					
	p-value	0.220	0.665					
E.nitens	$\bar{R}_i - \bar{R}_j$	0.500	-1.417	-0.917				
	z	0.433	-1.227	-0.794				
	p-value	0.665	0.220	0.427				
GCG	$\bar{R}_i - \bar{R}_j$	0.667	-1.250	-0.750	0.167			
	z	0.577	-1.083	-0.650	0.144			
	p-value	0.564	0.279	0.516	0.885			
GUA	$\bar{R}_i - \bar{R}_j$	-0.833	-2.750	-2.250	-1.333	-1.500		
	z	-0.722	-2.382	-1.949	-1.155	-1.299		
	p-value	0.470	**0.017**	0.051	0.248	0.194		
GUW	$\bar{R}_i - \bar{R}_j$	3.333	1.417	1.917	2.833	2.667	4.167	
	z	2.887	1.227	1.660	2.454	2.309	3.608	
	p-value	**0.004**	0.220	0.097	**0.014**	**0.021**	**0.000**	

The Eucledean distance matrix associated with the ranks in Table 5.11 is presented in Table 5.13 below and a corresponding nearest neighbour hierarchical dendogram is presented in Figure 5.2 which shows that the most mixable genotypes are GCG, GUA and Edunnii with Esmithii closely linked with these three. Egrandis and GUW are the next closely related genotypes with Enitens being isolated from the other genotypes. This means that it might not be a good idea to mix GUW or E.grandis with any of GCG, GUA, E.dunnii or E.smithii. These results are very similar to those of the Friedman's test with the exceptation that the clustering method placed E.smithii as the most different genotype in terms of overall distance from the other genotypes.

Table 5.13. Eucledean distances based on genotype PLR slope ranks

Genotype	PLR Euclidean Distances						
	Edunnii	Egrandi	E.smithi	E.nitens	GCG	GUA	GUW
Edunnii							
Egrandis	8.441						
Esmithi	6.801	6.708					
Enitens	9.434	7.566	10.404				
GCG	6.325	7.089	6.265	11.619			
GUA	6.083	7.297	8.139	7.746	6.083		
GUW	8.124	6.265	6.801	10.817	7.874	10.149	

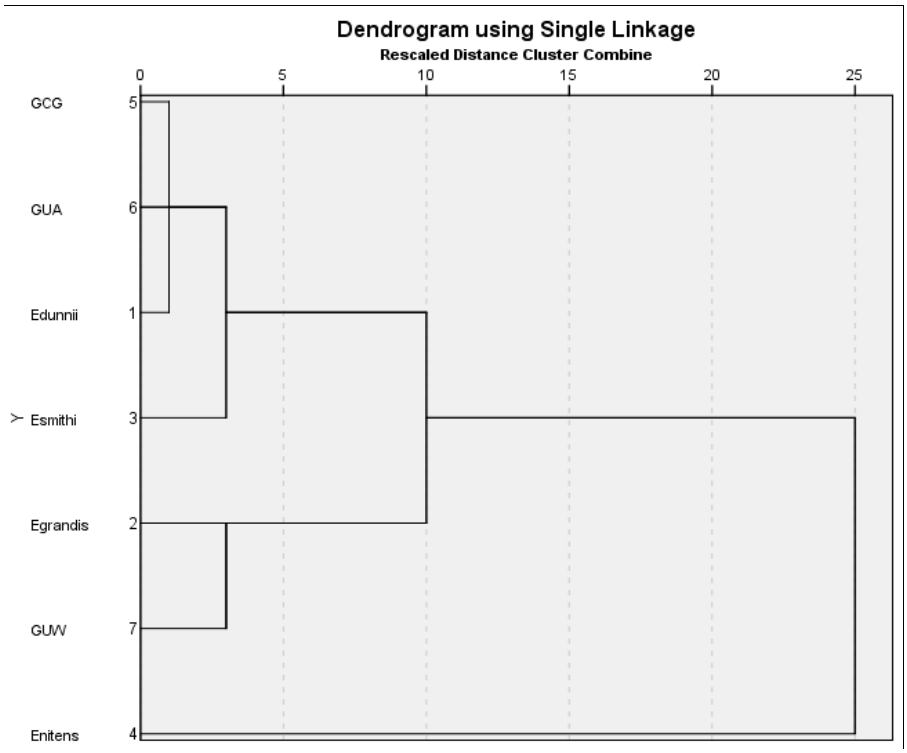


Figure 5.2. Nearest neigbhour dendogram based on the PLR slope ranks.

5.4. Conlusion

The results in this chapter have indicated that E.nitens and GUW are generally far removed from the other genotypes. When considering which genotypes to mix during processing Enitens and GUW should be processed on their own as much as possible as they have shown to be different from the other genotypes in both the random coefficient and the piecewise linear regression models. Slightly different clusters are obtained for the two models and the results for the piecewise linear regression model should take precedence as they take into consideration the three sub-processes of chemical pulping.

It was also noted that the two models generally identify genetypes with similar evolutions over the processing stages when the overall rates of changes are considered. It must be pointed out that the purpose of the study is not to identify which genotype is superior, but rather to identify genotypes with similar chemical evolutions over the processing stages.

Chapter 6

Classification of Timber Genotypes Using Piecewise Linear Regression and Kernel Density Based Clustering

6.1. Introduction

The classification or sorting of raw materials that feed into a manufacturing process is an important exercise if the process is to be optimised. Berget and Næs (2002a) looked at methods of optimally classifying or sorting raw materials that feed into a manufacturing process with particular focus on the quality of the end-product. If a manufacturing process depends on several sources and varieties of raw materials, it is important to sort such raw materials into homogenous groups in order to improve the stability and quality of the end–product. Fuzzy clustering is one of the methods suggested by Berget and Næs (2002b) who also mentioned other methods that minimise the distance between predicted and targeted end-product in a manufacturing system. These methods aim at minimising variability in the final product hence improving and ascertaining its quality. This study looks at a situation where various timber genotypes are used as raw materials in chemical pulping with a view of finding an optimal way to group such genotypes. Chapter 5 has shown how to come up with clusters, based on overall rates of change in chemical properties, using the random coefficient and the piecewise linear regression models. This chapter suggests an alternative method that makes full use of the rates of change in chemical properties at the sub-process level.

This study suggests a statistical method that can be used to classify different wood genotypes into classes of genotypes that exhibit similar response behaviour to chemical processing. Chemically bleached wood pulp (dissolving pulp) has a cellulose content of more than 90% and the changes in its chemical properties, over the processing stages, depend on the genotype of the tree being pulped. Raw pulp, which comes after acid bi-sulphite pulping, goes through a number of bleaching processing stages, each with a specific role, to produce dissolving pulp. These processing stages have different effects on the pulp depending on the type of wood genotype that is being

processed. The bleaching processing stages can be considered as time points for repeated measurements of the following chemical properties viz., viscosity, lignin, γ-cellulose, α-cellulose, copper number, glucose and xylose. Piecewise regression models were used to compare the changes of the chemical properties of seven pulping tree genotypes throughout the bleaching stages (see Chapter 4). In order to cut costs on the chemicals used for processing, it is important to identify species/genotypes that have similar chemical properties under the chemical pulping process in order to mix them together for optimised processing in case one genotype does not have enough volumes for processing. The piecewise regression model that was described in Chapter 4 was used with kernel density estimation to develop a “mixing matrix” for the seven genotypes. The method could be adopted for any situation were an industrial process depends on several types of raw materials. Using the methods developed in this study, it can be determined which genotypes or types of raw materials are optimally mixable for processing.

The classification of raw materials that feed into a manufacturing process is an important exercise if the process is to be optimised. Large variations in wood physical and anatomical characteristics among different Eucalyptus genotypes are well documented (Zbonak, Bush and Grzeskowiak, 2007). These variations justify the need to group genotypes considering that they behave differently during chemical processing. Non-statistical methods of materials classification have been discussed in literature, such as the work of Gu and Liu (2012), who proposed the use of coded illumination to directly measure discriminative features of raw materials for material classification. Their method can be used for a variety of materials that include iron, plastic and wood. From a statistical point of view, classification of such raw materials can be based on clustering methods. Lodi et al (2006) presented a novel algorithm for clustering streams of multidimensional points based on kernel density estimates and their work inspired this study. This study uses kernel density estimation as a tool for clustering.

In this study, a statistical method of classifying wood genotypes is proposed. The method is based on three statistical procedures, namely, piecewise regression, statistical simulation and kernel density estimation. If any two genotypes are found to be very similar in their behaviour then they will belong to the same class of genotypes

that can be mixed during processing. A method for developing a similarity or "mixability" matrix is also proposed.

6.2. Kernel Density Estimation and Clustering

The histogram and scatter plots have been traditionally used to get some insight into underlying distributions of observed data (Everitt et al, 2011). These tools give rough descriptions of the underlying distributions of observed data without assuming some specific parametric forms like the normal distribution. Parametric density estimators assume that the underlying density function, $f(\boldsymbol{x}, \boldsymbol{\theta})$, has a set of parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ where $\boldsymbol{\Theta} \subset \mathbb{R}^d$ and d is finite, that is, it is assumed that the parameter space is finite. An example of such fixed form parametric densities is the normal density which has only two parameters, the mean (μ) and standard deviation (σ). The kernel density estimator is a nonparametric estimator which does not assume any parametric form hence it has no limitation on the number of parameters the target density function can have. Kernel density estimators can be univariate or multivariate with the multivariate case being a density function of a random vector. The density function is estimated in a way that is similar in principle to the construction of a histogram. Density estimators of this nature were first introduced by Rosenblatt (1956) followed by Parzen (1962). Kernel density estimators have since been discussed extensively in the literature and include books by Silverman (1986), Simonoff (1996) and many others. Nonparametric methods of density estimation have the advantage of not depending on the sometimes incorrectly specified parametric models whose bias cannot be removed, even by the use of large sample sizes. This class of density estimation techniques can easily deal with data that are multimodal and hence difficult to fit any classical parametric form density function (Wang et al., 2004).

Kernel density smoothing techniques have also been used in classification problems in the social sciences (Shu et al., 2003). Earlier, Cheng established the connection between mean shift clustering and kernel density estimation (Cheng, 1995). This places kernel density estimation as a viable tool for classifying timber genotypes according to their behaviour under processing. Graphical presentations of kernel densities generated from chemical pulping data can highlight clusters in the data hence making it possible to identify genotypes that behave similarly during chemical processing.

Clusters in density based clustering have been defined as regions in the data space, where objects are concentrated and such regions are separated from each other by regions of low density (Mushdholifah et al, 2013). Density based clustering methods have advantages over hierarchical and partition clustering methods, especially for data with several clusters of different densities as they do not just get the clusters but also the concentrations or density levels of each cluster (Tran, Wehrens and Buydens, 2006). Kernel density estimation results can be analysed visually to allow for relationships or clusters to be determined and optimal graphical representations must be achieved through appropriate bandwidth selection and adjustment (Schwarz, 2005).

This study aims to exploit the clustering capabilities of kernel density estimators to optimally group different timber genotypes into clusters of genotypes with similar behaviours under chemical pulp processing. A novel, simple genotype “mixability” matrix was developed and can be used to decide if it is optimal to mix any two timber genotypes for chemical processing.

6.2.1. The kernel density estimator

Suppose a random sample of size N is taken on a *p*-dimensional random variable, $\mathbf{X}^{\mathrm{T}} = [X_1, X_2, \dots, X_p]$, for which the multivariate distribution is not known beforehand. The distribution of the random variable **X** will be estimated empirically using the observed vectors $\mathbf{x_i}^{\mathrm{T}} = [x_{i1}, x_{i2}, \dots, x_{ip}]$ for $i = 1, 2, \dots, N$. The probability density of the random variable **X** can be estimated by

$$\hat{f}_{\boldsymbol{H}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{h_1 h_2 \dots h_p} K\left(\frac{x_1 - x_{i1}}{h_1}, \frac{x_2 - x_{i2}}{h_2}, \dots, \frac{x_p - x_{ip}}{h_p}\right) \qquad (6.1)$$

where *K* is a multivariate kernel operating on *p* arguments and h_i, for i=1,2,…,*p,* is the optimum bandwidth corresponding to variable X_i. The bandwidths in equation (6.1) are multiplicative in this case but they can take any other form. For multiplicative kernels, equation (6.1) can also be written as

$$\hat{f}_{\boldsymbol{H}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \left\{ \prod_{j=1}^{p} h_j^{-1} K\left(\frac{x_j - x_{ij}}{h_j}\right) \right\} \qquad (6.2)$$

Let $\mathbf{X}^{\mathrm{T}} = [X_{1,}X_2, \ldots, X_p]$ be the multivariate variable for which a joint probability density is to be estimated and $\mathbf{H} = \mathrm{diag}(h_1, h_2, \ldots, h_p)$ be the diagonal matrix of bandwidths, then equation (6.2) can be written in a more compact form as

$$\hat{f}_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\det(\mathbf{H})} K[\mathbf{H}^{-1}(\boldsymbol{x} - \mathbf{X}_i)] \tag{6.3}$$

As a rule of thumb the bandwidth matrix **H** is generally made proportional to $\boldsymbol{\Sigma^{-1/2}}$ where **Σ** is the covariance matrix of the observed data (Härdle et al, 2004). If we let $K_H(\bullet) = \frac{1}{\det(\mathbf{H})} K[\mathbf{H^{-1}}(\bullet)]$ then

$$\hat{f}_{\boldsymbol{H}}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} K_H(\boldsymbol{x} - \boldsymbol{x}_i) \tag{6.4}$$

The underlying assumption of kernel estimation is that

$$\int K(\boldsymbol{u})\, d\boldsymbol{u} = 1 \tag{5.5}$$

this makes $\hat{f}_{\boldsymbol{H}}(\boldsymbol{x})$ a density function since

$$\int \hat{f}_{\boldsymbol{H}}(\boldsymbol{x}) = \int \frac{1}{N}\sum_{i=1}^{N} K_{\boldsymbol{H}}(\boldsymbol{x} - \boldsymbol{x}_i) \; = \frac{1}{N}\sum_{i=1}^{N} \int K(\boldsymbol{u}) d\boldsymbol{u} = \frac{1}{N}\sum_{i=1}^{N} 1 = 1$$

According to Härdle et al $\hat{f}_{\boldsymbol{H}}(\boldsymbol{x})$ is also a consistent estimator of $f(\boldsymbol{x})$ that is

$$\hat{f}_{\boldsymbol{H}}(\boldsymbol{x}) \xrightarrow{\;p\;} f(\boldsymbol{x}) \text{ or } \lim_{n\to\infty} \hat{f}_{\boldsymbol{H}}(\boldsymbol{x}) = f(\boldsymbol{x})$$

6.2.2. Kernel Functions

The kernel function determines how sample observations ($\boldsymbol{x_i}$) in the vicinity of a point $\boldsymbol{x}$ are going to contribute to the frequency or probability of that point. Some kernels assign equal weights to all values in the vicinity of $\boldsymbol{x}$ while others give higher weights to those sample observation that are closer to $\boldsymbol{x}$ than those that are further away. The uniform kernel gives a weight of 1/2 to every observed value that is in the vicinity of the point $\boldsymbol{x}$ while the other kernels give less weight to observed values further away from $\boldsymbol{x}$. The choice of a particular bandwidth determines the boundary of the vicinity of the point $\boldsymbol{x}$.

The multivariate kernel $K(\boldsymbol{u})$, where $\boldsymbol{u} = \mathbf{H}^{-1}(\mathbf{x} - \mathbf{x_i})$, can have any of the many well used functional forms. Some of the commonly used univariate kernels are listed in Table 6.1 below (Härdle et al). The simplest way to use such univariate kernels for multivariate cases is to use a multiplicative kernel function which is a product of the same kernel function operating on the *p*-random variables with different bandwidths. A particular multivariate kernel, the multivariate Epanechnikov kernel is such that

$$K(\boldsymbol{u}) \propto (1 - \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u})\mathrm{I}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} \leq 1). \tag{6.6}$$

In general multivariate kernels can be obtained from univariate kernels by taking

$$K(\boldsymbol{u}) \propto K(\|\boldsymbol{u}\|)$$

Where $\|\boldsymbol{u}\| = \sqrt{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}}$ is the Eucledian norm of the vector $\boldsymbol{u}$. In Table 6.1 the indicator function I(•) operating on *u*, where $u = \frac{(x - x_i)}{h}$, is such that

$$\mathrm{I}(u) = \begin{cases} 1 & \text{if } |u| < 1 \ \ (\text{or} - h_i < x - x_i < h) \\ 0 & \text{otherwise} \end{cases}. \tag{6.7}$$

The sum of the $\mathrm{I}(u)$'s will give the frequency of those observed values of X that are in the vicinity of *x* and this will give an estimate of $f(\boldsymbol{x})$. For the multivariate case values of u_i for $i = 1, 2, \ldots, p$, are put together to form the vector $\boldsymbol{u}$ in equation 6.6 above.

Table 6.1. Some common kernel functions

Kernel	K(u)
Uniform	$\frac{1}{2}\mathrm{I}(\lvert u\rvert \leq 1)$
Triangle	$(1 - \lvert u\rvert)\mathrm{I}(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2)\mathrm{I}(\lvert u\rvert \leq 1)$
Quartic (biweight)	$\frac{15}{16}(1 - u^2)^2\mathrm{I}(\lvert u\rvert \leq 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3\mathrm{I}(\lvert u\rvert \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$
Cosine	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)^2 \mathrm{I}(\lvert u\rvert \leq 1)$

6.2.3. Multivariate bandwidth selection

Bandwidth selection has a pivotal role in determining the accuracy of the kernel density function $\hat{f}_{\boldsymbol{H}}(\mathbf{x})$ as a predictor of the density function $f(\mathbf{x})$ for the random vector $\boldsymbol{X}$. As

indicated by Simonoff (1996), bandwidth selection has a more important role in determining the performance of a kernel density estimator than the choice of a kernel function. It also controls the smoothness of the resulting density estimate (Chiu, 1991). The bandwidth matrix ***H*** can be determined by a general rule of thumb which depends on the kernel function chosen, or cross-validation methods (Marron, 1987). When the bandwidth matrix is restricted to a class of positive definite diagonal matrices, the corresponding kernel function is known as a product kernel. Bandwidth matrices can also be determined from observed data using Markov Chain Monte Carlo (MCMC) algorithms (Zhang et al, 2004).

In this study product kernels based on diagonal bandwidth matrices will be used for their relative simplicity as compared to the MCMC derived ones. If the kernel function is based on a multivariate Gaussian distribution and a diagonal **H** matrix is assumed then the optimal bandwidths, that is, the elements of the diagonal bandwidth matrix, can be estimated by

$$h_i = \sigma_i \left[\frac{4}{(p+2)n}\right]^{1/(p+4)} \tag{6.8}$$

for $i = 1,2,\dots,p$, where σ_i is the standard deviation of the i^{th} variable, p is the dimension of the multivariate random vector. According to Zhang (2004), this method of bandwidth selection can be used if more complex methods are to be avoided even though the data observed might not be Gaussian. In the SAS statistical software's KDE procedure, different bandwidths can be tried and the one which the best smoothing effect is chosen (SAS/STAT, 2008).

6.3. Kernel density estimation as a clustering tool

Erdoğmus, Carreira-Perpñán and Özertem (2006) outlined the usefulness of kernel density estimation as a clustering tool. Their work was built on the cut clustering algorithm of Blatt et al (1997). They start by the classical kernel density estimation formula and show how it links to the cut clustering algorithm.

Suppose we have a set of observations $S = (\mathbf{x_1}, \dots, \mathbf{x_N})$ where $\mathbf{x_i} \in \mathbb{R}^{\boldsymbol{p}}$**.** The kernel density estimate for $f(\mathbf{x})$ is given in equation (6.4). Suppose that the set of observation

can be divided into clusters (s_i's) such that $S = (s_1, \ldots, s_q)$. The kernel density estimate for cluster s_j can be written as

$$\hat{f}_j(\mathbf{x}) = \frac{1}{N_j} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)\delta_{ij} \tag{6.9}$$

where N_j is the number of observations in cluster s_j, $\delta_{ij} = 1$ if $\mathbf{x}_i \in s_j$ otherwise $\delta_{ij} = 0$, j=1,…,q (j is the cluster index). The kernel density estimate for the whole set of observations is a combination of these partial estimates which can be written as

$$\hat{f}_H(\mathbf{x}) = \frac{1}{N} \sum_{j=1}^{q} \frac{1}{N_j} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)\delta_{ij}$$

$$= \sum_{j=1}^{q} \frac{N_j}{N} \hat{f}_j(\mathbf{x}) = \sum_{j=1}^{q} \pi_j \hat{f}_j(\mathbf{x})$$

where π_j is the proportion of the set of all observations that fall in cluster s_j. If two clusters overlap (typical of observations at boundaries) then the density overlap between the two clusters, s_r and s_t say, as outlined by Jensen et al. (2004), is given by

$$C_{rt}(s) = \int \hat{f}_r(\mathbf{x})\, \hat{f}_t(\mathbf{x}) d\mathbf{x}$$

and using equation (6.9) this becomes

$$C_{rt}(s) = \int \left(\frac{1}{N_r} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)\delta_{ir} \right) \left(\frac{1}{N_t} \sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x} - \mathbf{x}_i)\delta_{it} \right) d\mathbf{x}$$

According to Erdoğmus et al. (2006), the collective density overlap for all clusters can be shown to be

$$C(s) = N^2 \sum_{r \neq t}^{q} \pi_r \pi_t C_{rt}(s) \tag{6.10}$$

According to Erdoğmus et al., minimising equation (6.10), which is the mincut objective function in affinity based clustering, will lead to optimal clustering of the N observations. This attests to the role played by kernel density estimation in clustering procedures.

A visual inspection of contours of the estimated kernel density will be used for the determination of clusters in the data in this study. Li, Ray and Lindsay (2006) had a

more detailed study of such mode based clustering methods. The modes can be identified easily as regions of high density in the scatter and contour diagrams.

6.4. Data simulation and kernel density estimation

The data for this study comprises of the piecewise regression slope parameters that were obtained in Chapter 4 and their standard errors for the seven genotypes. It was found out, in Chapter 4, that the delignification and bleaching processes had significant effects on all response variables with the finishing stage not having a significant effect (see Table 4.3). Based on this general finding, the slope parameters for the delignification and bleaching processes were used to classify the seven genotypes.

Kernel density estimation, being an estimation method for a probability distribution, requires a sizeable sample size but the piecewise slope parameters estimated in Chapter 4, were not replicated hence there was need to use these slope parameters, with the assumption of normality, to generate more data. According to Silverman (1986), if a fairly small (<0.1) relative mean square error (MSE), given by $MSE = E\{\hat{f}(x) - f(x)\}^2 / f(x)^2$, is to be achieved, then for bivariate data a minimum sample size of 19 is required. Here $\hat{f}(x)$ is the estimated density while $f(x)$ is the true density of the variable of interest. Our variables of interest are the delignification and bleaching slopes hence a bivariate kernel density is to be estimated. Fifty sets of delignification and bleaching values will be generated for each genotype in a manner that is described in Section 6.4.1 that follows. The availability of the slope parameter estimates for each genotype and their standard errors makes it much easier to generate more variates around them. The general view is that if we know parameter estimates of a distribution then we can simulate the envisaged distribution hence we can produce its visualization by means of graphs. Since we have two slope parameters, which are individually normal and are also correlated, we can simulate a bivariate normal distribution for such slope parameters.

6.4.1. Simulating the bivariate normal distribution

The objective is to generate correlated variates, Y_1 and Y_2 say, where

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \right).$$

The parameters β_1 and β_2 are the delignification and bleaching slopes respectively, ρ is the correlation between the two slopes, σ_1 and σ_2 are the standard errors for the slope parameters. The variates Y_1 and Y_2, in this regard, are the variable slope parameters for the delignification and bleaching slopes respectively. To generate such data, it is necessary to obtain the Cholesky decomposition (Gentle, 1998; Wicklin, 2013) of the covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$ which can be shown to be

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \mathbf{\Sigma^{1/2}}(\mathbf{\Sigma^{1/2}})^{\mathbf{T}} = \begin{bmatrix} \sigma_1 & 0 \\ \rho\sigma_2 & \sqrt{\sigma_2^2(1-\rho^2)} \end{bmatrix} \begin{bmatrix} \sigma_1 & \rho\sigma_2 \\ 0 & \sqrt{\sigma_2^2(1-\rho^2)} \end{bmatrix}.$$

The correlated variates Y_1 and Y_2 are then generated as

$$Y_1 = \beta_1 + \sigma_1 \times Z_1$$

$$Y_2 = \beta_2 + \rho\sigma_2 Z_1 + \sqrt{\sigma_2^2(1-\rho^2)} \times Z_2$$

where Z_1 and Z_2 are standard normal variates that can be generated easily in most statistical packages like SAS. Wicklin (2013) provides a comprehensive description of how to simulate various kinds of data and simulation codes used in this study borrow heavily from his documentation of simulation methods. Estimates of ρ, σ_1 and σ_2 for the seven genotypes were obtained from sample data collected in laboratory experiments and were discussed in Chapter 4.

6.4.2. Density estimation from simulated data

The SAS procedure Proc KDE (SAS/STAT, 2008) was used to obtain the kernel density estimates. The procedure produces contour and three dimensional graphs which aid in the identification of peaks in the estimated density functions. The simulated delignification and bleaching slopes data for the seven genotypes were mixed together into a single data set with genotype identifiers. Scatter plots with genotype markers were then produced to help in the identification of the genotypes in the scatter plots. Genotype markers were then used to identify genotype that were closer together. This helps in grouping genotypes that can be mixed together during

processing. The three-dimensional plots help to identify peaks in the distributions and the number of peaks would indicate the number of genotype groups or clusters.

6.5. Results and discussions

The results are organised as follows: first the summary table of parameters used in the simulation is presented and then the kernel density estimation results are presented in tables and figures with discussions and descriptions. The parameters used for the simulation were obtained from piecewise linear regression models which were fitted to the chemical pulp processing data in Chapter 4. The pulping process was divided into three stages, namely, delignification, bleaching and finishing. Piecewise linear regression models were then fitted to the data for each chemical property analysed. The models had nodes at the changeover points from delignification to bleaching thus the parameters of the two line segments representing delignification and bleaching were obtained and used to characterise the genotypes. These parameters were then used to simulate the data that was needed for kernel density estimation. Densities estimated in this way are more accurate when the sample size is large (Seaman et al., 1999).

6.5.1. Kernel density estimation and genotype classification using lignin

The piecewise linear regression parameter estimates for lignin are presented in Table 6.2 below. These parameters, which differ across the different genotypes, can be used to group timber genotypes into clusters of those genotypes that respond to the pulping process in a similar way.

The negative delignification and bleaching parameter estimates indicate that lignin levels decline during the two sub-processes (Delignification and Bleaching) of chemical pulping, and the correlation of -0.7776 indicates that bleaching tends to reduce lignin levels at a higher rate if delignification leaves higher levels of lignin. This conforms to the fact that lignin has to be reduced down to some product specific levels.

Table 6.2 Slope parameters for Lignin

***Geno*type**	**Chemical property: Lignin**				
	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	-2.073	0.286	-0.449	0.128	-0.7776
Egrandis	-2.157	0.286	-0.284	0.128	-0.7776
Esmithii	-2.673	0.202	-0.556	0.091	-0.7776
Enitens	-1.52	0.286	-0.227	0.128	-0.7776
E gc	-2.453	0.286	-0.69	0.128	-0.7776
EguA	-2.467	0.286	-0.428	0.128	-0.7776
EguW	-1.538	0.286	-0.396	0.128	-0.7776

The kernel density estimation for lignin is shown Figures 6.1(a), which is a two-dimensional representation with contour lines indicating that there are three modes of different densities in the data. This is indicative of the existence of three possible groupings of the genotypes. Figure 6.1(c), which is a three-dimensional representation of the same estimated density, also shows that there are three distinct peaks in the data. It must be mentioned that optimum bandwidth selection makes it possible to bring out the peaks in the data.

The optimal bandwidths as calculated using equation (6.8) are h_1=0.19 for the delignification and h_2=0.017 for the bleaching slopes. Figures 6.1(b) an 6.1(d) show that when the bandwidths were doubled (h_1=0.38, h_2=0.14) it was not possible to discern the three peaks in the estimated density evident when the original optimal bandwidths were used (h_1=0.19, h_2=0.07). it must be borne in mind that all density estimates will be based on the optimal bandwidth. A reduced bandwidth would produce a more rugged (less smooth) density estimate with many peaks of no apparent importance while an oversmoothed density will not distinguish any clusters in the simulated data.

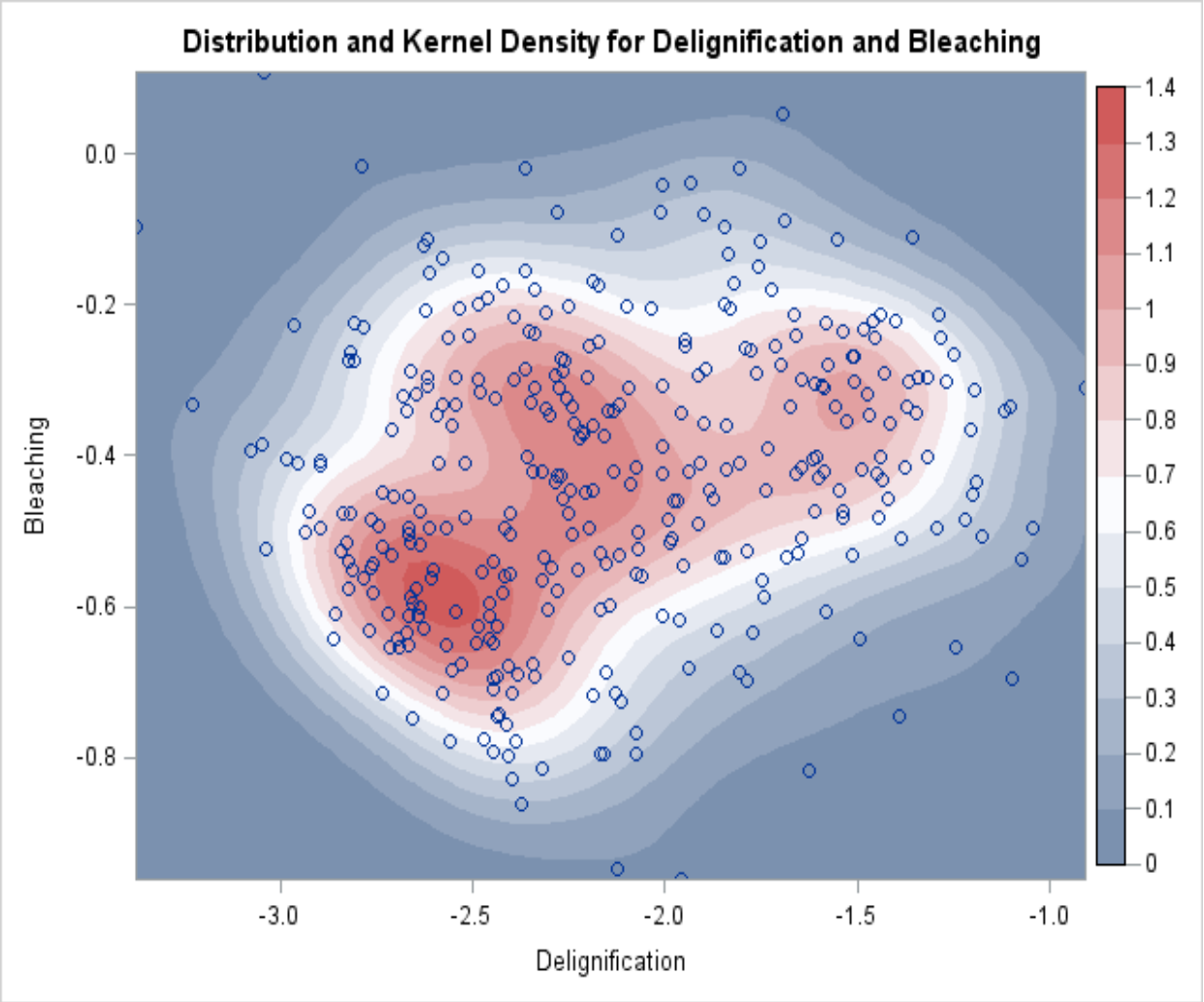


Figure 6.1(a). Scatter/contour plot for lignin (Optimal bandwidths: Delignification (h1)= 0.19 Bleaching (h2) = 0.07).

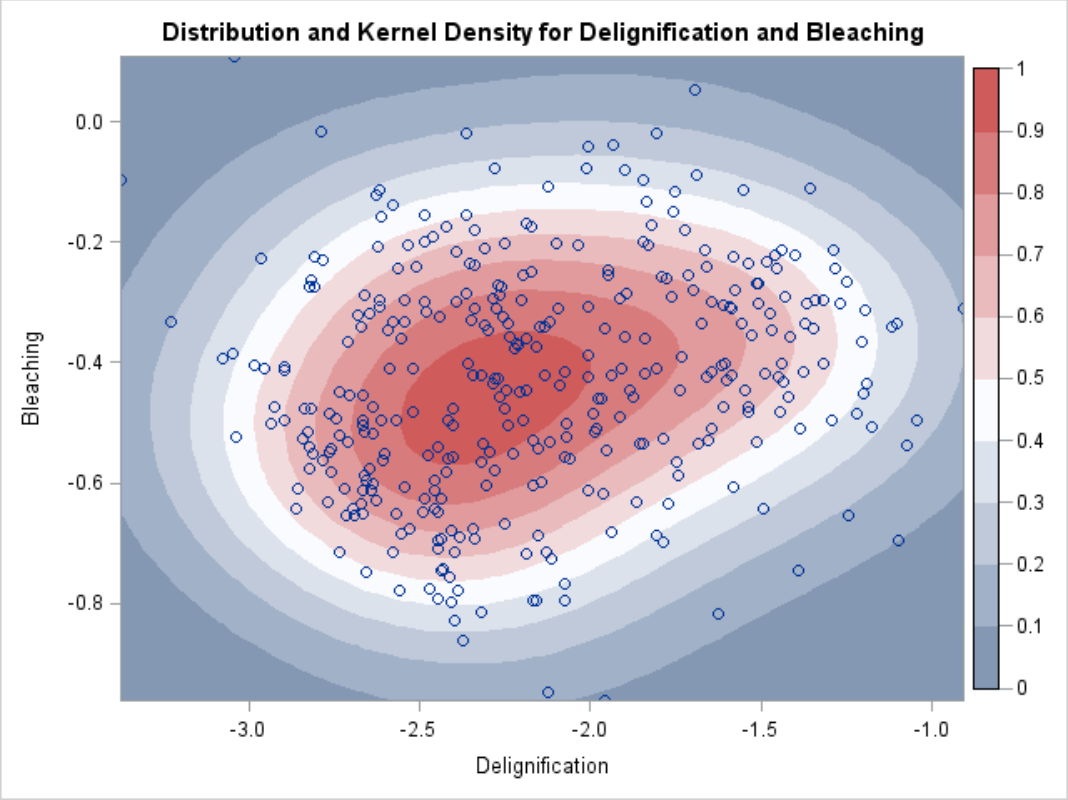


Figure 6.1(b). Scatter/contour plot for lignin (Optimal bandwiths×2)

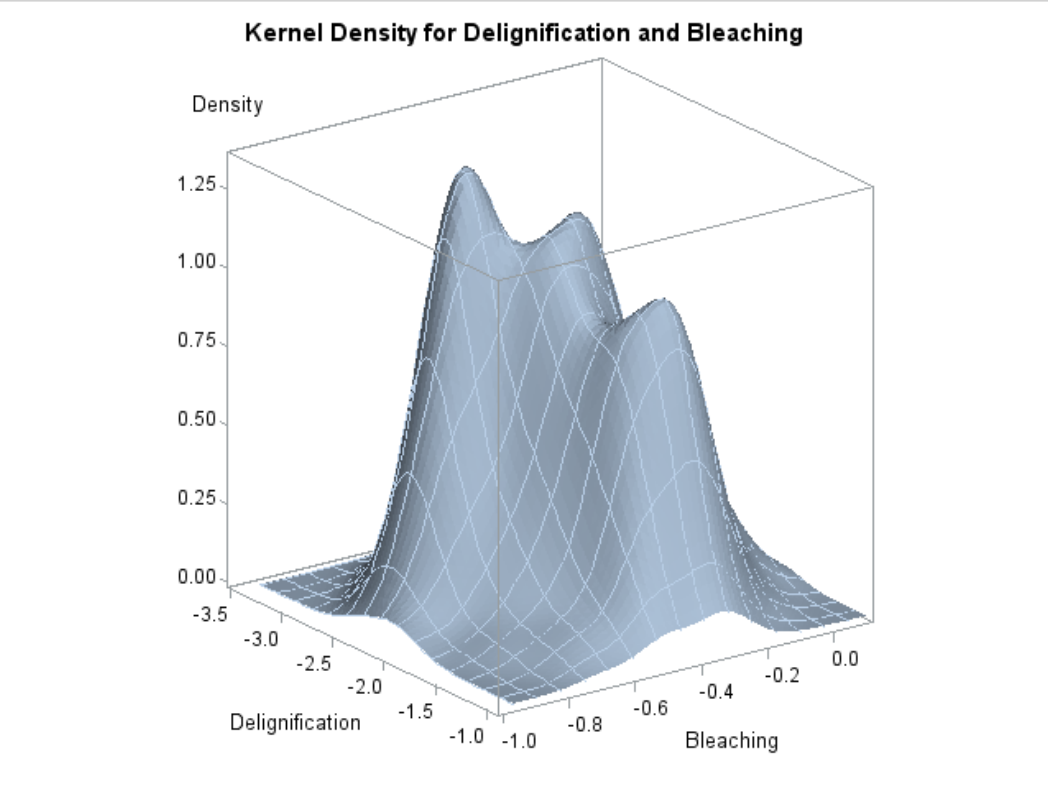


Figure 6.1(c). Surface plot for lignin (Optimal bandwidths: Delignification (h1)= 0.19 Bleaching (h2) = 0.07).

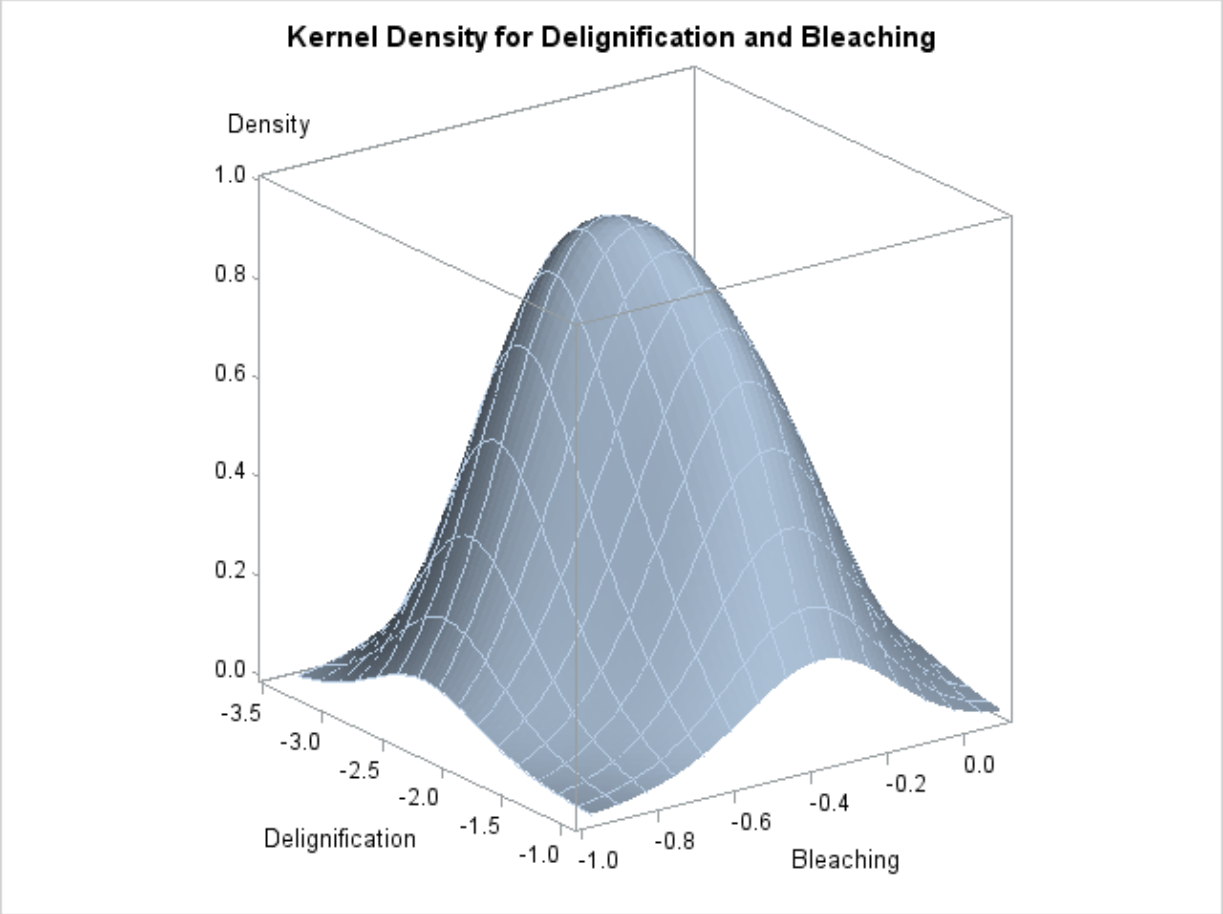


Figure 6.1(d). Surface plot for lignin (Optimal bandwidths×2).

