Bayesian Hierarchical Methods for Non-standard Compositional Data

Catherine Holland, MSci

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

SCHOOL OF MATHEMATICS AND STATISTICS

COLLEGE OF SCIENCE AND ENGINEERING

MARCH 2025

Dedicated to my Dad,

I hope you would be proud of this.

Abstract

Compositional data take the form of parts of some whole, consisting of sets of non-negative components. Compositional data can appear as proportions, percentages, general non-negative values or counts. The inherent characteristics of compositional data, i.e. non-negativity and the constraint to some total, pose unique challenges for traditional statistical techniques. Compositional data arise across many real-world applications such as health, environmental, forensic, financial and sports science. Further challenges occur when compositional data also include other advanced data challenges such as multilevel hierarchical structure, non-smooth time series or a spatial structure.

The main technique in the literature to overcome the complexities of compositional data is to transform the components from the simplex (the sample space of compositional data) into Euclidean space (the standard statistical space) using a log-ratio transformation. Once transformed, standard statistical models can be applied. However, while this transformation is powerful, it is not always suitable in practice. There are many features commonly found within compositional data that prohibit log-ratio transformations. For example, when compositional data contain zeros, the log-ratios become undefined. Similarly, when the components contain missing values, some or all of the log-ratio transformations may not produce sensible results. Lastly, when compositional data consist of counts, applying a log-ratio transformation may discard information on how the total count may impact the variance and the possible values the counts can take. Thus, there is a need for frameworks that can handle compositional data containing these features, as well as addressing advanced data challenges.

This thesis presents novel Bayesian hierarchical frameworks designed to overcome the limitations of log-ratio transformations in these instances. We apply and evaluate our proposed frameworks to three applications of compositional data containing both a feature which prevents log-ratio transformations and an advanced data challenge. These include: compositional data containing many zeros and a multilevel hierarchical structure, applied to forensic elemental glass data; non-smooth time series containing a count structure and zero values, applied to COVID-19 variant counts; and compositional data with a spatial pattern containing zeros, applied to tree species proportions across a spatial grid. We assess the performance of our frameworks through both in-sample and out-of-sample predictive experiments, comparing with commonly used models. The results from the predictive experiments demonstrate the effectiveness of our approaches, highlighting their contribution to compositional data analysis and offering a robust alternative for handling real-world compositional data.

All the code and any supplementary material produced for each of the proposed frameworks is available on GitHub: https://github.com/catherineholland1/PhD.git. We do not have permissions to share the data used within this thesis.

Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to express my deepest gratitude to my supervisors, Oliver Stoner and Tereza Neocleous. Your insights, guidance and support throughout this journey have made this research possible. I would also like to acknowledge the support from the College of Science and Engineering for funding my research.

I am incredibly grateful to all the data contributors. Thank you to Professor Grzegorz Zadora from the Institute of Forensic Research, Krakow for proving the forensic elemental glass data. To Dr Anastasia Frantsuzova, I appreciate your contribution in introducing and providing the spatial tree species data from Fera Science UK. I am also grateful to everyone involved in creating the GISAID Initiative COVID-19 variant data. These contributions have been vital in developing and testing the methodologies presented in this thesis.

On a personal note, I extend my heartfelt thanks to all my friends and family who have all helped me get to where I am today. To my friends who have become my Glasgow family, I am forever grateful for QM halls bringing us all together. To my brother William, your attitude to life and helping others is infectious, inspiring me to be a better person everyday. Craig, thank you for all your kindness, love and support throughout this whole journey - I could not have done this without you by my side. Lastly, to my mum, thank you for everything you have done for me throughout my life. Your unwavering support and guidance have shaped me into the person I am today, and for that, I am forever grateful.

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work.

Catherine Holland

Chapter 1

Introduction

In this chapter, we begin by introducing compositional data and the ways in which it is challenging to model using traditional methods. We then discuss the standard approach in dealing with compositional data and why this is not always suitable. We then introduce various data challenges that can naturally occur in compositional data. Finally, the chapter ends with an overview of the content of the thesis.

1.1 Motivation

Compositional data refer to multivariate sets of non-negative components, in the context where we are primarily interested in the size of the components relative to their total and relative to each other. Such data might be measured directly as proportions summing to a total of one (100%), e.g. the proportions of each ingredient in a smoothie, or measured in absolute terms with a different total, e.g. the number of votes cast for each political candidate. Sometimes, real-world measurement of compositional data may set a total first, e.g. collecting a total mass of soil to analyse. Other times, the components might arise in the real world as essentially independent processes, and we construct the compositional context and total later. For example, we could consider the number of times cars made by different manufacturers crash per year, and then consider the relative share of the different manufacturers in the total crashes. Compositional data arise in a wide range of fields, as seen in forensic science (Napier et al., 2015), environmental statistics (Zuo et al., 2013) or health data (Janssen et al., 2020).

A classical definition of compositional data, given by Aitchison (1982), considers non-negative vectors $\mathbf{x}$ with elements $x_1,\ldots,x_D$ that are subject to a unit sum constraint, i.e. $\sum x_i = 1$. The sample space for such compositional data are defined as the simplex (Aitchison, 1982):

$$S^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D) : x_i > 0 \ (i = 1, 2, \ldots, D), \sum_{i=1}^{D} x_i = 1\}. \tag{1.1.0.1}$$

The early work of compositional data analysis has been dominated by the approach of Aitchison (1982, 1986). This has followed a relatively restrictive framework of defining compositional data on the simplex and guiding the methods researchers have utilised to analyse compositional data. However, in some practical cases, compositional data do not initially reside on the simplex, such as count data where the total varies across observations. If required, we can transform these data onto the simplex by dividing each component by the total sum. By adhering too strictly to the early restrictive definition of compositional data, we risk overlooking more intuitive or practical approaches that better capture the true nature of the data.

Given the complex characteristics of compositional data, it cannot be modelled using standard statistical techniques. Some traditional statistical methods often assume independence between variables and do not account for the inherent constraints present in compositional data. Applying standard techniques could lead to misleading relationships and biased results due to the relative nature of the data. For example, changes in one component affect the values of others, violating key assumptions of many classical statistical models. Therefore, we require specialised methods to analyse and interpret compositional data while respecting its underlying structure.

The main technique to deal with compositional data is to apply log-ratio transformations, which map the data from the constrained simplex space to an unconstrained Euclidean space. This allows standard statistical methods to be used without violating the compositional structure. Log-ratio transformations have been proposed to deal with compositional data:

additive log-ratio (ALR), centered log-ratio (CLR) (Aitchison, 1982) and isometric log-ratio (ILR) (Egozcue et al., 2003). These log-ratio transformations have been widely used across various fields. For example, ALR has been used for air pollution time series data (Al-Dhurafi et al., 2018), CLR for modelling bacterial data (Sisk-Hackworth et al., 2020) and ILR for modelling birth population data (Martinez et al., 2020). However, while log-ratio transformations have been shown to be powerful tools for analysing compositional data and are the most straightforward technique for most compositional data problems, they may not be suitable in all cases, as compositional data can often have features that prevent their use - this is where we will focus our work.

One such feature is the presence of zeros. There are two different types of compositional zeros: rounded and structural. Rounded zeros represent components that fall below a detection limit and therefore, are not true zero values. In contrast, structural zeros are considered true zeros that can represent actual zero values or indicate that a component belongs to a different group. As structural zeros carry an informative value, they cannot simply be ignored. In the presence of zeros, log-ratio transformations are unsuitable as they are undefined for zero values. A common approach to address this issue is to impute the zeros with a small value so that a log-ratio transformation can be fitted. However, imputation contradicts the idea that structural zeros are informative. Additionally, if we impute our zero values, this can lead to further challenges, such as the choice of imputation value/method, potential distortion of the covariance structure and potential violations of the principles of compositional data.

Compositional data can also often involve a count structure that make log-ratio transformations less appropriate. Applying a log-ratio transformation to the compositional counts results in discrete variables in the real space that may not be suitable for modelling using standard methods. Thus, we likely cannot assume a smooth continuous distribution (e.g. a Normal distribution) for modelling, unless the total is sufficiently large so that the gaps between possible values are negligible. Additionally, it potentially discards information on

how the total count may impact the variance and the possible values the counts could take - for example, if the total is 10, the scaled count can only be 0, 0.1, 0.2 etc. These issues become more problematic the smaller the total count is, increasing the likelihood of zeros and reducing the number of unique values the compositional counts can take.

Another area of compositional data analysis where log-ratio transformations are unsuitable is where the data include missing or unobserved values for some components. Log-ratio transformations require complete data for all components in order to be correctly defined. When a component value is missing, some or all of the log-ratio transformations will not produce sensible results, as the necessary relative proportions cannot be properly computed. Imputation methods that do not account for the compositional nature of the data may be inappropriate as they can introduce artificial relationships and distort the true compositional structure. A compositional imputation approach could be considered, but they would carry risks of bias and overconfidence, particularly if missing values are not missing at random. If certain components are more likely to be missing under specific conditions, applying a log-ratio transformation without properly addressing these missing values could lead to systematic biases.

In this thesis, our research scope is expanding methodology for modelling compositional data with features that prohibit the use of log-ratio transformations. Meanwhile, compositional structures can naturally arise in the same situations as other advanced data challenges. One of these data challenges is when compositional data feature a multilevel (hierarchical) grouping structure, e.g. percentage of students achieving different grade classification within their class, within their school and within their county. In such situations, the components are correlated in a structured way.

Compositional data can also appear in the context of time series analysis, requiring specific techniques to account for both the compositional nature and underlying temporal structure. Traditional time series methods typically do not account for the compositional nature of the data, so custom methods are required. Further challenges occur when the time series is non-smooth, i.e. when the data exhibit abrupt changes or irregular fluctuations, rather than following a continuous and predictable trend over time. Similarly, compositional data can be observed over a spatial or geographical dimension, such that we likely need to account for spatial structure/dependence when modelling the relationships between components.

As such, further pinpointing the scope of work in this thesis, we will focus on developing solutions for situations where the data both have at least one feature that generally prohibits a log-ratio transformation and has at least one other modelling complexity. Specifically, in each chapter we will address the following combinations shown in Table 1.1.

Table 1.1: Compositional data situations addressed within each chapter in this thesis, including the prohibitive feature that generally prevents a log-ratio transformation approach and the advanced data challenge posed by the application.

Chapter	**Prohibitive feature w.r.t. log-ratio transformation**	**Advanced data challenge**
3	zeros	multilevel hierarchical groupings missing covariate prediction
4	zeros count data	non-smooth time series
5	zeros missing values	spatial structure

To address these, we will propose novel approaches combining Bayesian hierarchical models for compositional data that are less restrictive than the log-ratio transformation approach, with advanced modelling structures at the latent level, including latent data clustering, hidden Markov models and spatial penalised regression splines.

For the work in this thesis, we opt for Bayesian hierarchical frameworks as the fundamental support for our development because we believe that they grant us the most flexibility to account for different forms of compositional structure (e.g. counts) and to incorporate other advanced data structures in whatever way seems fit. Bayesian inference provides several key advantages over alternative approaches. One of the biggest strengths of Bayesian analysis is the ability to incorporate prior knowledge into the model framework. Bayesian hierarchical models also allow for flexible modelling of complex, multilevel dependencies, such as nested or grouped data. Bayesian inference provides the ability to quantify uncertainty in parameter estimates and predictions directly. Additionally, Bayesian models can naturally accommodate missing data by treating missing values as unknown quantities, which are estimated along with the parameters of interest. Finally, Bayesian analysis facilitates thorough model checking through posterior predictive checks, offering a powerful tool for assessing the adequacy of the model fit.

Creating new frameworks using custom Bayesian hierarchical methods would be challenging without recent developments in software. Throughout this thesis, we implement our frameworks using the *NIMBLE* package (Valpine et al., 2017), which allows for flexible implementation of Bayesian models. NIMBLE models are written in the BUGS language, like JAGS (Plummer, 2003), and then compiled automatically into C++ (Stroustrup, 1986) for fast execution. Within NIMBLE, the user can choose any combination of samplers for different model parameters, either chosen from pre-included samplers or straightforwardly adding their own sampling methods to the algorithm. Additionally, NIMBLE allows the user to create and include their own functions, algorithms and probability distributions. This allows freedom to choose different sampling algorithms and add modifications to the model. Using NIMBLE has given us the opportunity to produce effective Bayesian hierarchical methods that are widely applicable.

1.2 Thesis Overview

The remainder of this thesis is structured as follows:

Chapter 2: Provides an overview of compositional data concepts, along with some current common methods. Note that the subsequent chapters offer a deeper critical review of literature relevant to each chapter's topic.

Chapter 3: Presents a Bayesian hierarchical model for compositional data containing a large number of structural zeros. We develop a system for automation in the splitting of presence and absence of elements, requiring less strenuous expert input and making it more practical in real-world analysis. The proposed framework is examined through an experiment to predict the classification of glass items from a forensic elemental glass database.

Chapter 4: Explores a compositional time series, which contains counts over time and exhibits non-smooth behaviour. Here, we outline a GDM-HMM framework for compositional time series. We created and tested our methodology using COVID-19 variant data consisting of weekly counts from countries worldwide for each of the COVID-19 variants as defined by the World Health Organisation. We examine our proposed framework against simpler common time series models including a Random Walk and Dynamic Linear Model.

Chapter 5: Addresses compositional data that include spatial locations or coordinates. We construct a framework for spatial compositional data using the Generalised-Dirichlet-Multinomial distribution to model compositional counts arranged over a spatial grid, allowing for zero counts and missing values. We design a posterior predictive experiment where there are missing values in the compositions and test performance of the proposed framework against a benchmark Generalised Additive Model approach.

Chapter 6: Summarises the work, presents concluding remarks and potential further research.

Chapter 2

Overview of Compositional Data

2.1 Introduction

Compositional data refer to multivariate non-negative components ($\boldsymbol{x} = x_1, \ldots, x_D$), in the context where we are primarily interested in the size of the components relative to their total and relative to each other. Compositional data can appear as proportions, percentages, general non-negative values (e.g. object masses), or counts, and examples appear within many different statistical areas such as forensic science (Napier et al., 2015), geochemical statistics (Zuo et al., 2013), geology (Qiu et al., 2024), health data (Janssen et al., 2020), sports science (Lobo et al., 2025) and financial data (Carreras et al., 2020). Table 2.1 presents two simple examples of compositional data. In both cases, the total count of each row is shown in the final column.

Table 2.1: Examples of compositional data.

(a) Proportion data.

a	b	c	Row sum
0.10	0.49	0.41	1
0.25	0.55	0.20	1
0.12	0.03	0.85	1

(b) Count data.

x	y	z	Row sum
56	4	40	100
25	5	12	42
85	10	24	119

Compositional data analysis has largely been shaped by the approach developed by Aitchison (1982, 1986). Here, compositional data are defined as non-negative vectors that represent proportions of some whole that is subject to a unit sum constraint. The sample space of these compositional proportions is the simplex, S^D, and is defined as (Aitchison, 1982):

$$S^D = \left\{ \boldsymbol{x} = (x_1, x_2, \ldots, x_D) : x_i > 0 \, (i = 1, 2, \ldots, D), \sum_{i=1}^{D} x_i = 1 \right\}, \tag{2.1.0.1}$$

The simplex has a $(D-1)$ dimensional real space. As the $\boldsymbol{x}$ are proportions that are constrained to be positive and sum to one, they can be interpreted as forming a linear combination within the simplex space.

This definition of compositional data, introduced by Aitchison (1982), builds upon the earlier work of Pearson (1897), who first documented spurious correlation. In compositional data, this issue arises because the components are constrained to sum to a constant (typically one), meaning they lie within a simplex. As a result, an increase in the proportion of one component necessarily implies a decrease in the others, inducing negative correlations among components even when no such relationships exist in the absolute values. This dependency is a central challenge within compositional data analysis. Recognising the problem outlined by Pearson (1897), Tanner (1949) suggested the use of log-ratio transformations could help overcome this. Later, Chayes (1960) formally linked Pearson's concept to compositional data, though no methods were introduced to remove the effect of the constraint. Chayes's connection spurred further developments in compositional data analysis, ultimately leading to the adoption of log-ratio transformations. These transformations are explored in Section 2.2.

However, this issue of negative correlations is not necessarily present when compositional data reside outside this narrow definition. For example, the number of Ford car crashes is not negatively correlated with the number of Volkswagen cars crashing. However, when the number of crashes is expressed as proportions of the total number of crashes, a negative correlation can emerge. This is because an increase in the proportion of Ford crashes relative to the total would automatically reduce the proportion of Volkswagen crashes, even if the actual number of Volkswagen crashes remains unchanged.

Aitchison developed principles for compositional data residing on the simplex within Aitchison (1992) and Aitchison et al. (2005). The scale invariance principle means that multiplying all components within a composition by a positive constant does not change the relative information contained within the data, i.e. the compositional ratios remain unchanged under rescaling. A subcomposition is a subset of the full composition where the subcompositional coherence principle states that relationships between the parts remain valid even when analysing only a subset of the components, ensuring that relationships observed within the subcomposition remain consistent with those in the full composition. The subcompositional

dominance principle states that if one component dominates in the full composition, it should dominate in any subcomposition. Lastly, the permutation invariance principle states that the order of components should not affect the analysis, reinforcing the idea that the structure of the data - not the ordering of its parts - shapes the analysis.

Standard distance measures, such as the Euclidean distance, should not be computed directly for compositional data residing on the simplex. Aitchison (1986) developed a distance measure for such compositional data, defined as the Aitchison distance, to quantify the dissimilarity between two compositions. It is defined in terms of the log-ratio transformations of the components and satisfies the properties developed by the same author outlined above.

$$d_a(\boldsymbol{x},\boldsymbol{y}) = \left\{ \sum_{i=1}^{D} \left[\log\left(\frac{x_i}{g_m(\boldsymbol{x})}\right) - \log\left(\frac{y_i}{g_m(\boldsymbol{y})}\right) \right]^2 \right\}^{1/2}, \tag{2.1.0.2}$$

where $g_m(\boldsymbol{x}) = (x_1,\ldots,x_D)^{1/D}$ and $g_m(\boldsymbol{y}) = (y_1,\ldots,y_D)^{1/D}$ are the geometric means of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

Graphical visualisations of compositional data, as discussed by Aitchison (1986), are influenced by Aitchison's principles of compositional data and Aitchison's geometry. One widely used technique is the ternary diagram which provides a two-dimensional projection of a three-part composition. As shown in Figure 2.1, for the data in Table 2.1 (a), the ternary diagram represents the simplex as an equilateral triangle, where the sum of the distances remains constant for any chosen components of $\boldsymbol{x}$ (Filzmoser et al., 2008). Since the entire simplex must fit within the ternary diagram, its borders conceptually represent infinity, a direct consequence of the relative scale property of compositions. Points near a vertex (e.g. the bottom right near component c) indicate a high proportion of that component (such as the 0.85 value in Table 2.1 (a)). Conversely, points near the centre represent compositions

with nearly equal proportions of all three components, as seen with the two points near the middle of Figure 2.1. A point located directly on a vertex corresponds to a composition where that component equals 1, while the other two components are 0. The further a point is from a vertex, the lower its contribution is to that component.

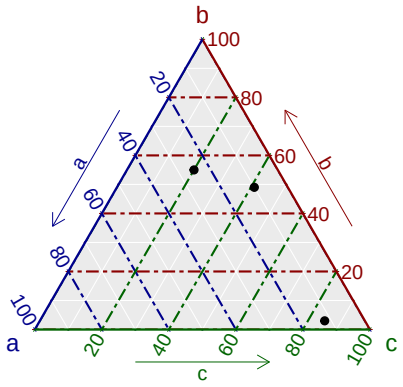


Figure 2.1: Ternary diagram for the example three-part composition from Table 2.1.

The main limitation of compositional graphical approaches is that data are typically visualised and interpreted using Euclidean distance, whereas Aitchison distance better reflects the compositional structure. This distinction must be considered when interpreting ternary diagrams. More recently, visualisations have been extended to four-part compositions, represented as a regular tetrahedron instead of a triangle. However, beyond four-part compositions, visual representation becomes increasingly challenging.

This initial, narrow definition of compositional data, which confines the data to values constrained to the simplex (e.g. proportions or percentages that sum to one), is too restrictive. It assumes that compositions must be analysed through their relative rather than absolute values. However, in practice, compositional data often originate from counts or absolute values, such as the number of people voting for each political party. While the primary focus is typically on the relative information between the parts, the absolute values remain important as they influence the variance and the overall dynamics of the data. For instance, scaling raw counts to sum to one can artificially alter correlations, sometimes transforming a positive correlation between components into a negative one. This highlights the importance of considering compositional data in its broader context. By broadening the definition,

as discussed within Firth et al. (2023), to include raw or unscaled compositions, we gain a deeper understanding of the underlying relationships – including correlations that may not be apparent under the traditional definition of compositional data. Firth et al. (2023) propose modelling the compositions directly in their original scale, allowing for both absolute and relative information to be considered. This is particularly useful when total size matters (e.g. the total number of votes cast in an election).

Since the foundation work of Aitchison (1982), the understanding of compositional data has largely followed a restrictive framework, with Aitchison's principles governing the techniques used. This has potentially prevented other alternative approaches being developed within a compositional framework, such as directly modelling compositional counts. By adhering too strictly to these conventions, we risk overlooking more intuitive or practical approaches that better capture the data. Expanding beyond these constraints opens new possibilities for analysing compositional data in ways that preserve both absolute and relative information.

In Chapters 3, 4, and 5, we will follow this rationale and motivation to expand methodology for modelling compositional data. First, though, in the remainder of this chapter, we will outline the current transformations proposed for compositional data in Section 2.2. Following this, we will address the issue of zeros in compositional data and discuss methods from the literature that have been proposed in order to handle these zeros.

2.2 Transformations

To address the inherent constraints and produce meaningful analysis of compositional data, various transformations have been proposed in the literature to convert compositions existing on a constrained simplex to an unconstrained Euclidean space. Using such transformations allow for standard statistical approaches to be applied to the transformed compositions without violating the compositional structure. Once the analysis is complete, the compositions can be transformed back to the simplex by applying the inverse of the chosen transformation. The primary transformation developed for compositional data is the log-ratio family of transformations, described in the next section.

2.2.1 Log-ratio Transformations

The simplest transformation is the additive log-ratio transformation (ALR) introduced by Aitchison (1982). ALR involves taking the logarithm of the $D-1$ components to the remaining component, represented as:

$$\text{ALR}(\boldsymbol{x}) = \left(\log\left(\frac{x_1}{x_D}\right), \ldots, \log\left(\frac{x_{D-1}}{x_D}\right)\right). \tag{2.2.1.1}$$

This transformation uses the compositional ratios relative to one component (x_D), removing the constraint and enabling the use of standard statistical techniques, as the data are mapped to an unbounded space. Here, any component can be chosen to be the common divisor (x_D). This means that the resulting transformation is asymmetric as it depends on the choice of

the divisor. The results of the transformation will change depending on which component is selected as the divisor, making it sensitive to this choice. This could be challenging if there is not one component that is always non-zero, i.e. $x_1, \ldots, x_D > 0$, which is problematic because there would then be no suitable component to serve as the divisor (x_D).

Examples of ALR transformation used in practice include: Greenacre et al. (2021) who apply the ALR to microbiome compositional data; Yoo et al. (2022) where the ALR is used to transform immune cellular compositions before fitting a generalised linear model with Dirichlet distribution; Al-Dhurafi et al. (2018) use the transformation for time series air pollution index data prior to fitting a Vector Autoregressive (VAR) model, and Leininger et al. (2013) who apply the ALR to land use data along with multivariate Conditional Autoregressive (CAR) model.

The centered log-ratio transformation (CLR), developed by Aitchison (1986), addresses the limitations of ALR by eliminating the need for a single reference component. Like ALR, CLR transformation adheres to Aitchison's principles of compositional data, as outlined earlier. However, instead of singling out one component as a common divisor for the compositional ratios, CLR transformation uses the geometric mean as the divisor, thus preserving the dimension of $\boldsymbol{x}$. By avoiding the selection of a specific component as the divisor, CLR treats the composition as symmetric. CLR is given as:

$$\mathrm{CLR}(\boldsymbol{x}) = \left(\log\left(\frac{x_1}{g(\boldsymbol{x})}\right), \ldots, \log\left(\frac{x_D}{g(\boldsymbol{x})}\right) \right), \tag{2.2.1.2}$$

where $g(\boldsymbol{x})$ is the geometric mean of $\boldsymbol{x}$.

An advantage of CLR is that it is an isometric transformation of the simplex and can be visualised over all the D-parts of the composition when performing exploratory analysis of the data. However, the covariance matrix of CLR is singular due to the inherent linear dependence among the components. Since the sum of the CLR-transformed components is always zero, they are perfectly linearly dependent, meaning that the transformed data lies in a $(D-1)$-dimensional subspace of the D-dimensional space. This linear dependence results in a covariance matrix where the determinant is equal to zero. Consequently, the covariance matrix is singular because all D components are involved in the transformation, but the dependency is preserved in the transformed space. This singularity makes it challenging to apply certain standard statistical techniques, especially in multivariate settings (e.g. discriminant analysis).

The CLR transformation has also been applied widely in the literature: Sisk-Hackworth et al. (2020) use the transformation to analyse bacterial data; Shang et al. (2022) apply the CLR before fitting multivariate and multilevel functional time series methods to forecast age-specific death counts across populations; and recently Bennett et al. (2025) employ the CLR with a scale uncertainty/information model to comparative glycomics data containing glycan concentrations.

The isometric log-ratio transformation (ILR), introduced by Egozcue et al. (2003), overcomes the limitations of the previous log-ratio transformations by projecting the compositional data into real coordinates with respect to an orthonormal basis in the simplex. This transformation ensures that the geometry of the simplex, as defined in Aitchison's geometry (Aitchison, 1986) is preserved, creating a direct link between the angles and distances in the simplex and also in Euclidean space. This is the most complex log-ratio transformation within the literature as it allows the angles and distances in the simplex to be linked with angles and distances in real space. Within Egozcue et al. (2003), the ILR transformation is given as:

$$\mathrm{ILR}(\boldsymbol{x}) = (y_1, \ldots, y_{D-1}), \tag{2.2.1.3}$$

where

$$y_i = \sqrt{\frac{i}{i+1}} \cdot \ln\left(\frac{g(x_1,\ldots,x_i)}{x_{i+1}}\right), \quad i = 1,\ldots,D-1,$$

and $g(x_1,\ldots,x_i)$ denotes the geometric mean of the first i components of the composition $\boldsymbol{x}$.

The ILR transformation results in a $(D-1)$ dimensional real space; the dimensionality of the composition is reduced by one, as the transformation uses a combination of the components for the divisor. The ILR avoids the arbitrariness of selecting a divisor for the ALR and the singularity from the CLR. The relationships between the three log-ratio transformations proposed are described in Egozcue et al. (2003). Therefore, the ILR transformation has some conceptual advantages over the other log-ratio transformations as it preserves compositional relationships. However, constructing the orthonormal basis in the simplex requires careful consideration, and the resulting coordinates may be less intuitive to interpret, particularly for large datasets.

Examples of ILR applied in the literature include: Karacan et al. (2018), who use sequential Gaussian simulation of isometric log-ratio transformed compositions to map the chemical properties of coal; Martinez et al. (2020), who transform birth population data using the ILR before modelling the spatial random effect with a CAR structure; Nguyen et al. (2021), who analyse election vote share data using a spatial autoregressive (AR) model; Oh et al. (2024), who use the transformed compositions to produce a groundwater pollution index from robust principal component analysis (RPCA); and Egozcue et al. (2024), who use geochemical river compositional data to illustrate how an ILR along with compositional techniques help explore compositions and detect patterns and outliers in the data.

Other transformations stemming from the log-ratio transformations that have been developed include the complementary log-log transformation (Neocleous et al., 2011). This transformation involves taking the logarithm of the negative of the log-ratio transformed data. Initially, we take the log-ratio transformation of the components $\boldsymbol{x}$:

$$u_i = \log\left(\frac{x_i}{x_D}\right), \tag{2.2.1.4}$$

for $i = 1,\ldots,D-1$ where x_D is the reference component. The complementary log-log transformation is then defined as:

$$v_i = \log(-u_i + c), \tag{2.2.1.5}$$

for $i = 1,\ldots,D-1$ where c a small positive constant added to ensure that $(-u_i + c)$ remains positive before taking the logarithm. In this case, all components must be strictly positive in order to compute the transformation, i.e. if $-u_i + c \leq 0$, the logarithm becomes undefined. An advantage of the complementary log-log transformation is that the resulting transformed data may resemble a Normal distribution more closely, especially in cases where the log-ratio transformed data are skewed. This transformation can help make the data more symmetric and better suited to apply standard statistical techniques. However, this transformation is sensitive to the choice of the positive constant c. Similarly to the ALR, the complementary log-log transformation is asymmetric because it is dependent on the choice of the reference component x_D.

Overall, despite log-ratio transformations being the most common and straightforward approach for most compositional data problems, they may sometimes not always be the most suitable choice. For example, if the compositions contain zero values, or there are any missing values in the compositions, the log-ratio transformations may not be directly applicable without prior work conducted on the compositions. Furthermore, compositional data can of-

ten include counts that sum to some total, but log-ratio transformations of count data could fail to yield continuous data suitable for modelling with standard methods, as explained in Chapter 1. These features motivate the development of alternative transformations for compositional data that do not rely on log-ratio methods.

2.2.2 Alternative transformations

This section explores alternative transformations from the literature that may be more suitable than the log-ratio in certain instances of compositional data.

The simplest transformation to apply is taking the square root of the compositional ratios, with a common divisor chosen (Stephens, 1982).

$$u_i = \sqrt{\frac{x_i}{x_D}}, \tag{2.2.2.1}$$

for $i = 1, \ldots, D-1$ and for the divisor component x_D. The square root transformation can handle zeros present in the data, making it computationally easier to implement as no values need to be modified, although it does require a divisor component that does not contain any zeros. In some applications this transformation has been shown to stabilise the variability effectively, as in Napier (2014).

Wang et al. (2007) avoid the complication of zeros within the components by applying a hyperspherical transformation to the compositional data. Firstly, the square root is applied to all D-parts of a composition $\boldsymbol{x}$:

$$u_i = \text{sqrt}(\boldsymbol{x}) = (\sqrt{x_1}, \ldots, \sqrt{x_D}), \tag{2.2.2.2}$$

transforming onto the surface of the $(D-1)$-dimensional hypersphere. The coordinates of u_i are then mapped to their polar coordinates using a recursive relationship. The polar coordinate system is a two-dimensional system in which each point on a plane is determined by a distance from a reference point and an angle from a reference direction. Wang et al. define the recursive relationship for computing for $s_i = \sqrt{x_i}$ is

$$\begin{aligned} \omega_1 &= \arccos u_1, \\ \omega_2 &= \arccos \frac{u_2}{\sin \omega_1}, \\ &\vdots \\ \omega_{P-1} &= \arccos \frac{u_{P-1}}{\sin \omega_1 \sin \omega_2 \cdots \sin \omega_{P-2}}. \end{aligned} \tag{2.2.2.3}$$

The dimension of the resulting transformed ω-vector is $d = (D-1)$ and the zeros map to $\arccos 0 = \pi/2$. An advantage to using the hypersphere transformation over the log-ratio transformations outlined above, is that it can handle zero values. Scealy et al. (2011) use the hypersphere transformation to allow for directional data distributions - such as the Kent distribution (Mardia et al., 2000) - to be used when modelling compositional data.

More recently, the power or α-transformation has been developed to avoid the use of log-ratio transformations. This is given in Tsagris et al. (2011) as:

$$u_i = \left(\frac{x_i \alpha}{\sum_{j=1}^{D} D x_j^{\alpha}} \right). \tag{2.2.2.4}$$

The use of this α-transformation allows greater flexibility by allowing a choice between the approaches of Aitchison's geometry (i.e. the simplex) and the Euclidean space, where the decision is made depending on the choice of α (Tsagris et al., 2016). This allows for tailored transformation methods depending on data characteristics. When $\alpha \to 0$, the transformation tends to Aitchison's geometry, whereas when $\alpha \to 1$, the transformation tends to the Euclidean space. An advantage in using an α-transformation is that when $\alpha > 0$ the transformation is well-defined even when there are zero values present. However, this trans-

formation is dependent on the value of α chosen, and may require optimisation methods to determine the correct value of α. Additionally, this transformation could make the interpretation of the original compositions more difficult, as the transformed data might lose some of the original compositional structure.

2.3 Compositional Zeros

Handling zeros in compositional data pose unique challenges, as they can represent an absence of a component, a measurement error or sampling limitations, and require specialised techniques to ensure accurate analysis. In these instances, log-ratio transformations would be problematic to apply as they lead to undefined log-ratios or potential loss of information. Within compositional data, there are two different types of compositional zeros: rounded and structural. Different approaches have been adopted in the literature for both types of zeros.

2.3.1 Rounded Zeros

Rounded zeros represent values that falls below some detection limit, and therefore are not true zero values. Addressing these zeros typically involve missing data techniques; particularly, not missing at random (NMAR) methods, since values below the detection limit ε remain unobserved. Here, in order to be able to apply a log-ratio transformation, these values would need to be replaced.

The techniques to address rounded zeros include both parametric and non-parametric methods to replace the zeros with a constant value that is at or below the detection limit. To maintain the constraint of compositional data, the non-zero components must be adjusted. Tsilimigras et al. (2016) compare the approaches noting that while non-parametric methods can be robust, they may lose statistical power in some cases. In contrast, parametric methods provide more accurate variance estimates for meaningful inference. However, when the proportion of rounded zeros is small, both methods tend to yield similar results.

The simplest method to deal with rounded zeros is simple replacement (Martín-Fernández et al., 2006). In this method, each rounded zero value is replaced with a fixed value and then the entire composition is rescaled to maintain the constraint. As a result, the imputed values depend not only on the chosen imputation threshold $\boldsymbol{\delta}$ but also on the number of rounded zeros in the composition $\boldsymbol{x}$. For a D-part composition $\boldsymbol{x}$ containing rounded zeros, the composition is replaced by a non-zero composition $\boldsymbol{u}$ using:

$$u_d = \begin{cases} \frac{c}{c+\sum_{\{k:x_k=0\}}\delta_k}\delta_d, & \text{if} x_d = 0, \\ \\ \frac{c}{c+\sum_{\{k:x_k=0\}}\delta_k}x_d, & \text{if} x_d > 0, \end{cases} \tag{2.3.1.1}$$

where $c = \sum x_d$ ensures the unit sum constraint holds, $\boldsymbol{\delta}$ is a chosen replacement value below the limit of detection $\boldsymbol{\varepsilon}$.

Aitchison (1986) proposed an additive replacement method for handling rounded zeros. For a D-part composition $\boldsymbol{x}$ containing Z zeros is replaced by a non-zero composition $\boldsymbol{u}$:

$$u_d = \begin{cases} \frac{\delta(Z+1)(D-Z)}{D^2}, & \text{if} x_d = 0, \\ \\ x_d - \frac{\delta(Z+1)Z}{D^2}, & \text{if} x_d > 0, \end{cases} \tag{2.3.1.2}$$

where $\boldsymbol{\delta}$ is a chosen replacement value below the limit of detection $\boldsymbol{\varepsilon}$. A limitation of this method is that it fails to preserve the ratios between the components in $\boldsymbol{x}$ and $\boldsymbol{u}$. Consequently, this violates Aitchison's principle of subcompositional coherence, meaning that relationships within subcompositions may not be consistent with the original data.

This led to the introduction of the multiplicative replacement method by Martín-Fernández et al. (2000), who extend the simpler predecessors by ensuring that Aitchison's principles are preserved. After zeros are replaced with a fixed value $\boldsymbol{\delta}$, the non-zero components are adjusted multiplicatively to preserve the original ratios between them. This adjustment does not affect the relative nature of the data. Unlike the previous methods, this approach does not rely on the number of components D or the number of rounded zeros Z, but only on the threshold value $\boldsymbol{\delta}$:

$$u_d = \begin{cases} \delta_d, & \text{if} x_d = 0, \\ \\ x_d - \frac{x_d}{c} \sum_{\{d:x_d=0\}} \delta_d, & \text{if} x_d > 0, \end{cases} \tag{2.3.1.3}$$

where $c = \sum x_d$ is the unit sum constraint. However, the question arises which value to select for $\boldsymbol{\delta}$ as this must be lower than the detection limit $\boldsymbol{\varepsilon}$.

A comparison of the performance of additive and multiplicative replacement methods can be found in Martín-Fernández et al. (2003), where the multiplicative replacement method is recommended for imputation, as it is simpler, computationally efficient, and more coherent when handling rounded zeros.

While non-parametric methods provide a simple approach to handling rounded zeros, they may lack efficiency and sensitivity in certain cases. This limitation has led to the development of parametric approaches that aim to address rounded zeros more effectively. Palarea-Albaladejo et al. (2007) propose a parametric approach to treat rounded zeros that fall below the limits of detection. Specifically, Palarea-Albaladejo et al. (2007) use a modified

version of the Expectation–Maximisation (EM) algorithm, allowing zero values to be treated as missing data. This modification of the EM algorithm is implemented using ALR (Equation (2.2.1.1)) which produces suitable estimates for the values below the detection limit ε. This process is independent of the component selected as the divisor in the transformation. In this method, the unobserved values (i.e. rounded zeros) are replaced by small values that are conditionally estimated based on the observed data using a probabilistic model. Palarea-Albaladejo et al. (2007) demonstrate that the EM algorithm performs better than the non-parametric multiplicative replacement method, particularly when the number of zeros increases. Unlike the non-parametric methods, this parametric approach takes into account information from the covariance structure, reducing the artificial correlation and providing a better estimation of the variability within the composition.

2.3.2 Structural Zeros

Structural or essential zeros are zeros that are considered to be true zeros. This could represent an actual zero value or an indication that it belongs to a different group or component. For example, when considering food group intakes, the component of meat in a vegetarian diet would always be zero, as it is excluded from the diet. Unlike rounded zeros, which are imputed based on assumptions, structural zeros carry significant informative value within the component and cannot be ignored. These zeros can sometimes serve as indicators of underlying patterns or structural features in the data. Many difficulties come with structural zeros, as they are more complex and the zero value is informative and cannot be ignored. As a result, various modelling techniques have been developed to address the complexities of handling structural zeros correctly. When zeros are considered true values in compositional data, the standard approach of applying a log-ratio transformation becomes problematic. Since the log of zero is undefined, zero values in the composition must be replaced with small positive values before applying the transformation. This undermines the assumption that these zeros are informative and represent true values. Such limitations highlight the need

for alternative approaches that avoid taking the logarithm of the compositional ratios, preserving the integrity of the zero values while still allowing for meaningful analysis. Within the compositional data analysis literature, less research has been conducted to deal with structural zeros compared with rounded zeros. Below we present some of the approaches proposed to model structural zeros.

Aitchison et al. (2003) introduce a two-stage model to handle structural zeros. The first stage identifies where zeros occur, while the second determines how the remaining values are distributed among non-zero components. To facilitate this, the data are first organized into an incidence matrix, $\boldsymbol{I}$. The first row of $\boldsymbol{I}$ represents the full $\boldsymbol{D}$-part composition, while the following $\boldsymbol{D}$ rows contain binary indicators (0 or 1), denoting whether each component corresponds to a structural zero. Aitchison et al. noted that computational challenges arise when estimating parameters via maximum likelihood estimation. While likelihood expressions exist, the complexity lies in identifying the various subcompositions within the likelihood function. These challenges are difficult to resolve analytically and often require computational methods such as Markov chain Monte Carlo (MCMC). Building on this framework, Zadora et al. (2010a) also apply a two-stage model, treating the presence of zeros with an independent binary model, as originally suggested by Aitchison et al. (2003).

Butler et al. (2008) and Leininger et al. (2013) propose modelling structural zeros using a latent Gaussian random variable, allowing constrained data to be analysed in $\mathbb{R}^D$ via a Multivariate Normal distribution. This method orthogonally projects points outside the simplex onto its faces and vertices, but it tends to assign excessive probability to vertices. As dimensionality increases, identifying the correct projection regions becomes more complex, affecting maximum likelihood estimation. A key limitation of the latent Gaussian model is its violation of scale invariance and subcompositional coherence principles. However, Butler et al. (2008) argue that any method attempting to model zero and non-zero proportions together will inevitably break these principles, as ratios become infinite along simplex boundaries. Developing on this further, Tsagris (2018) proposes an alternative projection

method, moving points along the line and thus connecting them to the simplex centre. This model assumes a latent multivariate normal distribution, where zero values indicate latent values outside the simplex. Unlike Butler et al. (2008), this approach simplifies likelihood estimation in any dimension, but it can only handle zeros in one component, limiting its applicability in real-world data.

To deal with the structural zeros present, another approach explored is to split the data into subsets according to the pattern of the presence and absence of zeros. Neocleous et al. (2011) and Napier et al. (2015) employ this prior to fitting random effects models to each subset of the data.

Another approach in the literature uses zero-inflated models to address compositional structural zeros. Stewart (2013) explores zero-inflated distributions which are particularly useful as they preserve the informative nature of zeros without modification and presents two zero-inflated distributions. The Zero-Inflated Logistic-Skew Normal (ZILS) distribution (Stewart, 2013) is given as:

$$f_{SL}(p) = \begin{cases} \theta, & p = 0, \\ (1-\theta)\mathrm{LS}(p;\mu,\sigma^2,\alpha), & 0 < p < 1. \end{cases} \tag{2.3.2.1}$$

The Zero-Inflated Beta (ZIB) distribution (Stewart, 2013), given as:

$$f_B(p) = \begin{cases} \theta, & p = 0, \\ (1-\theta)B(p;\mu,\phi) & 0 < p < 1, \end{cases} \tag{2.3.2.2}$$

where for $0 < 0 < 1$, $B(p;\mu,\phi)$ denotes the Beta distribution and $0 < \mu < 1$ and $\phi > 0$. An advantage of the ZIB model is its ability to provide direct estimation without data transformation, making it easier to implement.

Another common approach for modelling in compositional data analysis is the Dirichlet distribution, which naturally models proportions that sum to one, preserving the relative nature of the components (Connor et al., 1969). However, the Dirichlet is undefined when one or more components in the composition are zero. This is due to the behaviour of the term $x_i^{\alpha_i - 1}$ from the probability density function (PDF):

$$f(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\boldsymbol{\alpha_i})} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, \tag{2.3.2.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ are the concentration parameters with $\alpha_i > 0$ and $\sum_{i=1}^{k} x_i = 1$. Each $\boldsymbol{\alpha_i}$ influences the mean and variance of each x_i. If $x_i = 0$, the corresponding term $x_i^{\alpha_i - 1}$ becomes problematic. Specifically for $\alpha_i > 1$, this term would tend towards 0 (i.e. $0^{\alpha_i - 1} = 0$), but for $\alpha_i \leq 1$, is undefined. To address this, Tsagris et al. (2018) propose Zero-Adjusted Dirichlet Regression (ZADR), which modifies the log-likelihood function to link the precision parameter ϕ to the covariates to accommodate zero values without replacing them. The standard Dirichlet log-likelihood is defined as:

$$\ell = n \log \Gamma(\phi) - \sum_{j=1}^{n} \sum_{i=1}^{D} \log \Gamma(\phi \boldsymbol{a}_i^*) + \sum_{j=1}^{n} \sum_{i=1}^{D} (\phi \boldsymbol{a}_i^* - 1) \log y_{ij}, \tag{2.3.2.4}$$

where $\Gamma(\cdot)$ is the Gamma function, ϕ is the precision parameter and $\boldsymbol{a}_i^*$ are the transformed regression parameters. The proposed modified log-likelihood is given as:

$$\ell = \sum_{b=1}^{B} \left[n_b \log \theta_b + n_b \log \Gamma(\phi_b) - \sum_{j \in S_b} \sum_{i=1}^{D_b} \log \Gamma(\phi_b \boldsymbol{a}_{bij}^*) + \sum_{j \in S_b} \sum_{i=1}^{D_b} (\phi_b \boldsymbol{a}_{bij}^* - 1) \log y_{bij} \right] \tag{2.3.2.5}$$

where B is the number of groups with different zero patterns, n_b is the number of observations in group b, θ_b is the probability of an observation belonging to group b, with $\sum_{b=1}^{B} \theta_b = 1$, ϕ_b is the precision parameter for group b, S_b is the set of observations belonging to group b, D_b is the number of non-zero components in group b, $\boldsymbol{a}_{bij}^*$ are the transformed regression parameters and $\Gamma(\cdot)$ is the Gamma function. Here, the model adjusts for the presence of zeros by explicitly incorporating them into the likelihood function, through creating partitions of the dataset into subpopulations based on which components contain zero values. This

means that, instead of assuming all compositions following a single Dirichlet distribution, the model assumes that different groups (defined by the zero patterns) follow different conditional Dirichlet distributions. This fitted model enables estimation of compositional values for new predictor variables. However, predicting exact zero values remains difficult due to the modification of the Dirichlet likelihood, which adjusts the probability of observing a zero value rather than predicting a zero value directly. Moreover, zero inflation is arbitrary - for example, replacing zeros with an arbitrary number (e.g. 31) in both the data and likelihood would not alter the results. This raises concerns about the interpretability of zero-inflated approaches, as the zero-generating process is treated separately from the continuous distribution, which may not always align with the analysis objectives.

2.4 Summary

Compositional data can take many different forms, each with unique characteristics, and none of the aforementioned approaches are suitable for all cases. Therefore, the next three chapters explore alternative methods tailored to specific applications of compositional data, where the data are non-standard – such as containing a large proportion of structural zeros, non-smooth compositional count time series and spatial compositional data. In each chapter, we present and critically assess novel tailored approaches for handling these types of data that remove the need for log-ratio transformations, which are unsuitable for the unique characteristics of each application.

Chapter 3

Bayesian Hierarchical Model for Compositional Data

We investigate Bayesian hierarchical approaches to modelling compositional data with a large proportion of structural zeros in the compositions and a multilevel hierarchical structure. The typical approach of applying a log-ratio transformation is unsuitable here, as log-ratios are undefined for zeros. Additionally, we will need to ensure that we suitably account for correlation arising from the hierarchical structure.

Here, we propose a flexible integrated clustering approach within a Bayesian hierarchical model framework for compositional data with structural zeros; we apply the methodology to a forensic elemental glass database that poses this challenge, where zeros are considered to be true zeros or values below some detection limit that have been rounded down to zero. We assess our approach and compare it to others in terms of use-type classification of glass items, using a five-fold cross-validation approach.

3.1 Introduction

As explored in Chapter 2, compositional data are a unique type of data that requires special consideration when conducting statistical analysis. Difficulties arise due to the constrained nature of the data, which prevents them from being treated as independent observations and necessitates a tailored approach to account for the underlying compositional structure.

Compositional data analysis has emerged as a powerful tool in forensic glass analysis as it can help account for the dependencies between the chemical elements in glass. Using the proposed methods to deal with compositional data can aid in forensic decision-making and quantifying the strength of the accumulated glass evidence found at a crime scene.

Previous approaches to modelling compositional data with structural zeros include model based approaches (Zadora et al. (2010b), Neocleous et al. (2011), Napier (2014)), which implement multivariate random effects models. These models split the data based on the presence and absence of the compositional elements prior to fitting hierarchical mixed-effect models. However, executing this split involves manual intervention, requiring some level of prior and expert knowledge to decide which compositional elements the data should be split by. This may not be practical for all settings where this approach is being utilised. Therefore, we aim to improve upon this approach to make it more functional and applicable.

The compositional data examined in this chapter consist of forensic elemental glass measurements obtained from an experimental setting. The data contain four fragments, each with three replicate measurements, from 320 glass items giving a total of 3,840 data points. Each of the glass items falls into one of five different use types: bulbs, car windows, headlamps, containers and building windows. Figure 3.1 illustrates the hierarchical structure of the data for a single glass item. The elements in the data under consideration are oxygen (O), sodium (Na), magnesium (Mg), aluminium (Al), silicon (Si), potassium (K), calcium (Ca) and iron (Fe). The differences in each compositional element can be explored using boxplots within Figure 3.2 of the item-level means grouped by use type. Understanding and modelling these compositional differences may improve the classification of new glass items.

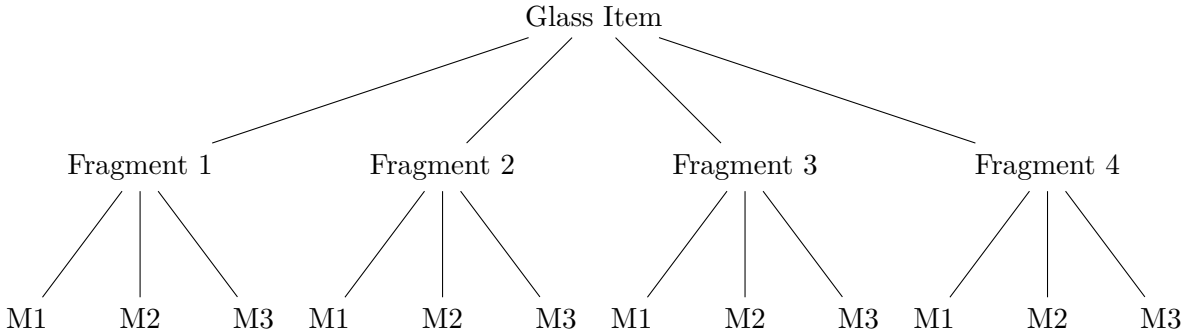


Figure 3.1: Example of the hierarchical structure for a single glass item in the forensic elemental glass dataset. Each glass item consists of four fragments and each fragment is measured three times (M1, M2 and M3).

