University of Cape Town

Longitudinal Analysis of Platelet Count Data

Author:
Mahdi Marcus

Supervisor:
Assoc. Prof.
Freedom Gumedze

A thesis submitted in partial fulfilment of the requirements for the

Master's Degree in Advanced Analytics

in the

Department of Statistical Sciences

June 10, 2025

Abstract

Platelet transfusions are critical in managing bleeding risks in patients with low platelet counts or dysfunctional platelets. This research explores the dynamics of platelet count levels.

The primary aim is to understand when and why platelet products are failing, by investigating differences in platelet count trajectories among donor groups, exploring seasonality, identifying donor clusters with similar behaviours, and establishing connections between platelet count dynamics and product failures.

Using longitudinal data from the South African National Blood Service (SANBS), I employed linear mixed-effect models to analyse platelet count trajectories and latent class mixed models to uncover donor clusters with distinct patterns.

The findings reveal evidence of seasonal fluctuations in platelet counts, with highs in winter months, though deviations were observed in specific branch zones. Functional principal component analysis (FPCA) further confirmed these seasonal patterns and revealed inter-year variability.

Critical to this study is the identification of two primary donor clusters, one with stable or elevated platelet counts and another showing a declining trend post-2018. Notably, these clusters did not significantly correlate with demographic factors like gender or location, suggesting other factors influencing platelet dynamics. The research also uncovered parallels between donor clusters and branch zones, highlighting variability in platelet profiles and product pass rates, particularly during periods of observed declines.

This research provides insights into the temporal dynamics of platelet counts and their role in the quality and reliability of platelet products. By understanding these dynamics, we can better identify the factors contributing to product failures, ultimately improving the safety and efficacy of platelet transfusion practices.

Plagiarism Declaration

I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this project from the work(s) of other people has been attributed, and has been cited and referenced.

Signature: _____

Acknowledgements

I am grateful to Prof. Jane Hutton, Prof. Francesca Little, and Assoc. Prof. Freedom Gumedze for their guidance and insights. I also want to extend my appreciation to the Department of Statistical Sciences at the University of Cape Town for their support throughout my studies.

Lastly, my heartfelt thanks go to my family, friends, and colleagues for their encouragement and support during this journey.

Contents

1 Introduction **9**
- 1.1 Background . . . 9
- 1.2 Aims and Objectives . . . 10

2 Literature Review **12**
- 2.1 Introduction . . . 12
- 2.2 Platelet Count Trends . . . 12
- 2.3 Methods for Analysing Longitudinal Data . . . 14
- 2.4 Conclusion . . . 15

3 Methodology **16**
- 3.1 Introduction . . . 16
- 3.2 Linear Mixed-Effect Models . . . 16
- 3.3 Latent Class Linear Mixed Modelling . . . 17
- 3.4 Functional Principal Component Analysis . . . 19
- 3.5 K-means Clustering . . . 22

4 Data Wrangling and Exploratory Data Analysis **24**
- 4.1 Introduction . . . 24
- 4.2 The Apheresis Donor Dataset . . . 24
- 4.3 The Platelet Products Dataset . . . 31
- 4.4 Concluding Remarks . . . 36

5 Level and Temporal Dynamics in Platelet Count **37**
- 5.1 Introduction . . . 37
  - 5.1.1 Building the Model . . . 37
- 5.2 Model Results . . . 41
- 5.3 Model Analysis . . . 41
- 5.4 Conclusion . . . 42

6 Trajectory Clusters **44**
- 6.1 Introduction . . . 44
- 6.2 Latent Class Linear Mixed Modelling . . . 44
- 6.3 FPCA . . . 49
  - 6.3.1 Clustering Based on Principal Curves . . . 52
- 6.4 Splitting each profile into separate yearly profiles . . . 56

6.5 Conclusion 58

7 Discussion and Conclusion **60**

A Linear Mixed Model Summary **64**

List of Figures

4.1 Histogram of the number of donations per donor (left) and a line graph of the number of donations on each donation date over time (right). The histogram highlights the data sparsity, showing that most donors have relatively few donations, while the line graph illustrates the low number of donations relative to the total number of donors in the dataset. . . . 25
4.2 Spaghetti plot depicting the profiles of 30 apheresis donors. Each line represents an individual donor's platelet count over time, illustrating the variability in donation patterns and platelet dynamics among the donors. . . . 26
4.3 Top: Histrograms of platelet count grouped by sex (left), branch zone (middle), and ethnicity (right). Bottom: Average platelet profiles grouped by sex (left), branch zone (middle), and ethnicity (right). . . . 26
4.4 Boxplots of platelet count on each date. . . . 28
4.5 Decomposition of average platelet count time series. . . . 29
4.6 ACF plot of the detrended average platelet count profile. . . . 29
4.7 Proportion of products passed through time. Across all branch zones (left) and separated by branch zone (right). . . . 34
4.8 Decomposition of percentage of products passing time series. . . . 35

5.1 Separate linear models for each year: Seasonal curves. . . . 38
5.2 Separate linear models for each donor: Intercepts (left) and seasonal components (right). . . . 39
5.3 Average fitted profiles from the fixed effects of the linear mixed effects model, showing seasonal variations and level differences in platelet count among female donors from 2016 to 2020. The grid format presents each year separately, with 2016 in the top left, 2017 in the top middle, 2018 in the top right, 2019 in the bottom left, and 2020 in the bottom middle. . . . 42
5.4 Residual analysis: Final Model. The plots assess variance consistency across individuals and check for patterns or heteroscedasticity with respect to month, year, branch zone, and sex. The inclusion of random intercepts and seasonal components addresses earlier model issues, ensuring residuals are centered and free from trends. . . . 43
5.5 Standardised residuals vs. Fitted values: Final Model. . . . 43

6.1 Spaghetti plot of reduced data with mean profile overlaid (black). . . 45
6.2 Cluster performance metrics for 2-5 clusters. 46
6.3 Cluster trajectories for data with more than 48 observations (left) and cluster trajectories for data with more than 30 observations (right). . 47
6.4 FPCA: fitted correlation (left) and covariance (right). 49
6.5 Top 4 principal curves. 50
6.6 Platelet Count curves when FPC curves are extreme. The left side displays the lower 2nd percentile, representing individuals with the lowest observed FPC values, while the right side shows the upper 98th percentile, capturing the highest FPC values. For PCs 1,2 and 3. 50
6.7 Scatter plots of loadings of platelet count curves onto principal curves coloured by covariate data. Gaussian ellipses representing data distribution, capturing 90% coverage are overlayed. 51
6.8 (left) Average platelet profiles for clustering based on loadings onto first 3 principal curves, and (right) average platelet profiles: Clustering based on raw loadings. 53
6.9 Average platelet count profiles: clustering based on normalised loadings (left), and clustering based on loadings onto principal curves 2 and 3 (right). 54
6.10 FPCA: fitted correlation (left) and covariance (right). 56
6.11 Top 3 principal curves. Separate yearly profiles. 57
6.12 Platelet count curves when FPC curves are extreme. The left side displays the lower 2nd percentile, representing individuals with the lowest observed FPC values, while the right side shows the upper 98th percentile, capturing the highest FPC values. For PCs 1,2 and 3. Separate yearly profiles. 57
6.13 Scatter plots of loadings of platelet count curves onto principal curves coloured by year. Gaussian ellipses representing data distribution, capturing 90% coverage are overlayed. 58

List of Tables

4.1 Variable descriptions: Apheresis Donor Dataset. 25
4.2 Number of observations in each group: Categorical Variables Crosstabulation. 27
4.3 Summary of final platelet count data by year. This table presents the mean and standard deviation of platelet counts for each year, along with the number of male and female donors. Additionally, it includes the count of donations from each branch zone by year. 30
4.4 Summary of pooled platelet products data. This table presents the number of products categorised by test outcomes—pass, fail, and not tested—for both volume and yield tests across different years. It also includes the number of products produced by year and branch zone. . 32
4.5 Pass and failure rates of platelet products by branch zone and year. The table summarizes the proportion of products that passed and failed across different branch zones over the specified years. 33
4.6 Summary output of the generalised linear mixed-effects model (GLMM) analysing seasonality in platelet product failures. The table presents estimates, standard errors, z-values, and p-values for the fixed effects 35

5.1 Summary of model comparison metrics for different models, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Log Likelihood (LogLik), Deviance, and χ^2 statistics. The table evaluates the balance between model complexity and fit, supporting the selection of Model 2A as the best fit for the data. 40

6.1 Cluster classification by gender for smaller dataset (at least 48 observations in each profile). 46
6.2 Cluster classification by gender for larger dataset (at least 30 observations in each profile). 47
6.3 Cluster classification by branch zone for smaller dataset (at least 48 observations in each profile). 47
6.4 Cluster classification by branch zone for larger dataset (at least 30 observations in each profile) 48
6.5 Contingency Table - Gender vs. Cluster. Clustering based on raw loadings. 52
6.6 Contingency Table - Branch zone vs. Cluster. Clustering based on raw loadings. 52

6.7 Contingency Table - Gender vs. Cluster. Clustering based on standardised loadings. 54
6.8 Contingency Table - Branch zone vs. Cluster. Clustering based on standardised loadings. 55
6.9 Contingency Table - Gender vs. Cluster. Clustering based loadings onto second and third principal curves. 55
6.10 Contingency Table - Branch zone vs. Cluster. Clustering based loadings onto second and third principal curves. 55

A.1 Linear mixed model summary 66

Chapter 1

Introduction

1.1 Background

Platelet transfusions are essential for preventing and treating bleeding in patients with low platelet counts or platelet function issues. Whether used as a proactive measure or to address active bleeding, these transfusions help restore platelet levels or correct platelet function abnormalities. Platelet products and transfusions are vital for managing bleeding risks and ensuring patient safety, underscoring their importance in medical practice [36].

The South African National Blood Service (SANBS) functions as a non-profit organisation, serving a crucial function within South Africa by providing necessary blood transfusion and related services. With a mandate to deliver blood products safely, SANBS is dedicated to maintaining the quality and integrity of its offerings.

Platelet components play a crucial role in transfusion medicine, and there are two primary methods for their collection: Pooled Buffy Coat Platelet Concentrates and Single Donor Apheresis Platelet Concentrates. Pooled Buffy Coat Platelet Concentrates are obtained from the buffy coat layers of multiple whole blood donations (a tissue from which necessary components are processed. In the case of this research - platelets), pooled together, and then re-suspended in plasma or a platelet additive solution. These concentrates, typically comprising of platelets from 4-5 donors, are filtered to produce a concentrate. On the other hand, Single Donor Apheresis Platelet Concentrates are collected from a single donor using apheresis systems, which allow for the collection of larger platelet numbers that can be divided into multiple bags.

While both types of concentrates exhibit therapeutic equivalence in terms of post-transfusion increments and efficacy, Single Donor Apheresis Concentrates are preferred for patients needing prolonged support, like those undergoing haemopoietic stem cell transplantation or chemotherapy for haematological cancers [36].

Apheresis donors provide platelets or plasma every two weeks. Their red cells are promptly returned to them post-procedure, allowing for frequent donations. Donors'

platelet counts are tested through pre- and post-procedures, ensuring the quality of the collected platelets. These platelets are amalgamated into larger bags, which rarely experience failures - as per SANBS. In contrast, whole blood donors contribute by providing a bag of blood every 56 days. Pooled platelet products are tested after they are produced and have to adhere to specified norms. Specifically, success is measured against the following standards: (1) platelet yield, should be at least 2.4×10^{11} platelets, reflecting the yield of platelets per unit; (2) sterility, where a negative result is non-negotiable, ensuring the product's purity and safety; and (3) volume, falling within the range of 200 to 800 millilitres. Any deviation from these criteria constitutes a failure. It is in this pooling process that what seems to be seasonal failures is observed, a phenomenon deserving of our attention. Given the mandate of SANBS, and the increasing importance and demand of platelet products in the healthcare system [15], it is important to understand the dynamics affecting the quality and reliability of the products produced.

1.2 Aims and Objectives

Data are received from SANBS relating to platelet donations and products. Two datasets are provided (1) relating to apheresis donors and (2) relating to pooled products. Detailed descriptions of the contents of the data provided can be seen in Chapter 4. The apheresis data has information on donors' race, ethnicity, gender, location, and platelet count. While the pool products data only has information on each product's test outcome and location. We highlight the most important limitation in the data, as it informs the aim of this research: we do not have platelet count information of the individual whole blood donors pooled to produce products.

Through answering the following research questions, we hope to further understand trends of product failures, and aim to decipher patterns in platelet count and potential causal factors behind pooled product failures.

Research Question 1: Is there a difference in platelet count levels and trajectories among empirical groupings?
Research Question 2: Is there seasonality in platelet count?
Research Question 3: Are there distinct groupings of donors that behave similarly?
Research Question 4: Is there a connection between average platelet trajectories and products failing?

To address the research questions posed, relevant literature concerning the modelling of platelet count levels and trajectories was first reviewed, as presented in Chapter 2. In Chapter 4, the donor data were explored using exploratory data analysis techniques to provide a comprehensive understanding of the dataset. The Platelet Product Dataset was also examined in Chapter 4 to identify patterns of platelet product failures, assess any seasonal trends, and confirm the observed increase in failure rates during the latter half of the analysis period. The methodologies employed in subsequent analyses were introduced in Chapter 3, including those used in Chapter 5, where platelet count trajectories were modelled using linear mixed-effect

models, and Chapter 6, where clustering was performed to identify groups of donors exhibiting similar platelet count trajectories. Finally, the discussion and conclusions were presented in Chapter 7.

Chapter 2

Literature Review

2.1 Introduction

The primary data in this thesis relates to platelet count profiles. These profiles are repeated measures from individual donors through time. In this chapter literature relating to factors implicating platelet levels and trajectories, and techniques used to analyse platelet levels and trends are explored. Furthermore, methods for analysing longitudinal data are reviewed.

2.2 Platelet Count Trends

The normal platelet count in humans ranges from $150 \times 10^9/L$ and $400 \times 10^9/L$ [8]. Interestingly, there is evidence to suggest that difference in platelet count is inherited, i.e related to genetic makeup [2] [16]. Some research into uncovering trends and differences in platelet count is reviewed. A study was done on a population in the United States of America which aimed to explore seasonal trends in complete blood count [23]. This study uses data from seven cross-sectional surveys conducted as part of the National Health and Nutrition Examination Survey (NHANES) between 1999 and 2012 to explore the seasonal trends in complete blood count (CBC) and C-reactive protein (CRP) within the non-institutionalised US population, encompassing both children and adults.

The investigation employs linear regression models and Wilcoxon tests to compare CBC and CRP levels between the winter-spring (November-April) and summer-fall (May-October) seasons, while accounting for various demographic factors, personal behaviours, and chronic disease conditions.

The final dataset includes 27,478 children and 36,644 adults (greater than 18 years). The analysis reveals notable differences in neutrophils, WBC, CRP, red blood cell components, and platelets between the two seasons, highlighting a more pro-inflammatory immune system during winter-spring. A similar study was done in [17] which aims to investigate the presence of a seasonal pattern in platelet count. The study uses the database of the Italian Association of Blood Volunteers (AVIS) and covers

the period from 2001 to 2010. The data used contains 16,422 donors. Categorising the data into seasonal and monthly intervals, the analysis employs partial Fourier analysis to identify significant seasonal trends. The results indicate a substantial increase in platelet count during the winter-autumn period. In [5], linear mixed effect models were used to assess variations in platelet count using data from de-identified plasma donor records at the Australian Red Cross Blood Services in Victoria. The authors treated the individuals as random effects and had the interaction between year variable and month variable as fixed effects. Seasonal variation was tested using a linear combination of the month effects. The authors also found significantly higher levels of platelet counts in the winter months relative to summer months. A significantly higher platelet count level was also found in the female cohort.

Furthermore, studies consider differences in average levels of platelet count based on demographics. Platelet count variability is heritable and dependent on genetic factors. Studies identify age, sex and ethnicity as major contributing variables to platelet count variation [10]. In [18] the link between cigarette smoking and platelet count in a cohort of 5017 Israeli industrial workers aged 20–64 is explored. Among females, smokers exhibited lower platelet counts. Conversely, among males, smokers showed a slightly higher platelet count, though not statistically significant. The gender difference in platelet count persisted even after adjusting for smoking status - it found that females, on average, have a significantly higher platelet count level. Multiple regression analysis emphasised a significant negative association between smoking and platelet count in women. A common finding in existing literature is the disparity between males and females – females have, on average, a higher platelet count than males – and is supported in [33][25][5] [2]. It is also commonly found that ethnicity is a contributor to platelet count variability and was studied in [33].

The purpose of this study, which uses data from the Third National Health and Nutrition Examination Survey, is to demonstrate that observed differences in platelet counts among various ethnicities, sexes, and age groups cannot be attributed to environmental factors. The study includes 12,142 participants, with 65% women, 27% non-Hispanic blacks, and 27% Mexican Americans. The findings reveal that platelet counts differ significantly among ethnic groups, with the lowest counts in whites and the highest in non-Hispanic blacks. In [26] it was also found that Non-Hispanic Whites had significantly lower level platelet counts compared to non-Hispanic Blacks, Hispanics or all other ethnicities.

Moreover, a study aimed at analysing platelet count levels in patients referred to the Emergency Department of the Sant'Orsola-Malpighi Hospital in Bologna was done. The study aimed to investigate platelet count variations in individuals with COVID-19 in comparison to a non-COVID-19 control group, examining potential correlations with respiratory parameters and clinical outcomes. Using Wilcoxon rank-sum tests, findings revealed a significant decrease in mean platelet values in COVID-19 patients as opposed to those with negative SARS-CoV-2 swabs. Additionally, the study observed links between platelet count and patient outcomes, including hospitalisation and mortality. The data suggests that SARS-CoV-2 infection may lead to a decrease in platelet count, and such counts could serve as crucial

prognostic indicators for assessing and stratifying COVID-19 patients [3].

2.3 Methods for Analysing Longitudinal Data

Although I am specifically dealing with repeated platelet count measures over time, exploring methods for analysing any longitudinal data is valuable for advancing my research. The literature on platelet count trend analysis employs limited methodologies. Since the method used is agnostic to the context in which it is applied, other methods commonly found in existing literature dealing with data of a similar structure are reviewed.

A study using data from longitudinal clinical data obtained from kidney transplant recipients focuses on tracking the progression of kidney function, measured through estimated glomerular filtration rates at various time points following kidney transplantation. Functional principal component analysis is used to explore major sources of variation in the data. Specifically, the study deals with sparse functional data. Traditional functional principal component analysis is not applicable in this context and, therefore, the Principal Analysis by Conditional Estimation (PACE) algorithm is used to estimate functional principal curves. It is found that functional principal component analysis is a useful tool for understanding dominant modes of variation.

Additionally, clustering is done based on the functional principal component scores. The authors also use functional principal component analysis to predict future scores, as well as detect outlying glomerular filtration rate curves [13]. A similar study was done which introduces a joint model employing functional principal component analysis (FPCA) to extract informative features from longitudinal trajectories (glomerular filtration rate curves), coupled with a competing risk model (survival model that handles competing events) to handle multiple time-to-event outcomes (kidney transplant outcomes). The connection between longitudinal trajectories and multiple time-to-event outcomes is established through shared functional features. Application of the model on real kidney transplant data stresses the significance of these functional features. In the application, 5,654 kidney transplant recipients are included. The outcomes of the experiment are (1) kidney transplant failure, (2) death, and (3) remain healthy. The primary outcome of the experiment is transplant while the competing event is death prior to failure of transplant [12]. In [20] FPCA was used to find the pattern of longitudinal growth trajectory. In the paper it is argued that Longitudinal Data Analysis (LDA) plays a vital role in recognising growth patterns, and FPCA provides a useful methodology for analysing these trends [35].