Figure 6.1(e), which has all genotypes clearly marked, can be used to identify which genotypes are closer together and hence form a cluster of genotypes that can be mixed during processing. Genotypes GCG (Egc) and Esmithii form a region with the highest density (highest peak) hence forming a cluster which means that the two genotypes can be mixed together during processing based on the behaviour of lignin during delignification and bleaching.

Egrandis, GUA (EguA) and Edunnii form the second cluster with the second highest density and lastly Enitens and GUW (EguW) form a cluster of their own although they seem to have minimal mixing across the line segment AB. The two genotypes can only be mixed if it is necessary but line (AB) seems to suggest that they may be processed separately.

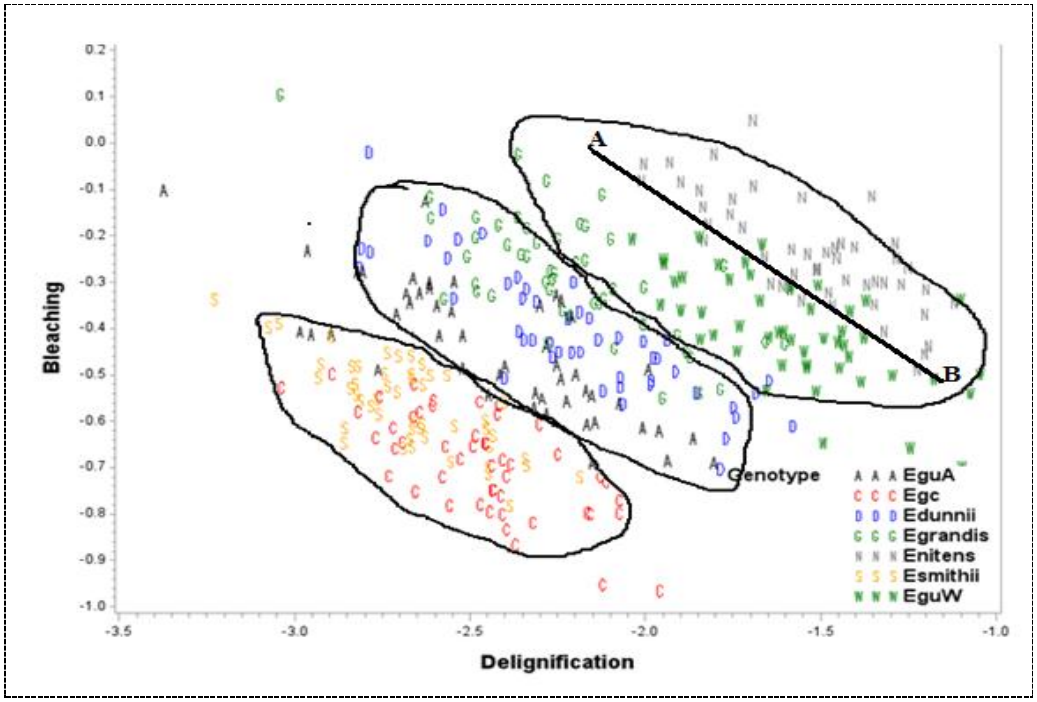


Figure 6.1(e). Genotype classification based on identified peaks for lignin data

6.5.2. Kernel Density estimation and genotype classification using α-Cellulose

The piecewise linear regression parameter estimates for α-Cellulose, which were used for the simulations in kernel density estimation, are presented in Table 6.3 below.

Table 6.3. Slope parameters for α-cellulose

Genotype	Chemical Property: α-cellulose				
	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	0.361	0.833	1.271	0.286	0.001
Egrandis	2.074	0.833	0.899	0.286	0.001
Esmithii	0.202	0.589	0.964	0.202	0.001
Enitens	1.393	0.833	1.216	0.286	0.001
E gc	1.663	0.833	0.843	0.286	0.001
EguA	1.474	0.833	1.094	0.286	0.001
EguW	0.923	0.833	1.031	0.286	0.001

The α-Cellulose results in Figure 6.2 (a) and (b) show that the kernel density estimate has only one peak which suggests that all the seven genotypes form one cluster. This

means that they do not behave differently as far as changes in α-Cellulose is concerned. Since the slope parameters of α-Cellulose under delignification and bleaching do not have more than one genotype clusters it then follows that α-Cellulose cannot be used a clustering variable under this procedure. Only those variables that produce distinct clusters become important clustering variables while those that do not can be considered non-essential clustering variables.

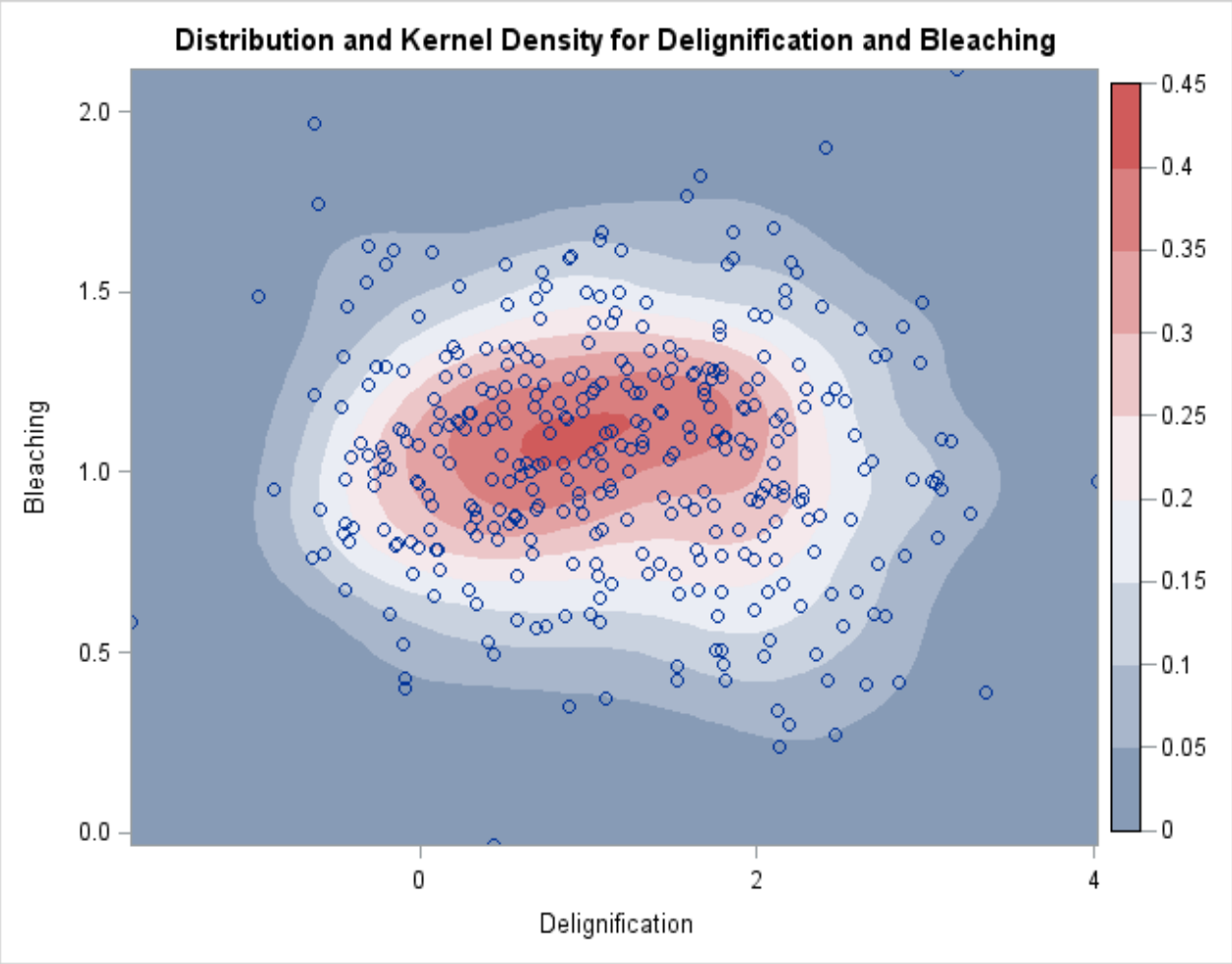


Figure 6.2(a) Contour plot of α-Cellulose (Optimal bandwidths: Delignification (h1)= 0.37 Bleaching (h2) = 0.12)

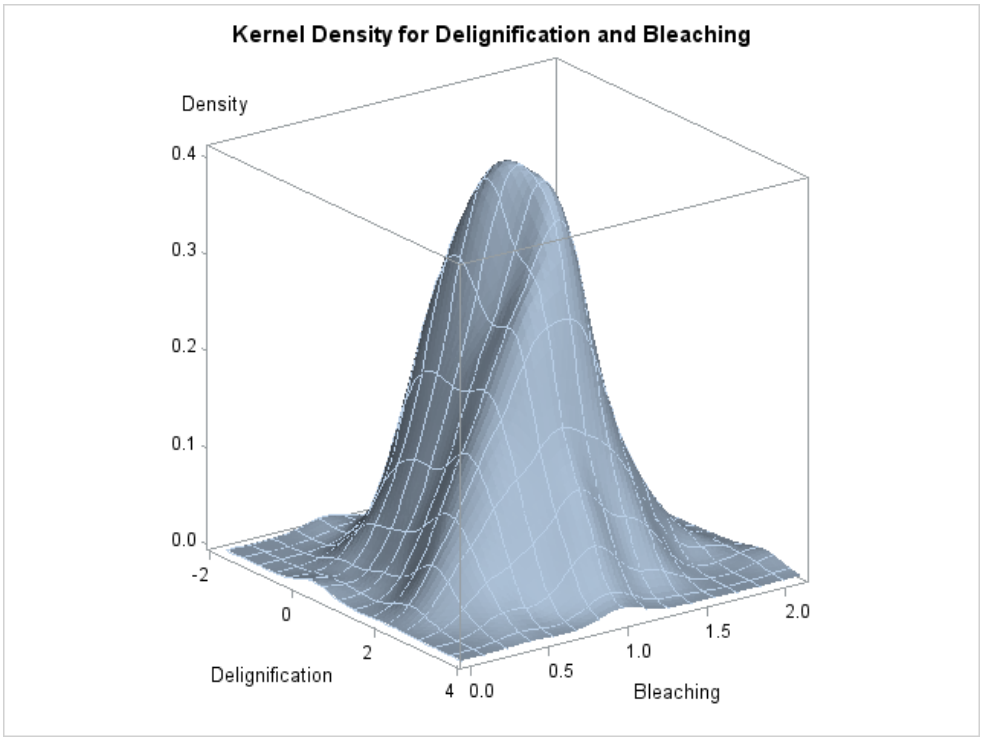


Figure 6.2(b). Surface plot of α-Cellulose (Optimal bandwidths: Delignification (h1)= 0.37 Bleaching (h2) = 0.12)

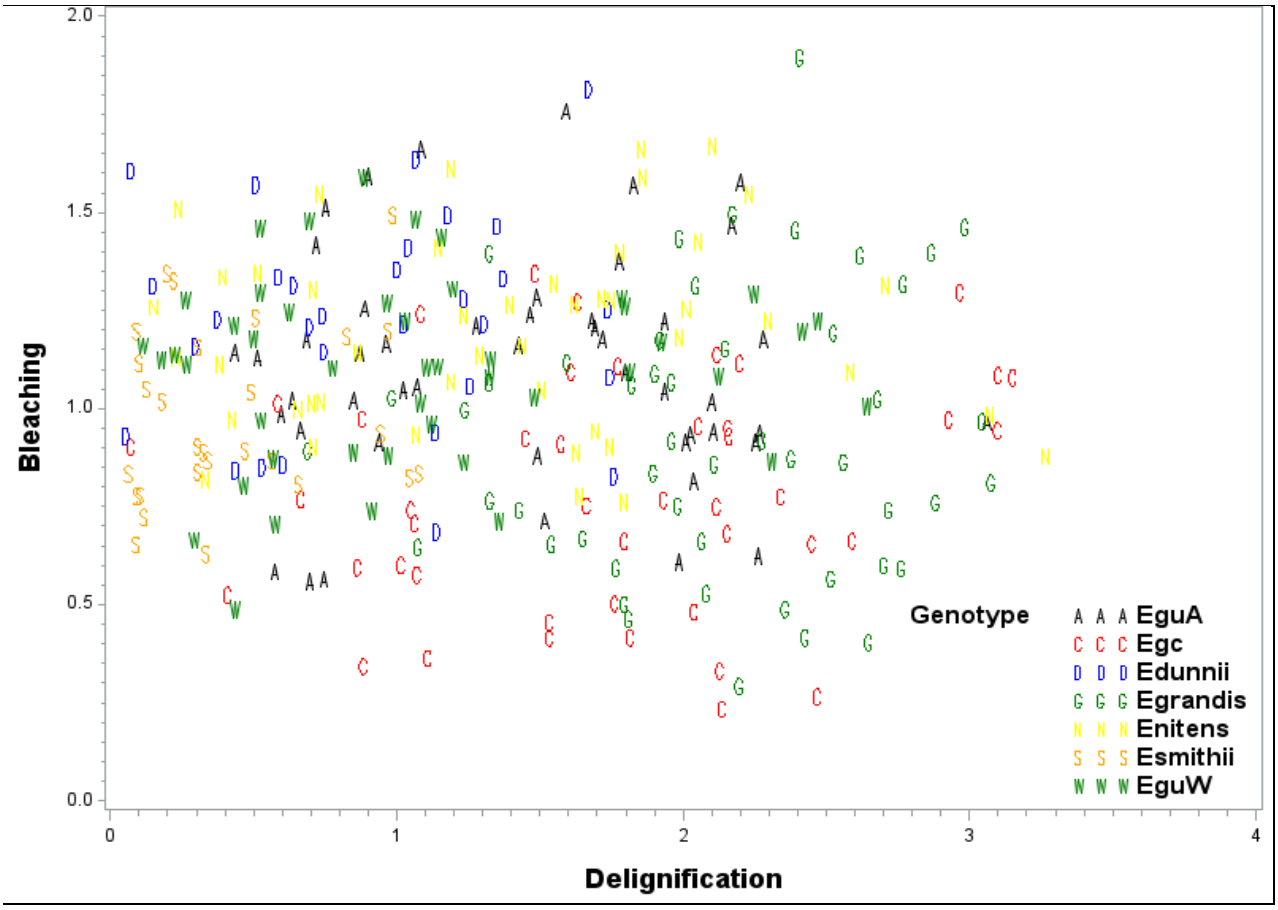


Figure 6.2(c). Genotype classification based on identified peaks for α-Cellulose data

6.5.3. Kernel density estimation and genotype classification using viscosity

Viscosity is a traditional measure of the length of the cellulose molecule chains. It is an important property of dissolving pulp. Delignification and bleaching slopes for the viscosity of the seven genotypes are presented in Table 6.4 below and the corresponding kernel density estimate in presented in Figures 6.3(a), (b) and (c) below.

Table 6.4. Slope parameters for viscosity

Chemical Property: Viscosity					
Genotype	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	-10.681	10.516	-2.427	5.114	0.000
Egrandis	4.501	10.516	0.019	5.114	0.000
Esmithii	2.473	7.436	-5.471	3.616	0.000
Enitens	3.062	10.516	-2.696	5.114	0.000
E gc	-2.143	10.516	-7.016	5.114	0.000
EguA	0.592	10.516	-9.878	5.114	0.000
EguW	2.986	10.516	-7.718	5.114	0.000

The contours in the density function estimate in Figure 6.3(a) and the surface plot in Figure 6.3.(b) show that there is one dominant region of high density in the data which means than the data forms one cluster. However in Figure 6.3(c) the genotypes E.grandis and E.nitens seem a bit detached from the other genotypes but do not in themselves from a region of high density. In general viscosity is not an important genotype clustering variable for the chemical pulping process. It is noted however, that Figure 6.3c seem to suggest some linear form of clustering. This is an indication of the weakness of this clustering method. It doesn't seem to pick out irregular clusters other that circular ones.

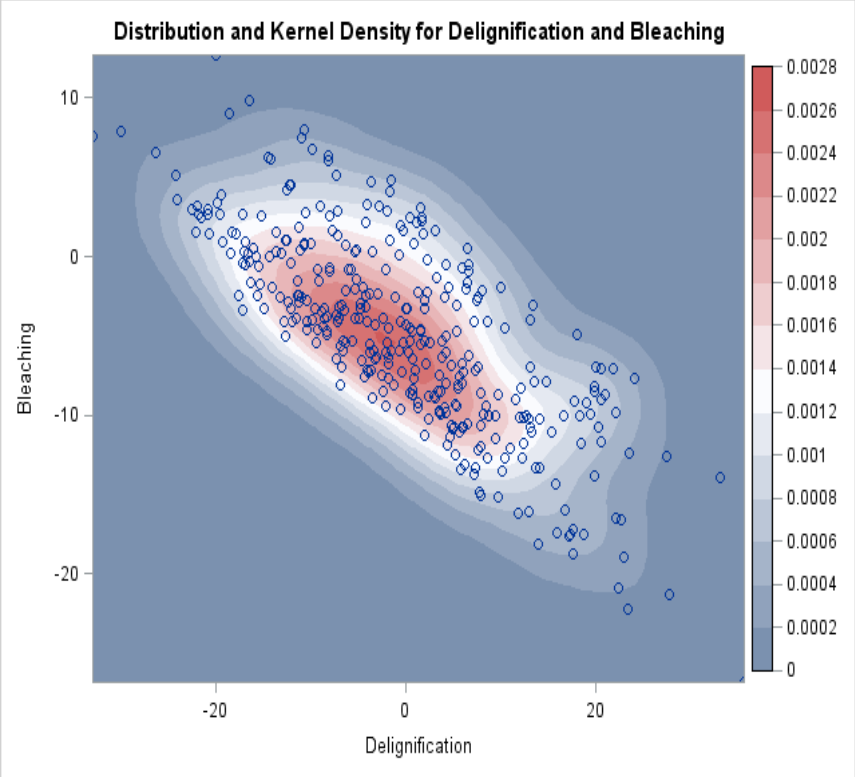


Figure 6.3(a). Contour plot of viscosity (Optimal Bandwidths: Delignification (h1)= 4.44, Bleaching (h2) = 2.29)

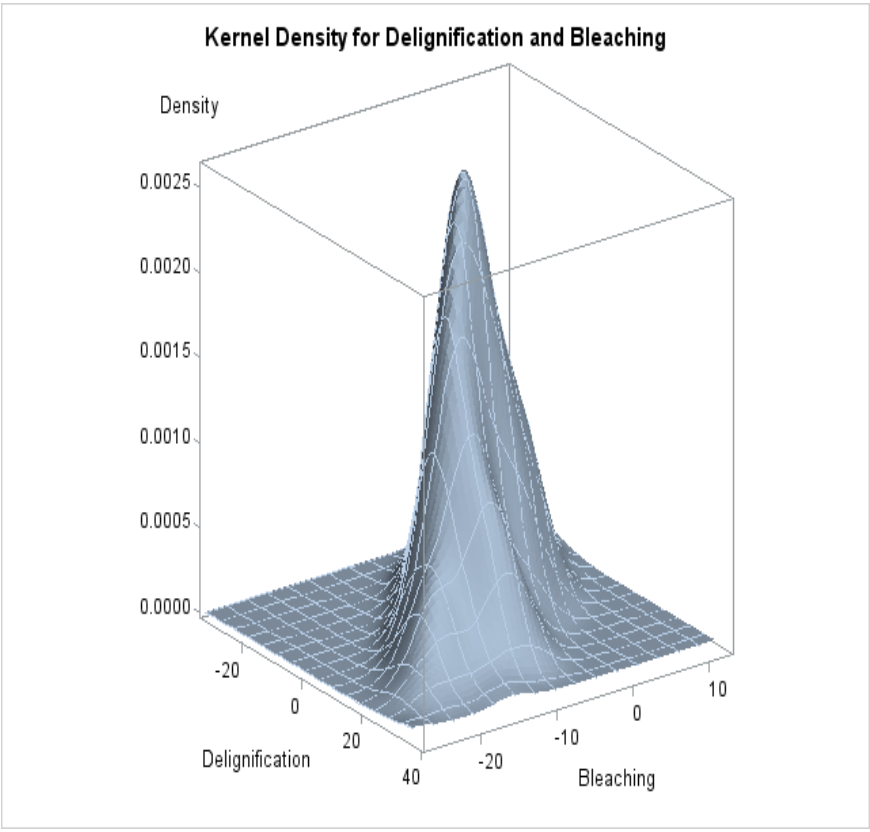


Figure 6.3(b). Surface plot of viscosity (optimal bandwidth).

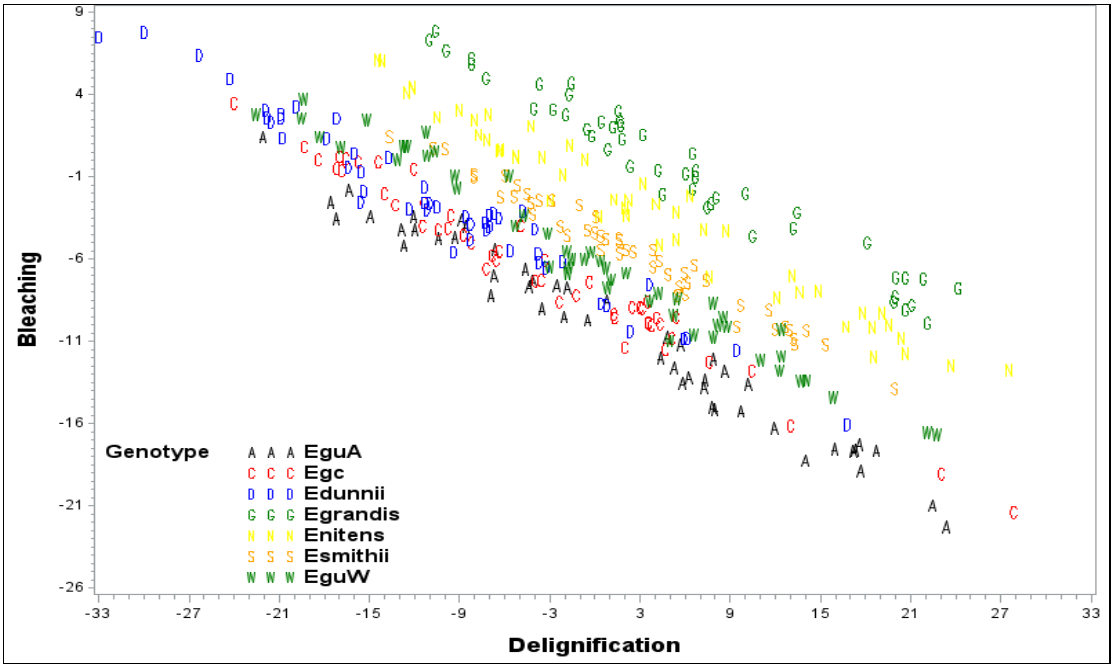


Figure 6.3(c). Genotype classification based on identified peaks for viscosity

6.5.4. Density estimation and genotype classification based on γ-Cellulose results

For the γ-Cellulose results delignification and bleaching slopes for the seven genotypes are presented in Table 6.5 below and kernel density estimation in Figures 6.4(a), (b) and (c).

Table 6.5. Slope parameters for viscosity

Genotype	**Chemical *Property: γ-cellulose***				
	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	0.283	0.56	-1.24	0.197	0.000
Egrandis	-2.117	0.56	-0.795	0.197	0.000
Esmithii	-1.17	0.396	-0.97	0.14	0.000
Enitens	-1.446	0.56	-1.103	0.197	0.000
E gc	-1.707	0.56	-0.816	0.197	0.000
EguA	-1.401	0.56	-1.086	0.197	0.000
EguW	-0.794	0.56	-0.894	0.197	0.000

Figures 6.4 (a) and (b) show that the kernel density estimate for γ-cellulose has two regions of high density. One region of high density has a much lower peak and comprises of the genotype E.dunnii on its own. As far as γ-cellulose is concerned E.dunnii behaves in a way that does not conform to the general behaviour of the other genotypes. The other region of high density is made up of the remaining six genotypes which form one cluster. This means that changes in γ-cellulose during delignification and bleaching tend to be similar for all genotypes except E.dunnii.

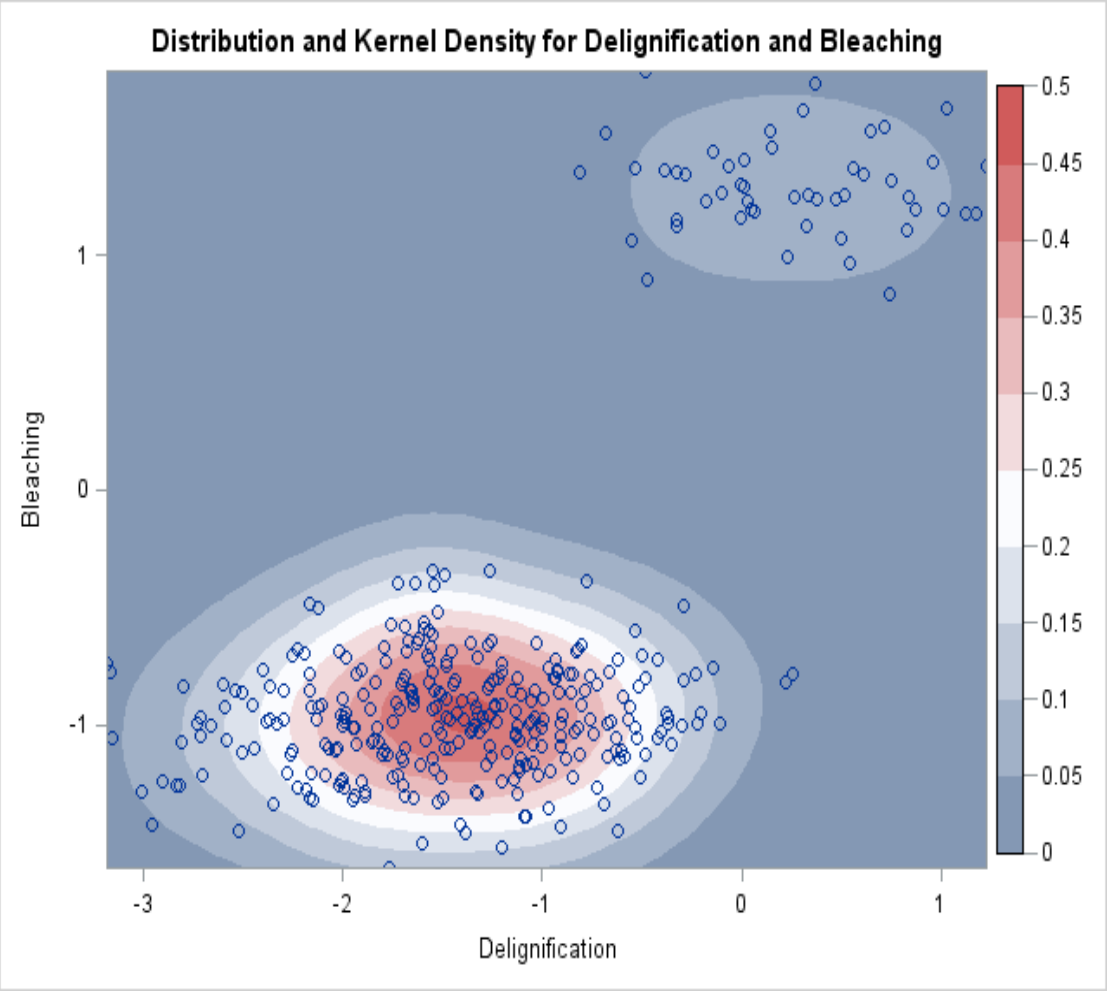


Figure 6.4(a). Contour plot of γ-Cellulose (Optimal Bandwidths: Delignification (h1)= 4.44, Bleaching (h2) = 2.29)

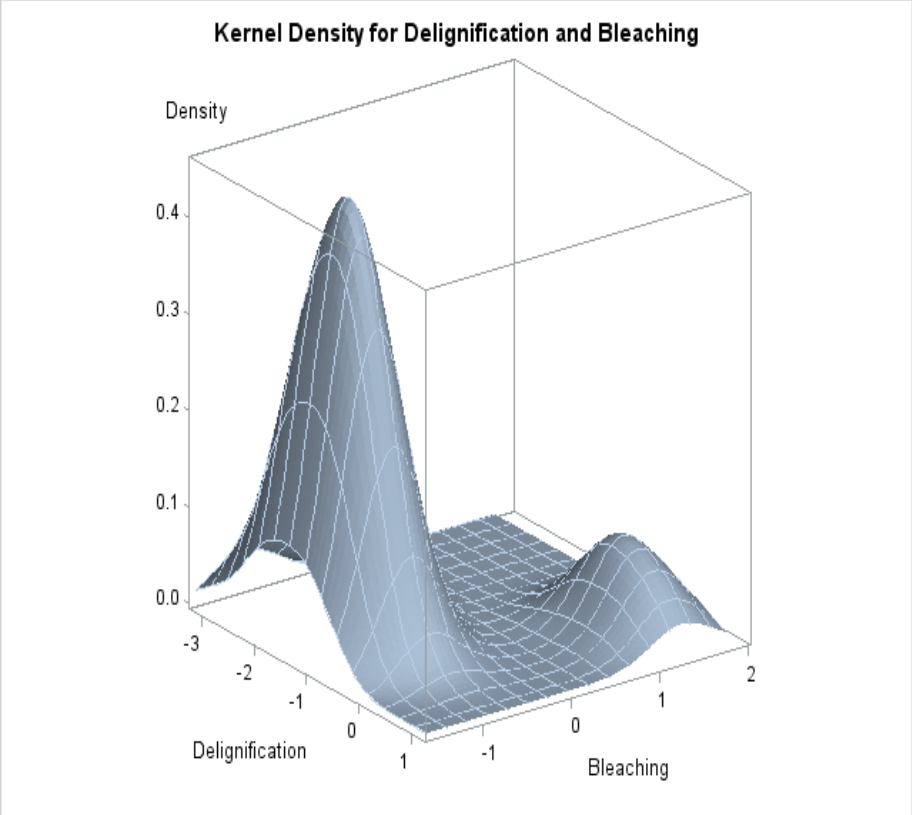


Figure 6.4(b). Surface plot of γ-Cellulose (optimal bandwidth).

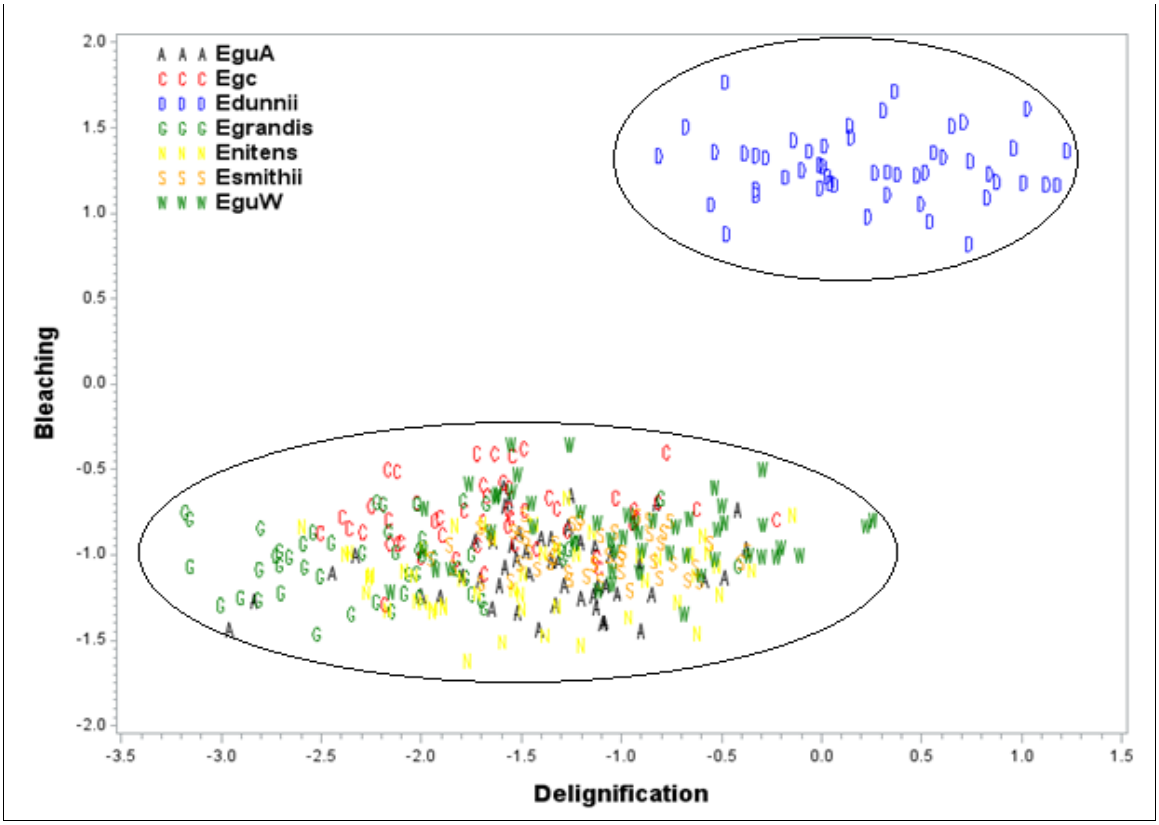


Figure 5.4(c). Genotype classification based on identified peaks for γ-Cellulose

6.5.5. Density estimation and genotype classification for Copper Numbers

As indicated in the slope parameter estimates in Table 6.6, copper numbers decline during the two sub-processes of delignification and bleaching. The correlation between the two slope parameters was found to be very low (<0.0001) and set to zero in the statistical simulations.

Table 6.6. Slope parameters for copper numbers

Chemical Property: Copper Numbers					
Genotype	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	-1.064	0.241	-0.514	0.083	0.000
Egrandis	-1.277	0.241	-0.414	0.083	0.000
Esmithii	-1.245	0.17	-0.417	0.053	0.000
Enitens	-0.657	0.241	-0.452	0.083	0.000
E gc	-1.534	0.241	-0.418	0.083	0.000
EguA	-1.397	0.241	-0.41	0.083	0.000
EguW	-0.881	0.241	-0.408	0.083	0.000

Results in Figures 6.5 (a), (b) and (c) show that the kernel density estimate for copper numbers, as shown in the contour scatter diagram and the surface plot, have only one region of high density and the colour coded scatter scatter diagram indicate that there is a fair mix of the genotypes. This suggests that all the seven genotypes fall into one cluster hence copper numbers can not be used as a grouping variable when determining which genotypes can be mixed together during chemical pulping.

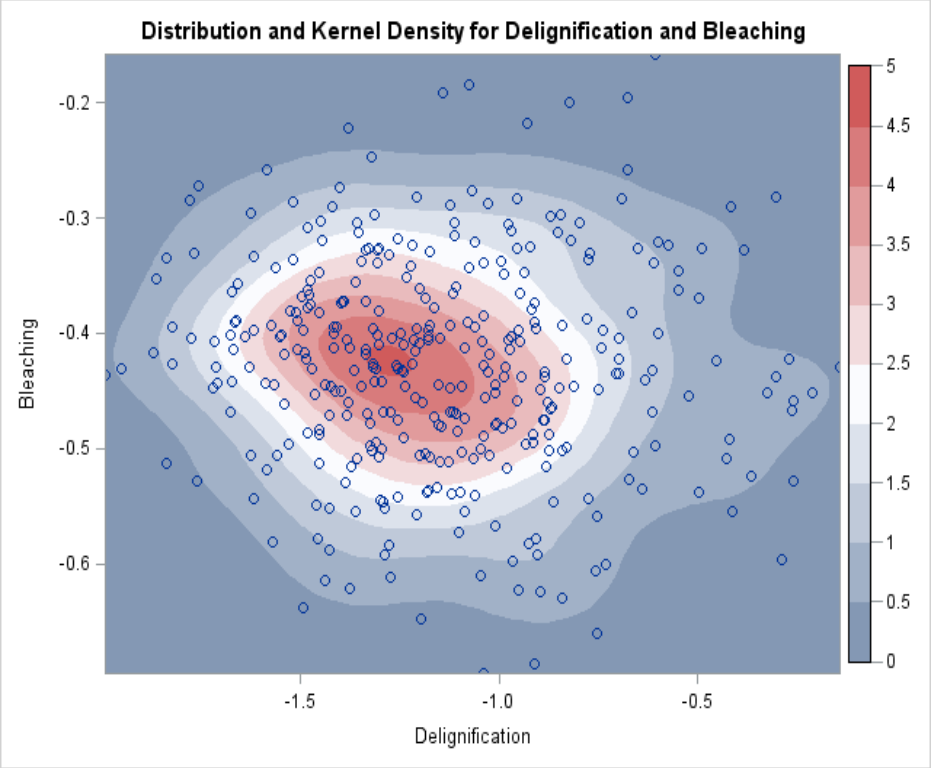


Figure 6.5(a). Contour plot of copper numbers (Optimal Bandwidths: Delignification (h1)= 0.13, Bleaching (h2) = 0.035).

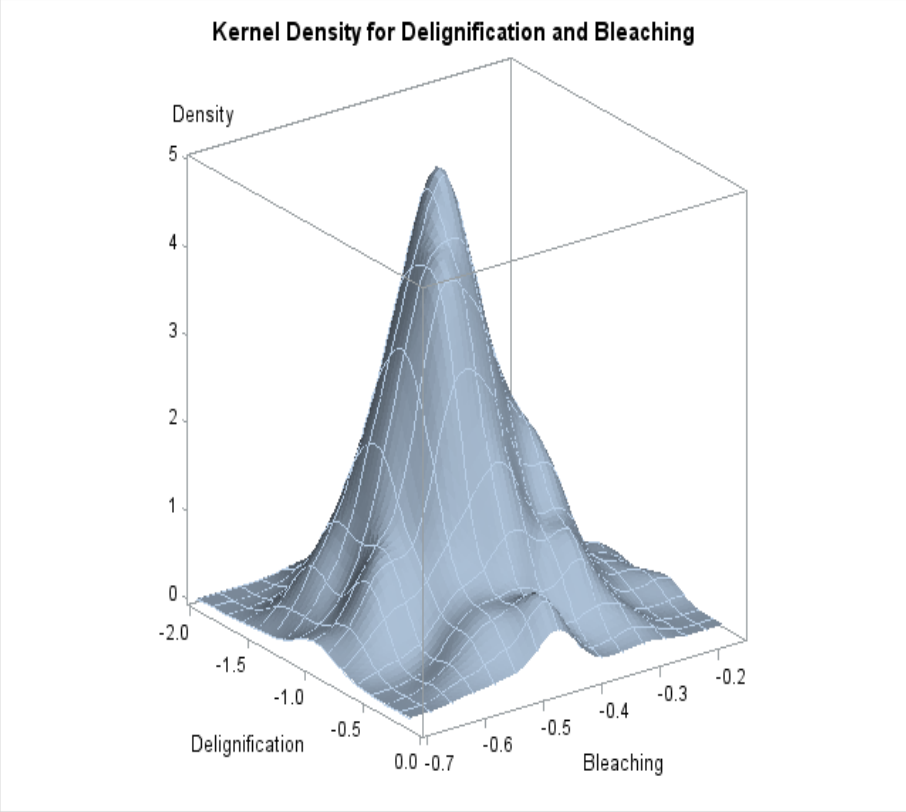


Figure 6.5(b). Surface plot of copper numbers (optimal bandwidth).

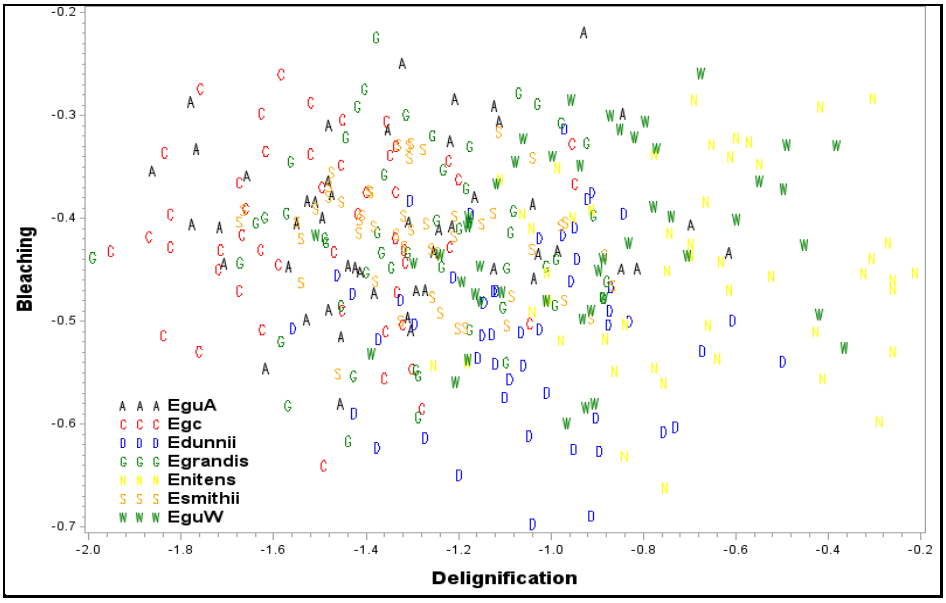


Figure 6.5(c). Genotype classification based on identified peaks for copper numbers

6.5.6. Density estimation and genotype classification for Glucose

Glucose, like *α*-cellulose, has positive slopes both during delignification and bleaching. The slope parameter estimates and their standard errors are listed in Table 6.7. The correlation between the delignification and bleaching parameters was not signficant (<0.0001) and it was set to zero during the data simulations.