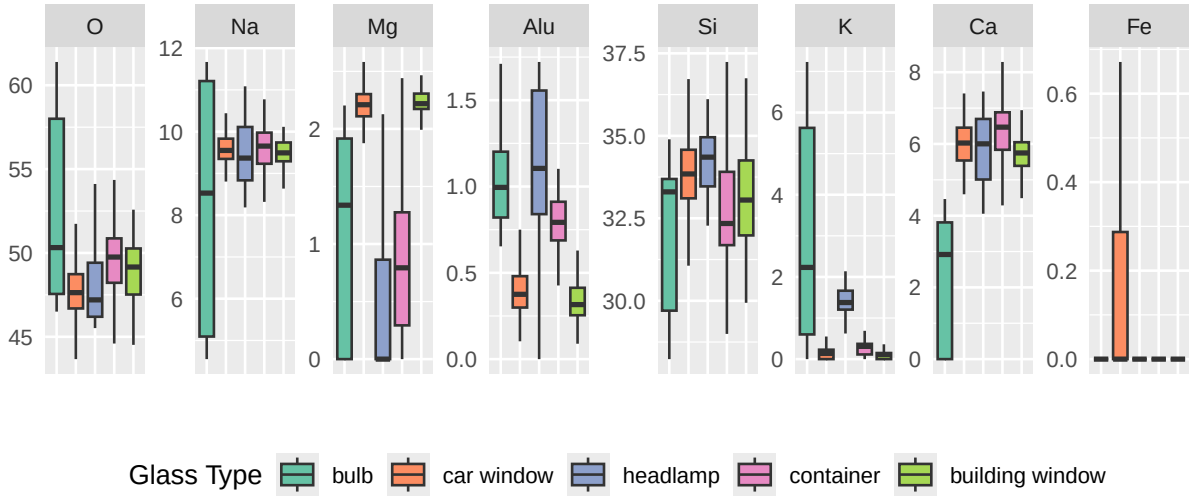


Figure 3.2: Boxplots of the item means for all 320 glass items in the forensic elemental glass data. The different coloured boxplots correspond to each of the use type groups: bulb, car window, headlamp, container and building window.

The aim of the work presented in this chapter is to develop Bayesian hierarchical models which build on previous work for compositional data with structural zeros, in four ways:

Aim 1: Develop a computationally efficient implementation of the framework, to allow for practical real-world analysis.

Aim 2: Investigate alternative data-driven approaches to splitting the data based on presence or absence of different elements, that require less strenuous expert input.

Aim 3: Explore a model-based approach to classification of new items, minimising computational demands after the hierarchical model has been fitted.

Aim 4: Allow for more general versions of the framework by implementing it using flexible Markov Chain Monte Carlo (MCMC) software.

In this chapter, we explore these aims through an application of compositional data analysis to a forensic elemental glass database. We will develop a computationally efficient implementation of a Bayesian hierarchical model in R using NIMBLE (Valpine et al., 2017), which is a flexible and efficient package for fitting a wide range of statistical models, particularly those that are computationally intensive and involve complex hierarchical structures (Aim

4). NIMBLE models are written in the BUGS language and then compiled automatically into C++, which allows for fast execution. Using NIMBLE results in efficient Markov chain Monte Carlo (MCMC) sampling, making the model suitable for real-world analysis (Aim 1). Furthermore, NIMBLE allows greater flexibility when applying a Bayesian hierarchical model, as it provides a wide range of statistical modelling capabilities, allowing users to specify and customise complex hierarchical structures (including spatiotemporal structures and penalised regression splines) and easily incorporating new functions, probability distributions, and sampling algorithms. We propose splitting the data by the presence and absence of the compositional elements through the use of clustering algorithms that automatically perform this task. This approach greatly alleviates the need for manual intervention, such as in Napier (2014) (Aim 2). We investigate hierarchical clustering and k-means clustering algorithms as two alternatives to the manual approach, which act as a baseline. We will compare all approaches in terms of performance in out-of-sample classification tasks. Additionally, we will propose a new integrated framework where clusters are included in the model as a latent quantity and where predictions of unknown forensic item types are generated simultaneously with model fitting (Aims 2 & 3).

The chapter is structured with Section 3.2 exploring previous approaches to modelling compositional data containing many structural zeros, focusing on the one which we will extend in this chapter. Section 3.3 gives an overview of the general methodology we adopt in this chapter. This section presents the proposed framework and the approaches we take to extend the current literature, including fitting our model to the full data comparing untransformed and square root transformed ratios, a manual approach to splitting the data, two pre-clustering algorithms to automate splitting the data and finally our proposed integrated clustering approach. Within Section 3.3.4 we describe how our proposed framework will be implemented in practice. Section 3.4 introduces the forensic glass database used to test the proposed approaches. In Sections 3.4.2 to 3.4.5, an overview of how we implement our framework to each of the approaches is outlined. Section 3.4.6 presents our experiment

to examine and compare each of the approaches outlined, through a cross-validation experiment to classify each new glass item into one of the five use types. This is evaluated in Section 3.4.7. Finally, in Section 3.5 we critically evaluate the work carried out in this chapter and discuss potential avenues for future research on this topic.

3.2 Background

Compositional data analysis has emerged as a valuable tool for the classification of forensic data such as the elemental compositions of glass. However, as explored in Chapter 2 Section 2.3.2 modelling compositional data becomes challenging when structural zeros are present, as these zeros are informative and cannot be ignored. Forensic elemental data often contain such zeros, representing the absence of a particular component. To address this issue, various modelling techniques have been proposed in the literature to handle structural zeros appropriately. These approaches serve as alternatives to log-ratio transformations, which are unsuitable in this context since the zeros carry meaningful information.

Previous modelling strategies for analysing compositional forensic glass data include Aitken et al. (2004), Zadora et al. (2010b) and Neocleous et al. (2011), all of which explore a frequentist approach to model log-ratio transformed compositional forensic data for classification purposes. This includes random effects models that incorporate two levels of variation: between-item and within-item. The between-item level variability is captured by a random effect associated with individual glass items, and the within-item variability by a random effect associated with individual fragments from the same glass item. Napier (2014) built on the previous work of Aitken et al. (2004) and Neocleous et al. (2011) to propose a new Bayesian approach to modelling forensic elemental glass data. This approach is outlined in more detail below in Section 3.2.1.

Meanwhile, Tsagris et al. (2016) apply the approach of using the α-transformation (Chapter 2, Section 2.2.2, Equation (2.2.2.4)) to a compositional forensic glass data set. This approach transforms the data using the α-transformation and then classifies the transformed data via regularised discriminant analysis and the k-nearest neighbours algorithm. Within this paper the value of α is varied, allowing a comparison between working in the standard Euclidean space, the compositional data Aitchison space or a value between these competing approaches. Tsagris et al. applied this to forensic elemental glass data that contain a large proportion of zero values. However, poor correct classification rates are presented for each of the approaches, with $\alpha \in [-1, 1]$ providing the best correct classification. As long as $\alpha > 0$, then the transformation is well-defined for any compositions containing zeros.

The same forensic glass data was examined in Tsagris et al. (2018) using a Dirichlet regression model. This approach modifies the Dirichlet distribution to accommodate a compositional response variable with zero values, eliminating the need for data modification. The log-likelihood of the Dirichlet distribution is modified to account for these zero values without the need to replace them. The modified log-likelihood is given in Chapter 2, Section 2.3.2, Equation (2.3.2.5). Here, the model adjusts for the presence of zeros by explicitly incorporating them into the likelihood function, through creating partitions of the dataset into subpopulations based on which components contain zero values. This means that instead of assuming all compositions following a single Dirichlet distribution, the model assumes that different groups (defined by the zero patterns) follow different conditional Dirichlet distributions. The fitted model can be used to estimate new values of the compositional elements. However, any zero value will be difficult to predict because the Dirichlet regression model inherently generates strictly positive values within the simplex and does not naturally produce exact zeros. A further limitation is that the model does not explicitly model when and where new zeros should occur. Additionally, a well-known drawback of the Dirichlet distribution is that it can only produce negative covariances/correlations between the compositional elements.

In contrast to the previous work outlined, other literature explores the use of a log-ratio transformation for forensic elemental glass data. Comas Cufí et al. (2016) fit a log-ratio transformation to the forensic glass data after replacing the zeros in the data with new values. This work centres around fitting a mixture of Normal and skew-Normal distributions to the log-ratio coordinates of the compositional data. However, imputing the zero values of the forensic glass data could remove the information they could have otherwise provided in the modelling procedure.

3.2.1 Napier approach

Building on the previous work of Aitken et al. (2004) and Neocleous et al. (2011), Napier (2014) accounts for the structural zeros in the data by modelling the presence and absence of the compositional elements. This work utilises the forensic elemental data outlined in Section 3.1. This relies on the manual separation of the elements that are absent, resulting in subsets of the data called "configurations". To illustrate this with a simple example, consider a case with five components (A, B, C, D, E). We could manually separate these based on the presence or absence of components B and D, as shown in Table 3.1. This results in four configurations: one where all components are present, one where both B and D are absent, and two where either B or D are absent.

Table 3.1: Simple example of configurations for five components: A, B, C, D, E. An entry of 1 represents the component is present and 0 represents an absent component.

Configuration	**Component**				
	A	B	C	D	E
1	1	1	1	1	1
2	1	0	1	1	1
3	1	1	1	0	1
4	1	0	1	0	1

Napier argued that analysing the configurations separately would reduce the impact that the large proportion of zeros may have on the data. Specifically, a high proportion of zeros can skew statistical models, so by modelling only the resulting subsets, the absent components are removed, leading to a more reliable analysis. Depending on how many compositional elements are present in the data, the number of possible configurations will vary. The maximum number of configurations possible is 2^p where p is the number of compositional elements in the data with values equal to zero. Not all of the potential presence-absence combinations will be exhibited in the data, so Napier solely modelled the configurations that are present in the data.

In this approach, it is possible to have cases where the configuration contains a very small number of data points. If this occurs, then Napier proposed combining these configurations with the configuration where all elements are present. This can allow the examination of specific elements of interest depending on the application. It is noted, the main limitation of this approach is that it requires some manual input by the user or even expert knowledge of which elements should be examined.

3.2.1.1 Bayesian hierarchical model

After constructing a final set of configurations, Napier (2014) proposed a Bayesian hierarchical mixed-effects model to be applied to each configuration separately. Hierarchical models can capture complex relationships in data where observations are organised in groups. Information can also be shared across different levels of the hierarchy, aiding the estimates where there may be little information for that group. This borrowing of strength can lead to more robust and flexible statistical analysis by reducing overfitting in small groups, improving parameter stability and allowing for pooling of information across groups, which is particularly valuable in settings with imbalanced or sparse data.

Napier applied the model to a forensic elemental glass data comprising of I glass items, each representing a distinct source of glass (e.g. from a bulb or window). Each glass item contains J fragments, which are pieces from each glass item I. For each fragment J, we have K replicate measurements taken to ensure accuracy. Each item belongs to one of T glass use types.

Then, Napier defines the compositional ratios of each element to oxygen for each glass measurement be denoted by $\boldsymbol{z}_{tijk}$, corresponding to the k-th replicate from the j-th fragment of the i-th glass item of use type t. The model for $\boldsymbol{z}_{tijk}$ is then assumed to be the sum of a fixed effect $\boldsymbol{\theta}_t$, random effects $\boldsymbol{b}_{ti}$ and $\boldsymbol{c}_{tij}$, and error term $\boldsymbol{\epsilon}_{tijk}$:

$$\boldsymbol{z}_{tijk} = \boldsymbol{\theta}_t + \boldsymbol{b}_{ti} + \mathbf{c}_{tij} + \boldsymbol{\epsilon}_{tijk}. \tag{3.2.1.1}$$

Here, $\boldsymbol{\theta}_t$ is a fixed effect term capturing the mean compositional ratio values for use type t; $\boldsymbol{b}_{ti}$ is a random effect capturing item-level variability; $\boldsymbol{c}_{tij}$ is a random effect capturing fragment-level variability; and $\boldsymbol{\epsilon}_{tijk}$ captures measurement error for individual pieces. Each of the random effects are assumed to have Multivariate Normal distributions, with covariance matrices $\boldsymbol{\Omega}_t^{-1}$, $\boldsymbol{\Psi}^{-1}$ and $\boldsymbol{\Lambda}^{-1}$:

$$\boldsymbol{b}_{ti} \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Omega}_{\boldsymbol{t}}^{-1}), \qquad \boldsymbol{c}_{tij} \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Psi}^{-1}), \qquad \boldsymbol{\epsilon}_{tijk} \overset{iid}{\sim} N_p(\mathbf{0}, \boldsymbol{\Lambda}^{-1}). \tag{3.2.1.2}$$

For a glass item $\boldsymbol{z}$ of use type $\boldsymbol{T}_z = t$ with JK measurements, the model (3.2.1.1) implies that the distribution of item $\boldsymbol{z}$ is

$$\boldsymbol{z} | \boldsymbol{T}_z = t, \boldsymbol{\xi} \sim N_{JKp}(\mathbf{1}_{JK} \otimes \boldsymbol{\theta}_t, \boldsymbol{\Sigma}_t), \tag{3.2.1.3}$$

where $\boldsymbol{\xi} = \{\boldsymbol{\theta}, \boldsymbol{\Omega}, \boldsymbol{\Psi}, \boldsymbol{\Lambda}\}$ collectively combines the model parameters. The covariance matrix $\boldsymbol{\Sigma}_t$ is given by

$$\Sigma_t = (\mathbf{1}_{JK}\mathbf{1}_{JK}^{'}) \otimes \boldsymbol{\Omega}_t^{-1} + [\mathbb{I}_J \otimes (\mathbf{1}_K \mathbf{1}_K^{'})] \otimes \boldsymbol{\Psi}^{-1} + \mathbb{I}_{JK} \otimes \boldsymbol{\Lambda}^{-1}, \tag{3.2.1.4}$$

where $\mathbf{1}_d$ is a column vector of d 1's and $\mathbb{I}_d$ is the $d \times d$ identity matrix.

Napier (2014) further assumes Multivariate Normal prior distributions for the fixed effects $\boldsymbol{\theta}_t$:

$$\boldsymbol{\theta}_t \overset{iid}{\sim} N_p(\mathbf{0}, \Phi^{-1}), \tag{3.2.1.5}$$

with $\boldsymbol{\theta}_t > \mathbf{0}$ for $t = 1, \ldots, T$. The non-negative restriction is imposed due to the compositional nature of the elements which must be greater than zero. The prior for the covariance matrix, Φ^{-1}, of the fixed effect $\boldsymbol{\theta}_t$ is fixed and set equal to $s \cdot \mathbb{I}_p$, where s is set equal to 1,000. This assumes a large variance a priori for $\boldsymbol{\theta}_t$, to establish a weakly informative prior such that the posterior modes of $\boldsymbol{\theta}_t$ will be close to the corresponding sample means from the data.

Next, Napier assumed conjugate Wishart hyperpriors for each of the random effect covariance matrices:

$$\boldsymbol{\Omega}_t \sim W_p(d_{1t}, A_t), \qquad \Psi \sim W_p(d_2, B), \qquad \Lambda \sim W_p(d_3, C), \tag{3.2.1.6}$$

where d_{1t}, d_2 and d_3 are the degrees of freedom and A_t, B and C are the precision matrices. Conjugate priors are used as they allow for efficient sampling from the posterior distributions due to the resulting distribution having the same form as the prior distribution. This allows for closed-form solutions that can reduce the computational burden and simplify the calculations in comparison to when non-conjugate priors are used. The degrees of freedom of the Wishart distribution need to be greater than the data dimension minus one, so weakly informative prior values for the degrees of freedom are set equal to p, $d_{1t} = d_2 = d_3 = p$ (DeGroot, 2005). These values are chosen to produce a weakly informative prior that provides a limited constraint on the parameters. The purpose of weakly informative priors is to balance between incorporating some prior information and allowing the data to have an influence on the posterior distribution. Using p as the degrees of freedom results in a minimally informative prior. As the degrees of freedom increase, the distribution becomes less spread out, leading to a more concentrated shape around the mean with reduced variance. A higher

degree of freedom results in a stronger prior that exerts more influence on the posterior. Meanwhile, increasing the scale matrix increases the expected value of the precision matrix. The scale matrices A_t, B and C are set equal to $(1/1000) \cdot \mathbb{I}_p$, so that the precision is expected to be small a priori. The Wishart priors thus suggest the random effect variance could be very large but the high prior uncertainty from setting the degrees of freedom equal to p means that posterior inference for these matrices is largely driven by the data.

This model can be directly applied to the raw compositional ratios z_{tijk}, but in practice, these ratios are unlikely to have normally distributed residuals due to their constraint on the simplex. One can instead apply the model to transformed values $z'_{tijk} = g(z_{tijk})$, where $g(\cdot)$ is some function, e.g. square root or log. Napier found that the square root of the compositional ratios improved the assumptions of normality and stability in the variability of the data more than a logarithmic transformation.

3.2.1.1.1 Classification

Once the model described in Section 3.2.1.1 (proposed by Napier (2014)) has been fitted, there is interest in classifying new glass items into the different glass types: bulbs, car windows, headlamps, containers and building windows. This task is motivated by the forensic setting where a glass fragment may be found and it is useful to find which glass item the fragment may have come from. The full classification process is outlined within Napier (2014) and summarised below.

Denote $\boldsymbol{y}$ as the transformed elemental composition of a newly observed glass item, with its unknown use type denoted by $\mathscr{T}_y$. The elemental configuration of $\boldsymbol{y}$ is $\mathscr{C}_y = m$, which is known if $\boldsymbol{y}$ is conditioned upon. The use-type probability for a newly observed glass item $\boldsymbol{y}$:

$$p(\mathscr{T}_y = t|\boldsymbol{y}, D) \propto p(\mathscr{T}_y = t)\frac{\alpha_{tm} + N_{tm}}{\sum_{r=1}^{M}(\alpha_{tr} + N_{tr})}E_{\xi|D_m}[p(\boldsymbol{y}|\mathscr{T}_y = t, \mathscr{C}_y = m, \xi_m)], \qquad (3.2.1.7)$$

where α_{tm} is the shape parameter vector of an assumed Dirichlet prior distribution for the classification probabilities, set to 0.1 for all t, resulting in a weakly informative prior which allows the classification probabilities to be driven by the data. N_{tm} defines the number of each glass type within each configuration. ξ_m is the posterior distribution corresponding to $\boldsymbol{y}$ is $\mathscr{C}_y = m$. $\boldsymbol{D}$ is the reference database which is not informative about the glass use type of a newly observed glass item, since the number of items of each type does not reflect the prevalence of these use types in a real-world setting. $E_{\xi|D_m}$ denotes the expectation with respect to the posterior distribution of ξ_m. This is estimated by taking the mean of the densities of $p(\boldsymbol{y}|\mathscr{T}_y = t, \mathscr{C}_y = m, \xi_m)$ with the ξ_m given by the MCMC samples. Napier found that the choice of the value of α did not seem to have that much of an effect on the classification results for values between 0.1 and 0.5, indicating that the data itself provides strong information for classification - so in this work α_{tm} is set to 0.1. In this case, the posterior classification probabilities are dominated by the likelihood. If we had a smaller dataset, we might want to incorporate a more informative prior so that, a priori, we have more knowledge about the type probabilities.

The glass use-type probabilities $p(\boldsymbol{T}_y = t|\mathscr{C}_y = m, \boldsymbol{D})$ from Equation (3.2.1.7) can be computed for the configurations of $\boldsymbol{y}$. It could occur that within some of the configurations there could be some absent glass use types, therefore, the use-type probability for that type would be much lower, meaning that it is more unlikely that the glass item will be classified into that glass use type.

The real-world practicability of Napier's approach is limited in the first instance by the manual aspect of splitting the data by the presence and absence of the compositional elements. This could require prior or expert knowledge to carry out, which could be avoided if we had a way of automating the split (Aim 2). Moreover, this model used hand-coded MCMC algorithms in R that were not optimally efficient (e.g. compared to compiled C++ code) and could be difficult to adapt. This could be improved by utilising a flexible MCMC software package which would reduce the computational burden and make the method more practical for real-life applications, in particular where frequent database updates may impact the model (Aim 4). Additionally, while the classification task occurs without requiring the model to be rerun, setting up the classification use-type probabilities for each glass item is a time-consuming process. An improvement to the method would be to introduce a new step in the modelling that directly predicts the glass use type, making the implementation more efficient and user-friendly (Aims 1 & 3).

3.3 Methodology

First, we detail a simple generalised version of the model given in Section 3.2.1 for a vector of measured elemental compositions, $\boldsymbol{z}_{tijk}$, with nested groupings $j \subset i \subset t$ (e.g, fragment, item, type), and with k indexing repeated measurements. We introduce an additional index cl_i, which denotes the configuration membership of the glass item. In Napier (2014), all parameters are assumed to be independent across the configurations, and indeed, the models are fitted independently. However, in the proposed model, this assumption will be relaxed to allow for dependencies between configurations.

$$\begin{aligned} \boldsymbol{z}_{tijk} \sim &\quad N(\boldsymbol{\mu}_{tij}, \boldsymbol{\Sigma}(cl_i)), \\ \boldsymbol{\mu}_{tij} \mid \boldsymbol{\beta}_{tij}, \boldsymbol{c}_{tij} = &\quad \boldsymbol{X}_{tij}\boldsymbol{\beta}_{tij} + \boldsymbol{c}_{tij}. \end{aligned} \tag{3.3.0.1}$$

Here, $\boldsymbol{X}_{tij}$ represents generic group-level covariates. These could be simple linear/factor effects or more complex terms such as penalised regression spline basis functions, with smoothing penalisation handled within the priors for $\boldsymbol{\beta}_{tij}$. Then, $\boldsymbol{c}_{tij}$ is a structured random effect for object tij. This could be an additive combination of e.g. type-, item-, and fragment-level random effects as in Equation (3.2.1.1).

The final assumption in this general framework is that the prior models for $\boldsymbol{\beta}_{tij}$ and $\boldsymbol{c}_{tij}$ may differ fully (like in Napier (2014)) or partially (e.g. through some hierarchical structure, as proposed later) depending on the configuration/grouping index cl_i.

3.3.1 Proposed framework

First, we propose an alternative hierarchical version of the random effect structure in Equation (3.2.1.1) for more efficient MCMC sampling, motivated by the forensic glass data. Recall that Napier (2014) captured structured variability in elemental compositions at type/item/fragment level through an additive combination of random effects:

$$\boldsymbol{z}_{tijk} = \boldsymbol{\theta}_t + \boldsymbol{b}_{ti} + \mathbf{c}_{tij} + \boldsymbol{\epsilon}_{tijk}. \tag{3.3.1.1}$$

However, this additive structure can induce substantial correlation between the random effect terms during MCMC sampling, due to the shared modelling of variability across multiple hierarchical levels. This correlation can slow convergence, making it more challenging for the sampler to explore efficiently the posterior distribution. In contrast, a nested formulation of hierarchical random effects can be more efficient when the data are "naturally nested" (Schielzeth et al., 2013), such as when multiple observations exist per group. In our case, we have multiple measurements for each glass item. Using a nested structure can account for the dependencies within the data and provide more accurate estimates of the effects at each level (Gelman, 2007). Furthermore, in nested data structures, the interaction

variance is combined with the main effect variance of the nested factor. This means that in cases where separating interaction effects is not a priority, a nested formulation simplifies the model and interpretation (Schielzeth et al., 2013). Therefore, structuring our model in a nested way can simplify the model and improve the MCMC mixing, leading to more efficient and faster convergence.

We therefore propose re-expressing these terms hierarchically rather than additively, as follows:

$$\begin{aligned} \boldsymbol{c}_{tij} &\sim N(\boldsymbol{b}_{ti}, \boldsymbol{\Psi}^{-1}(cl_i)), \\ \boldsymbol{b}_{ti} &\sim N(\boldsymbol{\theta}_t(cl_i), \boldsymbol{\Omega}_{\boldsymbol{t}}^{-1}(cl_i)), \\ \boldsymbol{\theta}_t(cl_i) &\sim N(\mathbf{0}, \boldsymbol{\Phi}^{-1}(cl_i)). \end{aligned} \tag{3.3.1.2}$$

It can be shown through relabelling that this is exactly the same parametric model.

We can treat model parameters ($\boldsymbol{\theta}_t(cl_i)$, $\boldsymbol{b}_{ti}$, $\boldsymbol{c}_{tij}$ etc.) as random quantities within the Bayesian framework, and obtain inference based on posterior distributions. Here, Bayesian inference has multiple advantages over alternatives:

- It can easily handle the various Multivariate Normal structures in this model.
- Models can be extended and/or adapted without changing the inferential framework.
- It enables thorough model checking through posterior predictive checking.
- It provides full predictive inference for new items, allowing future classification.

Given the complexity of the model (Equation (3.2.1.3)) and the high-dimension of the set of model effects, direct posterior inference is challenging. Instead, we can use MCMC methods (Appendix B) to simulate samples from the joint posterior distribution.

Next, instead of assuming that the type $\boldsymbol{t}_i$ of item i is a known constant, we treat it as a random quantity that can be unknown. Here, we assume that the glass use type each item belongs to can fall into one of t use type categories, with probabilities $\boldsymbol{\delta}_t(cl_i)$. We thus assume a categorical distribution for t_i:

$$\boldsymbol{t}_i \sim \text{Categorical}(\boldsymbol{\delta}_t(cl_i)). \tag{3.3.1.3}$$

The categorical distribution is a multi-class generalisation of the Bernoulli distribution, which would only allow for two possible outcomes.

Recall from Section 3.2.1.1.1, Napier assumed a Dirichlet prior for the classification probabilities, reflecting a belief that, before observing any data, each glass use type is equally likely. Specifically, the prior was set to be weakly informative, ensuring that the classification remains primarily data-driven. Following this, we assume a Dirichlet prior for the type probabilities $\boldsymbol{\delta}_t(cl_i)$:

$$\boldsymbol{\delta}_t(cl_i) \sim \text{Dirichlet}(\boldsymbol{\tau}_t), \tag{3.3.1.4}$$

where $\boldsymbol{\tau}_t = 1/t$ for all glass use types t, placing equal weight on each possible glass use type, reflecting a belief that a priori, each type is equally likely for each glass use type t. This results in the same Dirichlet prior assumption as in the previous work by Napier. The parameter $\boldsymbol{\delta}_t(cl_i)$ defines the probability of each glass use type t for each glass item i, allowing the probability of each glass use type to vary across the different configurations. Since the presence or absence of elements defines each configuration cl_i, the probability of glass use types may vary across configurations. Thus, incorporating configuration-dependent probabilities can improve the prediction of glass use types.

Comparing this with a previous approach outlined in Section 3.2.1.1.1, we observe that both processes follow a similar approach in determining the glass use type for each item. In the previous approach, the probability of each glass use type was determined for each configuration through Equation (3.2.1.7) and the glass use type is classified to the highest

probability. In both approaches, the Dirichlet distribution is assumed as a prior for the probabilities of each glass use type. Additionally, both approaches incorporate configuration-specific probabilities, accounting for the fact that some glass use types may appear more frequently in certain configurations than in others. However, our proposed framework offers advantages in that it streamlines the steps for the classification of each item and allows measurements from items of unknown type to be included in the model. By setting t_i as "NA" in the data, the MCMC algorithm will automatically sample it as an unknown quantity.

This framework can be used in conjunction with the manual strategy for splitting the data into configurations so that cl_i are known constants in advance of modelling. We will refer to this as the "manual configurations" approach.

In the following subsections, we detail alternatives to splitting the data into configurations manually, including automated clustering algorithms and ultimately an integrated Bayesian clustering approach.

3.3.2 Pre-clustering approach

Recall that Napier (2014) proposed splitting the data up by the presence and absence of the compositional elements to reduce the impact the zero values have on the model. This is limited by the manual and highly subjective nature of the separation process. It potentially requires prior knowledge of which elements to use to split the data, and the manual effort required may be impractical in operational settings.

Here we propose alternative approaches that automate the splitting procedure. We will investigate the use of standard clustering methods, in place of this manual separation, potentially to make the method more generally applicable. We will focus on two very common clustering methods: hierarchical clustering and k-means clustering. However, other clustering methods could also be implemented.

As our aim is to cluster the data based on the presence and absence of the compositional elements, we conducted our pre-clustering approaches on an indicator matrix of presence and absence of the mean across all measurements for each glass item. We followed this approach so that the clusters might potentially reflect similar information captured by the configurations based on the presence and absence of the elements.

3.3.2.1 Hierarchical clustering

Hierarchical clustering is an algorithm that progressively groups or divides objects based on a measure of similarity, resulting in a hierarchy of clusters (Nielsen, 2016). Hierarchical clustering has two main approaches: agglomerative, a "bottom-up" method where each observation starts as its own cluster and clusters are progressively merged, and divisive, a "top-down" method where all observations start in a single cluster and are recursively split. The choice between these methods depends on the characteristics of the data and the specific problem, as explored in Roux (2018). This research employs the agglomerative approach, explored in Murtagh et al. (2012), due to its computational feasibility and ease of use. In contrast, divisive clustering is often more computationally expensive, as it requires solving complex optimisation problems at each step.

The steps of the agglomerative algorithm (Day et al., 1984) are:

1. Each observation is considered to be its own cluster.
2. The distances between clusters are computed.
3. The two clusters with the smallest pairwise distance are combined into a single cluster. The distances between all clusters are then updated.
4. Steps 2 and 3 are repeated until there is only a single cluster containing all observations.

In hierarchical clustering, the distance between points is a crucial component for grouping similar points together. The most commonly used measure to compute the distance is the Euclidean distance. This calculates the distance between two points in a multi-dimensional space. However, if there are solely binary values or an indicator matrix being used to cluster upon, the Euclidean distance does not make sense to apply here. Instead, in these cases the binary distance can be used. This calculates the dissimilarity between the binary values by considering the presence or absence of elements. The resulting distance values range from 0 to 1, with 0 indicating complete similarity and 1 indicating complete dissimilarity. Here, we decided to use the binary distance as we aim to cluster the presence and absence of the elements, expressed as an indicator matrix. The indicator matrix is a $I \times p$ matrix containing 0 or 1s, where p is the number of elements.

After the distance matrix is computed, a linkage method is applied to measure the dissimilarity between clusters. The linkage method calculates the distances between all objects. Different linkage methods can lead to different clustering results, as they capture different aspects of the relationships between clusters. The four most common linkage methods are complete, single, average and Ward linkage. This research adopts the Ward linkage, introduced in Ward Jr (1963), which merges the two clusters that minimise the total within-cluster variance. The minimum between-cluster distance is computed, and the pair with the smallest distance is merged, ensuring that newly formed clusters remain compact and homogeneous (Sharma et al., 2019).

Ward's method offers several advantages over other hierarchical clustering techniques, as highlighted in Sharma et al. (2019). Unlike single linkage, which often results in elongated clusters by linking distant points, Ward's method maintains cluster compactness by minimising the increase in variance. Compared to complete linkage, which can struggle when clusters overlap, Ward's approach ensures that within-cluster variance remains low, leading to better-defined and more evenly sized clusters. While all methods perform well when clusters are clearly separated, Ward's linkage consistently produces more distinct and balanced groupings, making it particularly effective for identifying meaningful patterns. These properties make Ward's linkage especially suitable for this work, as it enables the discovery of patterns in the presence and absence of compositional elements, while avoiding clusters of largely dissimilar size, which could complicate statistical analysis.

3.3.2.2 k-means clustering

k-means clustering is a clustering algorithm that groups points together into k clusters in which each observation belongs to the cluster with the nearest cluster centre or centroid (Jain, 2010). This algorithm aims to minimise the within-cluster sum-of-squares. Unlike hierarchical clustering, outlined above in Section 3.3.2.1, this algorithm requires the number of clusters, k, to be specified prior to the clustering being fitted. Here, we applied k-means to the same $I \times p$ presence and absence indicator matrix that was used for the hierarchical clustering.

The iterative steps carried out when using the k-means clustering method (Hartigan et al., 1979) are:

1. Choose a k value. This is used as the initial set of k centroids.
2. Assign each data point to the cluster with the nearest centroid.

3. Determine the new centroids of the k clusters, by computing the mean of the cluster members.
4. Repeat step 3 until there is no change in the criterion after an iteration.

The k-means clustering algorithm is often run multiple times with different values of k to identify the optimal number of clusters. To assist in choosing k in Step 1, methods such as the elbow method or knowledge from the application or data can be used.

The k-means clustering algorithm is not guaranteed to find the global minimum, as it can get stuck in a local minimum due to its dependence on the initial choice of centroids. To mitigate this, the algorithm is typically run multiple times with different initialisations. This increases the likelihood of finding a better local minimum and, in some cases, the global minimum, but there is no guarantee of achieving the optimal solution (Ahmed et al., 2020).

3.3.3 Bayesian integrated clustering approach

As discussed in Section 3.3.2, although the pre-clustering approach automates the manual splitting of the data, a key limitation is that cluster labels are fixed before modelling, since a clustering algorithm is applied in advance.

Here, we propose a new integrated Bayesian clustering model for the compositional data, where we explore a latent variable for the cluster of glass item i within the model. By incorporating this into our model, we aim to reduce the burden on users of the approach to make important subjective decisions around splitting the data based on the presence and absence of the compositional elements, while also including clustering uncertainty into the predictions. Meanwhile, we might also achieve a better clustering structure, as the clustering will be more data-driven determined by its effect on the data likelihood and random effects

models, not solely with respect to presence or absence in an indicator matrix. This addresses the original Aims 1 and 2 from Section 3.1 as our proposed framework clusters and classifies within the model based on the input data. However, this integrated clustering approach requires the number of clusters to be defined prior to the modelling.

To achieve this, we treat the cluster $cl_i \in (1,\ldots,N_{CL})$ that item i belongs to as an unknown latent categorical variable:

$$cl_i \sim \text{Categorical}(\boldsymbol{\zeta}), \tag{3.3.3.1}$$

where the probability of belonging to cluster cl_i is given by $\boldsymbol{\zeta}$. A Dirichlet prior is placed upon the vector of probabilities $\boldsymbol{\zeta}$, i.e.

$$\boldsymbol{\zeta} \sim \text{Dirichlet}(\boldsymbol{\iota}), \tag{3.3.3.2}$$

where ι_{cl} is set equal to $1/N_{CL}$ for all clusters $cl = (1,\ldots,N_{CL})$. This indicates that we believe each cluster is equally likely for each glass item a priori. Assuming a Dirichlet prior on the probability of belonging to each cluster means ζ_{cl} defines the prior probability of any item belonging to each cluster cl.

Then, in addition to the Multivariate-Normal model for z_{tijk}, we include a Bernoulli model for the presence and absence of compositional elements:

$$u_{ie} \sim \text{Bernoulli}(q_{cl_i,e}), \tag{3.3.3.3}$$

where e is the index for the compositional element, for $e \in (1,\ldots,p)$. Here, $q_{cl_i,e}$ represents the probability of the presence and absence for each element depending on what cluster the item is in (cl_i). We assume a Uniform prior on $q_{cl,e}$, i.e.

$$q_{cl,e} \sim \text{Uniform}(0,1), \tag{3.3.3.4}$$

independently for each cluster and element. Assuming this Uniform prior for the probability of presence or absence of each element means that each outcome (either present or absent) is equally likely within each cluster a priori. This allows for flexibility in updating these probabilities based on observed data without imposing strong prior assumptions. This is a non-informative prior as it does not favour any value within the range $[0,1]$.

The posterior probability for the cluster cl_i that item i belongs to is given by

$$p(cl_i|z_{tijk},u_{ie}) \propto p(z_{tijk}|cl_i)p(u_{ie}|cl_i)p(cl_i), \tag{3.3.3.5}$$

where $p(cl_i|z_{tijk},u_{ie})$ is the posterior probability of item i belonging to cluster cl_i given the observed compositions. $p(z_{tijk}|cl_i)$ is the likelihood of the compositions given the cluster cl_i, $p(u_{ie}|cl_i)$ is the Bernoulli likelihood for the presence and absence of elements given the cluster cl_i and $p(cl_i)$ is the prior probability for each cluster assignment.

We could, if desired, use more informative priors to incorporate specific prior knowledge or beliefs about the parameters. For example, the Beta distribution could be used as a more flexible prior for $q_{cl,e}$, where the shape parameters could be chosen to reflect known tendencies in the compositions, such as higher probabilities for certain elements being present.

The complete definition of the proposed model is as follows:

$$
\begin{aligned}
\boldsymbol{z}_{tijk} &\sim N(\boldsymbol{c}_{tij}, \boldsymbol{\Sigma}(cl)),\\
\boldsymbol{c}_{tij} &\sim N(\boldsymbol{b}_{ti}, \boldsymbol{\Psi}^{-1}(cl)),\\
\boldsymbol{b}_{ti} &\sim N(\boldsymbol{\theta}_t(cl), \boldsymbol{\Omega}_t^{-1}(cl)),\\
\boldsymbol{\theta}_t(cl) &\sim N(\mathbf{0}, \boldsymbol{\phi}^{-1}(cl)),\\
\boldsymbol{\Sigma}(cl) &\sim W(p, \xi(cl)),\\
\boldsymbol{\Psi}(cl) &\sim W(p, B(cl)),\\
\boldsymbol{\Omega}(cl) &\sim W(p, A(cl)),\\
u_{ie} &\sim \text{Bernoulli}(q_{cl_i,e}),\\
q_{cl,e} &\sim \text{Uniform}(0,1),\\
t_i &\sim \text{Categorical}(\boldsymbol{\delta}_t(cl_i)),\\
\boldsymbol{\delta}_t(cl) &\sim \text{Dirichlet}(\boldsymbol{\tau}),\\
cl_i &\sim \text{Categorical}(\boldsymbol{\zeta}),\\
\boldsymbol{\zeta} &\sim \text{Dirichlet}(\boldsymbol{\iota}),
\end{aligned}
$$

where $\xi(cl) = A(cl) = B(cl) = I_p/1000$, $\boldsymbol{\phi}^{-1}(cl) = 1000 \cdot I_p$, $\tau_t = 1/t$ and $\iota_{cl} = 1/N_{CL}$.

In this model all parameters and prior distributions pertaining to the clusters cl_i are equal across the different cluster labels, leading to concerns surrounding "label-switching". This occurs when the labels of the clusters can be exchanged without altering the overall likelihood of the model parameters, leading to ambiguity in identifying which parameters correspond to which cluster (Stephens, 2000). Running the model across multiple chains and training datasets can yield different cluster structures or labels, preventing a definitive cluster structure due to the symmetry in the model's clustering structure. Since prior distributions treat the clusters equivalently, labels such as "Cluster 1" or "Cluster 2" are arbitrary and the model has no inherent preference for one label over another unless constraints are imposed. However, this is not a concern in our case as our goal is to use clusters for classifying new

glass items rather than interpreting them. If addressing label-switching is a requirement within a different application, we can add and tailor constraints to fit the specific needs. For example, one possible approach would be to order the clusters based on the mean of the presence and absence of the compositional elements, i.e. mean(u_{ie}).

If the above weakly informative prior distributions are used and no label-switching constraints are included in modelling, the only necessary choice for the user in relation to clustering is choosing the number of clusters. Strategies for choosing the optimal number of clusters is not explored here but some potential avenues are discussed in Section 3.5.

3.3.4 Implementation

A successful MCMC implementation is dependent on choosing appropriate sampling methods. Coding our MCMC algorithm manually, e.g. using the `R` (R Core Team, 2021) statistical programming language as in Napier (2014), gives the greatest freedom in choosing any combination of MCMC sampling algorithms, e.g. Metropolis-Hastings (Hastings, 1970) or Gibbs sampling (Casella et al., 1992). However, doing so is cumbersome in terms of the effort and expertise required, especially when making major changes to the model or adapting it for new applications. The computation may also be prohibitively slow for practical use if the code is not optimised or relies on inefficient software architectures.

A more accessible option is to use a software package that automates much of the mechanics such as assigning samplers to model parameters and running the MCMC algorithm. However, the most established packages, e.g. WinBUGS (Lunn et al., 2009) or JAGS (Plummer, 2003), offer no flexibility in choosing alternative sampling methods if the default samplers are not working well in a given situation.

Here, we aim to implement the model in a flexible way, both in terms of freedom to choose different sampling algorithms and modifications to the model, while also achieving practical computation times compared to previous work. To achieve this, we base our implementation on the *NIMBLE* package (Valpine et al., 2017) which allows for flexible implementation of Bayesian models using MCMC, among other algorithms. NIMBLE models are written in the BUGS language, like JAGS (Plummer, 2003), and then compiled automatically into C++ (Stroustrup, 1986) for fast execution. By default, NIMBLE uses a combination of Metropolis-Hastings random walk sampling algorithms and multivariate random walks, or conjugate relationships where possible, but the user can choose any combination of samplers for different model parameters. An extensive list of samplers are pre-included in the package, but the user can straightforwardly add their own sampling methods to the algorithm, as part of a general feature of NIMBLE that allows user-defined functions and probability distributions to be added. In Section 3.4.7 we will show the computational efficiency of our models that is achieved using NIMBLE, while only relying on the default random walk, blocked random walk and slice samplers. Moreover, modifications to the model can be made far more easily than in manual MCMC implementations.

The `R NIMBLE` model code to produce the Bayesian hierarchical model for the compositional data is given in Listing 1. The additional code required to predict the clusters for our proposed integrated clustering approach is given in Listing 2 - added to Listing 1 within the cluster loop in line 4. To ensure that the clusters are still based on the presence and absence of the compositional elements we also add the code given in Listing 3 within the item loop in line 25.

```
model_code <- nimbleCode({

    ## LOOP OVER NUMBER OF CLUSTERS ##
    for (cl in 1:N_CL) {

      # TYPE PROBABILITY #
      delta[1:t, cl] ~ ddirch(tau[1:t])

      # PRIORS #
      xi[1:p, 1:p, cl] <- diag(p)/1000
      Psi[1:p, 1:p, cl] ~ dwish(df = p, B[1:p, 1:p, cl])
      Psi_inv[1:p, 1:p, cl] <- inverse(Psi[1:p, 1:p, cl])
      Sigma[1:p, 1:p, cl] ~ dwish(df = p, xi[1:p, 1:p, cl])
      Sigma_inv[1:p, 1:p, cl] <- inverse(Sigma[1:p, 1:p, cl])

      ## LOOP OVER NUMBER OF GLASS USE TYPES ##
      for (k in 1:t){
        theta[1:p, k, cl] ~ dmnorm(mean = mean_zero[1:p],
                                   cov = inv_phi[1:p, 1:p, cl])
        Omega[1:p, 1:p, k, cl] ~ dwish(df = p, A[1:p, 1:p, k, cl])
        Omega_inv[1:p, 1:p, k, cl] <- inverse(Omega[1:p, 1:p, k, cl])
      }

     ## LOOP OVER NUMBER OF ITEMS ##
      for(i in 1:I){
        b[i, 1:p] ~ dmnorm(theta[1:p, t[i], cl],
                           Omega[1:p, 1:p, t[i], cl])
        ## LOOP OVER NUMBER OF PIECES ##
        for(j in 1:J){
          c[i, j, 1:p] ~ dmnorm(b[i, 1:p], Psi[1:p, 1:p, cl])
        }}}

    ## COMPUTE ITEM TYPE ##
    for (i in 1:I){
      t[i] ~ dcat(delta[1:t, cl[i]])
    }

    ## MODEL ELEMENTAL COMPOSITIONS ##
    for(i in 1:N){
      z[i, 1:p] ~ dmnorm(mean = c[item[i], piece[i], 1:p],
                         Sigma[1:p, 1:p, cl[i]])
    }
  })
```

Listing 1: Custom `R NIMBLE` model code to implement the Bayesian hierarchical model outlined in Section 3.3 where the configurations or clusters are provided to the model.

```
1     ## LOOP OVER NUMBER OF CLUSTERS ##
2     for (cl in 1:N_CL) {
3
4         ## CLUSTER PROBABILITY ##
5         zeta[1:N_CL] ~ ddirch(iota[1:N_CL])
6
7         ## COMPUTE ITEM CLUSTER ##
8         for (i in 1:I) {
9             cl[i] ~ dcat(zeta[1:N_CL])
10        }}
11
12        for (cl in 1:N_CL) {
13            for (e in 1:p){
14                ## PRIOR PROBABILITY PRESENCE / ABSENCE ##
15                q[cl, e] ~ dunif(0,1)
16    }}
```

Listing 2: Additional `R NIMBLE` code to implement the Bayesian hierarchical model outlined in Section 3.3.3 where the clusters are modelled within the NIMBLE model. This code is added to Listing 1 within the cluster loop at line 4.

```
1     for (i in 1:I) {
2         for(e in 1:p) {
3             ## MODEL ITEM'S PRESENCE / ABSENCE OF ELEMENTS ##
4             u[i, e] ~ dbern(q[cl[i], e])
5     }}
```

Listing 3: Additional `R NIMBLE` code to implement the Bayesian hierarchical model outlined in Section 3.3.3 where the clusters are modelled within the NIMBLE model where we model the presence and absence of each item's elements. This code is added to Listing 1 within the loop across items at line 25.

All computations were carried out on an Ubuntu Linux desktop computer with an Intel Core i9-13900K processor with 24 physical cores (32 logical cores) with 128GB system memory.

3.4 Application to Forensic Glass

In this section, we assess models using no splitting/clustering of the data, as well as using the manual configurations, pre-clustering and Bayesian integrated clustering approaches with respect to their application to forensic glass data, detailed in the next subsection, as a tool for predicting (classifying) the type of new glass items.

In Section 3.4.2, we will briefly compare posterior inference from applying the Bayesian hierarchical model fitted to the full data with square root transformed ratios to the same data without the transformation, and discuss implications for out-of-sample performance. In the subsections that follow, we will then present and interpret results from each different approach to separating/clustering the data. To summarise, this section will present the application of the following models to the forensic glass data:

1. No splitting: untransformed ratios
2. No splitting: square root transformed ratios
3. Manual configurations approach
4. Pre-clustering approach: hierarchical clustering
5. Pre-clustering approach: k-means clustering
6. Bayesian integrated clustering approach

All these models will ultimately be compared based on out-of-sample type classification performance, through a five-fold cross-validation experiment. We explain the design of this experiment and present results in Section 3.4.6.

3.4.1 Data set

This chapter examines forensic elemental glass data that were introduced in Section 3.1 and previously analysed in Napier (2014). Recall, the data contain four fragments, each with three replicate measurements, from 320 glass items giving a total of 3,840 data points. Each of the glass items falls into one of five different use types: bulbs, car windows, headlamps, containers and building windows. The percentage weights of each fragment are compositional, non-negative and sum to 100%. The number of elements in a fragment's composition can be denoted as D with the percentage weights $\boldsymbol{w} = (w_1, \ldots, w_D)$, with $w_d \geq 0$ and $\sum_{d=1}^{D} w_d = 100$. In this work, to remove the sum constraint imposed by the compositional nature of the data the composition is transformed into a $(D-1)$ dimensional vector of the ratios of the $(D-1)$ elements to the D^{th} element. Dividing by one of the compositional elements helps to remove the scale dependency and expresses the relative proportions in a meaningful way. The transformed vector is defined as

$$\boldsymbol{w}^* = \left(\frac{w_1}{w_D}, \ldots, \frac{w_{D-1}}{w_D} \right), \tag{3.4.1.1}$$

where oxygen is chosen to be the divisor W_D. Here, oxygen is chosen as the divisor as it is present in all the glass items, allowing the division in the ratio.

The number of zeros in the data varies by element, with oxygen (O), sodium (Na) and silicon (Si) containing no zero measurements, but the element iron (Fe) has mostly zero values with 79% of measurements being equal to 0. Table 3.2 presents the frequency of these zero values within each element and their respective percentage of zeros.

Table 3.2: Frequency of zeros present in the forensic glass data by compositional element.

Element	O	Na	Si	Ca	Al	Mg	K	Fe
Frequency	0	0	0	108	205	265	1168	3036
Percentage %	0	0	0	2.8	5.3	6.9	30.4	79.1

We can visualise this within the top row of Figure 3.3 which shows boxplots of each glass use type for each compositional element for the untransformed compositional ratios to oxygen. A large proportion of zeros can be identified from the plot with the median of the boxplots close to 0, seen within iron (Fe) and potassium (K), where these boxplots exhibit a small spread indicating that most values are around 0. If there are a large number of zeros in the data, this could skew or distort the mean and variance leading to inaccurate conclusions if not properly accounted for. We could transform the compositional ratios to reduce the effect of the zero values, allowing for a clearer visualisation of the distributions among glass types. Furthermore, applying a transformation could stabilise the variance across the different elements. The bottom row of Figure 3.3 displays the square root transformed compositional ratios to oxygen. Improvements can be detected through the shape of the boxplots, e.g. the shape and spread of values is more apparent, in particular for potassium (K), showing clearer variations and distinguishable features among glass types.