Another common technique used to uncover patterns in trajectory in longitudinal studies is latent class mixed models or growth mixture models [22]. Longitudinal trajectories of child abuse in a Chinese community sample is explored in [6]. The study uses three waves of data from 521 caregivers with children aged 4–7 years. Employing growth mixture models, the research examines cluster trajectories, to understand the developmental patterns and predictors associated with child abuse. It was found that child abuse trajectories are heterogonous in the sample; implying that different subtypes of abuse have different trajectories and therefore interven-

tions should be case-specific. A study focusing on finding homogenous groupings of students whose behaviours in some way predict whether that student would graduate high school or not was conducted on a sample of students from the U.S. National Center for Education Statistics. Applying growth mixture models, this study used the national dataset to investigate if there are distinct subgroups of dropouts and to understand the factors influencing each subgroup [4]. Growth mixture models proved to be effective in the study, detecting *homogenous groups within the heterogonous group* of dropouts.

Linear mixed effect models is a common technique used in the analysis of longitudinal data and is widely used in medical and biological studies. The structure of medical studies lends itself to the use of mixed-effects models as they generally present with repeated measures [32]. This requires a more flexible approach that allows for dependent and correlated observations. In [24] the authors aim to model time-course gene expression data with a mixed-effects model using B-splines, representing repeated measures of genes as the sum of smooth splines. Since there are many genes and biological processes are complicated, scientists often group genes together to make sense of all the data. When genes have similar patterns of activity, it can help predict what unknown genes do and decipher which genes work together in the same way. For this reason, the paper clusters genes using mixture models, and models gene expression trajectory – within cluster – using a mixed-effects model. The method proved to do well at clustering noisy data, outperforming the normal-mixture model, in simulation studies. The method was also tested on real data from yeast cell cycles and fibroblast responses to serum, both yielding biologically relevant results. As part of the study, the authors show that B-spline basis functions work well with periodic trajectories.

2.4 Conclusion

In this chapter, it is observed that studies on platelet count levels and trajectories employ limited statistical methods for modeling platelet count data. A common approach involves using non-parametric tests (such as the Wilcoxon test) to compare means. In [5], the authors employed linear mixed-effects models to capture variation in platelet count. No studies were found on a South African group of donors. Research on analysing longitudinal data to identify methods used in similar studies are also reviewed. Although various techniques are available for longitudinal data analysis, a few are focused on, including linear mixed-effects models, growth mixture models, and functional data analysis techniques, particularly functional principal component analysis. These methods hold promise for application in our research.

Chapter 3

Methodology

3.1 Introduction

In this chapter, the methodology employed in subsequent chapters is explored. Initially, linear mixed-effect models and their relevance to the nature of the data in this researchis examined. Additionally, latent class linear mixed modelling as a means to unveil heterogeneous clusters within the sample is investigated. Furthermore, functional data analysis, particularly functional principal component analysis using the PACE algorithm, as a method to reveal temporal patterns in functional data is discussed. Finally, anticipating the necessity in Chapter 6 to cluster platelet profiles based on their load onto principal curves, we explore K-means as a clustering method.

3.2 Linear Mixed-Effect Models

In this research, we worked with longitudinal data, specifically, looking at records of platelet counts from the same individuals over time. The fact that these measurements are taken repeatedly from the same people added a layer of complexity to the analysis. Unlike regular regression models, one can't assume that these measurements are independent and identically distributed [38].

To address this issue, we turned to mixed-effects models. These models incorporate both fixed and random effects to handle the repeated measurements. A fixed effect is something that would remain consistent if one repeated the experiment with a different group of people – think of it as a population-level parameter. On the other hand, random effects are associated with levels that result from a random selection of individuals from the population. In this case, we have a random sample of donors, and "donors" is considered a random effect. If the experiment were repeated, we would have a different set of donors in the sample. The data obtained and discussed in Chapter 4 and modeled in Chapter 5 contains longitudinal platelet count records from unique individuals over time. To address the issue of nonindependence in repeated measurements, random effects related to each individual

were incorporated into the model. Literature suggests that an individual's baseline platelet count varies inherently, making it necessary to account for these variations using random intercepts. Furthermore, as observed in Chapter 5, the model also integrates random slopes, allowing for individual-specific seasonal trends. Some individuals exhibit pronounced seasonal patterns, whereas others display more subtle or distinct fluctuations. By including both random intercepts and slopes, the model effectively captures person-specific dynamics while also isolating fixed effects—those consistent influences that persist beyond this specific dataset.

Equation 3.1 shows a general form of a mixed effect model with N observations. Where Y_i denotes the response value for observation i; $\mathbf{X}_i$ contains information on the k fixed effects for individual i - the i^{th} row of the design matrix relating to fixed effects $\mathbf{X}_{n\times k}$; $\beta_{k\times 1}$ is the vector of fixed effect parameters. $\mathbf{Z}_i$ contains information on the m fixed effects for individual i - the i^{th} row of the design matrix relating to random effects $\mathbf{Z}_{n\times m}$; $\mathbf{b}_{m\times 1}$ is the vector of random effects for observation i [1].

$$Y_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + e_i, \quad for\ \ i = 1, ..., N \tag{3.1}$$

where

$$e_i \sim N(0, \sigma^2) \tag{3.2}$$

and

$$\mathbf{b}_i \sim N(\mathbf{0}, \mathbf{\Sigma}). \tag{3.3}$$

Estimation is carried out using Restricted Maximum Likelihood (REML), which provides unbiased estimates of variance components [1]. Several key assumptions underpin the model, including: (1) Linearity between explanatory variables and the response variable, (2) Homoscedasticity, ensuring that errors have constant variance, and (3) Independence and normality of residuals.

These assumptions are primarily validated through graphical methods, such as boxplots for categorical variables to assess homoscedasticity and plots of standardized residuals against fitted values to check for other potential violations.

3.3 Latent Class Linear Mixed Modelling

Linear mixed models as described in the previous section assume that the population is homogeneous and can be described by $\mathbf{X}_i\beta$. Latent class mixed models assumes the population is heterogeneous and trajectories are governed by a set of latent classes of subjects which are characterised by the associated mean profiles [29]. We are interested in uncovering if there are distinct groups that have different trajectories through time, hence making this method appropriate – i.e finding homogeneous groups in a heterogeneous sample.

Each profile belongs to only one of G latent classes. Profiles are assigned to latent class $g \in \{1, 2, ..., G\}$ by probability - π_{ig} - which is obtained using a multinomial regression model [27] [14]. Essentially, finding the latent grouping that the profile is most associated with. If the variable c_i is defined to be a discrete random variable which equals g if profile i belongs to latent class g, then probability π_{ig} is given by Equation 3.4.

$$\pi_{ig} = P(c_i = g|x_i) = \frac{e^{\mathbf{x}_i'\mathbf{b}_g}}{\sum_{j=1}^{G} e^{\mathbf{x}_i'\mathbf{b}_j}} \tag{3.4}$$

where class membership is determined by elements of design vector x_i, and b_g are parameters associated with class g. A posterior probability for class membership is calculated for each subject. Membership is determined by class associated with the greatest posterior probability.

Equation 3.5 shows the general form of a latent class linear mixed model with N observations for class g:

$$Y_{i|c_i=g} = \mathbf{X}_i\beta + \mathbf{H}_i\lambda_g + \mathbf{Z}_i\mathbf{b}_{ig} + e_i, \quad for \ \ i = 1, ..., N. \tag{3.5}$$

where λ_g are the class-specific fixed effects for group g and $\mathbf{H_i}$ is the design matrix for the class-specific fixed effects. Y_i denotes the response value for observation i; $\mathbf{X}_i$ contains information on the k common fixed effects for individual i - the i^{th} row of the design matrix relating to fixed effects $\mathbf{X}_{n\times k}$; $\beta_{k\times 1}$ is the vector of fixed effect parameters. $\mathbf{Z}_i$ contains information on the m random effects for individual i - the i^{th} row of the design matrix relating to random effects $\mathbf{Z}_{n\times m}$; $\mathbf{b}_{m\times 1}$ is the vector of random effects for observation i in class g [29].

The R package `lcmm` is used to fit these models, and parameter estimation was performed using maximum likelihood estimation with an iterative Marquardt approach [29]. A critical aspect of latent class modeling is determining the optimal number of clusters. To assess this, we employed the following criteria: (1) **Akaike Information Criterion (AIC)**: AIC is used to compare models, penalizing excessive complexity. Lower AIC values indicate better-fitting models with fewer unnecessary parameters, (2) **Bayesian Information Criterion (BIC)**: Similar to AIC, BIC evaluates model fit while applying a stronger penalty for model complexity, making it more conservative in selecting the optimal number of clusters, (3) **Weighted Root Mean Squared Error (WRMSE)**: This measure quantifies the discrepancy between observed and predicted values while accounting for sample size and distribution of data points. A lower WRMSE signifies a more accurate model fit, and (4) **Weighted Mean Absolute Error (WMAE)**: Unlike WRMSE, WMAE focuses on the absolute differences between observed and predicted values, minimizing the influence of extreme deviations. A lower WMAE indicates better model performance.

Together, these metrics guided the selection of the optimal number of latent classes, ensuring a balance between model accuracy and complexity.

3.4 Functional Principal Component Analysis

A technique that is growing in popularity in the analysis of longitudinal data is functional principal component analysis (FPCA). In traditional longitudinal data analysis (LDA), expected values are typically represented as simple functions of time (like polynomials or non-linear functions). More specifically in this context (Chapter 5) sine and cosine functions are used to capture the seasonality of platelet counts - while functional methods offer more flexibility. In this section we begin by explaining FPCA through a brief exploration of Functional Data Analysis and Principal Component Analysis, before investigating the intricacies of FPCA and its underlying mechanisms. In Chapter 4 it is shown that the data exhibits sparsity. There is, therefore, the need for a technique that can effectively estimate principal components for sparse functional data. FPCA can be used to find the most significant trends in the data - principal curves.

Principal Component Analysis

Traditional principal component analysis in multivariate statistics involves reducing the dimensionality of data by projecting higher-dimensional data to the direction with the greatest variation in the data. Representing high-dimensional data in lower-dimensional space has numerous benefits and uses; most importantly, in the context of this paper, we are able to remove the noise and retain only the most important information. By extension, this encourages the ability to understand and visualise the complex structure of the data.

Principal components are linear combinations of the variables in a dataset. These linear combinations are aimed at depicting maximum variance. Essentially, we are projecting $\mathbf{X}$ to the direction $\mathbf{a}$ - the direction which captures the most variance is captured. Suppose you have some $\mathbf{X}_{k\times 1} = [X_1, X_2, ..., X_k]'$ with covariance $\mathbf{\Sigma}$, then we wish to maximise the variance of $Y = \mathbf{a'X} = a_{11}X_1 + ... + a_{1k}X_k$ subject to $\mathbf{a'a} = 1$ [1], where a_{ij} is the j^{th} loading of the i^{th} principal component. Through finding $\mathbf{a}$ that maximises $Var(Y) = \mathbf{a'\Sigma a}$, would mark the first principal component. Subsequent components are computed in the same way, however, it requires that each linear combination is independent of the others. Thus an additional constraint that $\text{Cov}(\mathbf{a}_i'\mathbf{X}, \mathbf{a}_k'\mathbf{X}) = 0$, $\forall k < i$ is required.

The result is a transformation of correlated features - X_i - into linearly independent components - principal components.

Functional Data

Functional data refers to data that are represented as continuous functions rather than discrete points. A key concept of FDA is that the profiles are smooth curves and values for the profile exist at any time point, but are only captured at discrete time points. In the context of this paper, we have platelet count profiles that are recorded only when a donor donates blood. The profiles of individual donors will

[1] This constraint exists so that the variance of Y has an upper limit. The components of $\mathbf{a}$ can increase without bound and, by extension, the variance of Y would be arbitrarily large

vary in both the number and frequency of donation, and the specific dates on which they donated blood. Using FDA, it is presumed that each profile - P_n, recorded at discrete time points $t_{n,j}$ (where $t_{n,j}$ denotes the discrete time points for donations by donor n) - can be viewed as smooth curves - $P_n(t)$ [21].

Since it is presumed that the data are derived from a smooth function, it is intuitive to think that each profile could be represented as a linear combination of smooth functions. It is both intuitive and convenient in the sense that there is not necessarily the time nor resources to explore known functional forms that best fit the problem. Thus, a collection of smooth basis functions to use as building blocks to represent each curve is needed:

$$P_n(t) \approx \sum_{k=1}^{K} \alpha_{kn}\phi_k(t), \forall n. \tag{3.6}$$

The building blocks - ϕ_k - are *basis functions*. Common basis functions used are splines and Fourier series. Each curve is then characterised by α_n.

Functional Principal Component Analysis

Functional principal component analysis, first introduced in [31], is an important tool for understanding functional data and uncovering important temporal patterns. Similar to principal component analysis, the top N functional principal components $\phi_1(t), ..., \phi_N(t)$ are to be found. These curves will summarise the major sources of variations among multiple curves $X_i(t)$. Each $X_i(t)$ is approximated by a linear combination of the top N functional principal components

$$X_i(t) = \mu(t) + \sum_{n=1}^{N} s_{in}\phi_n(t) \tag{3.7}$$

where s_{in} refers to the FPC of the i^{th} profile associated with the n^{th} FPC [2].

The data used for FPCA are centered (the mean curve - $u(t)$ - is subtracted from each $X_i(t)$) to emphasise the difference in variance from the mean curve - to detect the most important modes of variation in the n curves. The mean curve is computed by calculating the average of the n curves at each time t. Similar to PCA, the goal is to maximise

$$Var(\int (X_i(t) - \mu(t))\phi(t)dt) \tag{3.8}$$

subject to

$$\int \phi(t)^2 dt = 1. \tag{3.9}$$

[2]Sometimes referred as the loading of the i^{th} profile onto the n^{th} FPC.

This would provide the first principal curve - $\phi_1(t)$ - which represents the most important mode of variation. Subsequent principal curves are computed in the same way, however, we require that the principal curves are orthogonal. Thus an additional constraint is required

$$\int \phi_i(t)\phi_j(t)dt = 0. \tag{3.10}$$

The data in this research are sparse, thus a method for sparse functional data is required.

PACE Algorithm

Traditional FPCA requires a large amount of regularly spaced data. However, sparse data is observed. Therefore, we require a method to deal with this. In [40] this issue is addressed and a method termed principal component analysis through conditional expectation (PACE) is developed.

As per [40] FPCs and FPC scores are estimated in the following way. Denote Y_{ij} to be the j^{th} observation of profile i made at time point T_{ij}. ϵ_{ij} are the measurement errors and are assumed to be i.i.d. with a mean of 0 and a variance of σ^2.

$$Y_{ij} = X_i(T_{ij}) + \epsilon_{ij} \tag{3.11}$$

$$X_i(T_{ij}) = \mu(T_{ij}) + \sum_{n=1}^{N} s_{in}\phi_n(T_{ij}) + \epsilon_{ij}. \tag{3.12}$$

$X_i(t)$ is an independent realisation of a smooth random function $X(t)$ with unknown common mean function $\mu(t)$. First, $\hat{\mu}(T_{ij})$ is estimated using local linear smoothers [3]. Now define sample (raw) covariance matrix

$$G_i(T_{ij}, T_{il}) = (Y_{ij} - \hat{\mu}(T_{ij}))(Y_{il} - \hat{\mu}(T_{il})) \tag{3.13}$$

i.e for each time point T_{ij} and T_{il} we can compute the raw covariance G_i by Equation 3.13. Since we assume the errors are i.i.d. with mean 0 and variance σ^2, from Equation 3.11 we can see that $cov(Y_i(T_{ij}), Y_i(T_{ij})) = cov(X_i(T_{ij}), X_i(T_{ij})) + \delta_{jl}\sigma^2$. Where $\delta_{jl} = 1$ if $j = 1$ and 0 otherwise. It can also be seen that $E(G_i(T_{ij}, T_{il})) = cov(X_i(T_{ij}), X_i(T_{ij})) + \delta_{jl}\sigma^2$. Since $E(G_i(T_{ij}, T_{il}))$ is equal to $cov(X_i(T_{ij}), X_i(T_{ij}))$ ($\delta_{jl} = 0$ off diagonal), the off diagonal entries of the sample covariance matrix - $G_i(T_{ij}, T_{il})$ - are used in the estimate of the smoothed covariance surface $\hat{G}_i(T_{ij}, T_{il})$ using local quadratic smoother. The estimate of σ^2 is

$$\hat{\sigma^2} = \frac{2}{T}\int_{T_1} \hat{V}(t) - \hat{G}(t,t)dt \tag{3.14}$$

[3] mean, covariance, and eigenfunctions are assumed smooth

where T is the time frame range, T_1 is a subset of T with boundary values cutoff to mitigate boundary effects, and $\hat{V}(t)$ is a smoothed version (estimated using local linear smoother) of the diagonal elements of $G_i(T_{ij}, T_{il})$.

The FPCs are the eigenfunctions of the following eigenequation

$$\int \hat{G}(s,t)\phi_k(s)ds = \lambda_k\phi_k(t) \tag{3.15}$$

with additional constrains as before $\int \phi_k^2 dt = 1$ and $\int \phi_k\phi_n dt = 0$ for $n < k$.

Traditionally, FPC scores are estimated using integration:

$$\hat{s}_{ik} = \int (X_i(t) - \hat{\mu}(t))\hat{\phi}_k(t)dt. \tag{3.16}$$

However, with sparse functional data the above does not result in a reasonable explanation since observed $X_i(t)$ are not available. Scores are given by the conditional expectation of the score on the observed data Y_i:

$$\hat{s}_{ik} = E(s_{ik}|Y_i) = \lambda_k\phi_{ik}\Sigma_{Y_i}^{-1}(Y_i - \mu_i). \tag{3.17}$$

3.5 K-means Clustering

In Chapter 6, donors are grouped based on scores obtained from the method in the previous section. K-means is a non-parametric algorithm to classify observations into groups. Clustering is a fundamental technique in unsupervised learning, used to group similar observations based on shared characteristics. It aims to uncover hidden structures in data, identifying natural groupings without predefined labels. Among various clustering methods, K-means stands out as a robust, nonparametric approach that is particularly useful when the underlying data distribution is unknown or does not adhere to strict parametric assumptions. K-means operates by minimising within-cluster variance, making it flexible and computationally efficient in many practical scenarios. Its simplicity and speed allow for scalability, especially when clustering high-dimensional data. K-means excels in clear-cut partitioning, ensuring interpretable and well-separated clusters without requiring prior knowledge of distributional properties.

The K-means algorithm requires the selection of several clusters – k. Ultimately, each observation is assigned to one of these k clusters, to minimise the total within-cluster variation in equation 3.18.

$$SS_{within} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}d(x_j, \mu_i) \tag{3.18}$$

where k denotes the number of clusters chosen, n_i is the number of observations allocated to cluster i, z_{ij} is a binary variable associated with cluster allocation and takes on the value of 1 if observation j is allocated to cluster i, and $d(x_j, \mu_i)$ is the distance between observation j – x_j – and the mean value of all points in cluster i – μ_i. Through minimising this measure, it is ensured that observations within each cluster are similar – the observation is relatively close to the mean observation. In the context of this paper, the distance measure is defined as Euclidean distance.