Table 6.7. Slope parameters for glucose

Genotype	**Chemical Property: Glucose**				
	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	2.010	0.461	1.226	0.157	0.000
Egrandis	2.467	0.461	0.792	0.157	0.000
Esmithii	1.884	0.345	1.035	0.111	0.000
Enitens	3.493	0.461	0.987	0.157	0.000
E gc	3.640	0.461	0.701	0.157	0.000
EguA	2.908	0.461	0.949	0.157	0.000
EguW	2.493	0.461	0.675	0.157	0.000

The Results in Figures 6.6(a), (b) and (c) show that the kernel density estimate for glucose had two regions of high density which split the genotypes into two clusters.

The cluster comprising of the genotypes GUW, Egrandis, GCG, GUA and Enitens has the higher density of the two while the cluster comprising of Edunnii and Esmithii has a lower density.

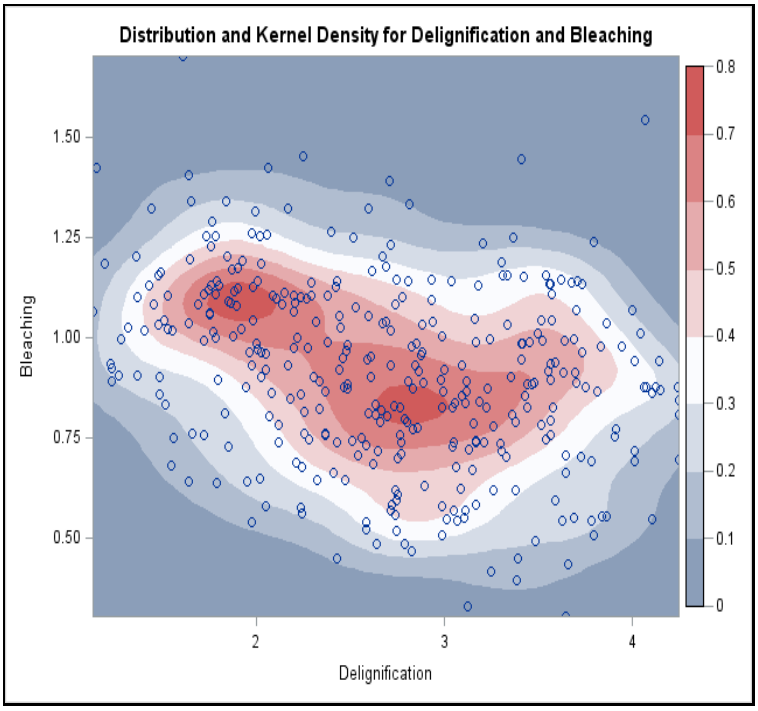


Figure 6.6(a). Contour plot of Glucose (Optimal Bandwidths: Delignification (h1)= 0.13, Bleaching (h2) = 0.035).

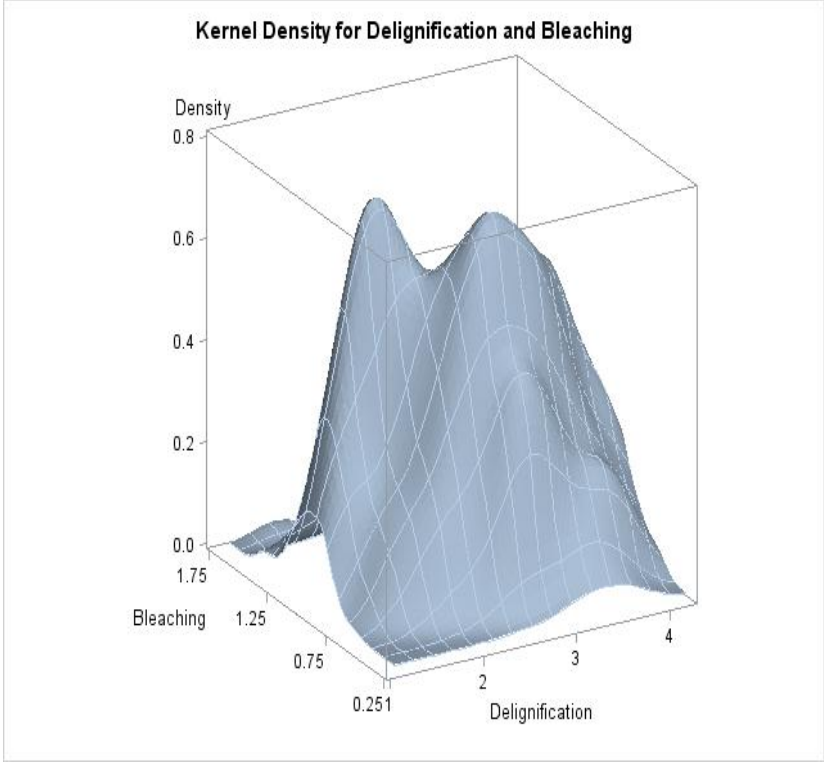


Figure 6.6(b). Surface plot of glucose (optimal bandwidths).

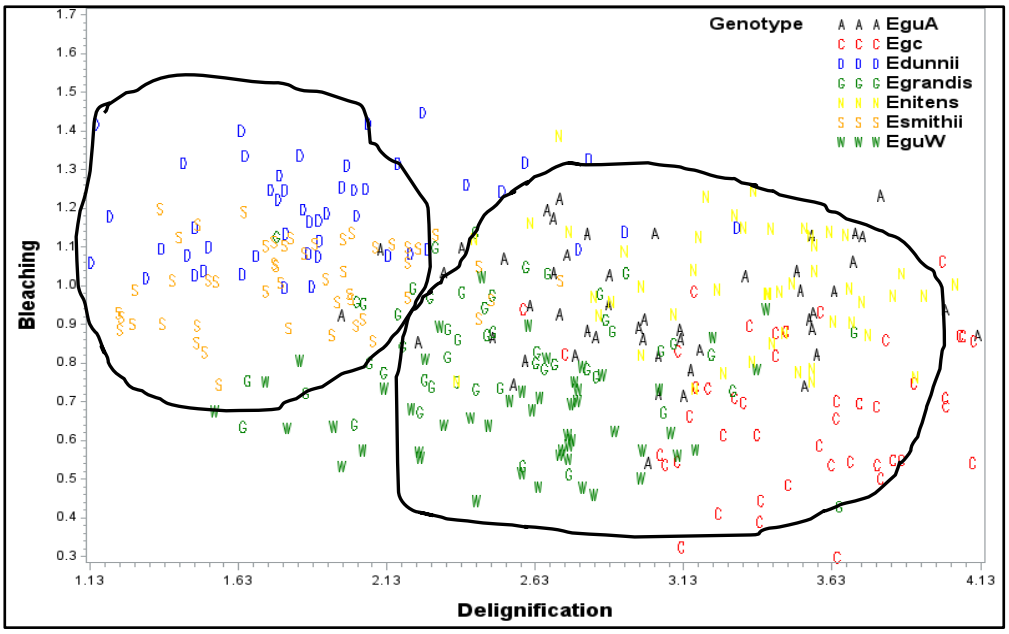


Figure 6.6(c). Genotype classification based on identified peaks for glucose

6.5.7. Density estimation and genotype classification for Xylose

The sub-processes of delignification and bleaching, progressively lower the levels of xylose in the dissolving pulp as indicated by the negative slope parameter estimates in Table 6.8. The delignification and bleaching slopes are not correlated which suggests that the two sub-processes reduce xylose independently.

Table 6.8. Slope parameters for xylose

Genotype	**Chemical Property: Xylose**				
	β_1 (Delignification)		β_2 (Bleaching)		Correlation(β_1, β_2)
	Slope Estimate	Standard error	Slope Estimate	Standard error	
Edunnii	-0.857	0.322	-0.565	0.096	0.000
Egrandis	-0.545	0.322	-0.402	0.096	0.000
Esmithii	-1.032	0.237	-0.528	0.068	0.000
Enitens	-2.279	0.322	-0.484	0.096	0.000
E gc	-0.817	0.322	-0.291	0.096	0.000
EguA	-0.626	0.322	-0.516	0.096	0.000
EguW	-0.951	0.322	-0.280	0.096	0.000

Figures 6.7(a), (b) and (c) show that the kernel density estimate for xylose has one cluster of high density compring of the genotypes genotypes GUW, Egrandis, GCG, GUA, Edunnii and Esmithii while Enitens forms a second cluster of low density.

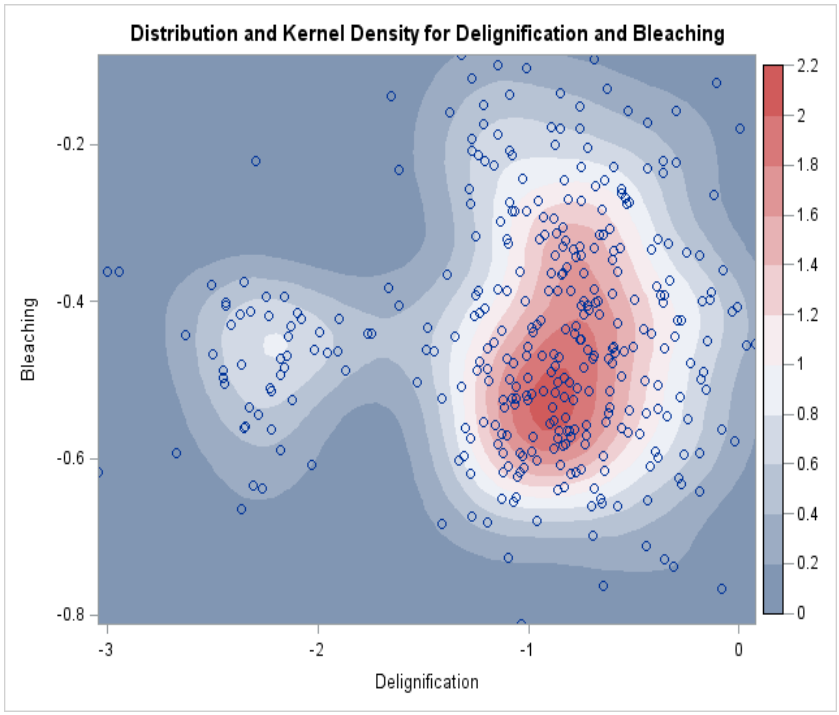


Figure 6.7(a). Contour plot of Xylose (Optimal Bandwidths: Delignification (h1)= 0.13, Bleaching (h2) = 0.035).

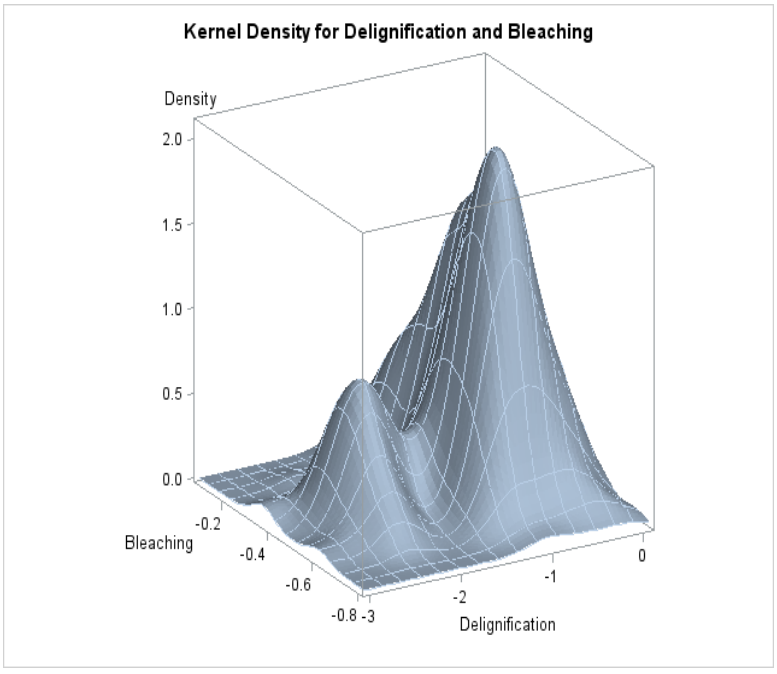


Figure 6.7(b). Surface plot of Xylose (optimal bandwidths).

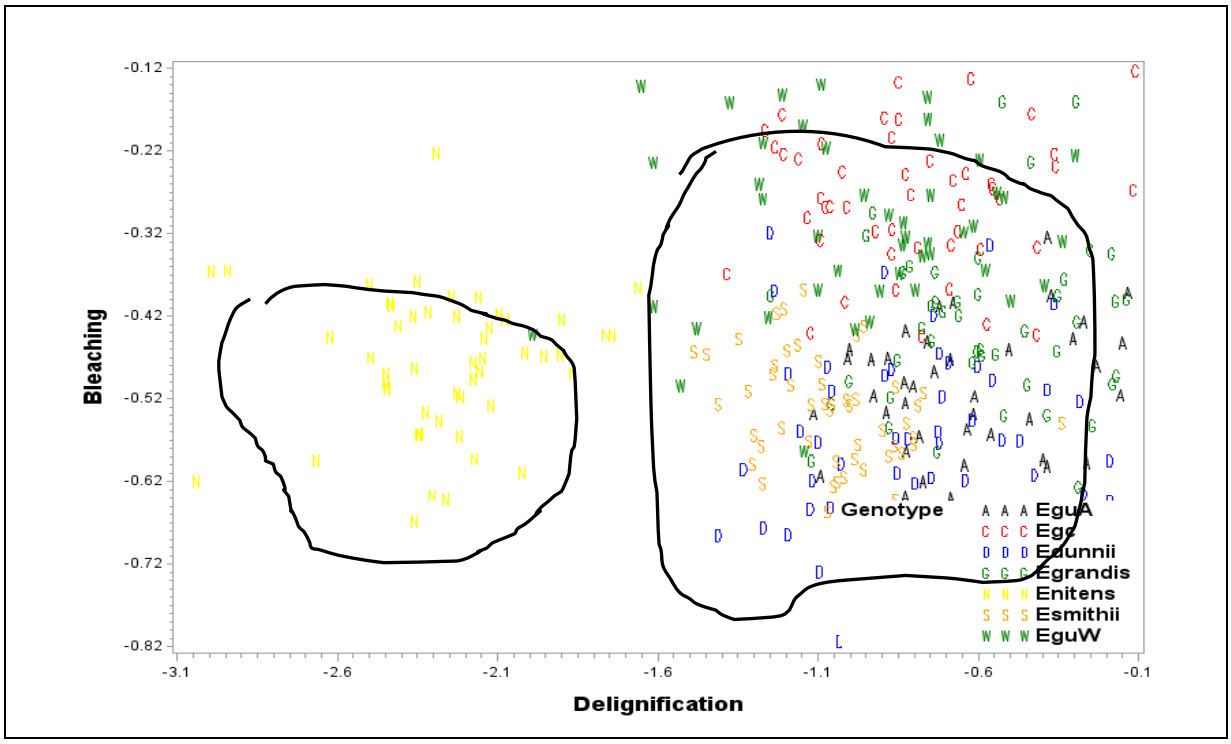


Figure 6.7(c). Genotype classification based on identified peaks for glucose

6.6. Summary of kernel density estimation and clustering results

Kernel density estimation for the seven genotypes produced different clusters as the chemical properties tended to respond to the processing stages differently. There is need to use these different clusters to come up with one overall clustering scheme. The number of times any two genotypes have been found to be in the same cluster, according to the seven chemical properties, can be used as a measure of similarity. The summary of the clustering results generated from the seven chemical properties using kernel density estimation as a clustering tool is presented Table 5.9 below. Based on this matching principle a similarity matrix for the seven genotypes was compiled and is presented in Table 6.10.

A score of 7 between any two genotypes indicates that, out of the seven chemical properties, the two genotypes always fell into the same cluster hence they are highly mixable. A score of 0, on the other hand, indicates that the pair of genotypes fell in

different clusters on all seven chemical properties hence not suitable for mixing together during processing.

Table 6.9. Summary of clusters generated by chemical properties under KDE

Chemical Property	Cluster 1	Cluster 2	Cluster 3
Lignin	Esmithii, E.gc	E.guA, Edunnii, Egrandis	Enitens, E.guW
α-Cellulose	E.smithii, E.gc, E.guA, E.dunnii, E.grandis, E.nitens, E.guW		
Viscosity	E.smithii, E.gc, E.guA, E.dunnii, E.grandis, E.nitens, E.guW		
γ-Cellulose	E.smithii, E.gc, E.guA, E.grandis, E.nitens, E.guW	E.dunnii,	
copper numbers	E.smithii, E.gc, E.guA, E.dunnii, E.grandis, E.nitens, E.guW		
Glucose	E.gc, E.guA, E.grandis, E.nitens, E.guW	E.dunnii, E.smithii	
Xylose	E.smithii, E.gc, E.guA, E.dunnii, E.grandis, E.guW	E.nitens	

Table 6.10. Number of times any two genotypes belonged to the same cluster

	E.smithii	E.gc	E.guA	E.dunnii	E.grandis,	E.nitens	E.guW
Esmithii	-						
Egc (GCG)	6	-					
EguA (GUA)	5	6	-				
Edunnii	5	4	5	-			
Egrandis,	5	6	7	5	-		
Enitens	4	5	5	3	5	-	
EguW (GUW)	5	6	6	4	6	6	-

The scores in Table 6.10 can therefore be used as some form of a mixability indicators for the seven genotypes. The higher the index between any two genotypes the more appropriate it is to mix them if necessary. A score of 7 would mean the two genotypes concerned are absolutely mixable. The two most mixable genotypes are Edunnii and Enitens with GUA and Egrandis with a score of 7.

6.7. Conclusion

The study managed to develop a form of scale that can be used to determine if any two genotypes can be mixed during processing based on their response to the two key sup-processes of delignification and bleaching. The behaviour of the genotypes, as measured by the rates of change in the chemical properties of lignin, *α*-cellulose,

γ-cellulose, viscosity, copper numbers, glucose and xylose, were used to develop a mixability matrix.

The underlying assumption in this study is that all the chemical properties are of equal importance and no single chemical property can override the values in the other properties. The chemical pulping process targets higher levels of *α*-cellulose but this chemical property did not produce different genotype clusters indicating that all genotypes had similar response profiles to the sub-processes of bleaching and delignification as far as α-cellulose is concerned. This means that, for the purpose of determining mixable genotypes, α-cellulose is not an important variable. Viscosity and copper numbers also came up as not important when determining which genotypes can be mixed during processing. This is in contract to results of Chapters 3 and 4 which made full use of all chemical properties. It can thus be said that the method discussed in this chapter uses information less efficiently than the methods of Chapters 3 and 4.

The behaviour of lignin, γ-Cellulose, glucose and xylose were deemed important determinants of how mixable the timber genotypes could be. The seven chemical properties studied are not the only ones involved in chemical pulping. A more comprehensive study incorporating as many chemical properties as possible is a possible area of further studies hence this study suggests an area of study that can be adopted for the optimisation of chemical pulping processes especially when there is a huge variety of timber genotypes available for processing. Mixing of timber varieties during processing will always occur especially when one genotype alone does not have economic quantities for processing. It is reasonable to mix genotypes for processing after considering how individual genotypes behave during chemical processing so that only those genotypes that have many properties in common are optimally mixed during processing.

There is still scope for further studies in this subject. More chemical properties can be added to make a more comprehensive study. The importance of each chemical property can also be considered in future studies so that some weights can be attached to them in the analysis. This study assumed equal importance on the chemical properties.

This chapter successfully developed a grouping mechanism for different genotypes for the purpose of chemical pulping, however, the variables where not considered together, especially as far as their inter-correlation are concerned. It is worthwhile to attempt to understand how the chemical properties behave together by looking at the correlations between their evolutions through the chemical pulping process. Chapter 7 looks at how the variables evolve through the processing stages together and how they interact with each other though a correlation analysis of their evolutions using joint modelling.

Table 6.11. Percentiles for the KDE estimates for the seven chemical properties.

	Percentiles generated by kernel density estimation (the β parameters as variables)													
	Lignin		**α-Cellulose**		**Viscosity**		**γ–cellulose**		**Copper Numbers**		**Glucose**		**Xylose**	
Percentile	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2	β_1	β_2
0.5	-3.230	-0.950	-0.960	0.240	-30.010	-22.190	-3.160	-1.520	-1.950	-0.690	1.150	0.330	-3.000	-0.770
1.0	-3.050	-0.830	-0.640	0.300	-24.280	-20.860	-3.010	-1.460	-1.860	-0.650	1.220	0.420	-2.670	-0.740
2.5	-2.950	-0.790	-0.560	0.400	-21.800	-17.540	-2.810	-1.420	-1.780	-0.620	1.320	0.480	-2.450	-0.680
5.0	-2.840	-0.750	-0.400	0.490	-19.330	-15.130	-2.590	-1.310	-1.680	-0.590	1.480	0.540	-2.350	-0.650
10.0	-2.750	-0.680	-0.170	0.610	-15.680	-12.430	-2.250	-1.250	-1.580	-0.550	1.670	0.580	-2.150	-0.620
25.0	-2.550	-0.560	0.410	0.830	-8.580	-8.920	-1.800	-1.090	-1.400	-0.490	2.050	0.750	-1.210	-0.550
50.0	-2.220	-0.430	1.080	1.050	-0.340	-4.530	-1.310	-0.920	-1.180	-0.430	2.710	0.920	-0.870	-0.460
75.0	-1.760	-0.300	1.920	1.260	6.690	-0.480	-0.680	-0.700	-0.910	-0.370	3.330	1.080	-0.610	-0.350
90.0	-1.440	-0.200	2.420	1.470	16.300	2.700	0.000	1.190	-0.660	-0.310	3.700	1.180	-0.340	-0.230
95.0	-1.300	-0.120	2.770	1.580	20.340	4.560	0.490	1.350	-0.490	-0.280	3.940	1.260	-0.180	-0.180
97.5	-1.200	-0.077	3.070	1.670	22.720	6.520	0.830	1.460	-0.310	-0.260	4.100	1.340	-0.082	-0.140
99.0	-1.100	-0.019	3.190	1.820	27.510	7.990	1.030	1.610	-0.260	-0.200	4.240	1.440	-0.009	-0.100
99.5	-1.050	0.050	3.360	1.970	33.200	9.810	1.180	1.730	-0.210	-0.180	4.240	1.540	0.036	-0.091
Mean	-2.15	-0.43	1.15	1.04	-0.30	-4.84	-1.22	-0.64	-1.15	-0.43	2.69	0.91	-1.00	-0.44
Standard Dev	0.49	0.19	0.98	0.11	11.77	6.09	0.88	0.82	0.36	0.09	0.77	0.23	0.62	0.14
Correlation	0.170		-0.057		-0.770		0.670		-0.056		-0.300		0.053	
Bandwidth used	0.190	0.017	0.370	0.120	4.440	2.290	0.330	0.310	0.130	0.035	0.250	0.074	0.200	0.046

Chapter 7

Joint Modelling of the Evolution of Pulp Chemical Properties During Chemical Processing

7.1. Introduction

There has been a wide usage of mixed models for the analysis of single outcome variables measured repeatedly over time but there are many practical situations that require extensions of such univariate mixed models to deal with multivariate longitudinal data that arise when a set of different responses on the same unit are measured repeatedly over time. Guo and Karlin (2004) noted that many well-established methods exist for analyzing such data separately, including linear mixed effects models for longitudinal data, and Weibull or semiparametric (Cox) proportional hazards models for survival data. However, the separate use of such methods may be inappropriate when the longitudinal variables are highly correlated. The association between different variables, as they evolve over time, can reveal the mechanism that drives such an evolution. Liu, Daniels and Marcus (2009) studied models of a longitudinal binary variable (smoking cessation) and a longitudinal continuous variable (weight change) and modelled the evolutionary association between the two variables. Joint modelling of such multivariate data is necessary to quantify, firstly, the relationship between evolutions of different responses and, secondly, the evolution of the relationships between different response variables over time. With joint modelling comes the problem that, as the number of response variables goes up, issues of convergence become more troublesome (Rizopoulos, 2012). To help resolve convergence complexity problems, a pairwise fitting approach has been proposed in the literature (Fieuws and Verbeke, 2006, 2007). In this study, the pairwise fitting approach is used to analyse the chemical pulping process and how it affects the evolution of six chemical variables (properties) of different timber genotypes in order to compare their behaviour under chemical pulping and to evaluate how the variables affect each other throughout the process. The choice of the six variables studied was made because inclusion of the seventh variable led to the non-convergence of the modelling procedure. A particularly interesting feature of multivariate data is the

possibility to make prediction and/or inference on one variable conditional on the others.

Joint modelling of multivariate outcomes, particularly bivariate ones, has been used extensively in recent years owing to researchers' desire for more insight into multivariate data using a single statistical model. Studies have been done on joint models of one continuous and one binary response (Faes, Geys and Catalano, 2008; Agresti, 1997; Iddi and Molenberghs, 2012). Joint modeling espouses the broad objective of formalising a framework within which relationships between outcomes of a multivariate nature and the factors affecting them can be scientifically probed (Verbeke and Davidian, 2008). The analysis of such multivariate phenomenon under the framework of joint modelling allows for more accurate calculations of Type I errors when the multivariate response variates are considered together in multiple tests (Gueorguieva, 2001). Gueorguieva (2001) outlined the analysis of multivariate repeated measurements for variables in the exponential family of distributions. Fieuws and Verbeke (2007) pointed out that, when fitting multivariate linear mixed models, there are computational challenges that can be overcome by analysing multivariate outcomes using a pairwise modelling approach. The pairwise modelling approach proceeds by first fitting all possible bivariate mixed models to the response variables of interest, then combining the bivariate mixed models to form an overall multivariate analysis using pseudo-likelihood arguments.

There has been a lot of discussion on the joint modelling of continuous and longitudinal outcomes and time to event variables that are dependent on some fixed and random effects, see for example discussions by Tsiatis and Davidian (2004) and Diggle et al (2008). Tsiatis and Davidian noted that, precise statements of underlying assumptions typically made for these models, has been rare. Their review focussed on the development of joint models and how they offer insight into the structure of the likelihoods for model parameters that clarifies the nature of common assumptions. Sousa (2011) gave a comprehensive and insightful review of developments in the work done on the joint modelling of longitudinal outcomes and time to events variables.

Joint modelling of multivariate responses takes into account the interrelationships between the variables comprising the response vector in order to produce more

accurate inferences. Wu and Carroll (1988) pointed out that in many longitudinal studies, the analysis of the main outcome must be linked to the dropout mechanism as ignoring dropout may cause bias since it is expected that the dropout mechanism carries some information about the main outcome. The same can be said about multiple responses that are measured simultaneously on the same subjects, as is the case with pulp chemical properties considered in this study. This calls for the use of at least bivariate joint modelling. However, in the absence of any association between the measurement variables, the model developed using joint modelling reduces to separate models for the measurements (Henderson et al., 2000).

The data discussed in this study comprises of several chemical properties which are interrelated. Chemical properties of wood genotypes observed when dissolving pulp goes through the six stages of chemical processing are multivariate in nature as several of them are measured at each stage of the process. They are multivariate repeated measurements with two types of correlations, that is, correlations between observations made on the same subject at different stages of chemical pulp processing and correlations between different chemical properties that are jointly measured on the same subject at each stage. The associations between these chemical properties are of interest as they can reveal the overall chemical evolution of dissolving pulp across the processing stages.

7.2. The Univariate Model

To better understand the build up to the joint model, it is essential to start by outlining the model for the single longitudinal continuous response. Assume that there are *N* subjects indicated by *i* =1, 2,…, *N*. In this study there are six time points indicative of the number of stages in the chemical pulping process, thus each subjects has 6 sequential measurements indicated by *j*=1, 2, …, 6. All the seven chemical property variables studied are continuous and the linear mixed model for each of the response variates is specified as

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{b} + \boldsymbol{\varepsilon}_i \qquad (7.1)$$

where $\boldsymbol{X}_i$ is vector of *p*-fixed effects covariates and $\boldsymbol{Z}_i$ is a vector of *q*-random effects covariates. Both $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are observed at time or stage *j* for the i^{th} subject. The vector $\boldsymbol{\beta}$ comprises of *p*-parameters for the fixed effects while the random effects vector is

$\boldsymbol{b}$~N(0,$\boldsymbol{G}$). Conditional on the random effects, the residual terms, $\boldsymbol{\varepsilon}_i$~N(0,$\boldsymbol{R}$), are assumed to be independent of each other at all time points. The expectation of Y_i is such that

$$\mathrm{E}(Y_i) = \mathrm{E}[\mathrm{E}(Y_i|\boldsymbol{b})] = \boldsymbol{X}^T\boldsymbol{\beta} \tag{7.2}$$

which means that the marginal and conditional fixed effects parameters are equal. The marginal model can be expressed as

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i^* \tag{7.3}$$

where the correlated residuals have the distribution $\boldsymbol{\varepsilon}_i^*$~N(0,$\boldsymbol{V}_i$). The covariance matrix $\boldsymbol{V}_i$ takes into account the correlations between the redefined residuals contained in vector $\boldsymbol{\varepsilon}_i^*$. The linear mixed model implies a marginal model with $\boldsymbol{\varepsilon}_i^*$~N(0,$\boldsymbol{V}$) where

$$\boldsymbol{V} = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^{\mathrm{T}} + \boldsymbol{R} \tag{7.4}$$

The marginal model affords greater flexibility on the restrictions on $\boldsymbol{G}$ and $\boldsymbol{R}$. Maximum likelihood and restricted maximum likelihood estimation methods can be used for the estimation of the linear mixed-model parameters (Laird and Ware, 1982).

7.3. Joint Multivariate Models

Suppose a study comprises of *p*-continuous response variables that make up the response vector $\boldsymbol{Y}^{\mathrm{T}} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_p)$, where $\boldsymbol{Y}_r = \left(Y_{r1}, Y_{r2}, \ldots, Y_{rn_r}\right)$ for r=1,…, p and n_r is the number of observations for individual r. Fitting a multivariate mixed model to such a multivariate response will generate a vector of parameters which we shall denote by $\boldsymbol{\Phi}^*$. Fieuws and Verbeke (2007) state that all parameters in the full multivariate model can be identified from all pairwise models, that is, all bivariate models for each pair of outcomes. This means that, the fitting of the full model is replaced by maximum likelihood estimation of each bivariate model separately with full model parameters estimates calculated as means of those obtained in the bivariate models.

If the pairwise modelling approach discussed by Fieuws and Verbeke is used, then the number of bivariate distributions that can be developed from the this response vector is $\binom{p}{2} = p(p-1)/2$, that is, the bivariate distributions for

$$(\boldsymbol{Y}_1, \boldsymbol{Y}_2), (\boldsymbol{Y}_1, \boldsymbol{Y}_3,), \cdots, \left(\boldsymbol{Y}_1, \boldsymbol{Y}_p\right), (\boldsymbol{Y}_2, \boldsymbol{Y}_3), \cdots, \left(\boldsymbol{Y}_2, \boldsymbol{Y}_p\right), \cdots, \left(\boldsymbol{Y}_{p-1}, \boldsymbol{Y}_p\right). \tag{7.5}$$

Consider any two continuous longitudinal outcomes taken on the response vectors $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_s$ and take their observations on the i^{th} subject at the j^{th} period to get Y_{rij} and Y_{sij}, where the two continuous outcomes are assumed to be Gaussian, $i = 1, 2, 3, \ldots, N$ indicates the subjects in the *N*-sample and $j = 1, 2, \ldots, 6$ indicates the time of the repeated. The vector observations ($\boldsymbol{Y}_r$) for subject *i* over the various time points of the repeated measurements are given as $\boldsymbol{Y}_{ri} = [Y_{ri1}, Y_{ri2}, Y_{ri3}, Y_{ri4}, Y_{ri5}, Y_{ri6}]$ and $\boldsymbol{Y}_s$ is defined in a similar way. It is desired to develop an appropriate model for the joint distribution of the two continuous, longitudinal variates $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_s$, that is $f(\boldsymbol{Y}_r, \boldsymbol{Y}_s)$. The likelihood function corresponding to $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_s$ is given by

$$\ell(\boldsymbol{Y}_r, \boldsymbol{Y}_s | \boldsymbol{\Phi}_{sr}) \tag{7.6}$$

where $\boldsymbol{\Phi}_{sr}$is the vector of all parameters of the bivariate mixed model involving $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_s$. Molenberghs and Verbeke (2005) used the so called shared parameter model to estimate the parameter vector ($\boldsymbol{\Phi}_{sr}$) of such bivariate joint distributions.

When all the $p(p-1)/2$ models are fitted, the parameters that result from the models can be presented in a stacked vector of parameters given by

$$\boldsymbol{\Phi}^{\mathrm{T}} = \left[\boldsymbol{\Phi}_{12}^{\mathrm{T}}, \boldsymbol{\Phi}_{13}^{\mathrm{T}}, \cdots, \boldsymbol{\Phi}_{1p}^{\mathrm{T}}, \boldsymbol{\Phi}_{23}^{\mathrm{T}}, \cdots, \boldsymbol{\Phi}_{2p}^{\mathrm{T}}, \cdots, \boldsymbol{\Phi}_{(p-1)p}^{\mathrm{T}}\right] \tag{6.7}$$

where $\boldsymbol{\Phi}$ is obtained by separately maximizing the likelihood functions $\ell(\boldsymbol{Y}_r, \boldsymbol{Y}_s | \boldsymbol{\Phi}_{sr})$, for all $r, s \in \{1,2,\cdots,p\},\ r < s$. While the purpose of the vector of parameters $\boldsymbol{\Phi}$ is to estimate $\boldsymbol{\Phi}^*$**,** it must be pointed out that $\boldsymbol{\Phi}$ has some parameters of each variable repeated several times. It is obvious that some fixed effects parameters from the single outcome will be repeated *p*-1 times. The covariances of random effects between different outcomes only appear once in $\boldsymbol{\Phi}^*$ so they are not affected by this multiplicity**.** This means that the parameter vectors, $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^*$**,** are not equivalent. For those parameters that are estimated several times, their representatives in $\boldsymbol{\Phi}^*$ are found by averaging all the pair specific estimates in $\boldsymbol{\Phi}$. Standard errors of $\widehat{\boldsymbol{\Phi}}$, the estimate of $\boldsymbol{\Phi}$ which translate into estimates of Φ^*, can be obtained from pseudo-likelihood methods.

7.3.1. Fitting the bivariate model

When two responses are fitted pairwise, say Y_r and Y_s, each with the mixed effects model described in equation (7.1), we get the bivariate version of equation (1) given as

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{ri} \\ Y_{si} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{ri}^T\boldsymbol{\beta_r} + \boldsymbol{Z}_{ri}^T\boldsymbol{b_r} + \boldsymbol{\varepsilon}_{ri} \\ \boldsymbol{X}_{si}^T\boldsymbol{\beta_s} + \boldsymbol{Z}_{si}^T\boldsymbol{b_s} + \boldsymbol{\varepsilon}_{si} \end{bmatrix} \tag{7.8}$$

The fixed effects covariate matrices for the two responses are $\boldsymbol{X}_{ri}$and $\boldsymbol{X}_{si}$ with the vectors of fixed effects for the two responses being $\boldsymbol{\beta}_r$ and $\boldsymbol{\beta}_s$. The random effects covariate matrices are $\boldsymbol{Z}_{ri}$and $\boldsymbol{Z}_{si}$ with the corresponding random effects vectors being $\boldsymbol{b_r}$ and $\boldsymbol{b_s}$. In most cases $\boldsymbol{X}_{ri} = \boldsymbol{X}_{si}$ and $\boldsymbol{Z}_{ri} = \boldsymbol{Z}_{si}$since the two response variables depend on the same factors. The vectors that make up equation (7.8) and their distributional assumptions follow from the description of equation (7.1). The task at hand is to develop a likelihood function for the bivariate vector $\boldsymbol{Y}_\mathrm{i}$ which will then be used to obtain maximum likelihood estimates for the parameters of model (7.8). The two responses, Y_r and Y_s, can also be presented as:

$$\boldsymbol{Y}_{ri} = [\boldsymbol{X}_{ri}^\mathrm{T} \quad \boldsymbol{Z}_{ri}^\mathrm{T}]\begin{bmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{b}_r \end{bmatrix} + \boldsymbol{\varepsilon}_{ri} \text{ and } \boldsymbol{Y}_{si} = [\boldsymbol{X}_{si}^\mathrm{T} \quad \boldsymbol{Z}_{si}^\mathrm{T}]\begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{b}_s \end{bmatrix} + \boldsymbol{\varepsilon}_{si}. \tag{7.9}$$

Without loss of generality we can also write $\boldsymbol{X}_r^\mathrm{T} = [\boldsymbol{X}_{ri}^\mathrm{T} \quad \boldsymbol{Z}_{ri}^\mathrm{T}]$ and $\boldsymbol{X}_s^\mathrm{T} = [\boldsymbol{X}_{si}^\mathrm{T} \quad \boldsymbol{Z}_{si}^\mathrm{T}]$ thus the fixed and random effects parameter matrices for the two responses can be presented as

$$\boldsymbol{\Phi}_r = \begin{bmatrix} \boldsymbol{\beta}_r \\ \boldsymbol{b}_r \end{bmatrix} \text{ and } \boldsymbol{\Phi}_s = \begin{bmatrix} \boldsymbol{\beta}_s \\ \boldsymbol{b}_s \end{bmatrix}, \text{ with } \boldsymbol{\Phi}_{rs} = \begin{bmatrix} \boldsymbol{\Phi}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Phi}_s \end{bmatrix}.$$

Using this notation, the formulation of the two outcomes presented in (7.9) can be written as

$$\boldsymbol{Y}_{ri} = \boldsymbol{X}_r^\mathrm{T}\boldsymbol{\Phi}_r + \boldsymbol{\varepsilon}_{ri} \text{ and } \boldsymbol{Y}_{si} = \boldsymbol{X}_s^\mathrm{T}\boldsymbol{\Phi}_s + \boldsymbol{\varepsilon}_{si}$$

and equation (7.9) can now be written as

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{ri} \\ Y_{si} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_r^\mathrm{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_s^\mathrm{T} \end{bmatrix}\begin{bmatrix} \boldsymbol{\Phi}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Phi}_s \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_{ri} \\ \boldsymbol{\varepsilon}_{si} \end{bmatrix} \tag{7.10}$$

and if we let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_r^\mathrm{T} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}_s^\mathrm{T} \end{bmatrix}$, we can then express equation (7.10) in a more compact form as

$$\boldsymbol{Y}_i = \boldsymbol{X}\boldsymbol{\Phi}_{rs} + \boldsymbol{\varepsilon}_{(rs)i}. \tag{7.11}$$

where $\boldsymbol{\varepsilon}_{(rs)i} = \begin{bmatrix} \boldsymbol{\varepsilon}_{ri} \\ \boldsymbol{\varepsilon}_{si} \end{bmatrix}$.

7.3.2. The number of parameter estimates in a multivariate mixed model

The number of parameters to be estimated in the multivariate model grows exponentially with the increase of the number of response variables. To illustrate this estimation problem consider the bivariate case. This study mainly seeks to profile the chemical evolution of some chemical properties of dissolving pulp over six processing stages. The problem in its simplest form takes the fixed effects part of the model as a mean function with the random effects part constituting the slope parameters (because of computational problems the intercept was not estimated in the multivariate problem, instead, the data was intercept corrected).

The linear mixed model for bivariate random variables (p=2) at each time point can be stated as:

$$\begin{aligned} Y_{1it} &= \mu_{1t} + a_{1i} + b_{1i}t + \varepsilon_{1it} \\ Y_{2it} &= \mu_{2t} + a_{2i} + b_{2i}t + \varepsilon_{2it} \end{aligned} \tag{7.12}$$

where μ_{1t} and μ_{2t} are the fixed effects and the random effects are jointly distributed as $\boldsymbol{b}_i \sim MVN(\mathbf{0}, \boldsymbol{G})$, or more specifically

$$\begin{bmatrix} a_{1i} \\ a_{2i} \\ b_{1i} \\ b_{2i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2} & \sigma_{a_1 b_1} & \sigma_{a_1 b_2} \\ & \sigma_{a_2}^2 & \sigma_{a_2 b_1} & \sigma_{a_2 b_2} \\ & & \sigma_{b_1}^2 & \sigma_{b_1 b_2} \\ & & & \sigma_{b_2}^2 \end{bmatrix} \right), \tag{7.13}$$

and the error components which are independent from the random effects are distributed as $\boldsymbol{\varepsilon}_i \sim MVN(\mathbf{0}, \boldsymbol{R})$ or

$$\begin{bmatrix} \varepsilon_{1i} \\ \varepsilon_{1i} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{\varepsilon_1}^2 & \sigma_{\varepsilon_1 \varepsilon_2} \\ & \sigma_{\varepsilon_2}^2 \end{bmatrix} \right). \tag{7.14}$$

The two error components are not necessarily independent as observations for the two response variables on the same subject can be correlated at any time point, thus $\sigma_{\varepsilon_1 \varepsilon_2}$ is not necessarily equal to zero. The correlations between the two intercept and two slope parameters for the two variables are given respectively as

$$r_{a_1 a_2} = Corr(a_{1i}, a_{2i}) = \frac{\sigma_{a_1 a_2}}{\sqrt{\sigma_{a_1}^2 \times \sigma_{a_2}^2}}$$

and

$$r_{b_1 b_2} = Corr(b_{1i}, b_{2i}) = \frac{\sigma_{b_1 b_2}}{\sqrt{\sigma_{b_1}^2 \times \sigma_{b_2}^2}}.$$

The marginal correlation between the two response variables Y_1 and Y_2 is given as

$$r_{Y_1Y_2} = Corr(Y_1, Y_2) = \frac{Cova(\mu_{1t} + a_{1i} + b_{1i}t + \varepsilon_{1it}, \mu_{2t} + a_{2i} + b_{2i}t + \varepsilon_{2it})}{\sqrt{var(\mu_{1t} + a_{1i} + b_{1i}t + \varepsilon_{1it}) \times var(\mu_{2t} + a_{2i} + b_{2i}t + \varepsilon_{2it})}}$$

$$= \frac{\sigma_{a_1a_2} + t\sigma_{a_1b_2} + t\sigma_{a_2b_1} + t^2\sigma_{b_1b_2} + \sigma_{\varepsilon_1\varepsilon_2}}{\sqrt{\left(\sigma_{a_1}^2 + t^2\sigma_{b_1}^2 + 2t\sigma_{a_1b_1} + \sigma_{\varepsilon_1}^2\right) \times \left(\sigma_{a_2}^2 + t^2\sigma_{b_2}^2 + 2t\sigma_{a_2b_2} + \sigma_{\varepsilon_2}^2\right)}}$$

For a problem with *p*-response variables there will be 2*p* slope and intercept parameters to be estimated. The matrix ***G*** will consist of $\binom{2p}{2} + 2p$ covariance parameters and ***R*** will consist of $\binom{p}{2} + p$ covariance parameters to be estimated. The chemical pulping problem has *p*=7, thus there will be $\binom{2p}{2} + 2p + \binom{p}{2} + p = \binom{14}{2} + 14 + \binom{7}{2} + 7$=133 parameters to be estimated. Compounded with the fact that there is need to estimate slope and intercept parameters for each of the seven genotypes under study, there will be 931 parameters to be estimated. The need to reduce the number of parameters to be estimated requires that we we fit models with only the slope and the intercept set at zero. This can be achieved by correcting the data for the intercept so that only the slope parameter, which is a measure of the changes in the chemical properties over time is estimated. An attempt to fit a model with an intercept and a slope parameter could not converge hence the need to intercept correct the data.