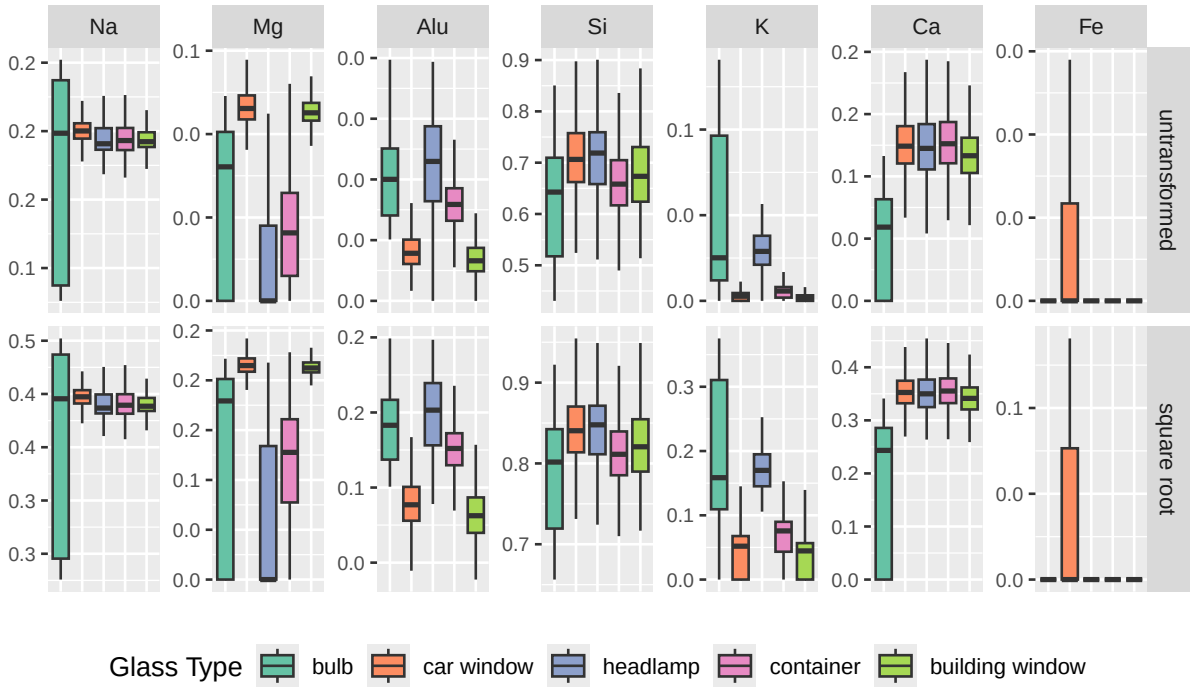


Figure 3.3: Boxplots of the untransformed (top row) and square root transformed (bottom row) compositional ratios, Equation (3.4.1.1), with oxygen as the divisor for all the glass item means. The different coloured boxplots correspond to each of the use type groups: bulb, car window, headlamp, container and building window.

However, even when applying square root transformation to the ratios, there are still correlation patterns between elements. For instance, Figure 3.4 presents a scatterplot of silicon (Si) against calcium (Ca) where a strong positive correlation for all glass use types can be identified.

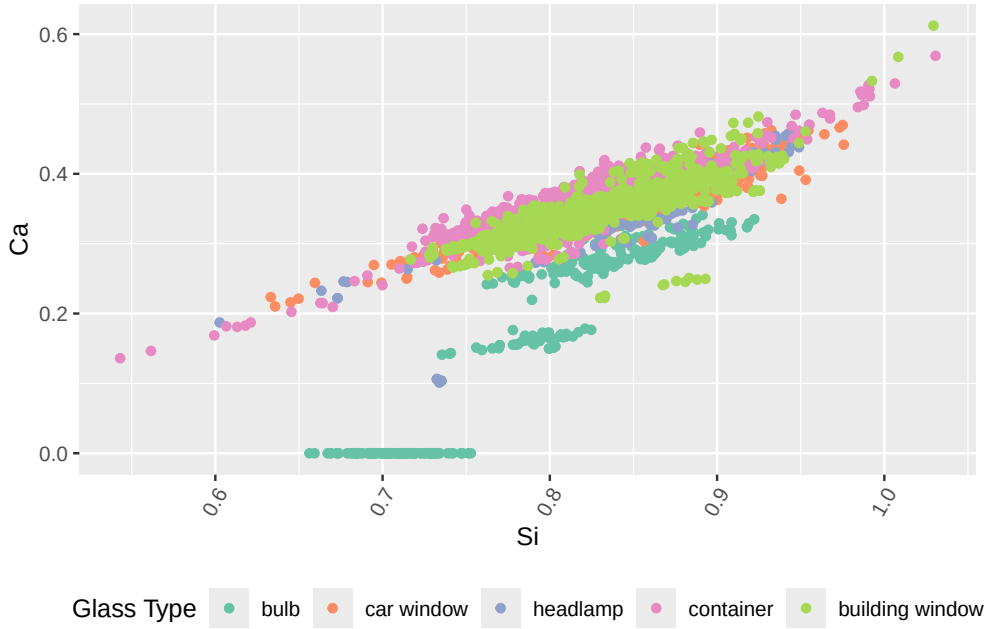


Figure 3.4: Scatterplot of the elements silicon (Si) and calcium (Ca) of the square root transformed compositional ratios, Equation (3.4.1.1), with oxygen as the divisor for all the glass item means. The different coloured boxplots correspond to each of the use type groups: bulb, car window, headlamp, container and building window.

3.4.2 Model with no splitting

Initially, we applied the Bayesian hierarchical model outlined in Section 3.3 to the full data for both untransformed and square root transformed compositional ratios. This comparison allows us to assess whether applying a transformation, as shown in Figure 3.3, helps to spread the data for each element away from zero, improving the predictive accuracy. For each model, the untransformed and square root transformed ratios, four chains were run in parallel, for 250,000 iterations each, with 150,000 discarded as burn-in and thinning each chain by 100. The time taken to run each model was approximately 3 hours.

For each version of the model that was run, the diagnostics outlined in Appendix B were examined to assess whether the model had adequately converged. Posterior samples were obtained for the sampled $\boldsymbol{\theta_t}$ and the resulting traceplots indicated good convergence across all glass use types. Each traceplot exhibited sufficient mixing, with no trends or periods of no movement, suggesting that the chains have explored the parameter space effectively. In addition, the multiple chains overlap indicating that they have all converged to the same posterior distribution. We computed the Gelman diagnostic (Appendix B) for each parameter and 96% of the PSRFs were less than or equal to 1.05, with a median of 1.00, indicating convergence.

Figure 3.5 displays boxplots of the posterior samples of $\boldsymbol{\theta_t}$ for the untransformed and square root transformed compositional ratios. Ideally, we would want to see a differentiation between each glass use type for each element. This could indicate that the model could learn the differences between the glass use types, to inform in a better way the classification of new items. From the top row of Figure 3.5, we can see that for the untransformed model the associated boxplots are very narrow and similar across the glass use types. Additionally, for a number of elements the boxplots for all glass use types are very narrow around zero. On the other hand, from the bottom row of Figure 3.5, across each element the posterior median of $\boldsymbol{\theta_t}$ appears to differ for each glass use type. However, for sodium (Na) and silicon (Si) there is a similarity within the boxplots for each glass use type. Overall, across the elements, the glass use types of car and building windows ($\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_5$) are the most similar, with the boxplots closely overlapping within each element. This is unsurprising given the likeness of the compositional elements within these glass use types. Thus, this could make prediction of either of these types more difficult given the close structure of each type.

Upon comparing the models, it is evident that applying a square root transformation enhances modelling the compositional ratios, as the posterior median of each glass use types differ. As a result, this could improve the prediction accuracy of new glass items. Hence, for the remainder of this chapter, the results will be based on applying a square root transformation to the compositional ratios.

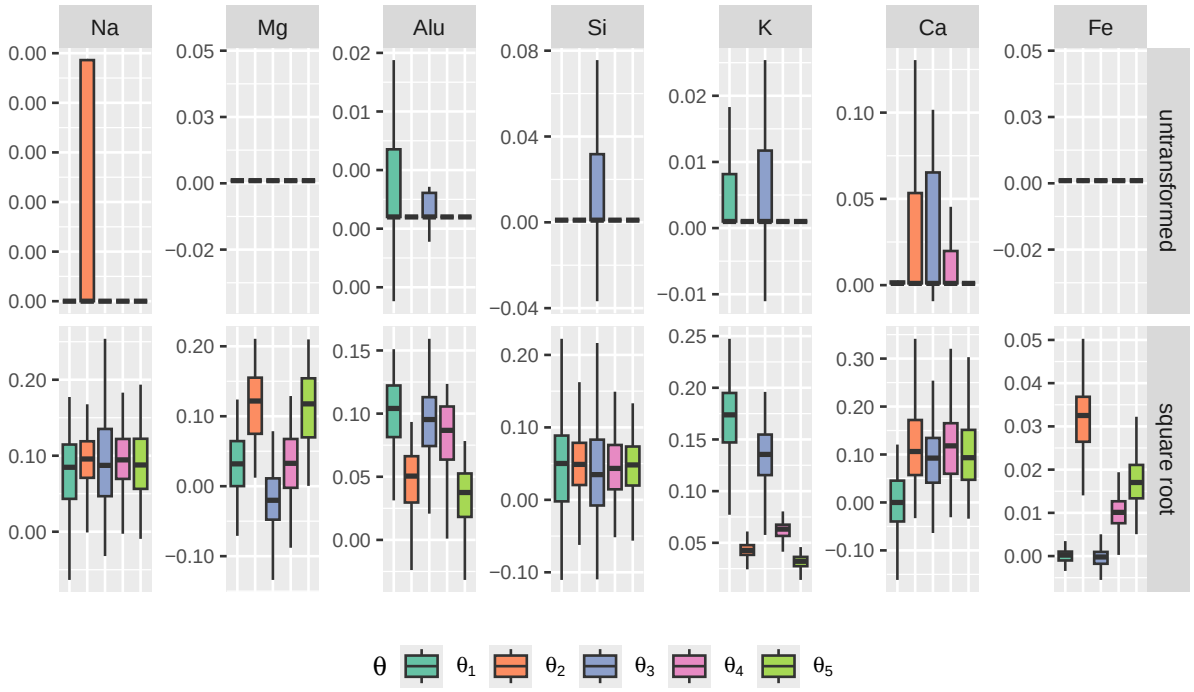


Figure 3.5: Boxplots of the posterior samples of $\boldsymbol{\theta_t}$ for the untransformed (top row) and square root transformed (bottom row) compositional ratios for the model with no splitting. The different coloured boxes represent the item types: bulb, car window, headlamp, container and building window.

3.4.3 Configuration models

Recall from Section 3.2.1, a previous approach to account for structural zeros is to split the data according to the presence or absence of the elements; called configurations. The *zCompositions* package (Palarea-Albaladejo et al., 2015) can be utilised to plot all the possible combinations of the presence and absence of the elements exhibited in the data. As oxygen, silicon and sodium are always present, from Table 3.2, the remaining five elements can either be present or absent, resulting in $2^5 = 32$ possible configurations. However, only

11 of the possible 32 are observed in the data, shown in Figure 3.6. Absence of an element is shown by the shaded blue boxes and the percentage of zeros of each element in each configuration is displayed in the teal bars on the top of the plot. The orange bars down the right side represent the percentage of the number of data points within each configuration. It can be noticed that some of these configurations contain a very small number of data points, Table 3.3 quantifies the frequency and percentage of data points in each of the observed configurations. For example, configuration 8, where solely magnesium (Mg) is absent, contains only one data point. This could present challenges when statistical methods are applied to each of the configurations.

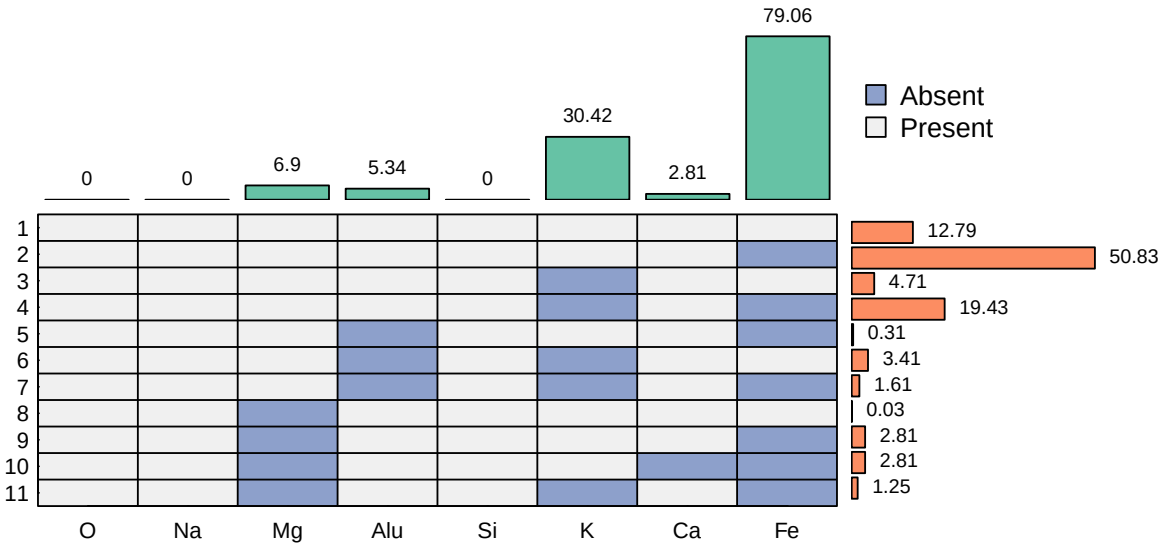


Figure 3.6: Plot of the presence and absence of the compositional elements for each observed configuration present in the forensic glass data. Absence of an element is shown by the shaded blue boxes in the grid and the percentage of zeros of each element in each configuration is displayed in the teal bars on the top of the plot. The orange bars down the right side represent the percentage of the number of data points within each observed configuration.

Table 3.3: Frequency and percentage of the number of data points within each configuration of the forensic glass data, from Figure 3.6.

Configuration	1	2	3	4	5	6	7	8	9	10	11
Frequency	491	1952	181	746	12	131	62	1	108	108	48
Percentage (%)	12.79	50.83	4.71	19.43	0.31	3.41	1.61	0.03	2.81	2.81	1.25

Napier (2014) overcame the issue of having a small number of data points within some of the configurations by only examining the presence and absence of the elements potassium (K) and iron (Fe). These elements account for the highest number of zeros in the data with a total of 88%. By only examining this presence and absence combination, the number of configurations to study reduces to four. Table 3.4 outlines the four configurations along with their glass use type.

Table 3.4: Number of glass items within each configuration for the presence and absence of the elements iron (Fe) and potassium (K) by glass use type.

Glass Type	**Configuration** m				**Total**
	$m = 1$: Fe present, K present	$m = 2$: Fe absent, K present	$m = 3$: Fe present, K absent	$m = 4$: Fe absent, K absent	
bulb	0	25	0	1	26
car window	23	40	11	20	94
headlamp	0	14	0	2	16
container	12	48	0	19	79
building window	10	52	15	28	105
Total	45	179	26	70	320

Upon examination, it was found that only eight of the 320 items have measurements that do not match the pattern seen for the other measurements for each item, accounting for 0.9% of the zeros in the data. In these cases, if an element is absent in one measurement but present in the others, it is recorded as a presence within the item.

Each of the configurations can be visually displayed using boxplots of the item mean compositional ratios to oxygen. Figure 3.7 presents this for each of the four configurations from Table 3.4. Here, we can see the elemental differences across the configurations for each glass use type. The absent elements within each configuration are clearly seen where the boxplot is a thin black line at zero, e.g. in Configuration 4 for iron (Fe) and potassium (K). Not all glass use types are present in each configuration (Table 3.4). This can be visualised where the boxplots for these glass use types are omitted within each panel, e.g. for Configuration 3 which only contains the two window types: car and building.

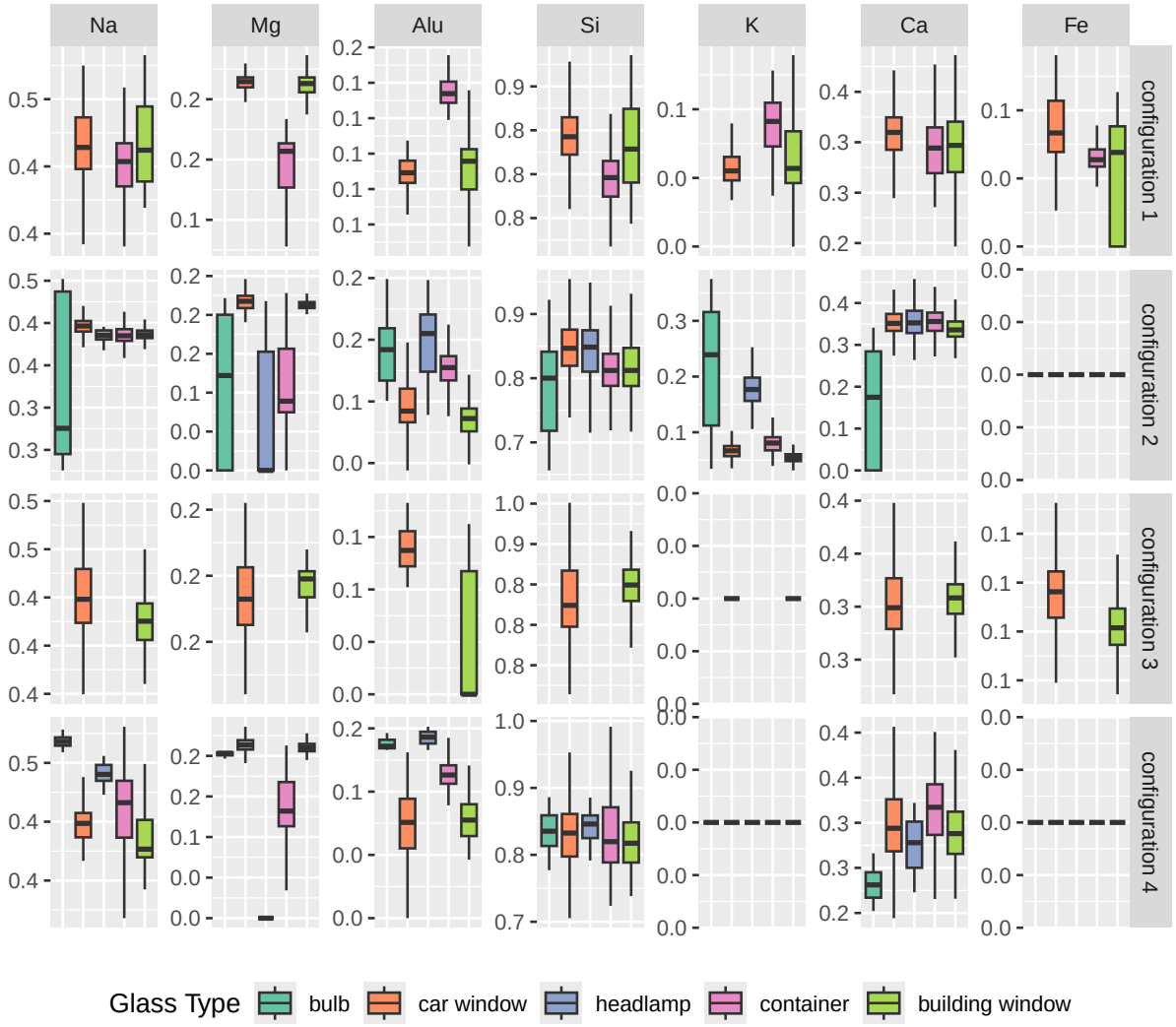


Figure 3.7: Boxplots of the square root transformed compositional ratios to oxygen for all the glass item means for each of the manual configurations. The different coloured boxes represent the item types: bulb, car window, headlamp, container and building window.

To fit our model to the manual configuration approach, we ran four chains in parallel for 200,000 MCMC iterations, discarding the first 100,000 as burn-in and thinning each chain by 100. The time taken to obtain the model results was approximately 3 hours and 30 minutes. The Gelman diagnostic for each of the $\boldsymbol{\theta_t}$ was equal to 1.00. When considering all parameters, the Gelman diagnostic was computed for each parameter and 90.4% of the PSRFs were less than or equal to 1.05, with a median of 1.01, indicating adequate convergence.

We can inspect the posterior samples of $\boldsymbol{\theta_t}$ for Configuration 2 (the largest configuration) in Figure 3.7. The boxplots for the elements silicon (Si) across all glass use types exhibit overlapping, which may pose challenges when classifying new glass items. Additionally, this can be seen for sodium (Na) and calcium (Ca) for all the glass use types except bulbs which have a lower point estimate. For these elements, the model may struggle to distinguish between the different glass use types due to the overlap. Additionally, the wide spread across these elements suggests greater variability in the posterior samples of $\boldsymbol{\theta_t}$, further complicating classification.

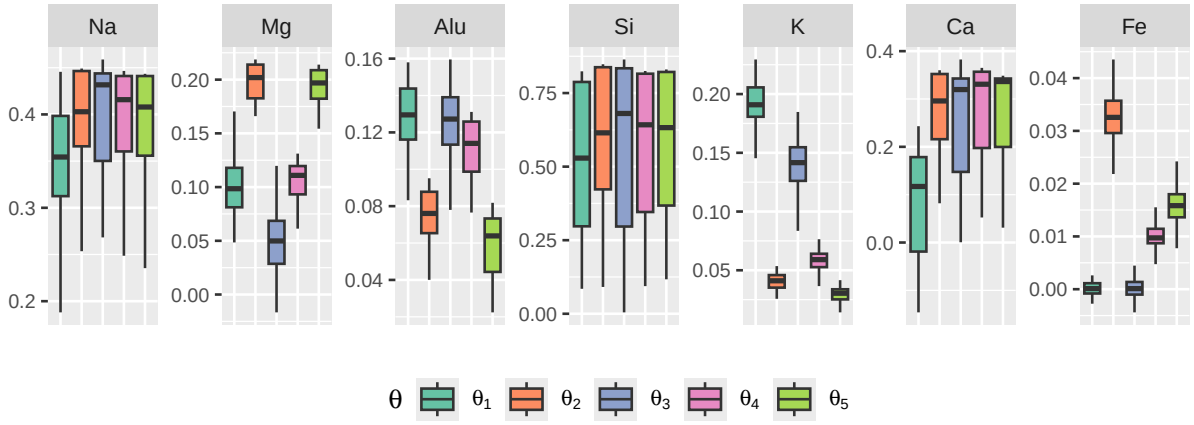


Figure 3.8: Boxplots of the posterior samples of $\boldsymbol{\theta_t}$ for the square root transformed manual configuration approach for Configuration 2 - K present and Fe absent. The different coloured points represent the item types: bulb, car window, headlamp, container and building window.

3.4.4 Pre-clustering: hierarchical clustering models

One of the main difficulties in clustering is the choice of the optimal number of clusters. In hierarchical clustering, this decision is made after the clustering has been completed. Several techniques have been developed to assist with this subjective choice, one of which is the elbow method, as detailed within Humaira et al. (2020). This approach plots the total within-cluster sum of squares as a function of the number of clusters. Since increasing the number of clusters will naturally reduce the total within-cluster sum of squares, the optimal number of clusters is typically identified at the "elbow" of the curve - where the rate of decrease in the total within-cluster sum of squares levels off. However, choosing too many clusters may lead to overfitting, capturing noise rather than meaningful structure.

Figure 3.9 presents the elbow plot of the indicator matrix of the presence and absence of each glass item. The "elbow" of this curve occurs at around five clusters. Therefore, we make the decision that for the hierarchical clustering algorithm using Ward linkage and the binary distance, five clusters are used to explain the presence and absence of the compositional elements.

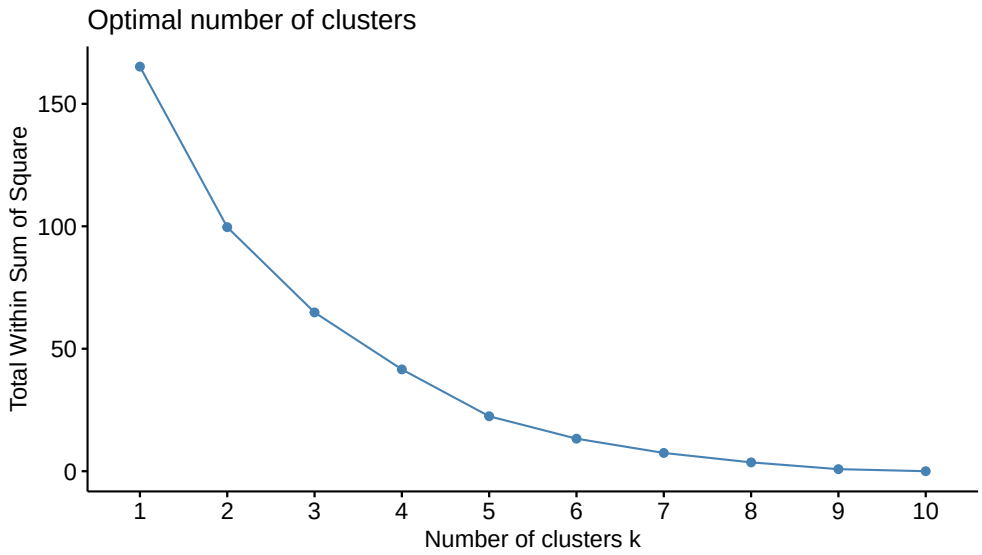


Figure 3.9: Pre-clustering hierarchical clustering elbow plot for the indicator matrix of the presence and absence of each glass item.

To understand what each cluster represents, the number of items of each glass use type in each cluster is shown in Table 3.5, along with the presence and absence of each compositional element displayed in Table 3.6. This allows for any similarities between the clusters and configurations to be identified. Cluster 1 is the largest and only cluster containing all five glass use types, resembling Configuration 2 as both have iron (Fe) absent. Cluster 2, the second largest cluster, excludes headlamps and is most similar to Configuration 4, both lacking iron (Fe) and potassium (K) and only containing one bulb. Cluster 3 represents the portion of Configuration 2 not included in Cluster 1. It also lacks iron (Fe) but additionally magnesium (Mg), omitting both window types.

Clusters 4 and 5 contain all elements, suggesting a subdivision of Configuration 1, composed of containers, car windows and building windows. Cluster 4 retains these three types, while Cluster 5 includes only the two window types and a single headlamp. This suggests the algorithm distinguishes headlamps from containers and identifies a similarity between headlamps and windows that was not apparent in manual analysis.

Overall, the hierarchical clustering algorithm aligns with previous analyses but introduces one new presence-absence distinction - the absence of magnesium (Mg) which had not been previously examined.

Table 3.5: Number of glass items within each hierarchical clustering cluster, for $k = 5$, by glass use type.

Glass Type	**Cluster**					**Total**
	1	2	3	4	5	
bulb	16	1	9	0	0	26
car window	40	18	0	23	13	94
headlamp	5	0	10	0	1	16
container	47	17	3	12	0	79
building window	55	25	0	7	18	105
Total	163	61	22	42	32	320

Table 3.6: Presence and absence of the compositional elements within each hierarchical clustering cluster, for $k = 5$. Presence in a cluster is defined by a 1 and absence by 0, with the total number of elements in each cluster given in the last column.

Cluster	**Element**							**Total**
	Na	Mg	Alu	Si	K	Ca	Fe	
Cluster 1	1	1	1	1	1	1	0	6
Cluster 2	1	1	1	1	0	1	0	5
Cluster 3	1	0	1	1	1	1	0	5
Cluster 4	1	1	1	1	1	1	1	7
Cluster 5	1	1	1	1	0	1	1	6

We ran four chains in parallel for 250,000 MCMC iterations, discarding the first 150,000 as burn-in and thinning each chain by 100 to fit the model with hierarchical clustering. The time taken to obtain the model results was approximately 4 hours and 20 minutes. The Gelman diagnostic for each of the $\boldsymbol{\theta_t}$ was equal to 1.00. When considering all parameters, the Gelman diagnostic was computed for each parameter and 94% of the PSRFs were less than or equal to 1.05, with a median of 1.01, indicating convergence.

We can examine the posterior samples of $\boldsymbol{\theta_t}$ from the largest cluster, Cluster 1, in Figure 3.10. This shows a clear separation between the glass use types across the elements. If the glass use types have different point estimates, we can assume that this may aid the classification of new glass items. The exception to this can be seen for sodium (Na) and calcium (Ca), where all the boxplots for all the glass use types except for bulbs have similar point estimates. For these boxplots we can see a narrow spread indicating that there is a smaller variability associated with the samples of $\boldsymbol{\theta_t}$ for these types. Within all the elements, the point estimates for the two window types (car and building) are overlapping which could potentially impact the classification of new glass items.

We applied the same approach using the k-means clustering algorithm instead of hierarchical clustering. Since k must be predefined, we ran the algorithm multiple times with different k values to determine the optimal number of clusters, including fitting the Bayesian hierarchical model for each k. Based on this process and the elbow plot assessment, we again

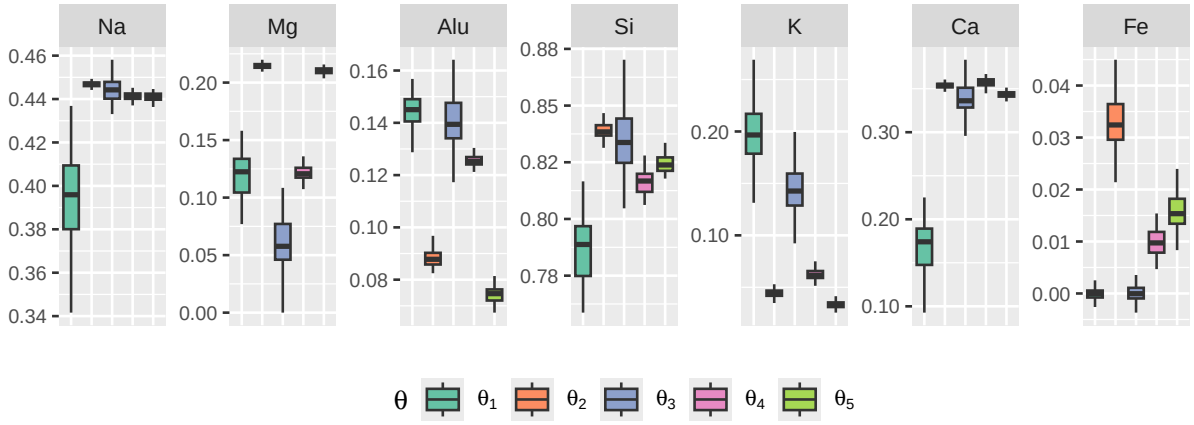


Figure 3.10: Boxplots of the posterior samples of $\boldsymbol{\theta_t}$ for the square root transformed compositional ratios for the pre-clustering hierarchical clustering Cluster 1. The different coloured boxes represent the item types: bulb, car window, headlamp, container and building window.

arrived at five clusters which closely resembled those from hierarchical clustering. For example, the largest cluster, Cluster 1, includes all glass use types and compositional elements except iron (Fe), while, Cluster 5 contains all compositional elements with car windows, containers, and building windows - which was also detected in hierarchical clustering. Additionally, fitting the Bayesian hierarchical model to both pre-clustering approaches yielded highly comparable results. We will present the results for both pre-clustering approaches within Section 3.4.7.

3.4.5 Integrated clustering

For the proposed integrated clustering approach, we ran eight chains in parallel for 400,000 MCMC iterations, discarding the first 300,000 as burn-in and thinning each chain by 100. The time taken to run this model was approximately 16 hours 40 minutes. This is significantly longer than we observed within our other approaches as our new model is having to estimate each glass item's cluster information.

As outlined in Section 3.3.3, we do not have definitive cluster labels for each glass item but instead have potentially different cluster labels for each chain. Therefore, we are unable to plot and examine the posterior samples of $\boldsymbol{\theta_t}$ as given for the previous approaches. Furthermore, we cannot examine the Gelman diagnostic for all relevant parameters as the different chains will not have converged to one distribution due to the differing cluster labels. Instead, we can explore the PSRF of the types of the glass items to check they have converged to one glass use type. We found that 94% of the PSRFs were less than or equal to 1.05 with a median of 1.01, indicating that the chains have converged to the same glass use type for each glass item. We can also visually inspect the parameters to see if the chains have converged to some distribution even if that will not be the same across chains. An illustrated example of this is given in Figure 3.11 which represents the traceplot for the prior cluster probability for Cluster 5. It can be seen that each chain has a distinct trajectory and appears to have converged even though this may be a different distribution to the other chains. However, as there are multiple overlapping chains, suggesting that these chains may have converged to the same distribution.

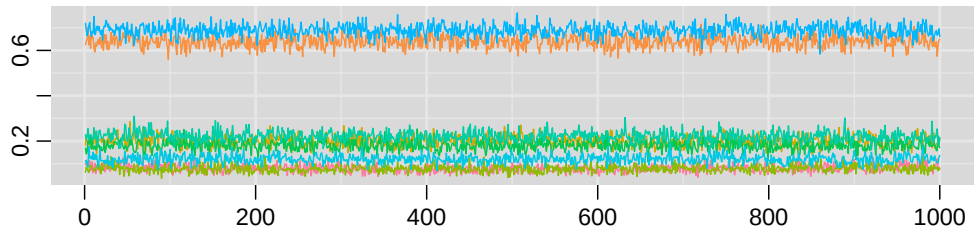


Figure 3.11: Traceplot of `p_cluster[5]`, the probability of Cluster 5 from the integrated clustering approach. The different coloured lines correspond to each of the eight chains.

3.4.6 Experimental design

To investigate the classification performance of the different Bayesian hierarchical models we tested and designed a five-fold cross-validation experiment. We randomly split the data into five equal parts by item, each containing 64 glass items. One at a time, each of the five parts is selected as the "test" data and the remaining four parts are combined to form

the "training data". Note that "test" data usually refers to data for which the value of the response is treated as unknown, in this case this would be the compositions. Here, however, the compositions can be seen by the model and the item type is the unknown quantity to be predicted. Moreover, the five-fold design means that all 320 glass items have an unknown glass use type to be classified exactly once.

To assess how well the different models perform at classifying the unknown types of glass items, each glass item was classified in the model as one of the five glass use types, outlined in Section 3.3. When we treat an item's type as unknown (i.e. to be predicted), we choose the type with the highest posterior probability as our "best" prediction. We obtain these by taking the mode across all posterior samples for each glass item. Using these predictions, we compute correct classification rates for each glass use type and modelling approach, given in Table 3.8. The performance of the different approaches can be compared and assessed based on their classification performance and accuracy.

We can also assess the uncertainty of each approach to compare the predictions from each model. The Brier score (Brier, 1950) is a commonly used tool to assess and compare the accuracy of binary predictions or prediction models, which can be thought of as a cost function. For a set of N predictions, the Brier score measures the mean squared difference between the predicted probability assigned to the possible outcomes for item i and the actual outcome o_i:

$$\text{Brier Score} = \frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2, \tag{3.4.6.1}$$

where f_i is the predicted probability of the i^{th} item for N predictions. Therefore, the lower the Brier score the better the predictions are calibrated to the original values.

Another measure of calibration we can examine is the Expected Calibration Error (ECE) (Naeini et al., 2015). Here, calibration aims to align the predictions of the model with the true probabilities to ensure that the predictions are reliable and accurate. The ECE measures how well a model's estimated probabilities match the true probabilities by taking a weighted average over the absolute difference between accuracy and confidence.

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |accuracy(B_m) - confidence(B_m)|, \tag{3.4.6.2}$$

where B represents the M equally spaced "bins" the data are split into, the $accuracy(B_m)$ represents proportion of correct predictions in the bin B_m and $confidence(B_m)$ is the average predicted probability in the bin B_m. An ECE of 0 indicates a perfectly calibrated model with higher ECE values suggesting that the predicted probabilities do not match actual outcomes well.

3.4.7 Classification results

Table 3.8 displays the correct classification rates for each approach outlined. The number of each glass item classified into each glass type for the proposed integrated clustering approach is given in Table 3.7.

Table 3.7: Classification of each glass item into one of the five glass use types, for the integrated clustering approach. The rows represent the observed glass use type for each item and the columns represent the glass use type each item has been classified into.

Classification	bulb	car window	headlamp	container	building window	**Total**
bulb	26	0	0	0	0	26
car window	0	71	1	2	20	94
headlamp	2	0	2	11	1	16
container	2	4	0	69	4	79
building window	0	21	2	6	76	105
Total	30	96	5	88	101	320

Initially, we find that not splitting the data - i.e. fitting the model on the entire data with untransformed ratios - results in the poorest overall correct classification rate of 38%. This poor performance is largely due to the model failing to classify any items as headlamps. We then find that transforming the compositional ratios using a square root transformation improves the classification performance, to 66% overall. However, the two window types are commonly misclassified as each other resulting in poor classification for both glass use types, with rates of 59% and 57% for car window and building window, respectively. This highlights the need to account for the presence and absence of the compositional elements to improve the classification of new glass items.

Across all approaches where we split the data based on the presence and absence of the elements, whether through manual or automated methods, the glass use type bulb has the highest correct classification with each approach correctly classifying all 26 glass items as bulbs.

The pre-clustering hierarchical clustering approach performs best at classifying headlamps, the smallest glass use type consisting of only 16 glass items. This approach correctly classifies 15 of the 16 glass items, achieving a success rate of 94%. The manual configuration approach performs second best here, classifying 13 of the glass items correctly. Although, it performs worse than k-means with respect to only correctly classifying 11 of the 16 headlamps. However, these are far superior in comparison with the integrated clustering approach which only results in a correct classification rate of 13%. While this result is notably poor it could potentially be improved if there were more headlamps in the data, to enhance the number of this glass use type the model can train on. Interestingly, within the integrated clustering approach most of the misclassified headlamps have been classified as containers, with 11 of the 16 classified as such.

Among all the approaches, except the untransformed compositional ratios, the glass use type container tends to be classified correctly, with correct classification rates of 86%, 90%, 89%, 90% and 87%, for the no split square root transformation, manual configurations, hierarchical clustering, k-means clustering and the integrated clustering approach, respectively. The performance is highest for the manual and pre-clustering approaches which correctly classify over 70 of the 79 new containers.

The most commonly misclassified glass use types are car and building windows, due to their similar elemental structure. However, the integrated clustering approach outperforms the other methods at classifying each of the window types. For the car window, the integrated clustering approach has a correct classification rate of 76%, which is 15% higher than the next highest classification rate for this type. For building windows, the integrated clustering approach has a correct classification rate of 72%, which is 5% higher than the manual and pre-clustering approaches. Despite the integrated clustering approach still misclassifying each window type as the other, the frequency of such is lower compared to the other approaches, with only about 20 glass items misclassified as the other window type.

Table 3.8: Correct classification rates for each approach examined in this section. The highest correct classification rate for each glass use type is highlighted in green, with a 2% tolerance applied in cases of near ties.

Approach	bulb	car window	headlamp	container	building window	**Overall**
No spilt: untransformed	26.9%	42.6%	0%	12.7%	61.9%	38.1%
No spilt: square root	76.9%	58.5%	43.8%	86.1%	57.1%	65.6%
Manual: configurations	100.0%	61.7%	81.2%	89.9%	69.5%	75.3%
Pre-clustering: hierarchical	100.0%	64.9%	93.8%	88.6%	68.6%	76.2%
Pre-clustering: k-means	100.0%	64.9%	68.8%	89.9%	68.6%	75.6%
Integrated clustering	100.0%	75.5%	12.5%	87.3%	72.4%	76.2%

In general, the overall correct classification rates are similar for all approaches that partition the data based on the presence and absence of the elements. The highest correct classification rate occurs for the integrated clustering approach and the hierarchical clustering with an overall rate of 76%. However, when allowing a 2 percentage point tolerance, the manual configuration and k-means clustering also have the highest classification rate. Therefore, des-

pite the poor classification of the headlamps, the integrated clustering approach successfully classified 244 of the 320 new glass items in the database. This demonstrates that applying an automated clustering approach, either before or during the modelling, leads to a high classification accuracy. Therefore, incorporating a fully automated clustering process into the model, minimising user decisions prior to fitting, delivers advantageous classification performance while being widely applicable.

We can visually examine the classification within the integrated clustering approach in Figure 3.12, which displays the two highest posterior probabilities for each new glass item. Each panel in this plot is the observed glass use type of each item with the shape and colour representing the glass use type each item has been classified into. The perfect classification of the bulb glass use type can be clearly seen with all the highest points in the bulb panel shaped and coloured as a bulb. Additionally we can see that each bulb item has a high posterior probability associated with it, meaning there is a strong likelihood that the item is, in fact, a bulb.

Recall that we saw that headlamps had the poorest correct classification rate. This is evident in the plot where the posterior probabilities for correctly classified headlamps are lower, with the highest posterior probability for headlamps is 0.86, highlighting the uncertainty the model has when it classifies glass items as headlamps. However, it appears that when a headlamp has been misclassified, the second highest probability corresponds to a headlamp - though this probability is low, less than 0.5. As shown earlier, the two window types are commonly misclassified as each other. We can see that both the window panels exhibit significant overlap in the points and lack a distinct structure or pattern as observed with the other glass use types. Within both these panels, a substantial proportion of points of the other window type are present, with some exhibiting a posterior probability greater than 0.75.

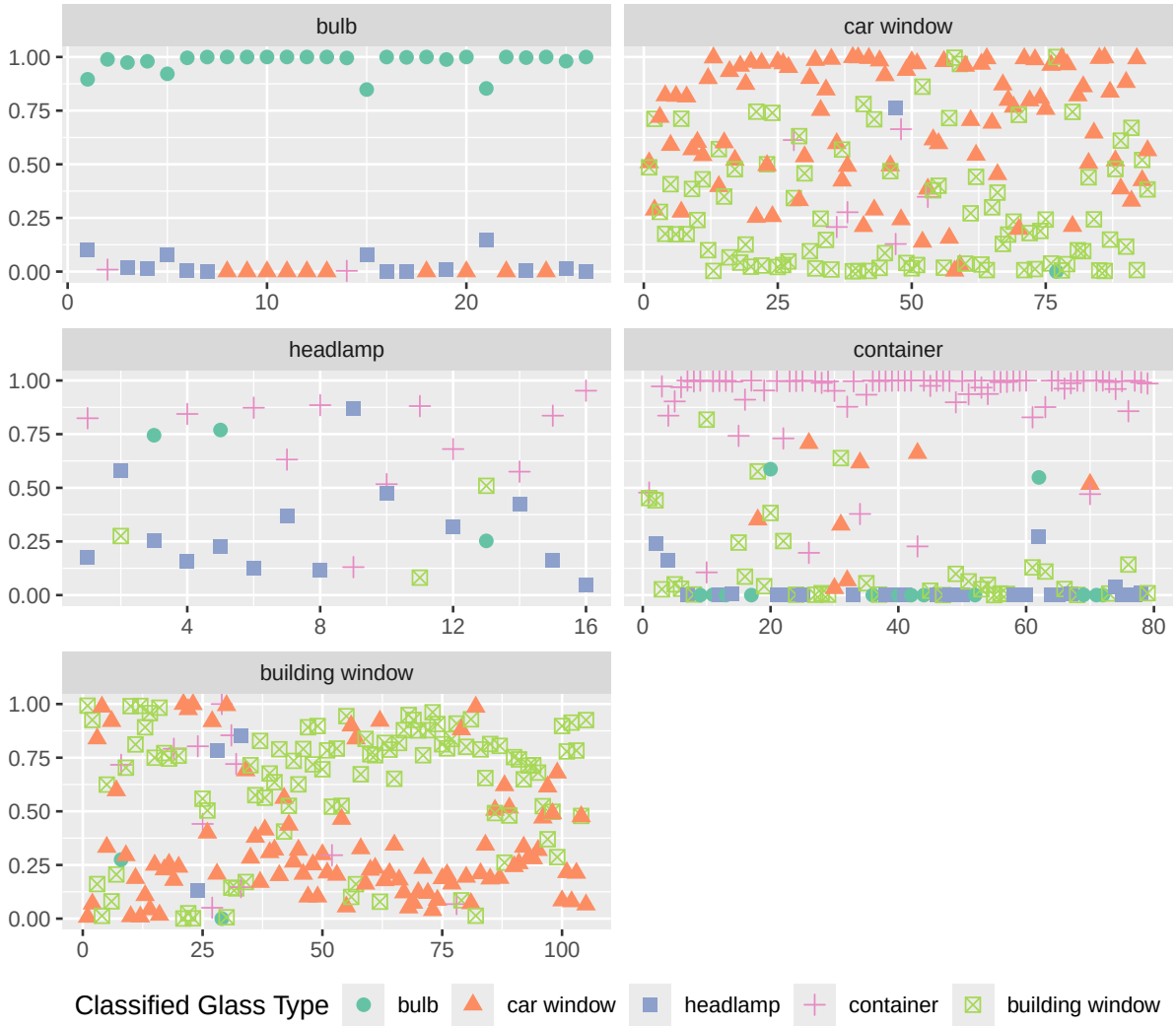


Figure 3.12: Posterior probabilities for the classification of each glass item into one of the five glass use types for the integrated clustering approach. The two largest posterior probabilities are displayed for each glass item. The panels refer to the actual glass type of each item, with the shape and colour representing the glass use type the item was classified into.

Instead of computing the mode across all samples, we can consider calculating the correct classification rates within each sample and summarising through the densities for each type of glass use. Although this was conducted for all the approaches considered, we illustrate this only for our most flexible proposed method. Figure 3.13 displays density plots of the classification rates across all samples for each glass use type. Here, we can see how the classification rate varies for each type across all 8,000 samples. Overall, the classification varies across the different glass types, with some having a large spread out distribution. For the glass use type bulb, we can detect a highly skewed distribution towards 100%, suggesting that most bulbs across the samples are correctly classified. Both the car and building windows exhibit a normal curve, with a peak occurring around 70%. This indicates that there is variability in the classification accuracy across the samples for the window types, with classification rates of 25% to 90%. The glass use type headlamp presents a multimodal distribution, with the highest peak occurring between 0-20% suggesting very poor classification performance. However, we can see that within some of the samples, headlamps have been well classified, with a small peak in the classification rates at 70% and 100%. Lastly, the glass use type container presents a left-skewed distribution with the number of correctly classified containers greater than 50%, with a peak round 75%.

To quantify these classification densities, we can examine the mean across each density for each glass use type, given in Table 3.9. In this case since we are averaging over the total number of samples (8,000), the correct classification percentages have decreased, as expected due to the increased variability with the number of samples. However, the values are still sufficient to suggest the integrated clustering approach is performing well at classifying out-of-sample glass items. Notably, the correct classification rate for headlamps has increased to 28%. Therefore, when we evaluate classification within samples rather than summarising the result across all samples, the classification accuracy for headlamps improves.

Figure 3.13: Density plots of the correct classification rates for each glass use type across all samples (8,000) for the Bayesian integrated clustering approach.

Table 3.9: Mean correct classification rates across all samples (8,000), represented in the density plots from Figure 3.13, for each glass use type for the Bayesian integrated clustering approach.

Glass Type	**Mean correct classification**
bulb	98.1%
car window	68.6%
headlamp	28.3%
container	86.3%
building window	60.2%
Overall	68.3%

We can quantify some measures of uncertainty as another way to compare our predictions from each model. Table 3.10 presents the individual Brier scores for each glass use type within each approach, in addition to an overall Brier score for each approach, and Table 3.11 displays the ECE values for each approach. Firstly, from Table 3.10 and Table 3.11, we can see that the worst performance in both the Brier scores and ECE occurs in the no splitting approach for both the untransformed and square root transformed compositional ratios. This outcome is expected as the model lacks any information accounting for the presence and absence of compositional elements, leading to more uncertainty in classifying new items. This aligns with the poor correct classification rates associated with this approach.

Among the methods that account for the presence and absence of zeros in the data, the manual approach and hierarchical clustering exhibit the lowest Brier score for the classification of new bulbs, at 0.002. Since this value is very close to 0, we can conclude that the models associated with these approaches are highly certain when classifying new bulbs. These same two approaches also have the lowest uncertainty associated with predicting new headlamps, closely followed by the k-means clustering approach. Unsurprisingly, the integrated clustering approach has a Brier score for headlamps 59% higher compared to the lowest score of 0.019. Given that the integrated clustering approach performs poorly in classifying new headlamps (Table 3.7), it is expected that this approach would exhibit higher uncertainty in these predictions. Notably, the integrated clustering approach has the lowest Brier score for both the window types, with scores that are 6% lower for car windows and 9% lower for building windows, aligning with this approach performing best at classifying these glass use types.

Overall, the lowest Brier score occurs for hierarchical clustering with k-means clustering, manual and the integrated clustering approach following closely behind. Comparing this with the ECE values in Table 3.11 presents a similar pattern. Here, the lowest ECE values occur for the manual approach and when allowing for a 2 percentage point difference also the integrated clustering approach. This indicates that the predicted probabilities are

well-aligned with the actual outcomes, and the model demonstrates confidence in its predictions, reflecting better overall performance. In addition to the correct classification rates and uncertainty examined here, a number of classification performance metrics can be considered. These reflected a similar performance to that presented throughout this section and included in the Appendix C.

Table 3.10: Brier score quantifying the classification uncertainty for each approach examined in this section. These scores correspond to the classification results presented in Table 3.8. The lowest (optimal) Brier Score is highlighted in green, with a 2% tolerance applied in cases of near ties.

Approach	**Brier Score**					
	bulb	car window	headlamp	container	building window	**Overall**
	n = 26	n = 94	n = 16	n = 79	n = 105	n = 320
No spilt: untransformed	0.060	0.203	0.049	0.187	0.215	0.714
No split: square root	0.035	0.163	0.034	0.069	0.170	0.472
Manual: configurations	0.002	0.140	0.019	0.043	0.157	0.361
Pre-clustering: hierarchical	0.002	0.132	0.019	0.042	0.154	0.348
Pre-clustering: *k*-means	0.004	0.133	0.023	0.043	0.155	0.358
Integrated clustering	0.006	0.124	0.035	0.062	0.141	0.369

Table 3.11: Expected Calibration Error (ECE) quantifying the classification uncertainty for each approach examined in this section. These scores correspond to the classification results presented in Table 3.8. The lowest (optimal) ECE is highlighted in green, with a 2% tolerance applied in cases of near ties.