To achieve a minimum, the algorithm starts with the chosen number of clusters – k – and chooses k random points to act as the centroids for each cluster. All remaining observations are then assigned to the closest centroid based on the distance measure chosen. The new centroids for each cluster – μ_i – are then computed, and observations are reassigned based on proximity to the newly calculated centroids to minimise SS_{within}. This process with recalculating μ_i and reassignment of observations is iterated until there is no longer a significant difference in the clustering [19].

Chapter 4

Data Wrangling and Exploratory Data Analysis

4.1 Introduction

The Apheresis Donor Dataset contains information on individual donations through time. In this chapter, data relating to individual donor platelet counts are explored. This is done by first describing the contents of the dataset. Platelet trajectories are explored by plotting mean curves through time. In addition to the plotting of mean platelet profiles, time series methods to assess the presence of seasonality are used. Additionally, for subsequent analysis in Chapter 5 and Chapter 6 we required that the data are approximately normally distributed. In this chapter, necessary transformations to the data are detailed, as well as any other data issues and manipulations.

The platelet products dataset is also examined, analysing the trajectory of the percentage of products passing by plotting the pass rate over time. Additionally, time series methods and Generalized Linear Mixed-Effects models are employed to assess trends and seasonal patterns in product failures, conducting this analysis to confirm insights received from SANBS.

4.2 The Apheresis Donor Dataset

Table 4.1 summarises the variables available in the Apheresis Donor Dataset. Platelet count pre-donation is used to determine platelet profiles. A number of factors influence the platelet count post-donation. Therefore, it is not useful to use post-donation platelet count to determine platelet profiles [9]. Other important variables available in the table that are relevant to this research include donor ID, sex, ethnicity, donation date, and branch zone. While the donation product is available in the data provided by SANBS, it is not relevant for this analysis.

Table 4.1: Variable descriptions: Apheresis Donor Dataset.

Variable	**Description**
donor_id	Unique identifier for individual donors
sex	Sex of donor
ethnic	Race of donor
donation_product	Product donated (PLCA, PLASPLCA)
PLTPRE	Platelet count pre-donation
PLTPOST	Platelet count post-donation
donation_date	Date of donation
branch_zone	Branch at which donation was made

The raw dataset contains information on 7174 unique donors over a 5-year period (2 January 2016 to 31 December 2020). It is decided to remove all donations that do not have a donation date, and to remove donations with other missing information - 5714 unique donors remain. There are 1587 individual donation dates. Naturally, each individual does not donate on every donation date. Therefore, the data are sparse. Figure 4.1 shows a histogram of the number of donations (left) as well as the number of donations on each donation date (right).

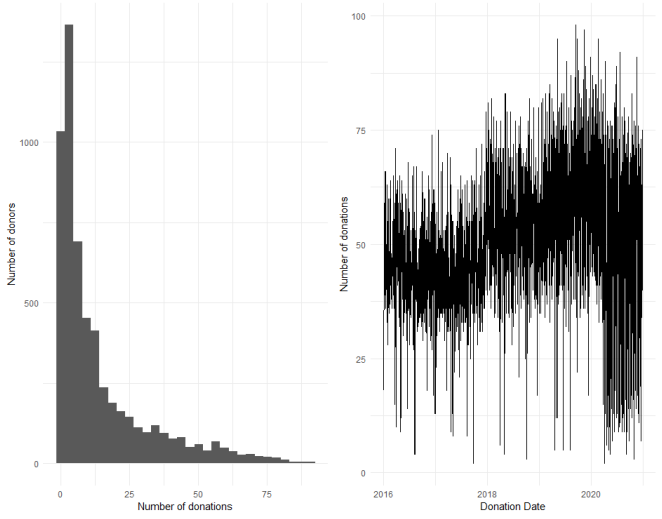

Figure 4.1: Histogram of the number of donations per donor (left) and a line graph of the number of donations on each donation date over time (right). The histogram highlights the data sparsity, showing that most donors have relatively few donations, while the line graph illustrates the low number of donations relative to the total number of donors in the dataset.

It is clear that the data are sparse – donors are not donating frequently. Furthermore, when considering the mean platelet profile across all donors, noise is observed - the mean profile is choppy. To combat this, it is decided to group platelet counts by month and donor, and average within groups in an effort to smooth the profile. This leaves 5×12 donation dates ranging from January 2016 to December 2020 (12 months for every 1 of 5 years). A spaghetti plot for the first 30 donors in the dataset

(donors are in no specific order) can be seen in Figure 4.2.

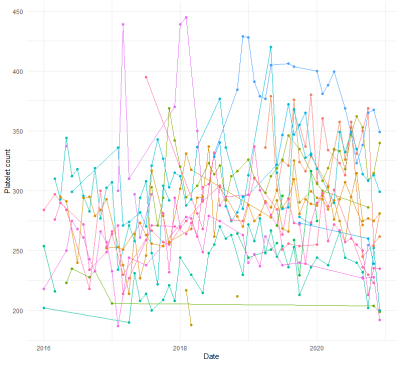

Figure 4.2: Spaghetti plot depicting the profiles of 30 apheresis donors. Each line represents an individual donor's platelet count over time, illustrating the variability in donation patterns and platelet dynamics among the donors.

Of interest is factors affecting the level and trajectory of platelet count. Therefore, we explore the average platelet count for each of the categorical variables present in the data. Namely, sex, ethnicity, and branch zone.

Interestingly, when considering plots related to sex in Figure 4.3, it can be seen that the average platelet count for males and females follows almost the same trajectory. However, females, on average, have a higher platelet count than males. Furthermore, when considering plots related to ethnicity, the mean platelet count is similar in both trajectory and level. Lastly, considering the plots related to the branch zone, the mean platelet count curves generally have the same trajectory and level.

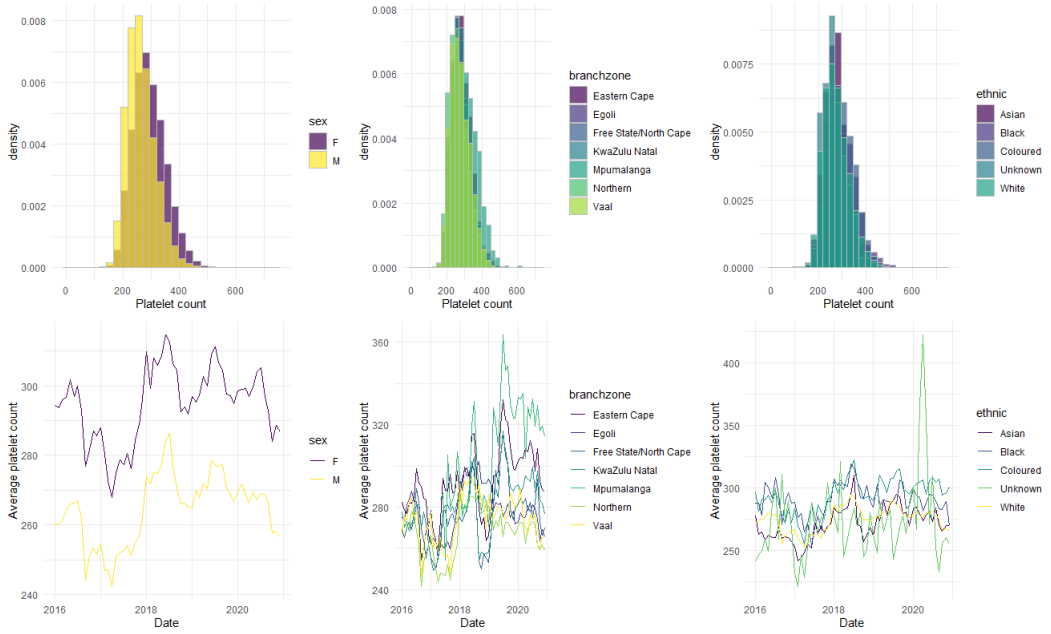

Figure 4.3: Top: Histrograms of platelet count grouped by sex (left), branch zone (middle), and ethnicity (right). Bottom: Average platelet profiles grouped by sex (left), branch zone (middle), and ethnicity (right).

However, in the latter half of the period considered, there seems to be a greater difference in the average level of platelet count across the branch zones. Considering the average platelet count graphs in Figure 4.2 and Figure 4.3, a cyclical pattern

emerged in the data. It seemed that the average platelet counts spike in the middle of the year. However, there is one year that is an exception to that rule – 2017. In 2017 a different shape is seen relative to other years. It is evident that average platelet count levels were lower in that year. Zooming in on the average profiles of the branch zones, it can be seen that in the first year, the average profiles seem mostly homogeneous without any notable, obvious differences.

In year 2, however, a difference in the trajectories can be seen. For example, Northern and Vaal regions depict lower average platelet count levels in the middle of the year while Eastern Cape and Mpumalanga have higher average levels. The clearest observation, however, is in the last year where greater variability exists among branch zones. Moreover, upon examining the updated data, a positive skew is observed. Consequently, the data is transformed using a logarithmic scale [1]. It should be noted that we are working with count data. Log-transforming the data addresses distributional and continuity assumptions necessary for modelling in later chapters.

In Table 4.2, the donors by ethnicity, sex, and branch zone are tabulated. Due to the unbalanced nature of the data, it is decided not to consider ethnicity as a factor in the subsequent analysis – some groups have no profiles. By focusing on groups with sufficient data, we ensure the reliability and interpretability of the analysis. The only categorical donor information considered in future analysis is the donors' sex and branch zone.

Table 4.2: Number of observations in each group: Categorical Variables Crosstabulation.

		Asian	Black	Coloured	Unknown	White
F	Eastern Cape	2	8	14	5	114
	Egoli	18	73	31	8	766
	Free State/North Cape	1	8	13	1	198
	KwaZulu-Natal	20	18	11	4	294
	Mpumalanga	1	2	0	0	55
	Northern	6	32	13	1	723
	Vaal	1	14	4	0	319
M	Eastern Cape	0	6	18	2	89
	Egoli	52	101	54	4	786
	Free State/North Cape	3	8	17	0	165
	KwaZulu-Natal	66	29	17	7	267
	Mpumalanga	0	10	0	1	41
	Northern	18	54	17	0	706
	Vaal	7	32	11	0	348

[1] This transformed variable is hereinafter referred to as platelet count, or average platelet count when discussing means.

An aim of this research is to answer if seasonality in platelet count exists. To further explore seasonality, in Figure 4.4, boxplots of platelet count for each donation date are plotted. It is evident that platelet counts, on average, stay around the same level. There is a noticeable pattern: platelet counts do not fluctuate randomly from month to month; instead, a cyclical pattern emerges. The levels oscillate around the mean value, indicating a consistent trend rather than random variation.

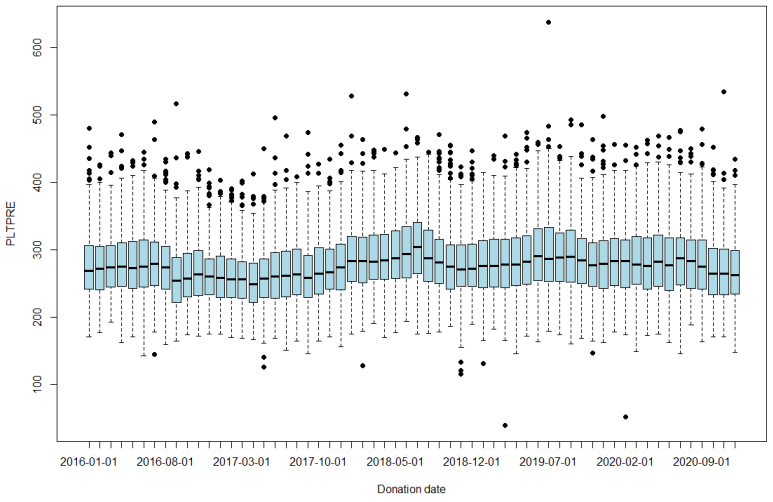

Figure 4.4: Boxplots of platelet count on each date.

To further explore the seasonality in the data, time series methods are used. The time series of average platelet count – the mean profile – is decomposed into distinct components, namely, trend , seasonality and error (Equation 4.1). A simple additive model is used in the decomposition as there is no evidence to suggest a multiplicative structure

$$Y(t) = T(t) + S(t) + e(t). \tag{4.1}$$

The trend component is derived through a moving average calculation, capturing the overall direction or pattern in the data. The seasonal component is computed by averaging values across all periods for each time unit, and subsequently centered. The residual component, representing the remaining variability is encapsulated by the error term. Figure 4.5 shows the decomposition of the average platelet count time series. Considering the seasonal component, it is seen that a seasonal high is evident in the winter months June/July. Additionally, we look at the autocorrelation function (ACF) plot of the detrended series. This can be seen in Figure 4.6. The ACF plot shows the linear correlation between observations at lagged time points – answering the question: can one linearly predict the value of a future time point given the value of a previous time point [34].

Based on literature and previous findings in this chapter, it is hypothesised that platelet levels exhibit annual seasonal variations, aligning with changes in the calendar year. Consequently, in the presence of seasonality within the dataset, observing

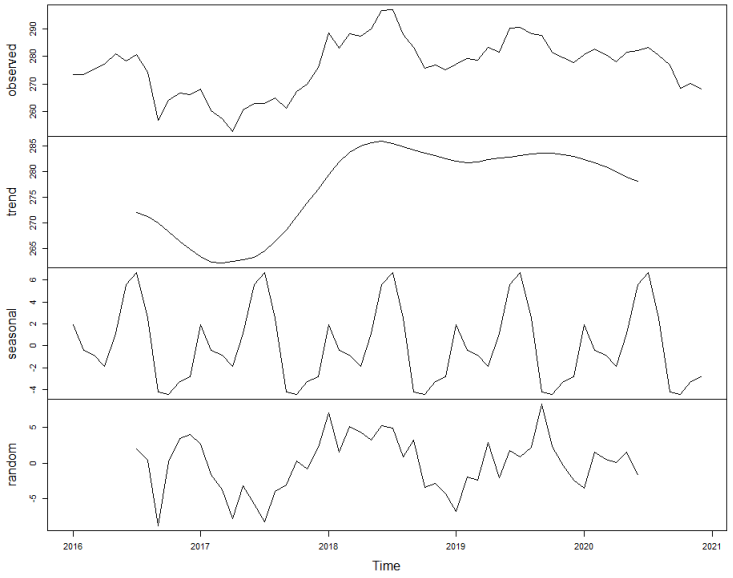

Figure 4.5: Decomposition of average platelet count time series.

periodic spikes occurring at yearly intervals is anticipated, typically every 12 months, within the ACF plot. From the ACF plot in Figure 4.6, evidence of yearly seasonality can be seen.

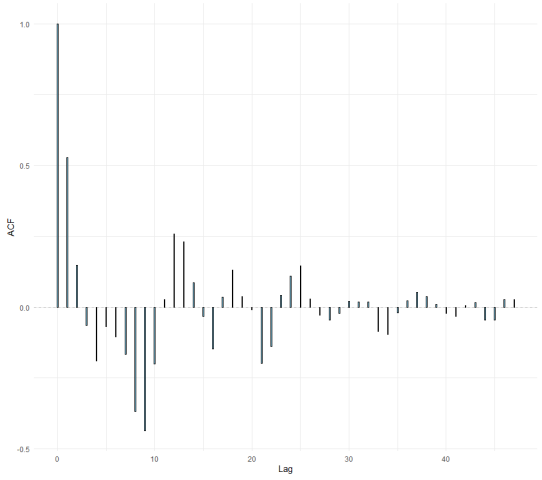

Figure 4.6: ACF plot of the detrended average platelet count profile.

A summary of the final data to be used can be seen in Table 4.3.

Table 4.3: Summary of final platelet count data by year. This table presents the mean and standard deviation of platelet counts for each year, along with the number of male and female donors. Additionally, it includes the count of donations from each branch zone by year.

Year	2016	2017	2018	2019	2020
(# obs)	(2509)	(2546)	(2512)	(2505)	(2341)
Average Platelet Count					
Mean	5.58	5.55	5.63	5.61	5.60
(SD)	(0.194)	(0.188)	(0.193)	(0.209)	(0.194)
Sex					
Female	1128 (45.0%)	1158 (45.5%)	1117 (44.5%)	1110 (44.3%)	1065 (45.5%)
Male	1381 (55.0%)	1388 (54.5%)	1395 (55.5%)	1395 (55.7%)	1276 (54.5%)
Branch Zone					
Eastern Cape	195 (7.8%)	181 (7.1%)	165 (6.6%)	151 (6.0%)	133 (5.7%)
Egoli	838 (33.4%)	815 (32.0%)	669 (26.6%)	793 (31.7%)	699 (29.9%)
Free State/North Cape	189 (7.5%)	183 (7.2%)	248 (9.9%)	218 (8.7%)	225 (9.6%)
KwaZulu Natal	349 (13.9%)	332 (13.0%)	365 (14.5%)	385 (15.4%)	297 (12.7%)
Northern	709 (28.3%)	690 (27.1%)	649 (25.8%)	602 (24.0%)	652 (27.9%)
Vaal	229 (9.1%)	316 (12.4%)	366 (14.6%)	295 (11.8%)	277 (11.8%)
Mpumalanga	0 (0%)	29 (1.1%)	50 (2.0%)	61 (2.4%)	58 (2.5%)

4.3 The Platelet Products Dataset

SANBS pools whole blood donor donations to form platelet products. A second dataset is made available comprising of 189,527 rows and 15 columns, documenting information about products derived from whole blood donors, distinct from those obtained through apheresis[2]. The dataset contains information on the outcome of three tests: (1) sterility, (2) volume, and (3) yield.

An objective was to analyse whether information in the apheresis donor dataset could explain product failures within this dataset. A product was deemed to have failed if its sterility, volume, or yield did not meet specific predefined criteria. Specifically, success is measured against the following standards: (1) platelet yield, should be at least 2.4×10^{11} platelets, reflecting the yield of platelets per unit; (2) sterility, where a negative result is non-negotiable, ensuring the product's purity and safety; and (3) volume, falling within the range of 200 to 800 milliliters. Importantly, the only common variables shared between this dataset and the platelet count dataset was the branch zone and time. Sterility, which denotes contamination of the product post-donation or during production, was omitted from this analysis. This decision was made because a failed sterility test is typically associated with product contamination rather than issues stemming from donors themselves. The final dataset comprised of variables associated with production date, volume outcome, yield outcome, and branch zone. If both yield outcome and volume outcome were successful, the product was marked as passed; otherwise, it was recorded as a failure.

Table 4.4 presents a summary of the pooled products dataset. The data indicates that product failures in the volume test are exceedingly rare, with only two instances of failure recorded over the 5-year period. Additionally, the data reveals that a higher volume of products is produced in the Egoli, Northern, and KwaZulu-Natal regions. This trend is consistent with the distribution of donations from the apheresis donor dataset, where these branch zones exhibit higher donation frequencies. It is also noted that a small percentage of products are tested – roughly around 5%.

Of interest is the trend in proportion of products that were tested passing. Figure 4.7 shows the percentage of products passed on each collection date [3]. On the left of Figure 4.7 the proportion passed of all products can be seen, and on the right the proportion passed within each branch zone can be seen. It is evident that the proportion passed is generally high ($> 90\%$), however, post 2019 a decreasing trend in proportion passed can be seen.

Table 4.5 summarises the number and percentage of products that passed and failed over time across the different branch zones. Table 4.5 and Figure 4.7 both show relatively stable and high pass rates up until 2019. The most notable decrease in products failing is observed in the Eastern Cape, Free State, and Northern branch zones. KwaZulu Natal and Vaal regions seem to have experienced no decrease in the pass rate. Evidence of seasonality can be seen in Figure 4.7, where the pass rate

[2]No information is available for individual whole blood donors.