7.3.3. Fitting the bivariate model using conditioning

The joint distribution for (Y_r, Y_s) can be specified by factorising its joint density as a product of a marginal and a conditional density as shown in equations (7.15) and (7.16) below;

$$f(y_r, y_s) = f(y_r|y_s)f(y_s) \quad (7.15)$$

$$= f(y_s|y_r)f(y_r) \quad (7.16)$$

In the determination of $f(y_r, y_s)$, the density function $f(y_s)$ can be obtained directly if we assume the Gaussian distribution. It is the conditional density, $f(y_r|y_s)$, that

requires a careful consideration of the association between Y_r and Y_s where Y_s would be playing the role of a time varying covariate with different values of Y_s giving different results and conclusions (Verbeke and Davidian, 2008). In (7.15) marginal inferences about Y_s are direct but inferences about Y_r would require additional calculations. Marginal expectations of Y_r , for example, would require the computation of

$$E(\boldsymbol{Y}_r) = E[E(Y_r|Y_s)] = \int \left\{ \int \boldsymbol{y}_r f(\boldsymbol{y}_r|\boldsymbol{y}_s)\, d\boldsymbol{y}_r \right\} f(\boldsymbol{y}_s)\, d\boldsymbol{y}_s.$$

One might avoid having to calculate this not so straightforward integral by fitting both models (7.15) and (7.16) and obtaining the marginal distribution of one variate at a time. Verbeke and Davidson (2008) argue that when Y_r and Y_s are highly correlated, as with the variables in this study, and are thought to be affected by a common treatment effect, then conditioning on one of the two variables will diminish the effect of the treatment factor on the other response variable.

7.3.4. Fitting the bivariate model using shared-parameter models

Suppose the two variates Y_r and Y_s have a common random effects vector $\boldsymbol{b}$ and are independent, conditionally on $\boldsymbol{b}$. The joint density of (Y_r, Y_s) can then be found by

$$f(y_r, y_s) = \int f(y_r, y_s|\boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b} = \int f(y_r|\boldsymbol{b}) f(y_s|\boldsymbol{b}) f(\boldsymbol{b}) d\boldsymbol{b}. \qquad (7.17)$$

where $f(\boldsymbol{b})$ denotes the density function of the random effects which is usually assumed to be the normal density function. Equation (7.17) is what is called the shared-parameter model as the response variables depend on a common random effects vector. The joint dependency of Y_r and Y_s on $\boldsymbol{b}$ induces some correlation between the two variables but, conditional on $\boldsymbol{b}$, the two variates are considered independent. In this study the variables Y_r and Y_s are both assumed to have normal densities and it follows that their conditional densities, that is $f(y_r|\boldsymbol{b})$ and $f(y_s|\boldsymbol{b})$ are also normal. It must be noted that the variables Y_r and Y_s, under the framework of repeated measurements, are actually multivariate as each of the two vectors comprises of the repeated measurements for each subject (six repeated measurements in this study). The likelihood function for the two random variables will be based on their joint density function outlined in (7.17) above. Other approaches that have been suggested for the bivariate model include the random-effects models formulation described by Verbeke and Davidson (2008).

7.3.5. Fitting the full joint multivariate model using pairwise fitting

For a dataset with *p*-response variables, fitting the $\binom{p}{2} = p(p-1)/2$ bivariate joint models is equivalent to maximizing the pseudo-likelihood function which is the product of all pairwise pseudo-likelihood functions given by

$$\begin{aligned} pl(\boldsymbol{\Phi}) &= l(Y_1, Y_2|\boldsymbol{\Phi}_{12}) \times l(Y_1, Y_3|\boldsymbol{\Phi}_{13}) \times \ldots \times l\left(Y_{(p-1)}, Y_p|\boldsymbol{\Phi}_{(p-1)p}\right) \\ &= \prod_{r=1, s=r+1}^{p} l(y_r, y_s|\boldsymbol{\Phi}_{rs}) \end{aligned} \tag{7.18}$$

where $pl(.)$ denotes a pseudo-likelihood function, $l(.)$ denotes a likelihood function, $\boldsymbol{\Phi}$ is the parameter vector of all p-response variables and $\boldsymbol{\Phi_{rs}}$ is the parameter vector for randoms variables Y_r and Y_s. The maximization of $pl(\boldsymbol{\Phi})$ is done through its log to obtain the pseudo-log-likelihood

$$p\ell(\boldsymbol{\Phi}) = \log[pl(\boldsymbol{\Phi})]$$

$$p\ell(\boldsymbol{\Phi}) = p\ell\left(y_1, y_2, \cdots, y_p|\boldsymbol{\Phi}\right) = \sum_{r=1, s=r+1}^{p} pl(y_r, y_s|\boldsymbol{\Phi}_{rs}) \tag{7.19}$$

The pseudo-likelihood function (7.18) makes the assumption that all pairwise parameter vectors $\boldsymbol{\Phi}_{rs}$ are distinct from each other which is actually not the case as parameter estimates for Y_1 made jointly with Y_2 might differ slightly with those of Y_1 made jointly with Y_3 and so on. The way around this lack of uniqueness of the pairwise-estimated parameters is to find some form of average of all parameter estimates for a particular variable. The parameter estimates based on this pseudo-likelihood function are called pseudo-likelihood estimates with certain asymptotic statistical properties (Fieuws and Verbeke, 2007). According to Fieuws and Verbeke, the parameter estimate $\widehat{\boldsymbol{\Phi}}$ asymptotically satisfies

$$\sqrt{N}\left(\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}\right) \sim N(0, \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}) \tag{7.20}$$

where $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$ is a "sandwich type" robust variance estimator derived from the components of $\boldsymbol{\Phi}$. A detailed discussion of $\boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}$ is given by Liang and Zeger (1986). The matrices $\boldsymbol{J}$ and $\boldsymbol{K}$ are based on the partial derivatives of the parameter vector $\boldsymbol{\Phi}$ as follows:

$$\boldsymbol{J} = \begin{bmatrix} J_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \ddots \\ \mathbf{0} & \cdots & J_p \end{bmatrix}, \quad \boldsymbol{K} = \begin{bmatrix} K_{11} & \cdots & K_{1p} \\ \vdots & \ddots & \ddots \\ K_{p1} & \cdots & K_{pp} \end{bmatrix}$$

where

$$\boldsymbol{J}_r = -\frac{1}{N}\sum_{i=1}^{N} E\left[\frac{\partial^2 \ell(y_{ri}|\boldsymbol{\Phi}_r)}{\partial \boldsymbol{\Phi}_r \partial \boldsymbol{\Phi}_r^{\mathrm{T}}}\right], \qquad K_{rs} = \frac{1}{N}\sum_{i=1}^{N} E\left[\frac{\partial \ell(Y_{ri}|\boldsymbol{\Phi}_r)}{\partial \boldsymbol{\Phi}_r}\ \frac{\partial \ell(Y_{si}|\boldsymbol{\Phi}_s)}{\partial \boldsymbol{\Phi}_s^{\mathrm{T}}}\right].$$

Since $\boldsymbol{\Phi}^*$ is estimated from $\boldsymbol{\Phi}$ by some form of averaging of the parameters in $\boldsymbol{\Phi}$, of which some appear more than once**,** there is need to obtain some weight matrix $\boldsymbol{A}$ such that $\widehat{\boldsymbol{\Phi}}^* = \boldsymbol{A}\widehat{\boldsymbol{\Phi}}$. Using the fact that $\widehat{\boldsymbol{\Phi}}^* = \boldsymbol{A}\widehat{\boldsymbol{\Phi}}$ and following from equation (7.20), we have

$$\sqrt{N}\left(\widehat{\boldsymbol{\Phi}}^* - \boldsymbol{\Phi}^*\right) = \sqrt{N}\boldsymbol{A}\left(\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}\right) \sim N(0, \boldsymbol{A}\ \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}\boldsymbol{A}^{\mathrm{T}}). \tag{7.21}$$

According to Arnold and Strauss (1991), the principal idea in the use of pseudo-likelihood estimators is to replace a computationally challenging joint density by a simpler function.

The matrices $\boldsymbol{J}$ and $\boldsymbol{K}$ can be estimated as shown in the steps that follow. Consider equation (7.4) above and let the estimate of $\boldsymbol{V}$ for the i^{th} subject and rth variable be $\widehat{\boldsymbol{V}}_{i,r}$ where

$$\widehat{\boldsymbol{V}}_{ir} = \boldsymbol{Z}_{i,r}\widehat{\boldsymbol{G}}_r\boldsymbol{Z}_{i,r}^{\mathrm{T}} + \widehat{\boldsymbol{R}}_{i,r}. \tag{7.22}$$

To estimate the matrices $\boldsymbol{J}$ and $\boldsymbol{K}$ we need first to define the following matrices:

$$\hat{\boldsymbol{J}}_r = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_{i,r}^{\mathrm{T}}\widehat{\boldsymbol{V}}_{i,r}^{-1}\boldsymbol{X}_{i,r} \text{ and } \widehat{\boldsymbol{K}}_{rs} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_{i,r}^{\mathrm{T}}\widehat{\boldsymbol{V}}_{i,r}^{-1}\boldsymbol{e}_{i,r}(\boldsymbol{X}_{i,s}^{\mathrm{T}}\widehat{\boldsymbol{V}}_{i,r}^{-1}\boldsymbol{e}_{i,s})^{\mathrm{T}}$$

The vectors $\boldsymbol{X}_{i,r}$ and $\boldsymbol{Z}_{i,r}$ are, respectively, the fixed and random effects for the i^{th} subject on the r^{th} response variable, $\boldsymbol{e}_{i,s}$ is the corresponding error component and N is the number of subjects. The matrix $\hat{\boldsymbol{J}}$ and $\widehat{\boldsymbol{K}}$ can then be estimated as

$$\hat{\boldsymbol{J}} = \begin{bmatrix} \hat{\boldsymbol{J}}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\boldsymbol{J}}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \hat{\boldsymbol{J}}_p \end{bmatrix} \text{ and } \widehat{\boldsymbol{K}} = \begin{bmatrix} \widehat{\boldsymbol{K}}_{11} & \widehat{\boldsymbol{K}}_{12} & \cdots & \widehat{\boldsymbol{K}}_{1p} \\ \widehat{\boldsymbol{K}}_{21} & \widehat{\boldsymbol{K}}_{22} & \cdots & \widehat{\boldsymbol{K}}_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ \widehat{\boldsymbol{K}}_{p1} & \cdots & \widehat{\boldsymbol{K}}_{p(p-1)} & \widehat{\boldsymbol{K}}_{pp} \end{bmatrix}$$

7.4. Fitting the Joint Multivariate Model to the Pulp data

Changes, over time, in any of the seven chemical properties studied, is indicative of how the chemical processes affect the raw pulp. If such changes can be quantified then they will form a basis on which to compare the behavior of pulps from different genotypes. These changes, over time, are represented by the slope parameters obtained by fitting a joint multivariate random coefficient model to the seven chemical properties (for the seven genotypes) under study. Because of computational

complexities the joint multivariate model will be fitted using the pairwise fitting approach described above. It is also important to note that it was not possible to fit a more complex model that the simple linear regression model with the data corrected so that the intercept is set at zero.

7.4.1. Intercept corrected data

Initial fitting of the pairwise bivariate models showed that it was not possible to fit both the intercept and the slope parameters for the various genotypes because of the number of parameters involved in the model. The SAS PROC MIXED procedure would estimate the intercept and set the slope parameter estimates to zero and when forced to fit the slope parameters, it would set the intercept parameter estimates to zero. A way to get around this problem was to first fit univariate regression models to the chemical properties in order to estimate their intercepts. The intercept estimates were then subtracted from the original data thereby obtaining some form of intercept corrected data $\left(x - \hat{\beta}_0\right)$ in a way that is similar to that of obtaining mean corrected data $(x - \bar{x})$. This allowed for the setting of the fixed effects part of the intercept values to zero and hence allow for the computation of only the fixed effects part of the slope parameters in the random coefficients model. The random effects components of the random coefficients model can still be estimated for the intercept and the slope parameters with fixed effects part of the intercept set to zero. To illustrate how the intercept corrected model works, consider the case of viscosity data for the seven genotypes. Figure 7.1 below, shows the random coefficients models for Viscosity for the seven chemical properties before the intercepts were set to zero (using intercept corrected data). The fixed parts of the intercept estimates obtained using the univariate random coefficient models are presented in Table 7.1 below. As an example, the estimated intercept for the genotype GUA is 70.895 and to effect intercept correction, all viscosity values of GUA will have 70.895 subtracted thereby making the expected value of the intercept zero. This will justify the setting to zero of the intercept value for GUA in the fitting of bivariate joint models which are required in the pairwise fitting approach.

The same process of setting the fixed parts for the intercept to zero is followed for the other six chemical properties that are considered in the pairwise fitting process.

Table 7.1. Univariate intercept estimates

Genotype	Intercept Estimates by Chemical Property						
	Viscosity	Lignin	α-cellulose	γ-cellulose	Copper number	Glucose	Xylose
E.dunnii	62.165	4.036	89.865	8.131	3.231	89.629	5.005
E.grandis	33.340	2.905	91.128	7.274	2.847	92.197	3.560
E.smithii	52.212	4.249	91.136	8.150	2.954	90.317	5.085
E.nitens	46.078	2.123	90.368	8.046	2.621	89.989	5.657
GCG	63.853	4.615	91.153	7.480	3.050	90.113	3.873
GUA	78.821	3.501	90.344	8.367	2.910	90.020	4.662
GUW	63.337	2.745	91.317	6.754	2.549	92.834	3.189

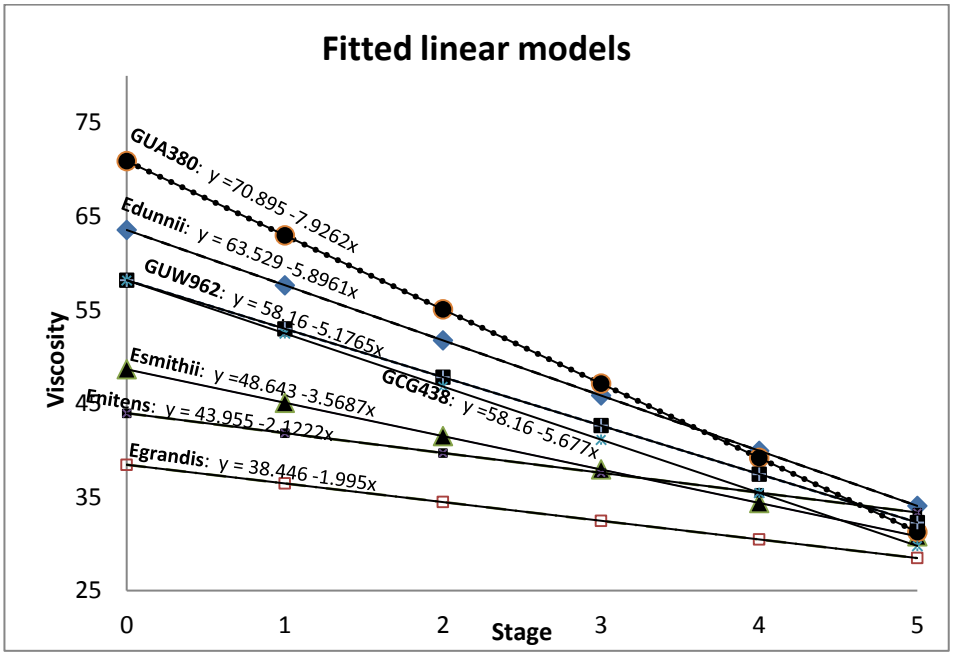


Figure 7.1. Random coefficients regression models for viscosity of the seven genotypes.

7.4.2. Pairwise fitting of the 21 possible pairs of variables

In trying to fit the multivariate model for the seven chemical properties Xylose could not be included since the pairwise fitting procedure could not converge for some pairs of variables that involved Xylose. Working with the other remaining six variables, the pairwise fitting of the 15 possible pairs of the 6 chemical property variables is presented in this section.

The pairwise parameter estimates have subscripts as listed in Table 7.2 below, from which, for example, the pairwise slope parameter estimates for Viscosity and Lignin would be $\widehat{\boldsymbol{\theta}}_{12}$.

Table 7.2. Variable codes

Variable	Variable number	Variable	Variable number
Viscosity	1	γ-cellulose	4
Lignin	2	Copper number	5
α-cellulose	3	Glucose	6

The vectors for the pairwise slope parameter estimates are presented as

$$\widehat{\boldsymbol{\theta}}_{ij} = [\text{Variable } i \text{ slope parameter}, \quad \text{Variable } j \text{ slope parameter}]$$

and these pairs of parameters are estimated for all possible pairs derived from the seven genotypes. No intercepts are estimated as these are set to zero by making use of intercept corrected data as discussed in Section 7.4.1. In total 14 parameters are estimated for each pair of variables in seven pairs as shown

$$\widehat{\boldsymbol{\theta}}_{ij} = \begin{matrix} \textbf{Variable } \boldsymbol{i} \textbf{ slope parameter} & \textbf{Variable } \boldsymbol{j} \textbf{ slope parameter} \\ [\text{EDunnii} & \text{EDunnii}] \\ [\text{EGrandis} & \text{EGrandis}] \\ [\text{ESmithii} & \text{ESmithii}] \\ [\text{Enitens} & \text{Enitens}] \\ [\text{GCG} & \text{GCG}] \\ [\text{GUA} & \text{GUA}] \\ [\text{GUW} & \text{GUW}] \end{matrix}$$

The matrix $\widehat{\boldsymbol{\theta}}_{ij}$ is made up of the fixed parts of the slope parameters of the two variables (i and j) whose parameters are being estimated pairwise. For computational convenience the seven genotypes are assumed to have identical covariance matrices for the random effects. It was not possible to compute individual covariance matrices for each genotype.

$$\widehat{\mathbf{G}}_{ij} = \boldsymbol{Cov}\begin{bmatrix} \boldsymbol{Rs\ of\ Y}_i \\ \boldsymbol{Rs\ of\ Y}_j \end{bmatrix} = \begin{bmatrix} \sigma^2_{Slope\ i} & \sigma_{Slope\ i,slope\ j} \\ & \sigma^2_{Slope\ j} \end{bmatrix}$$

The matrix $\widehat{\mathbf{G}}_{ij}$ is the covariance matrix of the random effects part of the slope parameters of the two random variables concerned. The term "R_s of Y_i " refers to the random effect part of the slope parameter of variable Y_i. The random error component of the bivariate model has the covariance matrix estimate $\widehat{\mathbf{R}}_{ij}$ given by ;

$$\widehat{\boldsymbol{R}}_{ij} = \begin{bmatrix} \sigma^2_{Variable\ i} & \sigma_{Variable\ i,Variable\ j} \\ & \sigma^2_{Variable\ j} \end{bmatrix}$$

were $\sigma^2_{Variable\ i}$ is the variance of variable i and $\sigma_{Variable\ i,Variable\ j}$ is the covariance of variables i and j.

7.4.3. Estimation of model parameters using the pairwise method

The parameter estimation in this section was mainly done using an adaptation of the SAS code developed by Kundu (2011). For the six variables that were modelled using pairwise fitting, fifteen pairwise datasets were created (${}_6C_2$). Suppose we wish to fit a bivariate model to two variables Y_i and Y_j. We would combine these two variables into one variable appropriately named Y_{ij}. It is important to keep track of which values in Y_{ij} belong to variable Yi and variable Yj and this is done by creating index variables using the time (stage) variable in the data.

The combination of variables Y_i and Y_j into variable Y_{ij} proceeds by creating two dummy variables for time as follows:

Set $Y_{ij}= Y_i$, variable numbe $r= i$;

$Time_i$ = 1;

$Time_j$ = 0;

Output;

Set $Y_{ij}= Y_j$, variable numbe $r= j$;

$Time_i$ = 0;

$Time_j$ = 1;

Output;

Keep all variables;

The data set that that results from the above procedure would have all the variables in the original dataset and the new dummy variables $Time_i$ and $Time_j$ which would indicate to which original variable a row in the combined data belongs to. If $Time_i$ =1 then the concerned row belongs to variable *i* and the same applies to $Time_i$, otherwise if $Time_i$=0 then the row belongs to the other variable in the pairwise fitting. Once the dataset for the pair is created, the parameter estimates can obtained using Proc Mixed in SAS.

7.4.4. Pairwise slope parameter estimates

The slope parameters estimates of the six variables (except xylose) for the seven genotypes obtained using the pairwise fitting method are presented in Tables 7.3(a), 7.3(b) and 7.3(c) below. There are fifteen pairs of estimates for which averages will be calculated a the final model estimates.

Table 7.3(a). Pairwise parameter estimates

Variables	Slope parameter estimates by genotype		Random Effects Parameter estimates	
	Genotype	$\widehat{\boldsymbol{\theta}}_{ij}^{T}$	$\widehat{\boldsymbol{G}}_{ij}$	$\widehat{\boldsymbol{R}}_{ij}$
Viscosity (i=1) and Lignin (*j*=2)	EDunnii	$[-14.7300 \quad -2.2304]$	$\begin{bmatrix} 0.6283 & 0.0001 \\ & 0.6264 \end{bmatrix}$	$\begin{bmatrix} 131.06 & 9.0200 \\ & 1.9178 \end{bmatrix}$
	EGrandis	$[-0.2163 \quad -1.8193]$		
	ESmithii	$[-12.4908 \quad -2.6885]$		
	ENitens	$[-7.4280 \quad -1.3104]$		
	GCG	$[-19.8691 \quad -2.9209]$		
	GUA	$[-27.7414 \quad -2.2713]$		
	GUW	$[-18.1180 \quad -1.7214]$		
Viscosity (*i*=1) and γ-cellulose (*j*=3)	EDunnii	$[-14.7300 \quad -2.8093]$	$\begin{bmatrix} 0.6271 & 0.0001 \\ & 0.6264 \end{bmatrix}$	$\begin{bmatrix} 131.07 & 10.8395 \\ & 3.0621 \end{bmatrix}$
	EGrandis	$[-0.2163 \quad -2.9566]$		
	ESmithii	$[-12.4908 \quad -2.9132]$		
	ENitens	$[-7.4280 \quad -3.2886]$		
	GCG	$[-19.8691 \quad -2.8587]$		
	GUA	$[-27.7414 \quad -3.2988]$		
	GUW	$[-18.1180 \quad -2.5204]$		
Viscosity (*i*=1) and α-cellulose (*j*=4)	EDunnii	$[-14.7300 \quad 3.1036]$	$\begin{bmatrix} 0.6274 & -0.0001 \\ & 0.6264 \end{bmatrix}$	$\begin{bmatrix} 131.06 & -10.3872 \\ & 3.6538 \end{bmatrix}$
	EGrandis	$[-0.2163 \quad 3.1652]$		
	ESmithii	$[-12.4908 \quad 2.4100]$		
	ENitens	$[-7.4280 \quad 3.5200]$		
	GCG	$[-19.8691 \quad 2.8387]$		
	GUA	$[-27.7414 \quad 3.2634]$		
	GUW	$[-18.1180 \quad 2.8136]$		
Viscosity (*i*=1) and Copper Number (*j*=5)	EDunnii	$[-14.7300 \quad -1.7644]$	$\begin{bmatrix} 0.6273 & 0.00003 \\ & 0.6264 \end{bmatrix}$	$\begin{bmatrix} 131.07 & 5.7503 \\ & 0.9340 \end{bmatrix}$
	EGrandis	$[-0.2163 \quad -1.6313]$		
	ESmithii	$[-12.4908 \quad -1.6325]$		
	ENitens	$[-7.4280 \quad -1.4055]$		
	GCG	$[-19.8691 \quad -1.7749]$		
	GUA	$[-27.7414 \quad -1.6734]$		
	GUW	$[-18.1180 \quad -1.4051]$		
Viscosity (*i*=1) and Glucose (*j*=6)	EDunnii	$[-14.7300 \quad 4.0108]$	$\begin{bmatrix} 0.6272 & 0.0083 \\ & 2.5053 \end{bmatrix}$	$\begin{bmatrix} 131.02 & -12.6581 \\ & 4.4274 \end{bmatrix}$
	EGrandis	$[-0.2163 \quad 2.9768]$		
	ESmithii	$[-12.4908 \quad 3.4879]$		
	ENitens	$[-7.4280 \quad 4.0638]$		
	GCG	$[-19.8691 \quad 3.5983]$		
	GUA	$[-27.7414 \quad 3.8891]$		
	GUW	$[-18.1180 \quad 2.5968]$		

Table 7.3(b). Pairwise parameter estimates (Continued)

Variables	Fixed Effects Parameter estimates		Random Effects Parameter estimates	
	Genotype	$\widehat{\boldsymbol{\theta}}_{ij}^{\mathrm{T}}$	$\widehat{\boldsymbol{G}}_{ij}$	$\widehat{\boldsymbol{R}}_{ij}$
Lignin (*i*=2) and γ-cellulose (*j*=3)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.2304 \quad -2.8093]$ $[-1.8193 \quad -2.9566]$ $[-2.6885 \quad -2.9132]$ $[-1.3104 \quad -3.2886]$ $[-2.9209 \quad -2.8587]$ $[-2.2713 \quad -3.2988]$ $[-1.7214 \quad -2.5204]$	$\begin{bmatrix} 0.6263 & 0.0000 \\ & 0.6263 \end{bmatrix}$	$\begin{bmatrix} 1.9180 & 1.9676 \\ & 3.0622 \end{bmatrix}$
Lignin (*i*=2) and α-cellulose (*j*=4)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.2304 \quad 3.1036]$ $[-1.8193 \quad 3.1652]$ $[-2.6885 \quad 2.4100]$ $[-1.3104 \quad 3.5200]$ $[-2.9209 \quad 2.8387]$ $[-2.2713 \quad 3.2634]$ $[-1.7214 \quad 2.8136]$	$\begin{bmatrix} 0.6263 & 0.0000 \\ & 0.6263 \end{bmatrix}$	$\begin{bmatrix} 1.9180 & -1.8854 \\ & 3.6540 \end{bmatrix}$
Lignin (*i*=2) and Copper Number (*j*=5)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.2304 \quad -1.7644]$ $[-1.8193 \quad -1.6313]$ $[-2.6885 \quad -1.6325]$ $[-1.3104 \quad -1.4055]$ $[-2.9209 \quad -1.7749]$ $[-2.2713 \quad -1.6734]$ $[-1.7214 \quad -1.4051]$	$\begin{bmatrix} 0.6263 & 0.0000 \\ & 0.6263 \end{bmatrix}$	$\begin{bmatrix} 1.9180 & 1.2172 \\ & 0.9340 \end{bmatrix}$
Lignin (*i*=2) and Glucose (*j*=6)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.2304 \quad 4.0108]$ $[-1.8193 \quad 2.9768]$ $[-2.6885 \quad 3.3248]$ $[-1.3104 \quad 4.0638]$ $[-2.9209 \quad 3.5983]$ $[-2.2713 \quad 3.8891]$ $[-1.7214 \quad 2.5968]$	$\begin{bmatrix} 0.6263 & 0.0028 \\ & 2.4933 \end{bmatrix}$	$\begin{bmatrix} 1.9180 & -2.6945 \\ & 4.5336 \end{bmatrix}$
γ-cellulose (*i*=3) and α-cellulose (*j*=4)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.8093 \quad 3.1036]$ $[-2.9566 \quad 3.1652]$ $[-2.9132 \quad 2.4100]$ $[-3.2886 \quad 3.5200]$ $[-2.8587 \quad 2.8387]$ $[-3.2988 \quad 3.2634]$ $[-2.5204 \quad 2.8136]$	$\begin{bmatrix} 0.6264 & 0.0000 \\ & 0.6264 \end{bmatrix}$	$\begin{bmatrix} 3.0620 & -3.1214 \\ & 3.6538 \end{bmatrix}$

Table 7.3(c). Pairwise parameter estimates (Continued)

Variables	Fixed Effects Parameter estimates		Random Effects Parameter estimates	
	Genotype	$\widehat{\boldsymbol{\theta}}_{ij}^{\mathrm{T}}$	$\widehat{\boldsymbol{G}}_{ij}$	$\widehat{\boldsymbol{R}}_{ij}$
γ-cellulose (*i*=3) and Copper Number (*j*=5)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.8093 \quad -1.7644]$ $[-2.9566 \quad -1.6313]$ $[-2.9132 \quad -1.6325]$ $[-3.2886 \quad -1.4055]$ $[-2.8587 \quad -1.7749]$ $[-3.2988 \quad -1.6734]$ $[-2.5204 \quad -1.4051]$	$\begin{bmatrix} 0.6263 & 0.0000 \\ & 0.6263 \end{bmatrix}$	$\begin{bmatrix} 3.0622 & 1.6174 \\ & 0.9340 \end{bmatrix}$
γ-cellulose (*i*=3) and Glucose (*j*=6)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-2.8093 \quad 4.0108]$ $[-2.9566 \quad 2.9768]$ $[-2.9132 \quad 3.4055]$ $[-3.2886 \quad 4.0638]$ $[-2.8587 \quad 3.5983]$ $[-3.2988 \quad 3.8891]$ $[-2.5204 \quad 2.5968]$	$\begin{bmatrix} 0.6264 & -0.0053 \\ & 2.5117 \end{bmatrix}$	$\begin{bmatrix} 3.0622 & -3.3881 \\ & 4.4173 \end{bmatrix}$
α-cellulose (*i*=4) and Copper Number (*j*=5)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[3.1036 \quad -1.7644]$ $[3.1652 \quad -1.6313]$ $[2.4100 \quad -1.6325]$ $[3.5200 \quad -1.4055]$ $[2.8387 \quad -1.7749]$ $[3.2634 \quad -1.6734]$ $[2.8136 \quad -1.4051]$	$\begin{bmatrix} 0.6263 & 0.0000 \\ & 0.6263 \end{bmatrix}$	$\begin{bmatrix} 3.6540 & -1.6043 \\ & 0.9340 \end{bmatrix}$
α-cellulose (*i*=4) and Glucose (*j*=6)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[3.1036 \quad 4.0108]$ $[3.1652 \quad 2.9768]$ $[2.4100 \quad 3.6128]$ $[3.5200 \quad 4.0638]$ $[2.8387 \quad 3.5983]$ $[3.2634 \quad 3.8891]$ $[2.8136 \quad 2.5968]$	$\begin{bmatrix} 0.6265 & -0.0081 \\ & 2.4965 \end{bmatrix}$	$\begin{bmatrix} 3.6540 & 3.8341 \\ & 4.6570 \end{bmatrix}$
Copper number (*i*=5) and Glucose (*j*=6)	EDunnii EGrandis ESmithii ENitens GCG GUA GUW	$[-1.7644 \quad 4.0108]$ $[-1.6313 \quad 2.9768]$ $[-1.6325 \quad 3.6128]$ $[-1.4055 \quad 4.0638]$ $[-1.7749 \quad 3.5983]$ $[-1.6734 \quad 3.8891]$ $[-1.4051 \quad 2.5968]$	$\begin{bmatrix} 0.6263 & 0.0053 \\ & 2.5054 \end{bmatrix}$	$\begin{bmatrix} 0.9340 & -1.9626 \\ & 4.5241 \end{bmatrix}$

The slope parameter of each variable for each of the seven genotypes is estimated 5 times hence a final estimate would be an average of the five pairwise estimates. The average slope parameter estimates are presented in Table 7.4 below. The results in Table 7.4 show mean slope parameters (rates of change of the of the six chemical properties) and how they compare to each other (rank). The values are ranked in order

of magnitude or rate of change. The steepest rate of change is given the rank of 1 while the least steep rate of change is given the rank of 7. After considering all the rankings, the average rank of each genotype over the six variables is calculated and this can be used to see which genotype, on average, has the slowest or highest rate of change (or reactivity) due to the various chemical pulping processes. Genotypes with closer ranks are more likely to be mixed during processing than those with ranks that are furthest apart.

If all the genotypes behaved exactly the same during processing then they would have similar average ranks. If some of the genotypes are less reactive to chemical processing than others then there would be marked differences in their mean ranks. The rank total for each variable is $\frac{n(n+1)}{2} = \frac{7(7+1)}{2} = 28$, which means that each genotype should have an average rank of 4 if the genotypes had equal rates of changes during chemical processing. The average ranks can be crudely used to group the seven genotypes according to their reactivity to chemical pulping.

Table 7.4. Mean slope parameters for the seven genotypes

Genotype	Viscosity (Y_1)		Lignin (Y_2)		γ-Celluose (Y_3)		α-Cellulose (Y_4)		Copper Number (Y_5)		Glucose (Y_6)		Average Rank
	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	Slope	Rank	
EDunnii	-14.7300	4	-2.2304	4	-2.8093	6	3.1036	4	-1.7644	2	4.0108	2	3.667
EGrandis	-0.2163	7	-1.8193	5	-2.9566	3	3.1652	3	-1.6313	5	2.9768	6	4.833
ESmithii	-12.4908	5	-2.6885	2	-2.9132	4	2.4100	7	-1.6325	4	3.4888	5	4.500
ENitens	-7.4280	6	-1.3104	7	-3.2886	2	3.5200	1	-1.4055	6	4.0638	1	3.833
GCG	-19.8691	2	-2.9209	1	-2.8587	5	2.8387	5	-1.7749	1	3.5983	4	3.000
GUA	-27.7414	1	-2.2713	3	-3.2988	1	3.2634	2	-1.6734	3	3.8891	3	2.167
GUW	-18.1180	3	-1.7214	6	-2.5204	7	2.8136	6	-1.4051	7	2.5968	7	6.000
Average	-14.3705		-2.1375		-2.9494		3.0164		-1.6124		3.5178		
Friedman's test χ^2 = 9.000, *df*=6, *p*-value=0.1736													

It is clear that GUA (average rank=2.167) is the least reactive genotype followed by GCG. In order of reactivity from the most reactive to the least reactive the genotypes can be arranged as 1.GUW, 2.EGrandis, 3.ESmitthii, 4.ENitens, 5.EDunnii, 6.GCG and 7.GUA. When mixing these genotypes for processing it is proposed that consideration be made to genotypes which are not too far apart in order of reactivity to chemical processing. It would be appropriate for example to mix GUW with EGrandis than GUW and GUA. It is important, however, to note that the slope

parameters obtained using the random coefficient models differ a great deal from the ones obtained using joint modelling. This might be because the joint model, which could not converge when the intercept was also fitted, is less accurate in its parameter estimates but all the same useful in the understanding of the intercorrelations between evolutions of the response variables.

In order to test is there are significant differences in the ranking of the genotypes across the six chemical properties listed in Table 7.4 a Friedman's test for a non-parametric randomized block design was carried out. In this case the chemical properties were considered as blocking factor levels. The Friedman's test results indicate that there is significant genotype effect on slope parameters (Friedman's χ^2 = 9.000, *df*=6, *p*-value=0.1736), meaning that the Friedman's test is not able to separate the different genotypes in terms of their response to the chemical process. However, the differences in average ranks of the slopes can still be used as some indicators how mixable certain genotypes are are.

7.4.5. Slope Covariances

The covariance matrix of the slope parameters (***G*** matrix) is estimated from the asymptotic distribution of the parameter estimates which was specified as;

$$\sqrt{N}\left(\widehat{\boldsymbol{\Phi}}^{*}-\boldsymbol{\Phi}^{*}\right)\sim N(\mathbf{0},\boldsymbol{A}\ \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}\boldsymbol{A}^{\mathrm{T}}).$$

In order to obtain the covariance matrix estimate $\widehat{\boldsymbol{G}}=\boldsymbol{A}\ \boldsymbol{J}^{-1}\boldsymbol{K}\boldsymbol{J}^{-1}\boldsymbol{A}^{\mathrm{T}}$, there is need to first find the ***J***, ***K*** and ***A*** matrices. Using the SAS code presented in Appendix A1.5, the $\widehat{\boldsymbol{G}}$ matrix was obtained as

$$\widehat{\boldsymbol{G}}=\begin{bmatrix} 0.7257 & 0.0171 & 0.0476 & -0.0334 & -0.0053 & -0.0234 \\ 0.0171 & 0.0010 & 0.0013 & -0.0006 & -0.0002 & -0.0012 \\ 0.0476 & 0.0013 & 0.0056 & -0.0030 & -0.0008 & -0.0016 \\ -0.0334 & -0.0006 & -0.0030 & 0.0037 & 0.0002 & 0.0001 \\ -0.0053 & -0.0002 & -0.0008 & 0.0002 & 0.0003 & 0.0002 \\ -0.0234 & -0.0012 & -0.0016 & 0.0001 & 0.0002 & 0.0032 \end{bmatrix}$$

It is assumed, for computational convenience, that all seven genotypes have the same covariance matrix for the six slope parameter estimates. This assumption could be relaxed for larger data sets which allow for the computation of such individual genotype covariance matrices. From the covariance matrix, we obtain the correlation matrix for the slope parameters calculated as

$$\boldsymbol{Corr}(\widehat{\boldsymbol{\Phi}}^{*}) = \left[\boldsymbol{diag}(\widehat{\boldsymbol{G}})\right]^{-\mathbf{1/2}} \widehat{\boldsymbol{G}} \left[\boldsymbol{diag}(\widehat{\boldsymbol{G}})\right]^{-\mathbf{1/2}}$$

to give

$$\boldsymbol{Corr}(\widehat{\boldsymbol{\Phi}}^{*}) = \begin{bmatrix} 1 & 0.6513 & 0.7434 & -0.6402 & -0.3788 & -0.4847 \\ 0.6513 & 1 & 0.5793 & -0.2928 & -0.4451 & -0.6990 \\ 0.7434 & 0.5793 & 1 & -0.6548 & -0.6749 & -0.3741 \\ -0.6402 & -0.2928 & -0.6548 & 1 & 0.1836 & 0.0212 \\ -0.3788 & -0.4451 & -0.6749 & 0.1836 & 1 & 0.2152 \\ -0.4847 & -0.6990 & -0.3741 & 0.0212 & 0.2152 & 1 \end{bmatrix}.$$

The covariance matrix for the residual terms ***R***, is obtained from the results in Tables 7.3(a), (b) and (c) above by averaging values in the $\widehat{\mathbf{R}}_{ij}$ matrices that correspond to particular variables since every value is estimated five times in the pairwise fitting process. After averaging out the ***R*** values for each variable the covariance matrix for the residual terms for the six variables is given as

$$\widehat{\boldsymbol{R}} = \begin{bmatrix} 131.06 & 9.0200 & 10.8995 & -10.3872 & 5.7503 & -12.6581 \\ 9.0200 & 1.9180 & 1.9676 & -1.8854 & 1.2172 & -2.6945 \\ 10.8995 & 1.9676 & 3.0621 & -3.1214 & 1.6174 & -3.3881 \\ -10.3872 & -1.8854 & -3.1214 & 3.6539 & -1.6043 & 3.8341 \\ 5.7503 & 1.2172 & 1.6174 & -1.6043 & 0.934 & -1.9626 \\ -12.6581 & -2.6945 & -3.3881 & 3.8341 & -1.9626 & 4.5119 \end{bmatrix}.$$

7.5. Discussion of results and conclusions

It is essential to outline the value addition of the joint modelling procedure to the study at hand and to thet understanding of the chemical evolution of dissolving pulp. A key result in the joint modelling procedure is the understanding of the interrelations or correlations of the evolutions of the chemical properties. This is a useful insight into the understanding of the interdependence of the chemical properties as they evolve through the processing stages. It is noted from the matrix $\boldsymbol{Corr}(\widehat{\boldsymbol{\Phi}}^{*})$, that the evolution of viscosity is positively correlated to that of lignin (*r*=0.6513) and γ-cellulose (*r*=0.7434) and negatively correlated to that of α-cellulose (r=-0.6401) and to a lesser extent negatively correlated to the evolution of copper number (r=-0.3788) and glucose (r=-0.4847). This means that when targeting viscosity in the chemical process it is inevitable that lignin and γ-cellulose will also be targeted while α-cellulose copper number and glucose will be moving in the opposite direction.