Approach	**ECE**
No spilt: untransformed	0.985
No spilt: square root	0.730
Manual: configurations	0.629
Pre-clustering: hierarchical	0.683
Pre-clustering: *k*-means	0.682
Integrated clustering	0.631

3.5 Summary & Discussion

In this chapter, we investigated the issue of compositional data which contain a large proportion of zeros. The main challenges we identified were that the common approach of applying a log-ratio transformation to the data are unsuitable as this would be undefined for zeros. Previous approaches looked to tackle this issue through a manual approach to split the data based on the presence and absence of the compositions. However, this requires expert knowledge and user intervention to undertake.

We set out four aims to create a flexible and efficient framework requiring less input from the user; to develop a computationally efficient framework which would be practical in real-world analysis; investigate data-driven approaches to splitting the data based on presence and absence; explore a model-based approach for the classification of glass items; and implementing the framework using flexible MCMC software.

To address these aims, we proposed a general Bayesian hierarchical framework for modelling compositional data that seek to account for the presence and absence of the compositional elements. The general framework includes a categorical distribution to predict the glass use type of each glass item to streamline the steps for the classification. We implemented our models using NIMBLE which results in computationally efficient frameworks which makes for more practical application in real-world analysis. Here, modifications to the model can be made more easily than if we had a manual MCMC implementation, adding flexibility in the modelling structure. This enables any specific samplers or distributions to be modified by the user to fit the specific application, showing a practical advantage over previous work.

We included a clustering approach to automate the manual aspect of splitting the data. This reduces the need for any expert knowledge required. Furthermore to enhance automation, we proposed a flexible approach which includes a latent variable for the cluster of each glass item within the Bayesian hierarchical model, automating the framework further. Due to this addition, the method becomes more broadly applicable as it can accommodate a wide range of applications without requiring extensive expert knowledge or manual intervention.

When examining the clusters produced during the pre-clustering approaches, we found that clustering on the indicator matrix of presence and absence of the elements preserved much of the key information highlighted during the manual splitting of the data. For example, in both clustering approaches the largest cluster does not contain the element iron (Fe) matching the largest configuration from the manual approach. Moreover, the pre-clustering approaches offer the advantage of exploring elemental combinations that may not have initially been considered relevant for modelling. In practice, we believe that unless there were specific elemental combinations to be examined, the clustering approach would be an easier method to implement. In addition, even if there were specific elemental combinations to be examined, from expert knowledge or key aims of the modelling, it is likely that the pre-clustering approaches would identify these. However, one thing to note is that there are still user decisions to be made prior to fitting a pre-clustering approach, e.g. which clustering algorithm to apply and how many clusters to use - although tools have been developed to help with this. Furthermore, it would be important to make sure all clusters have a suitable number of data points to ensure the model has enough information to estimate accurately the parameters. If there is a limited number of data in a cluster when fitting the model, the prior could have a stronger influence on the estimates rather than the data. Additionally, the framework requires at least one component to be greater than zero to compute the compositional ratios.

Despite incorporating an integrated clustering variable which creates a practical framework, we still need to define the number of clusters prior to modelling. This could be seen as a limitation not only of this approach but of all clustering methods. For a more formal treatment of the choice of the number of clusters, one could vary the number of clusters and find an optimum based on out-of-sample performance or a general criterion, such as the Widely Applicable Information Criterion (Watanabe et al., 2010), which can be automatically generated by NIMBLE if desired.

As highlighted earlier our integrated clustering proposed approach also does not include any constraints to address any label-switching of the cluster labels or structures, as it is not the focus of our analysis. This could be seen as a limitation if the goal is to determine the optimal clustering structure for the presence and absence of the glass items. In such cases, constraints to avoid any label-switching of the clusters could be incorporated into the model, using NIMBLE's "constraint" distribution function.

We evaluated different approaches within our proposed framework in their application to a forensic elemental glass data. Our assessment involved examining the proposed integrated clustering approach through out-of-sample classification performance via five-fold cross-validation. We then compared this approach: to an instance where the data was not split, containing either untransformed and square root transformed compositional ratios; to a manual approach to splitting the data and to pre-clustering methods of hierarchical and k-means clustering to automate the split. When we considered not splitting the data, both the untransformed and square root transformed compositional ratios resulted in a poor performance at classifying the glass items, indicating that any approach that splits the data is potentially advantageous. All the approaches that split the data based on the presence and absence of the compositional elements have comparable classification results. Overall, our proposed framework yielded the best overall correct classification, this could be due to the added potential our framework has to find more optimal clusters for modelling. Additionally, our approach also performed best when classifying car and building windows,

where the other approaches struggled given their close compositional elemental structure. However, our proposed approach was less effective at classifying headlamps due to the limited number of these glass items present in the data. With further research another model or specification could address the poor classification of the headlamps. Nevertheless, our model performs well in our task of classifying new glass items marginally beating the other approaches.

Additionally, a future model could be extended to include covariates which could aid in the predictions of the classifications of each glass item. Despite only evaluating the framework with the forensic glass data, we believe it could be applied more broadly to compositional data beyond forensic science, particularly in cases where there are numerous structural zeros present.

Chapter 4

Methods for Compositional Time Series

In this chapter, we explore Bayesian approaches to modelling compositional data that evolve over time. This requires methods that account for both the compositional structure and temporal dependence. When the time series consists of count data, applying log-ratio transformations can obscure meaningful information about the overall dynamics and variability of the counts. This motivates an approach that can directly handle raw compositional counts, in particular alongside a non-smooth time series.

Here, we propose a novel Bayesian hierarchical modelling framework that combines a Generalised-Dirichlet-Multinomial (GDM) distribution for compositional counts with a latent hidden Markov model (HMM) structure to capture non-smooth temporal dependence. We apply the methodology to a COVID-19 variant data including non-smooth count over time containing zeros. We assess our approach and compare the effectiveness to other common time series models through a posterior predictive experiment.

4.1 Introduction

As discussed in the preceding chapters, compositional data pose distinctive challenges across different statistical domains. The compositional structure introduces specific modelling challenges, including the requirement that the components sum to a total. As in many statistical analysis contexts, compositional data can also be arranged over time, which we refer to as compositional time series. Traditional time series methods do not typically account for compositional structures, thus methods specifically tailored to handling compositional time series data are likely to offer a more rigorous solution for inference and/or prediction. Within this focus, the most appropriate methods may also depend on finer characteristics of the data, such as the smoothness of the temporal structure, or the nature of the data (e.g. continuous or counts).

Within this chapter the real-world motivation is the challenge of modelling the composition of disease cases into different variants of a virus causing infections, with a specific focus on capturing the evolution of the compositions over time. Here, we will study data on COVID-19 case counts by variant, in an international context.

The emergence of the novel coronavirus, SARS-CoV-2, in late 2019 marked the onset of a global health crisis that had profound implications across all areas of life. The associated illness, COVID-19, quickly escalated into a pandemic, leading to an unprecedented international effort to understand, control and mitigate its impact. While considerable progress has been made in characterising the primary dynamics of the virus and developing preventative measures, the ongoing evolution of the virus through genetic variants continues to pose an ongoing challenge. Further information about the COVID-19 pandemic can be found in Atzrodt et al. (2020). The SARS-CoV-2 virus is an RNA virus, a type known for extreme mutations. These genetic alterations result in the formation of distinct variants, each carrying unique genetic signatures. The notable variants from SARS-CoV-2 are alpha, beta, gamma, delta and omicron (Harvey et al., 2021). Each of these variants is thought to have originated from different areas of the world where the first genetic mutation of that variant was found. The development of each new variant led to varying degrees of severity and concern among public health officials, prompting differing levels of interventions. Leading the classification of each new variant was the World Health Organisation (WHO) who identified each new strain as either "variant of concern" (VOC) or "variants of interest" (VOI). These categorisations signify variants that exhibit notable characteristics such as increased transmissibility, severity of illness and potential impact on vaccines, all of which warrant closer monitoring and public health responses. Through ongoing surveillance and assessment, WHO aimed to identify and respond swiftly to emerging variants to mitigate their potential impact globally.

Where records of case counts with a breakdown by variant are collected over time, we can conceptualise the resulting data as compositional time series. Figure 4.1 shows COVID-19 case counts for five VOCs (alpha, beta, gamma, delta, omicron) and an aggregated total for the VOIs, for one example country per continent. The different variants are coloured within each panel and the plots illustrate the emergence of new variants over the course of the pandemic. From the plots we can see diverse dynamics (e.g. how quickly the variant spreads) across the different variants and different countries. For instance, not all variants are present in all countries. Meanwhile, in the evolution of the case counts over time, we can see very sudden changes as new variant of concern emerge through mutations and take over as dominant strain.

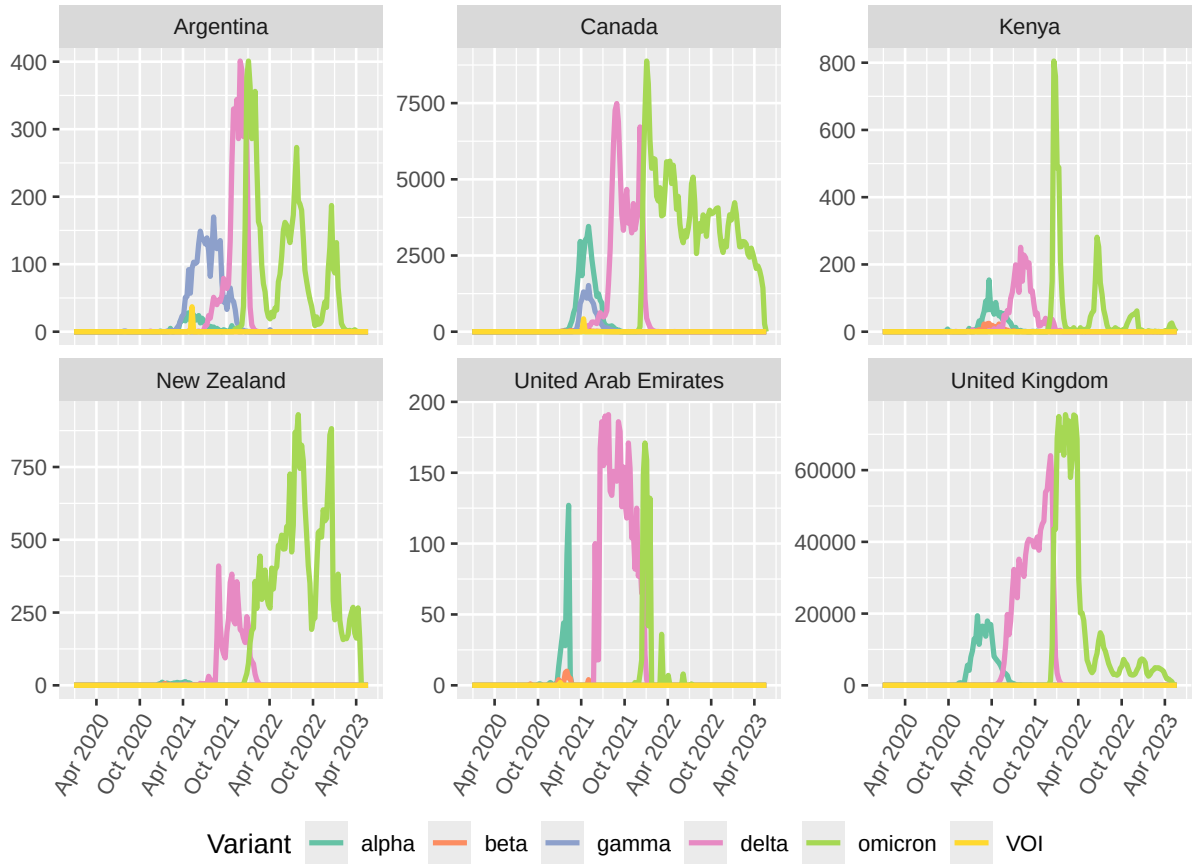


Figure 4.1: Time series of the weekly COVID-19 case count attributed to each variant, for one country per continent, from Jan 2020 to May 2021. The different coloured lines correspond to each of the COVID-19 variants: alpha, beta, gamma, delta, omicron and variants of interest (VOI).

In summary, Figure 4.1 points to three main characteristics of these data which we will aim to capture through developing a new modelling framework:

i. Compositional structure, specifically count data.
ii. Systematic differences between variants and between countries.
iii. Non-smooth temporal structure often characterised by sudden changes in the dominant strain.

In this chapter, we develop a multivariate hierarchical framework for non-smooth compositional time series data, combining a flexible family of distributions for compositional counts with a hidden Markov model (HMM) for the non-smooth temporal structure. In the context of the challenge of capturing the above three characteristics of COVID-19 variant data well, the HMM will characterise the lifetime of a new variant as a progression through multiple discrete states, from initial emergence through to dormancy.

The chapter is structured with Section 4.2 reviewing the previous approaches to modelling compositional time series. Section 4.3 provides an introduction to hidden Markov models (HMM), including examining the literature surrounding HMMs in the context of compositional data and disease data applications (including COVID-19). In Section 4.4 we present the proposed general framework, adapted for the COVID-19 variant case study seen in Section 4.5 with the results for the proposed framework presented in Section 4.6. To test effectiveness of the proposed compositional HMM, in Section 4.7.1 compares our HMM to variants of the hierarchical framework using alternative time series structures, based on posterior predictive model checking. Finally, in Section 4.8 we critically evaluate the work carried out in this chapter and discuss potential avenues for future research on this topic.

4.2 Compositional Time Series

Compositional time series refer to data that are both in the form of parts of some whole, as outlined in Chapter 2, and are arranged over time, usually with an expectation of a time-dependency structure. Compositional time series emerges in many applications, e.g. in health data (Ravishanker et al., 2001), environmental data (Al-Dhurafi et al., 2018), and labour markets (Brunsdon et al., 1998). Modelling compositional time series requires the development of robust statistical methods that effectively capture the temporal evolution of the constrained data. Such models could be developed for inference, prediction, or simulation and, in all these cases, failure to account for the compositional nature of the data could invalidate common modelling assumptions, e.g. independence of residuals.

4.2.1 Log-ratio transformations

Time series methods have been applied to compositional data for many years. As previously outlined, the majority of the proposed approaches aim to address the compositional nature of the data by applying a log-ratio transformation. The transformed data are then modelled using standard time series methods. For example, Brunsdon et al. (1998) applied the ALR to labour market time series data and then modelled the transformed time series using vector autoregressive moving-average (VARMA) models, to produce forecasts and associated measures of uncertainty. The combination of ALR transformations and VARMA models was also applied to mortality events in Ravishanker et al. (2001). Snyder et al. (2017) introduce a maximum likelihood approach to modelling market sales data using the ALR and exponential smoothing methods. More recently, Al-Dhurafi et al. (2018) apply the ALR

to air pollution index data. To account for zero values in the data (not permissible with log-ratio approaches), the authors took a multiplicative replacement approach, adding a small value to each zero value to eliminate any zero components, whilst maintaining the ratios of non-zero components, before fitting a VAR model.

Compositional time series can also take the form of a total count distributed across multiple categories. Many existing works have sought to model such data using log-ratio transformations. For instance, Sisk-Hackworth et al. (2020) examine microbial data to test if CLR is effective for analysing bacterial data. The compositional structure is addressed by taking the relative count with respect to the sample geometric mean. However, in order to do so, the zero values in the data had to be handled using the pseudo-counts method from the *zCompositions* package (Palarea-Albaladejo et al., 2015), which replaces zero values with a small positive number while maintaining the relationships between the components (i.e. preserving their unit-sum constraint). This is a limitation to using CLR as the zero values need to be altered in order to compute the transformation. Another instance of applying CLR to compositional time series with counts is Shang et al. (2022), who present an approach for modelling age-specific death counts across multiple populations. The counts are transformed using CLR before multivariate time series methods are applied for forecasting. The framework was tested using data from England, Wales and Sweden and the results were compared with several benchmarkers showing that the proposed method had superior performance in most cases.

Despite the most common approach being to transform the data using the log-ratio transformation and then apply standard statistical techniques, we argue there are at least three ways in which generality is limited. First, log-ratio transformations are undefined for zero values, which are often present in compositional data, including rounded and structural zeros, as outlined in Chapter 2, Section 2.3. Second, when there are missing values in the compositions, e.g. $y = (0.1, ?, 0.3, ?)$, log-ratios are usually undefined. Finally, considering compositional count data specifically, in addition to the problem of count data naturally

having zeros, through log-ratio transformations the count structure may translate into values which are simultaneously non-integer and non-continuous. These issues are likely to be more problematic the smaller the total count is, increasing the likelihood of zeros and decreasing the number of unique values the compositional counts can take.

4.2.2 Hierarchical approaches

A compelling alternative to log-ratio transformations, which can potentially overcome these limitations, is to develop general Bayesian hierarchical frameworks. These frameworks can feature probability distributions (e.g. Multinomial, Dirichlet, Generalised-Dirichlet) and mixture distributions (e.g. Dirichlet-Multinomial, Generalised-Dirichlet Multinomial, Logistic-Normal-Multinomial) that explicitly account for the compositional structure of count data. For example, Huston et al. (2012) propose a hierarchical multivariate conditional autoregressive (MVCAR) model applied to a compositional response vector of multinomial counts collected over time. The model is tailored for analysing compositional data with observed zero counts, particularly focusing on where the composition is discrete and based on small multinomial counts. The proposed model addresses limitations in existing approaches for handling count data, allowing for estimation even when zeros are observed in any component category. It also estimates a covariance matrix that the authors claim is not constrained by the limitations of Multinomial or Dirichlet-based models. The authors outline the importance of the proposed hierarchical model in reducing variance and smoothing proportion estimates through time, while also providing flexibility in adjusting the degree of smoothing. The proposed methods were applied to time series data on the migration patterns of salmon in the Fraser River, including information on the number of fish sampled daily, proportions of different stock groups and chronological day of sampling.

Meanwhile, Stoner et al. (2020c) also propose a multivariate hierarchical framework for modelling compositional count time series. The authors propose a Bayesian hierarchical model, based on the Generalised-Dirichlet–Multinomial (GDM) family of distributions, which they apply to the proportion of people in each country using different fuels for household cooking. The authors converted the available proportions into count data, for convenient modelling which they validated through a simulation study. Here, smooth non-linear trends in the use of eight of the key fuel types were captured through penalised regression splines in the parameters of the GDM. Applying the GDM distribution ensures the sum of proportions for all fuel types does not exceed 100%. The model has been adopted by the WHO for tracking worldwide progress from traditional solid fuels to greater use of cleaner fuels, and for estimating the global burden of disease from household air pollution.

Since we argue that hierarchical approaches to compositional time series have greater flexibility - i.e. addressing the limitations in the log-ratio approach we discussed in Section 4.2.1, we will not consider log-ratio approaches further beyond this point. However, the hierarchical methods in Huston et al. (2012) and Stoner et al. (2020c), relying on conditional autoregressive and spline structures respectively, may not be suitable for very non-smooth time series, such as the sudden emergence of variants in the COVID-19 case study. In the next section, we will discuss hidden Markov models and their potential integration into general frameworks for compositional data with non-smooth time series structure.

4.3 Hidden Markov models (HMM)

Hidden Markov models (HMMs), introduced by Rabiner et al. (1986) represent a widely used statistical modelling technique for data arranged over regular time intervals, e.g. days or weeks. In an HMM, we imagine that temporal dependence of an observable outcome y_t $(t = 1,\ldots,N)$ is handled by the temporal evolution of a "hidden" (unobserved) quantity

z_t. An HMM assumes that z_t transitions between a finite number of states $z \in \{1, \ldots, Z\}$, where state z at time t affects a conditional probabilistic model for y, $y_t \mid z_t$. The probability p of z_{t+1} being in state j at time $t+1$ depends only on the current state i at time t, i.e. $p_{i,j} = P(z_{t+1} = j \mid z_t = i)$. Notably, $p_{i,j}$ is independent of the state of z at time steps prior to t, meaning the "memoryless" assumption of the Markov property holds. Collecting the probabilities $p_{i,j}$ across all possible combinations of i and j results in a "transition matrix". For example, for an HMM with three states, this would look like:

$$\mathbf{p} = \begin{pmatrix} p_{1,1} & p_{1,2} & p_{1,2} \\ p_{2,1} & p_{2,2} & p_{2,3} \\ p_{3,1} & p_{3,2} & p_{3,3} \end{pmatrix}. \tag{4.3.0.1}$$

Here, each row represents the current state of z, and each column represents the possible future states, and the entries give the probabilities of transitioning from one state to another. For instance, $p_{i,j} = 0$ means a transition from state i to state j is impossible, and the other extreme $p_{i,j} = 1$ means it is guaranteed. Given a set of observed $\boldsymbol{y}$, we can learn about these transition probabilities to infer how likely transitions between states are, which can be based on prior information about the system where needed.

The conditional model for the variability of interest y, $y_t \mid z_t$, can be influenced by the state of z, usually through parameters in the model for y_t depending on z_t. For example, we could assume a Normal(μ_t, σ) model for y_t where the mean parameter μ_t is different depending on which state z is in. Typically, we assume that y_t are independent of one another given z_t (they are conditionally independent), i.e. we assume that all of the temporal dependence is accounted for by the hidden quantity z.

As an illustrative example of an HMM, consider a scenario where the probability of a person carrying an umbrella (U) or not (N) each day depends on the weather belonging to one of three states: sunny (S), cloudy (C), or rainy (R). The transition matrix is given below, and the HMM is illustrated graphically in Figure 4.2.

$$\mathbf{p} = \begin{pmatrix} \kappa_{S,S} & \kappa_{S,C} & \kappa_{S,R} \\ \kappa_{C,S} & \kappa_{C,C} & \kappa_{C,R} \\ \kappa_{R,S} & \kappa_{R,C} & \kappa_{R,R} \end{pmatrix}, \tag{4.3.0.2}$$

Here, $\kappa_{i,i}$ is the transition probability of remaining in state i when in the current state i and $\kappa_{i,j}$ is the transition probability of moving to state i when in the current state j. For example, $\kappa_{C,R}$ is the probability of transitioning from the cloudy state to the rainy state.

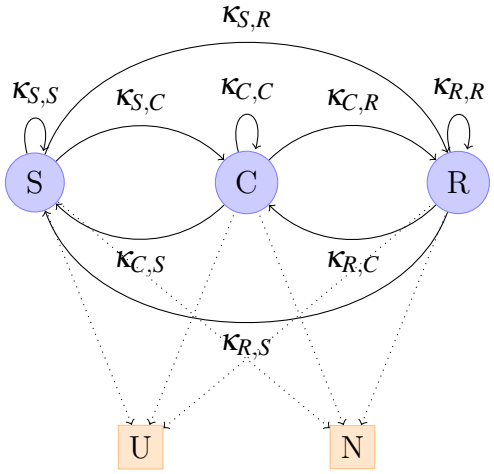


Figure 4.2: Example HMM with three weather states (sunny (S), cloudy (C), or rainy (R)) that affect the probability of whether a person carries an umbrella (U) or not (N) each day.

Here, the HMM can capture temporal structure in the umbrella-carrying behaviour through the hidden state sequence. By fitting this HMM, we can learn about the transition probabilities, infer the most likely sequence of weather states given the observed umbrella-carrying behaviour, simulate entirely new weather state sequences, or predict beyond the end point in time of the existing data.

A key challenge in designing HMMs is choosing the number of distinct states (Z), which often requires careful consideration of the problem in hand, in combination with trialling a few different values. Nonetheless, HMMs offer a flexible framework for capturing underlying temporal dynamics driving observed data, e.g. disease spread and seasonality, potentially without the need to include such mechanisms explicitly in the model as covariates/inputs.

In the domain of disease modelling, HMMs have been widely employed to monitor the progression of diseases at both individual and population levels. Nkemnole et al. (2023) apply an HMM to monitor the transmission of Lassa Fever, a viral disease in West Africa. Here, the authors use a susceptible-infected-recovered (SIR) model to construct the transition matrix, yielding insights into disease transmission patterns based on information from previous outbreaks. Notably, the estimated transition matrix indicated high probabilities of transitioning to the infected state from susceptibility and sustaining infection, and more generally the HMM approach was demonstrated to be a valuable tool for early detection and response to Lassa fever outbreaks. Meanwhile, Watkins et al. (2009) implement an HMM approach aiming to enhance the early detection of localised disease outbreaks. By developing a Bayesian HMM tailored for sparse small area count data, they assessed the performance of the HMM against established surveillance algorithms. This demonstrated that the HMM method was particularly effective with respect to low "false alarm" rates.

More recently, with the emergence of the COVID-19 pandemic at the start of 2020, researchers have used HMMs to analyse spatio-temporal COVID-19 data and gain insights into the disease's dynamics. One such study, conducted by Zhou et al. (2021) use HMMs to capture and model the complex dynamics of the COVID-19 pandemic over both space and time. Through the HMM framework, the authors aimed to understand better the virus's progression. There has also been some limited exploration of HMMs in the context of individual COVID-19 variants. For instance, Tahir et al. (2023) conduct a comparative analysis of the influence of COVID-19 variants, including the beta, gamma, delta, and omicron strains, on the behaviour of SARS-CoV-2 in both vaccinated and unvaccinated individuals. The goal

was to understand and forecast how effectively T-cells respond to the emerging COVID-19 variants in pre- and post-vaccinated individuals. The study compared a Bayesian neural network with the HMM, demonstrating similar performance in the prediction and classification of T-cell epitopes from SARS-CoV-2 among the different COVID-19 variants.

In summary, HMMs provide a versatile approach to modelling sequential data, with applications ranging from forecasting future trends to aiding decision-making in public health. For a full detailed introduction to HMMs, see Rabiner et al. (1986).

4.3.1 Compositional applications

To date, we believe use of HMMs for compositional time series data has been under-explored in the literature. One noteworthy exception is Fisher et al. (2022), which introduces a Dirichlet-HMM framework for detecting and modelling shifts in a time series of continuous proportions, measuring parts of a whole. As explored in Chapter 2, Section 2.3.2, the Dirichlet is a commonly used family of distributions for a vector of continuous proportions $\boldsymbol{x}$ summing to a constant, i.e. $\sim \text{Dirichlet}(\boldsymbol{\alpha})$. Here, $\boldsymbol{\alpha}$ is a vector of parameters, the same length as $\boldsymbol{x}$, which determines the mean and variance of $\boldsymbol{x}$. We can reparameterise $\boldsymbol{\alpha}$ as $\boldsymbol{\alpha} = \boldsymbol{\mu}\boldsymbol{\phi}$, such that $\boldsymbol{\mu}$ is the mean of $\boldsymbol{x}$ and $\boldsymbol{\phi}$ controls the variance.

In Fisher et al. (2022), the location and scale parameters ($\boldsymbol{\mu}$ and $\boldsymbol{\phi}$) of the Dirichlet are given by generalised linear models with coefficients that change depending on the HMM state, allowing for multiple regimes in the time series. This approach allows for modelling seasonality, trends, and inclusion of covariate information, while also detecting change points (i.e.

as characterised by HMM state transitions). The framework was assessed through simulation and demonstrated success when applied to lake phytoplankton data, with the authors suggesting that their approach has potential applications in various fields beyond ecology, such as economics and geography.

However, this approach is held back by the limited flexibility of the assumed Dirichlet distribution. A notable constraint of the Dirichlet is that the covariance/correlation between any pair of compositions (y_i, y_j) is strictly negative (Stoner et al., 2020b). Additionally, considering the Dirichlet expressed in terms of location and scale parameters $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$, note that the univariate/scalar $\boldsymbol{\phi}$ uniformly scales the variance of all components of $\boldsymbol{y}$. This means that the Dirichlet offers no flexibility to capture differing variances of the individual components of $\boldsymbol{y}$, or to capture more unusual covariance structures. This limitation motivates a new general framework that integrates HMM structures within a more flexible family of distributions for compositional data, which we will detail in the next section.

4.4 Proposed General Framework

The Generalised-Dirichlet-Multinomial (GDM) framework proposed in Stoner et al. (2020b) was shown to be an effective modelling tool for compositional count data. Here, the Generalised-Dirichlet (GD) distribution (Wong, 1998) extends the Dirichlet distribution (the limitations of which we noted in Section 4.3.1) by replacing the univariate/scalar variance parameter $\boldsymbol{\phi}$ with a set of variance parameters $\boldsymbol{\phi}$, the length of which is one less than the number of components in $\boldsymbol{y}$, allowing for more flexibility to capture different covariance structures encountered in real-world data problems.

Let y_t be some total count of interest arranged over discrete time points $t = 1,\ldots,N$, where y_t can be broken down into a series of V count compositions, $x_{t,v}$ ($\boldsymbol{x}_t = x_{t,1},\ldots,x_{t,V}$), such that $y_t = \sum_{v=1}^{V} x_{t,v}$. The GD distribution is given as:

$$\boldsymbol{p}_t \sim \text{GD}(\boldsymbol{\alpha};\boldsymbol{\beta}). \tag{4.4.0.1}$$

Here, the GD acts as an additional source of variability, introducing more degrees of freedom to capture more fully the variability of the compositions. The GD distribution is constructed as a series of independent scaled Beta distributions, for the proportions $p_{t,v}$ given $p_{t,1},\ldots p_{t,v-1}$:

$$\begin{aligned}
p_{t,1} &\sim \text{Beta}(\alpha_1,\beta_1),\\
\frac{p_{t,2}}{1-p_{t,1}} \mid p_{t,1} &\sim \text{Beta}(\alpha_2,\beta_2(1-p_{t,1}),\\
&\vdots\\
\frac{p_{t,v-1}}{1-\sum_{j=1}^{v-2} p_{t,j}} | p_{t,1},\ldots p_{t,v-1} &\sim \text{Beta}\left(\alpha_{v-1},\beta_{v-1}\left(\sum_{v=1}^{v-2} p_{t,v}\right)\right),\\
p_{t,v} &= 1-\sum_{v=1}^{D-1} p_{t,v},
\end{aligned} \tag{4.4.0.2}$$

where α_v and β_v are the shape parameters of the respective Beta distributions. The last component here is inferred from the Beta distributions for the other components.

We define the Multinomial distribution for the compositions counts $\boldsymbol{x}_t$ as:

$$\boldsymbol{x}_t | \boldsymbol{p}_t, y_t \sim \text{Multinomial}(\boldsymbol{p}_t, y_t), \tag{4.4.0.3}$$

for the compositional proportions $\boldsymbol{p}_t$ at time t and y_t the total count at time t. The Multinomial can be viewed as a series of conditional Binomial distributions:

$$\begin{aligned} x_{t,1}|p_{t,1},y_t &\sim \text{Binomial}(p_{t,1},y_t),\\ x_{t,2}|x_{t,1},p_{t,2},y_t &\sim \text{Binomial}(p_{t,2},y_t-x_{t,1}),\\ &\vdots\\ x_{t,v}|x_{t,v-1},p_{t,d},y_t &\sim \text{Binomial}\left(p_{t,d},y_t-\sum_{j=1}^{D-1}x_{t,j}\right). \end{aligned} \tag{4.4.0.4}$$

Mixing the GD with the Multinomial family of distributions yields the GDM, a flexible family of distributions for modelling compositional count data. This introduces greater flexibility for the compositional counts than a Multinomial distribution would provide alone. The GDM framework assumes that $\boldsymbol{x}_t$ arise from a Generalised-Dirichlet-Multinomial distribution, given the total y_t:

$$\boldsymbol{x}_t \sim \text{GDM}(\boldsymbol{\nu},\boldsymbol{\phi},y_t), \tag{4.4.0.5}$$

which can be expressed as a series of Beta-Binomial distributions. The GDM is parameterised in terms of $\boldsymbol{\nu}$ and $\boldsymbol{\phi}$, the mean and variance parameters for the series of conditional Beta-Binomial models for each count composition up to and including $x_{t,V-1}$ (the last count composition $x_{t,V}$ is given implicitly as $y_t-\sum_{v=1}^{V-1}x_{t,v}$), deriving directly from the GDM Stoner et al. (2020b):

$$\begin{aligned} x_{t,1}|y_t &\sim \text{Beta-Binomial}(\nu_1,\phi_1,r_{t,1}=y_t),\\ x_{t,2}|y_t,x_{t,1} &\sim \text{Beta-Binomial}(\nu_2,\phi_2,r_{t,2}=y_t-x_{t,1}),\\ &\vdots\\ x_{t,v}|y_t,x_{t,v} &\sim \text{Beta-Binomial}\left(\nu_v,\phi_v,r_{t,v}=y_t-\sum_{k<v}x_{t,k}\right). \end{aligned} \tag{4.4.0.6}$$

This is derived from the hierarchical mixture of scaled Beta distributions from the GD and a series of conditional Binomial distributions from the Multinomial. The interpretation of ν_v is the expected proportion of the remainder $r_{t,v}$ of y_t that will be made up of $x_{t,v}$ once counts before v in the order of compositions have been subtracted. These have previously been called "relative proportions".

Building on this, we then assume that each pair of ν_v and ϕ_v changes over time as a function of some latent HMM state sequence $z_{1,v},\ldots,z_{N,v}$, e.g. $\nu_{t,v} = \gamma_v(z_{t,v})$ and $\phi_{t,v} = \omega_v(z_{t,v})$. This is the most general form of our proposed GDM-HMM and tailoring for specific applications will occur in the design of the functions $\gamma_v(\cdot)$ and $\omega_v(\cdot)$, and/or in the design of the transition matrices. For example, $\gamma_v(\cdot)$ could be exclusively driven by the hidden state sequence, could include covariate effects, or could also include random effect terms.

Generally, the total counts y_t can be treated as fixed inputs or modelled using another layer of the hierarchy, for example Stoner et al. (2020b) model y_t with a Negative-Binomial. Moreover, where data are continuous compositional time series instead of counts, an equivalent Generalised-Dirichlet HMM model (GD-HMM) could be obtained by replacing the Beta-Binomial conditional models in Equation (4.4.0.6) with Beta conditional models.

We will set the general framework above within the Bayesian paradigm, which allows for complex hierarchical structures and provides rich posterior predictive inference, which is useful in the context of potential use of the GDM-HMM for simulation or forecasting.

4.4.1 Implementation

All code used to apply the framework was written and run using R (R Core Team, 2021) and the model was implemented using the *NIMBLE* package (Valpine et al., 2017). Recall from Chapter 3, NIMBLE is a facility for highly flexible implementation of MCMC models. Moreover, all computations were carried out on an Apple MacBook Air laptop with an Apple M3 chip (8 physical cores) and with 16GB system memory.

Our general framework is a model hierarchy consisting of a conditional Beta-Binomial model for each count composition $(x_{t,v} \mid z_{t,v}, y_t, x_{t,<v})$, an associated HMM for $z_{t,v}$, and optionally a further modelling layer for y_t - though we do not study the case where the latter is included here. Since $z_{t,v}$ is hidden/unobserved, it can be treated as an unknown quantity to be inferred. Within an MCMC implementation, this would typically involve using a categorical sampler to obtain posterior predictive samples of $z_{t,v}$. However, an alternative is to integrate out the unobserved $z_{t,v}$. This sacrifices the opportunity to store posterior predictive samples of $z_{t,v}$, which may or may not be useful depending on the application, but often reduces the computational complexity of fitting the overall model (Stoner et al., 2020a).

An analytical solution to integrating out the latent state sequence for HMMs is the forward algorithm (Scott, 2002), a dynamic programming technique used in HMMs to compute marginal likelihoods efficiently. The forward algorithm operates by recursively calculating the probability of reaching each state at each time step while taking into account the entire sequence of observations up to that point. Beginning with the initial state probabilities and transitioning through the model states based on transition probabilities, the algorithm accumulates probabilities of reaching each state at each time step. Simultaneously, it incorporates the likelihood of each state generating the observed variable of interest (in this case $x_{t,v}$). The algorithm avoids sampling the hidden states z_t, making the forward algorithm

more efficient, particularly for longer time series or larger state spaces. Here, we adapted software from Stoner et al. (2020a) that implements the forward algorithm within NIMBLE into a series of functions, which taken together allow for efficient computation of marginal likelihoods for the GDM-HMM.

The GDM-HMM is implemented within NIMBLE as a series of Beta-Binomials (one for each variant), from Equation (4.4.0.6). For each Beta-Binomial time series (i.e. for each composition), we compute a scalar marginal joint likelihood through the new custom function: "`dhmm_betabinomial`", given in Listing 4. We give `dhmm_betabinomial` the following inputs: `x` is the vector of count values for this composition over the entire time series (missing values are not permitted); `r` is the vector of count values for the remainder $r_{t,v}$ of y_t once counts before v in the order of compositions have been subtracted (as in Equation (4.4.0.6)), over the time series; `p0` is a vector of the initial state probabilities; `p` is the relevant transition matrix. Then, `N` is the number of time points; `S` is the number of states; ν and ϕ are the mean and variance parameters for the Beta-Binomial, respectively. The final input for any NIMBLE distribution function must always be `log`, which is an integer determining whether the function should return the output at the log scale or not - here, this argument is inoperative and the output is always returned on the log scale.

Within the `dhmm_betabinomial` function, `dens` is a matrix of probability density values, where the rows are the time points $t = 1,\ldots,N$ and the columns are the states $s = 1,\ldots,S$, such that `dens[t,s]` is the probability density for $x_{t,v}$ for the latent HMM quantity $z_{t,v}$ in state s. The values within `dens` are computed using the custom function "`dbetabinomial`", given in Listing 5 and obtained from Stoner et al. (2020c), which evaluates the log probability density for the Beta-Binomial. This function is required as NIMBLE does not currently contain built-in Beta-Binomial distribution functions.

```
dhmm_betabinomial = nimbleFunction(

    run = function(x = double(1),
                   r = double(1),
                   p0 = double(1),
                   p = double(2),
                   N = double(0),
                   S = double(0),
                   nu = double(1),
                   phi = double(1),
                   log = integer(0)) {

        # Initialise likelihood matrix
        dens = matrix(nrow = N, ncol = S)

        # Loop over states
        for(s in 1:S){

            # Loop over time points
            for (t in 1:N) {

                # Compute the likelihoods
                dens[t, j] = exp(dbetabinomial(x[t],
                                               nu[s],
                                               phi[s],
                                               r[t],
                                               log = TRUE))
            }
        }
        # Declare scalar output
        returnType(double(0))

        # Run the forward algorithm
        return(forward_alg(p0, p, N, S, dens))
    })
```

Listing 4: Custom `R NIMBLE` code for the `dhmm_betabinomial` function, which evaluates the marginal joint likelihood for the Beta-Binomial time series, integrating out the latent HMM quantity.

```
dbetabinomial = nimbleFunction(

    run = function(x = double(0),
                   nu = double(0),
                   phi = double(0),
                   y = double(0),
                   log = integer(0)){

        # return scalar
        returnType(double(0))

        # Upper limit to ensure stability in distribution
        phi <- min(phi, 1e+04)

        if(x >= 0 & x <= y){
            return(lgamma(y + 1) +
                   lgamma(x + nu * phi) +
                   lgamma(y - x + (1 - nu) * phi) +
                   lgamma(phi) -
                   lgamma(y + phi) -
                   lgamma(nu * phi) - lgamma((1 - nu) * phi) -
                   lgamma(y - x + 1) -
                   lgamma(x + 1))
        } else {return(-Inf)}
    })
```

Listing 5: Custom `R` "NIMBLE function" code for the `dbetabinomial` function, obtained from Stoner et al. (2020c), which computes the log probability density function for the Beta-Binomial.

The matrix of probability density values, `dens`, is then passed to the `forward_alg` function from Stoner et al. (2020a), given in Listing 6, which implements the forward algorithm to compute the marginal joint likelihood for the whole time series. The inputs to `forward_alg` are: the initial state probabilities `p0`; the transition matrix `p`; the number of time points `N`; the number of states `S`; and the matrix of the probability density values `dens`. Here, `columnsum` is a simple auxiliary function to compute the column sums of a matrix.

```
forward_alg = nimbleFunction(

    run = function(p0 = double(1),
                   p = double(2),
                   N = double(0),
                   S = double(0),
                   dens = double(2)){

        c = numeric(N)

        c[1] = sum(dens[1, ] * p0)

        alpha = (dens[1, ] * p0) / c[1]

        for(t in 2:N){

            delta = dens[t, ] *
                    columnsum(matrix(rep(alpha, S), ncol = S) * p)
            c[t] = sum(delta)
            alpha = delta / c[t]

        }

        returnType(double(0))
        return(sum(log(c)))
    }
)
```

Listing 6: `R` "NIMBLE function" code for the `forward_alg` function from Stoner et al. (2020c), which implements the forward algorithm.

We assessed convergence of MCMC chains for all models by following the procedure outlined in Appendix B. In summary, this includes visual inspection of traceplots and computing the PSRF (Gelman et al., 1992) for each parameter.

4.5 Application of the GDM-HMM to COVID-19 Variant Data

In this section, we apply our proposed GDM-HMM to a dataset of the weekly COVID-19 disease case counts by variant. As discussed in Section 4.1, this case study serves as the motivation for developing and evaluating new methodology for compositional time series, in this case featuring: (i) compositional count structure; (ii) diverse patterns of spread and mutation between variants and countries; and (iii) non-smooth temporal structure featuring rapid emergence of new variants and replacement of existing strains.

4.5.1 Data

The data studied in this chapter was sourced from GISAID (Khare et al., 2021) and contain COVID-19 case counts aggregated by country and week and dis-aggregated by variant. The data include recorded counts from the first known mutation of the SARS-CoV-2 virus onwards, and not the original strain of the virus - the L-strain (Vellingiri et al., 2020). Case counts are given for each week of data from 09/02/20 to 21/05/23, in addition to the total number of cases across all variants within that country that week. The dataset includes counts from 217 countries across the globe. In the original data, as it is available to us with posterity, we have a complete time series of counts for each variant and country, where there have been non-zero cases recorded for that variant in that country at some point in the time series. Where there are no recorded cases for a variant in a given country, there are instead no counts for that variant (i.e. there are no zero counts in the data to represent the lack of cases). In such situations, we add columns of zero counts for the variants and countries with no recorded cases, ensuring that each country has a complete time series of counts for all the variants.

Throughout the pandemic, the WHO, classified certain strains as either a variant of concern (VOC) or a variant of interest (VOI), based on a number of factors such as transmissibility and severity of disease. The emergence of new VOCs heightened unease, often prompting additional public health measures to be implemented at national levels. Table 4.1 outlines the five VOCs from this dataset defined by WHO, along with the date and country of their first detection. A more comprehensive examination of COVID-19 variants can be found in Gong et al. (2023). Due to the low counts observed across the VOIs, we combined them into a single aggregated VOI category.

Table 4.1: The COVID-19 variants of concern (VOC) as defined by the World Health Organisation (WHO).

WHO Name	**Lineage**	**Date of first detection**	**Country of first detection**
Alpha	B.1.1.7	Sep 2020	United Kingdom
Beta	B.1.351	May 2020	South Africa
Gamma	P.1	Nov 2020	Brazil
Delta	B.1.617.2	Oct 2020	India
Omicron	B.1.1.529	Nov 2021	South Africa

As an illustrative example, the first panel of Figure 4.3 shows the recorded time series of case counts by variants for the United Kingdom. This shows that cases of the alpha variant emerged in around late 2020, with a sharp rise around December 2020 - noting that the alpha variant originated in the United Kingdom in September 2020. The British Broadcasting Corporation reported that, in mid-December, it was estimated that almost 60 percent of cases in London involved alpha, highlighting its dominance (BBC News, 2020). Following this, the variant recedes, as the virus was constantly mutating over time and new variants were emerging. In around June 2021 it can be seen that the delta variant became the dominant variant spreading around the United Kingdom. Trobajo-Sanmartín et al. (2022) found that the delta variant was more transmissible than previous variants, especially among young adults. The Guardian reported in June 2021 that the new delta variant was causing more than 90% of all new COVID-19 cases in the United Kingdom. This dominance lasted until around December 2021 when omicron was first detected in the UK, with cases of omicron quickly soaring to case levels not seen previously. Omicron quickly became the most dominant variant across the world, prompting restrictions aiming

to curb the spread. Torjesen (2021), published in Nov 2021, stated that omicron may be more transmissible than other variants and partly resistant to existing vaccines, leading to the sudden domination of this variant. The decline of omicron from June 2022 onwards is unique and interesting too. The cases seem to decrease but then continue to fluctuate over the next 12 months. This is different from the previous three dominant variants which decrease and disappear completely, due to being replaced by a new VOC. The lifespan of omicron was by far the greatest of all the variants worldwide. In addition to the original counts, we can examine the corresponding proportions of each variant, as displayed in the second panel of Figure 4.3. Here we can clearly see that, in the UK, each new emerging VOC eventually dominates new cases, as indicated by percentages of cases reaching 100%. Moreover, in this case we can see a pattern of increasing duration across the alpha, delta and omicron variants.

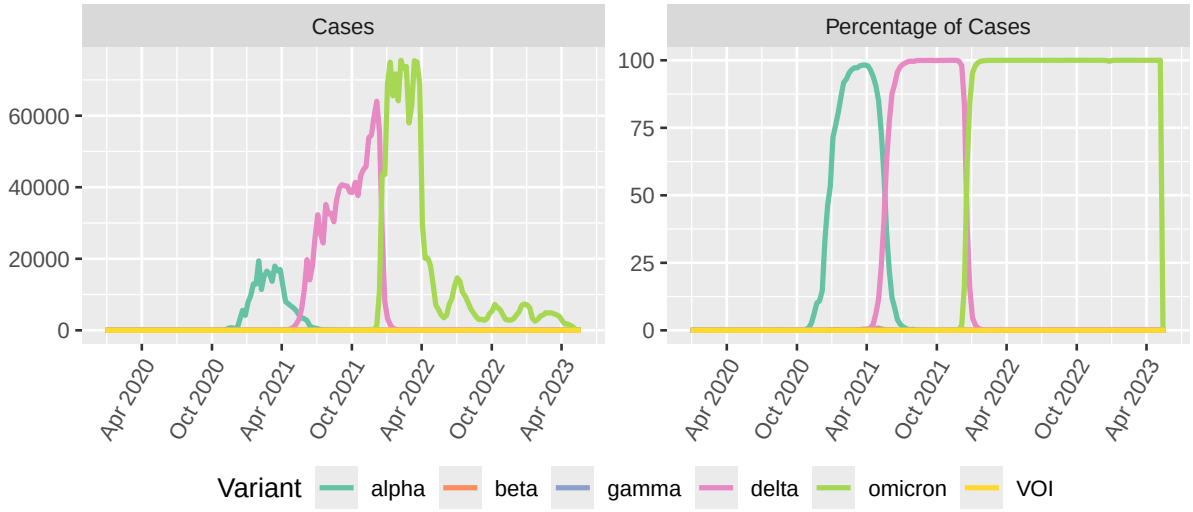


Figure 4.3: Time series of the weekly COVID-19 case count and percentage of cases attributed to each variant in the United Kingdom. The different coloured lines correspond to each of the five COVID-19 variants of concern: alpha, beta, gamma, delta and omicron; and the aggregated variants of interest count.

This was only a descriptive analysis for one country; modelling the evolution of case counts attributed to these variants over time and across a large number of countries could help us understand the characteristics of these variants (e.g. how quickly they tend to dominant). These insights could contribute to the overall understanding for the complex evolution of this pandemic and, in doing so, offer insights for future public health policy. Achieving

this requires a modelling framework that can capture the non-smooth temporal variation inherent to these data, including the rapid domination of new variants of concern. Naturally, we apply the novel GDM-HMM approach to investigate its potential as a candidate for this. Meanwhile, the aim of understanding the average characteristics of variants across countries motivates some level of pooling of information, which we discuss in the next subsection.

4.5.2 Clustering approach

During the pandemic, the trajectory of the different variants had unique patterns across continents and countries. This was dependent on where the variant originated, any travel restrictions or movement between countries and the population of the country. To pool information across groups of similar countries, for the purpose of estimating shared HMM parameters that capture the expected characteristics of each variant, we carried out a clustering exercise prior to fitting the model. Establishing clusters also allows for the exposition of hierarchical structures within our proposed general framework, as we will explore in Section 4.5.3.

We aimed to cluster countries based on the evolution of the variants over time. Since we are considering 169 time points and 6 variants per country, we have 1014 counts per country. Hence, we have more variables than we have countries to cluster. These counts will also have a strong correlation with nearby time points, potentially causing computational instabilities and/or inefficiencies in the clustering algorithm.

Instead, we used Generalised Additive Models (GAMs) (Wood, 2017) to reduce the dimensionality of the time series. Our approach is similar to that presented in Dejean et al. (2007) and Iorio et al. (2016), which use a spline model combined with a clustering algorithm on the resulting spline coefficients. By adopting this strategy, we aimed to address the com-

putational challenges posed by the large volume of observations in the time series, while also seeking to capture effectively underlying patterns within the data. Splines are piecewise polynomial functions that are smoothly joined together at specified points known as knots. A penalised regression spline approach offers a flexible approach to fitting non-linear covariate effects (with time being one possible example), without relying on strict assumptions about the degree of smoothness a priori. Here, we used the `gam` function from the *mgcv* package (Wood, 2003) to fit a one-dimensional thin plate regression spline of time to the log of the case counts plus 0.01, e.g. $\log(x_{t,v,m} + 0.01)$, with 10 knots, using the default Gaussian distribution with an identity link function, separately for each variant and country. In general, the chosen number of knots (10 in this case) represents an upper limit in the flexibility in the smooth function, and the actual degree of smoothness within this limit is determined by the smoothing penalty parameter, seeking to find an optimal trade-off between in-sample and out-of-sample predictive skill (Wood, 2017). We also fit spline models to the total COVID-19 case counts (the sum of all variants) for each country, with an offset for the total population size of that country. With this offset included, the spline then captures the change over time in the overall disease rate per capita. Each spline had nine coefficients plus an intercept term for the seven models (each VOC, the combined VOIs, and the total case count), resulting in 70 terms in total for each country, which we stored in a 217 x 70 matrix.