[3]The initial dataset had daily collection dates. It was decided to group collection dates by month as was done with the platelet count dataset.

Table 4.4: Summary of pooled platelet products data. This table presents the number of products categorised by test outcomes—pass, fail, and not tested—for both volume and yield tests across different years. It also includes the number of products produced by year and branch zone.

Year	2016(N=37729)	2017(N=39285)	2018(N=37423)	2019(N=37242)	2020(N=37847)
Volume Test Outcome					
Not tested	36119 (95.7%)	37594 (95.7%)	35685 (95.4%)	35474 (95.3%)	36037 (95.2%)
Pass	1610 (4.3%)	1691 (4.3%)	1736 (4.6%)	1768 (4.7%)	1810 (4.8%)
Fail	0 (0%)	0 (0%)	2 (0.0%)	0 (0%)	0 (0%)
Yield Test Outcome					
Fail	83 (0.2%)	101 (0.3%)	50 (0.1%)	155 (0.4%)	232 (0.6%)
Not tested	36119 (95.7%)	37595 (95.7%)	35699 (95.4%)	35476 (95.3%)	36052 (95.3%)
Pass	1527 (4.0%)	1589 (4.0%)	1674 (4.5%)	1611 (4.3%)	1563 (4.1%)
Branch Zone					
Eastern Cape	1595 (4.2%)	1849 (4.7%)	1697 (4.5%)	1641 (4.4%)	1632 (4.3%)
Egoli	6944 (18.4%)	6757 (17.2%)	6355 (17.0%)	5891 (15.8%)	5375 (14.2%)
Free State/North Cape	1754 (4.6%)	1764 (4.5%)	1813 (4.8%)	1961 (5.3%)	1909 (5.0%)
KwaZulu Natal	7977 (21.1%)	7328 (18.7%)	6709 (17.9%)	7373 (19.8%)	6824 (18.0%)
Mpumalanga	4003 (10.6%)	3852 (9.8%)	3652 (9.8%)	3405 (9.1%)	3879 (10.2%)
Northern	11019 (29.2%)	12493 (31.8%)	12283 (32.8%)	11185 (30.0%)	11098 (29.3%)
Vaal	4436 (11.8%)	5242 (13.3%)	4914 (13.1%)	5786 (15.5%)	7130 (18.8%)

Table 4.5: Pass and failure rates of platelet products by branch zone and year. The table summarizes the proportion of products that passed and failed across different branch zones over the specified years.

	Eastern Cape	Egoli	Free State/North Cape	KwaZulu Natal	Mpumalanga	Northern	Vaal
Outcome							
2016							
Fail	11 (3.5%)	20 (9.0%)	15 (7.7%)	6 (2.5%)	0 (0%)	11 (4.0%)	20 (11.6%)
Pass	300 (96.5%)	202 (91.0%)	181 (92.3%)	230 (97.5%)	196 (100%)	265 (96.0%)	153 (88.4%)
2017							
Fail	13 (4.0%)	23 (9.9%)	7 (3.6%)	4 (2.0%)	4 (2.0%)	26 (8.3%)	24 (11.1%)
Pass	313 (96.0%)	209 (90.1%)	190 (96.4%)	200 (98.0%)	199 (98.0%)	286 (91.7%)	192 (88.9%)
2018							
Fail	25 (8.4%)	4 (1.3%)	1 (0.5%)	6 (2.5%)	10 (4.0%)	3 (1.3%)	1 (0.5%)
Pass	271 (91.6%)	301 (98.7%)	209 (99.5%)	233 (97.5%)	238 (96.0%)	223 (98.7%)	199 (99.5%)
2019							
Fail	49 (19.1%)	27 (6.9%)	9 (4.5%)	10 (5.1%)	11 (5.1%)	39 (12.3%)	10 (5.2%)
Pass	208 (80.9%)	363 (93.1%)	190 (95.5%)	185 (94.9%)	206 (94.9%)	277 (87.7%)	182 (94.8%)
2020							
Fail	63 (21.5%)	49 (13.2%)	41 (18.8%)	9 (4.4%)	24 (10.3%)	44 (16.4%)	2 (1.0%)
Pass	230 (78.5%)	321 (86.8%)	177 (81.2%)	197 (95.6%)	208 (89.7%)	225 (83.6%)	205 (99.0%)

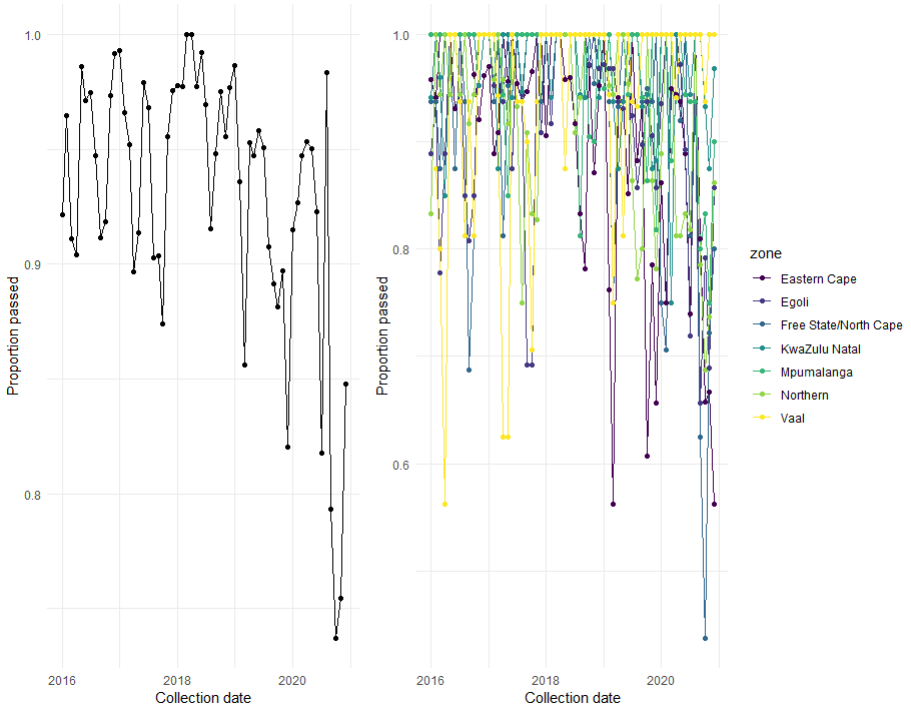

Figure 4.7: Proportion of products passed through time. Across all branch zones (left) and separated by branch zone (right).

appears to move up and down cyclically. To investigate this further, the time series of the percentage passed is decomposed, following the approach in Equation 4.1. This can be seen in Figure 4.8. A comparison can be made between the seasonal components in Figures 4.5 and 4.8. Interestingly, the trend components are also largely similar across the two figures, with the exception of the later years. In the percentage of products passed, a clear downward trend is observed, whereas this trend is not observed in the average platelet count.

To further investigate and confirm seasonality in product failures, we use a generalised linear mixed-effects model (GLMM) that accounts for the non-normal distribution of the binary response variable. Specifically, a logit link function is used to model the binary outcome. The primary focus in this model is to capture the overall seasonal pattern across different years. To achieve this, the year is treated as a random effect. This enables the recognition that seasonal trends may vary slightly from year to year. Using this approach, inter-year variability can be accounted for while still emphasizing the overarching seasonal trends. The model structure is detailed in Equation 4.2:

$$\begin{aligned}\text{logit}(\pi_{it}) = \beta_0 + \beta_1 \sin\left(\frac{2\pi \cdot \text{month}_t}{12}\right) + \beta_2 \cos\left(\frac{2\pi \cdot \text{month}_t}{12}\right) \\ + \left(\gamma_{1i} \sin\left(\frac{2\pi \cdot \text{month}_t}{12}\right) + \gamma_{2i} \cos\left(\frac{2\pi \cdot \text{month}_t}{12}\right)\right). \quad (4.2)\end{aligned}$$

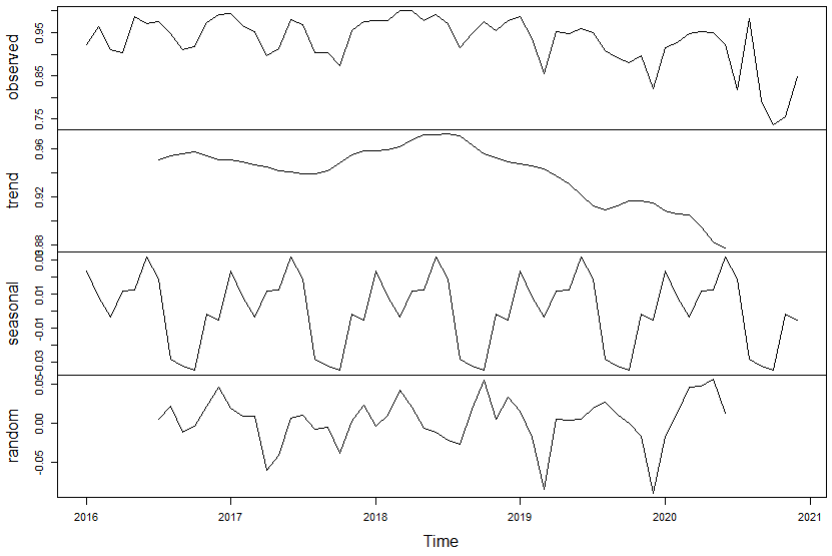

Figure 4.8: Decomposition of percentage of products passing time series.

In this equation, π_{it} represents the probability of the outcome for year i at time t, while $\text{logit}(\pi_{it}) = \log\left(\frac{\pi_{it}}{1-\pi_{it}}\right)$ denotes the log-odds of the outcome. The term β_0 is the fixed intercept, which serves as the baseline log-odds of a product passing. The coefficients β_1 and β_2 correspond to the fixed effects associated with the sine and cosine functions, respectively, capturing the seasonal patterns based on the month of the year. The random effects γ_{1i} and γ_{2i} account for the individual variation in these seasonal effects for each year i, allowing the model to flexibly represent the seasonal components that can change over time. The sine and cosine terms themselves, $\sin\left(\frac{2\pi \cdot \text{month}_t}{12}\right)$ and $\cos\left(\frac{2\pi \cdot \text{month}_t}{12}\right)$, encapsulate the cyclical seasonal patterns present in the data [4].

Table 4.6 presents the summary output from the fitted model. The coefficients $\beta_1 = 0.44$ and $\beta_2 = -0.15$ correspond to the peaks of the log-odds during the autumn-winter months. This suggests that there is a higher probability of a product passing during this period.

Table 4.6: Summary output of the generalised linear mixed-effects model (GLMM) analysing seasonality in platelet product failures. The table presents estimates, standard errors, z-values, and p-values for the fixed effects

	Estimate	Std. Error	z value	Pr(>|z|)
β_0	2.77	0.24	11.55	0.00
β_1	0.44	0.15	2.89	0.00
β_2	-0.15	0.12	-1.23	0.22

[4] As in Chapter 5, the angular frequency of sine and cosine terms are set to $\frac{2\pi}{12}$ to fix period to one year.

4.4 Concluding Remarks

In this chapter, an in-depth exploration of the apheresis donor dataset using various exploratory techniques was conducted. Key insights and considerations from the analysis include: (1) The data exhibits sparsity, prompting me to address this issue by aggregating observations based on donation dates into monthly intervals; (2) It was observed that the data deviate from a normal distribution and displays a positive skew. To address this, it was decided to utilise a log-transformed version of the platelet count variable for subsequent analysis; (3) the presence of data imbalance was identified, leading to the exclusion of the categorical variable associated with race from further analysis; (4) The examination also revealed apparent disparities in platelet count across different branch zones, although further investigation is required to explain these differences. Additionally, it was found that females have, on average, higher platelet counts than males; and (5) Finally, the analysis uncovered evidence of seasonality within the dataset.

The platelet products dataset was also analysed, and the key findings include: (1) evidence that the passing products exhibit similar seasonality to that observed in the apheresis donor dataset; (2) a sharp decrease in the percentage of products failing in the latter half of the dataset, a trend not reflected in the average platelet count profile; and (3) the reasons for the increased product failures in later years remain unclear based on the exploratory analysis of the platelet count data.

Chapter 5

Level and Temporal Dynamics in Platelet Count

5.1 Introduction

In this chapter, the objective is to uncover the temporal and level dynamics of platelet count using linear mixed-effect models. Linear mixed-effects models are used to uncover the relationship among covariate data, time, and the average platelet count level. In Chapter 3, linear mixed-effects models as a robust method for the analysis of longitudinal data is introduced, offering a comprehensive approach to understanding the interplay of variables over time. These models are used to unveil and interpret the patterns inherent in the data. This is done through exploring the model building and selection process, the model is then interpreted and tested for underlying assumptions associated with linear mixed-effect models.

5.1.1 Building the Model

From Chapter 4, it is noted that there is a seasonal component to trajectory of platelet count [1]. It is also noted that there is no trend element i.e platelet count oscillates about a mean level. To capture the seasonality in platelet count, sine and cosine curves are used:

$$\beta_{sin} sin(\omega \times \text{month}) + \beta_{cos} cos(\omega \times \text{month}). \tag{5.1}$$

Certainly, the aforementioned equation exhibits non-linearity concerning ω. Nevertheless, by holding ω constant, it reverts to a linear association. From an intuitive perspective, it is evident that the periodicity of the sine and cosine functions corresponds to an annual cycle. Consequently, the period (ω) of these functions is set to one year:

$$\omega = \frac{2\pi}{12} \tag{5.2}$$

[1] Recall that platelet count is log-transformed.

This assumption aids in the interpretation of the model, and simultaneously decreases the complexity.

At this point we recall Chapter 4. It was seen that generally, each year had a similar seasonal pattern. However, in 2017 a different pattern emerged. To further explore this, we fit separate linear models for each year:

$$\text{Platelet Count} \sim sin(\omega \times \text{month}) + cos(\omega \times \text{month}). \tag{5.3}$$

Plotting the mean curves for each of the years (Figure 5.1) , it is evident that there is interaction between year and the sine and cosine curves capturing the seasonality.

[h]

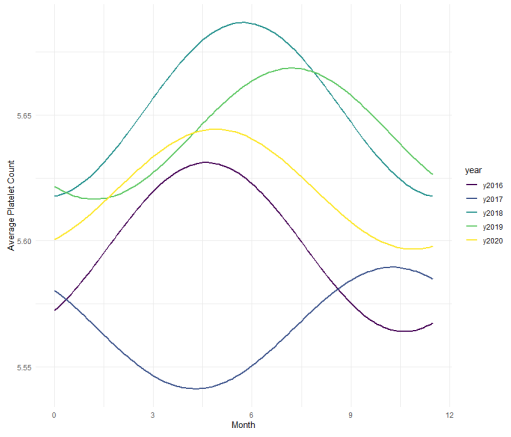

Figure 5.1: Separate linear models for each year: Seasonal curves.

At this point it is important to note that we are attempting to model why, how, and when platelet counts are changing. As well as determining if they change due to the covariates in the data, we also aim to uncover any differences between the years. It is understood that while there is a normal range of platelet count levels, average platelet count levels differ between individuals. Naturally, one would assume that the model should have a random intercept. Nevertheless, linear models for each donor are fit separately in the same it was done in equation 5.3.

It can be seen that intercept values vary largely as well as seasonal components among individual donors. It is, therefore, included random intercepts and random slopes by donor in the model. Similar analysis was done for branch zone and gender. No significant differences in seasonal components between sexes were seen. This covariate will be treated as only a fixed effect in the model. We found different seasonal components for branch zones as well. Additionally, from the exploratory data analysis it is noted that branch zones behave mostly the same in the first part of the period but have different behaviours in the latter half. From the analysis and patterns observable in the data, it shows that not only are different seasonal patterns every year evident, but seasonal patterns in branch zones also differ.

Due to the sparsity of the data we are unable to fit models with complex random effect structures. While the structure of the model is mostly decided on based on previous findings, 4 models with different fixed and random effect structures are explored – *good enough* models that use all of the data and is estimable with the data in hand [7].

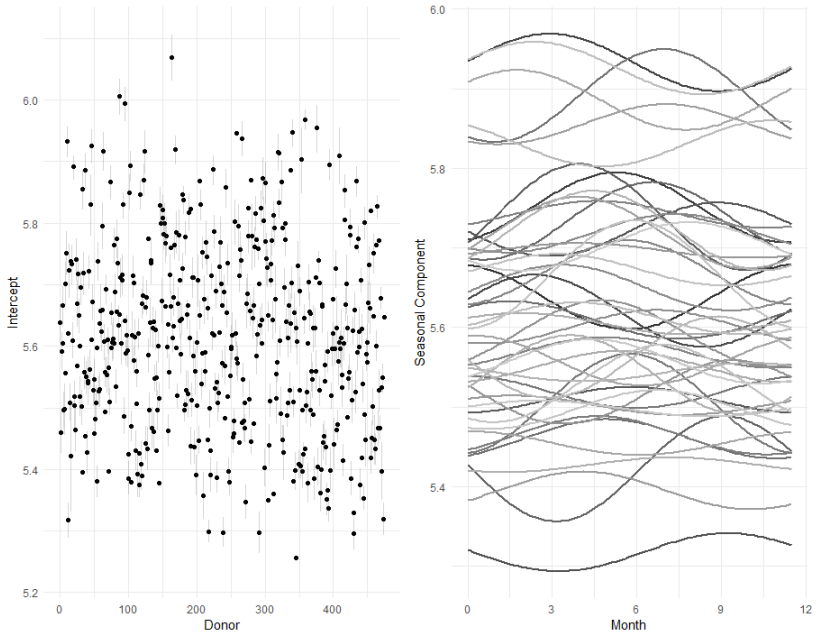

Figure 5.2: Separate linear models for each donor: Intercepts (left) and seasonal components (right).

Model 1

In this model, platelet count is modelled with fixed effects for the interaction between `year` and the seasonal component captured by sine and cosine curves (Equation 5.1). We are also interested in the effect of sex and branch zone, and therefore include these variables as fixed effects. This model includes only a random intercept.

Model 1A

This model has the same fixed effect structure as Model 1. However, random slopes for the seasonal components are added. [2]

Model 2

In this model, the fixed effect structure was changed to include interaction between year, branch zone, and the seasonal component (year$\times$branchzone$\times sin(\omega \times$month$)+$ year $\times$ branchzone $\times cos(\omega \times$ month$)$). The random effect structure is the same as in Model 1.

Model 2A

This model had the same fixed effect structure as Model 2 and the same random effect structure as Model 1A.

Based on the exploratory data analysis and comparisons it was expected that model 2A would best fit the data. This is supported in Table 5.1. where a test is done to

[2] As discussed previously, We could not fit a more complex structure.

address the balance between model complexity and fit, an important consideration in modeling. This approach is important in deciding whether the inclusion of additional parameters in a model significantly improves its explanatory power.

The output table summarises metrics such as Akaike Information Criterion, Bayesian Information Criterion, Loglik (Log Likelihood), Deviance, and the χ^2 statistic. It includes the associated p-value, testing the null hypothesis that the more complex model is not significantly better at capturing the data. These metrics collectively uncover the trade-off between model fit and complexity. It can be seen that Model 2A is significantly better than the other models.

Table 5.1: Summary of model comparison metrics for different models, including Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Log Likelihood (LogLik), Deviance, and χ^2 statistics. The table evaluates the balance between model complexity and fit, supporting the selection of Model 2A as the best fit for the data.