The other notable correlations in variable evolutions are the positive correlations between lignin and γ-cellulose (r=0.5793) and the negative correlation between lignin

and glucose (-0.6990). The variables γ-cellulose and α-cellulose (r=-0.6548), γ-cellulose and copper number (r=-0.6749) are also negatively correlated. The correlations between the other variable evolutions were not so large, for example, α-cellulose and glucose only had a correlation of r=0.0212 which means that the two variables cannot be targeted simultaneously in the chemical process.

Because the joint model attempts to estimate so many parameters at the same time, it has an unfortunate trade-off of compromising on accuracy as evident in the disparity between its slope parameter estimates and those obtained using the random coefficient models. One would need a fairly large amount of data to have more useful joint modelling results. In short joint modelling requires larger data sets than the other procudures discussed in this study. It presents more exciting further analysis provided more data can be collected. The main addition of joint modelling in this case in the analysis of the correlations of evolutions of different chemical properties.

Chapter 8

Discussions and Conclusion

This chapter presents a summary of all the work that was conducted in this study, highlighting significant findings and value of the work covered. Limitations of the study and possible improvements and extensions are also suggested.

The main focus of the study was to understand the behaviour of seven timber genotypes when going through the chemical pulping process with the prime objective of developing methods of grouping different timber genotypes into compatible groups of genotypes that can be optimally processed together. The main features of the data that presented a genotype mixing criteria are the evolution profiles of the chemical properties (variables) going chemical processing.

In order to understand the behaviour of the seven genotypes studied, four related statistical methods were used, namely, random coefficients models, under the mixed models framework, piecewise linear regression models, which made use of the three inherent sub-processes in chemical pulping, a combination of piecewise linear regression models and kernel density estimation as a clustering (grouping) tool and joint modelling which sought to understand the joint evolution of the chemical properties over the processing stages.

While the methods studied were specifically for timber chemical pulp data, they can also be extended to other materials under completely different industrial processes. It is a well-known fact that manufacturing systems can use raw materials from different sources with different characteristics that might affect the properties and quality of the final product. Such materials would need to be carefully scrutinised before they can me fed into the production system especially if there is need to mix them. There might be a need to identify source regions or varieties of the raw materials that can be optimally mixed during production.

The important results that came out of fitting random coefficient models to the data is that the higher the raw stage readings the higher the rates of change in the chemical properties over the processing stages where changes might be increases or decreases in the chemical properties studied. In a way, this meant that genotypes that started with similar readings at the raw stages tended to behave in a similar way during the processing stages hence such genotypes are highly mixable.

The random coefficient models also yielded a mixing criteriafor the different genotypes based on the average ranking of the slope parameters (rates of change) for the seven variables studied. The random coefficient results summarised in Table 7.1 indicate that the genotypes GUA and GUW are the least mixable if we are to compare their average rankings which are poles apart.

Table 8.1. Average slope genotype ranks based on random coefficient models.

Genotype	E.dunnii	E.grandis	E.smithii	E.nitens	GCG	GUA	GUW
Average Rank	4.71	3.29	3.86	4.57	4.14	5.57	1.86

For a much broader problem with more genotypes or categories of any raw materials that feed into a manufacturing process, the ranking criteria can be very useful in deciding which materials to mix optimally. In this problem, processing stage was considered as a time variable which might be problematic since the intervals between stages might not be uniform. The methods developed here might be more realistic in problems with an interval scaled time variable.

Chapter 4, which is a further development to Chapter 3, identified the three sub-processes that make up the whole process of chemical processing. This led to a time coding method to account for these sub-processes. This meant that the performance of each sub-process could now be evaluated individually. Differences in genotype behaviours per sub-process provided a deeper understanding of the different genotypes throughout the chemical pulping process hence a more accurate grouping mechanism. Piecewise linear regression modelling was used to model each sub-process as a linear component of a much bigger nonlinear process. The models had the capability to outline the effect of each sub-process of chemical pulping on the seven reactivity variables studied. The ability of the model to state, by the model

parameters, the effect of each sub-process on the chemical properties is a value addition to the study of chemical pulping processes. This can be extended to other types of pulp processing with known sub-processes, for example, kraft pulping and neutral sulphite pulping.

Based on the results from the piecewise linear regression models it was established that the six chemical properties lignin, γ-cellulose, α-cellulose, copper numbers, glucose and xylose were important classification variables for species/genotypes while viscosity was not. This means that when one wants to compare or group wood species/genotypes using their chemical properties for the purpose of deciding which ones are mixable during processing, they do not need to consider viscosity.

Using kernel density estimation, a mixing scale that can be used to determine if any two genotypes can be optimally mixed for processing was developed. The scale was based on the rates of response of the genotypes to the two key sup-processes of delignification and bleaching. The behaviour of the genotypes, as measured by the rates of change in the chemical properties of lignin, α-cellulose, γ-cellulose, viscosity, copper numbers, glucose and xylose, were used to develop a mixing matrix for the genotypes. This scale can be adopted for similar raw material mixing problems.

The underlying assumption in this study is that all the chemical properties are of equal importance and no single chemical property can override the values in the other properties. The chemical pulping process targets the production of higher levels of *α*-cellulose but this chemical property did not produce different genotype clusters indicating that all genotypes had similar response profiles to the sub-processes of bleaching and delignification as far as α-cellulose is concerned. This means that, for the purpose of determining mixable genotypes, α-cellulose is not an important variable. Viscosity and copper numbers also came up as not important when determining which genotypes can be mixed during processing.

The chemical properties of lignin, γ-Cellulose, glucose and xylose were deemed important determinants of how mixable timber genotypes could be. These are not the only chemical properties involved in chemical pulping hence there is need to

incorporate as many properties as possible in order to get a more accurate classification method. A study incorporating as many chemical properties as possible is a possible area of further studies.

Mixing of timber varieties during processing will always occur especially when one genotype alone does not have economic quantities for processing. The problem can also be compounded by the current drive for materials recycling. It is reasonable to mix genotypes or materials for processing after considering how they individually behave during chemical processing so that only those materials that have many properties in common are optimally mixed during processing. The random coefficient model and the piecewise linear regression model coupled with kernel density estimation provided a matrix that can be used for optimally mixing different genotypes for chemical processing. However, the variables where not considered together, especially as far as their inter-correlations are concerned. Joint modelling put an additional dimension to the study by considering all the variables together and how they interact with each other during processing. The coorelations of the evolutions of different chemical properties, as analysed using joint modelling, indicated which chemical properties can be targeted together, that is those with high correlations in their evolutions over the stages of chemical processing. The key result of Joint modelling was that it brought out a very insightful understanding of the interdependence of the variables as they evolve over the processing stages. Using joint modelling it was discovered that the evolution of viscosity is positively correlated with that of lignin (r=0.6513) and γ-cellulose (r=0.7434) and negatively correlated to that of α-cellulose (r=-0.6401). There was also notable positive correlations in variable evolutions between lignin and γ-cellulose (r=0.5793) and negative correlation between lignin and glucose (-0.6990). The variables γ-cellulose and α-cellulose (r=-0.6548), γ-cellulose and copper number (r=-0.6749) are also negatively correlated. The correlations between the other variable evolutions were not so large, for example, α-cellulose and glucose only had a correlation of r=0.0212. An understanding of this relationships in chemical changes helps the manufacturer to what would happen to other chemical properties when one particular property is being targeted.

The main limitation of the joint modelling method was its computational challenges. Because in joint modelling many parameters are estimated at the same time, there will

always be convergence problems. In fitting the joint model in this study, it was not possible to estimate both the intercept and slope parameters, hence the use of intercept corrected data. The results obtained using this approach, while providing a useful insight into the inter-correlations between response variables, may not be as accurate as those obtained using the other methods discussed. This is an unfortunate trade-off that compromises accuracy for more understanding of correlations between response variables. This is evident in the disparity between slope parameter estimates obtained using joint modelling and those obtained using the random coefficient and piecewise linear regression models. The main addition of joint modelling in this case is the analysis of the correlations of evolutions between different variables.

When all methods are considered it would seem tah the genotype Enitens is far removed from the other genotypes. When considering which genotypes to mix during processing Enitens should be processed on its own at all costs as it shown to be very different from the other genotypes in both the random coefficient and thye piecewise linear regression models.

A more comprehensive dataset with more experimental units would greatly improve the acturacy of these findings. The data expansion should also include a greater number of chemical properties and genotypes and even include recycling materials. The methods discussed in this study can also be adopted for other non-timber processes.

This study suggested a novel method of mixing raw materials optimally, for production systems that get their raw materials from various sources. Such sources might be producing raw materials of different quality and with different chemical properties. The use of Kernel density estimation as a clustering tool and the use of piecewise linear regression as a modelling tool for manufacturing processes with different sub-processes is novel to the best of my knowledge. The study also made use of various known data analysis techniques that have been applied in other fields, particularly in disease modelling, to model manufacturing processes. Methos such joint modelling, have been used extensively to model medical data and this study sought to make use of such methods for the modelling of manufacturing proceses.

References

Agresti, A. (1997). A Model for Repeated Measurements of a Multivariate Binary Response. *Journal of the American Statistical Association*, Vol 92 (437), pp 315-321.

Arnold, B.C. and Strauss, D. (1991). Pseudo-likelihood estimation: some examples. *Sankhya: The indian journal of Statistics-Series B,* Vol 53, pp 233-243.

Berget, I. and Næs, T. (2002a). Sorting of raw materials with focus on multiple end-product properties. *Journal of Chemometrics*, Vol. 16, pp 263-273.

Berget, I. and Næs, T. (2002b). Optimal Sorting of Raw Materials, Based on the Predicted End-Product Quality. *Quality Engineering*, Vol. 14(3), pp 459-478.

Biermann, C. J., (1993). *Handbook of Pulping and Papermaking*. Academic Press. San Diego.

Blatt, M., Wiseman, S. and Domany, E. (1997). Data clustering using a model granular. *Neural Computation*, Vol. 9 (8). Pp 1805-1842.

Bollen, K.A., and Curran, P.J., (2006). *Latent Curve Models: A Structural Equation Perspective*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, Vol. 52 (3), pp 345–370.

Brumback, B. A. and Rice, J.A. (1998). Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves. *Journal of the American Statistical Association*, Vol 93 (443), pp 961-976. DOI: 10.1080/01621459.1998.10473755

Bryk, A.S. and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methos,* Newbury Park, CA: Sage.

Burnhan, K.P., and Anderson, D.R., (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological methods and research*, Vol. 33, pp 261–304.

Casey, J.P. (1983). *Pulp and Paper, Chemisty and Chemical Technology*. 3rd Ed. Vol 4. John Wiley & Sons. Toronto.

Cheng, Y. (1995). Mean Shift, Mode seeking, and Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 17 (8), pp 790-799.

Chiu, S. T. (1991). Bandwith selection for Kernel density estimation. *The Annals of Statistics*, Vol. 19(4), pp 1883-1905.

Coull, B. A. (2011). A Random Intercepts – Functional Slopes Model for Flexible Assessment of Susceptibility in Longitudinal Designs. *Biometrics*, Vol. 67(2), pp 486–494.

Craven, P., and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerical Mathematics*, Vol. 31, pp 377- 403.

Davidian, M., Verbeke, G., Molenberghs, G. (2008). *Longitudinal Data Analysis*. Chapman & Hall/CRC, pp 219-226.

Davis, M.W. (1998). A Rapid Modified Method for Compositional Carbohydrate Analysis of Lignocellulosics by High pH Anion-Exchange Chromatography with Pulsed Amperometric Detection (HPAEC/PAD). *Journal of Wood Chemistry and Technology*, Vol. 18(2), pp 235–252.

Dechateau, L., Jansen, P. and Rowlands, G. J. (1998). Linear Models, An introduction with Applications in Veterinary Reserarch. International Livestock Research Institute, Nairobi, Kenya, pp 23-74.

Diggle, P., Sousa, I. and Chetwynd, A.G. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth Armitage lecture. *Statistics in Medicine*, Vol 27(16), pp 2981–2998.

Erdoğmus, D., Carreira-Perpñán M. Á., and Özertem, U. (2006). Kernel density estimation, affinity-based clustering, and typical cuts. Acoustic, Speech and Signal processing ICASSP proceedings. *IEEE international Conference* Vol 5, Toulouse 2006.

Everitt, B. S, Landau, S., Leese, M and Stahl D. (2011). *Cluster Analysis (5th Ed.)*, Wiley Series in Probability and Statistics, John Wiley and Sons Ltd. West Sussex.

Faes C., Geys. P. and Catalano, P. (2008). *Longitudinal Data Analysis*. Edited by Geert Verbeke, Marie Davidian, Garrett Fitzmaurice and Geert Molenberghs, Chapman and Hall/CRC, pp 327–348.

Fieuws. S. and Verbeke, G. (2007). Random-effects models for multivariate repeated measures. Statistical Methods in Medical Research Vol 16, pp 387-397.

Fieuws. S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modelling of multivariate longitudinal profiles*. Biometrics* Vol 62, 424–431.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). Applied Longitudinal Analysis. Wiley Series in Probability and Mathematical Statistics, Wiley, New Jersey, pp 147-150.

Funaoka, M., Chang, V. L., Kolppo, K., Stokke, D. D. (1991). Comparison of condensation reactions at Cα positions in kraft and acid sulfite delignification of Western hemlock. *Bulletin of the Faculty of Bioresources*, Mie University, No. 5, 37-44.

García, S., Fernández, A., Luengo, J. and Herrera, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences* Vol. 180, 2044–2064

Gelman, A. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics*, Volume 13 (4), pp 755–779.

Gentle, J. E. (1998). "Cholesky Factorization." *Numerical Linear Algebra for Applications in Statistics*. Berlin: Springer-Verlag, 93-95.

Geys, H., Molenberghs, G. and Ryan, L.M. (1999). Pseudo-likelihood Modelling of Multivariate Outcomes in Developmental Toxicology. *Journal of the American Statistical Association*, Vol. 94 (447), 734-745.

Gierer, J. (1985). Chemistry of delignification. *Wood Science and Technology*, Vol 19 (4), 289-312.

Gu, J. and Liu, C. (2012). Discriminative illumination: Per-pixel classification of raw materials based optimal projections of spectral brdfs. *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 797 – 804.

Gueorguieva, R. (2001). A multivariate generalized linear mixed model for joint modelling of clustered outcomes in the exponential family. *Statistical Modelling*, Vol 1, 177-193.

Guo, X. and Carlin, B. P. (2004). Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages. *The American Statistician*, Vol. 58 (1), pp 1-9. DOI: 10.1198/0003130042854 (Accessed 21 June 2016).

Hamlett, A., Ryan, L., Serrano-Trespalacios, P. and Wolfinger, R. (2003). Mixed Models for Assessing Correlation in the Presence of Replication. *Journal of the Air and Waste Management Association,* Vol 53 (4), 442-450.
DOI: 10.1080/10473289.2003.10466174.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*, Springer-Verlag, Berlin. ISBN 978-3-642-17146-8

Hastie, T.J. and Tibshirani, R.J. (1986). Generalized additive models. *Statistical Science*, Vol 1(3), 297-318.

Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized additive models*. Chapman and Hall. London.

He, J., Cui, S. and Wang S-y. (2007). Preparation and crystalline analysis of high-grade bamboo dissolving pulp for cellulose acetate. *Journal of Applied Polymer Science,* vol 107. pp1029-1038.

Hearne, E.M., Clark, G. M., Hatch, J. P. (1983). A Test for Serial Correlation in Univariate Repeated-Measures Analysis. BIOMETRICS Vol. 39, pp 237-243. (http://www.jstor.org/stable/pdf/2530823.pdf, accessed 17 March 2016).

Henderson, C. R. (1953). Estimation of Variance and Covariance Components *Biometrics Vol.* 9(2), pp 226–252.

Henderson, R., Diggle, P. and Dobson, A. (2000). Joint modelling of Longitudinal Measurements and Event Time Data. *Biostatistics*, Vol 1(4), pp 465-480.

Iddi, S. and Molenberghs, G. (2012). A Joint Marginalized Multilevel Model for Longitudinal Outcomes. *Journal of Applied Statistics*, Vol. 39 (11), pp 2413-2430.

Introduction to SAS. UCLA: Statistical Consulting Group.
http://www.ats.ucla.edu/stat/sas/notes2/ (accessed June 14, 2016).

Jahan, M.S., Ahsan, L., Noori, A. and Quaiyyum, M.A., (2008). Process for the production of dissolving pulp from *Trema Orientalis (Natalia)* by prehydrolysis kraft and soda-ethylenediamine (EDA) process. *BioResources,* vol. 3(3), pp 816-828.

Jensen, R., Erdoğmus, D., Principe, J. C. and Eltoft, T. (2004). The Laplacian PDF distance: A cost function for clustering in a kernel feature space. *Proceedings of the conference on Neural Information Processing Systems,* 2004.

Johnson, R. A. and Wichern, D. W. (1998). *Applied multivariate statistical Analysis,* 4th Ed. Prentice Hall, New Jersey. Pp 726-754.

Karlsson, H. (2006). Fibre guide: fibre analysis and process applications in the pulp and paper industry: A handbook. Lorentzen & Wettre.

Kulesz,P. A., Francis, D. F., Barr, C. D. (2016). Multi-Panel Scatter Plots and Scatter Plot Matrices. On http://www.scsug.org/wp-content/uploads/2012/11/Multi-Panel-Scatter-Plots-and-Scatter-Plot-Matrices1.pdf. (accessed on 17 June, 2016.)

Kundu, M. G. (2011). Implementation of Pairwise Fitting Technique for Analysing Multivariate Longitudinal Data in SAS. *PharmaSUG2011* - Paper SP09. Http://www.lexjansen.com/pharmasug/2011/SP/PharmaSUG-2011-SP09.pdf. (Accessed 12 November, 2015).

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data, *Biometrics*, Vol. 38 (4), 963-974.

Li, J., Ray, S. and Lindsay, B. G. (2006). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, Vol 8, 1687-1723.

Lina, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis using Generalised Linear Models. *Biometrika*, Vol. 73(1), 13-22.

Littell, R.C., Milliken, G.A., Stroup, W. W., Wolfinger, R. D. and Schabenberger, O., (2006). *SAS for Mixed Models,2nd Ed*. SAS Institute Inc. North Carolina.

Lodi, S., Moro, G. and Sartori, C. (2006). Stream Clustering Based on Kernel Density Estimation. *European Conference on Artificial Intelligence (ECAI 2006)*. IOS Press, Amsterdam. http://ebooks.iospress.nl/volumearticle/2835 **(**accessed on 17-11-2014**).**

Liu, C., Cao, D., Chen, P. and Zagar, T. (2007). *RANDOM and REPEATED statements - How to Use Them to Model the Covariance Structure in Proc Mixed*. Eli Lilly & Company. Indianapolis.

Liu, X., Daniels, M. J. and Marcus, B. (2009). Joint models for the association of longitudinal binary and continuous processes with application to smoking cessation trial. *Journal of the American Statistical Association,* Vol. 104, pp 429-438.

Marron, J. S. (1987). A Comparison of Cross-Validation Techniques in Density Estimation. *Annals of Statistics*, Vol. 15(1), pp 152-162.

Matange, S., Heath, D. (2011). Statistical Graphics Procedures by Example: Effective Graphs Using SAS. Cary, NY: SAS Institute Inc.

Mathews, J.H. and Fink, K.D. (2004). *Numerical Methods using MatLab, 4th Edition*. Prentice-Hall Inc, New Jersey, pp280-290.
Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, New York.

Musdholifah, A., Hashim, S. Z. M. and Nga, R. (2013). Robust Local Triangular Kernel Density-based Clustering for High-dimensional Data. *5th International Conference on Computer Science and Information Technology (CSIT)*, (ISBN: 978-1-4673-5825-5).

Parzen, E (1962). On estimation of probability density functions and mode. *Annals of Mathematical Statistics,* Vol. 33, pp 1065-1076.

Patrick, K., (2011). The Dissolving Pulp Gold Rush in High Gear. Paper 360º. http:www.tappi.org. [accessed on 1 September 2011}

Pohlert, T. (2016). The Pairwise Multiple Comparison of Mean Ranks Package (PMCMR). R package. http://CRAN.R-project.org/package=PMCMR (accessed on 1-04-2017).

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data With Application in R*. Chapman & Hall/CRC Biostatistics Series Boca Raton.

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, Vol. 6 (1), pp 15-32.

Röhrling, J., Potthast, A., Rosenau, T., Lange, T., Ebner, G., Sixta, H. and Kosma, P.(2002). A Novel Method for the Determination of Carbonyl Groups in Cellulosics by Fluorescence Labeling. 1. Method Development. *Biomacromolecules,* Vol. 3 (5), pp 959-968.

Rorres, C. and Howard A. (1984). *Applications of Linear Algebra 3rd Edition*. John Wiley and Sons ,New York.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics, Vol.* 27, pp 832-837.

Roy, A. (2006). Estimating Correlation Coefficient between Two Variables with Repeated Observations using Mixed Effects Model. *Biometrical Journal, Vol*. 48 (2), 286–301. DOI: 10.1002/bimj.200510192.

Rustagi, J.S. (1994). *Optimisation Techniques in Statistics*. Academic Press, New York.

SAS Institute Inc. (2008). SAS/STAT 9.2 User's Guide. Cary, NC: SAS Institute Inc

Schwarz, J. (2005). Clustering Analysis of Micro Array Data, SAS Institute Inc. Paper SA07_05, http://analytics.ncsu.edu/sesug/2005/SA07_05.PDF (accessed: 24 July, 2014).

Seaman, D. E., Millspaugh, J. J., Kernohan, B. J., Brundige, G. C., Raedeke, K. J. and Gitzen, R. A., (1999). Effects of Sample Size on Kernel Home Range Estimates. *The Journal of Wildlife Management*, Vol. 63(2), 739-747.

Searle, S. R. (1995). The Matrix Handling of BLUE and BLUP in the Mixed Linear Model. *The Fourth International Workshop on Matrix Methods for Statistics*, Montreal, Quebec, July 15-16, 1995.

Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, Vol. 7, 221-264.

Shu, G., Zeng, B., Chen, Y. P., Smith, O. H. (2003). Performance assessment of kernel density clustering for gene expression profile data. *Comparative and Functional Genomics*, Vol 4, 287-299.

Silverman, B.W., (1986). *Density Estimation For Statistics And Data Analysis.* Chapman and Hall. ISBN 0-412-24620-1

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer. ISBN 0-387-94716-7

Smyth, G.K. (2002). Optimisation. *Encyclopedia of Environmetrics*, Edited by El-Shaarawi A.H. and Piegorsch, W.W. John Wiley & Sons, Ltd, Chichester, 1481–1487.

Snijders, T.A. and Bosker, R. (1999). Multilevel Analysis: An introduction to Basic and Advanced Multilevel Modelling. Newbury Park, CA: Sage.

Soskolne, C. L. and Sieswerda, L. E. (2010). Cancer risk associated with pulp and paper mills: a review of occupational and community epidemiology. *Chronic Diseases in Canada,* Vol. 29 (2).

Sousa, I. (2011). A Review on Joint Modelling of Longitudinal Measurements and Time-to-Event. *REVSTAT–Statistical Journal*, Vol 9(1), 57–81.

Sundstrom, D.W., Klel, H.E. and Daubenspeck, T.H., (1983). Use of Byproduct Lignin as Externders in Asphalt. Industrial Engineering and Chemical Production, Research and Development, Vol. 22, 496-500.

Swamy, P.A.V.B. (1970). Efficient Inference in a Random Coefficient Regression Model. *Biometrica* Vol. 38(2), 311-323.

Tasman, J.E., and Berzins, V. (1957). The Permanganate Consumption of Pulp Materials. *Tappi Journal, Vol*. 40(9), 695-704.

Tran, T.N., Wehrens, R. and Buydens, L.M.C (2006). KNN-kernel density based clustering for high-dimensional multivariate data. *Computational Statistics and Data Analysis,* Vol 51, 513-525.

Tsiatis, A. A. and Davidian, M. (2004). Joint Modeling of Longitudinal and Time-To-Event Data: An Overview, Statistica Sinica Vol 14, pp 809-834.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed models for Longitudinal Data*. Springer-Verlag, New York.

Verbeke, G. and Davidian M (2008). Joint Models for Longitudinal Data: introduction and overview*. Longitudinal Data Analysis*, Edited by Fitzmaurice, G.

Viscosity of pulp, Tappi T230 om-94, 1994, Tappi Press, Atlanta.

Wahba, G. (1975). Optimal Convergence Properties of Variable Knot, Kernel, and Orthogonal Series Methods for Density Estimation. *The Annals of Statistics*, Vol 3(1), pp 15-29.

Wang W. J., Tan, Y. X., Jiang, J. H., Lu, J. Z., Shen, G. L. and Yu R. Q., (2004). Clustering based on kernel density estimation: nearest local maximum searching algorithm. *Chemometrics and Intelligent Laboratory Systems,* Vol 72, 1-8.

Wang, Y. (2007). Fisher scoring: An interpolation family and its Monte Carlo implementations. *Computational Statistics and Data Analysis*, Vol 54. 1744-1755.

Wedderburn, R. W. M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, Vol. 61(3), pp 439-447.

Wicklin, R. (2013). *Simulating Data with SAS*. SAS Institute Inc, Cary North Carolina.

Wood, S.N. (2006). *Generalized Additive Models, An introduction with R*. Chapman & Hall/CRC, Taylor Francis, Boca Raton.

Wooldridge, J. M. (1997). Quasi-likelihood methods for count data. *Handbook of applied econometrics,* Vol. 2, pp 352-406.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics,* Vol 44, 175-188.

Xiang, D. and Wahba, G. (1996). A genralized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, Vol. 6, 675-692.

Zbonak, A., Bush, T. and Grzeskowiak V. (2007). Comparison of tree growth, wood density and anatomical properties between coppiced trees and parent crop of six Eucalyptus genotypes. *Proceedings of the International Union of Forest Research Organizations* (IUFRO), Durban2007.

Zhang, X., King, M. L. and Hyndman R. J. (2004). Bandwidth selection for multivariate kernel density estimation using MCMC. *Econometric Society 2004 Australasian Meetings.* http://repec.org/esAUSM04/up.1603.1077410300.pdf

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2002). Variable selection and model building via likelihood basis pursuit, *Technical Report No. 1059*, Department of Statistics, University of Wisconsin.

Zobel BJ, Van Buijtenen JP (2012). Wood variation: its causes and control. Springer Science & Business Media; Vancouver.

Appendices

A.1. Model Diagnostics – Residual Analysis

A1.1. Residuals for Random Coefficient Models - Chapter 3

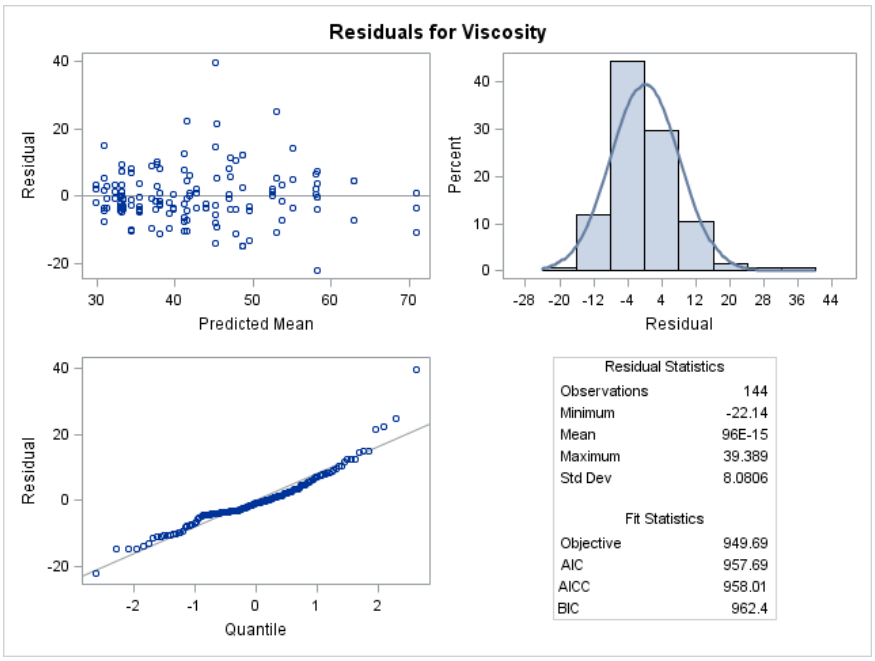


Figure A1.1. Residual plots for the random coefficient model for viscosity.

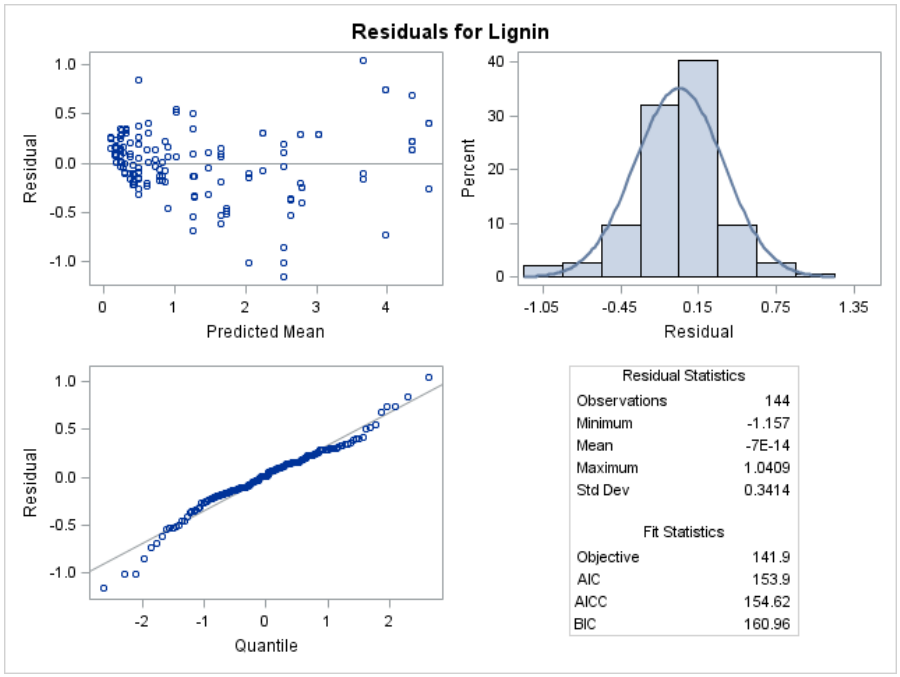


Figure A1.2. Residual plots for the random coefficient model for lignin.

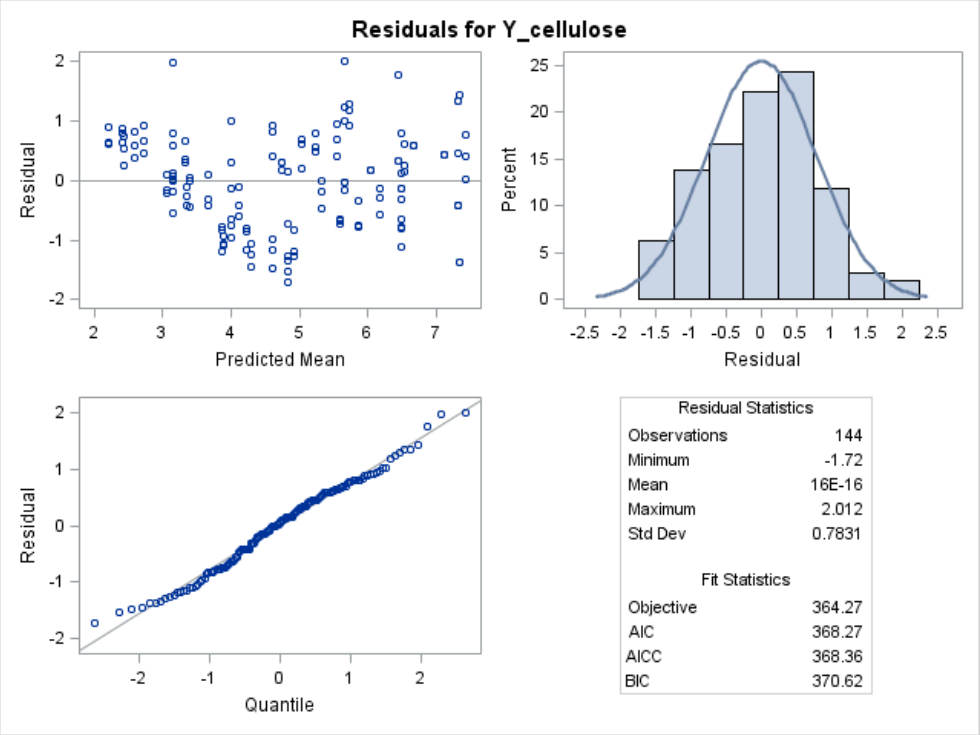


Figure A1.3. Residual plots for the random coefficient model for γ-cellulose.

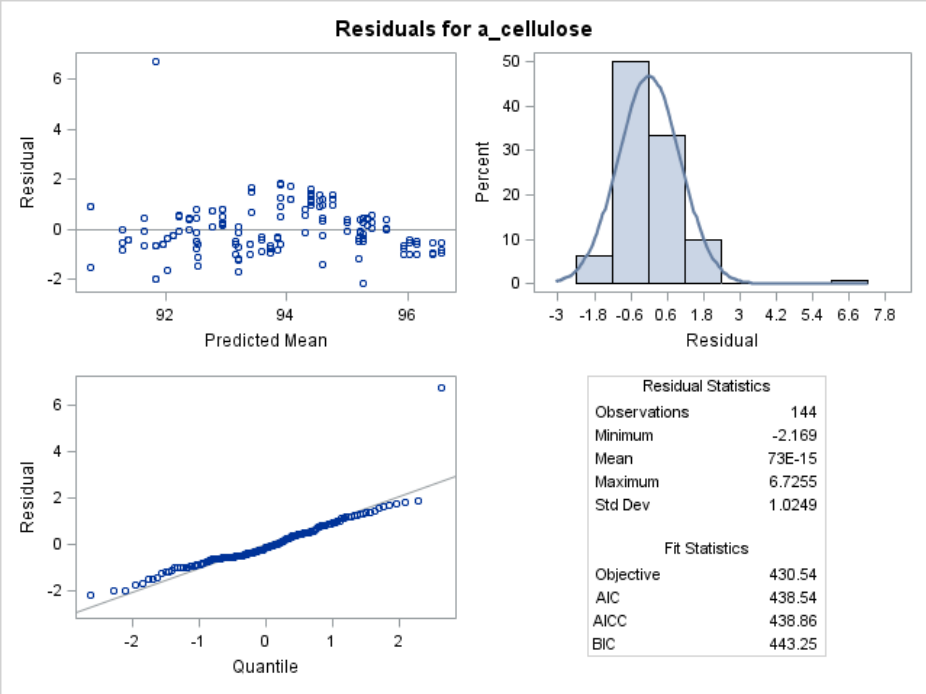


Figure A1.4. Residual plots for the random coefficient model for α-cellulose.

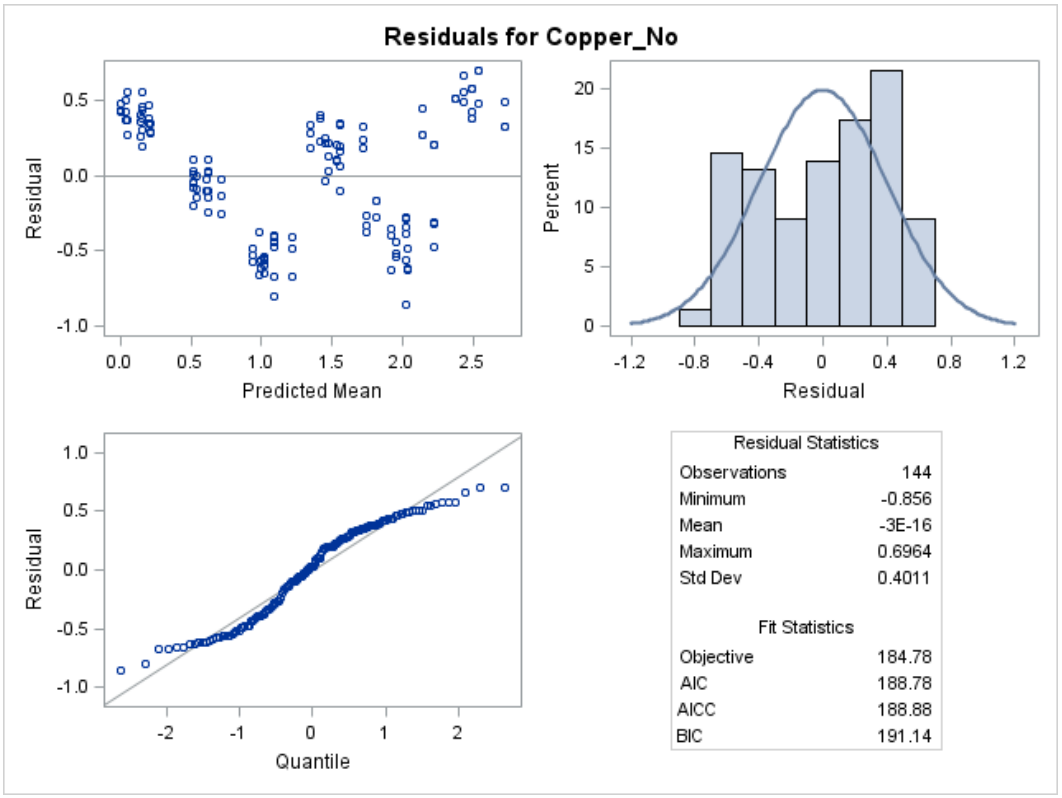


Figure A1.5. Residual plots for the random coefficient model for copper number.

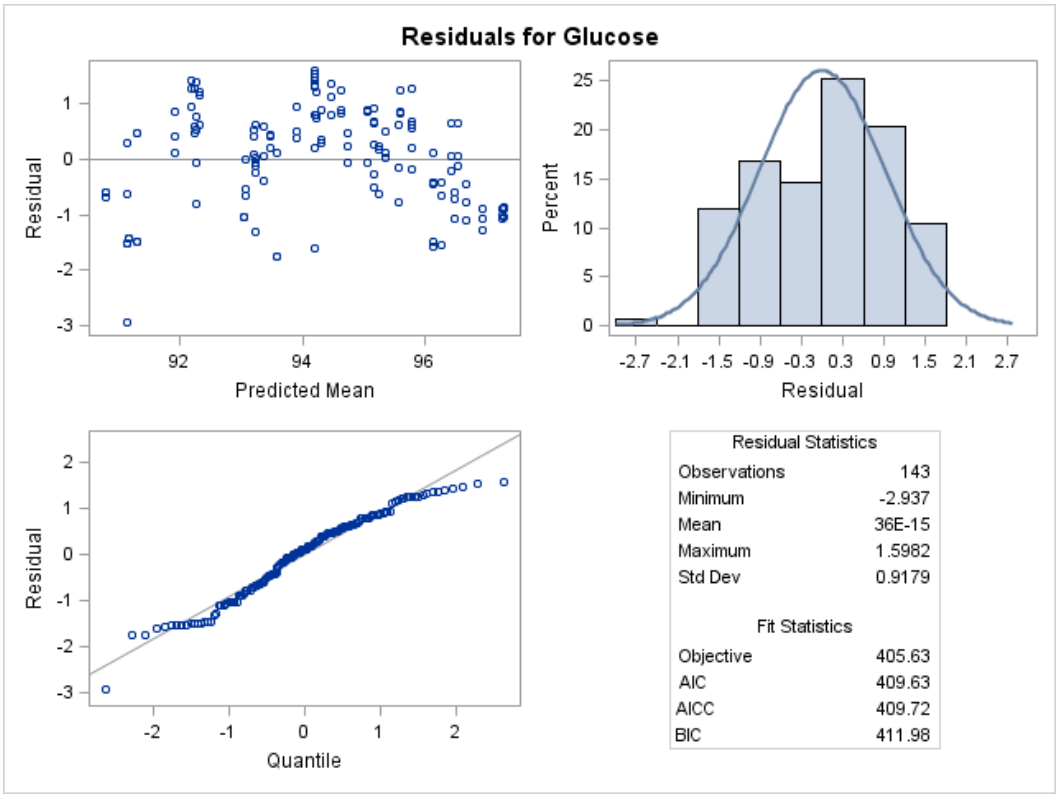


Figure A1.6. Residual plots for the random coefficient model for glucose.