We then applied hierarchical clustering (as outlined in Chapter 3, Section 3.3.2.1) to the spline coefficients, using Euclidean distance and Ward linkage. Based on the elbow method (Chapter 3, Section 3.4.4) we identified three clusters of countries: Cluster 1 with 86 countries, Cluster 2 with 40 countries, and Cluster 3 with 91 countries.

Figure 4.4 shows the cluster each country belongs to on a world map which shows similar patterns of countries across each cluster. For example, high income countries in Europe, North America and Australia are clustered into Cluster 3, which contains the largest number of countries. Cluster 1 contains a large portion of Africa, though this cluster also includes countries from Asia such as China and India. Cluster 2 is the the smallest cluster which has a less obvious geographic structure, combining countries from all six continents such as Algeria, Argentina, Dominican Republic, Norway, New Zealand and Pakistan.

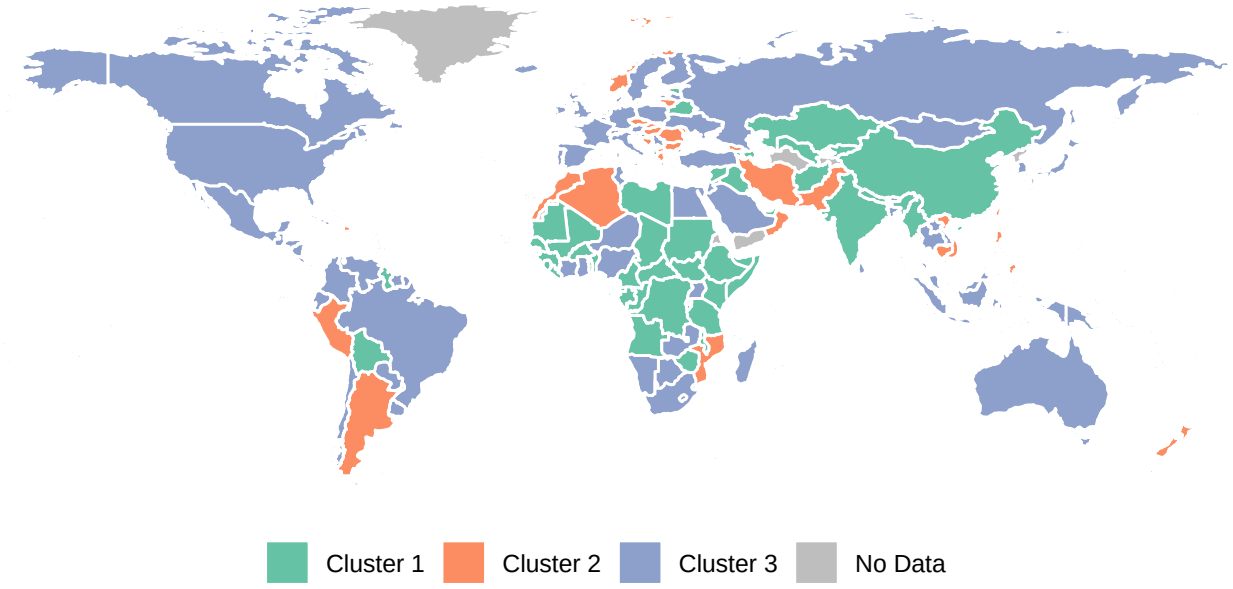


Figure 4.4: World map illustrating the three clusters of the COVID-19 variant data, produced from hierarchical clustering on the spline coefficients. The different clusters correspond to: Cluster 1, Cluster 2 and Cluster 3. No Data displays countries where no COVID-19 variant data are present.

To limit computational complexity whilst developing the new methodology, we decided to only examine ten countries from each cluster, as given in Table 4.2. We selected these specific countries to include a wide geographical range spanning multiple continents in each cluster. The time series of COVID-19 case counts for the countries in each cluster are shown in Figures 4.5 to 4.7 (one figure per cluster) with each of the five variants displayed showing the progress of the disease and variant composition over time.

Table 4.2: The 10 countries selected from each of the three clusters produced from hierarchical clustering on the spline coefficients.

Cluster 1	**Cluster 2**	**Cluster 3**
Armenia	Argentina	Australia
Azerbaijan	Bulgaria	Brazil
China	Dominican Republic	Canada
Dominica	Greece	France
Estonia	Jamaica	Germany
Fiji	Morocco	Italy
India	New Zealand	Mexico
Kenya	Pakistan	South Africa
Qatar	Peru	Spain
United Arab Emirates	Philippines	United Kingdom

First, we highlight some notable features of each cluster. In Figure 4.5, we can see that all the countries in Cluster 1 lack any recorded cases of gamma in these data, with many of them also having no instances of the beta variant. The countries in Cluster 2 (Figure 4.6) generally have very high omicron case levels, in relation to the other variants, persisting for well over a year from December 2021, when it was first detected. Many of these countries appear to have multiple peaks of the omicron variant, i.e. in Pakistan, where cases seem to decrease then cases soar again occurring twice. A number of the countries in Cluster 2 only have cases for four of the five VOCs recorded in the data but it is not always the same four, with some having no beta or gamma variant cases.

Lastly, countries in Cluster 3 (Figure 4.7) also tend to have large fluctuating numbers of omicron cases. The countries in Cluster 3 also generally had high case counts of the delta variant shortly before omicron took over. Alpha, the variant that originated in the UK, is also highly prevalent in most countries in Cluster 3 in the early stages of the time series. This is to be expected given the movement of people between the United Kingdom and Europe. Looking across the three clusters, omicron represents the majority of cases in most countries. This is to be expected, as since 2021, the omicron variant drove weekly case numbers to record levels worldwide, unlike any previous VOC (Taylor, 2022).

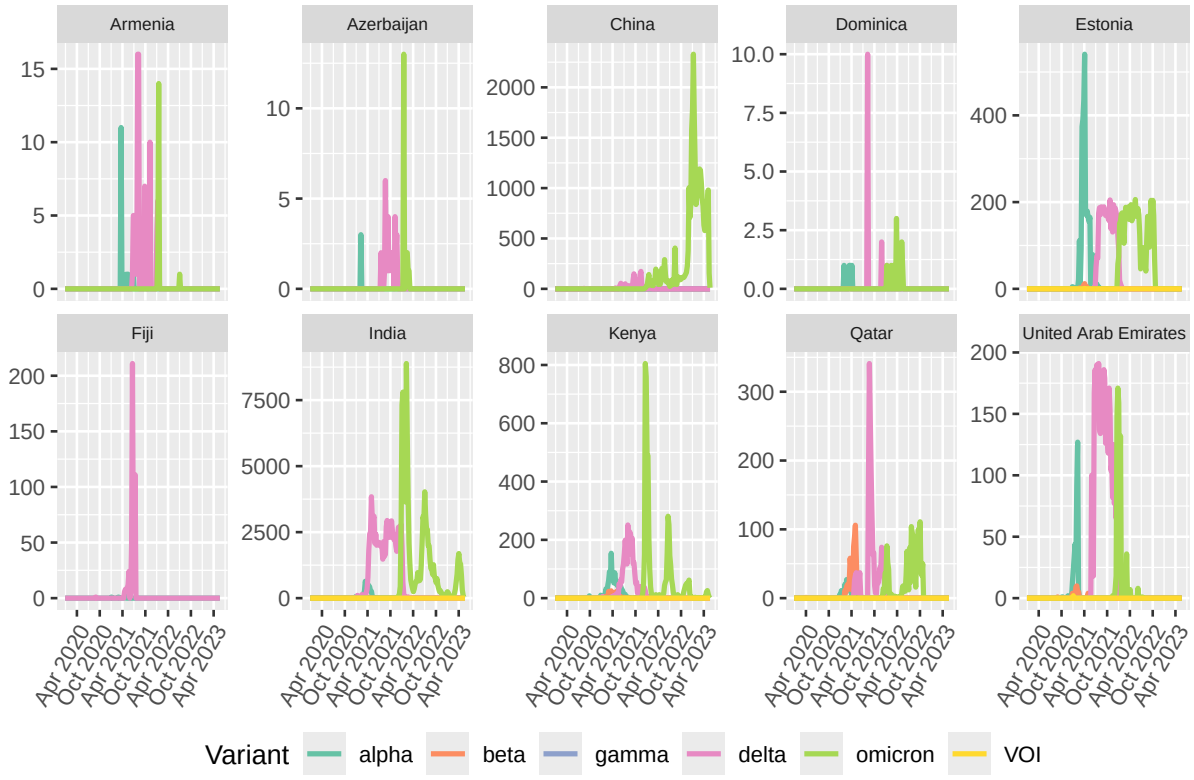


Figure 4.5: Time series of the weekly COVID-19 case count attributed to each variant for the 10 selected countries from Cluster 1 (Table 4.2). The different coloured lines correspond to each of the COVID-19 variants: alpha, beta, gamma, delta and omicron.

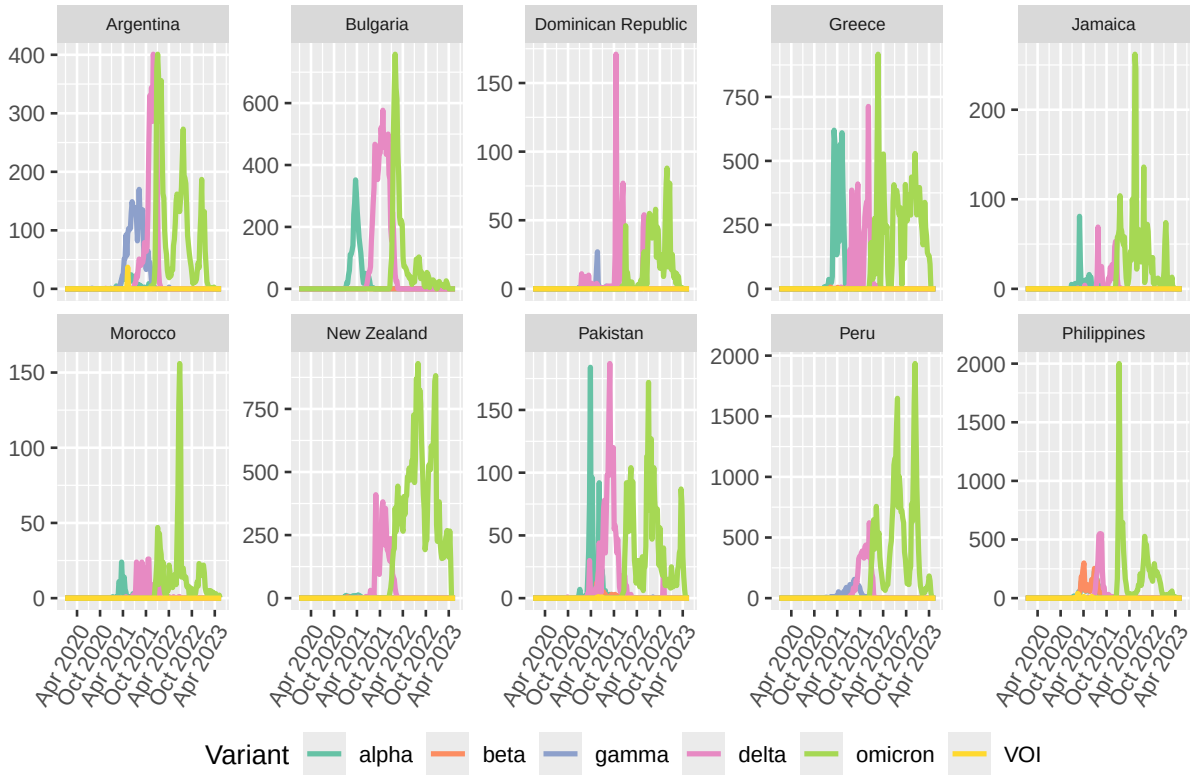


Figure 4.6: Time series of the weekly COVID-19 case count attributed to each variant for the 10 selected countries from Cluster 2 (Table 4.2). The different coloured lines correspond to each of the COVID-19 variants: alpha, beta, gamma, delta and omicron.

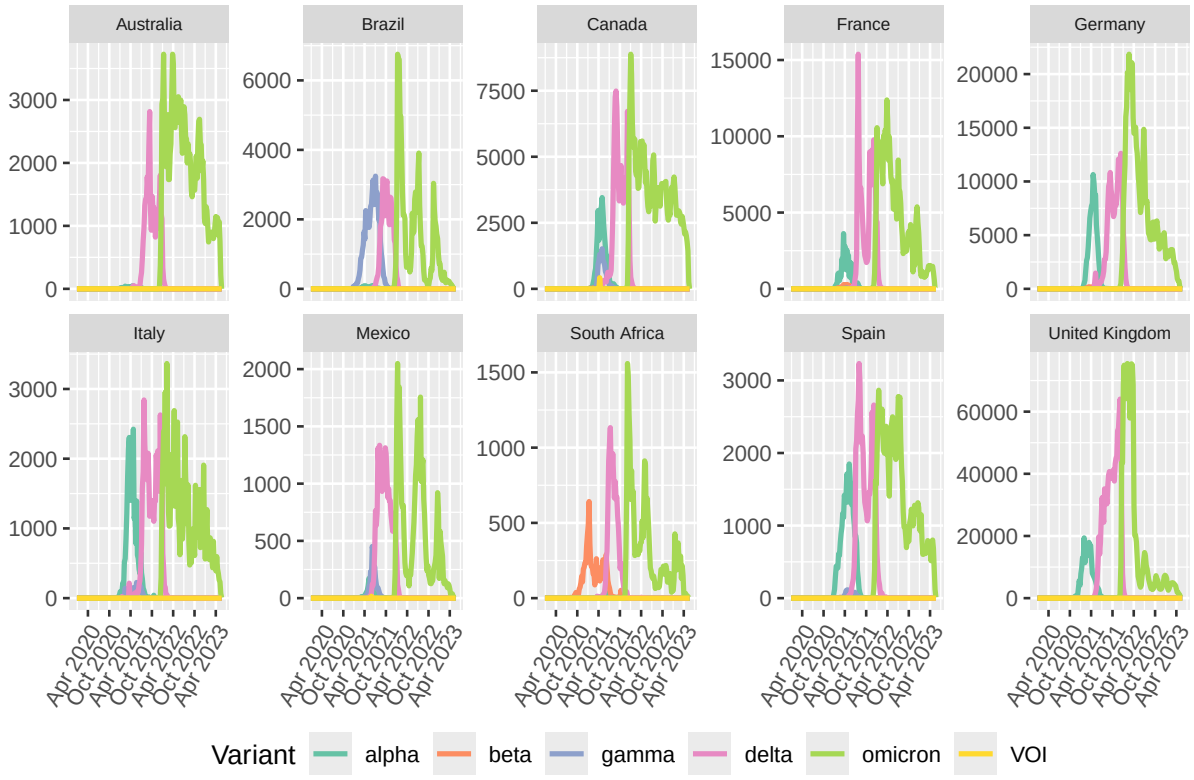


Figure 4.7: Time series of the weekly COVID-19 case count attributed to each variant for the 10 selected countries from Cluster 3 (Table 4.2). The different coloured lines correspond to each of the COVID-19 variants: alpha, beta, gamma, delta and omicron.

4.5.3 Model for COVID-19 Variants

Here we adapt the framework outlined in Section 4.4 into a tailored model for the COVID-19 variants. Let $y_{t,m}$ be the total COVID-19 cases recorded for week $t = 1,\ldots,N$ in country $m = 1,\ldots,M$, and let $x_{t,v,m}$ be the corresponding case count for variant $v = 1,\ldots,V$ from Table 4.1 ($\boldsymbol{x}_{t,m} = x_{t,1,m}, x_{t,2,m}, x_{t,3,m}, \ldots, x_{t,V,m}$), such that $y_{t,m} = \sum_{k=1}^{V} x_{t,k,m}$. We assume that the temporal structure in $\boldsymbol{x}_{t,m}$ can be captured by a hidden Markov state sequence for each variant up to and including $V-1$, $z_{1,v,m}, z_{2,v,m}, \ldots, z_{N,V-1,m}$. Recall from Section 4.4 that the last composition V is modelled implicitly and does not have an associated Beta-Binomial model.

As in the general framework, we assume a GDM model for $\boldsymbol{x}_{t,m}$ given the total COVID cases $y_{t,m}$, with the mean and variance parameters of the Beta-Binomial conditional models given by $\nu_{t,v,m}$ and $\phi_{t,v,m}$, respectively. To characterise the temporal evolution of each variant, we designed an HMM where $z_{t,v,m}$ progresses sequentially through five states, given in Table 4.3.

Table 4.3: Hidden state sequence ($\boldsymbol{z}_t$) for the GDM-HMM model for COVID-19 variants.

State	Description
State 1	Dormant before outbreak
State 2	Active, Increasing
State 3	Dominant
State 4	Active, Decreasing
State 5	Dormant after outbreak

The HMM is constrained such that $z_{t,v,m}$ can only move "forward" through the 5 states; it cannot revert to a previous state. For example, if the variant is dominant at the current time (State 3), it cannot return and be dormant before the outbreak as the virus is in circulation. It can, however, move forward to the decreasing state (State 4) and onto the last state, dormant after the outbreak (State 5), when the cases return to zero. This design is intended to reflect the behaviour seen in the variant case counts, as shown in Figures 4.5 to 4.7.

The constrained design of this HMM is constructed through the below transition matrix $\mathbf{p}_{v,c}$, for variant $v = 1,\ldots,V-1$ and country cluster $c = 1,\ldots,C$, where most entries/probabilities are equal to 0:

$$\mathbf{p} = \begin{pmatrix} 1-\kappa_{1,2,v,c} & \kappa_{1,2,v,c} & 0 & 0 & 0 \\ 0 & 1-\kappa_{2,3,v,c} & \kappa_{2,3,v,c} & 0 & 0 \\ 0 & 0 & 1-\kappa_{3,4,v,c} & \kappa_{3,4,v,c} & 0 \\ 0 & 0 & 0 & 1-\kappa_{4,5,v,c} & \kappa_{4,5,v,c} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where $\kappa_{i,j,v,c}$ defines the probability of moving from state i to state j, with assumed priors $\kappa_{i,j,v,c} \sim \text{Beta}(1,1)$. For the country cluster c, the initial state probability vector q_c is given as $q_c \sim \text{Dirchlet}(35,1,1,1,35)$. This means that, a priori, the probability of $z_{1,v,m}$ (the first week) being in each of the five states is $(0.35,0.1,0.1,0.1,0.35)$, i.e. a priori we believe it is more likely for the variant to be in one of the two dormant states (State 1 or 5), though posterior inference for q_c will learn from the observed data.

The HMM latent states drive changes in the GDM model through $\nu_{t,v,m} = \gamma_{t,v,m}(z_{t,v,m})$, where $\gamma_{t,v,m}(z_{t,v,m})$ is either fixed a priori or an unknown quantity to be inferred, depending on the state of $(z_{t,v,m})$. Using Equation (4.4.0.5) and (4.4.0.6) from Section 4.4 the extra conditions are placed for the model for the COVID-19 variants:

$$\begin{aligned} &\gamma_{v,s,m} = -10, && \text{for } s = 1,5; \\ &\gamma_{v,s,m} \sim N(0,1), && \text{for } s = 2,\ldots,4; \\ &\phi_{v,s,c} \sim \text{Gamma}(2,0.05). \end{aligned} \tag{4.5.3.1}$$

Constraining the first and last $\gamma_{v,s,m}$ to be equal to -10 means that corresponding mean parameters $\nu_{v,s,m}$ are very small values. The motivation for this is that when the variant is in the first or last state and is dormant, as shown in Table 4.3, i.e. the counts are 0 and the variant is not yet or no longer present in that country. A stronger constraint would be to set

$\nu_{v,s,m} = 0$ in the first and last states, but this means that the Beta-Binomial probability mass function is equal to 0 for non-zero counts, which causes the forward algorithm (Section 4.4.1) to return a log joint posterior value of negative infinity (since the values of the latent HMM quantity $z_{t,v,m}$ are not known a priori, negative infinity log probabilities spoil the calculation by design). For the remaining states, i.e. where the variant is generating non-zero counts, $\gamma_{v,s,m}$ can take on a value from a Normal distribution. For all states, we assume a Gamma prior for the GDM variance parameters $\phi_{v,s,c} > 0$.

A final constraint on $\gamma_{v,s,m}$ is needed to ensure identifiability between the three active states - increasing/dominant/decreasing states (States 2-4); issues with identifiability are common in HMMs due to "label switching" (recall from Chapter 3, Section 3.3.3), where the roles of different states swap such that the overall model doesn't change (Stoner et al., 2020a). Here, we address label switching through hard parametric constraints in the model, which is straightforward in this case due to the strong physical interpretation of our proposed five states: since State 3 is intended to capture periods where the variant is dominant, we strictly enforce in the MCMC algorithm that samples are rejected where $\gamma_{v,3,m} > \max(\gamma_{v,2,m}, \gamma_{v,4,m})$. This means that, for a given variant and country, $\nu_{v,s,m}$ is always highest in the dominant state (State 3). We also enforce that $\gamma_{v,2,m} > -10$ and $\gamma_{v,4,m} > -10$, to avoid conflicts with the dominant states.

For the GDM-HMM, we ran four chains in parallel for 2,000 MCMC iterations, discarding the first 1,000 as burn-in. The time taken to run this model was approximately 1 hour 47 minutes. We computed the PSRF for each parameter and 95% of the PSRFs were less than or equal to 1.05, with a median of 1.00, indicating convergence.

4.6 Results

Here we will present and discuss outputs and results from the GDM-HMM, demonstrating how parameter inference from this model can potentially show what variant characteristics are persistent both within and across clusters.

First, Figure 4.8 presents boxplots of the posterior samples of $\nu_{v,s,m}$, by COVID-19 variant and state, for a different country from each cluster (United Arab Emirates, New Zealand and United Kingdom, respectively). By examining these boxplots, we can see how $\nu_{v,s,m}$ varies across the different states for each variant. Overall, a higher $\nu_{v,s,m}$ value for a given variant translates to a higher expected proportion of COVID cases due to that variant. For simplicity, we can focus on the posterior medians (the horizontal black lines in the middle of the box plots) as our point estimates. The spread of the boxplots relates to the posterior uncertainty for each $\nu_{v,s,m}$, e.g. the posterior interquartile ranges (posterior 50% credible intervals) are indicated by the widths of the box plots.

Firstly, we can see for all clusters and variants, as outlined in Section 4.5.3, the fixed constraint of $\nu_{v,s,m}$ to be close to zero for States 1 and 5. As an illustrative example, we will now present the findings shown for the two most dominant VOCs: delta and omicron.

Delta was initially detected in India, from Cluster 1. Across all clusters, we can see that delta became highly transmissible during the dominant state (State 3) which is represented by a high posterior medians of all three clusters which is very close to 1. This means that during this time, the delta variant was contributing to almost all the COVID-19 cases across all countries. During this state, each boxplot across all clusters has a very small boxplot width indicating that there is little uncertainty in the posterior range. For Cluster 1 and 2, the posterior median is notably higher during the active increasing state (State 2), 0.4 and 0.5,

respectively, highlighting the transmissibility of the delta variant within these clusters from its initial detection. However, during the decreasing state (State 4) all the point estimates across each cluster are significantly reduced, with values of 0.16, 0.19 and 0.07 for each cluster respectively, each with a minimal interquartile range. This indicates that the prevalence of the delta variant decreased following its peak. This could be due to the emergence of the next variant overlapping with the evolution of delta.

The last VOC present in the data is omicron, which was first found in South Africa, a country in Cluster 3. Across all three clusters, a similar picture is presented where omicron quickly contributed to most of the COVID-19 cases. The posterior medians for the active increasing state (State 2) are 0.50, 0.49 and 0.63, respectively. This is significantly greater than the point estimates for State 2 for any of the other four VOCs. The supremacy of omicron is extremely evident in the dominant state (State 3) with all three clusters exhibiting considerably high posterior medians of 0.9, 0.98 and 1, respectively. Each boxplot in this state has a very small width signifying little uncertainty in the high values of $\nu_{v,s,m}$. Within the decreasing state (State 4) the posterior medians remain high, greater than 0.5, for all three clusters. However, the posterior interquartile range is at its largest for this state in comparison with the other VOCs, indicating that the uncertainty in the decreasing state is much higher. This could be representing the fluctuating pattern detected from Figure 4.5 to 4.7 after the peak of omicron.

In summary, these figures highlight the variation between the evolution in each of the variants. Here, the contrasts across the clusters are evident, underlining the need to account for the different cluster structures within the HMM.

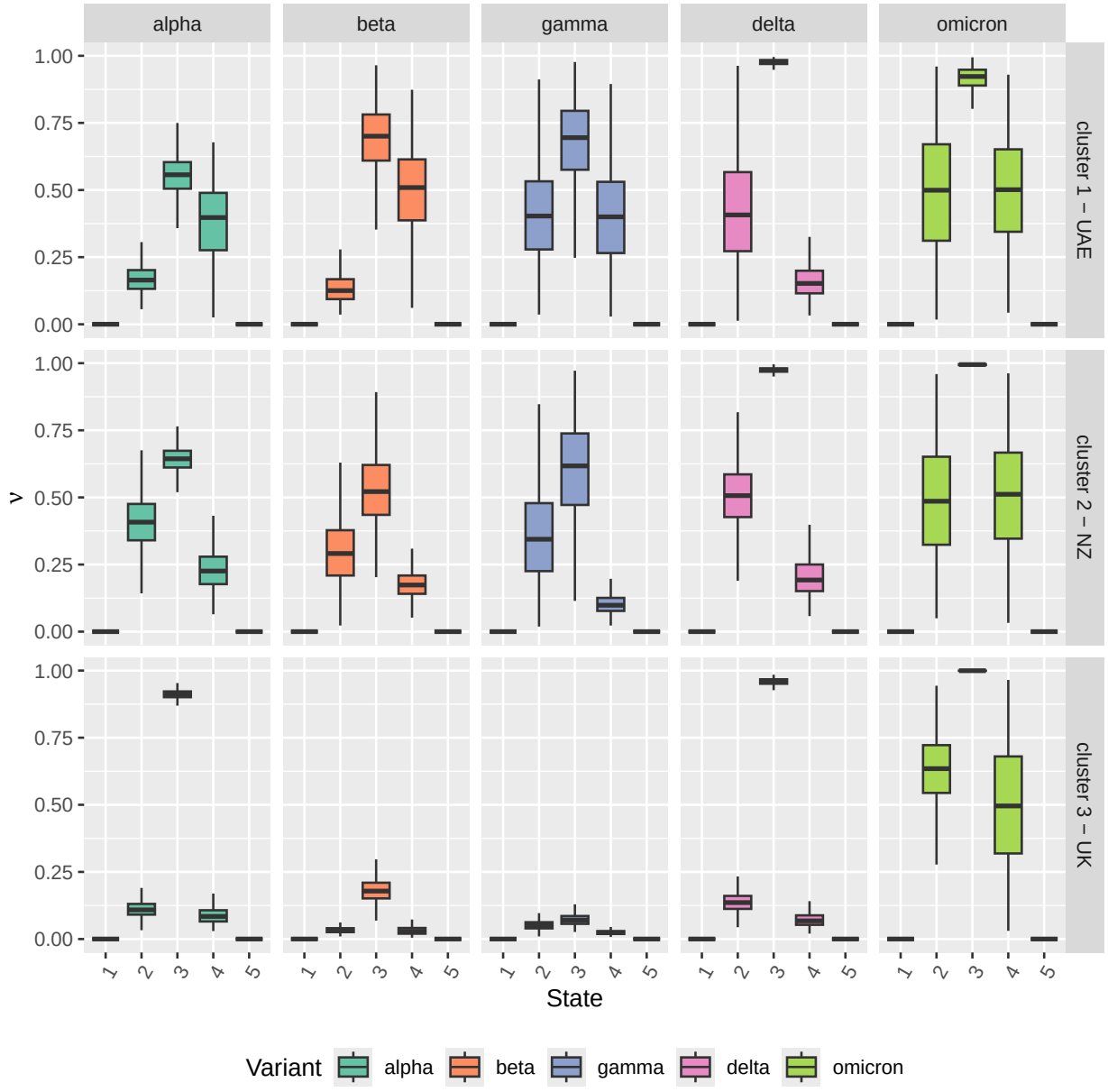


Figure 4.8: Boxplots of the posterior samples for the Beta-Binomial parameter $\nu_{v,s,m}$ for each of the five HMM states for a country from each of the three clusters: United Arab Emirates (Cluster 1), New Zealand (Cluster 2) and United Kingdom (Cluster 3). The different coloured boxplots correspond to each of the COVID-19 variants: alpha, beta, gamma, delta and omicron.

Another model output from the GDM-HMM framework that may provide insights into the characteristics of the different variants is inference based on the transition probabilities $\kappa_{i,j,v,c}$. We define the persistence lengths $L_{v,s,m}$ as the length of time (weeks) the HMM quantity $z_{t,v,m}$ persists in state s before transitioning to the next state. Since persistence is based on a series of Bernoulli trials with probability $\kappa_{i,j,v,c}$, the mean/expected persistence length is given by $E[L_{v,s,m}] = \kappa^{-1}_{i,j,v,c}$. Figure 4.9 shows the posterior median expected persistence lengths for each variant in the non-dormant HMM states (States 2, 3, and 4), across the different clusters. Here, if a variant has a longer expected persistence in the active state (State 3), this would indicate a longer period of dominance of the COVID-19 cases. Similarly, a longer expected persistence in the decreasing state for one variant could indicate that cases from that variant tend to decline more slowly. We might expect that variants showing both of these traits could have a more prolonged active period following initial emergence.

Notably, the expected persistence lengths tend to be very similar across clusters, despite these being estimated independently with no crossover in the data or model parameters. This provides some reassurance that we can potentially draw useful insights about the general characteristics of the different variants in terms of their progression through the different stages of outbreak. However, we will discuss the limitations of such conclusions in Section 4.8, including the potential confounding effect of the persistence of one variant being dependent on the timing of a new dominant variant emerging.

Here, we only describe in detail the characteristics presented in Figure 4.9 for the most dominant variant. Omicron is perhaps the most distinct variant in terms of its persistence characteristics across the states and clusters. Within the increasing state (State 2), omicron spends significantly more time in Cluster 1 compared to the other two clusters, with a duration of 66 weeks compared to just 3 and 4 weeks in Clusters 2 and 3, respectively. The short expected duration in the increasing states reflect the very rapid transition to the dominant state. The greatest difference, compared to other variants, lies in the expected persistence lengths for the dominant state (State 3), which are 210, 264 and 149 weeks for

each cluster respectively. These expected durations are not only much greater than those for the dominant state of other VOCs - reflecting the extended period during which omicron remained dominant compared to other VOCs detected since the COVID-19 outbreak in late 2019 - but they also have greater variance between three clusters. The higher variance between clusters can also been seen for the decreasing state (State 4), with persistence lengths of 1, 22 and 14 weeks, respectively.

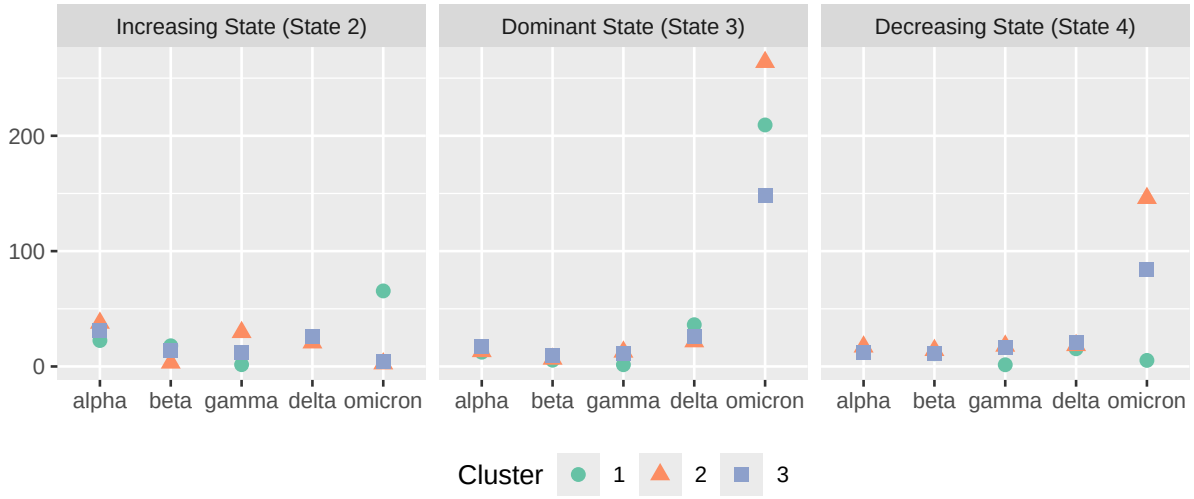


Figure 4.9: Expected persistence length $E[L_{v,s,m}] = \kappa_{i,j,v,c}^{-1}$ (in weeks) for each variant of concern (VOC), for the three active HMM states: State 2 (increasing), State 3 (dominant), State 4 (decreasing). The different coloured shapes correspond to each of the country clusters: Cluster 1, Cluster 2 and Cluster 3.

4.7 Posterior predictive experiment

4.7.1 Alternative methods for comparison

To examine the effectiveness of our proposed approach with respect to capturing non-smooth temporal variation, we chose to compare our GDM-HMM approach against two alternatives. In each, the HMM is replaced with a common structure for temporal dependence, a Random Walk and a Dynamic Linear Model, respectively. The next section explains how we developed these in the context of the COVID-19 variant data. In both comparison models, compositional count zeros are handled using the GDM, as described in Section 4.4; the only difference lies in how each model captures the temporal structure.

4.7.1.1 Random Walk Model (RW)

The Random Walk (RW) model - a simple model commonly used in time series analysis - assumes that each value in a series is a Normal random deviation from the previous observation. The principles of a random walk were first proposed in Pearson (1905). In general, it is defined by the equation:

$$\lambda_t \sim N(\lambda_{t-1}, \sigma^2), \tag{4.7.1.1}$$

where λ_t denotes the value at time t, λ_{t-1} is the value at the previous time $t-1$, and σ^2 represents a random error term. The RW model is memoryless, the range of likely values of λ_t at the next time step only depends on the value at the current time and is independent of any values prior to that.

Random walks are commonly used to model time series directly, but they can also be integrated within hierarchical models as random effects, to capture temporal structure at some latent level – this is how we will use them here.

To replace the HMM within the GDM-HMM for COVID-19 variants detailed in Section 4.5.3, we define the Beta-Binomial means as:

$$\log\left(\frac{\nu_{t,v,m}}{1-\nu_{t,v,m}}\right) = \lambda_{t,v,m}; \tag{4.7.1.2}$$

$$\lambda_{t,v,m} \sim \text{Normal}(\lambda_{t-1,v,m}, \sigma^2_{v,m}). \tag{4.7.1.3}$$

We make several assumptions about the RW part of our GDM-RW, so that it is comparable with our proposed GDM-HMM model. First, we assume that the value of the random walk at time step 1, $\lambda_{1,v,m}$, is a country-level random effect with cluster-level mean $\tau_{v,c}$ and standard deviation $\xi_{v,c}$:

$$\lambda_{1,v,m} \sim \text{Normal}(\tau_{v,c}, \xi^2_{v,c}). \tag{4.7.1.4}$$

This is comparable to assuming that the initial state probabilities p_0 in the GDM-HMM model are also controlled by a cluster-level distribution, with the prior for p_0 accounting for the variation within the clusters. We assume a Normal$(0, 10^2)$ prior for $\tau_{v,c}$ and a Half-Normal$_{>0}(0,1)$ prior (a Normal distribution truncated at 0, to ensure all values are positive) for each $\xi_{v,c}$. The equivalent prior for the variance parameter ϕ is incorporated in the GDM-RW model, ensuring comparability between the models, which vary across variants and clusters. Furthermore, we assume a country-level standard deviation parameter, σ, into the random-walk component of the GDM-RW model.

For the RW model, four chains were run in parallel for 400,000 MCMC iterations, with 300,000 discarded as burn-in and storing every 100^{th} sample. The computation time was approximately 3 hours and 25 minutes. We computed the PSRF for each parameter and 92% of the PSRFs were less than or equal to 1.05, with a median of 1.00, indicating convergence.

4.7.1.2 Dynamic Linear Model (DLM)

As the RW lacks a memory of any trends, either short-term or longer-term. An alternative which has an extra layer of "memory", are Dynamic Linear models (DLM), first denoted as such by Petris et al. (2009), however, different terminology and notation had been developed previously. A DLM can be a powerful framework for modelling time series data, offering flexibility in capturing complex temporal dynamics. DLMs allow for time-varying parameters, enabling the representation of evolving trends within the data. The specific DLM we will use here is given by:

$$\begin{aligned} \lambda_t &\sim N(\lambda_{t-1} + \alpha_t, \sigma_\lambda^2), \\ \alpha_t &\sim N(\alpha_{t-1}, \sigma_\alpha^2), \end{aligned} \tag{4.7.1.5}$$

where, as in the RW model, λ_t denotes the value at time t and λ_{t-1} is the value at the previous time $t-1$. Then, the DLM introduces a trend term α_t, which itself evolves as a random walk. The expected value of λ_t is thus a combination of the previous value λ_{t-1} and α_t, with the latter representing the short-medium term trajectory of the latent process. The standard deviation parameters σ_λ and σ_α determine the relative contributions of the random noise part and the trend term, respectively. Figure 4.10 shows two illustrative time series of simulated λ_t from this simple DLM, one where the relative contribution of σ_λ is larger ($\sigma_\lambda = 10$, $\sigma_\alpha = 0$), and one where the contribution of σ_α is larger ($\sigma_\lambda = 0$, $\sigma_\alpha = 10$). We can see that when the relative contribution of σ_λ is larger, the respective time series has a larger variance of values, thus showing the influence that σ_λ has on the values of λ_t.

Whereas when the relative contribution of σ_α is greater, the values of the time series are more closely related with smaller value for α_t affecting λ_t. With this in mind, note that the DLM reduces to the RW (Equation 4.7.1.1) in the limit that σ_α tends to zero, such that the additional trend term α_t drops out of Equation 4.7.1.5.

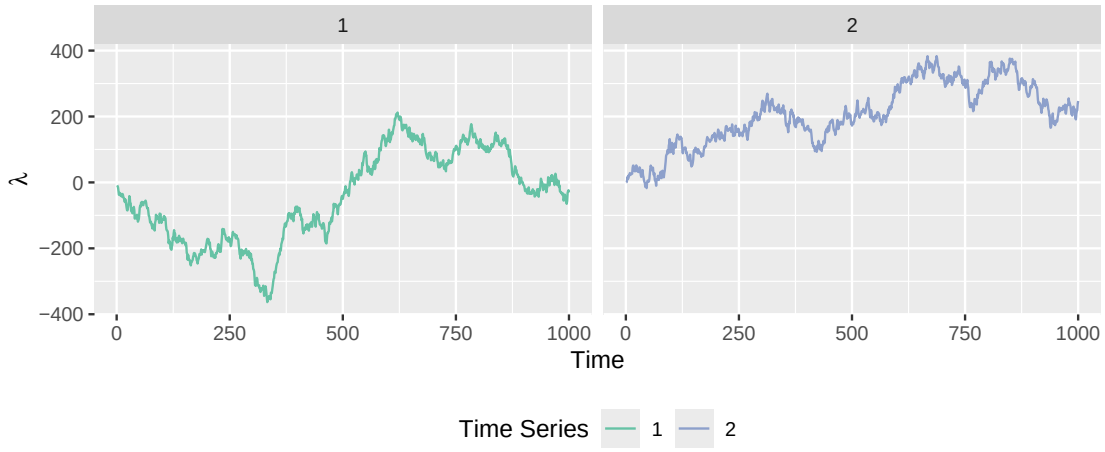


Figure 4.10: Example simulated time series from a simple DLM (Equation 4.7.1.5), where Time Series 1 is generated using $\sigma_\lambda = 10$, $\sigma_\alpha = 0.1$, and Time Series 2 corresponds to $\sigma_\lambda = 0.1$, $\sigma_\alpha = 10$.

As in the GDM-RW model, the DLM captures temporal structure in the GDM-DLM through Equation (4.7.1.2), and then:

$$\lambda_{t,v,m} \sim N(\lambda_{(t-1),v,m} + \alpha_{t,v,m}, \sigma^2_{v,m,\lambda}), \qquad \text{for}(t = 2,\ldots,N); \tag{4.7.1.6}$$

$$\alpha_{t,v,m} \sim N(\alpha_{(t-1),v,m}, \sigma^2_{v,m,\alpha}), \qquad \text{for}(t = 2,\ldots,N). \tag{4.7.1.7}$$

Similar to the GDM-RW model once again, we make several assumptions about the DLM part of the GDM-DLM for comparability with our proposed GDM-HMM model. At the initial time point $t = 1$, we assume that both the latent DLM quantity $\lambda_{1,v,m}$ and trend term $\alpha_{1,v,m}$ are Normal random effects:

$$\lambda_{1,v,m} \sim \text{Normal}(\tau_{v,c,\lambda}, \xi^2_{v,c,\lambda}); \tag{4.7.1.8}$$

$$\alpha_{1,v,m} \sim \text{Normal}(\tau_{v,c,\alpha}, \xi^2_{v,c,\alpha}), \tag{4.7.1.9}$$

where we assume Normal$(0, 10^2)$ priors for $\tau_{v,c,\lambda}$ and $\tau_{v,c,\alpha}$, and Half-Normal$_{>0}(0,1)$ priors for $\xi_{v,c,\lambda}$, $\xi_{v,c,\alpha}$, $\sigma_{v,m,\lambda}$, and $\sigma_{v,m,\alpha}$. As within the GDM-DLM model, the same prior is placed upon the variance prior ϕ to ensure positivity and comparability.

For the DLM model, we ran four chains in parallel for 400,000 iterations, discarding the first 300,000 as burn-in and storing every 100^{th} iteration. The time taken to run this model was approximately 6 hours and 45 minuates. Computing the PSRFs for each parameter resulted in 91% of the PSRFs being less than or equal to 1.05, with a 1.01 median indicating convergence.

4.7.2 Model Checking

Both the GDM-RW and GDM-DLM models offer a lot of flexibility as prior random effect models ($p(\boldsymbol{\lambda}_{v,m})$) to capture different temporal patterns. We would generally expect them both to be capable of capturing non-smooth or otherwise volatile patterns, learning from the data into the posterior $p(\boldsymbol{\lambda}_{v,m} \mid \boldsymbol{x}_{v,m})$. However, this is not a guarantee that such temporal patterns would be reproduced when generating new random effect time series from the posterior predictive distribution $p(\tilde{\boldsymbol{\lambda}}_{v,m} \mid \boldsymbol{x})$. Since $\boldsymbol{\lambda}_{v,m}$ are intended to drive the temporal dependency in the modelling framework, if predicted/simulated values $\tilde{\boldsymbol{\lambda}}_{v,m}$ do not reflect the temporal patterns of the original data well, then forecasted future values or new simulated time series of $\boldsymbol{x}_{v,m}$ are unlikely to either. Thus, we decided to compare our proposed GDM-HMM approach to the GDM-RW and GDM-DLM alternatives through a posterior predictive model checking exercise.

Posterior predictive model checking involves the simulation of new data from the posterior predictive model, e.g. in this case simulating new COVID-19 variant counts. These new simulated data, $\tilde{x}_{t,v,m}$ are called "replicates" of the original $x_{t,v,m}$, because they are generated using the exact same covariate values/inputs as $x_{t,v,m}$, in this case they are generated over the same time steps. We simulate a new replicate data set $\tilde{\boldsymbol{x}}$ of the original $\boldsymbol{x}$ once for each set of saved MCMC samples. For instance, if we have 1,000 MCMC samples for all model parameters, we would simulate 1,000 new replicate sets. We can then compare these to the original data by looking at the discrepancy between either:

- Individual data points $x_{t,v,m}$ and the corresponding distribution of replicates $\tilde{x}_{t,v,m}$, or;
- Summary statistics $S(\tilde{x})$ of the original data versus the distribution(s) of statistic(s) $S(\tilde{\boldsymbol{x}})$ from the replicates. Simple options are the sample mean and sample variance, but $S(\cdot)$ could be any statistic capturing features of the data that are important to us.

Essentially, we can investigate whether individual data points or summary statistics for the original data are an extreme value with respect to the corresponding posterior predictive distributions, e.g. $p(\tilde{x}_{t,v,m}|x_{t,v,m})$ or $p(S(\tilde{x}_{t,v,m})|x_{t,v,m})$. If this is the case, then the model does not capture the data well in this respect.

For each saved MCMC iteration and for each country, the procedure for generating replicate COVID-19 variant case counts from the GDM-HMM model is as follows:

1. Simulate a new time series of HMM latent states $\tilde{\boldsymbol{z}}_{v,m}$ for each variant, using the transition matrix and initial state probabilities corresponding to that MCMC iteration.
2. Simulate new case counts for the first variant from the first Beta-Binomial in Equation 4.4.0.6, using the latent states simulated in Step 1 and the samples of $\boldsymbol{\gamma}_{1,s,m}$ and $\boldsymbol{\phi}_{1,s,m}$ from that MCMC iteration.

3. For the next variant, we compute the remainder counts $r_{t,v}$ (Equation 4.4.0.6), i.e. the remainder of the total COVID cases $y_{t,m}$ not yet accounted for by simulated counts of the previous variants in the ordering. Then, simulate new case counts for this variant from the corresponding Beta-Binomial.
4. Repeat Step 3 until counts have been simulated from all Beta-Binomials in Equation 4.4.0.6. Compute the final count $x_{t,V,m}$ (in this case the aggregate count for all VOIs) as $x_{t,V,m} = y_{t,m} - \sum_{k=1}^{V-1} x_{t,k,m}$.

Then, for each country, the procedure for generating replicates from the GDM-RW and GDM-DLM models is the same for Steps 2 - 4 as above. In this case, Step 1 becomes: simulate new time series of RW/DLM random effects $\tilde{\boldsymbol{\lambda}}_{v,m}$ for each variant, using samples of the parameters associated with either the GDM-RW or GDM-DLM models (e.g. those denoted by τ, ξ, σ), from that MCMC iteration. Following these procedures resulted in 4,000 replicate time series for the GDM-HMM, GDM-RW, and GDM-DLM versions, respectively.

Here, to assess how well the replicate time series of COVID-19 variant counts imitate the non-smooth temporal patterns seen in the original data, we summarised time series variability using a moving window approach. This involves defining a window of length L (e.g. 15 weeks) and continuously moving this window through the replicate time series, generating many overlapping subsets $w = 1, \ldots, W = N - L + 1$ of the data. This approach is illustrated in Figure 4.11 where a simple time series is represented by the red line. The respective moving windows w_W of a length L are denoted by different coloured boxes for each w_W. In this specific example with $N = 12$ and $L = 5$, there are eight moving windows ($w = 1, \ldots, 8$).

We then calculate sample standard deviation values $\omega_w(\boldsymbol{x}) = {}^1/_{L-1}\sum_{t=w}^{w+L-1}(x_t - \bar{x}_w)$ for each window/subset w, where $\bar{x}_w$ is the sample mean of $x_w, \ldots, x_{w+L-1}$. Summarising these standard deviations with some statistic $S(\boldsymbol{x}) = S(\omega_1(\boldsymbol{x}), \ldots, \omega_W(\boldsymbol{x}))$ can then quantify how variable the time series are over different time scales. Here, we compute the means of the standard deviations, which tell us something about how variable the time series are on average, and upper quartiles, which relate to more volatile periods/windows. Given the large proportion of zero values in the data, the lower quartile did not yield informative results.

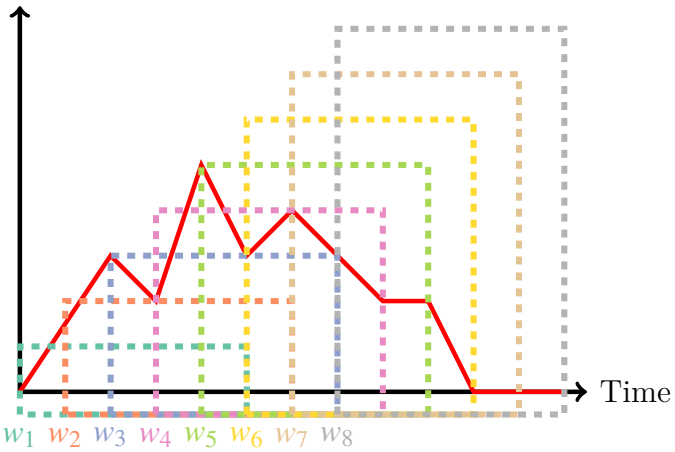


Figure 4.11: Diagram illustrating the moving window approach applied. For a given example time series shown in red, with $N = 12$, the overlapping moving windows ($w = 1, \ldots, W = N - L + 1 = 8$) are generated for a given window length $L = 5$.