	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
model1	-90793.45	-90575.36	45420.73	-90841.45			
model1A	-91220.96	-90957.43	45639.48	-91278.96	437.51	5	0.0000
model2	-94815.04	-93860.88	47512.52	-95025.04	3746.07	76	0.0000
model2A	-94865.83	-93866.24	47542.92	-95085.83	60.79	5	0.0000

The final model formulation can be seen in Equation 5.4.

$$\begin{aligned}\log(\text{Platelet Count})_{ijk} = \beta_0 + \beta_1 \text{branchzone}_i + \beta_2 \text{year}_k + \beta_3 \sin(\omega \times \text{month}_j) \\ + \beta_4 \cos(\omega \times \text{month}_j) + \beta_5(\text{branchzone}_i \times \text{year}_k \times \sin(\omega \times \text{month}_j)) \\ + \beta_6(\text{branchzone}_i \times \text{year}_k \times \cos(\omega \times \text{month}_j)) + \beta_7 \text{sex}_i \\ + \gamma_{0i} + \gamma_{1i} \sin(\omega \times \text{month}_j) + \gamma_{2i} \cos(\omega \times \text{month}_j) + \epsilon_{ijk}. \quad (5.4)\end{aligned}$$

In this model, β_0 represents the overall intercept for platelet count. The term $\beta_1 \text{branchzone}_i$ accounts for the effect of the geographic branch zone of the ith individual, while $\beta_2 \text{year}_k$ captures the effect of the year k. The terms $\beta_3 \sin(\omega \times \text{month}_j)$ and $\beta_4 \cos(\omega \times \text{month}_j)$ represent the seasonal effects using sine and cosine components, respectively. The interaction terms $\beta_5(\text{branchzone}_i \times \text{year}_k \times \sin(\omega \times \text{month}_j))$ and $\beta_6(\text{branchzone}_i \times \text{year}_k \times \cos(\omega \times \text{month}_j))$ model the combined effects of branch zone and year interacting with seasonal patterns. The term $\beta_7 \text{sex}_i$ reflects the effect of the sex of the ith individual. The random effects γ_{0i} capture individual-specific deviations in platelet count, while $\gamma_{1i} \sin(\omega \times \text{month}_j)$ and $\gamma_{2i} \cos(\omega \times \text{month}_j)$ account for individual-specific seasonal deviations. Finally, ϵ_{ijk} represents the residual error, which is assumed to be normally distributed with mean 0 and variance σ^2.

5.2 Model Results

In this section, the top-performing model from the previous section is fitted to the dataset and presented the findings. It is observed that, on average, males exhibit a lower platelet count compared to females ($\beta_{male} = -0.11$), a trend consistent with the insights gained from data exploration in Chapter 4. Given the presence of multiple categorical variables, and interactions in the model, there are lots of parameters estimated.

The detailed outcomes of the model are provided in Table A.1. We are particularly interested in investigating the seasonal variations in platelet count. To explore this aspect along with the effect of branch zones across different years, plots depicting average profiles are analysed. These visualisations are illustrated in Figure 5.3, showcasing the average profiles among female donors. It's worth noting that the estimated trajectories for males would be similar but shifted downward. Hence, only the average trajectories for females are presented.

Observing Figure 5.3, it becomes evident that the fitted seasonal component predominantly exhibits peaks closer to the middle of the year, particularly leaning towards the winter months rather than the summer months. However, one notable exception stands out. In 2017, profiles associated with branch zones Egoli, Vaal, and Northern demonstrate a dip in the middle of the year. Additionally, differences in seasonal effects among the respective branch zones are apparent. An intriguing observation emerges as two distinct groups of branch zones become evident: (1) Egoli, Northern, and Vaal, and (2) Free State/North Cape, KwaZulu-Natal, Mpumalanga, and Eastern Cape.

5.3 Model Analysis

In this section, a residual analysis is conducted to assess the adequacy of the fitted model from the previous sections. The primary objective was to ensure that the model assumptions are met, particularly focusing on the constant variance of residuals centered around zero.

To achieve this, a series of diagnostic plots is employed. First, individual-specific boxplots of residuals is presented, checking for consistent variance across all individuals - Figure 5.4 (top left). It should be noted that all donors could not be plotted and a subset is shown. It is evident from the plot that the errors of individual donors have similar variances and are centered close to 0 – the inclusion of the random intercept dealt with the issue seen in Figure 5.2. Subsequently, boxplots of residuals are generated with respect to temporal variables, namely month and year, as well as key covariates such as branch zone and sex. The aim is to identify any discernible patterns or heteroscedasticity that may compromise the model's performance. Considering the plot relating to boxplots by month (top middle), it can be seen that there is no significant trend element – the inclusion of sine and cosine functions has removed any trend in the residuals.

Considering boxplots relating to year, sex, and branch zone, it is noted that the

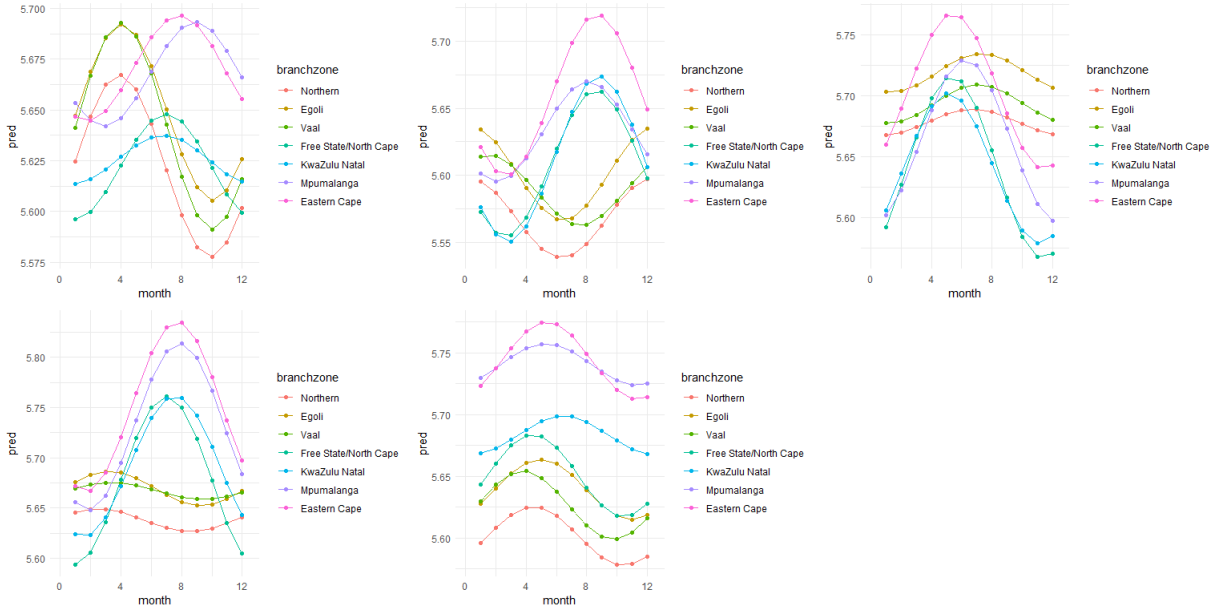

Figure 5.3: Average fitted profiles from the fixed effects of the linear mixed effects model, showing seasonal variations and level differences in platelet count among female donors from 2016 to 2020. The grid format presents each year separately, with 2016 in the top left, 2017 in the top middle, 2018 in the top right, 2019 in the bottom left, and 2020 in the bottom middle.

distribution of the errors is centered at 0. It can also be seen that heteroskedasticity is not present. The data appears consistent across the different groupings which suggests homoskedasticity. This also holds true for other plots shown (month and year).

Concurrently, a plot of standardised residuals against fitted values is constructed to detect any systematic deviations or patterns, with a well-fitted model expected to exhibit a random scatter. This can be seen in Figure 5.5. It is evident that no significant pattern occurs. Based on the visual inspections of the diagnostic plots, it is concluded that the model is adequate.

5.4 Conclusion

In this chapter, platelet count profiles were modelled using linear mixed-effect models. The nonlinear seasonal component was approached linearly by assuming yearly seasonality and setting the period of sine and cosine curves to one year, a methodology supported by insights from Chapter 4. Notable findings derived from the fitted models included: (1) a higher average platelet count in females compared to males, (2) the presence of seasonality in the data, with seasonal highs typically occurring in the middle of the year (winter months), albeit with the exception of 2017, where a seasonal low in the winter months was observed, primarily in Egoli, Vaal, and Northern branch zones, (3) the identification of two distinct groupings of branch zones, and (4) the observation of the widest spread of platelet count levels in 2020.

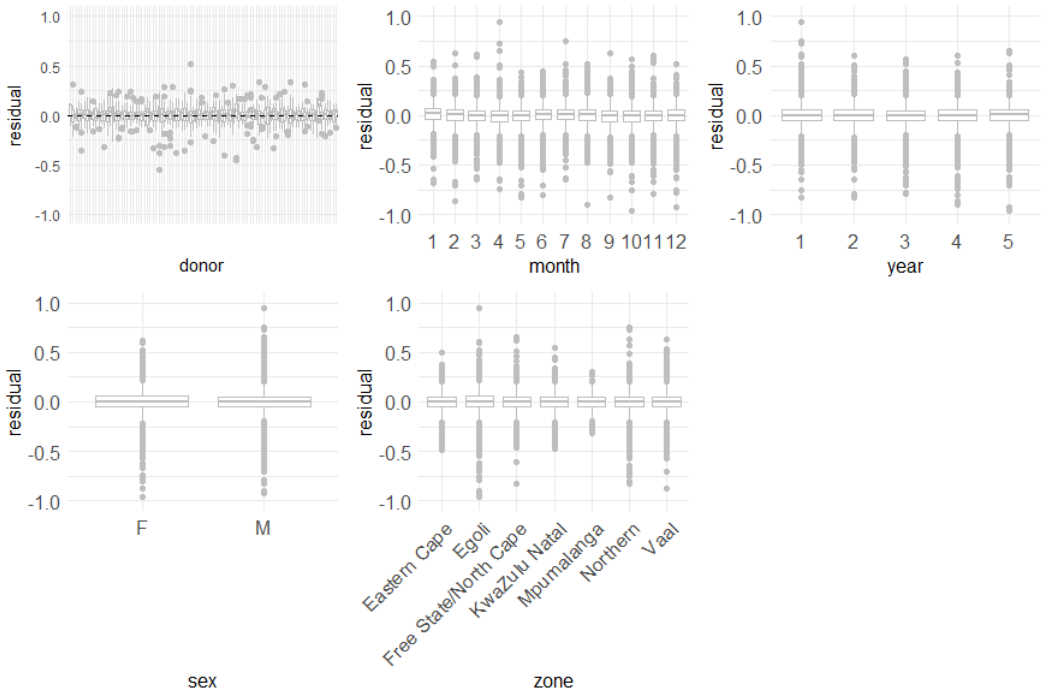

Figure 5.4: Residual analysis: Final Model. The plots assess variance consistency across individuals and check for patterns or heteroscedasticity with respect to month, year, branch zone, and sex. The inclusion of random intercepts and seasonal components addresses earlier model issues, ensuring residuals are centered and free from trends.

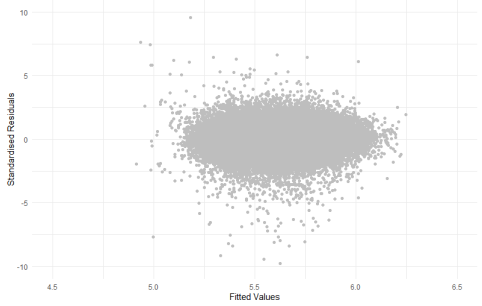

Figure 5.5: Standardised residuals vs. Fitted values: Final Model.

Chapter 6

Trajectory Clusters

6.1 Introduction

In this chapter, the focus is on finding groups of donors whose platelet trajectories are similar – to find homogeneous groups in the sample. To do this, we first adopt latent class mixed-effect models to uncover groupings of profiles in the data. Latent class mixed models assume that the population consists of several groups with distinct trajectories. Unique patterns are estimated for each latent class and individuals are probabilistically assigned to a class based on the likelihood that their data follows the trajectory of a particular class. Therefore, the result is clusters of individuals with similar longitudinal patterns. Next, we use functional principal component analysis to uncover dominant modes of variation in platelet count profiles. In Chapter 3, the PACE algorithm as a method to handle sparse functional data is introduced. Profiles are clustered, using K-means, based on functional principal curve scores in an effort to see if certain groups of profiles move differently through time. This analysis is also performed on the data after splitting the profiles into separate yearly profiles to capture dominant modes of variation within years – seasonality.

6.2 Latent Class Linear Mixed Modelling

To allow for fitting of models, the data used in this section is reduced to only contain donors with more than (1) 48 months of data (this reduces the number of donors to 139) and (2) 30 months of data (this reduces the number of donors to 657).

In this study, the dataset is reduced to address computational constraints. This approach introduces several potential biases. Firstly, the remaining sample may exhibit systematic differences from the excluded donors, leading to selection bias. For instance, donors with more extended records might have distinct health profiles or engagement levels compared to those with fewer records. Secondly, the reduced sample may not accurately represent the broader donor population, affecting the generalisability of the findings. Additionally, the sensitivity of the latent class mixed models to the sample size could lead to variations in cluster trajectories and model

performance. The two cluster trajectories are compared to see the sensitivity of the method to changing the sample. A spaghetti plot overlayed with the average platelet profile in the sample is shown in Figure 6.1.

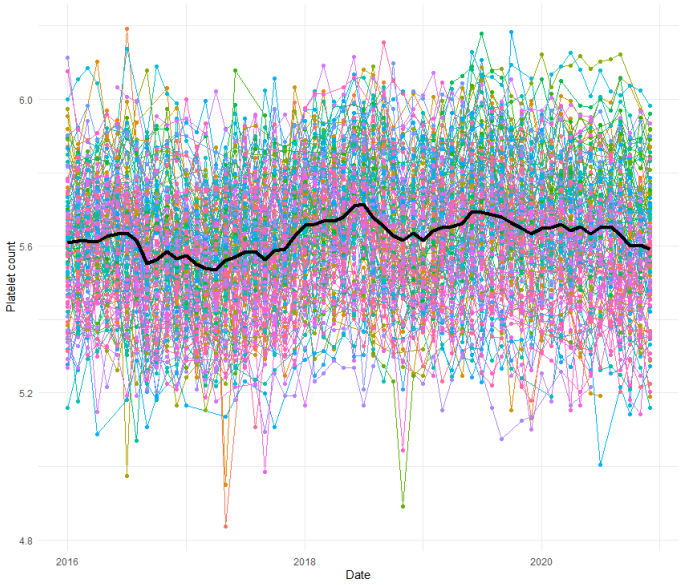

Figure 6.1: Spaghetti plot of reduced data with mean profile overlaid (black).

In this chapter we are not focused on interpretation of the mean curves and associated parameters, but rather on uncovering if there are distinct latent trajectories governing platelet count through time. Therefore, I decided to adopt a more flexible structure, as compared to the structure in the previous section. It is decided to use a B-spline with equally spaced half-yearly knots to measure temporal dynamics. It is included as a class-specific and fixed effect.

To determine the optimal number of clusters, models are fitted for 2-5 clusters, and their performance is evaluated using AIC, BIC, weighted root mean square error (WRMSE), and weighted mean absolute error (WMAE), as illustrated in Figure 6.2. AIC and BIC serve as measures for model selection, balancing goodness of fit and model complexity, where lower values indicate better fitting models adjusted for complexity. Additionally, lower values of WMAE and WRMSE are preferred for assessing predictive accuracy. Upon reviewing the metrics, it is decided that the model with 2 clusters is the best choice due to its lower AIC and BIC values compared to other models.

In Figure 6.3 cluster trajectories for each of the datasets across the 60 months are shown. Cluster trajectories for the smaller dataset (containing profiles with at least 48 observations) can be seen on the left, and on the right are the cluster trajectories for the larger dataset (containing profiles with at least 30 observations). It can be seen that there is no visible difference between the graphs in the two figures.

Furthermore, in both datasets the same split between the two clusters is seen, with one cluster containing approximately 40% of the data and the other approximately 60% of the data. Considering these in isolation lead me to conclude that the esti-

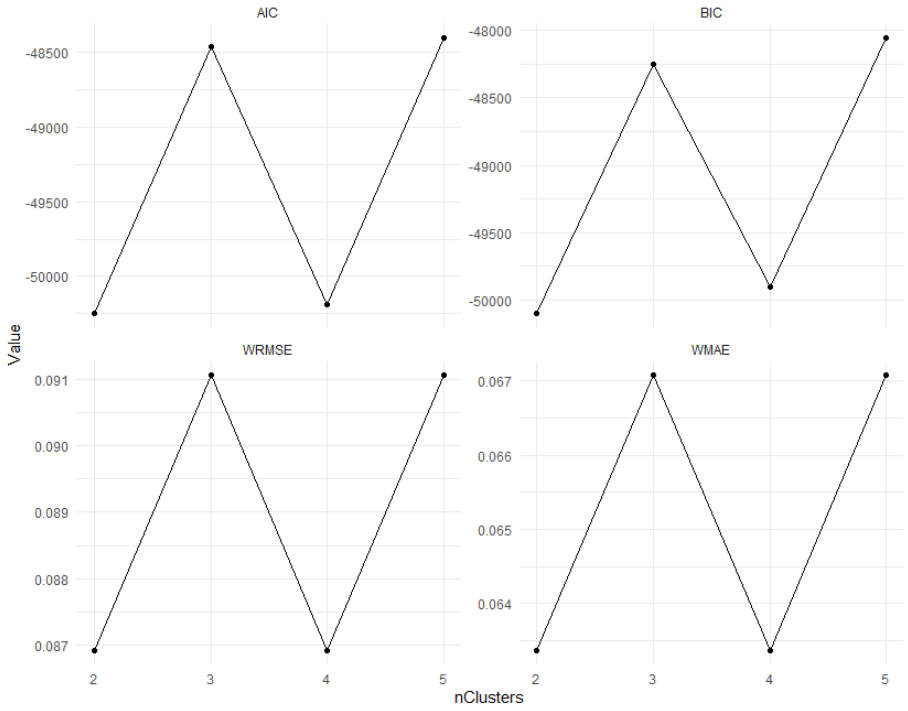

Figure 6.2: Cluster performance metrics for 2-5 clusters.

mated curves show no real sensitivity to the change in the number of observations. The estimated trajectories are very different when comparing the two clusters. Cluster A exhibits more pronounced cyclical variation relative to Cluster B. The most notable difference between the two cluster trajectories occurs in the latter half of the period. Cluster B exhibits a downward trend and lower level relative to Cluster A. Parallels can be drawn between the trajectories in Figure 5.3. In Chapter 5, it was seen that two distinct groups of branch zones become evident: (1) Egoli, Northern, and Vaal, and (2) Free State/North Cape, KwaZulu-Natal, Mpumalanga, and Eastern Cape. Interestingly, we see that Cluster B mostly tracks the trajectory of group (1) and Cluster A mostly tracks the trajectory of group (2).

To uncover if there are any patterns that arise in the clustering, the clusters grouped by the covariate data are tabulated. This is done for both datasets. Table 6.1 and Table 6.2 show the classifications relating to gender of donor. Interestingly, a proportional split between Clusters A and B is seen. Bearing in mind that Cluster A has 40% of all observations, and Cluster B has 60% of all observations. Cluster A contains approximately 40% of males and 40% of females, and cluster B contains the remaining 60%.

Table 6.1: Cluster classification by gender for smaller dataset (at least 48 observations in each profile).