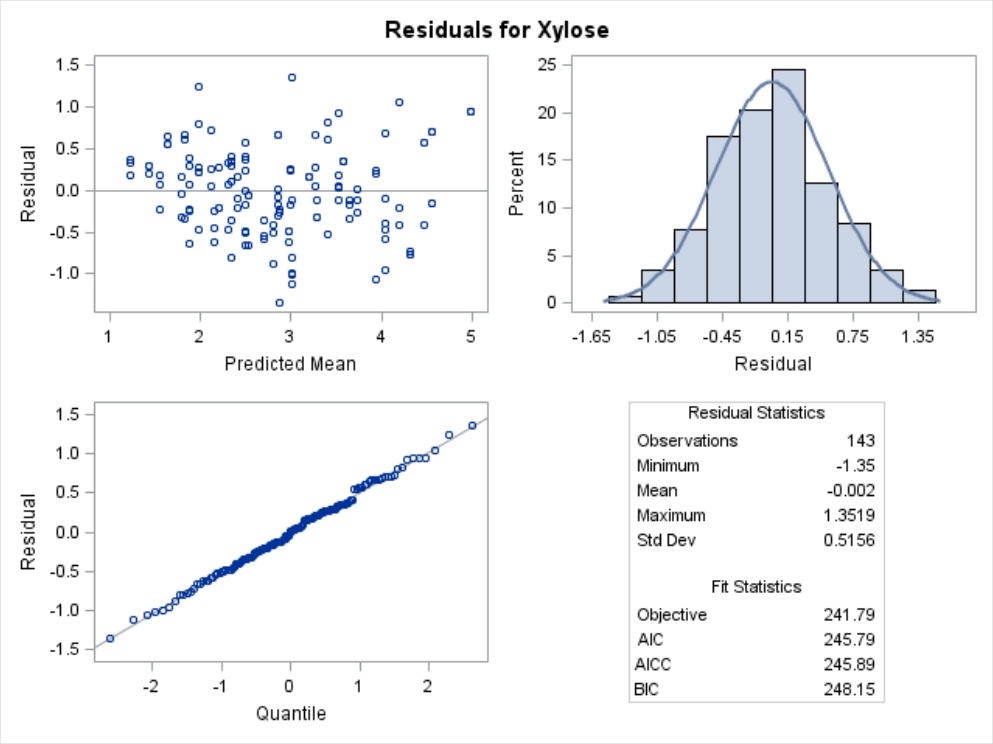


Figure A1.7. Residual plots for the random coefficient model for xylose.

A1.2. Residuals for Piecewise Linear Regression Models - Chapter 4

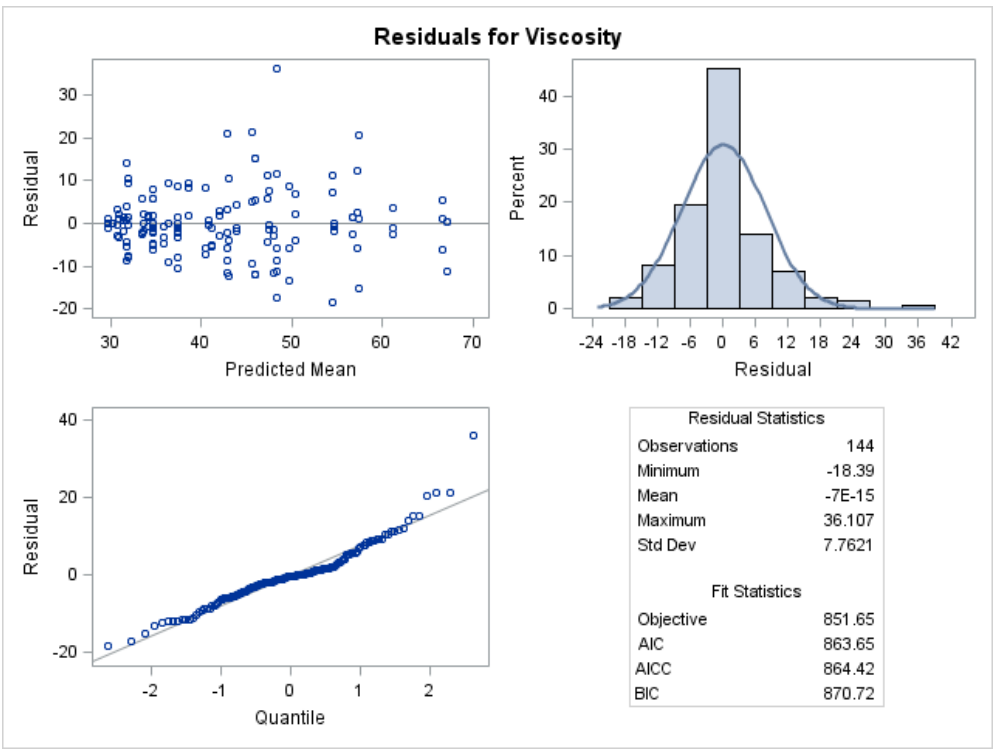


Figure A1.8. Residual plots for the piecewise linear regression mode for viscosity.

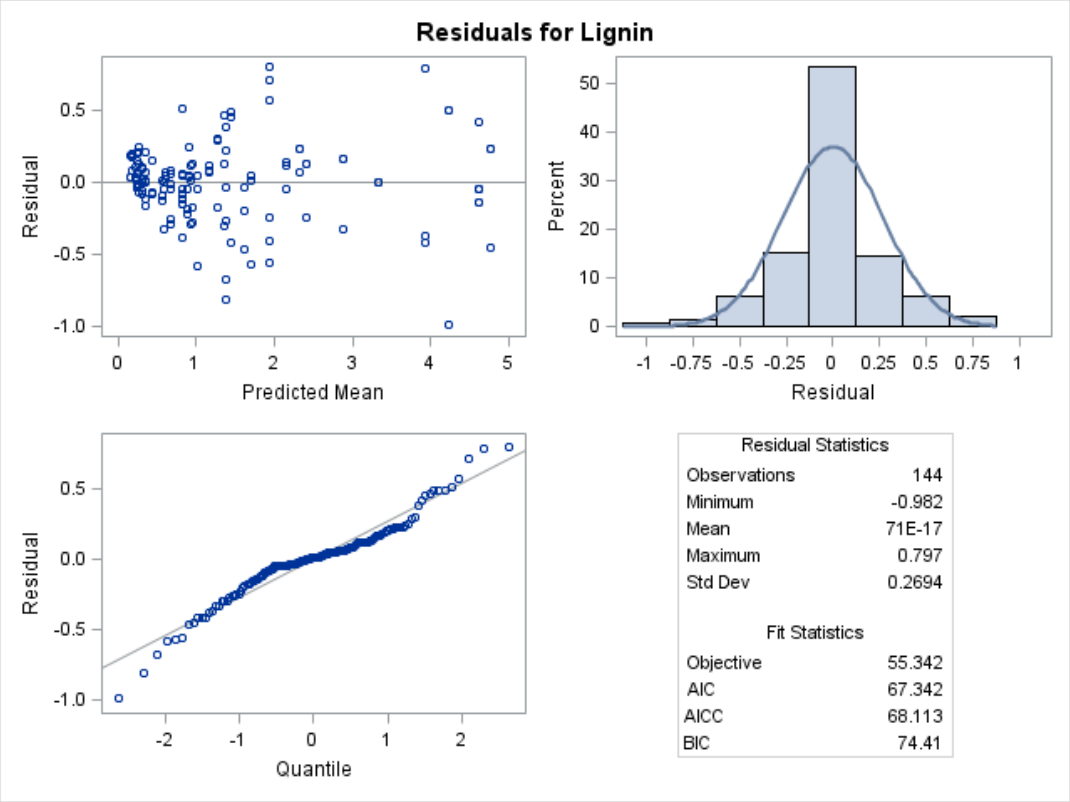


Figure A1.9. Residual plots for the piecewise linear regression mode for lignin.

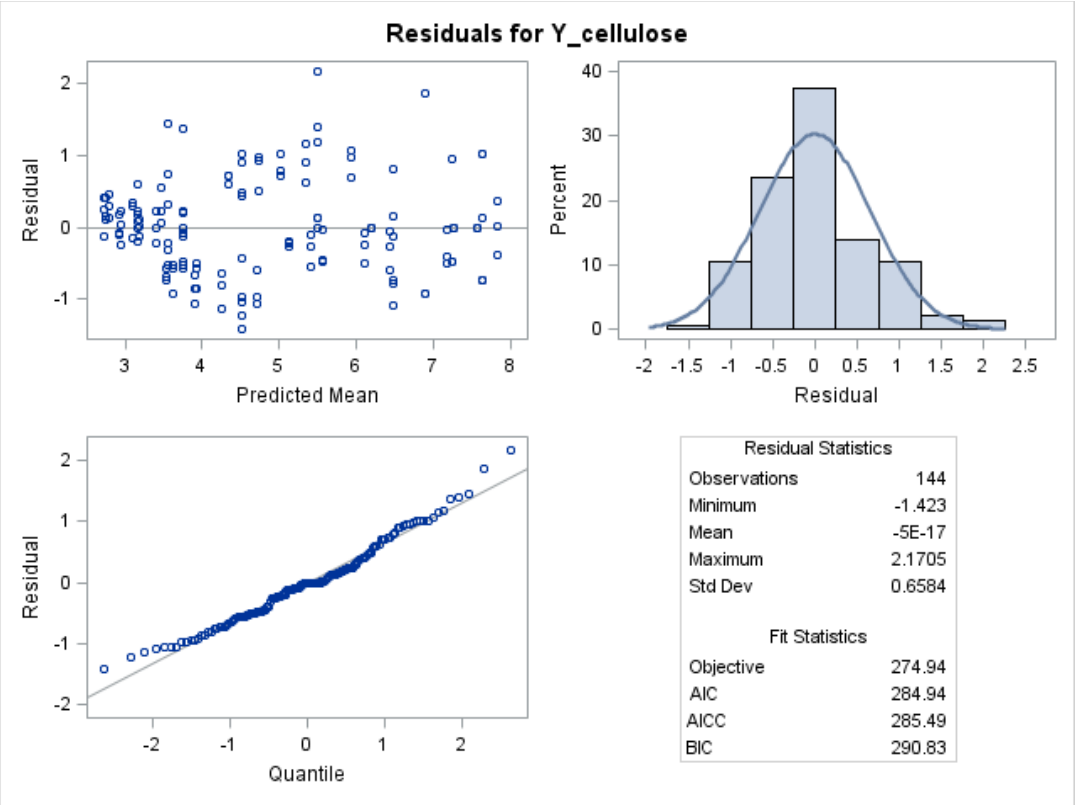


Figure A1.10. Residual plots for the piecewise linear regression mode for γ-cellulose.

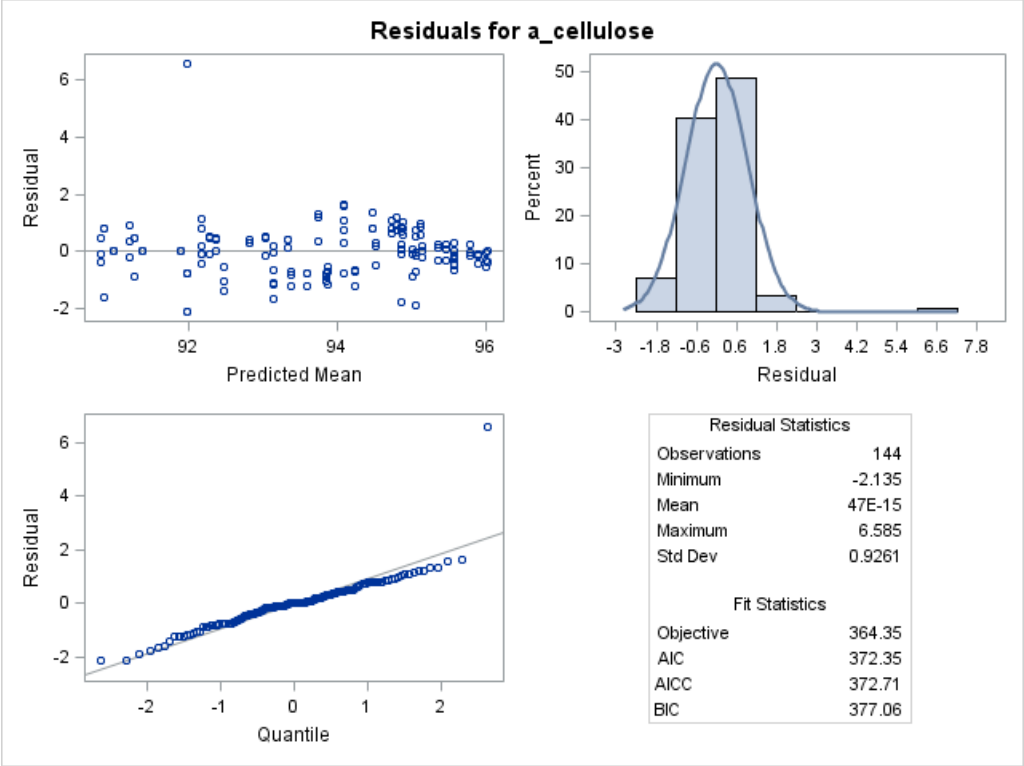


Figure A1.11. Residual plots for the piecewise linear regression mode for α-cellulose.

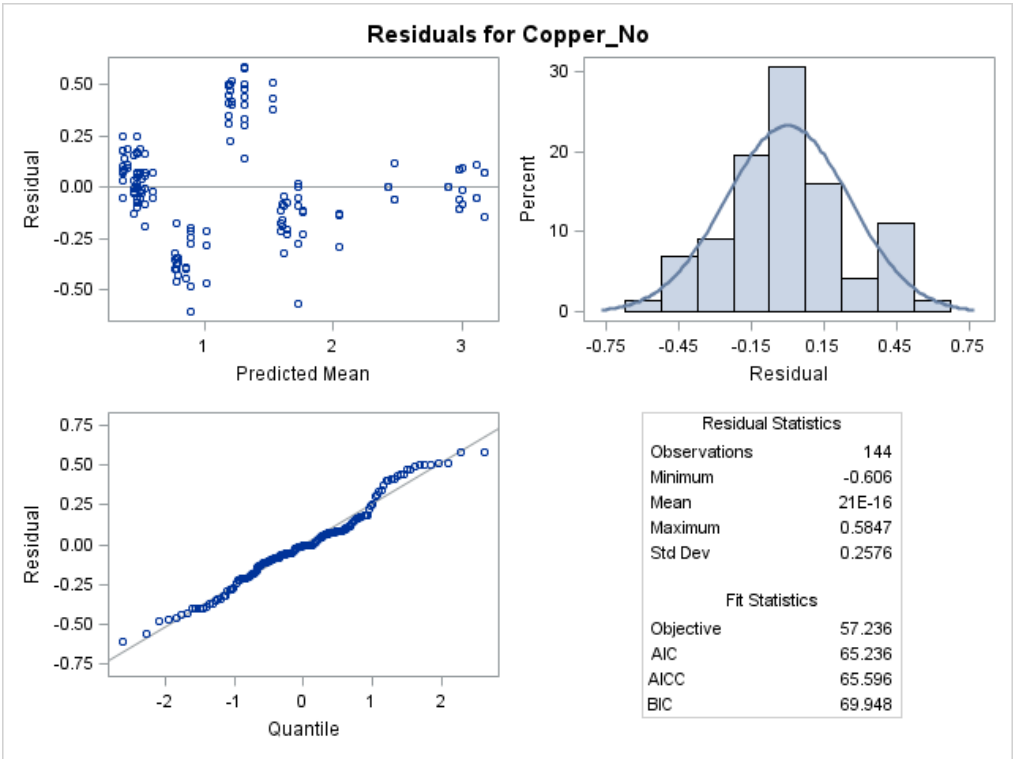


Figure A1.12. Residual plots for the piecewise linear regression mode for copper number.

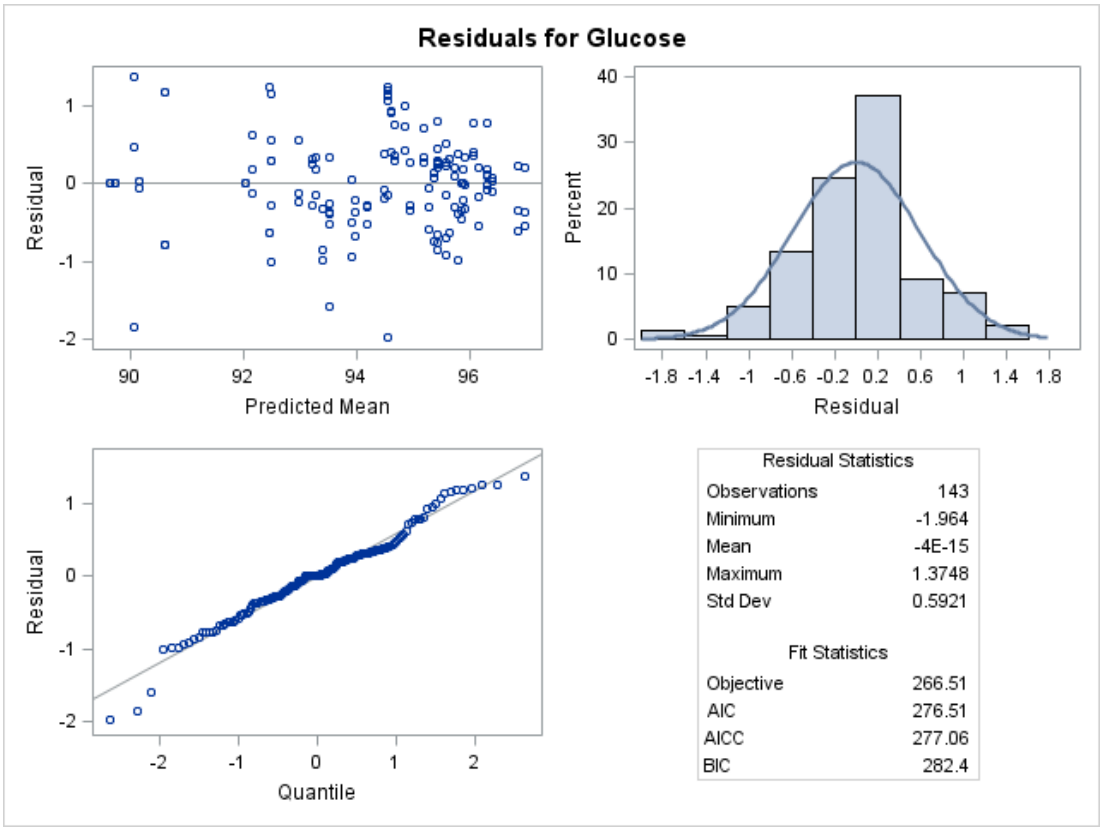


Figure A1.13. Residual plots for the piecewise linear regression mode for glucose.

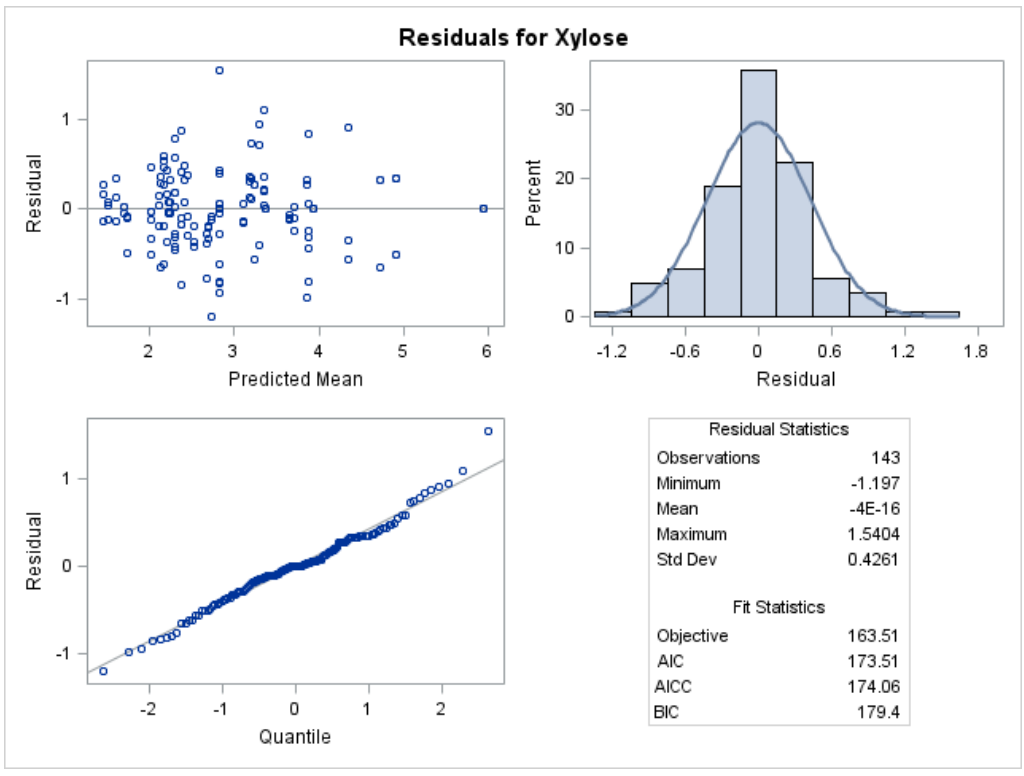


Figure A1.14. Residual plots for the piecewise linear regression mode for xylose.

A.2. SAS codes used

A2.1. SAS Codes for exploratory data analysis

```
/*************************************************************************/
/*************                                                **************/
/*************    EXPLORATORY DATA ANALYSIS PLOTS             **************/
/*************                                                **************/
/*************************************************************************/

Proc Sort data=new.alpha96joint;
by Genotype Stage1;
run;

ods html close;
ods html;
goptions reset=all HBY=2.3;
axis1  label=(a=90 h=1.9 'Viscosity');
axis2  label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=1 i=spline r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot Viscosity*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
quit;


goptions reset=all HBY=2.3;
axis1 label=(a=90 h=1.9 'Lignin');
axis2 label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=0.8 i=j r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot Lignin*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
quit;

goptions reset=all HBY=2.3;
axis1 label=(a=90 h=1.9 'Y_cellulose');
axis2 label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=0.8 i=j r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot Y_cellulose*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
quit;


goptions reset=all HBY=2.3;
axis1 label=(a=90 h=1.9 'a_cellulose');
axis2 label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=0.8 i=j r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot a_cellulose*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
```

```
quit;

goptions reset=all HBY=2.3;
axis1 label=(a=90 h=1.9 'Copper_No');
axis2 label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=0.8 i=j r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot Copper_No*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
quit;

goptions reset=all HBY=2.3;
axis1 label=(a=90 h=1.9 'Glucose');
axis2 label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=0.8 i=j r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot Glucose*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
quit;

goptions reset=all HBY=2.3;
axis1 label=(a=90 h=1.9 'Xylose');
axis2 label=(a=0 h=1.9 'Processing Stage');
symbol1 c=black v=star h=0.8 i=j r=10;
Proc Gplot data=new.alpha96joint;
by Genotype;
plot Xylose*Stage1=Sample / vaxis=axis1 haxis=axis2 nolegend;
symbol i=spline;
run;
quit;
ods html close;


/*************************************************************************/
/*************                                                **************/
/*************              SCORRELATIONS                     **************/
/*************                                                **************/
/*************************************************************************/

ods html;
ods graphics on;
title 'Correlations of Chemical properties';
Proc corr data=new.alpha96jointcorr plots=matrix(histogram); /* Variable
correlations*/
var Viscosity  Lignin a_cellulose Y_cellulose Copper_No Glucose Xylose;
run;
ods graphics off;
ods html close;

/* For stage correlations*/

Data new.Visco;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
60.20370192 55.98911859 51.58001603 58.58270833 34.04086539 34.07328526
67.49817308 67.47512821 69.39730769 43.08333333 29.66121795 27.73903846
71.84974359 67.57455128 59.88583333 52.92621795 37.31679487 30.55602564
```

```
54.1411859  53.97908654 55.1786218  48.95400641 39.22804487 33.13310897
57.99915064 52.62794872 47.19282051 34.69865385 31.45083333 32.11365385
57.99915064 54.64955128 46.23173077 33.43929487 30.92057692 27.77217949
36.01847756 42.14583333 36.4399359  44.73942308 35.82395833 33.58698718
61.735      77.95 58.375      39.01 35.925      35.58
65.48666667 58.445      43.91 43.74 33.355      31.675
58.64754808 46.43057692 46.56314103 36.15685897 30.72173077 27.2750641
59.95211539 57.26769231 36.42198718 50.73891026 38.97384615 36.3225641
64.64990385 52.37807692 45.14524039 66.74091346 53.61211539 45.83081731
30.18290064 34.52716346 30.08564103 40.42758013 32.67923077 29.82628205
30.18290064 34.69865385 33.43929487 42.55307692 29.66121795 28.73326923
30.18290064 36.42198718 28.48572115 40.58615385 32.97625    30.57673077
41.52985577 48.17592949 36.24541667 40.36274039 30.96097756 32.80891026
40.55725962 41.98967949 40.10064103 46.86141026 32.41192308 29.39608974
40.55725962 42.85134615 35.9248718  47.8225     35.09634615 32.0473718
34.04086539 39.58466346 37.12075321 35.46733974 31.47969551 32.4198718
34.04086539 31.12519231 34.313125   38.94076923 41.30600962 35.95850962
34.04086539 37.05543269 31.33086539 36.23274039 33.31903846 27.45735577
60.94778846 59.81658654 40.65471154 46.10504808 42.47149039 45.83081731
60.94778846 42.54004808 46.17360577 35.03298077 24.33798077 26.32615385
51.28115385 84.46307692 63.93004808 26.94317308 23.85807692 23.06966346
;
run;

ods html;
ods graphics on;
title 'Correlations of Viscosity by Stage of processing';
Proc corr data=new.Visco;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;

title 'Panel plots of Viscosity by Stage of processing';
Proc sgscatter data=new.Visco;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;


Data new.Lignin;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
3.55  1.0295      0.44375     0.26625     0.2485      0.19525
3.49675     1.93475     1.065 0.55025     0.355 0.30175
4.70375     1.89925     0.97625     0.58575     0.33725     0.19525
4.98775     2.53825     1.15375     0.639 0.23075     0.213
4.98775     2.53825     1.42  0.94075     0.3905      0.30175
4.31325     2.3785      1.57975     1.04725     0.44375     0.3905
2.556 1.04725     0.781 0.4615      0.19525     0.213
3.053 1.8105      1.08275     0.568 0.33725     0.30175
3.053 1.47325     0.6745      0.426 0.19525     0.19525
4.7215      2.28975     1.136 1.08275     0.69225     0.62125
4.7215      2.11225     1.75725     1.54425     0.76325     0.65675
3.24825     2.272 1.72175     1.562 0.86975     0.72775
3.31925     1.2425      0.65675     0.62125     0.2485      0.213
3.31925     1.278 0.83425     0.65675     0.33725     0.23075
3.31925     1.22475     0.69225     0.639 0.40825     0.30175
2.1655      0.923 0.37275     0.355 0.2485      0.26625
2.53825     1.136 0.7455      0.58575     0.3195      0.26625
2.53825     0.94075     0.40825     0.37275     0.23075     0.213
```

```
4.56175     1.3845      0.568 0.44375     0.2485      0.23075
4.56175     1.68625     1.11825     0.86975     0.47925     0.40825
4.56175     1.5265      0.71  0.6745      0.19525     0.1775
4.473 2.64475     1.75725     1.33125     0.51475     0.55025
4.473 2.7335      1.349 0.7455      0.3905      0.33725
5.02325     2.50275     1.5975      0.6745      0.3195      0.355
;
run;

ods html;
ods graphics on;
title 'Correlations of Lignin by Stage of processing';
Proc corr data=new.Lignin;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;

title 'Panel plots of Lignin by Stage of processing';
Proc sgscatter data=new.Lignin;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;



Data new.Y_cellulose;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
7.444246831 5.832626738 5.955764175 3.119661228 3.245773603 3.179965046
8.19255692  6.163319193 6.236572932 3.60873796  3.755838173 3.631465662
7.840911097 6.360080456 6.496855791 3.439669308 3.331154331 3.392082058
7.256298842 5.513821446 5.717047469 3.419030386 3.388697222 3.394597275
7.256298842 5.066362764 5.657436184 3.354482888 3.443507079 3.157075237
7.256298842 5.104294408 5.231438749 3.050955852 2.949838553 2.965278319
6.198448733 4.845267459 4.995266058 2.688864453 2.604643978 2.68863947
6.198448733 5.297829196 5.527803678 3.051283103 3.137998898 3.17810419
6.198448733 5.135322261 5.420963646 3.101300235 2.966009445 2.970555013
5.961152449 6.772441461 6.902590532 3.641435856 3.695177112 3.629082523
5.961152449 6.677611252 6.64055757  3.734667931 3.509637918 3.666294972
8.764591381 7.142079054 7.006361482 4.095157735 4.017181758 3.972624096
6.767787701 4.928199661 5.037396899 2.795258861 2.837204912 2.842491634
6.767787701 4.907339336 5.056747784 2.966669137 3.156446776 3.097500345
8.197444484 4.869369997 4.92452171  2.840620361 2.898422446 2.82183378
7.552976211 6.024302387 6.024302387 3.2235283   3.245141853 3.286146372
7.552976211 5.605301451 5.794238223 2.836387198 3.097253112 3.204134759
7.552976211 5.869847147 5.721715085 3.051370222 2.938606916 3.051796299
6.907996267 5.669087103 5.62962223  3.473588945 3.242199676 3.181670746
6.907996267 5.377767849 5.488206692 3.100693844 3.340529291 3.269020478
6.907996267 5.715927417 5.616701614 3.290932744 3.028993844 3.223461322
8.655234657 6.61856377  6.668895801 3.562375436 3.865833126 3.74305252
8.655234657 6.347720629 6.898545319 4.081749138 4.285530386 3.944236646
7.774692368 7.282940236 7.664950213 4.966289177 5.003707684 5.131144073
;
run;

ods html;
ods graphics on;
title 'Correlations of Y-cellulose by Stage of processing';
Proc corr data=new.Y_cellulose;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
```

```
run;

title 'Panel plots of Y_cellulose by Stage of processing';
Proc sgscatter data=new.Y_cellulose;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;


Data new.a_cellulose;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;

;
run;

ods html;
ods graphics on;
title 'Correlations of a-cellulose by Stage of processing';
Proc corr data=new.a_cellulose;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;

title 'Panel plots of a-cellulose by Stage of processing';
Proc sgscatter data=new.a_cellulose;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;



Data new.a_cellulose;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
91.2641823  92.79320488 92.66654215 95.82321219 95.41928861 95.31161168
90.43064346 92.71787511 92.56481755 95.27049343 94.92861695 94.95092844
90.72798054 92.16247062 92.16764054 95.28221354 95.27225808 95.1793523
91.3750945  92.86095517 92.73018731 95.3472533  95.09495488 95.04487271
91.3750945  93.51852914 93.15309989 95.51090967 95.15452438 95.47034567
91.3750945  93.52914099 93.31122396 95.82419004 95.59090526 95.58342429
91.89004605 93.12709256 93.05639772 95.90761123 95.73789338 95.52495078
91.89004605 93.08516485 92.79579597 95.69719286 95.49159433 95.15808949
91.89004605 93.22580558 92.97699217 95.42707898 95.78526221 95.68307781
91.64867082 92.10928996 91.95318284 94.93963263 94.8987425  94.80643907
91.64867082 91.55930208 91.40442943 95.0649821  95.08404954 94.6942019
89.2415862  91.00209894 91.06115555 94.12515455 94.15952493 94.03410724
91.69350867 93.44390974 93.00054441 95.7267094  95.67296495 95.61852301
91.69350867 93.76022529 93.54282156 95.96696361 95.64417    95.71901051
90.38069919 93.74327062 93.49397958 96.11705089 96.02973519 96.02916482
90.97547132 92.38342571 92.38342571 95.51565164 95.43856949 95.41233476
90.97547132 92.84822761 92.78332007 96.0037347  95.67298966 95.44342359
90.97547132 92.78341287 92.8184224  95.69141848 95.98355622 95.89515492
91.19360943 92.04570445 92.4941548  94.83605039 94.9077527  94.76984661
91.19360943 93.31393969 93.07854863 95.69825476 95.5812128  95.52675514
91.19360943 92.95134188 93.29850705 95.73282945 95.80475604 95.73745544
89.83018796 92.04589021 92.00510643 95.17754546 95.01780564 95.09080336
```

```
89.83018796 92.36795202 91.9648191  94.39389215 94.31448276 94.91841831
88.55022914 91.73376142 91.48265345 93.30608587 93.17038325 93.0984262
;
run;

ods html;
ods graphics on;
title 'Correlations of a-cellulose by Stage of processing';
Proc corr data=new.a_cellulose;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;

title 'Panel plots of a-cellulose by Stage of processing';
Proc sgscatter data=new.a_cellulose;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;



Data new.CopperNo;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
2.983941911 1.439746162 1.602044899 0.385404032 0.471631117 0.465488988
2.918496301 1.415934273 1.500919547 0.385375263 0.440994771 0.410842556
3.094960704 1.517613466 1.69410782  0.436319898 0.552831599 0.54236378
3.23949014  1.409714064 1.633605022 0.369659175 0.466380699 0.430492125
3.23949014  1.425652766 1.619612431 0.45619008  0.51745626  0.445978997
3.025387516 1.555870339 1.73206398  0.430840892 0.466464579 0.486641273
2.589785893 1.479763665 1.67605045  0.420886048 0.395587361 0.400471409
2.412331478 1.375879884 1.527531262 0.375065715 0.446041596 0.496938659
2.412331478 1.41683946  1.627195164 0.456798036 0.542240169 0.53769944
3.051515713 1.90790007  2.038284048 0.729982906 0.684449655 0.674263708
3.051515713 1.752823865 1.90615615  0.547206586 0.451415902 0.555620402
3.216982869 1.911039489 1.956659415 0.802046689 0.573397998 0.588209487
2.886419312 1.517794143 1.695777037 0.41114236  0.441608383 0.421270736
2.886419312 1.562234923 1.668327374 0.604229768 0.618696296 0.608090593
2.886419312 1.288362597 1.416387819 0.324730456 0.319507712 0.324762842
2.423291866 1.648328671 1.790369608 0.461550248 0.593264072 0.558302742
2.423291866 1.649051606 1.816061316 0.476270221 0.516463216 0.487096858
2.423291866 1.539060941 1.644306426 0.420445532 0.50713806  0.5020881
3.06170101  1.729027392 1.89036876  0.694447907 0.725300481 0.704690907
3.06170101  1.744547717 1.896427522 0.685087713 0.64956904  0.613401958
3.06170101  1.67943229  1.749946227 0.617644673 0.639491791 0.593520574
2.871251851 1.637032345 1.71521764  0.648822393 0.522701938 0.53241955
2.871251851 1.455776037 1.612394929 0.416014205 0.481756943 0.456338439
2.91504259  1.164998947 1.454816464 0.288739178 0.380257653 0.349393659
;
run;

ods html;
ods graphics on;
title 'Correlations of Copper Number by Stage of processing';
Proc corr data=new.CopperNo;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;
```

```
title 'Panel plots of Copper Number by Stage of processing';
Proc sgscatter data=new.CopperNo;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;
```

```
Data new.Glucose;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
91.41643025 93.49966788 93.9551139  95.83705704 96.18244143 96.24978945
90.51375655 92.72447405 92.96597755 95.27653184 94.81662328 95.91248391
88.19480858 92.8189642  93.40436608 95.5821175  95.41330985 95.59941718
89.6185200  93.60185059 93.27640545 95.42465017 95.50085434 95.63013473
89.618520   93.10553488 93.74053164 94.95250166 95.43837379 95.85668841
89.618520   93.4337923  93.60201527 95.01383588 94.61653607 94.75363788
93.69456541 95.21858953 95.93219928 97.06508746 97.18057746 96.22338605
91.83305118 94.60825693 94.98945815 96.48753039 96.59605814 96.27729676
91.83305118 94.66022753 95.93094021 96.41019407 96.42064804 96.39347197
90.17495995 92.03027965 92.40861469 95.54583806 95.4781843  95.80440866
90.17495995 92.33695647 92.5364037  95.01566174 95.38170551 95.42409311
90.08840244 92.775374   93.05792794 95.51676687 95.84630355 95.92018849
92.00859522 94.39896214 94.97272673 96.83877177 97.08850659 96.44690536
92.00859522 94.28884928 94.67431797 96.43121498 96.24706049 96.41175935
92.00859522 94.84849889 95.21881947 96.45590655 96.50347865 96.27617906
89.71206918 93.53102431 93.88900786 95.8808335  96.36433044 95.87611291
89.71206918 92.9215883  93.67574033 95.52413029 95.98776156 95.6658206
89.71206918 93.45143932 93.91396306 95.44730487 95.62775483 96.05787398
89.81251007 92.78364608 92.98963031 95.6796165  95.85869633 95.6793134
89.81251007 92.77554431 93.24778197 95.79514604 96.0960548  95.71141228
89.81251007 93.63946777 93.85067063 95.74400599 95.8158388  95.72971431
91.76843871 92.19385166 93.15202803 94.40185582 94.66289334 94.6600776
91.76843871 93.03737717 93.11693196 95.60825309 95.43535323 96.24078161
93.8543     91.46790392 91.93004408 92.58554837 94.89432911 94.57608046
;
run;

ods html;
ods graphics on;
title 'Correlations of Glucose by Stage of processing';
Proc corr data=new.Glucose;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;

title 'Panel plots of Glucose by Stage of processing';
Proc sgscatter data=new.Glucose;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;
```

```
Data new.Xylose;
Input Stage1      Stage2      Stage3      Stage4      Stage5      Stage6;
cards;
3.787674857 3.469972593 3.314063025 1.915584802 1.985317456 1.936538636
3.987495931 3.741126028 3.93367005  2.397594607 2.637797671 2.161451074
5.243348146 3.612751317 3.538670323 2.296422976 2.436025241 2.26324019
3.927035649 2.958743126 3.252347076 2.160562144 2.194415695 2.367839695
3.927035649 3.168177811 3.219334523 2.11715572  2.31207402  2.159905035
3.927035649 2.969320532 2.816539761 2.34063289  2.568071651 2.835219881
2.694322695 1.868190422 1.988545243 1.244346161 1.322481164 1.398130234
3.519228035 2.478753412 2.480860384 1.644209325 1.721155275 1.541378278
3.519228035 2.477071399 1.988912883 1.633094781 1.623709039 1.58279016
5.037537188 4.183264488 4.230234542 2.644178022 2.701796186 2.48338552
5.037537188 4.131835811 4.009873379 2.616049391 2.748658266 2.417618062
4.065873329 2.883776011 2.889005564 1.530774903 1.547074616 1.474842376
3.366807067 2.766836372 2.742864388 1.519347789 1.472191178 1.643881725
3.366807067 2.872402103 2.901187492 1.895191466 1.962895408 1.721683903
3.366807067 2.551617921 2.498394116 1.69326226  1.742862973 1.730657525
5.938887225 3.553621841 3.482621261 2.361681535 1.836481137 2.295950401
5.938887225 3.547106501 3.308687653 2.497848787 2.377123803 2.1882355
5.938887225 3.592749153 3.532841811 2.484703954 2.63784431  2.194821398
5.254289384 3.647750217 3.573849348 1.888180591 1.848699778 2.198961541
5.254289384 3.449216158 3.706059775 2.023018001 1.986503856 2.196629434
5.254289384 3.562388171 3.557082193 2.21172611  2.324747963 2.259633114
4.398619425 3.937768829 3.555235642 2.885982991 2.871018379 2.786300295
4.398619425 3.074284025 3.402150453 2.001729038 2.016323724 1.519058856
3.117236    4.715413319 4.446638148 4.365089648 3.073931629 3.237424665
;
run;

ods html;
ods graphics on;
title 'Correlations of Xylose by Stage of processing';
Proc corr data=new.Xylose;
var Stage1  Stage2      Stage3      Stage4      Stage5      Stage6;
run;

title 'Panel plots of Xylose by Stage of processing';
Proc sgscatter data=new.Xylose;
matrix Stage1     Stage2      Stage3      Stage4      Stage5      Stage6/
diagonal=(histogram normal) ellipse;
run; quit;
ods graphics off;
ods html close;
```