Note that the standard deviations of the compositional case counts $\boldsymbol{x}$ within the moving windows will in large part be determined by the scale of the total counts $\boldsymbol{y}$ (i.e. a larger $\boldsymbol{y}$ will increase the variance of $\boldsymbol{x}$). To regularise the variance across countries and variants, we applied the moving window summarisation approach to replicate proportions, calculated from the replicate counts, rather than the counts themselves, making it easier to assess how well each model performs across countries/clusters and variants using global summaries. In this case, we quantify how well replicates from each model capture the temporal structure of the original data through computing the Mean Absolute Error (MAE) (Willmott et al., 2005) of the medians and upper quartiles from the replicates, compared to the means and upper quartiles obtained when applying the moving window summarisation to the original

data. This is the mean magnitude of errors, i.e. $1/n \sum_{i=1}^{n} |S(\tilde{\boldsymbol{x}}) - S(\boldsymbol{x})|$, where n is the total number of replicates and $S(\boldsymbol{x})$ is either the mean or upper quartile of the standard deviations across the windows, for the original data values x or replicate values $\tilde{x}$. We compute these statistics for each variant, each country and for each of the models considered - GDM-HMM, GDM-RW and GDM-DLM. These will tell us how close the simulated replicate values are to the original data, on average, with respect to these statistics. A low MAE value indicates a closer match, which in this case may mean that the non-smooth temporal structure is being captured better.

4.7.2.1 Results

One way to visualise the distribution of the statistics of the replicate data sets is density plots. In each case we plot the logarithm of the statistic (mean or upper quartile), allowing for a clearer visualisation of the distributions. These are given in Figure 4.12 and 4.13, within each the corresponding statistic for the original data are given in each plot by the orange vertical line.

As an example, the row of Figure 4.12 corresponding to Cluster 3 shows multiple modes in the density curve for the mean of the standard deviations of the proportions across all variants. Overall, our proposed GDM-HMM consistently outperforms the GDM-RW and GDM-DLM models in accurately capturing the true value of the data suggesting that the original value of the data is not an extreme value with respect to the posterior predictive distribution. This is evident with the orange vertical line of the original data falling within a peak of the density curve of the GDM-HMM. An exception to this is shown in row of Figure 4.12 for United Arab Emirates (Cluster 1), where for the gamma variant the orange line is further left than the peaks of all three density curves. This could be due to Cluster 1 countries not having any incidences of the gamma variant.

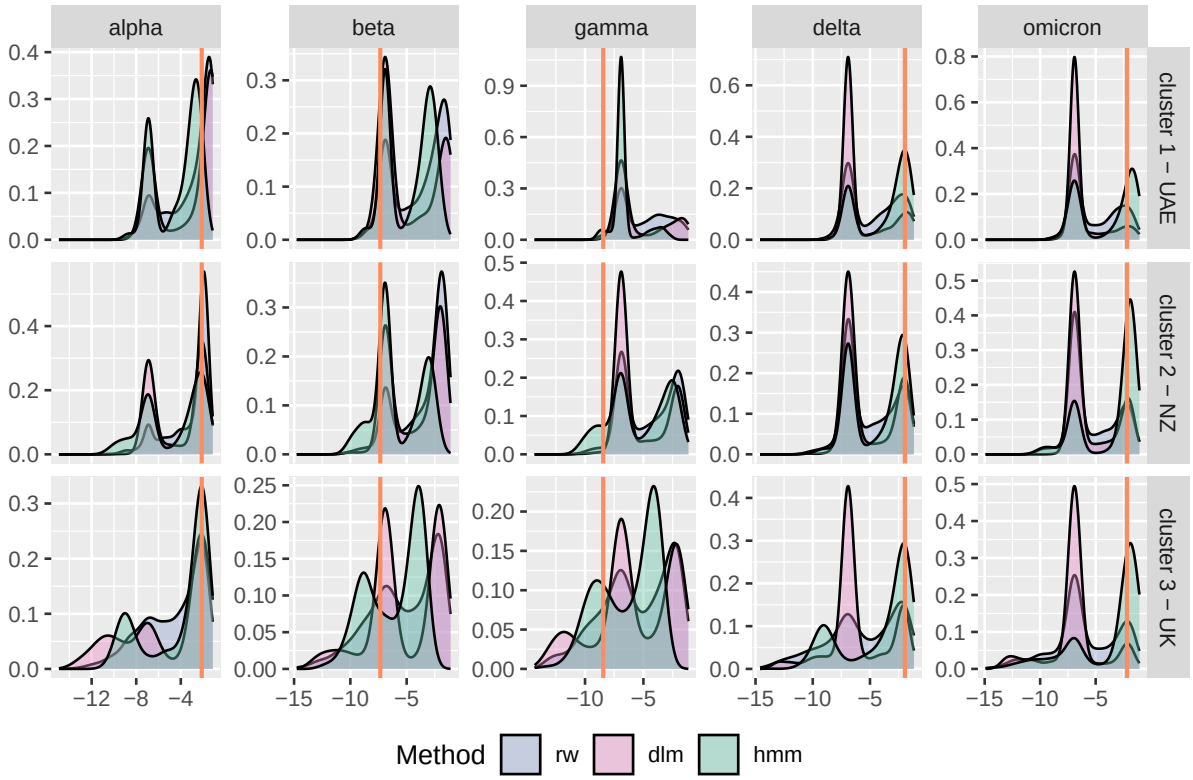


Figure 4.12: Density plots of the log of the mean standard deviation across the windows of length 15 for a country from each of the three clusters: United Arab Emirates (Cluster 1), New Zealand (Cluster 2) and United Kingdom (Cluster 3). The different coloured densities correspond to each of the methods: GDM-HMM, GDM-RW and GDM-DLM.

When assessing the equivalent plot for the upper quartile (Figure 4.13), it is less evident which method performs best. This is apparent as the original data line does not always closely correspond with the peak of a density curve for each of the methods. An example of this is for Cluster 1 where the original data line lies further left then the peaks of the curve for the gamma variant, as seen for the equivalent panel for the mean. However, on the whole, again the GDM-HMM appears to perform best in comparison to the other methods in producing replicates that closely mimic the original data and could be plausible in real-world scenarios. This is most prominently identified for omicron across all three clusters.

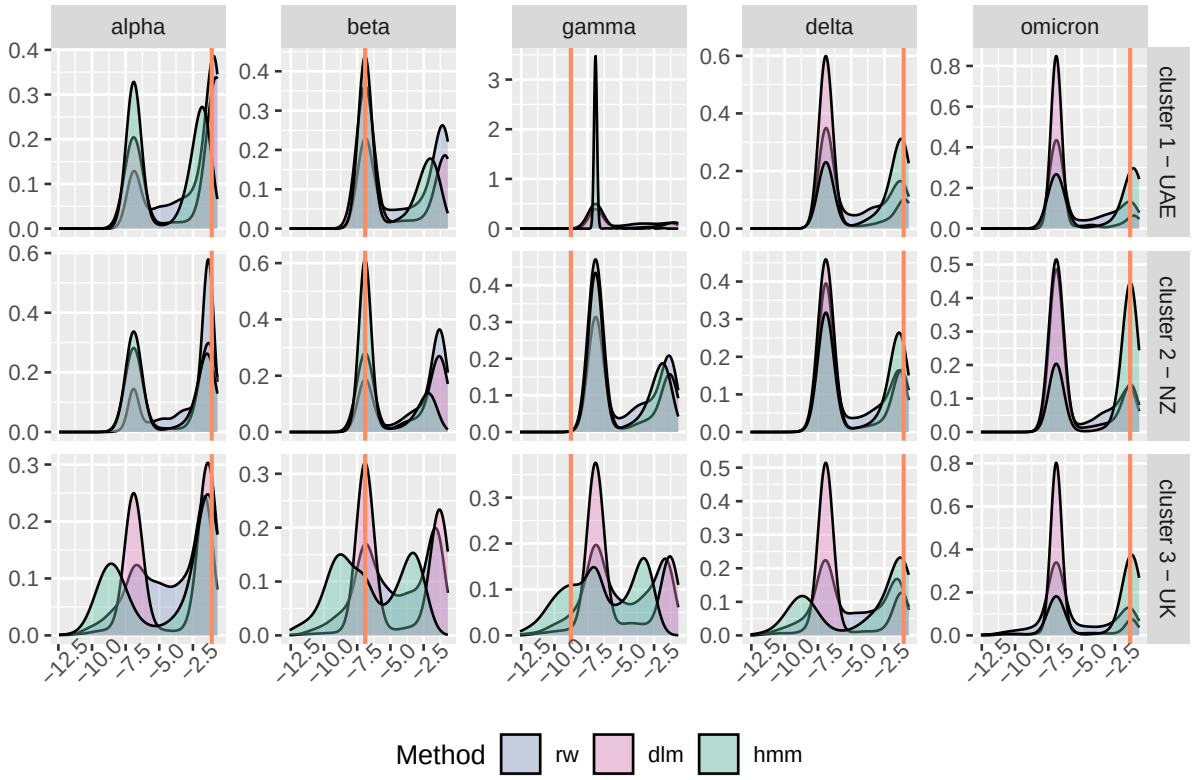


Figure 4.13: Density plots of the log of the upper quartile of the standard deviation across the windows of length 15 for a country from each of the three clusters: United Arab Emirates (Cluster 1), New Zealand (Cluster 2) and United Kingdom (Cluster 3). The different coloured densities correspond to each of the methods: GDM-HMM, GDM-RW and GDM-DLM.

The density plots offer a detailed view into assessing whether the model can accurately capture the true values of the data, suggesting that the original values are plausible within the distribution. However, it is challenging to draw broader conclusions about which method(s) perform best across the different variants and countries. The MAE was computed for all three models for both the mean and upper quartile of the standard deviations of the proportions of the moving windows of length 15. The MAE for each country and variant can be plotted in a barplot alongside the aggregated means across all the countries for each cluster.

Overall, for the mean of the standard deviations of the proportions across the windows, the GDM-HMM has the lowest MAE across all three clusters. The only exception for this is in Cluster 2 (in the second row of Figure 4.14) and the gamma variant, where the GDM-DLM has the lowest MAE at 0.036 whereas the MAE for the GDM-HMM is 12% higher. Although, this is still lower than seen for the GDM-RW model. This highlights that the GDM-HMM is performing better than the other two methods for the mean of the standard deviations of the proportions for the window length 15. A similar picture is seen for the upper quartile of the standard deviations across the windows, with the GDM-HMM consistently having the lowest MAE value across all variants and clusters. Thus, when evaluated using moving windows, the GDM-HMM consistently outperforms both the GDM-RW and GDM-DLM in terms of accurately producing realistic replicate data.

To ensure robustness, various window lengths were investigated, e.g. window lengths 5, 10, 15 and 20. We can visualise the median MAE value across each window length in Figures 4.16 and 4.17. Overall across all window lengths, the GDM-HMM demonstrates optimal performance in regards to minimising the MAE compared to alternative models. In general, as the window length increases this in turn also increases the MAE for all methods.

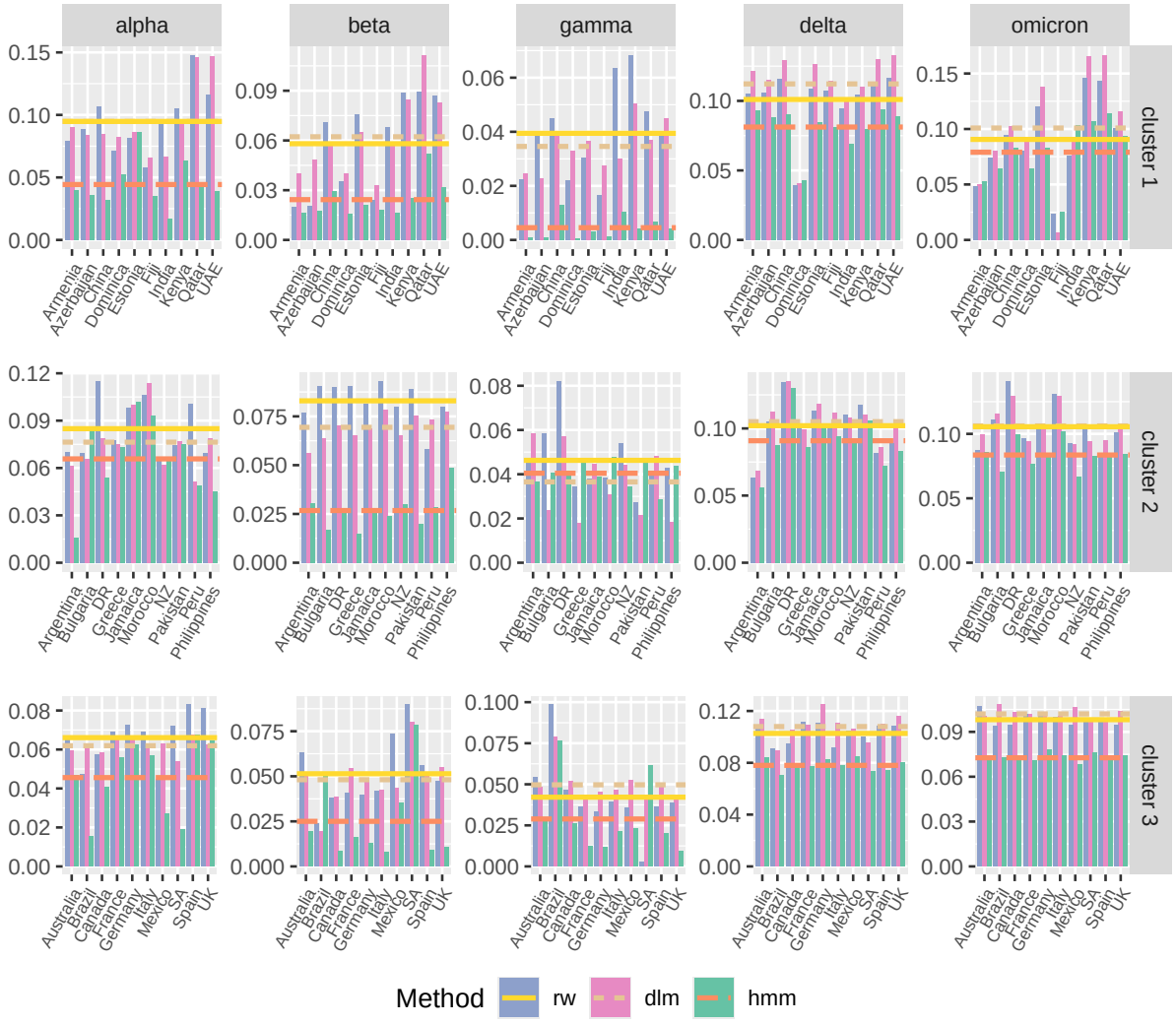


Figure 4.14: Barplots of the mean absolute error (MAE) of the mean standard deviation across the windows of length 15 for the countries selected from each cluster. The different coloured bars correspond to each of the methods: GDM-HMM, GDM-RW and GDM-DLM. The horizontal lines corresponding to the mean value across the clusters with each line corresponding to each of the methods: GDM-HMM, GDM-RW and GDM-DLM.

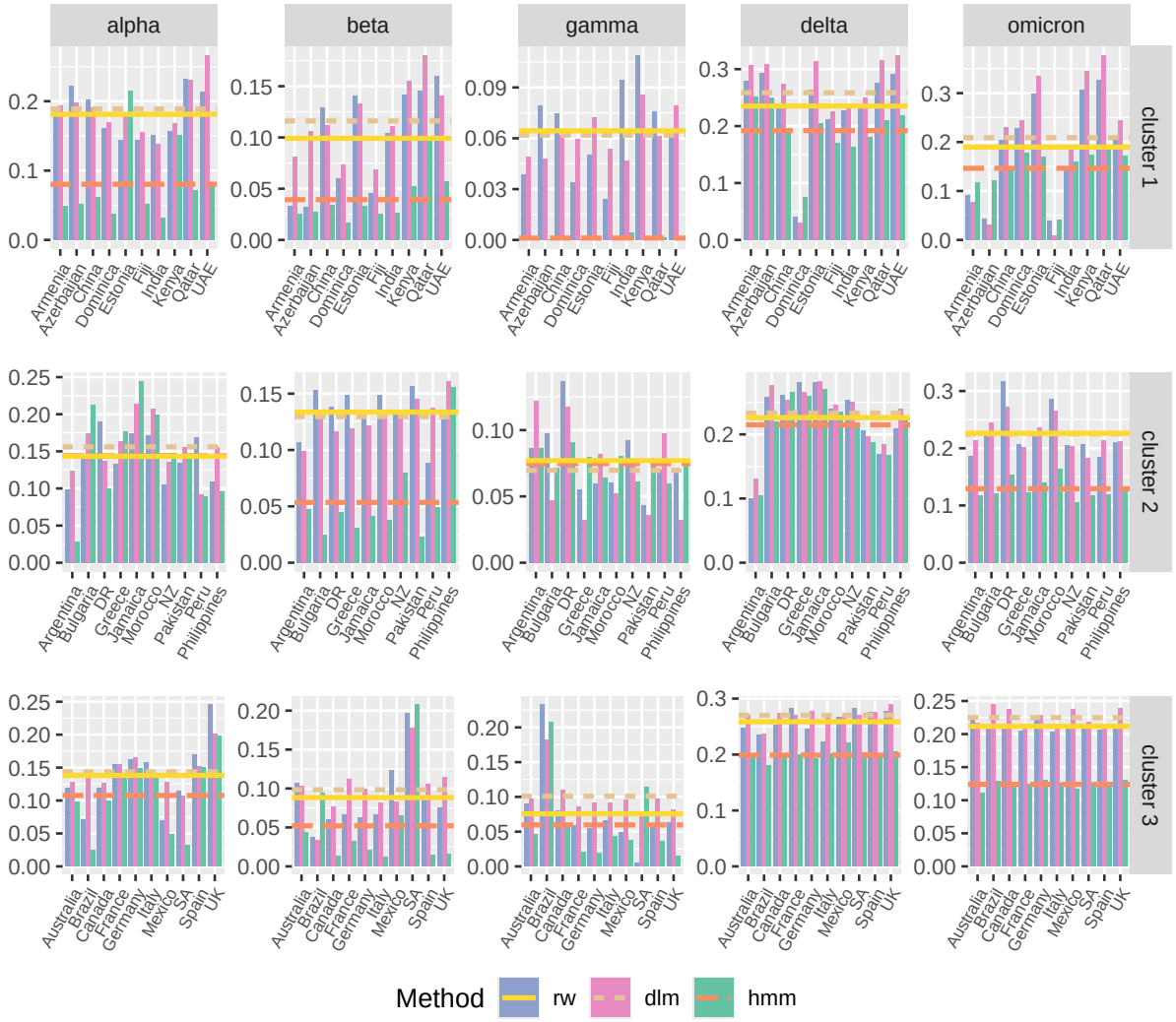


Figure 4.15: Barplots of the mean absolute error (MAE) of the upper quartile of the standard deviation across the windows of length 15 for the countries selected from each cluster. The different coloured bars correspond to each of the methods: GDM-HMM, GDM-RW and GDM-DLM. The horizontal lines corresponding to the mean value across the clusters with each line corresponding to each of the methods: GDM-HMM, GDM-RW and GDM-DLM.

As an illustrative example, for the mean of the standard deviation (Figure 4.16), we can see that for the smallest window length (5) for the omicron variant, the GDM-HMM has the highest MAE but by window length of 10 and larger, the MAE for the GDM-HMM is consistently the lowest. Here, the slope of the GDM-HMM line is much flatter, indicating a more gradual increase in the MAE for each increase of window length in comparison to both the GDM-RW and GDM-DLM lines. This pattern in the slopes holds for most of the variants in each cluster. A deviation to this is seen for the gamma variant within Cluster 2, where the GDM-DLM produces the minimum MAE values across all window lengths. This supports what was presented earlier for the gamma variant in Cluster 2 where the minimum MAE value for this variant occurs for the GDM-DLM.

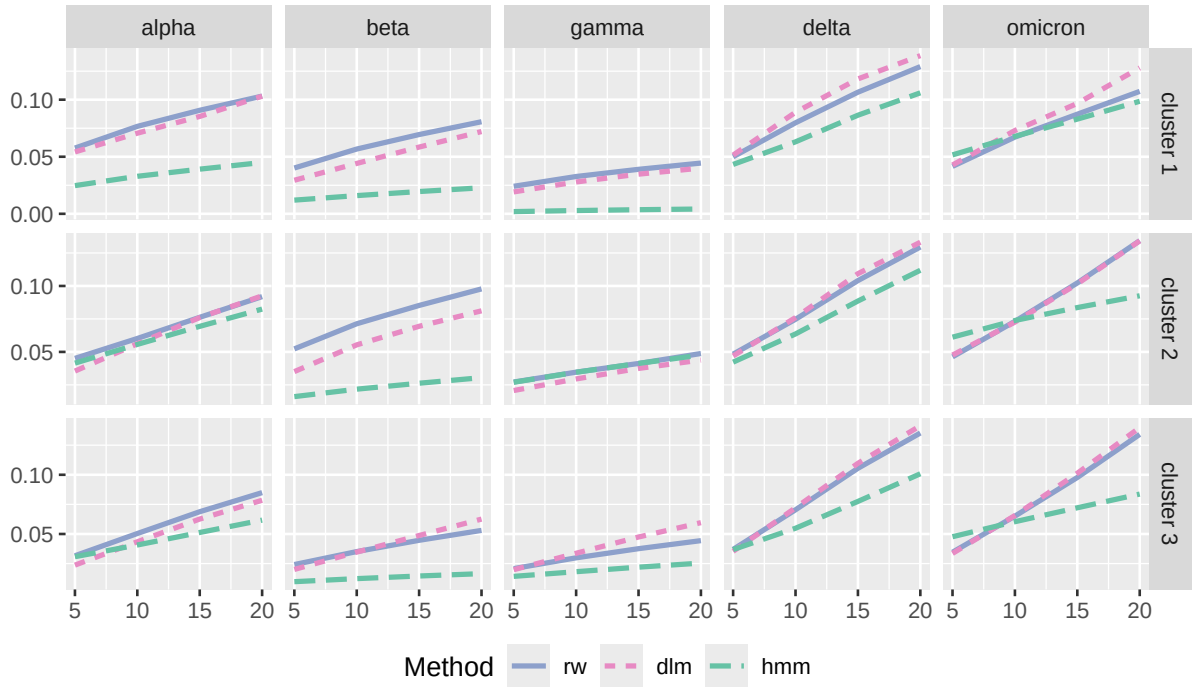


Figure 4.16: Median mean absolute error (MAE) across each cluster for the mean standard deviation across each windows length by variant. The different coloured bars correspond to each of the methods: GDM-HMM, GDM-RW and GDM-DLM.

On the other hand, the upper quartile of the standard deviation across the windows (Figure 4.17) has a less linear shape in comparison. This can be seen within the delta and omicron variants where the lines for each cluster are presented in a stepwise trajectory with a series or sharp increases or decreases. Here, the method with the lowest MAE value changes as the window length increases. Overall, within both these variants across each cluster, the GDM-HMM has the lowest MAE for window length 10 and larger.

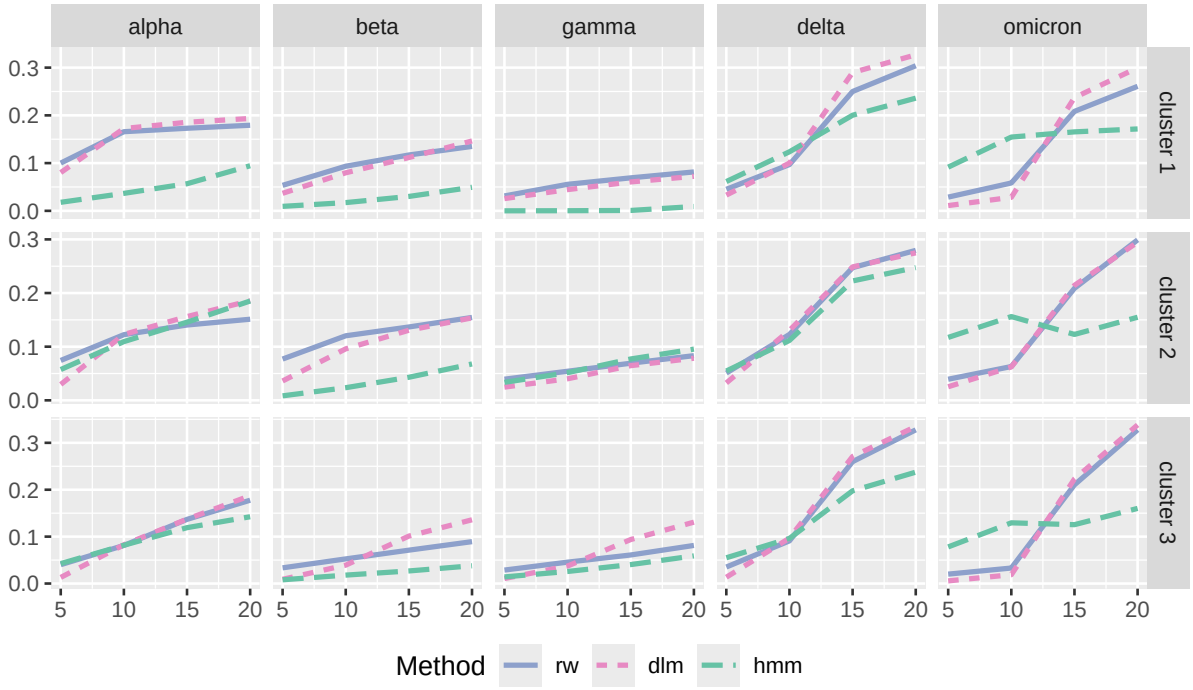


Figure 4.17: Median mean absolute error (MAE) across each cluster for the upper quartile of the standard deviation across each window length by variant. The different coloured bars correspond to each of the methods: GDM-HMM, GDM-RW and GDM-DLM.

In conclusion, after assessing both density curves and the MAE across each method, it can be deduced that the GDM-HMM demonstrates superior performance compared to other considered time series models.

4.8 Summary & Discussion

In this chapter, we introduced compositional time series, particularly focusing on non-smooth data. We reviewed previous methods used to address compositional time series, which mainly involve applying a log-ratio transformation. However, a log-ratio transformation is unsuitable for data which contain zeros or consists of count data, a common occurrence of compositional data. We also discussed other methods, including hierarchical approaches that apply a Bayesian hierarchical framework to the data, such as the work by Stoner et al. (2020c), who propose a multivariate hierarchical framework for modelling compositional count time series. Nevertheless, this approach may not be suitable for very non-smooth time series. We outlined the use of hidden Markov models (HMMs), a common modelling technique for time series. Although previous approaches applying HMMs to compositional time series were reviewed, we found that work in this area is limited, especially for non-smooth compositional time series. This gap paves the way for the work presented in this chapter.

The proposed GDM-HMM framework for compositional time series was outlined and assessed. This was developed by implementing a HMM with the flexible GDM distribution. This model is particularly suited for non-smooth time series data which is presented as counts. This approach addresses a notable gap in the literature for this specific type of compositional time series.

The COVID-19 variant data from GISAID motivates our application, consisting of a non-smooth compositional count time series. This dataset includes 169 weeks of COVID-19 counts from 217 countries worldwide. With many zero values across each country, these data demonstrate why a log-ratio transformation is unsuitable for this time series. The COVID-19 data was pivotal in the development and testing of the proposed GDM-HMM framework.

We demonstrated that treating the data as multivariate preserves information about how the total is distributed across each variant. For example, if one variant accounts for half the total count, the remaining variants must collectively account for the other half - something that separate univariate models would be unable to capture. This joint modelling ensures that compositional constraints are respected and provides a more coherent representation of the underlying structure in the data.

We showed that the GDM-HMM framework can effectively characterise the temporal evolution in the prevalence of each variant as a progression through a series of hidden states. This revealed that each variant differs across countries, highlighting the need to model each variant individually. Given this, we aimed to cluster countries based on the evolution of the compositions over time to capture the expected characteristics of each variant. This enables unique HMM parameters on a per-cluster basis. We applied a clustering approach prior to modelling which resulted in three unique clusters of countries. We then built the HMM using a series of independent Beta-Binomial distributions from the GDM distribution for each variant and created a transition matrix with five states to model the evolution of the variants. To ensure the HMM can only move forward through the states, we added constraints to the transition probabilities because each variant evolves without returning to a previous state. By implementing a HMM framework, we can gain insights into the characteristics of the different variants through the transition probabilities. The persistence length is the length of time (weeks) the GDM-HMM continues in that state before transitioning into the next state. Overall, the highest posterior median expected persistence occurs with the omicron variant in State 3 suggesting that this variant has a prolonged period of dominance which is detected across all clusters. Whereas the other variants have a similarity in persistence lengths between clusters suggesting a common transition matrix for all clusters in the GDM-HMM framework may have been adequate, the pronounced differences observed for omicron suggest that allowing for differences between clusters could be beneficial. A hierarchical approach, with some information pooling between clusters, may serve as an appropriate compromise.

We designed a posterior predictive model checking experiment to assess the effectiveness of the proposed framework. The GDM-HMM framework was examined by comparing it to alternative GDM models using more standard time series structures: a Random Walk model (RW) and a Dynamic Linear model (DLM). Model computation was fastest for our proposed GDM-HMM model (approximately 2 hours) compared to the GDM-RW model (approximately 3 ½ hours) and the GDM-DLM model (approximately 7 hours). We evaluated the performance of the replicate values for each method through a window summary to examine how the replicates behave over time. This was produced for the mean, upper and lower quartiles of the standard deviation across windows for multiple window lengths. We visually compared the density for the summary statistic of each of the replicates with the distribution of the same statistic of the original data. Overall, the GDM-HMM consistently outperforms the GDM-RW and GDM-DLM models by accurately capturing the corresponding statistic computed using the original data and producing replicate values that could be plausible in real-world scenarios. The three methods were also compared using the Mean Absolute Error (MAE) for both the mean and upper quartile of the standard deviation within each moving window. Visually examining the MAE per country and aggregated cluster means, the GDM-HMM consistently provided the lowest MAE for all three clusters. There was only one exception to this seen within the gamma variant for Cluster 2, where the GDM-DLM had a lower MAE for the mean of the standard deviations. When investigating multiple window lengths to ensure robustness, the same conclusions can be reached. Across all window lengths, the GDM-HMM demonstrates superior performance in minimising the MAE compared to the alternative models.

In conclusion, the proposed GDM-HMM framework was evaluated against alternative time series models using these data. The performance of the model was evaluated by carrying out a posterior predictive model checking experiment and examining the replicates using a moving window summary. The density plots of the GDM-HMM indicated that the GDM-HMM was producing sensible values that were a close match to the original data, which could explain the real-world evolution of the variants. It was found that the proposed method

largely produced a lower MAE when compared with the other GDM models examined. The reduced computational time required to run the GDM-HMM proves valuable alongside the superior model fit. This highlights the effectiveness of the GDM-HMM in comparison to alternative models for non-smooth compositional time series.

Further work could include incorporating covariates - such as vaccination rates and intervention measures - into the model to better inform the transition probabilities, making them more reflective of real-world dynamics. It should also be noted that the framework was only applied to the COVID-19 variant data and further work would be required to test the method within other applications.

Chapter 5

Methods for Spatial Compositional Data

In this chapter, we propose an approach to modelling compositional data arranged over a spatial domain, such that we need to account for both the compositional and spatial structures. Furthermore, we target the additional challenges of accounting for both zero and missing values in the spatial compositions.

Here, we propose a framework combining the Generalised-Dirichlet-Multinomial (GDM) family of distributions with two-dimensional penalised regression splines that capture spatial structure. We evaluate our approach through two posterior predictive experiments, one to assess a novel variance parameter specification and another to assess how well the framework can predict missing compositional counts.

5.1 Introduction

Spatial data are observations arranged over some spatial domain, often referring to information that identifies the geographical location, characteristics of that location or defined regions on the surface of the Earth.

There are three main forms of spatial data; areal, geostatistical and point-process data, each detailed within Cressie (2015). Areal data means that a region of interest is divided into non-overlapping areas with defined boundaries (e.g. states or counties), such that there is one aggregated measurement per unit. Areal data include counts of infected individuals in regions administered by different health boards. In contrast, geostatistical data are a set of observations taken at fixed spatial locations, e.g. soil moisture or air quality measurements made by a device installed at a single point in space – in these cases, the location of the device is potentially informative for understanding the soil moisture or air pollution levels but not an outcome of interest. Like geostatistical data, point process data are observations at

points in space but where the locations are themselves an outcome of interest. For example, the locations of crime incidents are an outcome of interest as studying them could provide information to identify patterns such as crime hotspots or links to socioeconomic conditions. Cressie (2015) discusses different types of spatial data in more detail.

A real-world quantity of interest arranged over a spatial dimension will often have spatial correlation (dependence), which means observations that are closer together in space are more likely to have similar characteristics than observations farther apart. Meanwhile, many statistical procedures (e.g. linear regression) do not give reliable results when dependence in the data is not accounted for. Thus, analysis of spatial quantities usually means we know of a potential source of dependence (spatial dependence) that should be considered carefully in our analysis and modelling. The study of methods for capturing spatial dependence and their applications is one of the most active areas of statistical research; Schabenberger et al. (2017) is a useful introduction to the general challenges and common approaches.

As explored in previous chapters, compositional data arise in many statistical contexts and this includes geographic contexts; in this chapter, we study and propose methods for compositional data observed over a spatial or geographical dimension, which we will call spatial compositional data. Such data can arise in physical science contexts, such as in soil compositions (Odeh et al., 2003), iron ore compositions (Tolosana-Delgado et al., 2019) and mineral rock compositions (Zuo et al., 2013), but also within social science contexts, including election voting (Nguyen et al., 2021) and population studies (Martinez et al., 2020).

Methods for capturing such data will naturally need to account for both the spatial and compositional structures, and potentially interactions between the two. Interactions between spatial and compositional structures can arise when the spatial locations induce similarity in the compositions, or when spatial trends in one component affect the relative proportions

of others due to the compositional constraint. These dependencies can bias inference if not properly accounted for. Developing these methods into models could then prove useful for prediction at locations with missing data or for covariate inference, adjusting for spatial dependence that might otherwise confound the covariate effects.

Many of the existing approaches for analysing spatial compositional data begin by applying a log-ratio transformation to convert the data from a constrained space to an unconstrained space to allow standard spatial methods or models to be applied. However, log-ratio approaches have their drawbacks, including challenges handling zeros in the data and missing values in the compositions (as discussed in more detail in Chapter 2, Section 2.2.1). For data without these features, log-ratio approaches may be a compelling option. However, in line with one of the main themes of the thesis, we will focus our study on alternative frameworks that account for the nature of compositional data without requiring transformation to an unconstrained space, such that they are suitable for a wider variety of real-world data problems.

5.1.1 Tree species data

A real-world example of spatial compositional data is tree species composition over a defined geographic area of interest (AOI). The data set we will investigate was collected and compiled by Fera Science UK, providing insight into a small mixed woodland area in North Yorkshire that contains both natural and plantation woodland within the AOI. Each grid cell in the data represent a 10m by 10m area on the ground, identified by its British National Grid easting and northing coordinates. The data was collected through a combination of drone (UAS) imagery, satellite imagery and ground survey data to classify the different tree species within the AOI. Each grid cell contains proportional estimates of tree species coverage, meaning that multiple tree species can be present within a single cell. The sum of

all proportions for tree species within a given grid cell equals one. Since these observations represent aggregated measures over defined spatial areas (each grid cell), the data are areal in nature. However, the simple grid structure may allow the use of non-areal models for spatial dependence. Full details of the data set and the collection methods can be found in Frantsuzova (2021).

The tree species found in the AOI are: ash, beech, larch, oak, scots pine, silver birch, sitka spruce, sweet chestnut, sycamore and a shadow class. The shadow class is thought to be a real shadow detected by the measuring equipment. The 10-part compositional element of tree species contains no missing values; however, there is a large proportion of structural zero values, as defined in Chapter 2, Section 2.3.2, over the 2,153 grid cells. Table 5.1 displays the mean proportion of each tree species and summarises the percentage of zero values for each tree species with an overall percentage of structural zeros of 62%.

Table 5.1: Mean proportion and percentage of structural zeros observed for each tree species.

Tree Species	**Mean proportion**	**Percentage of zeros (%)**
Ash	0.021	87.7
Beech	0.049	81.1
Larch	0.186	51.5
Oak	0.189	45.2
Scots Pine	0.052	75.3
Shadow	0.145	6.3
Silver Birch	0.092	64.0
Sitka Spruce	0.068	78.3
Sweet Chestnut	0.071	70.9
Sycamore	0.126	54.3

A heatmap of each tree species can be examined to give an indication of the spatial spread over the AOI in Figure 5.1. It can be noted that the majority of individual tree proportions across grid cells are equal to zero, identified by the pixels coloured in dark purple. It is evident from Figure 5.1 that there are substantial clusters of the tree species larch concentrated in the lower right-hand portion of the AOI, as indicated by the abundance of yellow pixels. In the same region, there is another smaller cluster of tree species of sitka spruce. The presence of the oak tree species appears widespread across various spatial locations within the AOI,

characterised by the diverse array of coloured points representing the proportions scattered throughout the region. The shadow class is also distributed across the entire region, with a lower proportion, typically between 0.25 and 0.5, as depicted by the dark teal pixels. In the lower left-hand region of the AOI, the most prominent cluster of tree species appears to be the beech tree, holding the highest proportion within this location. The heatmap for ash, which exhibits the largest proportion of structural zeros, is predominantly composed of purple pixels, indicating the absence of ash within these areas.

In general, the proportions of the tree species vary considerably across the whole AOI. Different species are distributed unevenly, creating distinct hotspots where certain tree species are more concentrated. In some cases, beyond the boundaries of the hotspots the presence of the dominant tree species drops suddenly, whereas elsewhere the proportion drops more gradually. The spatial structure is therefore not uniform both in terms of the overall distribution (since there are clear regions where specific tree species dominate) and the degree of smoothness or lack thereof.

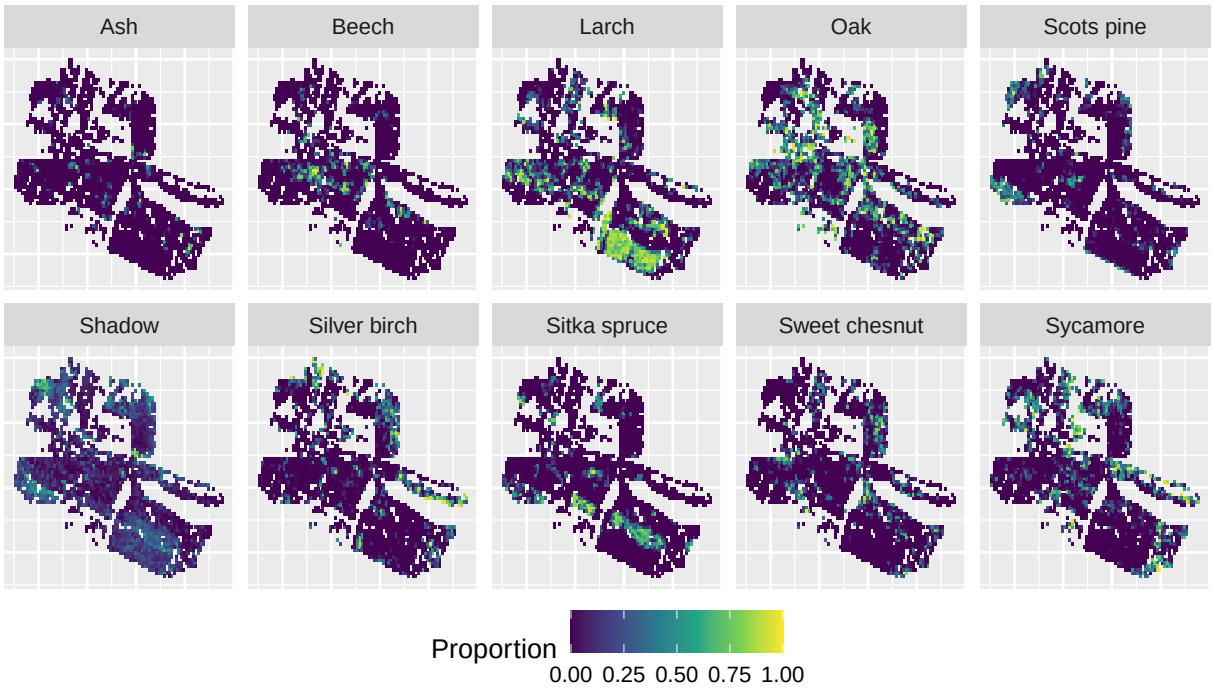


Figure 5.1: Heatmaps of the estimated proportions of tree species detected within each grid cell.

Given that the tree species compositions include a substantial proportion of structural zeros, common spatial compositional data techniques relying on log-ratio transformations are unlikely to be suitable for modelling these data (Chapter 2). Therefore, we believe that this is an interesting motivating application for developing and testing a general multivariate framework that directly accounts for compositional structure and accounts for spatial dependence through latent effects. For general applicability, we aim for our framework to allow for zeros in the data, as well as missing values in the spatial compositions. We will assess the latter through predictive experiments for an augmented version of the tree species data where some of the compositions are missing, potentially reflecting hypothetical measurement failures/gaps or unresolvable ambiguity in classifying tree species.

The chapter is structured with Section 5.2 reviewing previous approaches to modelling spatial compositional data. In Section 5.3 we present the methodology and proposed general framework, including exploring the spatial splines approach used. The implementation of the framework is also outlined in this section. The general framework is applied to the tree species data within Section 5.4. A single tree prediction experiment is carried out in Section 5.4.1 which examines the difference between a novel polynomial variance parameter versus a single fixed parameter. We investigate a multi-tree species prediction experiment in Section 5.4.2 which applies the general framework to the compositional data with missing values across the compositions. Finally, in Section 5.5 we summarise and critically evaluate the work carried out in this chapter.

5.2 Spatial Compositional Data

In this section, we review existing methods for modelling spatial compositional data and broadly arrange these into two groups: those that apply log-ratio transformations to compositional data before applying standard spatial methods, and alternatives that directly model the data in their original compositional form. While we have already laid out our case for why log-ratio transformations are not always suitable, a critical review provides useful insights, including previous successes and challenges in capturing spatial structure.

First, however, we will introduce some general notation that we will use for exposition of existing methods and those we will propose later. In many cases, this notation will differ from the notation used in the original cited works. Let $\boldsymbol{x}_s = (x_{s,1}, \ldots, x_{s,D})$ represent a set of proportions, for D compositions, summing to a total y_s, existing for some spatial location s from the set of locations $s \in 1, \ldots, S$. In most cases reviewed, $y_s = 1$, i.e. $\boldsymbol{x}_s$ exists on the simplex. At this stage suppose that s could refer either to areal regions or fixed points with known coordinates.

5.2.1 Log-ratio approaches

As for other kinds of compositional data studies in previous chapters, the predominant methods for handling spatial compositional data stem from the work of Aitchison (1982); that apply log-ratio transformations to map the compositions $\boldsymbol{x}_s$ to the unconstrained real space, resulting in $\boldsymbol{x}_s^* \in \mathbb{R}$, so that standard methods can be used.

A comprehensive review of log-ratio transformations for the context of spatial compositional data is given in Pawlowsky-Glahn et al. (2016). Within this, the log-ratio approach and the principle of working in coordinates using ILR (Chapter 2) log-ratio representation are outlined for spatial compositional data. However, the review is limited as Pawlowsky-Glahn et al. do not discuss instances when there are zeros in the data or whether the methods are suitable when zeros are present. We will review a few specific examples throughout this chapter.

Following a log-ratio transformation, standard spatial models have been applied to spatial compositional data. Autoregressive (AR) spatial models assume that the value at a location depends on the values at neighbouring locations. These often incorporate a spatial weight matrix to define the dependence structure. Spatial autocorrelation is captured in this model through the specification of how nearby locations influence each other. This is extended to the Conditional Autoregressive (CAR) models that specify the joint spatial distribution through conditional distributions. The value within each location is modelled as dependent on its neighbours. CAR models are particularly useful in Bayesian spatial analysis due to their compatibility with hierarchical modelling approaches. Leininger et al. (2013) uses a multivariate Conditional Autoregressive (CAR) specification following the application of ALR to analyse land use and land cover (LULC) data in the northeastern United States:

$$\boldsymbol{x}_s^* = \text{ALR}(\boldsymbol{x}_s), \tag{5.2.1.1}$$

$$\boldsymbol{x}_s^* = \left(\log\left(\frac{x_{s,1}}{x_{s,D}}\right), \ldots, \log\left(\frac{x_{s,D-1}}{x_{s,D}}\right)\right) \tag{5.2.1.2}$$

The proportions of land use types are observed in each 3 km $\times$ 3 km grid cell. Leininger et al. propose a spatial regression model that captures the flexible dependence among the components at each location in addition to the dependence across locations of the simplex-restricted measurements. The model is formulated to be able to handle a high incidence of zero values, that are treated as random occurrences rather than structural zeros in the compositional sense. This is conducted through a latent variable modification of an ALR,

allowing the model to accommodate zeros while transforming the data for regression. For the compositional data vector, $\boldsymbol{x}$ is generated from a latent multivariate Gaussian random variable, $\boldsymbol{z}$, where the transformation from $\boldsymbol{z}$ to $\boldsymbol{x}$ sets the d^{th} component $x_d = 1$ when $z_d \leq 0$ and $x_d > 0$ when $z_d > 0$.

$$x_d = \frac{\max(0, z_d)^\gamma}{1 + \sum_{d'=1}^{D-1} \max(0, z_{d'})^\gamma}, \tag{5.2.1.3}$$

where where $\gamma > 0$ is a power-scaling parameter and D the number of components. The denominator ensures the components sum to 1, satisfying the simplex constraint. The zeros are incorporated as part of the model by allowing for $\max(0, z_d)$ which maps negative values of the latent variable z_d to zero in x_d. However, this requires the data to contain at least one component (x_D) that does not contain any zeros to serve as the baseline classification for the transformation, which in real-world situations may not always be achievable. Although the framework can be extended to include a temporal structure, which could be advantageous in certain circumstances, it still faces limitations in cases with a significant proportion of zeros.

Yoshida et al. (2018) investigate the impact of spatial relationships on prediction accuracy within spatial compositional multivariate models, focusing on Compositional Multivariate Conditionally Autoregressive (CMCAR) models. The work is assessed using ALR transformed Japanese land-use data to evaluate how variations in the spatial weight matrix affect the model performance. The CMCAR model is given by

$$x_i \sim N_d(B^\top x_i, \eta_i V), \tag{5.2.1.4}$$

where B is a $(p+1) \times (D-1)$ parameter matrix, x_i is a $(p+1) \times 1$ explanatory variable vector, η_i is a $(D-1) \times 1$ random effect vector and V is a $(D-1) \times (D-1)$ variance-covariance matrix. The spatial autocorrelation is modelled by setting a prior distribution for η_i as:

$$\eta_i | \eta_{j\, j \neq i} \sim N_d \left(\frac{1}{U_i} \sum_{j=1}^{n} w_{ij} \eta_j, \frac{1}{U_i} \Sigma \right), \tag{5.2.1.5}$$

where U_i is the row sum of the i^{th} row of the spatial weight matrix W with elements w_{ij} and Σ is the $(D-1)\times(D-1)$ variance-covariance matrix. It was found that the choice of spatial weight matrix significantly influences the prediction accuracy of the CMCAR, hence the choice of this would be up to the user and requires extra work to deduce the optimal spatial weight matrix. The same spatial weight matrix was assumed for all categories, which may not be favourable since different compositions may have different spatial structures. Again, this adds complexity to model formulation prior to fitting the CMCAR. However, this method does not allow for zeros in the compositions and Yoshida et al. add a small value to all the compositional values. Therefore, this approach may not be suitable for applications where the zero value is informative.

More examples of commonly used spatial models applied to compositional data include Martinez et al. (2020), which use a standard CAR model to analyse ILR-transformed birth population data, and Nguyen et al. (2021) who present a spatial autoregressive (AR) model to analyse election vote share data which has also been transformed using ILR. Additionally, to tackle any zero values in the data, Nguyen et al. (2021) aggregate the election vote share parties into three blocks which leads to loss of detail, especially for the smaller / less dominant compositions.

Another type of spatial model that has been applied to spatial compositional data is kriging. Kriging is a standard statistical technique that estimates the values at unobserved locations using the values at nearby locations. This is conducted through weighted averages of the observed data. Odeh et al. (2003) apply a kriging approach to ALR transformed soil compositional data. The values at the unmeasured locations are then estimated through kriging by:

$$Z(x_0) = \sum_{i=1}^{S} \lambda_i Z(x_i), \tag{5.2.1.6}$$

where $Z(x_0)$ is the estimated value at the unknown point x_0, $Z(x_i)$ is the ALR transformed value at the i^{th} observed location x_i, λ_i is the weight assigned to each known value $Z(x_i)$ and S is the number of known values. The weights λ_s are calculated to minimise the estimation variance. Odeh et al. compare ALR with the case where no transformation is placed upon the compositional data. This resulted in ALR having a far superior performance showing the need to treat the compositional data specifically as without it the sum of the predictions was not constant at many of the locations breaking the compositional constraint.

Cokriging extends kriging to handle multiple variables within the prediction. This method estimates the values at unmeasured locations using the values of the other variables available at the spatial location. Then for two variables, Equation (5.2.1.6) can be extended to estimate a location x_0 for the primary variable Z_1 and secondary variable Z_2 as

$$Z_1(x_0) = \sum_{i=1}^{S_1} \lambda_i Z_1(x_i) + \sum_{j=1}^{S_2} \gamma_i Z_2(x_{ij}), \tag{5.2.1.7}$$

where λ_i and γ_j are the cokriging weights for the primary and secondary variables, respectively. $Z_1(x_i)$ and $Z_2(x_j)$ are the log-ratio transformed values of the primary and secondary variables at locations x_i and x_j. N_1 and N_2 are the number of observations for the primary and secondary variables. Again in this instance, the cokriging weights are determined by solving the following system of cokriging equations, derived from the spatial covariance structure of the variables. Tolosana-Delgado et al. (2013) utilise ILR to transform the compositional data before applying cokriging on the spatial data. Pawlowsky-Glahn et al. (2015b) also applied cokriging to spatial compositional data using the previous work of Pawlowsky-Glahn et al. (2015a) which used cokriging for the compositional ILR coordinates jointly with the coordinate of the total. This suggests that the chosen total can be included as an additional coordinate in addition to those coming from the composition. Cokriging has also been applied by Tolosana-Delgado et al. (2013), which use the log-ratio transformed compositions from an iron ore deposit dataset consisting of several mineralogical textural types. However, there is no evidence within these works on how cokriging handles any zeros present.