	Female	Male
Cluster A	28 (39%)	28 (41%)
Cluster B	43 (61%)	40 (59%)

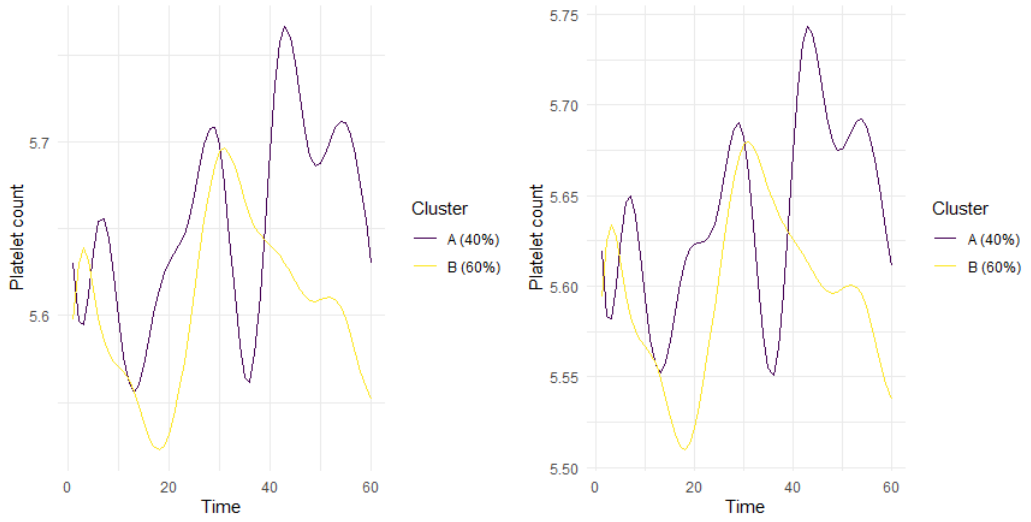

Figure 6.3: Cluster trajectories for data with more than 48 observations (left) and cluster trajectories for data with more than 30 observations (right).

Table 6.2: Cluster classification by gender for larger dataset (at least 30 observations in each profile).

	Female	Male
Cluster A	136 (40%)	120 (38%)
Cluster B	201 (60%)	200 (62%)

The same tabulation for branch zone is performed and shown in Tables 6.3 and 6.4. Fisher's exact test to test association between cluster and branch zone is performed. The test yields an insignificant result for the smaller dataset (p-value = .82) and a more significant result for the larger dataset (p-value = .04). This indicates that there is a significant association between cluster and branch zone in the larger dataset. However, considering the tables jointly, no clear or consistent pattern emerges. The tables provide contradicting information. For example, in Table 6.3, it noted Mpumalanga mostly represented in Cluster B, however, in Table 6.4 this is reversed. Similar observations can be made for Free State (FS), KwaZulu-Natal (KZN), and Vaal. It is worth noting as well that the biggest difference is seen in branch zones with the smallest number of donors i.e. Eastern Cape and Mpumalanga. Clusters do not represent branch zones nor do they represent gender.

Table 6.3: Cluster classification by branch zone for smaller dataset (at least 48 observations in each profile).

	Egoli	FS	KZN	Mpu	North	Vaal
Cluster A	20 (40%)	3 (23%)	3 (50%)	1 (33%)	21 (42%)	8 (47%)
Cluster B	30 (60%)	10 (77%)	3 (50%)	2 (67%)	29 (58%)	9 (53%)

Table 6.4: Cluster classification by branch zone for larger dataset (at least 30 observations in each profile)

	EC	Egoli	FS	KZN	Mpu	North	Vaal
Cluster A	3 (100%)	89 (37%)	28 (44%)	2 (13%)	11 (55%)	94 (41%)	29 (33%)
Cluster B	0 (0%)	150 (63%)	36 (56%)	13 (87%)	9 (45%)	134 (59%)	59 (67%)

This section explored reduced versions of the data due to the computational limitations of the method used. A more flexible method to cluster platelet profiles is required so that the analysis can include all profiles present in the data.

6.3 FPCA

In this section I applied FPCA to the apheresis donor dataset. Figure 6.4 shows the estimated covariance and correlation of platelet count. Examining the correlation matrix reveals strong correlations between platelet count readings within individuals, approaching 1 for adjacent readings and never dropping below 0.80. The covariance matrix shows high covariance among curves in the latter half of the period relative to the former half of the period. This correlates with findings in Chapter 4 and Chapter 5.

Figure 6.5 shows the first four principal curves. The first three principal curves account for 98.7% of the total variance in the data. The first curve explains 91.8% of the variance. The second curve explains 4.7% of the variance, and the third curve explains 2.2%. Figure 6.6 contains profiles of donors with extreme[1] FPC scores. These are used to aid the interpretation of the principal curves. The second principal curve is positive for the first 3 years and negative thereafter. This component is explaining the change in platelet count before year 3 in comparison to after year 3. Considering the second row of Figure 6.6, we can see that profiles with extreme FPC scores on the lower end trend upwards after 2019, while extreme values on the upper end trend downwards. The first principal component is strictly positive. It signifies the proportion of variance in platelet count curves attributed to a weighted average of platelet count patterns, where greater emphasis is placed on the latter half of the period (given the higher level). Parallels can be drawn between the shape of the first PC and the fitted covariance matrix – higher covariance values in the latter half of the period are seen. Looking at Figure 6.6, it is noted that the first principal curve seems to be separating platelet curves by their innate difference in level. Combining the interpretation of curves 1 and 2 shows that there is a lot of variance in the latter half of the period (post 2019), and specifically, some platelet curves trend upwards while other curves trend down causing an unusually high level of variance during that period.

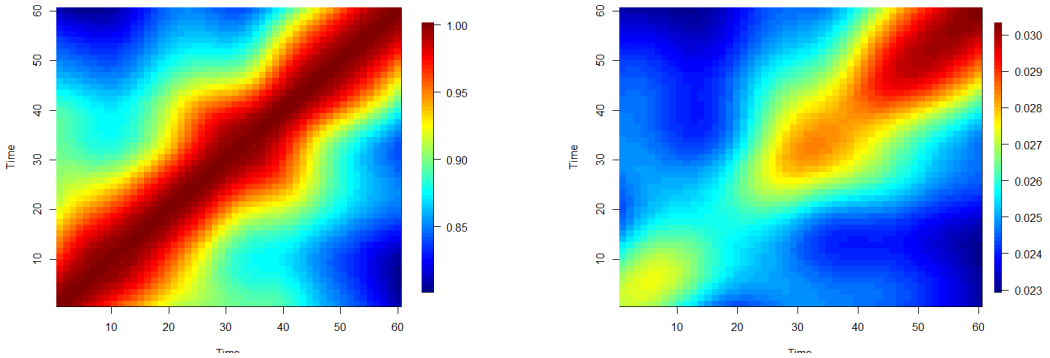

Figure 6.4: FPCA: fitted correlation (left) and covariance (right).

The third FPC demonstrates a positive trend in 2016, 2017 and 2020, while being negative in intervening years. It captures the variance associated with deviations

[1]Considering 2% and 98% percentiles.

between the middle of the period and other time points.

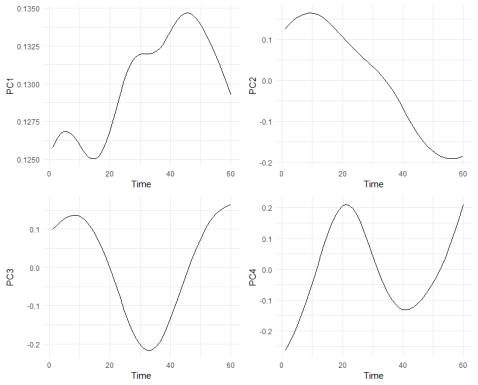

Figure 6.5: Top 4 principal curves.

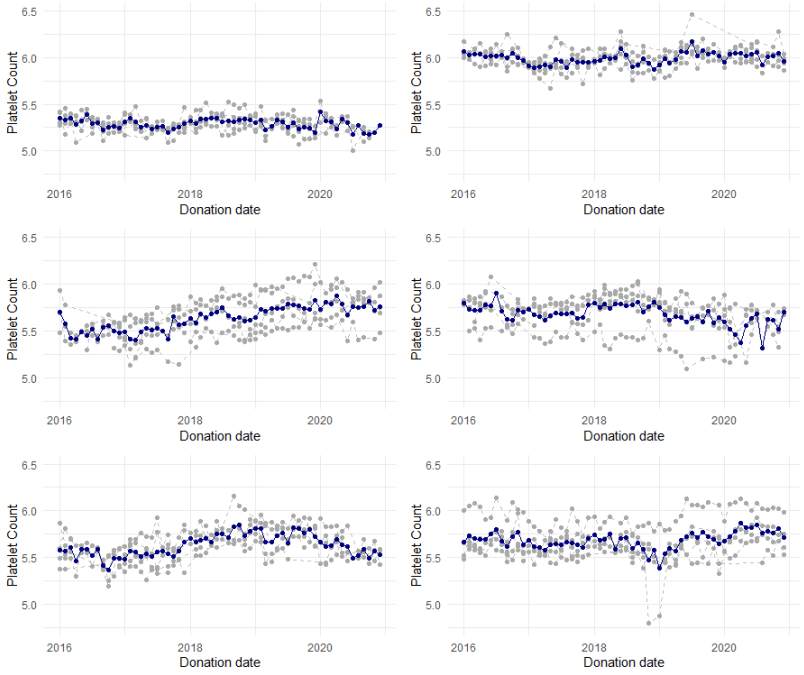

Figure 6.6: Platelet Count curves when FPC curves are extreme. The left side displays the lower 2nd percentile, representing individuals with the lowest observed FPC values, while the right side shows the upper 98th percentile, capturing the highest FPC values. For PCs 1,2 and 3.

The investigation was extended into the loadings of platelet count curves onto the principal curves, as depicted in Figure 6.7. Within the scatter plots presented in Figure 6.7, an examination of the covariate data is conducted, in an attempt to uncover patterns. The top three scatter plots mirror their counterparts in the bottom row; however, there is a distinction in the colouring applied. Specifically, the top three are differentiated by sex, uncovering potential gender-related variations, while the bottom set is differentiated by branch zone, exploring influence of this covariate on the observed platelet count dynamics. Furthermore, Gaussian ellipses representing data distribution, capturing 90% coverage, are overlaid for ease of interpretation.

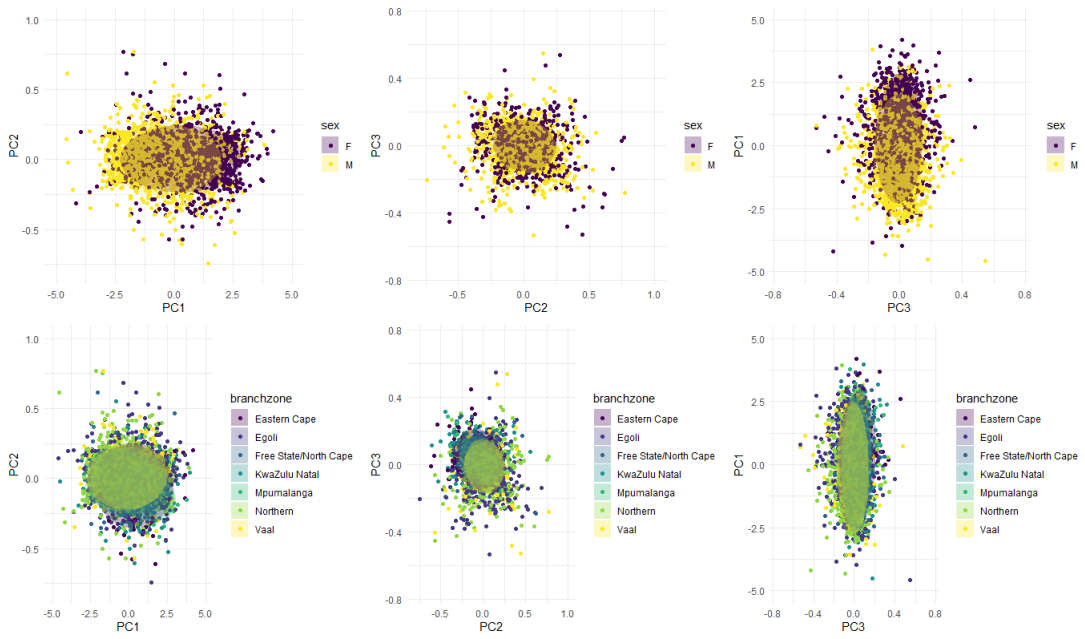

Figure 6.7: Scatter plots of loadings of platelet count curves onto principal curves coloured by covariate data. Gaussian ellipses representing data distribution, capturing 90% coverage are overlayed.

In the context of the top three plots differentiated by sex, a significant pattern emerges: females exhibit higher loadings onto the first principal curve. This observation correlates with prior findings. This is seen through females on average having higher loadings onto the first principal curve. Principal curves 2 and 3, however, capture variations in platelet count levels within specific periods relative to the remaining time frame. Interestingly, no distinct differentiation by sex is apparent in the loadings onto these latter principal curves – the ellipses are mostly overlapping. This concurs with previous findings, indicating that, although females exhibit an overall higher platelet count, the trajectory of average male profiles aligns with that of females across these specific periods.

Shifting attention to the bottom 3 plots differentiated by branch zone, there is no significant pattern or observation to be made. The most significant observation made is that there is some difference between the branch zones in connection with loadings onto the second principal curve. For example, Vaal and Northern tend to load more highly onto the second principal curve relative to Eastern Cape. This observation is consistent with earlier findings, as illustrated in Figure 4.3 and 5.3 – Vaal and Northern regions exhibit a downward trend relative to Eastern Cape.

6.3.1 Clustering Based on Principal Curves

In the previous section the behaviour of the platelet count curves based on their loadings onto the principal curves was explored visually and in an exploratory way. In this section we aimed to cluster profiles based on how they load onto the principal curves to find curves that vary similarly from the mean curve.

In the context of this research K-means clustering is used to cluster profiles based on their loadings onto principal curves. The loadings onto the first principal curve are larger than loadings onto principal curves 2 and 3. An initial clustering using 3 clusters is shown in Figure 6.8. Due to the dominance of loadings onto the first principal curve, it can be seen that only a level difference is detected in the clustering. The average profile trajectories in the respective clusters are almost identical. The only significant difference noted is a difference in the level.

Table 6.5: Contingency Table - Gender vs. Cluster. Clustering based on raw loadings.

	Female	Male	Total
Cluster 1	527 (19%)	1223 (42%)	1750 (30%)
Cluster 2	1279 (46%)	1238 (42%)	2517 (44%)
Cluster 3	972 (35%)	475 (16%)	1447 (26%)

Examining Table 6.5, a predominant trend emerges, with the majority of both males and females being assigned to Cluster 2. Specifically, 1 279 females, constituting 46% of the total female population, and 1 238 males, representing 42% of the total male population, fall into this cluster. In contrast, Cluster 3 exhibits a higher proportion of females compared to males, while Cluster 1 demonstrates the converse, having more males than females. This alignment is consistent with the earlier observation that females, on average, display higher platelet count levels relative to males. Fisher's exact test is performed indicating a significant association between cluster and gender (p-value $< .01$).

When looking at branch zones, the clustering approach does not yield as notable results. Table 6.6 attests to this observation.

Table 6.6: Contingency Table - Branch zone vs. Cluster. Clustering based on raw loadings.

	EC	Ego	FS/NC	KZN	Mpu	North	Vaal
1	48 (19%)	541 (29%)	128 (31%)	215 (29%)	25 (23%)	558 (36%)	235 (32%)
2	117 (45%)	859 (45%)	184 (44%)	331 (45%)	46 (42%)	671 (43%)	309 (42%)
3	93 (36%)	494 (26%)	103 (25%)	186 (25%)	39 (35%)	340 (21%)	192 (26%)

An observation is that a pattern emerges indicating that the Northern region is more inclined to be associated with Cluster 1 over Cluster 3. Conversely, Eastern Cape and Mpumalanga show a higher likelihood of affiliation with Cluster 3 relative to Cluster 1 [2].

[2]Recall that we expect 30% in cluster 1, 44% in cluster 2 and 26% in cluster 3.

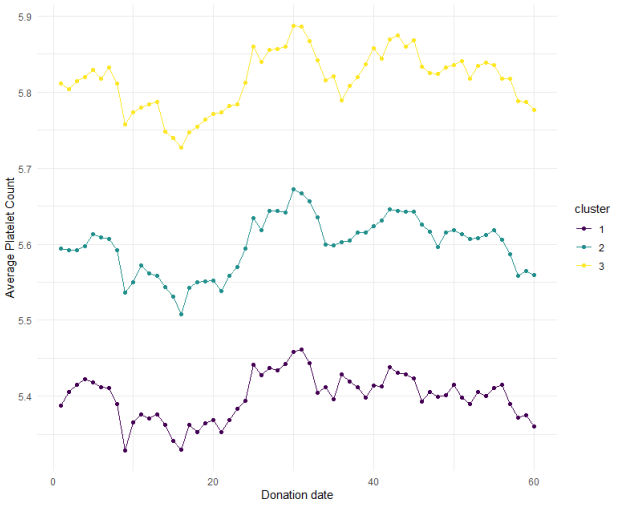

Figure 6.8: (left) Average platelet profiles for clustering based on loadings onto first 3 principal curves, and (right) average platelet profiles: Clustering based on raw loadings.

We aimed to understand both the average level and the differences in trajectories. Principal curves 2 and 3 exhibit the capability to capture distinctions in curves, as discussed earlier. However, to mitigate the influence of loadings onto the dominant first principal curve, clustering using two methods was attempted: (1) normalisation of loading values to ensure a comparable scale, and (2) clustering based solely on loadings onto principal curves 2 and 3. The resulting average profiles, as illustrated in Figure 6.9, demonstrates the efficacy of this clustering approach. A comparative analysis between Figure 6.8 and Figure 6.9 reveals that the clustering method associated with Figure 6.9 successfully identifies groups of profiles with divergent trajectories, particularly evident in the latter half of the observed period.

Examining the left plot in Figure 6.9, the average platelet count profiles for each cluster based on all three normalised component loadings can be observed. Cluster 1 is associated with a higher level and maintains this high trajectory without a decline in the latter half of the period. Cluster 2, on the other hand, largely mirrors the same trajectory; however, during the latter half of the period, the average platelet count exhibits a descending trend. Meanwhile, Cluster 3 is linked with a lower level of average platelet count, characterised by a *relatively* stable trajectory throughout the observed period.

The plot on the right tells a slightly different story - clustering based only on loadings onto second and third principal curves. Cluster 1 is associated with a relatively low average platelet count at the beginning of the period and shows an increasing trend all the way to the end of the period; Cluster 2 is associated with a relatively high average platelet count at the beginning of the period and displays a downward trend towards the end of the period; and Cluster 3 is associated with a relatively moderate level of platelet count with a relatively more stable trajectory. More specifically, the 3 clusters have the same trajectories up until a point (2019) and thereafter differ in their trajectories. Cluster 1 is associated with increasing average platelet count;

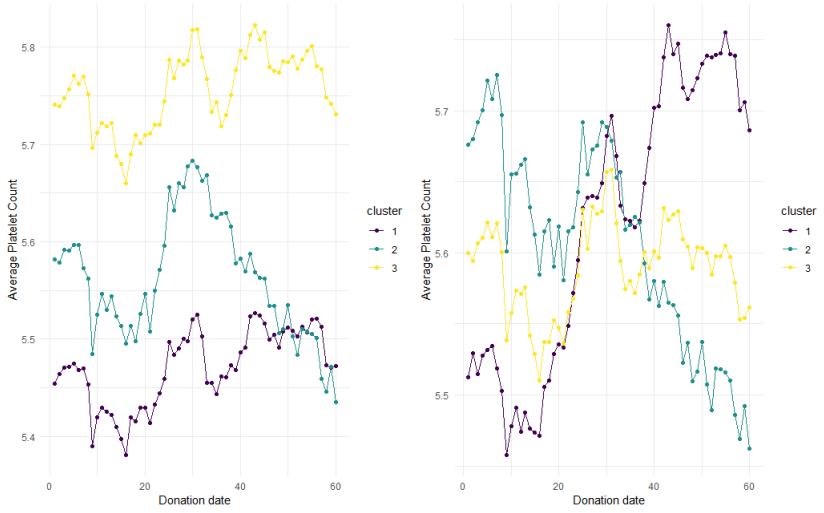

Figure 6.9: Average platelet count profiles: clustering based on normalised loadings (left), and clustering based on loadings onto principal curves 2 and 3 (right).