A2.2. SAS Codes for Random Coefficient Models

```
libname new 'C:\PulpData\SAS FILES\';
data new.Intercepts;
set new.alpha96;
Stage2=Stage1*Stage1;
Stage3=Stage1*Stage1*Stage1;
run;

ods html; /* opens new output content*/
proc mixed data = new.Intercepts covtest;/* to fit the RANDOM COEFFICIENT
model Lignin*/
class Genotype Sample;
model  Lignin = Genotype Genotype*Stage1 Genotype*Stage2/solution;
random intercept Stage1 Stage2 / type=un subject=sample G Gcorr;
ods output solutionf=fixed solutionr=random;
ESTIMATE "Intercept: EDunnii vs Egrandis" Genotype  1 -1  0  0  0  0  0;
ESTIMATE "Intercept: EDunnii vs ESmithii" Genotype  1  0 -1  0  0  0  0;
ESTIMATE "Intercept: EDunnii vs ENitens"  Genotype  1  0  0 -1  0  0  0;
ESTIMATE "Intercept: EDunnii vs GCG"      Genotype  1  0  0  0 -1  0  0;
ESTIMATE "Intercept: EDunnii vs GUA"      Genotype  1  0  0  0  0 -1  0;
ESTIMATE "Intercept: EDunnii vs GUW"      Genotype  1  0  0  0  0  0 -1;

ESTIMATE "Intercept: EGrandis vs ESmithii" Genotype 0  1 -1  0  0  0  0;
ESTIMATE "Intercept: EGrandis vs ENitens"  Genotype 0  1  0 -1  0  0  0;
ESTIMATE "Intercept: EGrandis vs GCG"      Genotype 0  1  0  0 -1  0  0;
ESTIMATE "Intercept: EGrandis vs GUA"      Genotype 0  1  0  0  0 -1  0;
ESTIMATE "Intercept: EGrandis vs GUW"      Genotype 0  1  0  0  0  0 -1;

ESTIMATE "Intercept: Smithii vs ENitens"  Genotype  0  0  1 -1  0  0  0;
ESTIMATE "Intercept: Smithii vs GCG"      Genotype  0  0  1  0 -1  0  0;
ESTIMATE "Intercept: Smithii vs GUA"      Genotype  0  0  1  0  0 -1  0;
ESTIMATE "Intercept: Smithii vs GUW"      Genotype  0  0  1  0  0  0 -1;

ESTIMATE "Intercept: Enitens vs GCG"      Genotype  0  0  0  1 -1  0  0;
ESTIMATE "Intercept: Enitens vs GUA"      Genotype  0  0  0  1  0 -1  0;
ESTIMATE "Intercept: Enitens vs GUW"      Genotype  0  0  0  1  0  0 -1;

ESTIMATE "Intercept: GCG vs GUA"      Genotype      0  0  0  0  1 -1  0;
ESTIMATE "Intercept: GCG vs GUW"      Genotype      0  0  0  0  1  0 -1;

ESTIMATE "Intercept: GUA vs GUW"      Genotype     0  0  0  0  0  1 -1;
/*-----------------------------------------------------------------------
-----*/
ESTIMATE "Slope: EDunnii vs Egrandis" Genotype*Stage1  1 -1  0  0  0  0  0;
ESTIMATE "Slope: EDunnii vs ESmithii" Genotype*Stage1  1  0 -1  0  0  0  0;
ESTIMATE "Slope: EDunnii vs ENitens"  Genotype*Stage1  1  0  0 -1  0  0  0;
ESTIMATE "Slope: EDunnii vs GCG"      Genotype*Stage1  1  0  0  0 -1  0  0;
ESTIMATE "Slope: EDunnii vs GUA"      Genotype*Stage1  1  0  0  0  0 -1  0;
ESTIMATE "Slope: EDunnii vs GUW"      Genotype*Stage1  1  0  0  0  0  0 -1;

ESTIMATE "Slope: EGrandis vs ESmithii" Genotype*Stage1 0  1 -1  0  0  0  0;
ESTIMATE "Slope: EGrandis vs ENitens"  Genotype*Stage1 0  1  0 -1  0  0  0;
ESTIMATE "Slope: EGrandis vs GCG"      Genotype*Stage1 0  1  0  0 -1  0  0;
ESTIMATE "Slope: EGrandis vs GUA"      Genotype*Stage1 0  1  0  0  0 -1  0;
ESTIMATE "Slope: EGrandis vs GUW"      Genotype*Stage1 0  1  0  0  0  0 -1;

ESTIMATE "Slope: Smithii vs ENitens"  Genotype*Stage1  0  0  1 -1  0  0  0;
ESTIMATE "Slope: Smithii vs GCG"      Genotype*Stage1  0  0  1  0 -1  0  0;
ESTIMATE "Slope: Smithii vs GUA"      Genotype*Stage1  0  0  1  0  0 -1  0;
ESTIMATE "Slope: Smithii vs GUW"      Genotype*Stage1  0  0  1  0  0  0 -1;
```

```
ESTIMATE "Slope: Enitens vs GCG"      Genotype*Stage1  0  0  0  1 -1  0  0;
ESTIMATE "Slope: Enitens vs GUA"      Genotype*Stage1  0  0  0  1  0 -1  0;
ESTIMATE "Slope: Enitens vs GUW"      Genotype*Stage1  0  0  0  1  0  0 -1;

ESTIMATE "Slope: GCG vs GUA"      Genotype*Stage1      0  0  0  0  1 -1  0;
ESTIMATE "Slope: GCG vs GUW"      Genotype*Stage1      0  0  0  0  1  0 -1;

ESTIMATE "Slope: GUA vs GUW"      Genotype*Stage1      0  0  0  0  0  1 -1;
/*---------------------------------------------------------------------------
*/
ESTIMATE "Curvature: EDunnii vs Egrandis" Genotype*Stage1 1 -1 0 0 0 0 0;
ESTIMATE "Curvature: EDunnii vs ESmithii" Genotype*Stage1 1 0 -1 0 0 0 0;
ESTIMATE "Curvature: EDunnii vs ENitens"  Genotype*Stage1 1 0 0 -1 0 0 0;
ESTIMATE "Curvature: EDunnii vs GCG"      Genotype*Stage1 1 0 0 0 -1 0 0;
ESTIMATE "Curvature: EDunnii vs GUA"      Genotype*Stage1 1 0 0 0 0 -1 0;
ESTIMATE "Curvature: EDunnii vs GUW"      Genotype*Stage1 1 0 0 0 0 0 -1;

ESTIMATE "Curvature: EGrandis vs ESmithii" Genotype*Stage1 0 1 -1 0 0 0 0;
ESTIMATE "Curvature: EGrandis vs ENitens"  Genotype*Stage1 0 1 0 -1 0 0 0;
ESTIMATE "Curvature: EGrandis vs GCG"      Genotype*Stage1 0 1 0 0 -1 0 0;
ESTIMATE "Curvature: EGrandis vs GUA"      Genotype*Stage1 0 1 0 0 0 -1 0;
ESTIMATE "Curvature: EGrandis vs GUW"      Genotype*Stage1 0 1 0 0 0 0 -1;

ESTIMATE "Curvature: Smithii vs ENitens"  Genotype*Stage1 0 0 1 -1 0 0 0;
ESTIMATE "Curvature: Smithii vs GCG"      Genotype*Stage1 0 0 1 0 -1 0 0;
ESTIMATE "Curvature: Smithii vs GUA"      Genotype*Stage1 0 0 1 0 0 -1 0;
ESTIMATE "Curvature: Smithii vs GUW"      Genotype*Stage1 0 0 1 0 0 0 -1;

ESTIMATE "Curvature: Enitens vs GCG"      Genotype*Stage1 0 0 0 1 -1 0 0;
ESTIMATE "Curvature: Enitens vs GUA"      Genotype*Stage1 0 0 0 1 0 -1 0;
ESTIMATE "Curvature: Enitens vs GUW"      Genotype*Stage1 0 0 0 1 0 0 -1;

ESTIMATE "Curvature: GCG vs GUA"      Genotype*Stage1     0 0 0 0 1 -1 0;
ESTIMATE "Curvature: GCG vs GUW"      Genotype*Stage1     0 0 0 0 1 0 -1;

ESTIMATE "Curvature: GUA vs GUW"      Genotype*Stage1     0 0 0 0 0 1 -1;
run;
ods html close;
```

A2.3. SAS Codes for Piecewise Regression Models

```
libname new 'C:\PulpData\SAS FILES\';
data alpha96;
set new.alpha96;
     if Stage=1 then t1=0;
else if Stage=2 then t1=1;
else if Stage=3 then t1=1;
else if Stage=4 then t1=1;
else if Stage=5 then t1=1;
else if Stage=6 then t1=1;
     if Stage=1 then t2=0;
else if Stage=2 then t2=0;
else if Stage=3 then t2=1;
else if Stage=4 then t2=2;
else if Stage=5 then t2=3;
else if Stage=6 then t2=3;
     if Stage=1 then t3=0;
else if Stage=2 then t3=0;
else if Stage=3 then t3=0;
else if Stage=4 then t3=0;
else if Stage=5 then t3=0;
else if Stage=6 then t3=1;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model viscosity = Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model Lignin = Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr;
Repeated/ subject=Sample;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model a_cellulose= Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr;
Repeated/ subject=Sample;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model Y_cellulose = Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr;
Repeated/ subject=Sample;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model Copper_No = Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr;
Repeated/ subject=Sample;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model Glucose= Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr Repeated/ subject=Sample;
run;

proc mixed data=alpha96 covtest ;/*To fit the PIECEWISE LINEAR REGRESSION MODEL*/
class Genotype BleachCond Sample;
model Xylose= Genotype Genotype*t1 Genotype*t2 Genotype*t3/solution noint;
```

```
RANDOM t1 t2 t3/ subject=Sample(Genotype) type=un gcorr;
Repeated/ subject=Sample;
Run;
```

A2.4. SAS Codes for Kernel Density Estimation

```
libname new 'C:\PulpData\SAS FILES\';
/* To generate bivariate normal Lignin data for the seven genotypes*/

/************************************************************************ ***/
/******                                                                  ******/
/******  To generate bivariate normal Lignin data for the seven genotypes ******/
/******                                                                  ******/
/************************************************************************/
data LigDunnii;
mean1=-2.073; /*mean delignification slope*/
mean2=-0.449; /*mean bleaching slope*/
sig1=0.286;   /*Standard deviation for delignification slope*/
sig2=0.128;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='D';
  z1 = rannor(32794);
  z2 = rannor(55647);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data LigGrandis;
mean1=-2.157; /*mean delignification slope*/
mean2=-0.284; /*mean bleaching slope*/
sig1=0.286;   /*Standard deviation for delignification slope*/
sig2=0.128;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(4774);
  z2 = rannor(687902);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data LigSmithii;
mean1=-2.673; /*mean delignification slope*/
mean2=-0.556; /*mean bleaching slope*/
sig1=0.202;   /*Standard deviation for delignification slope*/
sig2=0.091;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(67231984);
  z2 = rannor(8967451);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data LigNitens;
mean1=-1.520; /*mean delignification slope*/
mean2=-0.227; /*mean bleaching slope*/
```

```
sig1=0.286;   /*Standard deviation for delignification slope*/
sig2=0.128;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(78012);
  z2 = rannor(90847);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data LigGc;
mean1=-2.453; /*mean delignification slope*/
mean2=-0.690; /*mean bleaching slope*/
sig1=0.286;   /*Standard deviation for delignification slope*/
sig2=0.128;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(661128);
  z2 = rannor(975564);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data LigGua;
mean1=-2.467; /*mean delignification slope*/
mean2=-0.428; /*mean bleaching slope*/
sig1=0.286;   /*Standard deviation for delignification slope*/
sig2=0.128;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(569948);
  z2 = rannor(377628);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data LigGuw;
mean1=-1.538; /*mean delignification slope*/
mean2=-0.396; /*mean bleaching slope*/
sig1=0.286;   /*Standard deviation for delignification slope*/
sig2=0.128;   /*Standard deviation for bleaching slope*/
rho=-0.7776;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(786545);
  z2 = rannor(190354);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

/************************************************************************ ***/
/******                                                                  ******/
/******  Kernel density estimation of liginin                            ******/
```

```
/******                                                                      ******/
/*****************************************************************************/

Data KernelLig;
set LigDunnii LigGrandis LigSmithii LigNitens LigGc LigGua LigGuw;
run;

ods graphics on;
goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(-3.5 to -0.7 by
0.5);
axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(-1.0 to 0.2 by
0.1);
symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yeloow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(bottom right inside) mode=protect
             label=(f='arial/bo' h=1.4 "Genotype")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");
ods graphics on;
proc gplot data=KernelLig;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
proc kde data=KernelLig;
bivar (Delignification Bleaching) (Delignification (bwm=2) Bleaching (bwm=2))
/bivstats levels percentiles unistats plots=all;
run;
ods graphics off;


/*****************************************************************************/
/****                                                                     ***/
/**** To generate bivariate normal a-cellulose data for the seven genotypes ***/
/****                                                                     ***/
/*****************************************************************************/
data CelluDunnii;
mean1=0.361; /*mean delignification slope*/
mean2=1.271; /*mean bleaching slope*/
sig1=0.833;   /*Standard deviation for delignification slope*/
sig2=0.286;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='D';
  z1 = rannor(994645);
  z2 = rannor(245635);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CelluGrandis;
mean1=2.074; /*mean delignification slope*/
mean2=0.899; /*mean bleaching slope*/
sig1=0.833;   /*Standard deviation for delignification slope*/
sig2=0.286;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(4587878);
  z2 = rannor(987089898);
  Delignification = mean1+sig1*z1;
```

```
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CelluSmithii;
mean1=0.202; /*mean delignification slope*/
mean2=0.964; /*mean bleaching slope*/
sig1=0.589;   /*Standard deviation for delignification slope*/
sig2=0.202;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(6723688);
  z2 = rannor(98876);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CelluNitens;
mean1=1.393; /*mean delignification slope*/
mean2=1.216; /*mean bleaching slope*/
sig1=0.833;   /*Standard deviation for delignification slope*/
sig2=0.286;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(98564);
  z2 = rannor(9986764);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CelluGc;
mean1=1.663; /*mean delignification slope*/
mean2=0.843; /*mean bleaching slope*/
sig1=0.833;   /*Standard deviation for delignification slope*/
sig2=0.286;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(76998);
  z2 = rannor(20972);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CelluGua;
mean1=1.474; /*mean delignification slope*/
mean2=1.094; /*mean bleaching slope*/
sig1=0.833;   /*Standard deviation for delignification slope*/
sig2=0.286;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(4535356);
  z2 = rannor(904368);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
```

```
run;
data CelluGuw;
mean1=0.923; /*mean delignification slope*/
mean2=1.031; /*mean bleaching slope*/
sig1=0.833;   /*Standard deviation for delignification slope*/
sig2=0.286;   /*Standard deviation for bleaching slope*/
rho=0.001;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(98787);
  z2 = rannor(1890354);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

/************************************************************************** ***/
/******                                                                  ******/
/******  Kernel density estimation of a-cellulose                        ******/
/******                                                                  ******/
/**************************************************************************/
Data KernelCellu;
set CelluDunnii CelluGrandis CelluSmithii CelluNitens CelluGc CelluGua CelluGuw;
run;

goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(0 to 4 by 1.0);

axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(0 to 2 by 0.5);

symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yellow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(bottom right inside) mode=protect
             label=(f='arial/bo' h=1.4 "Genotype")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");

ods graphics on;
proc gplot data=KernelCellu;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
proc kde data=KernelCellu;
bivar (Delignification Bleaching) (Delignification (bwm=0.85) Bleaching (bwm=0.85))
/ bivstats levels percentiles unistats plots=all;
run;
ods graphics off;


/**************************************************************************/
/****                                                                  ***/
/****  To generate bivariate normal viscosity data for the seven genotypes  ***/
/****                                                                  ***/
/**************************************************************************/
data ViscDunnii;
mean1=-10.681; /*mean delignification slope*/
mean2=-2.472; /*mean bleaching slope*/
sig1=10.516;   /*Standard deviation for delignification slope*/
sig2=5.114;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
```

```
  genotype='D';
  z1 = rannor(99461245);
  z2 = rannor(24545635);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data ViscGrandis;
mean1=4.501; /*mean delignification slope*/
mean2=0.019; /*mean bleaching slope*/
sig1=10.516;   /*Standard deviation for delignification slope*/
sig2=5.114;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(89078);
  z2 = rannor(3089898);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data ViscSmithii;
mean1=2.4730; /*mean delignification slope*/
sig1=7.436;   /*Standard deviation for delignification slope*/
mean2=-5.471; /*mean bleaching slope*/
sig2=3.616;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(72368);
  z2 = rannor(98876);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data ViscNitens;
mean1=3.062; /*mean delignification slope*/
sig1=10.516;   /*Standard deviation for delignification slope*/
mean2=-2.696; /*mean bleaching slope*/
sig2=5.114;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(3464);
  z2 = rannor(75764);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data ViscGc;
mean1=-2.143; /*mean delignification slope*/
sig1=10.516;   /*Standard deviation for delignification slope*/
mean2=-7.016; /*mean bleaching slope*/
sig2=5.114;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(5676998);
  z2 = rannor(26790972);
  Delignification = mean1+sig1*z1;
```

```
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data ViscGua;
mean1=0.592; /*mean delignification slope*/
sig1=10.516;   /*Standard deviation for delignification slope*/
mean2=-9.878; /*mean bleaching slope*/
sig2=5.114;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(678956);
  z2 = rannor(453968);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data ViscGuw;
mean1=2.986; /*mean delignification slope*/
sig1=10.516;   /*Standard deviation for delignification slope*/
mean2=-7.718; /*mean bleaching slope*/
sig2=5.114;   /*Standard deviation for bleaching slope*/
rho=-0.9813;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(129087);
  z2 = rannor(998894);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;


/*************************************************************************/
/******                                                            *****/
/******  Kernel density estimation of Viscosity                    *****/
/******                                                            *****/
/*************************************************************************/
Data KernelVisc;
set ViscDunnii ViscGrandis ViscSmithii ViscNitens ViscGc ViscGua ViscGuw;
run;

goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(-33 to 36 by
6);

axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(-26 to 13 by 5);

symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yellow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(bottom left inside) mode=protect
             label=(f='arial/bo' h=1.4 "Genotype")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");

ods graphics on;
```

```
proc gplot data=KernelVisc;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
proc kde data=KernelVisc;
bivar (Delignification Bleaching) (Delignification (bwm=0.85) Bleaching (bwm=0.85))
/ bivstats levels percentiles unistats plots=all;
run;
ods graphics off;


/*****************************************************************************/
/****                                                                     ***/
/**** To generate bivariate normal Y-vellulose data for the seven genotypes ***/
/****                                                                     ***/
/*****************************************************************************/

data GammaDunnii;
mean1=0.283; /*mean delignification slope*/
sig1=0.560;   /*Standard deviation for delignification slope*/
mean2=1.240; /*mean bleaching slope*/
sig2=0.197;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='D';
  z1 = rannor(96235);
  z2 = rannor(21223);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GammaGrandis;
mean1=-2.117; /*mean delignification slope*/
sig1=0.560;   /*Standard deviation for delignification slope*/
mean2=-0.975; /*mean bleaching slope*/
sig2=0.197;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(7889078);
  z2 = rannor(3909898);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GammaSmithii;
mean1=-1.170; /*mean delignification slope*/
sig1=0.396;   /*Standard deviation for delignification slope*/
mean2=-0.970; /*mean bleaching slope*/
sig2=0.140;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(21768);
  z2 = rannor(65476);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GammaNitens;
mean1=-1.446; /*mean delignification slope*/
sig1=0.560;   /*Standard deviation for delignification slope*/
mean2=-1.103; /*mean bleaching slope*/
```

```
sig2=0.197;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(3464);
  z2 = rannor(75764);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GammaGc;
mean1=-1.707; /*mean delignification slope*/
sig1=0.560;   /*Standard deviation for delignification slope*/
mean2=-0.816; /*mean bleaching slope*/
sig2=0.197;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(786998);
  z2 = rannor(90972);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GammaGua;
mean1=-1.401; /*mean delignification slope*/
sig1=0.560;   /*Standard deviation for delignification slope*/
mean2=-1.086; /*mean bleaching slope*/
sig2=0.197;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(56433);
  z2 = rannor(97396);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GammaGuw;
mean1=-0.794; /*mean delignification slope*/
sig1=0.560;   /*Standard deviation for delignification slope*/
mean2=-0.894; /*mean bleaching slope*/
sig2=0.197;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(129087);
  z2 = rannor(998894);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;


/************************************************************************** ***/
/******                                                                   ******/
/******  Kernel density estimation for Y-cellulose                        ******/
/******                                                                   ******/
/*************************************************************************/
Data KernelGamma;
```

```
set GammaDunnii GammaGrandis GammaSmithii GammaNitens GammaGc GammaGua GammaGuw;
run;
goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(-3.5 to 1.5 by
0.5);
axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(-2.0 to 2.0 by
0.5);
symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yellow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(top left inside) mode=protect
             label=(f='arial/bo' h=1.4 "")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");
ods graphics on;
proc gplot data=KernelGamma;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
proc kde data=KernelGamma;
bivar (Delignification Bleaching) (Delignification (bwm=0.85) Bleaching (bwm=0.85))
/ bivstats levels percentiles unistats plots=all;
run;
ods graphics off;


/*************************************************************************/
/****                                                                 ***/
/**** To generate bivariate normal Copper Number data for the seven genotypes**/
/****                                                                 ***/
/*************************************************************************/
data CopperDunnii;
mean1=-1.064; /*mean delignification slope*/
sig1=0.241;   /*Standard deviation for delignification slope*/
mean2=-0.514; /*mean bleaching slope*/
sig2=0.083;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='D';
  z1 = rannor(239035);
  z2 = rannor(56721223);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CopperGrandis;
mean1=-1.277; /*mean delignification slope*/
sig1=0.241;   /*Standard deviation for delignification slope*/
mean2=-0.414; /*mean bleaching slope*/
sig2=0.083;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(934078);
  z2 = rannor(450781);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CopperSmithii;
```

```
mean1=-1.245; /*mean delignification slope*/
sig1=0.170;   /*Standard deviation for delignification slope*/
mean2=-0.417; /*mean bleaching slope*/
sig2=0.058;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(78848);
  z2 = rannor(76236);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CopperNitens;
mean1=-0.657; /*mean delignification slope*/
sig1=0.241;   /*Standard deviation for delignification slope*/
mean2=-0.452; /*mean bleaching slope*/
sig2=0.083;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(175658871);
  z2 = rannor(2005612);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CopperGc;
mean1=-1.534; /*mean delignification slope*/
sig1=0.241;   /*Standard deviation for delignification slope*/
mean2=-0.418; /*mean bleaching slope*/
sig2=0.083;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(231675);
  z2 = rannor(2312267);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CopperGua;
mean1=-1.397; /*mean delignification slope*/
sig1=0.241;   /*Standard deviation for delignification slope*/
mean2=-0.410; /*mean bleaching slope*/
sig2=0.083;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(12564);
  z2 = rannor(349731);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data CopperGuw;
mean1=-0.881; /*mean delignification slope*/
sig1=0.241;   /*Standard deviation for delignification slope*/
mean2=-0.408; /*mean bleaching slope*/
sig2=0.083;   /*Standard deviation for bleaching slope*/
```

```
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(34786);
  z2 = rannor(76588987);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

/*************************************************************************** ***/
/******                                                                    ******/
/******  Kernel density estimation for Copper number                       ******/
/******                                                                    ******/
/*************************************************************************** ***/
Data KernelCopper;
set CopperDunnii CopperGrandis CopperSmithii CopperNitens CopperGc CopperGua
CopperGuw;
run;

goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(-2.0 to -0.14
by 0.2);
axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(-0.7 to -0.15 by
0.1);
symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yellow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(bottom left inside) mode=protect
             label=(f='arial/bo' h=1.4 "")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");

ods graphics on;
proc gplot data=KernelCopper;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
proc kde data=KernelCopper;
bivar (Delignification Bleaching) (Delignification (bwm=0.85) Bleaching (bwm=0.85))
/ bivstats levels percentiles unistats plots=all;
run;
ods graphics off;



/*************************************************************************** ***/
/****                                                                      ***/
/**** To generate bivariate normal Glucose data for the seven genotypes    ***/
/****                                                                      ***/
/*************************************************************************** ***/
ods html close;/* closes previous output content*/
ods html; /* opens new output content*/

data GlucDunnii;
mean1=2.010; /*mean delignification slope*/
sig1=0.461;   /*Standard deviation for delignification slope*/
mean2=1.226; /*mean bleaching slope*/
sig2=0.157;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='D';
```

```
  z1 = rannor(6710001);
  z2 = rannor(6474646);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GlucGrandis;
mean1=2.467; /*mean delignification slope*/
sig1=0.461;   /*Standard deviation for delignification slope*/
mean2=0.792; /*mean bleaching slope*/
sig2=0.157;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(67673);
  z2 = rannor(8907765);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GlucSmithii;
mean1=1.884; /*mean delignification slope*/
sig1=0.345;   /*Standard deviation for delignification slope*/
mean2=1.035; /*mean bleaching slope*/
sig2=0.111;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(312676);
  z2 = rannor(18734);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GlucNitens;
mean1=3.493; /*mean delignification slope*/
sig1=0.461;   /*Standard deviation for delignification slope*/
mean2=0.987; /*mean bleaching slope*/
sig2=0.157;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(4387653);
  z2 = rannor(56435367);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GlucGc;
mean1=3.640; /*mean delignification slope*/
sig1=0.461;   /*Standard deviation for delignification slope*/
mean2=0.701; /*mean bleaching slope*/
sig2=0.157;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(785665);
  z2 = rannor(8876564);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
```

```
  output;
end;
keep genotype Delignification Bleaching;
run;
data GlucGua;
mean1=2.980; /*mean delignification slope*/
sig1=0.461;   /*Standard deviation for delignification slope*/
mean2=0.949; /*mean bleaching slope*/
sig2=0.157;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(675643);
  z2 = rannor(877652);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data GlucGuw;
mean1=2.493; /*mean delignification slope*/
sig1=0.461;   /*Standard deviation for delignification slope*/
mean2=0.675; /*mean bleaching slope*/
sig2=0.157;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(18875);
  z2 = rannor(5443548);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

/************************************************************************** ***/
/******                                                                  ******/
/******  Kernel density estimation for Glucose                           ******/
/******                                                                  ******/
/*************************************************************************/
Data KernelGluc;
set GlucDunnii GlucGrandis GlucSmithii GlucNitens GlucGc GlucGua GlucGuw;
run;
goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(1.13 to 4.3 by
0.5);
axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(0.3 to 1.71 by
0.1);
symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yellow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(top right inside) mode=protect
             label=(f='arial/bo' h=1.4 "Genotype")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");

ods graphics on;
proc gplot data=KernelGluc;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
```

```
ods html close;/* closes previous output content*/
ods html; /* opens new output content*/
ods graphics on;
proc kde data=KernelGluc out=Gluc;
bivar (Delignification Bleaching) (Delignification (bwm=0.85) Bleaching (bwm=0.85))
/ bivstats levels percentiles unistats plots=surface (rotate=-30);
run;
ods graphics off;


/*****************************************************************************/
/****                                                                     ***/
/**** To generate bivariate normal Xylose data for the seven genotypes    ***/
/****                                                                     ***/
/*****************************************************************************/
ods html close;/* closes previous output content*/
ods html; /* opens new output content*/

data XyloDunnii;
mean1=-0.857; /*mean delignification slope*/
sig1=0.322;   /*Standard deviation for delignification slope*/
mean2=-0.565; /*mean bleaching slope*/
sig2=0.096;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='D';
  z1 = rannor(986753);
  z2 = rannor(62368);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data XyloGrandis;
mean1=-0.545; /*mean delignification slope*/
sig1=0.322;   /*Standard deviation for delignification slope*/
mean2=-0.402; /*mean bleaching slope*/
sig2=0.096;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='G';
  z1 = rannor(543487);
  z2 = rannor(2123117);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data XyloSmithii;
mean1=-1.032; /*mean delignification slope*/
sig1=0.237;   /*Standard deviation for delignification slope*/
mean2=-0.528; /*mean bleaching slope*/
sig2=0.068;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='S';
  z1 = rannor(342980908);
  z2 = rannor(45986721);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data XyloNitens;
mean1=-2.279; /*mean delignification slope*/
```

```
sig1=0.322;   /*Standard deviation for delignification slope*/
mean2=-0.484; /*mean bleaching slope*/
sig2=0.096;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='N';
  z1 = rannor(16009876);
  z2 = rannor(897867);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data XyloGc;
mean1=-0.817; /*mean delignification slope*/
sig1=0.322;   /*Standard deviation for delignification slope*/
mean2=-0.291; /*mean bleaching slope*/
sig2=0.096;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='C';
  z1 = rannor(4568787);
  z2 = rannor(723987);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;
data XyloGua;
mean1=-0.626; /*mean delignification slope*/
sig1=0.322;   /*Standard deviation for delignification slope*/
mean2=-0.516; /*mean bleaching slope*/
sig2=0.096;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='A';
  z1 = rannor(543388);
  z2 = rannor(8243456);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

data XyloGuw;
mean1=-0.951; /*mean delignification slope*/
sig1=0.322;   /*Standard deviation for delignification slope*/
mean2=-0.280; /*mean bleaching slope*/
sig2=0.096;   /*Standard deviation for bleaching slope*/
rho=0.000;  /*correlation between delignification and bleaching*/
do i=1 to 50;
  genotype='W';
  z1 = rannor(64522341);
  z2 = rannor(8772237);
  Delignification = mean1+sig1*z1;
  Bleaching = mean2+rho*sig2*z1+sqrt(sig2**2-sig2**2*rho**2)*z2;
  output;
end;
keep genotype Delignification Bleaching;
run;

/*********************************************************************** ***/
/******                                                                 ******/
/******  Kernel density estimation for Glucose                          ******/
/******                                                                 ******/
```

```
/*****************************************************************************/
Data KernelXylo;
set XyloDunnii XyloGrandis XyloSmithii XyloNitens XyloGc XyloGua XyloGuw;
run;
goptions reset=global;
axis1 label=(f='arial/bo' h=1.5 "Delignification" justify=c) order=(-3.1 to 0.1 by
0.5);
axis2 label=(a=90 f='arial/bo' h=1.5 "Bleaching" justify=c)order=(-0.82 to -0.08 by
0.1);
symbol1 value='A' colour=black interpol=none h=1.2;
symbol2 value='C' colour=red interpol=none h=1.2;
symbol3 value='D' colour=blue interpol=none h=1.2;
symbol4 value='G' colour=Green interpol=none h=1.2;
symbol5 value='N' colour=yellow interpol=none h=1.2;
symbol6 value='S' colour=orange interpol=none h=1.2;
symbol7 value='W' colour=green interpol=none h=1.2;
legend1 across=1 down=2 noframe
        position=(bottom right inside) mode=protect
             label=(f='arial/bo' h=1.4 "Genotype")
             value=(f='arial/bo' h=1.4 "EguA" "Egc" "Edunnii" "Egrandis" "Enitens"
"Esmithii" "EguW");

ods graphics on;
proc gplot data=KernelXylo;
plot Bleaching*Delignification=genotype/haxis=axis1 vaxis=axis2 legend=legend1;
run;
ods html close;/* closes previous output content*/
ods html; /* opens new output content*/
ods graphics on;
proc kde data=KernelXylo;
bivar (Delignification Bleaching) (Delignification (bwm=0.85) Bleaching (bwm=0.85))
/ bivstats levels percentiles unistats plots=surface (rotate=-30);
run;
ods graphics off;
```