More recently, Clarotto et al. (2022) apply a kriging approach to land cover compositional data. Clarotto et al. introduce a new transformation - the Isometric α-transformation (α-IT) which combines ILR with the α-transformation (Chapter 2, Section 2.2.2, Equation (2.2.2.4)). The α-IT is defined for a compositional vector $\boldsymbol{x}_s \in S_D^0$ as:

$$\boldsymbol{z}_{\alpha\text{-IT}}(\boldsymbol{x}_s) = \alpha^{-1}\boldsymbol{H}_D\boldsymbol{x}_s^{\alpha}, \tag{5.2.1.8}$$

where $\boldsymbol{x}_s^{\alpha}$ is the component-wise power of $\boldsymbol{x}_s$, $\boldsymbol{H}_D$ is the $(D-1)\times D$ Helmert matrix, and $\alpha > 0$ allows for the presence of zeros in the compositions.

Similarly to other α-transformations, as α tends to 0, the α-IT becomes the ILR and adheres to the Aitchison geometry and when $\alpha = 1$ corresponds to a linear transformation of the data, adhering to the Euclidean geometry. An advantage of this proposed transformation over the original ILR, is in the transformation's ability to accept zero values in the compositions when $\alpha > 0$. However, a downside of the transformation is that a decision has to be made of what value α will take. An advantage of the α-IT over the traditional α-transformation is that is provides a direct connection to the spatial covariance structure, allowing geostatistical tools like cokriging to be applied more coherently in a compositional setting.

Clarotto et al. propose using maximum likelihood estimation (MLE) to determine the optimal value of α. The results demonstrate that the transformation performs particularly well when compositions include zeros, although a small value for α may not be optimal when zeros are present. The authors highlight that a potential bias may be introduced when back-transforming the predictions from the transformed space to the simplex, due to the non-linear nature of the α-IT transformation with the differences in geometry between the Euclidean space and the simplex. As prediction is conducted in the unbiased transformed space, the back-transformation process does not perfectly preserve the relationships among

the components of the compositional data, leading to potential inconsistencies in the predictions. The authors state that further research is needed to develop methods for correcting and reducing this bias, ensuring more accurate back-transformation of compositional data. Furthermore, the authors have not addressed the problem of incomplete compositions.

Another instance of applying log-ratios to spatial compositional data is within Tjelmeland et al. (2003). Here, the authors apply ALR to spatial compositional data of sediments in an Arctic lake, expanding on the work of Aitchison (1982), which did not account for spatial dependence. Tjelmeland et al. extended the logistic normal distribution by incorporating Gaussian processes to model the spatial structure. The authors state that this could be extended by using Gaussian Markov random fields, which was conducted in Pirzamanbein et al. (2018) explored in the next section.

Frantsuzova (2021) focuses on advancing the classification of tree species within a spatial region using a dataset that contains a large proportion of zero values. Note that we study the same data set in this chapter. The goal of this analysis was to address challenges with spatial imagery in monitoring woodlands. To address the goal, Frantsuzova performs regression analysis on ALR transformed compositional tree data, with a small number (0.005) added to each of the structural zeros present. Frantsuzova explored the hierarchical relationships among tree species compositions in the study area using clustering techniques to investigate whether certain tree species exhibited spatial or compositional clustering patterns. To do this, the original data compositions were transformed using CLR. Following fitting hierarchical clustering with Ward linkage, it was found that no clear pattern emerged, meaning that the clusters created didn't distinctly represent different tree types, suggesting that the tree types are mixed in a way that does not form distinct, identifiable clusters or patterns. The spatial coordinates of the compositional data were then included as predictors in random forest (RF) regression models, which aimed to predict the proportion of different tree species. By including the spatial coordinates as predictors, the aim was to capture any

spatial dependencies in tree type distributions. Frantsuzova found that the impacts of spatial variables on predictive accuracy varied, suggesting the presence of spatial structure that interacts differently across tree species. This highlights an opportunity to develop further the understanding of spatial structure in relation to tree species.

Whilst the most common approach to handling spatial compositional data involves applying a log-ratio transformation before using standard statistical techniques, we argue that this method has limitations in some cases. Notably, in Chapter 2, Section 2.2.1 we argued that log-ratio transformations can be inappropriate in cases where zeros are present in the data - a frequent occurrence in real-world compositional datasets. Furthermore, in the presence of any missing values log-ratio transformations are not appropriate, limiting their application in real-world spatial compositional data. Log-ratio transformations require complete data for all components in order to be correctly defined. When a component value is missing, some or all of the log-ratio transformations will not produce sensible results, as the necessary relative proportions cannot be properly computed. As outlined within this review, there is no approach that would be suitable to apply if there are zeros present in the data or any missing values in the compositions, providing a need to develop a methodology to tackle this gap in the literature.

5.2.2 Alternative approaches

Less work has been conducted on examining spatial composition without applying a log-ratio transformation. In this section, we will review these approaches and explain why they fall short of wholly addressing our aims.

Walvoort et al. (2001) introduce an alternative method for spatial compositional data to the log-ratio transformation - compositional kriging. The compositional kriging method proposed extends ordinary kriging (Equation (5.2.1.6)) by incorporating constraints necessary for compositional data, i.e. nonnegativity and the constant sum constraint. The compositional kriging approach considers all the compositional elements simultaneously by minimising the sum of their prediction error variances:

$$\min_{\lambda_i} \sum_{i=1}^{D} (\sigma_i^2 + \lambda_i^\top C_i \lambda_i - 2\lambda_i^\top r_i), \tag{5.2.2.1}$$

where λ_i is the vector of weights for i^{th} component, σ_i represents the variance of the kriging prediction for component i, C_i is the covariance matrix for for i^{th} component, r_i is the covariance vector between observed data and the prediction point x_i. Equation (5.2.2.1) is subject to: unbiasedness - $\Lambda^\top 1_{(n)} = 1_D$; nonnegativity - $\lambda_i^\top z_i(x_i) \geq 0$ for $i = 1, \ldots, D$; and constant sum - $\text{tr}(\Lambda^\top X) = 1$, where $\Lambda = (\lambda_1, \ldots, \lambda_D)$ is the matrix of weights, X is the data matrix and $\text{tr}(\cdot)$ denotes the trace of a matrix. This allows prediction of the values of the compositional variables at unsampled locations by accounting for the spatial relationships observed in the data. This ensures that the predictions will adhere to the compositional constraints. Unlike traditional kriging, that could result in predictions that are negative values or do not have a constant sum, the authors assert that compositional kriging guarantees compliance with the constraints. Walvoort et al. optimise the kriging weights using Lagrange multipliers and Kuhn-Tucker conditions. The authors state that this approach accounts for spatial covariance structures of the data without requiring transformations, unlike other methods that require log-ratio transformations. An advantage of the compositional kriging approach is in its ability to handle zero compositions. Yet, this was not examined within the case studies examined in Walvoort et al. where the zero components were few and removed for analysis. Pawlowsky-Glahn et al. (2016) state that this approach follows the Euclidean geometry rather than Aitchison's geometry. This assumption suggests that compositional data carry absolute rather than relative information, which conflicts with Aitchison's principles that compositions are relative measures.

Expanding on the work of Tjelmeland et al. (2003) discussed above, Pirzamanbein et al. (2018) constructed a Gaussian Markov random field (GMRF) within a hierarchical spatial model. This used the Dirichlet distribution to handle the compositional nature of the data with the log-ratio to link the GMRF to the compositional probabilities. The latent field is represented as an $N_d \times 1$ vector with $\eta_{\text{all}} = (\eta_{\text{all},1}^\top, \ldots, \eta_{\text{all},d}^\top)^\top$, where each $\eta_{\text{all},k}$ is a spatial field with N locations. The latent field is then connected to the observed spatial locations through:

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\eta}_{\text{all}}, \tag{5.2.2.2}$$

$$\boldsymbol{\eta}_{\textbf{all}} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{X}, \tag{5.2.2.3}$$

where $\boldsymbol{A}$ extracts observed elements, $\boldsymbol{B}$ is the matrix of covariates, $\boldsymbol{\beta}$ represents regression coefficients, and $\boldsymbol{X}$ is the spatially correlated multivariate field. The spatial dependence is modelled using a Gaussian Markov Random Field (GMRF):

$$X \sim N(0, \Sigma \otimes Q^{-1}(\kappa)), \tag{5.2.2.4}$$

where $Q(\kappa)$ is a precision matrix approximating a Matérn field (Matérn, 1960) with spatial scale κ, and Σ captures covariances among the components. The log-ratio link function is constructed using ALR to map between the latent field and the compositions is given as:

$$\eta_{s,k} = \log(z_{s,k}) - \log(z_{s,D}), \tag{5.2.2.5}$$

for $k = 1, \ldots, D-1$ for each spatial location s.

As outlined earlier (Chapter 2, Section 2.3.2) the Dirichlet distribution is limited due to its lack of flexibility. Specifically, as it only has one variance parameter regardless of the length of $\boldsymbol{x}$, it may not capture complex dependencies present in the compositional data. While the GMRF is computationally efficient due to the sparse precision matrices, it is important to note that real-world spatial compositional data may not always conform to Gaussian

distributions at the latent level. This assumption could lead to misrepresentation of the spatial dependence and potentially biased predictions. However, with large spatial grids, the computational demands could increase, greatly reducing the computational efficiency of the method.

Feng et al. (2017) propose a two-stage spatial mixture Dirichlet regression model to account for the compositional nature of spatial land cover data whilst allowing for zeros in the data. This model also allows for compositions to be missing, as certain land cover categories may not be observed at certain sampling locations. The two-stage model is formulated with the first stage predicting whether the response type (e.g. the land cover category) is present or absent using a spatial multivariate probit model from the raw compositions.

$$P(x_{ij} = 1) = \Phi(z_j^\top \beta_i), \tag{5.2.2.6}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution, z_j is the covariate vector and β_i are regression coefficients for response type i. This accounts for spatial dependence through a spatial latent process. The second stage then models the proportions of the land cover categories that are present, i.e. once a composition is predicted to be present at a location, the second stage applies a Dirichlet regression only to the remaining non-zero values. For locations with non-zero response compositions, the data are modelled using a Dirichlet distribution. Let $G_{ij} \sim \text{Gamma}(\kappa_{ij}, 1)$ where

$$\kappa_{ij} = \phi \mu_{ij}, \tag{5.2.2.7}$$

$$\mu_{ij} = \frac{\exp(z_j^\top \eta_i)}{1 + \sum_{i=1}^{I-1} \exp(z_j^\top \eta_i)}, \tag{5.2.2.8}$$

where ϕ is the precision parameter and η_i are the regression coefficients for the Dirichlet model. The compositional response X_{ij} is constructed as:

$$x_{ij} = \frac{G_{ij}}{\sum_{k-1}^{I} G_{kj}}, \tag{5.2.2.9}$$

ensuring the response is compositional and adheres to the unit-sum constraint. Through this two-stage model, Feng et al. hope the model captures both the presence and absence of the land cover types and their proportional compositions across the spatial space. However, this approach requires at least one composition to be present at all spatial locations which could be challenging in real-world applications if some categories are rarely observed or contain many zeros. Additionally, the spatial dependence is only modelled in the first stage of the model so some spatial structures might not be captured when forming the compositions in the second stage, potentially leading to loss of information. The data tested only contained three land cover categories, therefore, when there are a greater number of categories, the number of parameters to estimate will grow which could increase the computational burden and complexity of the model.

Therefore, as demonstrated in this section, there is currently no established method for handling spatial compositional data that simultaneously accounts for the general presence of zeros and missing values, which the novel approach presented in this chapter aims to address.

5.3 Proposed Methodology

Recall that our aim is to develop a general multivariate framework that directly accounts for compositional structure, accounts for spatial dependence through latent effects, allows for zeros in the data and allows for missing values in the spatial compositions.

If $\boldsymbol{x}_s$ are compositional proportions, one option that we could consider is to assume that compositions $\boldsymbol{x}_s$ come from a Generalised-Dirichlet (GD) distribution (Wong, 1998), i.e.

$$\boldsymbol{x}_s \sim \text{GD}(\boldsymbol{\alpha};\boldsymbol{\beta}). \tag{5.3.0.1}$$

This distribution arises from a series of Beta distributions, for $x_{s,i}$ given $x_{s,1},\dots x_{s,i-1}$:

$$x_{s,i}|x_{s,1},\dots x_{s,i-1} \sim \text{Beta}(\alpha_i,\beta_i), \tag{5.3.0.2}$$

where $i = 1,\dots,D-1$ and $x_{s,D} = 1-\sum_{i=1}^{D-1} x_{s,i}$. Here, α_i and β_i are the shape parameters of the respective Beta distributions, where for each component i are given as:

$$\alpha_i = \lambda_i, \tag{5.3.0.3}$$

$$\beta_i = (\lambda_{i+1}+\lambda_{i+2}+\cdots+\lambda_D), \tag{5.3.0.4}$$

where λ_i are the parameters for the Dirichlet distribution.

For a random vector $\boldsymbol{x}_s = x_{s,1},\dots,x_{s,D}$ where $x_{s,d} \in (0,1)$ and $\sum_{i=1}^{D} x_{s,i} = 1$, the probability density function (PDF) of the GD is given as:

$$f(\boldsymbol{x};\boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_{i=1}^{D-1} \frac{1}{B(\alpha_i,\beta_i)} x_i^{\alpha_i-1}\left(1-\sum_{j=1}^{i} x_j\right)^{\beta_i-\alpha_{i+1}}, \tag{5.3.0.5}$$

where $B(\alpha_i,\beta_i)$ is the beta function given as $\frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_i+\beta_i)}$.

When $\beta_i = \alpha_{i+1}, \forall i = 1,\dots,k-1$, the GD distribution reduces to the standard Dirichlet distribution (the limitations of which we noted in Chapter 2, Section 2.3.2). In all other cases, the GD has $2(D-1)$ free parameters, where D is the length of $\boldsymbol{x}_s$, compared to D for the Dirichlet. Thus, the GD has a more general covariance structure and is very flexible compared to the Dirichlet. Examples of the GD being applied in practice can be found

in Ankam (2019) and Bentahar (2015). Given that the GD can be expressed as series of conditional Beta distributions, the GD can handle missing values in the data, meaning that in a Bayesian framework we could predict the missing values given the others. However, one limitation of the GD which is not suitable for our aim is that it does not allow zero values in the compositions (or $x_i = 1$, for that matter).

Following on from this, another option we could consider is a Zero-Inflated Generalised-Dirichlet (ZIGD) (Tang et al., 2019). The ZIGD combines the GD distribution with a zero-inflation component to model data better when many zeros are present. For our proportions $\boldsymbol{x}_s = x_{s,1}, \ldots, x_{s,D}$, each $x_{s,d}$ can be zero with probability π_i or otherwise follows the GD distribution. For the ZIGD, the PDF is given as

$$f(\boldsymbol{x}_s; \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{i=1}^{D-1} (\pi_i \cdot \delta(x_{s,i}) + (1 - \pi_i) \cdot f_{GD}(x_{s,i})), \tag{5.3.0.6}$$

where $B(\alpha_i, \beta_i)$ is the beta function. The term $\pi_i \cdot \delta(x_i)$ introduces the probability mass at zero allowing x_i to take zero values with probability π_i. The term $(1 = \pi_i) \cdot f_{GD}(x_{s,i})$ ensures that the component follows the GD distribution when it is not zero. In this case, when $\beta_i = \alpha_{i+1}, \forall i$, the GD distribution reduces to the zero-inflated Dirichlet distribution. Tang et al. (2019) give an overview of the ZIGD along with an application to microbiome compositional data presenting advantages of the proposed method. However within a zero-inflated approach, the zero-generating process is separated from the continuous distribution, which may not always suit the purpose of the analysis. In the zero-inflated model, zeros are treated qualitatively as a distinct category and not explicitly as a value that is only slightly smaller than the smallest non-zero values (e.g. 0.01). The inflation of zeros is arbitrary, meaning that zeros could be replaced in both the data and likelihood by any number (e.g. 31) without changing the results. In most real-world contexts of spatial modelling, we believe it is preferable to create a continuum where the model recognises that a data point of 0.01 is more similar to 0 than it is to 0.5.

A third option is to convert the proportions into conceptual counts and assume a Generalised-Dirichlet-Multinomial (GDM) distribution, as proposed by Stoner et al. (2020c). Here, we transform the proportions $\boldsymbol{x}_s$ into counts $\boldsymbol{v}_s = v_{s,1}, \ldots, v_{s,D}$ through choosing an artificial total count N and computing $v_{s,d} = \lfloor N/y_d \cdot x_{s,d} \rfloor$. Note that in Stoner et al. (2020c), $y_s = 1$ and thus it is not written in the floor function. Use of the floor function instead of rounding ensures that the created counts do not sum to more than N. We can then assume a GDM distribution for $\boldsymbol{v}_s$:

$$\boldsymbol{v}_s \sim \text{GDM}(\boldsymbol{\mu}_s, \boldsymbol{\phi}_s, N). \tag{5.3.0.7}$$

If our original $\boldsymbol{x}_s$ are already compositional counts, we can leave them as they are and simply use the totals y_s in place of N. As explored in Chapter 4, the GDM (Stoner et al., 2020b) can in general be a versatile solution for adapting to different compositional count data structures.

Recall from Chapter 4, Section 4.4, that the GDM arises as a mixture of a Multinomial and a GD distribution such as:

$$\boldsymbol{v}_s | \boldsymbol{x}_s, N \sim \text{Multinomial}(\boldsymbol{x}_s, N), \tag{5.3.0.8}$$

$$\boldsymbol{x}_s \sim \text{GD}(\boldsymbol{\alpha}, \boldsymbol{\beta}), \tag{5.3.0.9}$$

for the compositional proportions $\boldsymbol{x}_s$ for spatial location s. The marginal probability mass function of the GDM is then

$$p(v_1, \ldots, v_d | \boldsymbol{\alpha}, \boldsymbol{\beta}, n) = \frac{\Gamma(n+1)}{\Gamma(v_d+1)} \prod_{i=1}^{d-1} \frac{\Gamma(v_i + \alpha_i)\Gamma\left(\sum_{j=i+1}^{d} v_i + \beta_i\right)}{B(\alpha_i, \beta_i)\Gamma(v_i+1)\Gamma\left(\alpha_i + \beta_i + \sum_{j=i}^{d} v_j\right)}. \tag{5.3.0.10}$$

The GDM has the ability to model sparse data where both zero and one outcomes occur with complex dependencies and overdispersion. The Multinomial part contributes some variability to $\boldsymbol{v}_s$, which can be flexibly added to through the GD component, to capture different variance patterns in real-world data well. Additionally, the GDM can handle missing values in the compositions.

Stoner et al. (2020c) applied the GDM to global household fuel data to estimate trends in the use of polluting and clean fuels for cooking. It has also been applied in Stoner et al. (2020b) to model reporting delays in time series of infectious disease counts. In the case of Stoner et al. (2020c), the flexibility of the GDM was intended to account for variability from non-representative household survey designs, and its generality allowed the integration of hierarchical structures to estimate trends in the fuel usage.

With respect to the creation of artificial counts $\boldsymbol{v}_s$ from continuous compositions $\boldsymbol{x}_s$, the choice of the total count N is to some extent arbitrary - it controls the "resolution" of predictions from the model (i.e. the number of decimal places), and smaller total counts will mean the Multinomial is contributing more variance in the model. Stoner et al. conducted a simulation experiment which demonstrated that as N increases, the accuracy of this conversion to counts converges to the accuracy achieved by modelling original data directly. The experiment suggested that an N of at least 10,000 yields nearly identical parameter inferences as using the original data. The choice of the total count here can also be seen to be equivalent to the decision made in other works of how much of a small number to add to any zeros in the data.

5.3.1 General framework

Recall that our plan is to create artificial counts $\boldsymbol{v}_s$ with total N from $\boldsymbol{x}_s$ if they are continuous compositions $\boldsymbol{x}_s$ (or alternatively leave them as they are if they are already counts).

Our general framework assumes that the counts $\boldsymbol{v}_s$ arise from a GDM distribution, given the total N (or y_s in the case where $\boldsymbol{x}_s$ are already count data):

$$\boldsymbol{v}_s \sim \text{GDM}(\boldsymbol{\mu}_s, \boldsymbol{\phi}_s, N). \tag{5.3.1.1}$$

As outlined in Chapter 4, Section 4.4, the GDM is parameterised in terms of $\boldsymbol{\mu}_s$ and $\boldsymbol{\phi}_s$, which relates to the expression of the GDM as a series of conditional Beta-Binomial models for each count composition up to and including $v_{s,D-1}$ (the last count composition $v_{s,D}$ is given implicitly as $y_s - \sum_{i=1}^{D-1} v_{s,i}$):

$$\begin{aligned}
v_{s,1}|N &\sim \text{Beta-Binomial}(\mu_{s,1}, \phi_{s,1}, r_{s,1} = N), \\
v_{s,2}|N, v_{s,1} &\sim \text{Beta-Binomial}(\mu_{s,2}, \phi_{s,2}, r_{s,2} = N - v_{s,1}), \\
&\vdots \\
v_{s,d}|N, v_{s,d} &\sim \text{Beta-Binomial}\left(\mu_{s,d}, \phi_{s,d}, r_{s,d} = N - \sum_{l<d} v_{s,l}\right).
\end{aligned} \tag{5.3.1.2}$$

Here, the $\mu_{s,1}, \ldots, \mu_{s,D-1}$ are the means of the Beta-Binomials and $\phi_{s,1}, \ldots, \phi_{s,D-1}$ the variance parameters. Each $\mu_{s,d}$ is strictly between 0 and 1 and there is no sum constraint. The derivation of the Beta-Binomial for the GDM is given in Chapter 4, Section 4.4.

Next, we assume that μ_d and ϕ_d vary over the spatial locations s according to general functions $f_d(\cdot)$ and $g_d(\cdot)$ of spatial location, i.e.

$$\text{logit}(\mu_{s,d}) = f_d(s), \tag{5.3.1.3}$$

$$\log(\phi_{s,d}) = g_d(s). \tag{5.3.1.4}$$

These functions capture spatial variation in the mean and covariance structure of the data, respectively. Either could include, for instance: spatial covariate effects, random effects, spatial autocorrelation models or point processes. In the next subsection, we will propose a version of this framework based on spatial penalised regression splines.

5.3.2 Spatial penalised regression splines

Penalised regression splines are a powerful tool for smoothing and modelling complex data structures. They extend traditional spline regression by incorporating a penalty term that controls the smoothness of the fitted curve. This penalty helps to avoid overfitting, ensuring that the model captures the underlying trend without being overly sensitive to noise in the data. This approach is particularly useful in Generalised Additive Models (GAMs) to model nonlinear effects of covariates on the response. Full details of regression splines can be found in Hastie (2017).

A spline is a piecewise polynomial function defined over a sequence of knots, with continuity constraints ensuring smooth transitions between segments. For example, a cubic spline can be constructed as:

$$S_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3. \tag{5.3.2.1}$$

A penalty term is then added to the spline function to penalise excessive wiggliness. The penalised least squares problem outlined in Hastie (2017) is to minimise:

$$\sum_{i-1}^{n}\left(y_i-\sum_{j=1}^{p}f_j(\boldsymbol{x}_{ji})\right)^2+\sum_{j=1}^{p}\lambda_j\int(f_j^{''}(\boldsymbol{x}_j))^2dx_j, \tag{5.3.2.2}$$

where the first term measures the goodness of fit, and the second term penalises the roughness of $f(x)$. The tuning parameter λ controls the trade-off between fit and smoothness. A small λ allows a more flexible wiggly function, whereas a large λ produces a smoother function, closer to a linear trend. The roughness penalty measures the integrated squared second derivative of $f(x)$ which reflects the curvature of the function. In the context of GAMs, the penalised regression spline approach models the additive predictor η as

$$\eta=\beta_0+\sum_j f_j(x_j), \tag{5.3.2.3}$$

where each $f_j(x_j)$ is a smooth function estimated using penalised splines. This framework is extensively discussed within Wood (2017). Penalised regression splines have been widely used for modelling spatial data such as Sangalli et al. (2013) to model irregularly shaped spatial domains and Fahrmeir et al. (2004) who apply a penalised spline GAM to analyse space-time regression data, incorporating both spatial and temporal effects.

A key advantage of applying this approach is in its flexibility in selecting basis functions to accommodate different data structures and modelling needs. These include commonly used spline bases, such as thin-plate and cubic regression splines, which assume isotropy and smoothness over a continuous spatial domain, as well as basis functions designed specifically for areal data, such as Gaussian Markov random field smooths.

The Bayesian framework for penalised regression splines offers a probabilistic perspective on smoothing by interpreting the smoothing penalty as a prior distribution over the spline coefficients, i.e.

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \lambda^{-1}\boldsymbol{K}^{-1}), \tag{5.3.2.4}$$

where $\boldsymbol{\beta}$ represents the spline coefficients, λ is the smoothing parameter and $\boldsymbol{K}$ is the penalty matrix. The penalty matrix $\boldsymbol{K}$ captures the complexity of the spline basis functions and penalises higher-order derivatives to enforce smoothness. This ensures that the spline fit is smooth by penalising large variations in the coefficients $\boldsymbol{\beta}$. This approach is detailed in Wood (2016), which provides a comprehensive Bayesian formulation for GAMs. Wood looks at automatically and reliably generating JAGS (Plummer, 2003) model specification code and implementing any GAMs in R (R Core Team, 2021). Generally, our function $f(x,y)$ can be expressed as

$$f(x,y) = \sum_{j=1}^{K} \beta_j b_j(x,y), \tag{5.3.2.5}$$

where the β_j are the spline coefficients and the $b_j(x,y)$ are the basis functions. Here, we assume $f_k(\cdot)$ is a penalised regression spline of space, i.e. $\text{logit}(\mu) = f_k(s) = Z_{s,r} \cdot b_r$, where r represents the basis functions.

Then, we assume a polynomial relationship between the means $\boldsymbol{\mu}_{s,d}$ and the variance parameters $\boldsymbol{\phi}_{s,d}$. The polynomial relationship for $\boldsymbol{\phi}_{s,d}$ is given as

$$\log(\phi_{s,d}) = \psi_{1,d} + \sum_{j=2}^{4} \psi_{j,d} \cdot (\mu_{s,d} - \bar{\mu}_{s,d})^{j-1}, \tag{5.3.2.6}$$

where η is the linear predictor computed using the design matrix and parametric effects from Equation (5.3.2.7). Normal priors are assigned to $(\psi_{1,d},\ldots,\psi_{4,d})$, for example, $\psi_{1,d} \sim N(2,2^2)$ and $\psi_{2,d},\ldots,\psi_{4,d} \sim N(0,1^2)$. We can give $\mu_{s,d}$ as a function of the design matrix Z along with b, the coefficients of the smooth terms for the spatial effect, e.g.

$$\begin{aligned} \mu_{s,d} &= \text{expit}(\eta_{s,d}), \\ \eta_{s,d} &= Z_{s,r} \cdot b_r, \end{aligned} \tag{5.3.2.7}$$

for $r = 1,\ldots,R$ which are the number of the basis used for the splines.

The main novel aspects of our proposed method can be summarised as (i) use of the GDM as a general modelling framework for spatial compositional data, (ii) integrating penalised regression splines of space within the GDM and (iii) designing a polynomial link between the mean and variance parameters, for an even more flexible (yet simple in terms of the number of parameters) mean-variance relationship to suit different data sets than would otherwise result from the GDM with non-varying variance parameters. Whether this improves model fit is tested later in Section 5.4.1.

In comparison to that outlined in Chapter 4, this is a general form of our proposed GDM spatial model and can be tailored for specific applications in constructing the mean and variance parameters of the GDM.

5.3.3 Tree species model

For the tree species application outlined in Section 5.1.1, each tree species is modelled by individual Beta-Binomial models. As outlined in Section 5.3 our tree species proportions ($\boldsymbol{x}_s$) can be transformed into counts ($\boldsymbol{v}_s$) (Stoner et al., 2020c). We decided to use a total count of 100 to transform the proportions into counts. This is intuitive as it directly represents percentages, making it easier to understand the transformed count data.

Let $N = 100$ be the total count of tree species in each spatial location $s = 1, \ldots, S$ and $v_{s,d}$ represent the corresponding count for the d^{th} tree species ($d = 1, \ldots, D$) at each spatial location s. Recall from Chapter 4, Section 4.4, the last composition D is modelled implicitly and does not have an associated Beta-Binomial model. When selecting four tree species - larch, oak, sitka spruce and sycamore - to model within our framework (Section 5.3.1), they are assigned to the Beta-Binomials within the GDM in that given order, with the remaining tree species being modelled implicitly as the last composition D.

From Listing 7, it can be seen that we have implemented a penalised regression spline (Section 5.3.2) with a thin plate regression basis function. This is produced through the specification of the spatial coordinates of the grid cell centroids within the `s(.)`. We also decided to use 400 basis functions for the regression spline to balance computational efficiency and flexibility, ensuring that the spline can capture the spatial patterns. This is defined inside the spline in Listing 7.

5.3.4 Implementation and prior distributions

As outlined in the previous chapters, all the code used to apply the proposed framework was written and run using R (R Core Team, 2021) and the model was implemented using NIMBLE (Valpine et al., 2017). In addition, all computations were carried out on a MacBook Air laptop with an Apple M3 chip (8 physical cores) and with 16GB system memory or on an Ubuntu Linux desktop computer with an Intel Core i9-13900K processor with 24 physical cores (32 logical cores) with 128GB system memory. It will be stated throughout the chapter which computer was used for each task.

Here, we detail how we constructed the smooth spatial regression spline $f(s)$ for integration within the general GDM framework. We do this using the *mgcv* package for R (Wood, 2003). Firstly, we fit a GAM (Hastie, 2017) to produce the design matrix `Z` for the spline basis functions. GAMs are a powerful tool to model spatial data as they are from the family of generalised linear models where the linear response variable depends linearly on some unknown smooth functions of the predictor variables. The link between these generalised linear models and the additive were originally developed by Hastie (2017) with a full overview of applying GAMs in R found within Wood (2017). However, note that we fit a GAM for convenience to generate `Z` for use in our Bayesian model, but the actual "fit" is irrelevant, `Z` depends only on the covariates and the model formula and does not depend on the values of the response variable or on the estimated coefficients.

We can fit a GAM in R using the *mgcv* package (Wood, 2003) using the code given in Listing 7. The `predict` function generates the matrix of linear predictors (`lpmatrix`) which corresponds to the smooth term enclosed in the `s(.)`. Each row of `Z` corresponds to a new observation, and each column corresponds to a different basis function used in the model. It essentially translates the smooth term in the model into a matrix form.

```
spatial_gam_model <- gam(count/total ~
                         s(X_coord, Y_coord, k=r),
                          family = "binomial",
                          weights = total,
                          data = spatial_data)

Z <- predict(spatial_gam_model, spatial_data, type = "lpmatrix")
```

Listing 7: Spatial `gam` model code using the Binomial distribution to predict the design matrix Z for the GDM where `X_coord`, `Y_coord` are the spatial coordinates and `r` the number of basis functions.

Next, the `jagam` function (from *mgcv* package) is run to extract the penalty matrix associated with the smooth term, `S1`. Unlike the `gam` model from Listing 7, the `jagam` function does not fit any model explicitly but returns standard spatial setup information, such as the penalty matrix. We fit the `jagam` using the Binomial distribution to model the counts as follows:

```
spatial_jagam_model <- jagam(count/total ~
                          s(X_coord, Y_coord, k=r),
                             family = "binomial",
                             weights = total,
                             data = data,
                             file = "spatial_model.jags")

S1 <- spatial_jagam_model$jags.data$S1
```

Listing 8: Spatial `jagam` model code using the Binomial distribution for compositional counts where `X_coord`, `Y_coord` are the spatial coordinates and `r` the number of basis functions.

However, the Binomial family is a placeholder, as the model matrix and penalty matrix only depend on the covariate values and the right-hand side terms specified in the formula, not the response variable values or the distribution family.

This function also writes a JAGS (Plummer, 2003) model specification to the `.jags` file listed in `file` argument of the function. This output file forms the basis for creating the spatial GDM model which we translate into our NIMBLE code. Further in-depth information about `jagam` can be found in Wood (2016).

We can therefore specify a NIMBLE model given in Listing 9 for the tree species compositional data. Note, the log probability density function for each Beta-Binomial is computed using the same custom "NIMBLE function" outlined in Chapter 4, Listing 5.

For the terms that account for the spatial structure in the data, λ_i for $i = 1,2$ have a Gamma$(0.05, 0.005)$ prior placed upon them, as can be seen in Line 9 from Listing 9. Here, $\boldsymbol{\lambda}$ controls the smoothness of the effects for each tree species. The parametric effects $b_{r,k}$ have Normal priors assigned, where $b_1 \sim N(0, 10^2)$ and $b_r \sim N((0), \mathrm{K1}_{r-1,r-1})$ for $r > 1$ for r the number of basis functions chosen for the spline. $K1$ is constructed from the penalty matrix $S1$ and the smoothing parameters $\boldsymbol{\lambda}$ which represent a regularised version of the smoothing penalty matrix that incorporates both the individual smoothing of the spline terms and their interactions. $K1$ is computed in Listing 9 in Lines 17-19 with the Normal distribution placed on $\boldsymbol{b}$ in Lines 14-15.

```
spatial_GDM_model <- nimbleCode({

    ## FOR EACH TREE SPECIES ##
    for (d in 1:N_types) {

        ## PRIORS ##
        for (i in 1:2) {lambda[i, d] ~ dgamma(0.05, 0.005)}
        psi[1, d] ~ dnorm(2,sd=2)
        for(j in 2:4){psi[j, d] ~ dnorm(0,sd=1)}

        ## SPATIAL LOCATIONS ##
        K1[1:(r-1), 1:(r-1), d] <- S1[1:(r-1), 1:(r-1)] *
                lambda[1, d] + S1[1:(r-1), r:((r-1)*2)] *
                lambda[2, d]
        ## PARAMETRIC EFFECT ##
        b[1, d] ~ dnorm(0, sd = 10)
        b[2:r, d] ~ dmnorm(zero[1:(r-1)], K1[1:(r-1), 1:(r-1), d])

        ## LINEAR PREDICTOR ##
        eta[1:S, d] <- Z[1:S, 1:r] %*% b[1:r, d]

        ## MEAN PARAMETER ##
        mu[1:S, d] <- expit(eta[1:S, d])
        mu_mean[d] <- mean(mu[1:S, d])

        ## LOOP OVER SPATIAL LOCATIONS ##
        for (i in 1:S) {
            ## VARIANCE PARAMETER ##
            log(phi[i, d]) <- psi[1, d] +
                psi[2, d]*(mu[i, d]-mu_mean[d]) +
                psi[3, d]*(mu[i, d]-mu_mean[d])^2 +
                psi[4, d]*(mu[i, d]-mu_mean[d])^2}
    }
    ## MODEL COUNTS V ##
    for (i in 1:S) {
        ## FIRST SPECIES ##
        v[i, 1] ~ dbetabinomial(mu[i, 1], phi[i, 1], y[i])
        ## LOOP OVER THE REMAINING SPECIES ##
        for (d in 2:N_types) {
            v[i, d] ~ dbetabinomial(mu[i, d], phi[i, d],
                                    y[i] - sum(v[i, 1:(d-1)]))
        }}
})
```

Listing 9: Custom `R NIMBLE` model code to fit the multivariate spatial GDM model which uses the `R` "NIMBLE function" code for the `dbetabinomial` function from Chapter 4 Listing 5.

Given the flexibility that using NIMBLE provides, we can select different samplers to use to aid the convergence of the parameters. Regular slice samplers were placed on the smoothing parameters λ_i for $i = 1, 2$ which work by iteratively sampling from a region under the target density function, ensuring efficient exploration of the parameter space without requiring gradient information (Neal, 2003). The Automated Factor Slice Sampler (AFSS) (Tibbits et al., 2014) is a variant of the slice sampler designed to sample efficiently from high-dimensional and structured distributions. The AFSS breaks up the target distribution into factors, allowing it to explore the parameter space more effectively by sampling along specific directions that align with the structure of the distribution. The AFSS is particularly useful for complex models with correlated parameters or hierarchical structures, where traditional slice samplers might struggle. We used the AFSS to sample efficiently the parametric effects $\boldsymbol{b}$ and the flexible mean-variance relationship $\boldsymbol{\phi}$ to reduce the number of MCMC iterations and overall computation time needed for convergence.

The convergence of MCMC chains for all models is assessed following the procedure outlined in Appendix B. In brief, this includes visually inspecting the traceplots to confirm that all chains are stationary and overlapping with each other, and computing the potential scale reduction factor (Gelman et al., 1992) for each parameter; where we assume that the chains have converged if PSRFs are less than 1.05.

5.4 Application to tree species data

In this section, we apply our proposed framework from Section 5.3 to a spatial tree species dataset. These data contain the estimated proportions of each tree species for each spatial location within the grid.

We assessed our model through a combination of in-sample and out-of-sample posterior predictive checking, using Monte Carlo simulation, in two experiments. These checks allow us to assess how good our model is at producing data values with similar characteristics to the original data values, some of which we remove from the model training data (i.e. we set them to NA) to gain an insight into out-of-sample prediction. The two experiments are detailed in the next two subsections. Before that, we will describe the posterior predictive checking process.

We created simulation functions that use the parameter samples from each iteration of the model to simulate a new data set for each MCMC iteration. For each sample from the model, we run the simulation function with the specific combination of parameters (e.g. $\boldsymbol{\mu}_{s,d}$ and $\boldsymbol{\phi}_{s,d}$) produced by the model for that iteration. This process is repeated for the total number of MCMC iterations, and the resulting simulated counts are referred to as replicates or replicate data. These replicates are synthetic datasets that reflect the uncertainty in the model's parameter estimates. Recall, this is the same process detailed in Chapter 4, Section 4.7.2 for posterior predictive checking. Within this instance, for each saved MCMC iteration the procedure for generating replicate tree species counts from the GDM model is as follows:

1. Simulate new case counts for the first tree species (larch) from the first Beta-Binomial in Equation (5.3.1.2), using the samples of $\mu_{s,1}$ and $\phi_{s,1}$ from that MCMC iteration.
2. For the next tree species, we compute the counts of $v_{s,d}$ (Equation (5.3.1.2)), i.e. the remainder of the total counts $y_{s,d}$ not yet accounted for by the simulated counts of the previous tree species in the ordering. Then, we simulate new counts for this tree species from the corresponding Beta-Binomial.
3. Repeat Step 2 until counts have been simulated for all tree species (oak, sitka spruce and sycamore) from all Beta-Binomials in Equation (5.3.1.2).

4. Lastly, the final count $v_{s,D}$ (in this case the aggregate count for the remaining tree species) can be computed as $v_{s,D} = N - \sum_{k=1}^{D-1} v_{s,k}$, where N is the total count. If the original compositions $v_{s,D}$ are counts and not proportions then $N = y_s$.

This results with the same number of replicate counts as we have MCMC iterations, i.e. if we have 1,000 MCMC iterations we will have 1,000 replicate datasets.

After producing the replicate data, we can assess how well the replicates match the original data by comparing their distributions. One way to do this is by computing a relevant statistic (e.g. mean, median or standard deviation) for each replicate dataset and visualising the distribution using density plots, with the corresponding statistic from the original data overlaid. This helps determine whether the observed statistic is plausible within the distribution of the replicates. Alternatively, we compute quantiles (from 0.01 to 0.99 in increments of 0.01) for each replicate dataset. Specifically, for each tree species within each MCMC iteration, we compute the quantiles of the replicate counts resulting in one sample per quantile per iteration. We then summarise these quantiles across all MCMC iterations by computing the median and the lower (2.5th percentile) and upper (97.5th percentile) bounds for each quantile level. This provides us with a comprehensive view of the distribution and spread. A quantile-quantile plot comparing the replicate data quantiles against those from the original data, along with 95% intervals, offers another visual assessment of the model fit. Ideally, the quantiles of the original data should fall within the 95% interval of the replicates, while the median posterior predictive quantiles should closely follow the original data quantiles, aligning with the diagonal line.

Additionally, we can compute summary statistics to quantify the point estimate (e.g. posterior median) prediction accuracy for out-of-sample data. This includes computing the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) of the mean across the replicates compared to the original data values at each spatial location. Furthermore, we calculated the MAE for the median posterior predictive quantile values with respect to the original data quantile values. This provides an indication of how well the quantiles of the replicates fit the original data, which is useful for comparing different model variations.

Next, we computed the uncertainty intervals for each data point and computed coverage rates to verify these intervals. In our Bayesian modelling context, by "coverage" we refer to the percentage of times that a posterior predictive interval with a given level of uncertainty contains the original data value. For example, ideally, if we compute 95% uncertainty intervals for all data points, 95% of them should include the original data value. If the coverage rate is below 95%, the model is too confident. If the rate is above 95%, the model is too uncertain. Following on, we can examine the mean width of the intervals. This is a useful measure for comparing models. If two models each have appropriate coverage rates, we prefer the model with lower mean widths as it indicates greater prediction precision while still providing correct coverage.

5.4.1 Single tree prediction experiment

We first tested our proposed model for a single tree species, larch, as an initial step in the development of the full multivariate model, in particular to inform a model for $\boldsymbol{\phi}_{s,d}$ that achieved a suitable model fit. We thus compared the model given in Section 5.3.2, where $\boldsymbol{\phi}_{s,d}$ are a polynomial function of $\boldsymbol{\mu}_{s,d}$ (Equation 5.3.2.6), to an alternative model where $\boldsymbol{\phi}_{s,d}$ does not vary in space, i.e. there is a single $\boldsymbol{\phi}_{s,d}$ for all spatial locations - $\boldsymbol{\phi}_{s,d} \sim \text{Gamma}(2, 0.05)$.

To compare the two models, we carried out an out-of-sample prediction experiment to assess our model's performance in predicting the unseen larch counts. Within our cross-validation setup, we split the data into two approximately equal parts, this resulted in 1,077 data points in the "train data" and 1,076 within the "test data". The training data are used for modelling and the test data are excluded. Heatmaps for the respective train and test data for the tree larch are given in Figure 5.2.

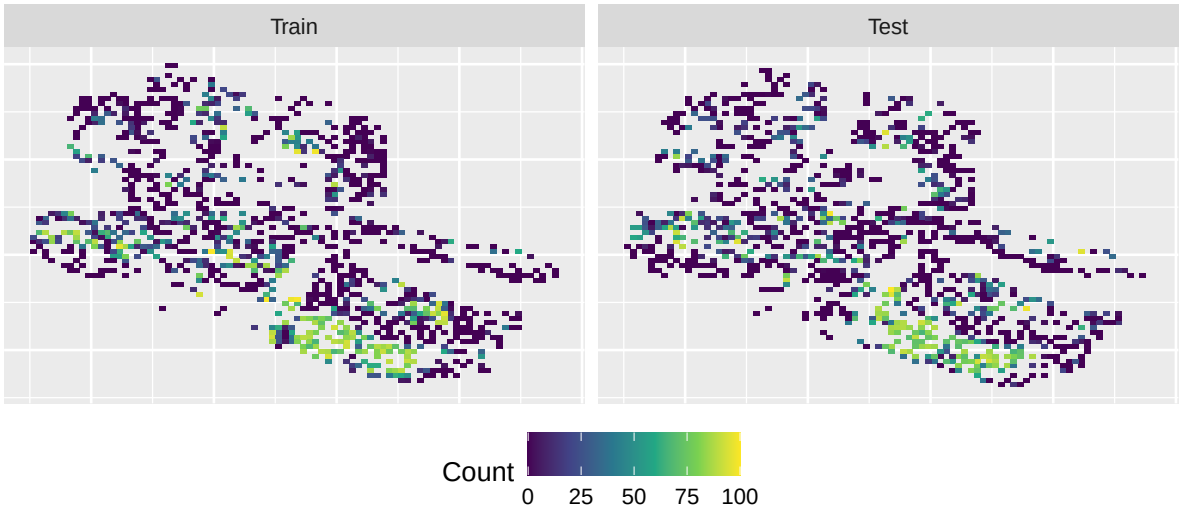


Figure 5.2: Heatmaps of the counts of Larch within each grid cell showing the split of grid cells by train and test.

Following fitting each model (GDM-fixed and GDM-polynomial), we can run the simulation function provided in Listing 10, where how the function is run to produce the replicate data shown in Lines 24-30. From the listing, the function given takes on the inputs: N, y, μ and ϕ, where N is the number of spatial locations (a vector of length 1), y is the total count (a vector of length N) and lastly, μ and ϕ are the outputted parameters from the fitted model (matrices with a dimension of the number of samples $\times$ N). Lines 11-15 show the function simulating the new counts from the Beta-Binomial distribution. The resulting matrix produced from the function is of dimension S $\times$ number of samples which can then be analysed to see how closely they fit the characteristics of the train and test data. This

function can be used with both the $\phi_{s,d}$ specifications with the only difference seen in the prior extraction from the model before running the function, as the two $\phi_{s,d}$ specifications differ in dimension. When producing replicates for the train and test data the only difference is seen in the number given to the function as N.

```r
### SIMULATE NEW COUNTS ####
simulate_tree_count <- function(S, y, mu, phi) {

    # INITILISE TO STORE SIMULATION
    simulated_data <- array(0, dim = c(S))

    # SIMULATE COUNTS
    observation <- rbinom(S, y, rbeta(S, mu*phi, (1-mu)*phi))

    # SAVE SIMULATED VALUES
    simulated_data <- observation

    return(simulated_data)
}

### RUN FUNCTION ####
simulation_replicates <- array(NA, dim = c(S, N_samples))
for (i in 1:N_samples){
    simulation_replicates[, i] <- simulate_tree_count(
                                                S = S,
                                                y = y,
                                                mu = mu_samples[i, ],
                                                phi = phi_samples[i, ])
}
```

Listing 10: Custom `R` function code to simulate replicates from a single tree species spatial compositional model.

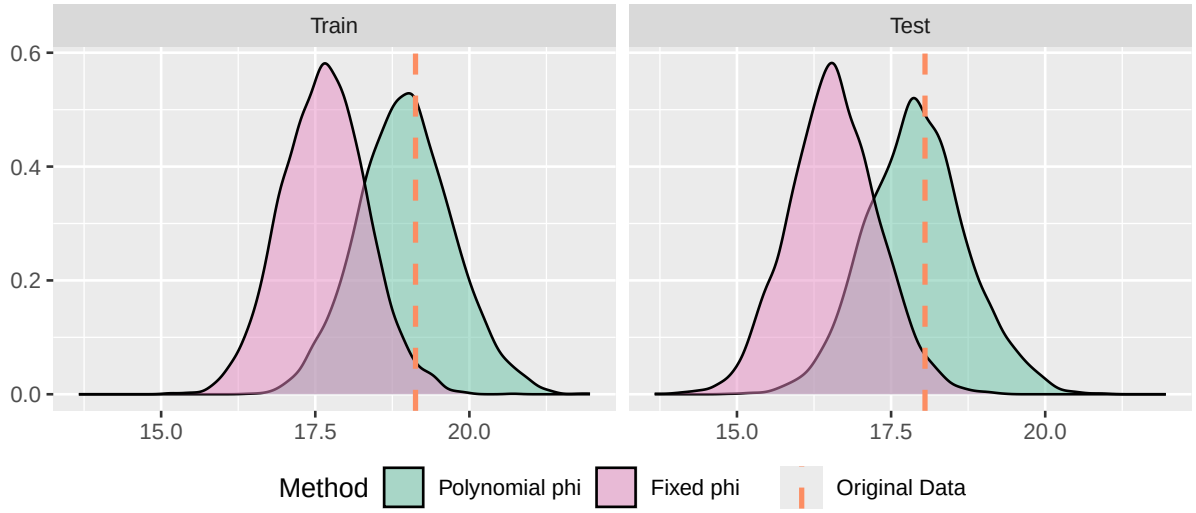
For both the single tree species spatial GDM-polynomial and GDM-fixed models, (run on an Apple MacBook Air M3), we ran four chains in parallel for 2,000 MCMC iterations, with a burn-in of 1,000 discarded in each chain. This took approximately 2 hours per model. We computed the PSRF for each parameter resulting in 88% of the PSRFs for both models being less than or equal to 1.05, respectively, both with a median of 1.01, indicating convergence.