Cluster 2 with decreasing average platelet count; and Cluster 3 with a relatively stable average platelet count.

Figure 6.9 tells a different story than that of Figure 6.8. In Figure 6.8 the average profiles follow the same trajectories. However, in Figure 6.9 the difference in trajectories of platelet count profiles is highlighted, specifically in the latter half of the period – there were distinct clusters that move differently towards the end of the period.

Table 6.7 and Table 6.8 show the sexes and branch zones associated with the clustering based on standardised values. It is noted that 49% of all donors allocated to Cluster 1, 12% of donors allocated to Cluster 2, and 39% of donors allocated to Cluster 3. The same pattern emerged as in Table 6.5 i.e males are more associated with Cluster 1 (lower platelet count) relative to Cluster 3 (higher platelet count) and vice versa for females.

Table 6.7: Contingency Table - Gender vs. Cluster. Clustering based on standardised loadings.

	Female	Male	Total
1	1002 (36%)	1777 (61%)	2779 (49%)
2	380 (14%)	294 (10%)	674 (12%)
3	1396 (50%)	865 (29%)	2261 (39%)

Considering Table 6.8, I noticed that Eastern Cape, Mpumalanga, Free state, and KwaZulu-Natal are more inclined to be associated with Cluster 2. Additionally, donors in Eastern Cape and Mpumalanga are less inclined to be associated with Cluster 1.

Table 6.9 and Table 6.10 shows the association between clusters based on only loadings onto the second and third principal curves and the covariate data available.

Table 6.8: Contingency Table - Branch zone vs. Cluster. Clustering based on standardised loadings.

	EC	Egoli	FS	KZN	Mpu	North	Vaal
1	79 (31%)	932 (49%)	186 (45%)	328 (45%)	41 (37%)	852 (54%)	361 (49%)
2	86 (33%)	138 (7%)	86 (21%)	153 (21%)	35 (32%)	114 (7%)	62 (8%)
3	93 (36%)	824 (44%)	143 (34%)	251 (34%)	34 (31%)	603 (39%)	313 (43%)

It can be seen that 29% of all observations are allocated to Cluster 1, 11% to Cluster 2, and 60% to Cluster 3.

Table 6.9: Contingency Table - Gender vs. Cluster. Clustering based loadings onto second and third principal curves.

	Female	Male	Total
1	744 (27%)	911 (31%)	1655 (29%)
2	270 (10%)	386 (13%)	656 (11%)
3	1764 (63%)	1639 (56%)	3403 (60%)

Table 6.10: Contingency Table - Branch zone vs. Cluster. Clustering based loadings onto second and third principal curves.

	EC	Egoli	FS	KZN	Mpu	North	Vaal
1	109 (42%)	478 (25%)	139 (34%)	270 (37%)	46 (42%)	427 (27%)	186 (25%)
2	8 (3%)	233 (12%)	30 (7%)	26 (4%)	3 (3%)	257 (16%)	99 (14%)
3	141 (54%)	1183 (63%)	246 (59%)	436 (59%)	61 (55%)	885 (57%)	451 (61%)

With reference to Table 6.10, It is noticed that Eastern Cape, Mpumalanga, KwaZulu-Natal, and Free State are all less inclined to be associated with Cluster 2 and relatively more associated with Cluster 1. The Northern region has a relatively larger percentage of donors allocated to Cluster 2.

6.4 Splitting each profile into separate yearly profiles

In the previous section FPCA was used to find dominant patterns of variation across the 60 time points. Naturally, given the trajectories through time (inter-year variability), the seasonal part of the variance is overpowered. In this section we separate each profile into 5 yearly profiles to capture variance within years and uncover seasonal patterns. Figure 6.10 shows the correlation and covariance plots. The covariance matrix shows high covariance among curves in the middle of the period (winter months) relative to the beginning and the end.

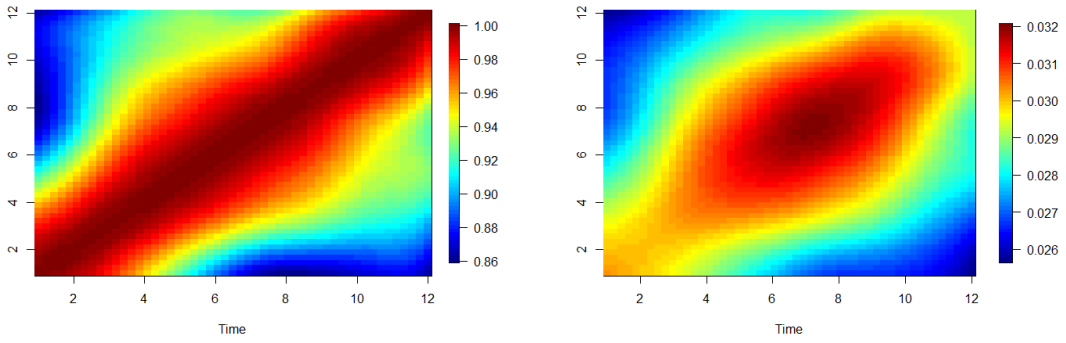

Figure 6.10: FPCA: fitted correlation (left) and covariance (right).

Figure 6.11 shows the first 3 principal curves. These 3 principal curves account for 99.8% of the total variance in the data. The first curve explains 95.6% of the variance. The second curve explains 2.9% of the variance, and the third curve explains 1.3%.

Figure 6.12 contains profiles of donors with extreme[3] FPC scores. These are used to aid the interpretation of the principal curves. The second principal curve is positive for the first 6 months and negative thereafter. This component is explaining the change in platelet count before month 6 in comparison to after month 6. Considering the second row of Figure 6.12, it is noted that profiles with extreme FPC scores on the lower end trend upwards after 2019, while extreme values on the upper end trend downwards.

It was noted that the first principal component is strictly positive. It signifies the proportion of variance in platelet count curves attributed to a weighted average of platelet count patterns, where greater emphasis is placed in the middle of the period. Parallels were then drawn between the shape of the first PC and the fitted covariance matrix – higher covariance values in the middle of the period is noted. Looking at Figure 6.12, the first principal curve seems to be separating platelet curves by their innate difference in level. The third principal curve is positive in the summer months and negative in the winter months. It captures the variance associated with deviations between summer and winter months. Considering the

[3]Considering 2% and 98% percentiles.

third row of Figure 6.12, the third principal curve detects the difference in seasonal pattern in the yearly profiles. Specifically, those profiles with extreme values on the lower end tend to have highs in the winter months and those with extreme values on the upper end tend to have lows in the winter months relative to the summer months.

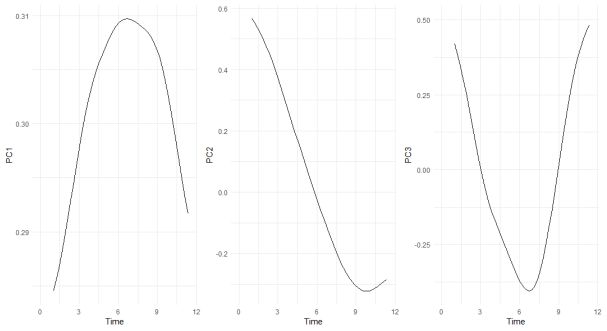

Figure 6.11: Top 3 principal curves. Separate yearly profiles.

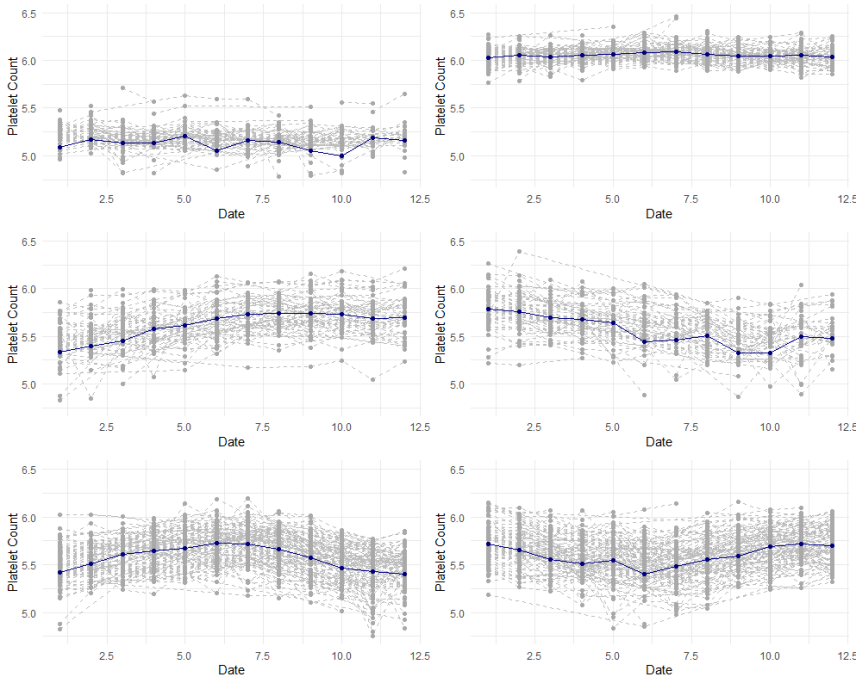

Figure 6.12: Platelet count curves when FPC curves are extreme. The left side displays the lower 2nd percentile, representing individuals with the lowest observed FPC values, while the right side shows the upper 98th percentile, capturing the highest FPC values. For PCs 1,2 and 3. Separate yearly profiles.

Combining interpretations of Figure 6.11 and Figure 6.12, it is evident that the prinicpal curves relating to the yearly profiles are detecting differences in level (first principal curve), trend (second principal curve), and seasonality (first and third principal curves) between the yearly profiles.

In Figure 6.13, scatter plots of the loadings of the yearly profiles onto the principal

curves are shown. Considering loadings onto the first principal curve, year 3 (2018) and year 4 (2019) are associated with higher loadings on average. While year 2 (2017) is more associated with lower loadings onto the first principal curve. This is consistent with what was found in the exploratory data analysis – years 3 and 4 have on average higher average platelet count levels relative to other years, with year 2 having the lowest average platelet count levels. With reference to loadings onto the second principal curve, it can be seen that year 5 (2020) is most associated with higher loadings. This indicates that in year 5 there is a dominant downward trajectory of platelet count. Conversely, year 4 and year 2 are associated with lower loadings, indicating an upward trend.

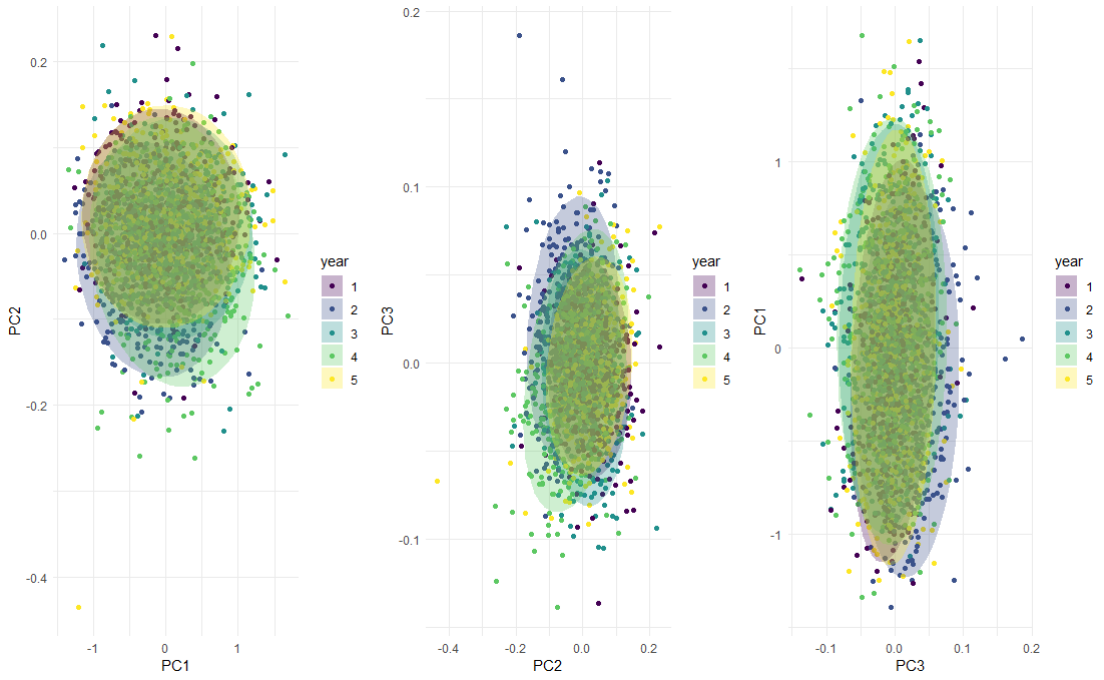

Figure 6.13: Scatter plots of loadings of platelet count curves onto principal curves coloured by year. Gaussian ellipses representing data distribution, capturing 90% coverage are overlayed.

Finally, considering the third principal curve detecting difference in seasonality, it can be seen that year 2 is associated with higher loadings relative to all the other years. This indicates that on average, a different seasonal pattern is evident in year 2.

6.5 Conclusion

This chapter aimed to identify groupings of donors that behave similarly or differently through the period. Initially, in Section 6.2 Latent Class Mixed Models were applied, revealing no significant associations between clusters and gender or branch zone. However, two distinct latent profiles were detected with the most notable differences in the latter half of the period. Different clustering methods, using results

from FPCA, were applied in Section 6.3. Clustering was performed based on donors' raw FPC scores. This method primarily highlighted level differences due to the dominance of the loadings onto the first principal curve. From this clustering method it was seen that females tended to be associated with higher platelet count levels, while regions like Eastern Cape and Mpumalanga exhibited higher levels compared to Northern regions. To mitigate the dominance of the loadings onto the first principal curve, two additional methods were employed. The first involved standardising the loadings onto principal components, while the second method entailed excluding loadings onto the first principal curve during clustering. These approaches effectively detected clusters exhibiting different trajectories. The clusters identified by these methods showcased inconsistencies in grouping patterns, with certain regions – Eastern Cape, Free State, Kwa-Zulu Natal, and Mpumalanga – which displayed decreasing trends in the latter half of the period. Conversely, the second method highlighted that donors from the same branch zones were more associated with clusters displaying increasing trends in the latter half of the period. Nevertheless, consistent latent trajectories were observed across methods, indicating heterogeneous trajectories in the latter half of the period, particularly post-2019, signifying distinct donor behaviours compared to the more consistent trajectories observed in the first half of the period.

In section 6.4 the donor profiles were separated by year and applied FPCA to detect intra-year variability. It is showed that this method is effective in detecting dominant modes of variation in the data.

Chapter 7

Discussion and Conclusion

The research questions guiding this study revolved around understanding the dynamics of platelet count and product failures within the context of platelet transfusion medicine. Firstly, the investigation sought to determine if there are different differences in platelet count levels and trajectories among various empirical groupings. Secondly, it aimed to explore the presence of seasonality in platelet count data, investigating potential patterns and fluctuations over time. Thirdly, the study aimed to identify distinct groupings of donors exhibiting similar behaviours regarding platelet count trajectories. Lastly, it sought to establish if there is a connection between average platelet trajectories and the occurrence of product failures, examining the relationship between platelet count dynamics and the reliability of platelet products.

Seasonality in platelet count is a well-documented phenomenon, as supported by various sources in the literature [23][17][5], which consistently highlight seasonal lows during summer periods. This study represents the first attempt to assess seasonality within a cohort of South African donors. Initial exploratory data analysis in Chapter 4 revealed a clear seasonal pattern in platelet count, further corroborated by modelling using linear mixed effect models. The analysis indicated that platelet count tends to exhibit seasonal highs closer to the middle of the year, aligning with winter months (autumn-winter or winter-spring). Interestingly, while most years exhibited a similar average seasonal pattern, 2017 appeared to deviate from this trend. However, based on the fitted trajectories in Figure 5.3 in Chapter 5, the difference in seasonality is only present in 3 of the 7 branch zones – Northern, Egoli, and Vaal. Furthermore, fitted trajectories track relatively closely over the period forming what seems to be one group. The remaining 4 branch zones – Free State, KwaZulu-Natal, Mpumalanga, and Eastern Cape – exhibit similar average seasonal pattern and level through time. Additionally, intra-year variability was investigated through FPCA, which not only confirmed the presence of seasonality, but also revealed differences in seasonal patterns between years. FPCA demonstrated potential as a method for identifying outlying observations and detecting curves with abnormal behaviours, highlighting its efficacy in uncovering insights into seasonal variations in platelet count data.

The examination of platelet count data revealed intriguing insights into the temporal dynamics of platelet profiles. As discussed in Chapter 4, there is evidence of seasonality in the product pass rate that aligns with the seasonal patterns observed in platelet count data. Additionally, a trend has emerged since 2018, marked by a decline in the proportion of platelet products successfully passing the volume and yield tests. However, an accompanying decrease in average platelet count levels was not observed. This observation prompted research question 3: Are there distinct groupings of donors that behave similarly? More specifically, are there specific groupings of donors that behave differently during that decline. It is important to note that the donors whose blood donations are pooled are different from the apheresis donors. Therefore, this research was limited to evaluating patterns in the respective datasets independently.

This question is addressed in Chapter 6, where the approach encompassed two primary methodologies: latent class mixed models and clustering based on loadings onto functional principal curves. Using latent class mixed models, a random intercept was introduced to accommodate variability in baseline platelet count levels among donors. Two distinct clusters within the data are identified. Interestingly, while these clusters exhibited somewhat similar trajectories pre-2019, differences are apparent post-2018. One cluster demonstrated sustained or elevated platelet count levels, while the other exhibited a distinct declining trend.

Furthermore, the investigation extended to the relationship between these identified clusters and available covariate data, specifically gender and sex. Surprisingly, the clusters appeared to show no significant associations with these demographic factors, suggesting the presence of inherent groupings within the data that were independent of gender or branch zone. This finding underscores the complexity of platelet count dynamics and hints at potentially novel factors influencing platelet profile variations. Using latent class mixed models can impose computational challenges. This is specifically true when datasets are large and/or complex. The dataset was reduced in order to fit the models in Section 6.2. Of course, this introduces its own issues, such as introducing biases into the analysis. Furthermore, even after reducing the data, fitting of these models is computationally expensive and, therefore, time-consuming.

FPCA was introduced as a method to uncover dominant modes of variation in functional data. Once these are uncovered, one is able to identify outlying curves by assessing curves that have extreme loadings associated with the respective principal curves. Furthermore, curves can be clustered based on how they load onto the principal curves. Since each observation is a linear combination of the principal curves as seen in Equation 3.7, intuitively, curves with similar scores should exhibit similar trajectories. Donor profiles are clustered based on how they load onto the principal curves. This method proved to be more efficient relative to latent class mixed models and showed efficacy in uncovering trends in the data.