A2.5. SAS Codes for Joint Modelling

```
libname new 'C:\PulpData\SAS FILES\';
/*********************************************************************/
/*********************************************************************/
/*STEP 1: To Create all Possible Bivariate Normal pairs for Pairwise */
/* fitting and to calculate the parameter estimates using Proc MIXED.*/
/*         Since there are 7 variables there will be 21 pairs        */
/*********************************************************************/
/*********************************************************************/

data new.alpha96jointcorr; /* To create Intercept Corrected Data*/
set new.alpha96joint;
if Genotype="EDunnii" then      Viscosity=Viscosity-62.165;
if Genotype="EDunnii" then             Lignin=Lignin-4.036;
if Genotype="EDunnii" then  a_cellulose=a_cellulose-89.865;
if Genotype="EDunnii" then   Y_cellulose=Y_cellulose-8.131;
if Genotype="EDunnii" then       Copper_No=Copper_No-3.231;
if Genotype="EDunnii" then          Glucose=Glucose-89.629;
if Genotype="EDunnii" then             Xylose=Xylose-5.005;

if Genotype="EGrandis" then     Viscosity=Viscosity-33.340;
if Genotype="EGrandis" then            Lignin=Lignin-2.905;
if Genotype="EGrandis" then a_cellulose=a_cellulose-91.128;
if Genotype="EGrandis" then  Y_cellulose=Y_cellulose-7.274;
if Genotype="EGrandis" then      Copper_No=Copper_No-2.847;
if Genotype="EGrandis" then         Glucose=Glucose-92.197;
if Genotype="EGrandis" then            Xylose=Xylose-3.560;

if Genotype="ESmithii" then      Viscosity=Viscosity-52.212;
if Genotype="ESmithii" then             Lignin=Lignin-4.249;
if Genotype="ESmithii" then  a_cellulose=a_cellulose-91.136;
if Genotype="ESmithii" then   Y_cellulose=Y_cellulose-8.150;
if Genotype="ESmithii" then       Copper_No=Copper_No-2.954;
if Genotype="ESmithii" then          Glucose=Glucose-90.317;
if Genotype="ESmithii" then             Xylose=Xylose-5.085;

if Genotype="Enitens" then      Viscosity=Viscosity-46.078;
if Genotype="Enitens" then             Lignin=Lignin-2.123;
if Genotype="Enitens" then  a_cellulose=a_cellulose-90.368;
if Genotype="Enitens" then   Y_cellulose=Y_cellulose-8.046;
if Genotype="Enitens" then       Copper_No=Copper_No-2.621;
if Genotype="Enitens" then          Glucose=Glucose-89.989;
if Genotype="Enitens" then             Xylose=Xylose-5.657;

if Genotype="GCG438" then      Viscosity=Viscosity-63.853;
if Genotype="GCG438" then             Lignin=Lignin-4.615;
if Genotype="GCG438" then  a_cellulose=a_cellulose-91.153;
if Genotype="GCG438" then   Y_cellulose=Y_cellulose-7.480;
if Genotype="GCG438" then       Copper_No=Copper_No-3.050;
if Genotype="GCG438" then          Glucose=Glucose-90.113;
if Genotype="GCG438" then             Xylose=Xylose-3.873;

if Genotype="GUA380" then      Viscosity=Viscosity-78.821;
if Genotype="GUA380" then             Lignin=Lignin-3.501;
if Genotype="GUA380" then  a_cellulose=a_cellulose-90.344;
if Genotype="GUA380" then   Y_cellulose=Y_cellulose-8.367;
if Genotype="GUA380" then       Copper_No=Copper_No-2.910;
if Genotype="GUA380" then          Glucose=Glucose-90.020;
if Genotype="GUA380" then             Xylose=Xylose-4.662;

if Genotype="GUW962" then      Viscosity=Viscosity-63.337;
if Genotype="GUW962" then             Lignin=Lignin-2.745;
if Genotype="GUW962" then  a_cellulose=a_cellulose-91.317;
if Genotype="GUW962" then   Y_cellulose=Y_cellulose-6.754;
if Genotype="GUW962" then       Copper_No=Copper_No-2.549;
if Genotype="GUW962" then          Glucose=Glucose-92.834;
```

```
if Genotype="GUW962" then             Xylose=Xylose-3.189;
keep Genotype BleachCond Tree Sample Stage Stage1 t1 t2 t3 Viscosity Lignin
Y_cellulose a_cellulose Copper_No Glucose Xylose;
output;
run;
ods html close;/* closes previous output content*/
ods html; /* opens new output content*/
/*********************************************************************/
data new.PairVisLig;   /*To create Joint model Data 1 for Viscosity and
             Lignin Pair */
set  new.alpha96jointcorr;
Y12=Viscosity;
outcomenum=1; /* Outcomenum=1 is Viscosity, 2 is Lignin*/
timeVis=1;
timeLig=0;
timeVis2=1;
timeLig2=0;
interceptVis=1;
interceptLig=0;
interceptVis2=1;
interceptLig2=0;
XVis=Genotyp;   /* XVis is Indicator for each Genotype's Viscosity slope*/
XLig=0;         /* XLig is Indicator for each Genotype's Lignin slope*/
output;
Y12=Lignin;
outcomenum=2;
timeVis=0;
timeLig=1;
interceptVis=0;
interceptLig=1;
XVis=0;
XLig=Genotyp;
keep Genotype BleachCond Sample Tree Stage1 t1 t2 Outcomenum timeVis timeLig
interceptVis interceptLig XVis XLig Y12;
output;
run;

Proc Mixed Data= new.PairVisLig covtest noclprint; /*To fit the joint     model for
the new variable Y12*/
Class Sample Genotype Stage1 outcomenum;
model Y12= timeVis*Genotype timeLig*Genotype/noint solution outpm=resid_1;
random interceptVis interceptLig timeVis timeLig/ subject=Sample type=un;
repeated outcomenum/subject=Sample*Stage1 type=un;
ods output covparms=cov_1 solutionF=Fixed_1;
run;
/*********************************************************************/
data new.PairVisGam; /* To create Joint model Data 2 for Viscosity and Y_cellulose
Pair */
set  new.alpha96jointcorr;
Y13=Viscosity;
outcomenum=1; /* Outcomenum=1 is Viscosity, 3 is Y_cellulose*/
timeVis=1;
timeGam=0;
interceptVis=1;
interceptGam=0;
output;
Y13=Y_cellulose;
outcomenum=3;
timeVis=0;
timeGam=1;
interceptVis=0;
interceptGam=1;
keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeVis timeGam
interceptVis interceptGam XVis XGam Y13;
output;
run;

Proc Mixed Data= new.PairVisGam  covtest METHOD=REML;
```

```
Class Sample Genotype Stage1 outcomenum;
model Y13= timeVis*Genotype timeGam*Genotype/noint solution outpm=resid_2;
random interceptVis interceptGam timeVis timeGam/ subject=Sample type=un;
repeated outcomenum/subject=Sample*Stage1 type=un;
ods output covparms=cov_2 solutionF=Fixed_2;
run;
/*****************************************************************/
data new.PairVisAlpha; /* To create Joint model Data 3 for Viscosity and
a_cellulose Pair */
set  new.alpha96jointcorr;
Y14=Viscosity;
outcomenum=1; /* Outcomenum=1 is Viscosity, 4 is a_cellulose*/
timeVis=1;
timeAlpha=0;
interceptVis=1;
interceptAlpha=0;
output;
Y14=a_cellulose;
outcomenum=4;
timeVis=0;
timeAlpha=1;
interceptVis=0;
interceptAlpha=1;
keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeVis timeAlpha
interceptVis interceptAlpha XVis XAlpha Y14;
output;
run;

Proc Mixed Data= new.PairVisAlpha; /* covtest METHOD=REML;*/
Class Sample Genotype Stage1 outcomenum;
model Y14= timeVis*Genotype timeAlpha*Genotype/ noint solution outpm=resid_3;
random interceptVis interceptAlpha timeVis timeAlpha/ subject=Sample type=un;
repeated outcomenum/subject=Sample*Stage1 type=un;
ods output covparms=cov_3 solutionF=Fixed_3;
run;
/*****************************************************************/
data new.PairVisCopp; /*To create Joint model Data 4 for Viscosity and Copper
Number Pair */
set  new.alpha96jointcorr;
Y15=Viscosity;
outcomenum=1; /* Outcomenum=1 is Viscosity, 5 is Copper Number*/
timeVis=1;
timeCopp=0;
interceptVis=1;
interceptCopp=0;
output;
Y15=Copper_No;
outcomenum=5;
timeVis=0;
timeCopp=1;
interceptVis=0;
interceptCopp=1;
keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeVis timeCopp
interceptVis interceptCopp XVis XCopp Y15;
output;
run;

Proc Mixed Data= new.PairVisCopp;  METHOD=REML;
Class Sample Genotype Stage1 outcomenum;
model Y15= timeVis*Genotype timeCopp*Genotype/ noint solution outpm=resid_4;
random interceptVis interceptCopp timeVis timeCopp/ subject=Sample type=un;
repeated outcomenum/subject=Sample*Stage1 type=un;
ods output covparms=cov_4 solutionF=Fixed_4;
run;
/*****************************************************************/
data new.PairVisGluc; /*To create Joint model Data 5 for Viscosity and Glucose Pair
*/
set  new.alpha96jointcorr;
```

```
Y16=Viscosity;
outcomenum=1; /* Outcomenum=1 is Viscosity, 6 Glucose*/
timeVis=1;
timeGluc=0;
interceptVis=1;
interceptGluc=0;
output;
Y16=Glucose;
outcomenum=6;
timeVis=0;
timeGluc=1;
interceptVis=0;
interceptGluc=1;
keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeVis timeGluc
interceptVis interceptGluc XVis XGluc Y16;
output;
run;

Proc Mixed Data= new.PairVisGluc;  /*METHOD=REML8*/;
     Class Sample Genotype Stage1 outcomenum;
        model Y16= timeVis*Genotype timeGluc*Genotype/ noint solution
outpm=resid_5;
        random interceptVis interceptGluc timeVis timeGluc/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_5 solutionF=Fixed_5;
run;
/*****************************************************************/
data new.PairVisXylo; /* Data 6 for Viscosity and Xylose Pair */
set  new.alpha96jointcorr;
  Y17=Viscosity;
  outcomenum=1; /* Outcomenum=1 is Viscosity, 7 Xylose*/
  timeVis=1;
  timeXylo=0;
  interceptVis=1;
  interceptXylo=0;
  output;
  Y17=Xylose;
  outcomenum=7;
  timeVis=0;
  timeXylo=1;
  interceptVis=0;
  interceptXylo=1;
keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeVis timeXylo
interceptVis interceptXylo XVis XXylo Y17;
output;
run;

Proc Mixed Data= new.PairVisXylo; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y17= timeVis*Genotype timeXylo*Genotype/ noint solution
outpm=resid_6;
        random interceptVis interceptXylo timeVis timeXylo/
subject=Sample(Genotype) type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_6 solutionF=Fixed_6;
run;
/*****************************************************************/
data new.PairLigGam; /* Data 7 for Lignin and Y-Cellilose Pair */
set  new.alpha96jointcorr;
  Y23=Lignin;
  outcomenum=2; /* Outcomenum=2 is Lignin, 3 Y-Cellilose*/
  timeLig=1;
  timeGam=0;
  interceptLig=1;
  interceptGam=0;
  output;
  Y23=Y_cellulose;
  outcomenum=3;
```

```
  timeLig=0;
  timeGam=1;
  interceptLig=0;
  interceptGam=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeLig timeGam
interceptLig interceptGam XLig XGam Y23;
output;
run;

Proc Mixed Data= new.PairLigGam; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y23= timeLig*Genotype timeGam*Genotype/ noint solution outpm=resid_7;
        random interceptLig interceptGam timeLig timeGam/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_7 solutionF=Fixed_7;
run;
/*************************************************************************/
data new.PairLigAlpha; /* Data 8 for Lignin and a-Cellilose Pair */
set  new.alpha96jointcorr;
  Y24=Lignin;
  outcomenum=2; /* Outcomenum=2 is Lignin, 4 a-Cellilose*/
  timeLig=1;
  timeAlpha=0;
  interceptLig=1;
  interceptAlpha=0;
  output;
  Y24=a_cellulose;
  outcomenum=4;
  timeLig=0;
  timeAlpha=1;
  interceptLig=0;
  interceptAlpha=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeLig timeAlpha
interceptLig interceptAlpha XLig XAlpha Y24;
output;
run;

Proc Mixed Data= new.PairLigAlpha; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y24=timeLig*Genotype timeAlpha*Genotype/ noint solution
outpm=resid_8;
        random interceptLig interceptAlpha timeLig timeAlpha/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_8 solutionF=Fixed_8;
run;
/*************************************************************************/
data new.PairLigCopp; /* Data 9 for Lignin and copper number pair*/
set  new.alpha96jointcorr;
  Y25=Lignin;
  outcomenum=2; /* Outcomenum=2 is Lignin, 5 copper number*/
  timeLig=1;
  timeCopp=0;
  interceptLig=1;
  interceptCopp=0;
  output;
  Y25=Copper_No;
  outcomenum=5;
  timeLig=0;
  timeCopp=1;
  interceptLig=0;
  interceptCopp=1;
 keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeLig timeCopp
interceptLig interceptCopp XLig XCopp Y25;
output;
run;
Proc Mixed Data= new.PairLigCopp; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
```

```
        model Y25= timeLig*Genotype timeCopp*Genotype/ noint solution
outpm=resid_9;
        random interceptLig interceptCopp timeLig timeCopp/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_9 solutionF=Fixed_9;
run;
/*****************************************************************************/
data new.PairLigGluc; /* Data 10 for Lignin and Glucose pair*/
set  new.alpha96jointcorr;
  Y26=Lignin;
  outcomenum=2; /* Outcomenum=2 is Lignin, 6 Glucose*/
  timeLig=1;
  timeGluc=0;
  interceptLig=1;
  interceptGluc=0;
  output;
  Y26=Glucose;
  outcomenum=6;
  timeLig=0;
  timeGluc=1;
  interceptLig=0;
  interceptGluc=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeLig timeGluc
interceptLig interceptGluc XLig XGluc Y26;
output;
run;

Proc Mixed Data= new.PairLigGluc; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y26= timeLig*Genotype timeGluc*Genotype/ noint solution
outpm=resid_10;
        random interceptLig interceptGluc timeLig timeGluc/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_10 solutionF=Fixed_10;
run;
/*****************************************************************************/
data new.PairLigXylo; /* Data 11 for Lignin and Xylose pair*/
set  new.alpha96jointcorr;
  Y27=Lignin;
  outcomenum=2; /* Outcomenum=2 is Lignin, 7 Xylose*/
  timeLig=1;
  timeXylo=0;
  interceptLig=1;
  interceptXylo=0;
  output;
  Y27=Xylose;
  outcomenum=7;
  timeLig=0;
  timeXylo=1;
  interceptLig=0;
  interceptXylo=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeLig timeXylo
interceptLig interceptXylo XLig XXylo Y27;
output;
run;
Proc Mixed Data= new.PairLigXylo; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y27=timeLig*Genotype timeXylo*Genotype/ noint solution
outpm=resid_11;
        random interceptLig interceptXylo timeLig timeXylo/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_11 solutionF=Fixed_11;
run;
/*****************************************************************/
data new.PairGamAlpha; /* Data 12 for Y-Cellulose and a-Cellulose pair*/
set  new.alpha96jointcorr;
  Y34=Y_cellulose;
  outcomenum=3; /* Outcomenum=3 is Y-Cellulose, 4 a-Cellulose*/
```

```
  timeGam=1;
  timeAlpha=0;
  interceptGam=1;
  interceptAlpha=0;
  output;
  Y34=a_cellulose;
  outcomenum=4;
  timeGam=0;
  timeAlpha=1;
  interceptGam=0;
  interceptAlpha=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeGam timeAlpha
interceptGam interceptAlpha XGam XAlpha Y34;
output;
run;


Proc Mixed Data= new.PairGamAlpha; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y34=timeGam*Genotype timeAlpha*Genotype/ noint solution
outpm=resid_12;
        random interceptGam interceptAlpha timeGam timeAlpha/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_12 solutionF=Fixed_12;
run;
/*****************************************************************/
data new.PairGamCopp; /* Data 13 for Y-Cellulose and Copper Number pair*/
set  new.alpha96jointcorr;
  Y35=Y_cellulose;
  outcomenum=3; /* Outcomenum=3 is Y-Cellulose, 5 Copper Number*/
  timeGam=1;
  timeCopp=0;
  interceptGam=1;
  interceptCopp=0;
  output;
  Y35=Copper_No;
  outcomenum=5;
  timeGam=0;
  timeCopp=1;
  interceptGam=0;
  interceptCopp=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeGam timeCopp
interceptGam interceptCopp XGam XCopp Y35;
output;
run;

Proc Mixed Data= new.PairGamCopp; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y35= timeGam*Genotype timeCopp*Genotype/ noint solution
outpm=resid_13;
        random interceptGam interceptCopp timeGam timeCopp/ subject=Sample type=UN;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_13 solutionF=Fixed_13;
run;
/*****************************************************************/
data new.PairGamGluc; /* Data 14 for Y-Cellulose and Glucose pair*/
set  new.alpha96jointcorr;
  Y36=Y_cellulose;
  outcomenum=3; /* Outcomenum=3 is Y-Cellulose, 6 Glucose*/
  timeGam=1;
  timeGluc=0;
  interceptGam=1;
  interceptGluc=0;
  output;
  Y36=Glucose;
  outcomenum=6;
  timeGam=0;
```

```
  timeGluc=1;
  interceptGam=0;
  interceptGluc=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeGam timeGluc
interceptGam interceptGluc XGam XGluc Y36;
output;
run;

Proc Mixed Data= new.PairGamGluc; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y36=timeGam*Genotype timeGluc*Genotype/ noint solution
outpm=resid_14;
        random interceptGam interceptGluc timeGam timeGluc/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_14 solutionF=Fixed_14;
run;
/*************************************************************************/
data new.PairGamXylo; /* Data 15 for Y-Cellulose and Xylose pair*/
set  new.alpha96jointcorr;
  Y37=Y_cellulose;
  outcomenum=3; /* Outcomenum=3 is Y-Cellulose, 7 Xylose*/
  timeGam=1;
  timeXylo=0;
  interceptGam=1;
  interceptXylo=0;
  output;
  Y37=Xylose;
  outcomenum=7;
  timeGam=0;
  timeXylo=1;
  interceptGam=0;
  interceptXylo=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeGam timeXylo
interceptGam interceptXylo XGam XXylo Y37;
output;
run;

Proc Mixed Data= new.PairGamXylo; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y37= timeGam*Genotype timeXylo*Genotype/ noint solution
outpm=resid_15;
        random interceptGam interceptXylo timeGam timeXylo/ subject=Sample type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_15 solutionF=Fixed_15;
run;
/*************************************************************************/
data new.PairAlphaCopp; /* Data 16 for a-Cellulose and Copper Number*/
set  new.alpha96jointcorr;
  Y45=a_cellulose;
  outcomenum=4; /* Outcomenum=4 is a-Cellulose, 5 Copper Number*/
  timeAlpha=1;
  timeCopp=0;
  interceptAlpha=1;
  interceptCopp=0;
  XAlpha=Genotyp;
  XCopp=0;
output;
  Y45=Copper_No;
  outcomenum=5;
  timeAlpha=0;
  timeCopp=1;
  interceptAlpha=0;
  interceptCopp=1;
  XAlpha=0;
  XCopp=Genotyp;
keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeAlpha timeCopp
interceptAlpha interceptCopp XAlpha XCopp Y45;
output;
```

```
run;

 Proc Mixed Data= new.PairAlphaCopp; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y45= timeAlpha*Genotype timeCopp*Genotype/ noint solution
outpm=resid_16;
        random interceptAlpha interceptCopp timeAlpha timeCopp/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_16 solutionF=Fixed_16;
run;
/*********************************************************************/
data new.PairAlphaGluc; /* Data 17 for a-Cellulose and Glucose*/
set  new.alpha96jointcorr;
  Y46=a_cellulose;
  outcomenum=4; /* Outcomenum=4 is a-Cellulose, 6 Glucose*/
  timeAlpha=1;
  timeGluc=0;
  interceptAlpha=1;
  interceptGluc=0;
  output;
  Y46=Glucose;
  outcomenum=6;
  timeAlpha=0;
  timeGluc=1;
  interceptAlpha=0;
  interceptGluc=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeAlpha timeGluc
interceptAlpha interceptGluc XAlpha XGluc Y46;
output;
run;
Proc Mixed Data= new.PairAlphaGluc; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y46= timeAlpha*Genotype timeGluc*Genotype/ noint solution
outpm=resid_17;
        random interceptAlpha interceptGluc timeAlpha timeGluc/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_17 solutionF=Fixed_17;
run;
/*********************************************************************************/
data new.PairAlphaXylo; /* Data 18 for a-Cellulose and Xylose*/
set  new.alpha96jointcorr;
  Y47=a_cellulose;
  outcomenum=4; /* Outcomenum=4 is a-Cellulose, 7 Xylose*/
  timeAlpha=1;
  timeXylo=0;
  interceptAlpha=1;
  interceptXylo=0;
  output;
  Y47=Xylose;
  outcomenum=7;
  timeAlpha=0;
  timeXylo=1;
  interceptAlpha=0;
  interceptXylo=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeAlpha timeXylo
interceptAlpha interceptXylo XAlpha XXylo Y47;
output;
run;
Proc Mixed Data= new.PairAlphaXylo; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y47= timeAlpha*Genotype timeXylo*Genotype/ noint solution
outpm=resid_18;
        random interceptAlpha interceptXylo timeAlpha timeXylo/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_18 solutionF=Fixed_18;
```

```
run;
/*****************************************************************************/
data new.PairCoppGluc; /* Data 19 for Copper Number and Glucose Pair*/
set  new.alpha96jointcorr;
  Y56=Copper_No;
  outcomenum=5; /* Outcomenum=5 is Copper Number, 6 Glucose*/
  timeCopp=1;
  timeGluc=0;
  interceptCopp=1;
  interceptGluc=0;
  Y56=Glucose;
  outcomenum=6;
  timeCopp=0;
  timeGluc=1;
  interceptCopp=0;
  interceptGluc=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeCopp timeGluc
interceptCopp interceptGluc XCopp XGluc Y56;
output;
run;

 Proc Mixed Data= new.PairCoppGluc; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y56= timeCopp*Genotype timeGluc*Genotype/ noint solution
outpm=resid_19;
        random interceptCopp interceptGluc timeCopp timeGluc/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_19 solutionF=Fixed_19;
run;
/*****************************************************************/
data new.PairCoppXylo; /* Data 20 for Copper Number and Xylose Pair*/
set  new.alpha96jointcorr;
  Y57=Copper_No;
  outcomenum=5; /* Outcomenum=5 is Copper Number, 7 Xylose*/
  timeCopp=1;
  timeXylo=0;
  interceptCopp=1;
  interceptXylo=0;
  output;
  Y57=Xylose;
  outcomenum=7;
  timeCopp=0;
  timeXylo=1;
  interceptCopp=0;
  interceptXylo=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeCopp timeXylo
interceptCopp interceptXylo XCopp XXylo Y57;
output;
run;
Proc Mixed Data= new.PairCoppXylo; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y57=timeCopp*Genotype timeXylo*Genotype/ noint solution
outpm=resid_20;
        random interceptCopp interceptXylo timeCopp timeXylo/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_20 solutionF=Fixed_20;
run;
/*****************************************************************************/
data new.PairGlucXylo; /* Data 21 for Glucose and Xylose Pair*/
set  new.alpha96jointcorr;
  Y67=Glucose;
  outcomenum=6; /* Outcomenum=6 is Glucose, 7 Xylose*/
  timeGluc=1;
  timeXylo=0;
  interceptGluc=1;
  interceptXylo=0;
```

```
  output;
  Y67=Xylose;
  outcomenum=7;
  timeGluc=0;
  timeXylo=1;
  interceptGluc=0;
  interceptXylo=1;
  keep Genotype BleachCond Sample Stage1 t1 t2 Outcomenum timeGluc timeXylo
interceptGluc interceptXylo XGluc XXylo Y67;
output;
run;

Proc Mixed Data= new.PairGlucXylo; /* covtest METHOD=REML;*/
     Class Sample Genotype Stage1 outcomenum;
        model Y67= timeGluc*Genotype timeXylo*Genotype/ noint solution
outpm=resid_21;
        random interceptGluc interceptXylo timeGluc timeXylo/ subject=Sample
type=un;
        repeated outcomenum/subject=Sample*Stage1 type=un;
        ods output covparms=cov_21 solutionF=Fixed_21;
run;
/*****************************************************************/
/*                                                               ***/
/*    STEP 2: Culculating the H and G matrix for each pair       ***/
/*                                                              ***/
/*****************************************************************/
proc iml symsize=10000 worksize=10000;   /*PAIR 1, VARIABLES 1 AND 2*/
 free H; free B;
/*Between subject covariance matrix*/

 G={3.371  -0.9227 -0.6225  0.0001,
     -0.9227  0.3390  0.0001 -0.6263,
     -0.6225  0.0001  0.6283  0.0001,
      0.0001 -0.6263  0.0001  0.6264};
/*Within subject covariance matrix*/
 R={131.0600 9.0200,
      9.0200 1.9178};
/*Extracting Y, X, Z and residuals*/
use resid_1;
read all var{Sample} into id;
read all var{Y12} into Y;
read all var{resid} into resid;
read all var{timeVis timeLig} into X;
read all var{interceptVis interceptLig timeVis timeLig} into Z;
close resid_1;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
```

```
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_1 from H; append from H;
 create B_1 from B; append from B;
quit;
/*=====================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 2, VARIABLES 1 AND 3*/
 free H; free B;
/*Between subject covariance matrix*/
 G={3.3693 -0.1941 -0.6248 0.0001,
-0.1941 0.3074 0.0001 -0.6263,
-0.6248 0.0001 0.6271 0.0001,
0.0001 -0.6263 0.0001 0.6264};
/*Within subject covariance matrix*/
 R={131.07  10.8395,
10.8395  3.0621};
/*Extracting Y, X, Z and residuals*/
use resid_2;
read all var{Sample} into id;
read all var{Y13} into Y;
read all var{resid} into resid;
read all var{timeVis timeGam} into X;
read all var{interceptVis interceptGam timeVis timeGAm} into Z;
close resid_2;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
```

```
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_2 from H; append from H;
 create B_2 from B; append from B;
quit;
/*=================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 3, VARIABLES 1 AND 4*/
 free H; free B;
/*Between subject covariance matrix*/
 G={3.3696 0.6013 -0.6243 -0.0001,
0.6013 0.6013 -0.0001 -0.6263,
-0.6243 -0.0001 0.6274 -0.0001,
-0.0001 -0.6263 -0.0001 0.6264};

/*Within subject covariance matrix*/
 R={131.06   -10.3872,
-10.3872  3.6538};
/*Extracting Y, X, Z and residuals*/
use resid_3;
read all var{Sample} into id;
read all var{Y14} into Y;
read all var{resid} into resid;
read all var{timeVis timeAlpha} into X;
read all var{interceptVis interceptAlpha timeVis timeAlpha} into Z;
close resid_3;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
```

```
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_3 from H; append from H;
 create B_3 from B; append from B;
quit;

/*=============================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 4, VARIABLES 1 AND 5*/
 free H; free B;
/*Between subject covariance matrix*/
 G={3.3695 -1.1391 -0.6245 0.00003,
-1.1391 0.4799 0.00003 -0.6263,
-0.6245 0.00003 0.6273 0.00003,
0.00003 -0.6263 0.00003 0.6264};
/*Within subject covariance matrix*/
 R={131.07  5.7503,
    5.7503  0.9340};
/*Extracting Y, X, Z and residuals*/
use resid_4;
read all var{Sample} into id;
read all var{Y15} into Y;
read all var{resid} into resid;
read all var{timeVis timeCopp} into X;
read all var{interceptVis interceptCopp timeVis timeCopp} into Z;
```

```
close resid_4;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_4 from H; append from H;
 create B_4 from B; append from B;
quit;
/*================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 5, VARIABLES 1 AND 6*/
 free H; free B;
/*Between subject covariance matrix*/
 G={5.2497 1.3382 -2.5012 0.0083,
1.3382 1.8714 0.0083 -2.5057,
-2.5012 0.0083 2.5075 0.0083,
0.0083 -2.5057 0.0083 2.5053};
/*Within subject covariance matrix*/
 R={131.02   -12.6581,
```

```
-12.6581   4.4274};
/*Extracting Y, X, Z and residuals*/
use resid_5;
read all var{Sample} into id;
read all var{Y16} into Y;
read all var{resid} into resid;
read all var{timeVis timeGluc} into X;
read all var{interceptVis interceptGluc timeVis timeGluc} into Z;
close resid_5;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_5 from H; append from H;
 create B_5 from B; append from B;
quit;

/*=========================================================================*/
```

```
proc iml symsize=10000 worksize=10000;   /*PAIR 6, VARIABLES 1 AND 7*/
 free H; free B;
/*Between subject covariance matrix*/

 G={3.3689   -0.7560   -0.6256   0.0039,
-0.7560   0.5709   0.0039   -0.6251,
-0.6256   0.0039   0.6267   0.0039,
0.0039   -0.6251   0.0039   0.6270};

/*Within subject covariance matrix*/
 R={131.08  5.5440,
5.5440   0.9878};

/*Extracting Y, X, Z and residuals*/
use resid_6;
read all var{Sample} into id;
read all var{Y17} into Y;
read all var{resid} into resid;
read all var{timeVis timeXylo} into X;
read all var{interceptVis interceptXylo timeVis timeXylo} into Z;
close resid_6;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
```

```
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_6 from H; append from H;
 create B_6 from B; append from B;
quit;
/*=================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 7, VARIABLES 2 AND 3*/
 free H; free B;
/*Between subject covariance matrix*/

 G={0.3390  -0.2824  -0.6263  0.0000,
-0.2824  0.3074  0.0000  -0.6263,
-0.6263  0.0000  0.6263  0.0000,
0.0000  -0.6263  0.0000  0.6263};

/*Within subject covariance matrix*/

R={1.9180  1.9676,
1.9676  3.0622};
/*Extracting Y, X, Z and residuals*/
use resid_7;
read all var{Sample} into id;
read all var{Y23} into Y;
read all var{resid} into resid;
read all var{timeLig timeGam} into X;
read all var{interceptLig interceptGam timeLig timeGam} into Z;
close resid_7;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
```

```
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_7 from H; append from H;
 create B_7 from B; append from B;
quit;

/*================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 8, VARIABLES 2 AND 4*/
 free H; free B;
/*Between subject covariance matrix*/
G={0.3390  0.2955  -0.6263  0.0000,
0.2955  0.1441  0.0000  -0.6263,
-0.6263  0.0000  0.6263  0.0000,
0.0000  -0.6263  0.0000  0.6263};
/*Within subject covariance matrix*/
R={1.9180  -1.8854,
-1.8854  3.6540};
/*Extracting Y, X, Z and residuals*/
use resid_8;
read all var{Sample} into id;
read all var{Y24} into Y;
read all var{resid} into resid;
read all var{timeLig timeAlpha} into X;
read all var{interceptLig interceptAlpha timeLig timeAlpha} into Z;
close resid_8;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
```

```
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_8 from H; append from H;
 create B_8 from B; append from B;
quit;
/*============================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 9, VARIABLES 2 AND 5*/
 free H; free B;
/*Between subject covariance matrix*/
G={0.3390  -0.2106  -0.6263  0.0000,
-0.2106  0.4799  0.0000  -0.6263,
-0.6263  0.0000  0.6263  0.0000,
0.0000  -0.6263  0.0000  0.6263};
/*Within subject covariance matrix*/
R={1.9180  1.2172,
1.2172  0.9340};
/*Extracting Y, X, Z and residuals*/
use resid_9;
read all var{Sample} into id;
read all var{Y25} into Y;
read all var{resid} into resid;
read all var{timeLig timeCopp} into X;
read all var{interceptLig interceptCopp timeLig timeCopp} into Z;
close resid_9;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
```

```
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_9 from H; append from H;
 create B_9 from B; append from B;
quit;
/*=============================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 10, VARIABLES 2 AND 6*/
 free H; free B;
/*Between subject covariance matrix*/
G={0.3390  0.3902  -0.6263  0.0028,
0.3902  0.0057  0.0028  -0.5999,
-0.6263  0.0028  0.6263  0.0028,
0.0028  -0.5999  0.0028  0.6396};
/*Within subject covariance matrix*/

R={1.9180  -2.6945,
-2.6945  4.5336};
/*Extracting Y, X, Z and residuals*/
use resid_10;
read all var{Sample} into id;
read all var{Y26} into Y;
read all var{resid} into resid;
read all var{timeLig timeGluc} into X;
read all var{interceptLig interceptGluc timeLig timeGluc} into Z;
close resid_10;
numobs=nrow(X);
/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
```

```
end;
nsubjects=nrow(vsize);
p=ncol(x);
H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;
 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;
 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_10 from H; append from H;
 create B_10 from B; append from B;
quit;

/*=========================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 11, VARIABLES 2 AND 7*/
 free H; free B;
/*Between subject covariance matrix*/

G={0.3390  -0.1570  -0.6263  -0.0016,
-0.1570  0.5769  -0.0016  -0.6130,
-0.6263  -0.0016  0.6263  -0.0016,
-0.0016  -0.6130  -0.0016  0.6330};


/*Within subject covariance matrix*/

R={1.9180  1.1659,
1.1659  1.0102};

/*Extracting Y, X, Z and residuals*/
use resid_11;
read all var{Sample} into id;
read all var{Y27} into Y;
read all var{resid} into resid;
```

```
read all var{timeLig timeXylo} into X;
read all var{interceptLig interceptXylo timeLig timeXylo} into Z;
close resid_11;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_11 from H; append from H;
 create B_11 from B; append from B;
quit;


/*================================================================*/
```

```
proc iml symsize=10000 worksize=10000;   /*PAIR 12, VARIABLES 3 AND 4*/
 free H; free B;
/*Between subject covariance matrix*/

G={0.3074  0.4183  -0.6263  0.0000,
0.4183  0.1441  0.0000  -0.6263,
-0.6263  0.0000  0.6264  0.0000,
0.0000  -0.6263  0.0000  0.6264};

/*Within subject covariance matrix*/

R={3.0620  -3.1214,
-3.1214  3.6538};

/*Extracting Y, X, Z and residuals*/
use resid_12;
read all var{Sample} into id;
read all var{Y34} into Y;
read all var{resid} into resid;
read all var{timeGam timeAlpha} into X;
read all var{interceptGam interceptAlpha timeGam timeAlpha} into Z;
close resid_12;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
```

```
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_12 from H; append from H;
 create B_12 from B; append from B;
quit;


/*============================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 13, VARIABLES 3 AND 5*/
 free H; free B;
/*Between subject covariance matrix*/

G={0.3074  -0.2980  -0.6263  0.0000,
-0.2980  0.4799  0.0000  -0.6263,
-0.6263  0.0000  0.6263  0.0000,
0.0000  -0.6263  0.0000  0.6263};

/*Within subject covariance matrix*/

R={3.0622  1.6174,
1.6174  0.9340};

/*Extracting Y, X, Z and residuals*/
use resid_13;
read all var{Sample} into id;
read all var{Y35} into Y;
read all var{resid} into resid;
read all var{timeGam timeCopp} into X;
read all var{interceptGam interceptCopp timeGam timeCopp} into Z;
close resid_13;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
```

```
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_13 from H; append from H;
 create B_13 from B; append from B;
quit;


/*=================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 14, VARIABLES 3 AND 6*/
 free H; free B;
/*Between subject covariance matrix*/

G={2.1864  0.5039  -2.5054  -0.0053,
0.5039  1.8778 -0.0053  -2.4928,
-2.5054  -0.0053 2.5054  -0.0053,
-0.0053  -2.4928  -0.0053  2.5117};

/*Within subject covariance matrix*/

R={3.0622  -3.3881,
-3.3881  4.4173};

/*Extracting Y, X, Z and residuals*/
use resid_14;
read all var{Sample} into id;
read all var{Y36} into Y;
read all var{resid} into resid;
read all var{timeGam timeGluc} into X;
read all var{interceptGam interceptGluc timeGam timeGluc} into Z;
close resid_14;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
```

```
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_14 from H; append from H;
 create B_14 from B; append from B;
quit;

/*=================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 15, VARIABLES 3 AND 7*/
 free H; free B;
/*Between subject covariance matrix*/

G={0.3074  -0.2412  -0.6263  0.0035,
-0.2412  0.5739  0.0035  -0.6191,
-0.6263  0.0035  0.6263  0.0035,
0.0035  -0.6191  0.0035  0.6300};

/*Within subject covariance matrix*/

R={3.0622  1.5778,
1.5778  0.9857};

/*Extracting Y, X, Z and residuals*/
```

```
use resid_15;
read all var{Sample} into id;
read all var{Y37} into Y;
read all var{resid} into resid;
read all var{timeGam timeXylo} into X;
read all var{interceptGam interceptXylo timeGam timeXylo} into Z;
close resid_15;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_15 from H; append from H;
 create B_15 from B; append from B;
quit;
```

```
/*=====================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 16, VARIABLES 4 AND 5/
 free H; free B;
/*Between subject covariance matrix*/

G={0.1441  0.2737  -0.6263  0.0000,
0.2737  0.4799  0.0000  -0.6263,
-0.6263  0.0000  0.6263  0.0000,
0.0000  -0.6263  0.0000  0.6263};

/*Within subject covariance matrix*/

R={3.6540  -1.6043,
-1.6043  0.9340};

/*Extracting Y, X, Z and residuals*/
use resid_16;
read all var{Sample} into id;
read all var{Y45} into Y;
read all var{resid} into resid;
read all var{timeAlpha timeCopp} into X;
read all var{interceptAlpha interceptCopp timeAlpha timeCopp} into Z;
close resid_16;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
```

```
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_16 from H; append from H;
 create B_16 from B; append from B;
quit;


/*============================================================*/

proc iml symsize=10000 worksize=10000;   /*PAIR 17, VARIABLES 4 AND 6/
 free H; free B;
/*Between subject covariance matrix*/

G={2.0231  -0.6118  -2.5054  -0.0081,
-0.6118  1.8626  -0.0081  -2.5232,
-2.5054  -0.0081  2.5054  -0.0081,
-0.0081  -2.5232  -0.0081  2.4965};

/*Within subject covariance matrix*/

R={3.6540  3.8341,
3.8341  4.6570};

/*Extracting Y, X, Z and residuals*/
use resid_17;
read all var{Sample} into id;
read all var{Y46} into Y;
read all var{resid} into resid;
read all var{timeAlpha timeGluc} into X;
read all var{interceptAlpha interceptGluc timeAlpha timeGluc} into Z;
close resid_17;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];
```

```
 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_17 from H; append from H;
 create B_17 from B; append from B;
quit;


/*================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 18, VARIABLES 4 AND 7/
 free H; free B;
/*Between subject covariance matrix*/

G={0.1441  0.3409  -0.6263  0.0034,
0.3409  0.5652  0.0034  -0.6366,
-0.6263  0.0034  0.6263  0.0034,
0.0034  -0.6366  0.0034  0.6212};

/*Within subject covariance matrix*/

R={3.6540  -1.8035,
-1.8035  1.0482};

/*Extracting Y, X, Z and residuals*/
use resid_18;
read all var{Sample} into id;
read all var{Y47} into Y;
read all var{resid} into resid;
read all var{timeAlpha timeXylo} into X;
read all var{interceptAlpha interceptXylo timeAlpha timeXylo} into Z;
close resid_18;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
```

```
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_18 from H; append from H;
 create B_18 from B; append from B;
quit;


/*================================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 19, VARIABLES 5 AND 6/
 free H; free B;
/*Between subject covariance matrix*/

G={0.4799  0.3293  -0.6263  0.0053,
0.3293  0.0054  0.0053  -0.6005,
-0.6263  0.0053  0.6263  0.0053,
0.0053  -0.6005  0.0053  0.6393};

/*Within subject covariance matrix*/
```

```
R={0.9340  -1.9626,
-1.9626  4.5241};

/*Extracting Y, X, Z and residuals*/
use resid_19;
read all var{Sample} into id;
read all var{Y56} into Y;
read all var{resid} into resid;
read all var{timeCopp timeGluc} into X;
read all var{interceptCopp interceptGluc timeCopp timeGluc} into Z;
close resid_19;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
```

```
 pnt=pnt+vsize[i];
 end;
 create J_19 from H; append from H;
 create B_19 from B; append from B;
quit;


/*============================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 20, VARIABLES 5 AND 7/
 free H; free B;
/*Between subject covariance matrix*/

G={0.4799  -0.1538  -0.6263  -0.0029,
-0.1538  0.5768  -0.0029  -0.6134,
-0.6263  -0.0029  0.6263  -0.0029,
-0.0029  -0.6134  -0.0029  0.6328};

/*Within subject covariance matrix*/

R={0.9340  0.8843,
0.8843  1.0072};

/*Extracting Y, X, Z and residuals*/
use resid_20;
read all var{Sample} into id;
read all var{Y57} into Y;
read all var{resid} into resid;
read all var{timeCopp timeXylo} into X;
read all var{interceptCopp interceptXylo timeCopp timeXylo} into Z;
close resid_20;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
```

```
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_20 from H; append from H;
 create B_20 from B; append from B;
quit;

/*============================================================*/
proc iml symsize=10000 worksize=10000;   /*PAIR 21, VARIABLES 6 AND 7/
 free H; free B;
/*Between subject covariance matrix*/

G={0.0002  0.2538  -0.6315  0.0014,
0.2538  0.5700  0.0014  -0.6270,
-0.6315  0.0014  0.6238  0.0014,
0.0014  -0.6270  0.0014  0.6260};

/*Within subject covariance matrix*/

R={4.4381  -2.0202,
-2.0202  0.9906};

/*Extracting Y, X, Z and residuals*/
use resid_21;
read all var{Sample} into id;
read all var{Y67} into Y;
read all var{resid} into resid;
read all var{timeGluc timeXylo} into X;
read all var{interceptGluc interceptXylo timeGluc timeXylo} into Z;
close resid_21;
numobs=nrow(X);

/*Generate Matrix vsize that contains number of observations for each subject*/
count=1;
free vsize;
do i=1 to (numobs-1);
     if id[i]=id[i+1] then
        do; count=count+1; end;
     else if id[i]^=id[i+1] then
        do; vsize=vsize//count; count=1; end;
     if i=numobs-1 then vsize=vsize//count;
end;
nsubjects=nrow(vsize);
p=ncol(x);

H=J(p,p,0);
do i=1 to nsubjects;
 if i=1 then pnt=1;

 /*Z, X, Y and resid matrix for i-th subject*/
 Zi=Z[pnt:pnt+vsize[i]-1,];
```

```
 Xi=X[pnt:pnt+vsize[i]-1,];
 yi=y[pnt:pnt+vsize[i]-1];
 residi=resid[pnt:pnt+vsize[i]-1];

 /*Generates Ri */
 ni=nrow(yi);
 I_ni=diag(J(ni/2,1,1));
 Ri=I_ni@R;

 /*Check for missing observation in Y and X
 Vector pr_Y: contains the list of non-missing observations in Y
 Vector pr_X: contains the list of non-missing observations in X.
 (If any of the covariate value is missing - it will be considered as missing
 in general)*/
 pr_Y=loc(residi^=.);
 tloc=ncol(Xi);
 fnd=(Xi^=.);
 fnd2=fnd[,+];
 pr_X=loc(fnd2=tloc);
 if (ncol(pr_Y)>0 & ncol(pr_X)>0) then
 do;
 Present=xsect(pr_Y, pr_X);
 Zi=Zi[Present,];
 Xi=Xi[Present,];
 yi=yi[Present,];
 Residi=Residi[Present,];
 Ri=Ri[Present,Present];
 Vi=Zi*G*t(Zi)+Ri; /*Vi*/
 Wi=ginv(Vi);      /*Wi*/
 H_i=t(Xi)*Wi*Xi;  /*Contribution to H from i-th subject*/
 H=H+H_i;          /*Accumulated Hessian matrix till i-th individual*/
 B_ik=t(Xi)*Wi*Residi; /*Contribution to G from i-th subject*/
 B=B||B_ik;        /*Accumulated G matrix till i-th individual*/
 end;
 else B=B||J(p,1,0);
 pnt=pnt+vsize[i];
 end;
 create J_21 from H; append from H;
 create B_21 from B; append from B;
quit;

/*****************************************************************/
/**********                                             **********/
/**********  STEP 3: Combining all Matrices into H and G     **********/
/**********                                             **********/
/*****************************************************************/
proc iml;
use H_1; read all into H_1; close H_1;
use H_2; read all into H_2; close H_2;
use H_3; read all into H_3; close H_3;
H_comb=block(H_1, H_2, H_3, H_4, H_5, H_6, H_7, H_8, H_9, H_10, H_11, H_12, H_13,
H_14, H_15, H_16, H_17, H_4, H_4, H_4, H_4, );
create H from H_comb;
append from H_comb;
use G_1; read all into G_1; close G_1;
use G_2; read all into G_2; close G_2;
use G_3; read all into G_3; close G_3;
G_comb=G_1//G_2//G_3;
create G from G_comb; append from G_comb;
quit;

/*****************************************************************/
/**********                                           ***********/
/**********   STEP 4: Combining all Matrices into H and G   ***********/
/**********                                           ***********/
/*****************************************************************/
proc iml;
use H_1; read all into H_1; close H_1;
```

```
use H_2; read all into H_2; close H_2;
use H_3; read all into H_3; close H_3;
use H_4; read all into H_4; close H_4;
use H_5; read all into H_5; close H_5;
use H_7; read all into H_7; close H_7;
use H_8; read all into H_8; close H_8;
use H_9; read all into H_9; close H_9;
use H_10; read all into H_10; close H_10;
use H_12; read all into H_12; close H_12;
use H_13; read all into H_13; close H_13;
use H_14; read all into H_14; close H_14;
use H_16; read all into H_16; close H_16;
use H_17; read all into H_17; close H_17;
use H_19; read all into H_19; close H_19;

H_comb=block(H_1, H_2, H_3, H_4, H_5, H_7, H_8, H_9, H_10, H_12, H_13, H_14, H_16,
H_17, H_19);
create H from H_comb;
append from H_comb;

use B_1; read all into B_1; close B_1;
use B_2; read all into B_2; close B_2;
use B_3; read all into B_3; close B_3;
use B_4; read all into B_4; close B_4;
use B_5; read all into B_5; close B_5;
use B_7; read all into B_7; close B_7;
use B_8; read all into B_8; close B_8;
use B_9; read all into B_9; close B_9;
use B_10; read all into B_10; close B_10;
use B_12; read all into B_12; close B_12;
use B_13; read all into B_13; close B_13;
use B_14; read all into B_14; close B_14;
use B_16; read all into B_16; close B_16;
use B_17; read all into B_17; close B_17;
use B_19; read all into B_19; close B_19;

B_comb=B_1//B_2//B_3//B_4//B_5//B_7//B_8//B_9//B_10//B_12//B_13//B_14//B_16//B_17//
B_19;
create B from B_comb; append from B_comb;
quit;

/*************************************************************************/
/*****                                                          *********/
/*****  STEP 5: Creating the Matrices J-hat and K-hat           *********/
/**********                                                     *********/
/*************************************************************************/
proc iml;
     use B; read all into B; close B;
     nsubjects=ncol(B);
     K_1=B*t(B);
     K=K_1#1/nsubjects;
     create K from K; append from K;
     use H; read all into H; close H;
     J=H#1/nsubjects;
     create J from J; append from J;
        print J K;
quit;
/*************************************************************************/
/***                                                             ******/
/***  STEP 6: Estimating the Covariance matrix: SIGMA_zero       ******/
/***                                                             ******/
/*************************************************************************/
proc iml;
     use J; read all into J; close J;
     use K; read all into K; close K;
     nsubjects=24;
     Sigma=inv(J)*K*inv(J);
     Sigma0=Sigma#1/nsubjects;
```

```
     create Sigma0 from Sigma0; append from Sigma0;
        print Sigma0;
        print J K;
quit;
/*****************************************************************/
/****                                                         ***/
/**** STEP 7: Estimating the Fixed parameter Matrix of estimates ***/
/****                                                         ***/
/*****************************************************************/
proc iml; /*Edunnii Slope parameters */
     ThetaED12={-14.7300, -2.2304};
     ThetaED13={-14.7300, -2.8093};
     ThetaED14={-14.7300, 3.1036};
        ThetaED15={-14.7300, -1.7644};
        ThetaED16={-14.7300, 4.0108};
        ThetaED23={-2.2304, -2.8093};
        ThetaED24={-2.2304, 3.1036};
        ThetaED25={-2.2304, -1.7644};
        ThetaED26={-2.2304, 4.0108};
        ThetaED34={-2.8093, 3.1036};
        ThetaED35={-2.8093, -1.7644};
        ThetaED36={-2.8093, 4.0108};
        ThetaED45={3.1036, -1.7644};
        ThetaED46={3.1036, 4.0108};
        ThetaED56={-1.7644, 4.0108};

ThetaED_est=ThetaED12//ThetaED13//ThetaED14//ThetaED15//ThetaED16//ThetaED23//Theta
ED24//ThetaED25//ThetaED26//ThetaED34//ThetaED35//ThetaED36//ThetaED45//ThetaED46//
ThetaED56;
     create ThetaED_est from ThetaED_est;
     append from ThetaED_est;
        print ThetaED_est;
quit;
/*****************************************************************/
/****                                                         ***/
/**** STEP 8: Culculation of the overall parameter estimates      ***/
/****   and their standard errors                             ***/
/****                                                         ***/
/*****************************************************************/
Proc iml;
A={0.2 0 0.2 0 0.2 0 0.2 0 0.2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0,
   0 0.2 0 0 0 0 0 0 0 0 0.2 0 0.2 0 0.2 0 0.2 0 0 0 0 0 0 0 0 0 0,
   0 0 0 0.2 0 0 0 0 0 0 0 0.2 0 0 0 0 0 0 0.2 0 0.2 0 0.2 0 0 0 0 0 0,
   0 0 0 0 0 0.2 0 0 0 0 0 0 0 0.2 0 0 0 0 0 0.2 0 0 0 0 0.2 0 0.2 0 0 0,
   0 0 0 0 0 0 0 0.2 0 0 0 0 0 0 0 0.2 0 0 0 0 0 0.2 0 0 0 0.2 0 0 0.2 0,
   0 0 0 0 0 0 0 0 0 0.2 0 0 0 0 0 0 0 0.2 0 0 0 0 0 0.2 0 0 0 0.2 0 0.2};
use Sigma0; read all into Sigma0; close Sigma0;
use ThetaED_est; read all into ThetaED_est; close ThetaED_est;
ThetaED_est_star=A*ThetaED_est;
G_hat=A*Sigma0*t(A);
temp=diag(A*Sigma0*t(A));
stderr=J(6,1,0);
do i=1 to 6;
stderr[i,1]=sqrt(temp[i,i]);
end;
create ThetaED_est_star from ThetaED_est_star; append from ThetaED_est_star;
create Stderr from Stderr; append from Stderr;
G_diag=diag(G_hat);
G_diagSQRT=sqrt(G_diag);
G_diagSRTInv=inv(G_diagSQRT);
G_corr=G_diagSRTInv*G_hat*G_diagSRTInv;
print ThetaED_est_star Stderr A G_hat G_corr;
/* ThetaED_est_star and Stderr contain parameter estimates and their Standard erros
*/
quit;
/*****************************************************************/
/****                                                         ***/
/****   STEP 9: Culculation of the G and R matrices           ***/
```

```
/****     and their standard errors                                ***/
/****    And Correlations between Slopes                          ***/
/*****************************************************************/
proc iml;
G = {0.62746 0.0001 0.0001 -0.0001 0.00003 0.0083,
     0.0001 0.62632 0.0000 0.0000  0.00000 0.0028,
     0.0001 0.0000 0.62636 0.0000  0.00000 -0.0053,
     -0.0001 0.0000 0.0000 0.62638 0.00000 -0.0081,
     0.00003 0.0000 0.0000 0.0000 .62632 0.0053,
     0.0083 0.0028  -0.0053 -0.0081 0.0053 2.5051486}; /* this is the G matrix*/

R={131.06   9.0200 10.8395 -10.3872 5.7503 -12.6581,
     9.0200 1.91796 1.9676 -1.8854  1.2172  -2.6945,
    10.8395 1.9676 3.06214  -3.1214 1.6174 -3.3881,
   -10.8395 -1.8854 -3.1214 3.65393 -1.6043 3.8341,
     5.7503 1.2172  1.6174  -1.6043  0.9340 -1.9626,
   -12.6581 -2.6945 -3.3881 3.8341  -1.9626 4.51188}; /* this is the R matrix*/
/*Association/Correlations between slopes*/
corr_bet_slopes=j(nrow(G), nrow(G), 0);
      do i=1 to nrow(G);
         do j=1 to ncol(G);
            corr_bet_slopes[i,j]=G[i,j]/sqrt(G[i,i]*G[j,j]);
         end;
      end;
print corr_bet_slopes R;
/*Marginal correlation at time 2*/
/* Marg_corr=j(nrow(D_marg), nrow(D_marg), 0);
do i=1 to nrow(D_marg);
do j=1 to ncol(D_marg);
Marg_corr[i,j]=D_marg[i,j]/sqrt(D_marg[i,i]*D_marg[j,j]);
end;
end;
print Marg_corr; */
run;
quit;
```

A3. Published articles from the study