Due to the size of the loadings onto the first principal curve, clustering based on the raw loadings detected clusters with the same trajectory, but different levels. It was found that females were associated with the cluster with a greater average platelet

count. This finding correlates with previous literature [33][25][5][2] as well as with the findings in Chapter 5. On average, females have higher platelet count levels compared to males. Interestingly, in Chapter 4, it was noted that while females have a higher average level, the trajectory of males' and females' average platelet count is almost identical. To uncover different trajectories, the dominance of scores related to the first principal curve needed to be dealt with. This was done in two ways: (1) standardise the loadings and, (2) remove loadings onto the first principal curve and cluster based on only loadings onto the second and third principal curves.

Interestingly, these methods of clustering uncovered cluster trajectories similar to those seen when applying latent class mixed models. Both clustering methods identified a groups of donors that have the same trajectory in platelet count in the first half of the period with different trajectories in the second. Most significantly, using latent class mixed models *and* clustering based on FPCA, a latent grouping associated with a decline in platelet count in the latter half of the period is found. This correlates with the observed decline in the proportion of products passing. This variability in the data is also highlighted in Chapter 5 where it was noted that average platelet profiles differ more significantly in the latter half of the period. It was also seen in Figure 6.4, where elevated levels of covariance in platelet count curves were apparent. It is a common theme that in the latter half of the period platelet count profiles behaved more differently relative to the first half of the period.

There is evidence in literature suggesting that genetic makeup/geographics affects platelet count [2]. A key question to be answered is if there is association between variables in the dataset and clusters associated with decreasing levels of platelet count in the latter half of the period. As discussed, in Chapter 5 two groupings of branch zones were found: (1) Egoli, Northern, and Vaal, and (2) Free State/North Cape, KwaZulu-Natal, Mpumalanga, and Eastern Cape. The first group displayed relatively stable average profiles and seasonal component through time, while the second groups seasonal component was relatively more volatile. In the second group, the seasonal relationship being reversed in 2017 is observed and, thereafter, less pronounced seasonal components and a decreasing average level through time was noted. Parallels with the trajectories of these groups and the latent profiles derived from latent class mixed models can be drawn. Two groups are derived using this method. The one group, Cluster A in Figure 6.3, appears to mirror the trajectory of the second grouping, and the other group (Cluster B) mirrors the first group. Cluster A is associated with more pronounced seasonal components and relatively higher levels in the latter half, while Cluster B is associated with a low in the middle of 2017 and decreasing trend post 2018 with a less apparent seasonal component. While this relationship between the group trajectories is apparent, when analysing contingency tables, it is found that there is no significant association between branch zone and cluster assignment. Interestingly, a similar pattern is found in Section 6.3. As discussed, the clustering methods identified groupings of platelet profiles with different behaviours in the latter half of the period. Considering the clustering based on standardised loadings, in Figure 6.9 it can be seen that Cluster 2 is associated with a decrease in average platelet count in the latter half of the period and, upon analysis of Table 6.8, it is seen that branch zones Eastern Cape, Free State, KwaZulu-Natal,

and Mpumalanga are more inclined to be associated with that cluster while Northern, Vaal, and Egoli zones are less inclined to be associated with that cluster. This contradicts discussed findings in Chapter 5, as well as findings based on clustering on only loadings onto the second and third principal curves. In Figure 6.9, cluster trajectories for this method can also be seen. Here, the opposite was seen, branch zones Eastern Cape, Free State, KwaZulu-Natal, and Mpumalanga were not associated with the cluster exhibiting a downward trend in the latter half of the period (Cluster 2) and were more inclined to being associated with the cluster exhibiting a higher level in the latter half of the period (Cluster 1). Furthermore, Northern, Vaal and Egoli regions were most associated with Cluster 2. This agreed with groupings found in Chapter 5. It is presumed that keeping the level effect in the analysis skews the intended result, therefore, there is a preference for considering the clustering based on only the loadings onto the second and third principal curves. Although, analysis points to distinct groups within *all* branch zones displaying a decreasing trend in the latter half of the period. With some branch zones exhibiting a greater propensity towards the declining trajectory.

While this downward trend appeared to be more apparent in some branch zones relative to others, there is not enough evidence to make a conclusive decision regarding association between clusters associated with a decreasing trend in the latter half of the period and branch zone in which blood was donated. It is presumed that there is some other factor not present in the data that is affecting platelet count levels and platelet product pass rates. However, it can conclusively be said that during the period in which declining platelet product pass rates were observed, there was more variability in platelet profiles and there were groupings of platelet profiles displaying a decreasing trend during that period.

Appendix A

Linear Mixed Model Summary

Covariate	Estimate	Std. Error	Pr(>|t|)
(Intercept)	5.67	0.01	0.00
branchzoneEgoli	-0.02	0.01	0.08
branchzoneFree State/North Cape	-0.05	0.02	0.00
branchzoneKwaZulu Natal	-0.04	0.01	0.00
branchzoneMpumalanga	-0.00	0.02	0.89
branchzoneNorthern	-0.05	0.01	0.00
branchzoneVaal	-0.03	0.01	0.04
year2	-0.01	0.00	0.01
year3	0.03	0.00	0.00
year4	0.08	0.00	0.00
year5	0.07	0.00	0.00
month.sin	-0.02	0.00	0.00
month.cos	-0.02	0.00	0.00
sexM	-0.11	0.00	0.00
branchzoneEgoli:year2	-0.04	0.00	0.00
branchzoneFree State/North Cape:year2	-0.00	0.01	0.71
branchzoneKwaZulu Natal:year2	-0.00	0.01	0.62
branchzoneMpumalanga:year2	-0.02	0.02	0.27
branchzoneNorthern:year2	-0.04	0.01	0.00
branchzoneVaal:year2	-0.04	0.01	0.00
branchzoneEgoli:year3	0.04	0.01	0.00
branchzoneFree State/North Cape:year3	-0.01	0.01	0.03
branchzoneKwaZulu Natal:year3	-0.02	0.01	0.00
branchzoneMpumalanga:year3	-0.04	0.01	0.00
branchzoneNorthern:year3	0.02	0.01	0.00
branchzoneVaal:year3	0.02	0.01	0.01
branchzoneEgoli:year4	-0.06	0.01	0.00
branchzoneFree State/North Cape:year4	-0.02	0.01	0.00
branchzoneKwaZulu Natal:year4	-0.01	0.01	0.02
branchzoneMpumalanga:year4	-0.02	0.01	0.12

branchzoneNorthern:year4	-0.06	0.01	0.00
branchzoneVaal:year4	-0.06	0.01	0.00
branchzoneEgoli:year5	-0.08	0.01	0.00
branchzoneFree State/North Cape:year5	-0.04	0.01	0.00
branchzoneKwaZulu Natal:year5	-0.02	0.01	0.01
branchzoneNorthern:year5	-0.09	0.01	0.00
branchzoneVaal:year5	-0.09	0.01	0.00
branchzoneEgoli:month.sin	0.06	0.00	0.00
branchzoneFree State/North Cape:month.sin	0.01	0.01	0.17
branchzoneKwaZulu Natal:month.sin	0.02	0.01	0.00
branchzoneMpumalanga:month.sin	-0.00	0.01	0.65
branchzoneNorthern:month.sin	0.06	0.00	0.00
branchzoneVaal:month.sin	0.06	0.01	0.00
year2:month.sin	-0.04	0.01	0.00
year3:month.sin	0.04	0.01	0.00
year4:month.sin	-0.04	0.01	0.00
year5:month.sin	0.03	0.01	0.00
branchzoneEgoli:month.cos	-0.01	0.00	0.10
branchzoneFree State/North Cape:month.cos	-0.01	0.01	0.23
branchzoneKwaZulu Natal:month.cos	0.00	0.01	0.45
branchzoneMpumalanga:month.cos	0.01	0.01	0.18
branchzoneNorthern:month.cos	-0.01	0.00	0.26
branchzoneVaal:month.cos	-0.01	0.01	0.06
year2:month.cos	0.00	0.01	0.42
year3:month.cos	-0.05	0.01	0.00
year4:month.cos	-0.04	0.01	0.00
year5:month.cos	-0.01	0.01	0.02
branchzoneEgoli:year2:month.sin	0.01	0.01	0.18
branchzoneFree State/North Cape:year2:month.sin	-0.00	0.01	0.76
branchzoneKwaZulu Natal:year2:month.sin	-0.02	0.01	0.02
branchzoneMpumalanga:year2:month.sin	0.03	0.03	0.31
branchzoneNorthern:year2:month.sin	0.00	0.01	0.65
branchzoneVaal:year2:month.sin	0.01	0.01	0.11
branchzoneEgoli:year3:month.sin	-0.09	0.01	0.00
branchzoneFree State/North Cape:year3:month.sin	-0.00	0.01	0.77
branchzoneKwaZulu Natal:year3:month.sin	-0.01	0.01	0.31
branchzoneMpumalanga:year3:month.sin	-0.02	0.02	0.13
branchzoneNorthern:year3:month.sin	-0.08	0.01	0.00
branchzoneVaal:year3:month.sin	-0.09	0.01	0.00
branchzoneEgoli:year4:month.sin	0.02	0.01	0.00
branchzoneFree State/North Cape:year4:month.sin	0.02	0.01	0.08
branchzoneKwaZulu Natal:year4:month.sin	-0.00	0.01	0.82
branchzoneMpumalanga:year4:month.sin	0.00	0.01	0.92
branchzoneNorthern:year4:month.sin	0.02	0.01	0.03
branchzoneVaal:year4:month.sin	0.01	0.01	0.29
branchzoneEgoli:year5:month.sin	-0.06	0.01	0.00

branchzoneFree State/North Cape:year5:month.sin	0.01	0.01	0.55
branchzoneKwaZulu Natal:year5:month.sin	-0.03	0.01	0.00
branchzoneNorthern:year5:month.sin	-0.05	0.01	0.00
branchzoneVaal:year5:month.sin	-0.05	0.01	0.00
branchzoneEgoli:year2:month.cos	0.05	0.01	0.00
branchzoneFree State/North Cape:year2:month.cos	0.01	0.01	0.43
branchzoneKwaZulu Natal:year2:month.cos	0.00	0.01	0.93
branchzoneMpumalanga:year2:month.cos	-0.02	0.02	0.30
branchzoneNorthern:year2:month.cos	0.04	0.01	0.00
branchzoneVaal:year2:month.cos	0.04	0.01	0.00
branchzoneEgoli:year3:month.cos	0.06	0.01	0.00
branchzoneFree State/North Cape:year3:month.cos	-0.00	0.01	0.79
branchzoneKwaZulu Natal:year3:month.cos	0.00	0.01	0.86
branchzoneMpumalanga:year3:month.cos	-0.02	0.02	0.24
branchzoneNorthern:year3:month.cos	0.06	0.01	0.00
branchzoneVaal:year3:month.cos	0.06	0.01	0.00
branchzoneEgoli:year4:month.cos	0.06	0.01	0.00
branchzoneFree State/North Cape:year4:month.cos	-0.01	0.01	0.19
branchzoneKwaZulu Natal:year4:month.cos	0.00	0.01	0.89
branchzoneMpumalanga:year4:month.cos	-0.01	0.01	0.61
branchzoneNorthern:year4:month.cos	0.06	0.01	0.00
branchzoneVaal:year4:month.cos	0.06	0.01	0.00
branchzoneEgoli:year5:month.cos	0.02	0.01	0.02
branchzoneFree State/North Cape:year5:month.cos	0.01	0.01	0.11
branchzoneKwaZulu Natal:year5:month.cos	0.01	0.01	0.20
branchzoneNorthern:year5:month.cos	0.02	0.01	0.01
branchzoneVaal:year5:month.cos	0.03	0.01	0.00

Table A.1: Linear mixed model summary

Bibliography

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Ginevra Biino, Carlo L Balduini, Laura Casula, Piergiorgio Cavallo, Simona Vaccargiu, Debora Parracciani, Donatella Serra, Laura Portas, Federico Murgia, and Mario Pirastu. Analysis of 12,517 inhabitants of a sardinian geographic isolate reveals that predispositions to thrombocytopenia and thrombocytosis are inherited traits. *Haematologica*, 96(1):96, 2011.

[3] Andrea Boccatonda, Damiano D'Ardes, Ilaria Rossi, Alice Grignaschi, Antonella Lanotte, Francesco Cipollone, Maria Teresa Guagnano, and Fabrizio Giostra. Platelet count in patients with sars-cov-2 infection: a prognostic factor in covid-19. *Journal of Clinical Medicine*, 11(14):4112, 2022.

[4] Alex J Bowers and Ryan Sprott. Examining the multiple trajectories associated with dropping out of high school: A growth mixture model analysis. *The Journal of educational research*, 105(3):176–195, 2012.

[5] Michael F Buckley, John W James, Dianne E Brown, Gordon S Whyte, Mark G Dean, Colin N Chesterman, and Jennifer A Donald. A novel approach to the assessment of variations in the human platelet count. *Thrombosis and haemostasis*, 83(03):480–484, 2000.

[6] Chen Chen. Trajectories and predictors of child abuse in chinese children aged 4–7 years: A growth mixture model analysis. *Children and youth services review*, 141:106628, 2022.

[7] Jing Cheng, Lloyd J Edwards, Mildred M Maldonado-Molina, Kelli A Komro, and Keith E Muller. Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in medicine*, 29(4):504–520, 2010.

[8] Martina E Daly. Determinants of platelet count in humans. *Haematologica*, 96(1):10, 2011.

[9] Sudipta Sekhar Das, Rajendra Chaudhary, Sunil Kumar Verma, Shashank Ojha, and Dheeraj Khetan. Pre-and post-donation haematological values in healthy donors undergoing plateletpheresis with five different systems. *Blood transfusion*, 7(3):188, 2009.

[10] Giovanni De Gaetano, Marialaura Bonaccio, and Chiara Cerletti. How different are blood platelets from women or men, and young or elderly people? *Haematologica*, 108(6):1473, 2023.

[11] Niek Den Teuling. *latrend: A Framework for Clustering Longitudinal Data*, 2023. R package version 1.5.1.

[12] Jianghu Dong, Haolun Shi, Liangliang Wang, Ying Zhang, and Jiguo Cao. Jointly modelling multiple transplant outcomes by a competing risk model via functional principal component analysis. *Journal of Applied Statistics*, 50(1):43–59, 2023.

[13] Jianghu J Dong, Liangliang Wang, Jagbir Gill, and Jiguo Cao. Functional principal component analysis of glomerular filtration rate curves after kidney transplant. *Statistical methods in medical research*, 27(12):3785–3796, 2018.

[14] Steven W Enck. *Latent class linear mixed models-a general approach implemented via SAS® macro with a tutorial for clinical researchers*. PhD thesis, The University of North Carolina at Chapel Hill, 2009.

[15] Lise J Estcourt, Janet Birchall, and Shubha Allard. Guidelines for the use of platelet transfusions a british society for haematology guideline.

[16] Cristina Fraumene, Enrico Petretto, Andrea Angius, and Mario Pirastu. Striking differentiation of sub-populations within a genetically homogeneous isolate (ogliastra) in sardinia as revealed by mtdna analysis. *Human Genetics*, 114(1):1–10, 2003.

[17] Massimo Gallerani, Roberto Reverberi, Raffaella Salmi, Michael H Smolensky, and Roberto Manfredini. Seasonal variation of platelets in a cohort of italian blood donors: a preliminary report. *European journal of medical research*, 18(1):1–4, 2013.

[18] Manfred S Green, Israela Peled, and Theodore Najenson. Gender differences in platelet count and its association with cigarette smoking in a large cohort in israel. *Journal of clinical epidemiology*, 45(1):77–84, 1992.

[19] John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

[20] Reka Karuppusami, Belavendra Antonisamy, and Prasanna S Premkumar. Functional principal component analysis for identifying the child growth pattern using longitudinal birth cohort data. *BMC Medical Research Methodology*, 22(1):76, 2022.

[21] Piotr Kokoszka and Matthew Reimherr. *Introduction to functional data analysis*. CRC press, 2017.

[22] Brett Laursen and Erika Hoff. Person-centered and variable-centered approaches to longitudinal data. *Merrill-Palmer Quarterly (1982-)*, pages 377–389, 2006.

[23] Bian Liu and Emanuela Taioli. Seasonal variations of complete blood count and inflammatory biomarkers in the us population-analysis of nhanes data. *PloS one*, 10(11):e0142382, 2015.

[24] Yihui Luan and Hongzhe Li. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics*, 19(4):474–482, 2003.

[25] Eric S Lugada, Jonathan Mermin, Frank Kaharuza, Elling Ulvestad, Willy Were, Nina Langeland, Birgitta Asjo, Sam Malamba, and Robert Downing. Population-based hematologic and immunologic reference values for a healthy ugandan population. *Clinical and Vaccine Immunology*, 11(1):29–34, 2004.

[26] Pavlos Msaouel, Anthony P Lam, Krishna Gundabolu, Grigorios Chrysofakis, Yiting Yu, Ioannis Mantzaris, Ellen Friedman, and Amit Verma. Abnormal platelet count is an independent predictor of mortality in the elderly and is influenced by ethnicity. *Haematologica*, 99(5):930, 2014.

[27] Ronald C Neath and Matthew S Johnson. Discrimination and classification. 2010.

[28] Cécile Proust-Lima, Viviane Philipps, Amadou Diakite, and Benoit Liquet. *lcmm: Extended Mixed Models Using Latent Classes and Latent Processes*, 2023. R package version: 2.1.0.

[29] Cécile Proust-Lima, Viviane Philipps, and Benoit Liquet. Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, 78(2):1–56, 2017.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[31] C Radhakrishna Rao. Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17, 1958.

[32] Patrick Schober and Thomas R Vetter. Repeated measures designs and analysis of longitudinal data: If at first you do not succeed—try, try again. *Anesthesia & Analgesia*, 127(2):569–575, 2018.

[33] Jodi B Segal and Alison R Moliterno. Platelet counts differ by sex, ethnicity, and age in the united states. *Annals of epidemiology*, 16(2):123–130, 2006.

[34] Robert H Shumway, David S Stoffer, and David S Stoffer. *Time series analysis and its applications*, volume 3. Springer, 2000.

[35] Helle Sørensen, Jeff Goldsmith, and Laura M Sangalli. An introduction with medical applications to functional data analysis. *Statistics in medicine*, 32(30):5222–5240, 2013.

[36] South African National Blood Service (SANBS). CLINICAL GUIDELINES FOR THE USE OF BLOOD PRODUCTS IN SOUTH AFRICA, 2014.

[37] AM Stolwijk, HMPM Straatman, and GA Zielhuis. Studying seasonality by using sine and cosine functions in regression analysis. *Journal of Epidemiology & Community Health*, 53(4):235–238, 1999.

[38] Toon W Taris. A primer in longitudinal data analysis. *A Primer in Longitudinal Data Analysis*, pages 1–176, 2000.

[39] Niek Den Teuling, Steffen Pauws, and Edwin van den Heuvel. latrend: A framework for clustering longitudinal data. *arXiv preprint arXiv:2402.14621*, 2024.

[40] Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005.

[41] Yidong Zhou, Satarupa Bhattacharjee, Cody Carroll, Yaqing Chen, Xiongtao Dai, Jianing Fan, Alvaro Gajardo, Pantelis Z. Hadjipantelis, Kyunghee Han, Hao Ji, Changbo Zhu, Hans-Georg Müller, and Jane-Ling Wang. *fdapace: Functional Data Analysis and Empirical Dynamics*, 2022. R package version 0.5.9.