5.4.1.1 Single tree prediction experiment results

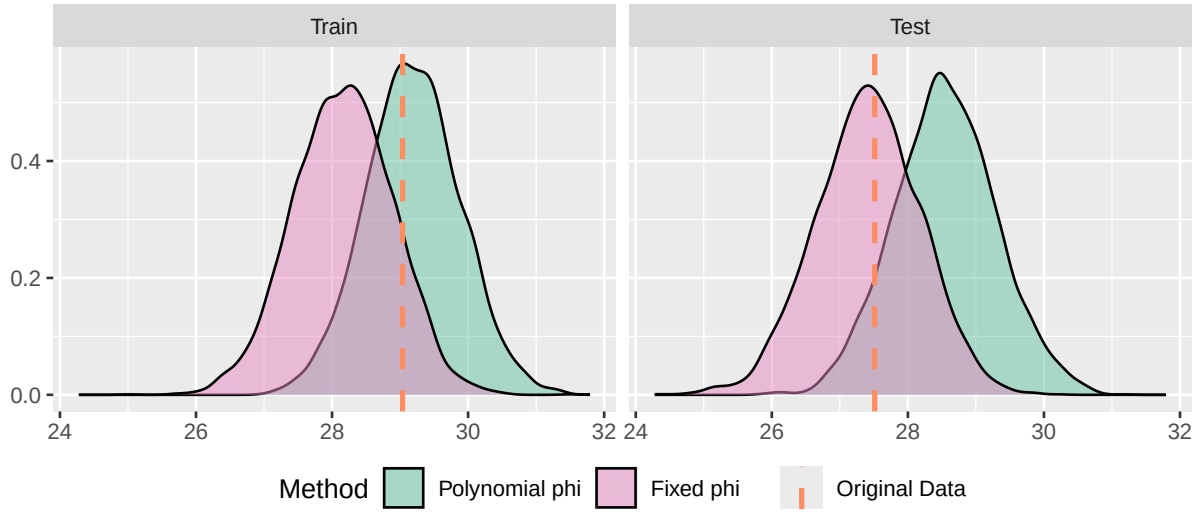
The visual inspection of the two models fitted, GDM-polynomial and GDM-fixed, is given in Figure 5.3, which shows the density of the distribution of both the mean and standard deviation of the replicate values against the original data values given by the vertical dashed lines. Overall, the GDM-polynomial outperforms the GDM-fixed for both the train and test data in accurately capturing the overall means and standard deviations, as its peak aligns with the original data statistics better. The only exception to this is the distribution of the standard deviation of the replicates from the test data, where the original data line occurs at the peak for the GDM-fixed model. However, for the GDM-polynomial model, the density curve still includes the original data lines within the bulk of the distribution, suggesting that the standard deviations of the original data are plausible values within the distribution.

This is supported by the quantile plots (Figure 5.4); for both the train and test data, the original data quantiles follow the quantiles from the GDM-polynomial more closely and remain within the 95% interval, with only a slight tail outwith this region seen for the test. It can be noticed that for the GDM-fixed, the original data quantiles depart from the 95% interval with a clearer deviation for the test data. This indicates that, when assessing visual comparisons, the GDM-polynomial fits the distribution of both the train and test data significantly better for the single tree model.

To quantify what can be seen visually, we compute various summary statistics given in Table 5.2. For the training data, the point prediction accuracy of the GDM-polynomial and GDM-fixed models is very similar with only a 0.27 and 0.04 difference in MAE and RMSE values, respectively, in favour of the GDM-fixed. A greater difference between the two models is observed when comparing the MAE of the median quantile values to the original data quantiles, with a percentage difference of 58%. This indicates that for the training data, the GDM-polynomial quantile values more closely align with the original data quantiles,



(a) Posterior predictive sample mean



(b) Posterior predictive sample standard deviation

Figure 5.3: Density plots for the GDM-polynomial and GDM-fixed models for the posterior predictive sample mean and standard deviation for the train and test data for Larch. The different coloured densities correspond to each of the methods: GDM-polynomial and GDM-fixed with the original data given by the vertical line.

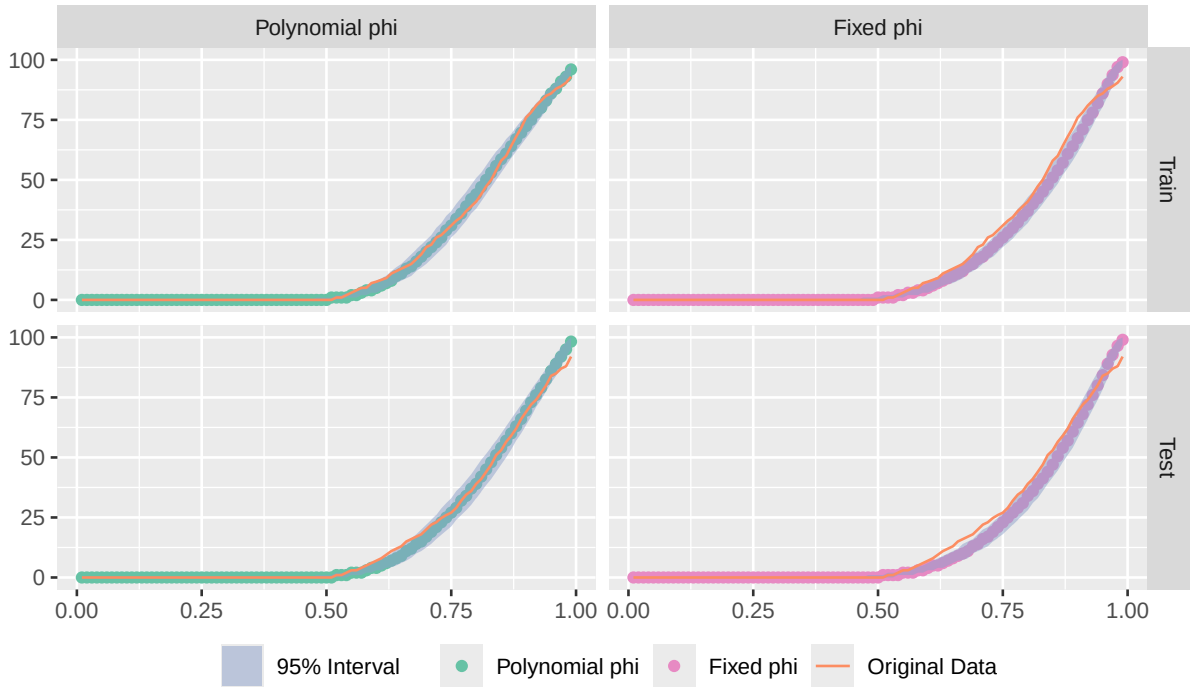


Figure 5.4: Quantile plots for the GDM-polynomial and GDM-fixed models for the train and test data for Larch. The points represent the median posterior predictive quantile values for each method: GDM-polynomial and GDM-fixed. The values for the original data quantiles are indicated by the lines. The shaded areas represent the 95% intervals associated with each quantile.

corroborating what is seen in Figure 5.4. A similar picture is noticed for the test data, where there are very similar MAE and RMSE values for both the GDM-polynomial and GDM-fixed when compared to the original data values. Again, when considering the MAE of the median quantile values, the GDM-polynomial has far superior performance with a MAE value 53% less than the GDM-fixed. Therefore, in terms of point prediction accuracy both models present a similar picture with the GDM-polynomial demonstrating a slight improvement.

To assess point prediction uncertainty, we examine coverage rates for the uncertainty intervals and the mean widths associated with each interval for 80%, 85%, 90%, 95% and 99%. Coverage rates, as presented in Table 5.3, indicate the percentage of times that a posterior predictive interval with a given level of uncertainty contains the original data value, defined in Section 5.4. We investigate how close the coverage rate is to the associated percentage of the uncertainty interval. Ideally this value should be similar otherwise the model could be too confident or uncertain. It can be observed that for all the intervals for both the train and test data, the coverage rate for both the GDM-polynomial and GDM-fixed models is far greater than their associated uncertainty percentage, i.e. for the 95% uncertainty interval, the coverage for the GDM-polynomial and GDM-fixed are 99.2% and 98.9%, respectively. This could mean that our models are too uncertain. Therefore, we can examine the mean width of each interval. It is evident that the GDM-polynomial has larger interval mean widths than the GDM-fixed, suggesting that the GDM-polynomial may be too uncertain as indicated by the wide intervals. However, for the two models the mean widths of the intervals for smaller percentage of uncertainty are more similar. For example, the mean widths of the 80% uncertainty interval are 34.3 and 33.1, for the GDM-polynomial and GDM-fixed models respectively. As we examine a higher percentage of uncertainty intervals, the difference in mean widths of the intervals between the two models also increases, i.e. for the 80% uncertainty interval for the train data, the difference in the mean widths of the intervals is

1.2 percentage points but for the largest interval (99%) the difference is 10.6. We can identify a similar pattern for the test coverage rates too. Overall, this implies that the GDM-fixed model has greater prediction precision and is effectively capturing the uncertainty in the predictions.

Table 5.2: Summary statistics quantifying the prediction accuracy for the GDM-polynomial and GDM-fixed models.

Statistic	**Train**		**Test**	
	GDM-polynomial	GDM-fixed	GDM-polynomial	GDM-fixed
MAE	10.11	9.84	12.08	11.88
RMSE	15.29	15.25	18.37	18.64
Quantile-MAE	0.84	1.98	0.96	2.06

Table 5.3: Bayesian coverage of the uncertainty intervals and their associated mean widths for the GDM-polynomial and GDM-fixed models.

Uncertainty Interval		**Train**		**Test**	
		GDM-polynomial	GDM-fixed	GDM-polynomial	GDM-fixed
80%	coverage	95.2%	94.9%	87.3%	86.1%
	mean width	34.3	33.1	33.5	31.9
85%	coverage	96.3%	96.5%	91.3%	88.9%
	mean width	39.9	37.7	39.1	36.4
90%	coverage	97.6%	97.9%	94.2%	92.6%
	mean width	47.7	43.8	46.8	42.7
95%	coverage	99.2%	98.9%	97.4%	96.2%
	mean width	60.7	53.8	59.6	52.8
99%	coverage	99.7%	99.9%	99.7%	99.4%
	mean width	83.5	72.9	82.6	72.4

After conducting a cross-validation prediction experiment to assess and compare the performance of a polynomial relationship for the variance $\phi_{s,d}$ versus a single fixed value, it can be concluded that both models provide similar fits both in-sample and out-of-sample. Therefore, both setups for $\phi_{s,d}$ could be sensible to use in practice but we will use the GDM-polynomial moving forward given the advantages it has by capturing the distribution better. We will now consider the GDM-polynomial model for a multi-tree setting, with more tree species included.

5.4.2 Multi-tree prediction experiment

As our aim is to model spatial compositional data effectively, here we move on a multi-tree problem, now only using the polynomial $\phi_{s,d}$ version of our model shown in Section 5.4.1.

To reduce the complexity of fitting the GDM to the full multivariate tree species data, we decided to select four tree species of interest - larch, oak, sitka spruce and sycamore. We also randomly select a reduced data set of 1,000 spatial locations from the original 2,153 locations. This resulted in a data matrix of dimension 1000×4. The four tree species have a varied levels of counts over the whole spatial area with each having a hotspot in different sections of the grid.

We then designed a prediction experiment to assess if the model is effective at predicting missing values in the compositional counts. As we had no missing values in the data originally, we decided to "poke holes" in our 1,000 randomly selected spatial locations by making some of the observed counts missing (NA). To setup this we follow these steps:

1. Randomly select 1,000 spatial locations from the original data.
2. From these 1,000 spatial locations:
    - Randomly sample 200 spatial locations to introduce one missing tree species.
    - From the remaining spatial locations, select another 200 spatial locations which will have two missing tree species.
    - Finally, from the remaining 600 spatial locations, randomly select another 200 to introduce three missing tree species.

This results in a dataset with 400 spatial locations which have no missing components, then 200 spatial locations with one missing component, 200 spatial locations with two missing components and 200 spatial locations with three missing components. The resulting data contain 50.2%, 43.6%, 79.7% and 53.8% of zeros within each tree species larch, oak, sitka spruce and sycamore, respectively. Additionally, as we have introduced missing components, we have 29.1%, 30.7%, 30.7% and 29.5% of missing counts for larch, oak, sitka spruce and sycamore, respectively. The associated heatmap for the four tree species with the missing counts is shown in Figure 5.5 with the missing spatial locations clearly outlined by the orange grid cells.

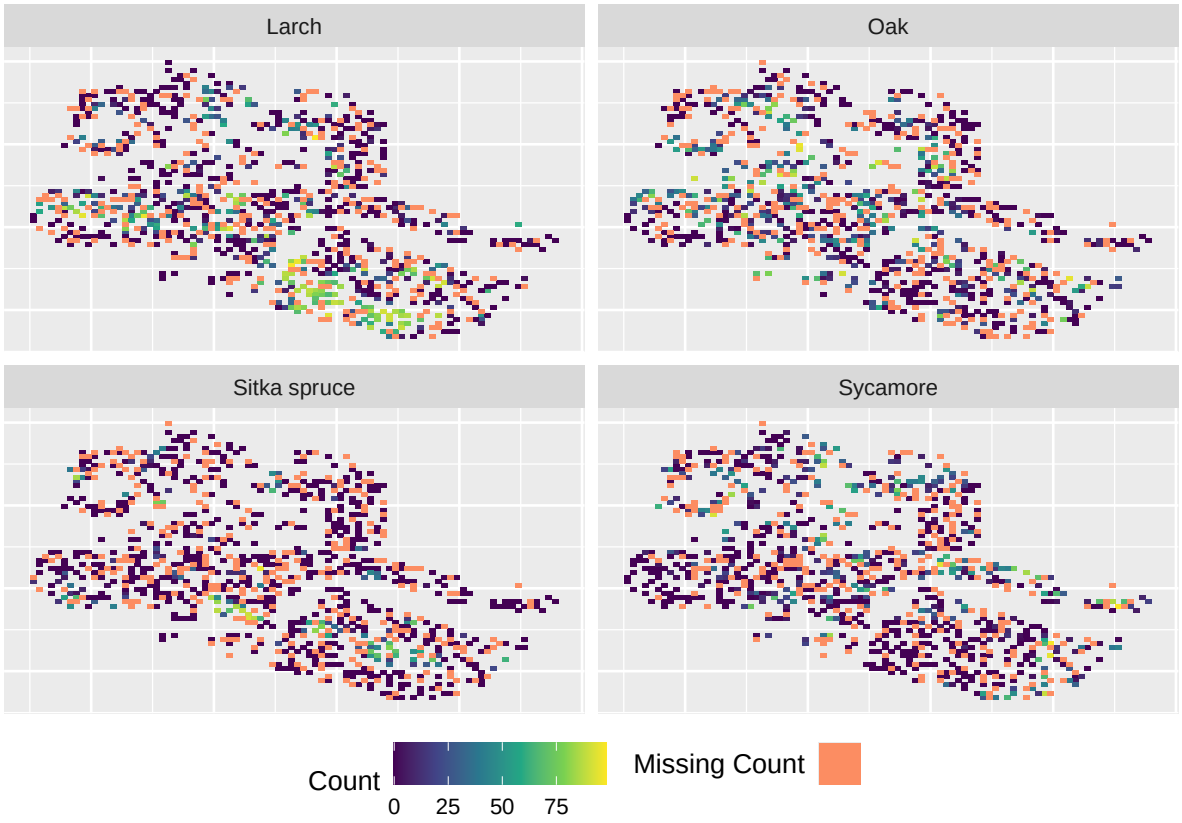


Figure 5.5: Heatmaps of the counts of the tree species: Larch, Oak, Sitka spruce and Sycamore; within each grid cell from the 1,000 randomly selected locations for the GDM. The spatial locations which contain a missing count is shown by the orange grid cells.

Once the multivariate spatial compositional data with missing values has been created, the GDM model can be implemented as in Section 5.3.4 with the custom `R NIMBLE` model code from Listing 9. One change to the model code when the response contains missing values is that unobserved $\boldsymbol{v}_{s,d}$ are sampled by the MCMC algorithm using slice samplers, and we

save the samples generated by the algorithm. This means that we no longer need to simulate new counts through a function (such as for the single tree experiment in Listing 10) but we extract them from the model output instead (since those samples are posterior predictive samples). In addition, as our data now contains NA values, the model requires additional information to sample these missing values. In this case, code is added to the GDM to compute initial values for $\boldsymbol{v}_{s,d}$ where the counts are missing and assigns NA values where data are available. When the model is compiled, it integrates the computed initial values with the data, resulting in a complete $\boldsymbol{v}_{s,d}$ for the GDM that will be updated with sampling.

Similarly to the single tree experiment, we assessed the fit by producing replicate values for each tree species using the same Monte Carlo simulation method outlined in Section 5.3.3. The adapted code for the multivariate case is given in Listing **??**, it should be noted here that the values of S and y are that of the reduced data of the randomly selected rows of the data that the model was fitted to, i.e. $S = 1000$.

We ran the GDM model on a computer with Intel Core i9-13900K processor for 10,000 iterations with 5,000 discarded for burn-in, with four chains run in parallel. This took approximately 16 hours. We computed the PSRF for each parameter resulting in 99% of the PSRFs in the two models being less than or equal to 1.05, respectively, both with a median of 1.00, indicating convergence.

5.4.2.1 Comparison Spatial Model

To examine the effectiveness of our proposed framework with respect to predicting missing counts within the components, we chose to compare our GDM approach against a commonly used alternative which does not take into account the compositional nature: GAM (Section 5.3.4).

Note that the missing values in the compositions effectively rule out applying log-ratio transformations to the data, which would align with other approaches reviewed in Section 5.2.2 and retain compositional information. We thus fit a series of single-tree GAMs to the counts.

These GAMs assume that the counts $v_{s,d}$ for tree species come from a quasi-binomial model:

$$v_{s,d} \sim \text{Quasi-Binomial}(n_s, p_{s,d}), \tag{5.4.2.1}$$

where $\eta_{s,d}$ is the inverse link function for the quasi-binomial, n_s is the total number of trials, i.e. the total count of trees in location s and $p_{s,d}$ is the probability of success, i.e. the proportion of tree species d in location s. For the quasi-binomial distribution, the logit link function is applied to $p_{s,d}$:

$$\text{logit}(p_{s,d}) = \beta_0 + s_1(\rho_{s,d}) + s_2(\delta_{s,d}), \tag{5.4.2.2}$$

where β_0 is the intercept, s_1 and s_2 are smooth functions of the coordinates fitted using splines with r basis functions and ρ and δ represent the x-coordinate and y-coordinate of the spatial location s. We used the quasi-binomial family as the Beta-Binomial family is not currently implemented in the *mgcv* package and the quasi-binomial also allows for overdispersion in the counts.

After constructing the missing tree species data in the same way as for the GDM model, the GAMs are fitted using the code in Listing 11. The predicted counts for each of the four GAMs are produced using the `predict` function from the *mgcv* package (Wood, 2003). It took approximately 12 minutes to run the four GAMs on an Apple MacBook Air M3 for each tree species for 400 basis functions for the 1,000 randomly selected spatial locations.

```
tree_gam <- gam(count / total ~ s(X_coord, Y_coord, k = r),
                family = "quasibinomial",
                weights = total,
                data = missing_tree_data)
```

Listing 11: `R GAM` code to fit spatial GAMs to each of the tree species where `count` will be replaced with larch, oak etc., for the count for each tree species. The GAM is produced using the "quasibiomial" distribution with `r` the number of the basis functions set to 400.

By comparing the GDM model and the GAMs, we can assess the difference between a model that accounts for the compositional nature of the data and another that does not, with respect to predicting missing values where information (counts) for the other tree species is available. The aim is that in addition to knowing the total count of each spatial location, which is the case for both models, the GDM is superior as it has the information of the counts of the tree species within that location, i.e. if the counts are (?, ?, 50, 10) the model can see that the first two missing values cannot exceed 40. Comparison with GAMs fitted using the *mgcv* package is advantageous because we can use the same basis function for both models.

To assess which model (GDM or GAM) performs better in terms of predicting the counts of tree species within each spatial location, we compared the counts both visually and using summary statistics. This involved plotting the predicted counts against the observed counts for each spatial location, and looking at how different the predicted counts for each model were in comparison to the $y = x$ line, which represents perfect agreement between predictions and observations. Any points that lie on the $y = x$ line indicate that the predicted count is the same as the observed count. Futhermore, any over-predicting or under-predicting in the model is given by points falling either above or below the $y = x$ line, respectively. We also inspected heatmaps for each model, which visually show how similar the predicted counts are to the original heatmaps from the observed data. Additionally, we produced some numerical comparisons to summarise which model was performing better on average: this included computing the MAE and RMSE, as done for the single tree species model comparison in Section 5.4.1. We also computed an R^2-type quantity, ξ, that quantifies GDM prediction

error variance relative to the baseline GAM model, using the Mean Square Error (MSE) of both models, defined as:

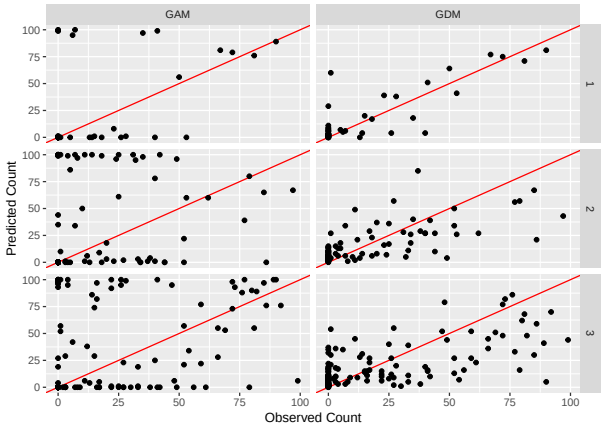
$$\xi = 1 - \frac{\text{MSE}_{GDM}}{\text{MSE}_{GAM}}. \tag{5.4.2.3}$$

This is a measure of how much better the GDM model reduces out-of-sample mean squared prediction errors compared to the GAM. Values of ξ close to 1 indicate that the GDM predictions reduce the MSE effectively compared to the GAM, whereas values near 0 indicate little difference in the models. Negative values of ξ indicate that the GDM model performs worse than the GAM.

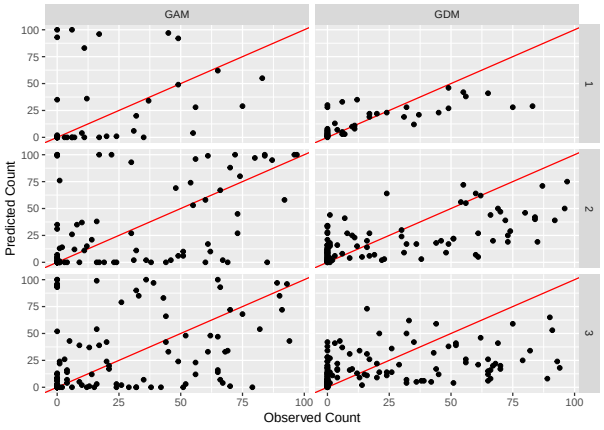
5.4.2.2 Multi-tree prediction experiment results

Figure 5.6 presents a side-by-side visual comparison of the predicted and observed counts for each tree species for the GAM and GDM models, with the grid on the right-hand side illustrating the number of components missing within each species.

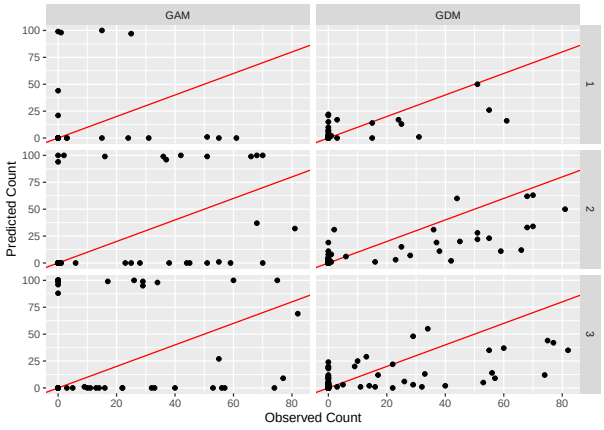
As an illustrative example, for the tree species larch (Figure 5.6 (a), the GDM model produces predictions that align reasonably well with the actual values, with the points within the GDM column following the red $y = x$ line. The GDM predicts the larch counts more accurately when it is the only missing tree species, as shown in the top grid. However, when the larch counts are missing alongside two other species, the model under-predicts the counts of larch, with points deviating further from the line. When larch is not the only tree species missing, the predictions tend to cluster around the origin, both above and below the $y = x$ line. In contrast to the GDM, the GAM consistently over-predicts larch counts across all levels of missing components (1, 2 and 3), as evident from the clusters of points concentrated in the top-left area of each plot in the figure. The largest difference between the two models is observed in Figure 5.6 (c) for the tree species sitka spruce. This tree species has the highest number of zero values compared to the other species, with 79.7%



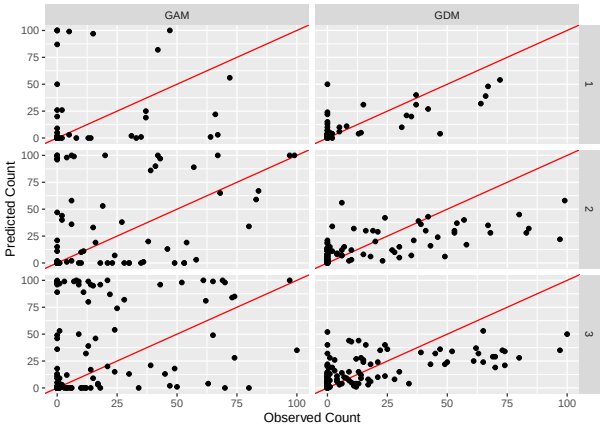
(a) Larch



(b) Oak



(c) Sitka Spruce



(d) Sycamore

Figure 5.6: Scatterplots of the predicted counts against the observed counts for each tree species for the GDM and GAM for the 1,000 randomly fitted grid cells. The y=x line is given in red which indicates perfect agreement between the predicted and observed counts.

of its values being zero. It can be seen that the GDM is doing fairly well at predicting the counts of sitka spruce with the points lying close to the $y = x$ red line. A vastly different picture is displayed for the GAM which has very few points lying on or close to the identity line, indicating that the predictions for sitka spruce do not align with the actual values in the data, both over and under predicting the counts. Moreover, it seems that that GAM has predicted counts close to the total value 100 over all cases where sitka spruce is missing even when the actual count of sitka spruce is zero.

Figure 5.7 presents heatmaps of predicted counts for each model alongside observed counts for 1,000 randomly selected spatial locations, providing clear evaluation of how effectively each model captures the counts of each tree species. Inspecting the GDM column in the figure reveals that overall the GDM has predicted counts that are reasonable when compared to the observed data. Overall for all tree species, the GDM predictions closely align with the observed data's colour grid cells, effectively capturing each species clusters of counts across the AOI. However, the GDM does seem to under-predict some counts, shown for sitka spruce where the counts do not exceed 50. On the other hand, the GAMs consistently predicts high counts (close to 100) for each tree species, as illustrated in the heatmaps where numerous yellow grids indicate counts around 85–100 across the spatial locations, a pattern that reflects what we identified in the GAM scatterplots. This contrasts significantly with the heatmap of the observed counts, which shows very few yellow grid cells for any tree species, highlighting the GAMs significant over-prediction of counts.

The summary statistics comparison in Table 5.4 further supports this, quantifying the performance between the two models, presenting the MAE, RMSE and ξ (Equation (5.4.2.3)) for each model in three ways: across all levels of missing components for each tree species, overall for each tree species and the model as a whole. The table is coloured with green cells

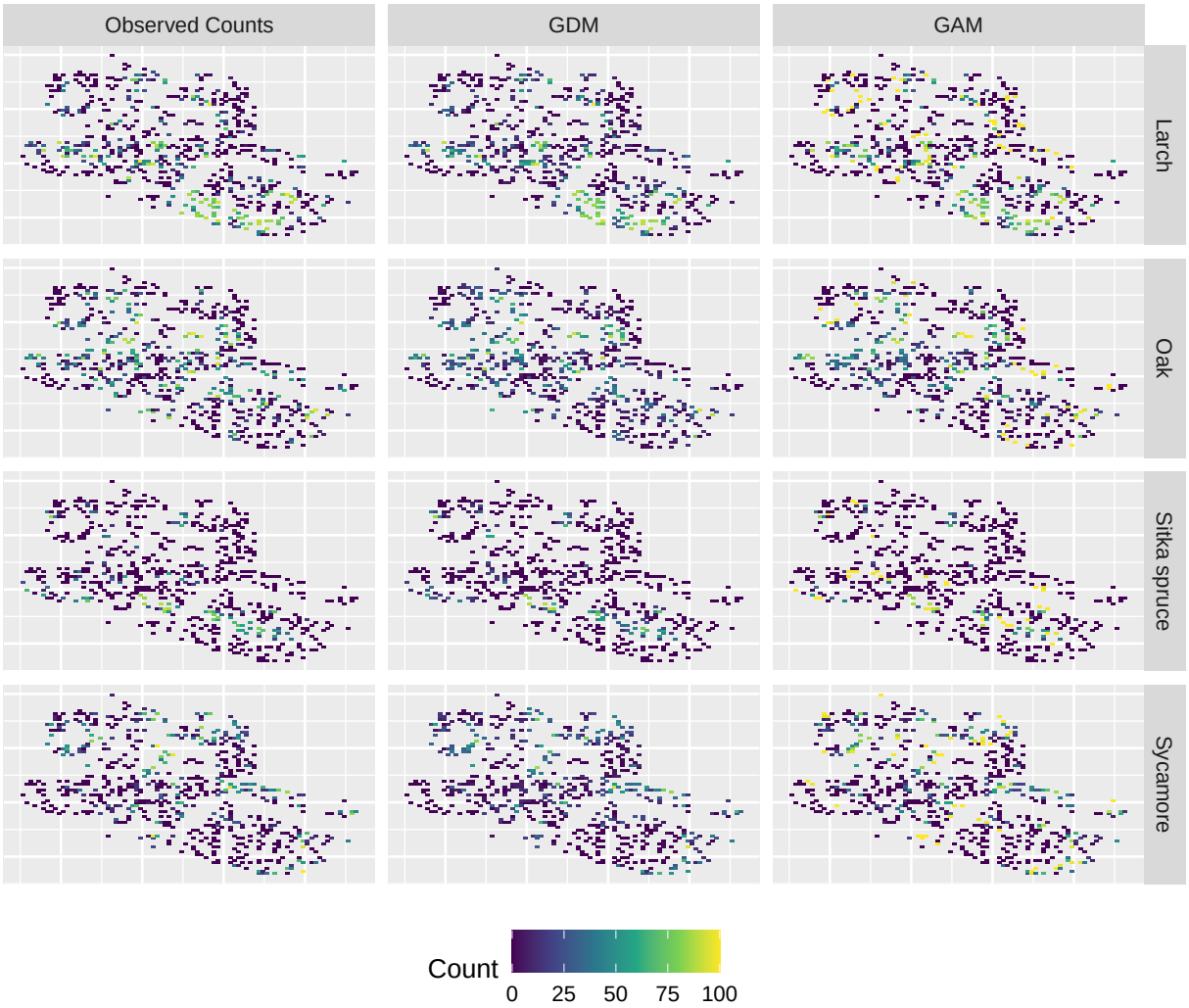


Figure 5.7: Heatmaps of the observed and predicted counts for each tree species for the GDM and GAM for the 1,000 randomly fitted grid cells which had one or more missing tree species count.

that represent the lowest MAE or RMSE value. A 2% tolerance has been applied, meaning that more than one value may be coloured green in the event of a near tie. The initial comparison will be conducted between the GDM and GAM-400 where both models have been fitted with 400 basis functions.

A similar picture is presented across each of these measures showing that the GDM has superior performance in predicting the counts of each compositional tree species. For example, for the MAE, the largest difference between the two models is presented when larch plus another tree species is missing, where the MAE for the GDM is 17.6 lower than the MAE for the GAM. This was shown in Figure 5.6 (a) where the GAM over-predicted the small values to be very high values, whereas the GDM predictions follow the trajectory of the $y = x$ line. Overall, based on the MAE results, the GDM predictions outperform the GAM for the tree species larch. When not considering the number of missing components, there is a 66% percentage difference between the MAE for the two models for larch in favour of the GDM. The tree species oak (not considering the number of missing tree species) appears to have the smallest difference observed between the MAE values for the two models at only 2.4 suggesting that for this tree species the models have a similar performance in predicting the counts. Overall, the GDM model performs better with a MAE value of 11.4 compared to the GAM MAE of 19.5, showing a 53% lower MAE for the GDM.

The RMSE also shows the largest difference between the models is for larch and the difference is minimal for oak. Sitka spruce, which contains the highest number of zeros, shows a 20.2 lower overall RMSE for the GDM compared to the GAM. This suggests that the GAM struggles to capture zeros in the data, as illustrated in Figure 5.7, where it tends to over-predict counts. Again, the GDM consistently outperforms the GAM, with an RMSE that is 17.4 lower.

Lastly, we computed ξ, an out-of-sample R^2 (Equation (5.4.2.3)), to quantify any improvement in out-of-sample MSE from the GDM model, relative to the GAM. Here we see from the final column in Table 5.4 that, as all values are greater than zero, the predictions from the GDM are closer to the actual values in comparison to the GAM. Any ξ values close to 1 indicate stronger evidence that the GDM is performing better than the GAM, which is seen with most values of ξ greater than 0.7. The lowest ξ values are observed for the tree species oak, particularly when there is more than one tree species missing alongside oak, which supports the findings seen within the MAE and RMSE. When evaluating each tree species without considering the number of missing components, the results strongly favour the GDM, as the values exceed 0.75 for three of the tree species with the only exception seen for oak that has an ξ of 0.5. Interestingly, sitka spruce has the highest overall ξ value of 0.84, further indicating that when there is a large number of zeros in the component the GDM outperforms the GAM in prediction performance.

Since the initial GAM with 400 basis functions did not adequately fit the spatial compositional data, we assessed the model's performance by fitting GAMs with 200 and 600 basis functions to the missing tree species data. The performance of each model was quantified using MAE and RMSE, as presented in Table 5.4. As the number of basis functions in the GAM increases, the model tends to overfit the tree species counts, with the GAM with 600 basis functions performing less adequately than the model with 400 basis functions. However, interestingly, the GAM with 600 basis functions performs better than the other two models for sitka spruce, as indicated by the lowest MAE and RMSE values across all GAMs, although this is on par with the GDM. On the other hand, the GAM with 200 basis functions shows better performance than the GDM in some cases, for example, when larch is missing along with two other tree species, and oak is not the only missing tree species. For both larch and oak, the overall tree species MAE favours the GAM with 200 basis

Table 5.4: Summary statistics quantifying the prediction accuracy of the missing compositions for the GDM and GAM. Each summary statistic is given for the GAM for 200, 400 and 600 basis functions for comparison. $\boldsymbol{\xi}$ (Equation 5.4.2.3) is computed for the GDM and GAM with 400 basis functions. The lowest Mean Absolute Error (MAE) or Root Mean Square Error (RMSE) is highlighted in green, with a 2% tolerance applied in cases of near ties.

Tree Species		**MAE**				**RMSE**				$\boldsymbol{\xi}$
	missing	GDM	GAM			GDM	GAM			
			200	400	600		200	400	600	
larch	1	7.8	9.2	18.0	18.9	13.1	17.4	35.5	35.2	0.864
	2	11.4	11.6	29.0	20.8	16.6	18.2	45.4	35.1	0.867
	3	14.3	12.8	23.3	22.1	20.6	20.1	38.9	36.8	0.721
oak	1	9.9	13.8	20.2	29.9	15.8	19.9	34.6	45.8	0.791
	2	16.5	15.1	18.0	27.2	22.8	22.5	31.1	43.4	0.461
	3	17.0	16.0	17.7	27.7	24.5	24.8	30.8	43.8	0.369
sitka spruce	1	5.1	16.2	12.5	7.2	10.5	35.1	29.1	17.6	0.871
	2	6.6	15.0	13.0	9.6	13.6	31.4	28.5	21.6	0.770
	3	5.5	10.1	14.2	5.9	12.3	25.8	33.5	17.0	0.865
sycamore	1	8.5	11.4	21.6	29.6	13.8	19.7	37.9	47.9	0.867
	2	12.5	12.8	21.3	20.6	18.8	20.6	36.1	35.5	0.728
	3	12.4	12.3	23.5	31.6	18.3	20.4	39.7	49.6	0.788
larch		12.3	11.8	24.3	21.2	18.2	19.0	40.7	36.0	0.799
oak		15.8	15.3	18.2	27.8	22.8	23.3	31.5	43.9	0.478
sitka spruce		5.8	12.6	13.6	7.2	12.6	30.9	32.8	19.4	0.842
sycamore		11.8	12.3	22.4	27.4	12.5	29.4	31.3	18.6	0.784
overall		11.4	13.1	19.5	20.8	18.2	23.4	35.6	37.4	0.738

functions, suggesting that a lower number of basis functions might be a better fit for the spatial data. However, when considering the model on the whole, the GDM performs the best in predicting the counts of the missing compositional tree species in terms of the MAE and RMSE.

5.5 Summary & Discussion

In this chapter, we introduced a framework to deal with spatial compositional data. We reviewed previous methods used to address spatial compositional data, which mainly involve applying a log-ratio transformation to the counts, percentages or proportions in the data. However, this transformation is unsuitable when the data contain zero or missing values, a common issue found within compositional data. We also examined other methods, including kriging and Conditional Autoregressive (CAR) models which have frequently been used to model spatial data. However, these approaches were applied to spatial compositional data but did not address cases where the data contained zeros or missing values. This limitation motivates the development of the framework presented in this chapter.

The tree species data collected and compiled by Fera Science motivated our application, consisting of a spatial grid of compositional count data. These data include 2,153 spatial locations which contain counts from 10 different tree species. As some tree species are sparse over the spatial grid, a large proportion of zeros is observed, hence making the log-ratio transformation unsuitable for this compositional data. If a small number were added to all the zeros, to allow for a log-ratio transformation to be used, the data would be dominated by the log of that value, thus still prohibiting the use of standard methods on the

transformed values that assume a continuous distribution. The tree species data was vital in the development and testing of the proposed Generalised-Dirichlet-Multinomial (GDM) framework. The framework is implemented using the *NIMBLE* package in `R` which allows for flexible and efficient fitting of MCMC models.

The proposed framework is advantageous over other current methods due to its ability to directly account for the compositional structure, to account for spatial dependence through latent effects, to allow for zeros in the data and to allow for missing values in the spatial compositions. This is conducted through the combination of the flexible Generalised-Dirichlet distribution which allows each composition to have its own variance and the Multinomial distribution which can model the compositional counts. Our approach uses penalised regression splines to model the spatial locations. This incorporates a penalty term that controls the smoothness of the fitted curve. This penalty helps to avoid overfitting, ensuring that the model captures the underlying trend without being overly sensitive to noise in the data. However, other spatial models could be applied within the framework such as a CAR model which is already specified within NIMBLE. We tested the framework using compositional proportions transformed into counts, but it is also suitable for continuous compositions summing to something other than one, and directly suitable for compositional counts, showing its suitability for a wide range of compositional data types and its versatility in various contexts.

Initially, to test the performance of the GDM framework with spatial compositional data and inform the variance parameter design, we investigated a model for a single tree species, larch, in a out-of-sample prediction experiment. We adapted the GDM framework for a single tree species and assessed the performance of two different variance terms, one which is fixed in space for each tree and the other which considers a flexible mean-variance relationship between $\boldsymbol{\mu}$ and $\boldsymbol{\phi}$. We trained both the GDM models using 50% of the data locations, and tested the model's prediction of the withheld spatial locations. We simulated replicate/predicted tree counts for the spatial grids within the test data and compared them to

the observed counts both visually, using density and quantile plots, and numerically, using the MAE, RMSE and assessing the Bayesian coverage of the uncertainty intervals. Both the single tree species GDM models were computationally efficient and ran in a timely manner. We concluded that both models have a similar fit when considering both the in-sample and out-of-sample results. The GDM-polynomial model, containing a flexible mean-variance relationship, is superior in capturing both the sample mean and standard deviation of the original data. Likewise, for the GDM-polynomial the predicted quantiles of the replicates are more similar to the values from the original data. Numerically, the prediction performance of both variance terms were similar but the GDM-polynomial model outperformed the GDM-fixed when considering the MAE of the quantiles. Examination of the Bayesian coverage for the uncertainty intervals showed that the coverage of both models exceed the uncertainty percentage of each interval. However, when investigating the mean widths of the uncertainty intervals, it was found that for the GDM-polynomial the intervals were considerably wider, suggesting that this model could be overly uncertain. Therefore, both variance terms could be sensible to use in practice, however, applying a flexible mean-variance relationship demonstrates a tighter fit to the distribution of the data, making it our preference to apply.

We then extended the GDM to fit the model to multiple tree species and assessed prediction performance in the context of counts for one or more tree species being missing from available compositional data. To do this, we randomly removed some count values, resulting in 20% of the locations having one, 20% having two and 20% having three tree species counts missing, respectively. To assess how well the GDM model was able to predict these missing counts, we compared it to an alternative spatial model, the Generalised Additive Model (GAM). For comparison we had to compare our GDM to four individual GAMs for each tree species. Although these ran much quicker than the GDM, the performance of the GDM is far superior. We visually inspected the predictions of both models against the observed counts in scatterplots which showed that the predicted counts from the GDM are much closer to

those observed in the data. Moreover, the heatmaps produced by the GDM predictions are a closer fit to the original counts than the GAM. This is further backed up from our numerical comparison summary which favour the GDM predictions. Therefore, we can conclude that the GDM is superior in predicting the compositional tree counts of the spatial data.

Here, we compared the GDM with the GAM where both were fitted with 400 basis functions for a fair comparison. However, after further investigation, we concluded that a GAM with a smaller number of basis functions might better fit our spatial tree data. We additionally fitted the GAMs for 200 and 600 basis functions to inspect the impact of the number of basis functions. We found that as the number of basis functions increased, the GAM produced predictions that were closer to the total count (100) even for low observed counts. Consequently, we discovered that a GAM with 200 basis functions produced more reasonable predictions. We examined the numerical summary of these results, which showed that for the tree species larch and oak, the GAM with 200 basis functions had a lower MAE than the GDM model. However, for sitka spruce and sycamore, which contain over 50% zeros in the component, the GDM demonstrated superior predictive performance. Further still, there is potential for the GDM to increase its lead through optimising the number of basis functions.

Avenues for future research to develop this framework further include exploring testing spatial areal models, such as GMRF smooths from *mgcv* package or a spatial CAR within the NIMBLE framework. Another approach could be to extend this framework by incorporating covariates into the model, allowing it to improve prediction accuracy from the added knowledge of other factors.

In conclusion, we have created a flexible and widely applicable framework which could be applied to other types of spatial compositional data such as the proportions or counts of different farming crops in different fields or disease prevalence across different spatial locations. The superior performance of the GDM compared to the GAM can be explained by the GDM's ability to predict missing values with compositional constraints. Specifically, within the GDM, if the model observes high counts of one tree species, it knows to predict low counts for the remaining species. In contrast, the GAM lacks this knowledge, leading to over-predictions of very high counts. Therefore, the GDM is an effective tool for modelling spatial compositional data.

Chapter 6

Conclusion

In this thesis, we studied statistical methods for modelling compositional data, which is so prevalent in our understanding of the world that we often overlook its uniqueness. We sought to build an understanding of the core concepts of compositional data analysis arising from decades of research; we then aimed to offer a fresh perspective that sets aside strict rules in favour of practical model design that reflects our understanding of the features and structures of the data.

In Chapter 2, we outlined the early definition of compositional data and evaluated the main approach used, log-ratio transformations, which map the relative information of compositional data to an unconstrained real space for analysis using more common statistical methods. These have been applied across many different statistical fields throughout the literature, as reviewed in Chapter 2, as well as in Chapter 3 in the context of compositional data with a large proportion of structural zeros, in Chapter 4 in the context of compositional data evolving over time and in Chapter 5 in the context of compositional data arranged over a spatial domain.

Since log-ratio approaches are so well studied, and their basic function is to bring compositional data into a form that most will know how to analyse using standard methods/tools they are familiar with, they may well be the most straightforward option for many, if not most, real world data problems. However, we have identified data features that can at least pose a real obstacle to using log-ratio transformations and at worst make them essentially unsuitable. Notably, when compositional data contain zeros, the log-ratio is undefined, and in cases where these are true zeros, we argued that we can't simply replace them with a small value to allow for the transformation to proceed without discarding important information or biasing results. Another instance where we argued that log-ratio transformations are unsuitable is when there are missing values in the compositions. Addressing missing values explicitly is an area that is underexplored in the literature, with these compositions typically either being removed or replaced with a small value. Furthermore, compositional data can often involve a count structure, where the components sum to a total count. In this

case, we argued that applying a log-ratio transformation to the counts results in discrete variables in the real space that may not be suitable for modelling using standard methods. This may also potentially discard information on how the total count may impact the variance and constrains the possible values the counts could take. We argued that these issues are especially pertinent when the total count is small.

We opted to focus our efforts on such situations, where core data characteristics deter the use of log-ratio transformations, as a fruitful ground for expanding methodology for modelling compositional data. Here, we accounted for these prohibitive features through automated splitting of the data in Chapter 3, and through use of the Generalised-Dirichlet-Multinomial (GDM) model for real count data in Chapter 4 and for artificial counts derived from proportion data in Chapter 5. A key advantage of using the GDM family of distributions is the flexibility of the Generalised-Dirichlet (GD) distribution's covariance structure between the components, which potentially allows a good fit in a range of applications. Mixing the GD with the Multinomial allows the GDM to model compositional count data, where the components do not need to be independent and the total count can vary. The GDM also has the ability to model sparse data where both zero and one outcomes occur, and using the Beta-Binomial conditional representation of the GDM we can easily handle missing values in the compositions.

To further strengthen our contribution, we sought out and tackled situations where compositional data exhibit features that prohibit the use of log-ratio transformations and interact with other advanced statistical challenges. Not properly accounting for these data challenges in our compositional modelling framework could have distorted the interpretation of the results and produced misleading conclusions. In Chapter 3, we developed an approach motivated by an application involving a large number of structural zeros and a multilevel hierarchical structure. To address the structural zeros in the compositions, we automated splitting the data based on the presence and absence of the components through latent clustering within our hierarchical model, reducing the need for expert knowledge to conduct

this split. The proposed framework was evaluated using a forensic elemental glass data, by examining the classification of new glass items into one of the five glass use types. Here, we found that our proposed approach performed well at correctly classifying out-of-sample glass items. Additionally, when we examined the uncertainty of our classification predictions, our integrated clustering approach performed well, with comparable low uncertainty for each glass type.

In Chapter 4, we explored an approach to compositional time series data, consisting of count data with real zeros. A further complexity we had to address was the non-smooth nature of time series, i.e. where data exhibited abrupt changes or irregular fluctuations, rather than following a continuous and predictable trend over time. To address these challenges, we developed a framework to account for both the compositional count and non-smooth temporal structure. Our proposed framework combines a GDM distribution with a latent hidden Markov model (HMM) structure to capture the non-smooth temporal dependence. Here, we developed and evaluated our proposed framework using compositional counts of COVID-19 variants globally, clustering the countries to produce different HMM parameters for each cluster and allowing specialised modelling for each variant. We found that our proposed GDM-HMM framework outperformed other commonly used time series models in place of the HMM in a posterior predictive experiment.

Finally, in Chapter 5, we investigated compositional data arranged over a spatial structure with both zero and missing values in the compositions. When a component value is missing, some or all of the log-ratio transformations will not produce sensible results. To address this challenge, we developed a general multivariate framework that accommodates the compositional structure, incorporates spatial dependence and can handle zeros and missing values in compositions. Here, we again created a framework that implemented the GDM, in this case with two-dimensional penalised regression splines that capture the spatial structure. Within our framework, we proposed a novel polynomial variance parameter $\boldsymbol{\phi}$ for the GDM instead of a strictly fixed variance term, which we found to be superior in terms of reproducing the

quantiles of the data. If we only had log-ratio methods at our disposal, we may have needed to impute the zeros. As an alternative, we could keep the data as counts and fit univariate models to each tree species, however we found that our proposed framework substantially outperformed this benchmark approach in predicting the counts of missing tree species. In particular, our framework performed best when a tree species had a high proportion of zero values. We showed the framework was effective when converting compositional proportions to counts to be able to apply the GDM distribution. This further extends the potential use of the framework for both spatial compositional data including proportions or a count structure.

Throughout all this work, we leaned exclusively on the Bayesian hierarchical approach to modelling due to its ability to handle complex data structures flexibly. This was particularly advantageous in addressing the issue of missing values in the compositions, as Bayesian models naturally accommodate missing data by treating them as unknown quantities that we can learn about through posterior inference, along with model parameters. By implementing Bayesian analysis, we were able to assess the success of each proposed framework through posterior predictive model checking experiments, comparing the performance with alternative, simpler models. Without using Bayesian hierarchical frameworks, these aspects would have been nearly impossible or significantly less reliable. However, our proposed frameworks also presented challenges. Perhaps the greatest of these was the computational cost associated with our complex models. MCMC sampling, while a powerful and general tool, was computationally intensive and required careful tuning and validation to ensure convergence. Although the comparison models often required less computational time, the superior model fit achieved through our novel frameworks justified the additional computational cost.

Throughout the thesis we implemented all the proposed frameworks using the *NIMBLE* package (Valpine et al., 2017). Without the recent development of modelling tools, such as NIMBLE, creating frameworks that address the challenges presented throughout the thesis would have been difficult. NIMBLE is a package that allows for flexible implementation of

Bayesian models using MCMC. Writing models in NIMBLE is advantageous due to the flexibility it adds in terms of the model specification, samplers for the model parameters and the ability to add functions and distributions alongside the wide range of already defined functions and distributions. Here, we extended NIMBLE by creating the functions to compute the Beta-Binomial distribution (Chapters 4 and 5) and the forward algorithm (Chapter 4) for the latent state sequence.

We highlighted several potential avenues for future research within each of the individual Chapters 3, 4 and 5. As outlined, we only tested each proposed framework on a single data application. Therefore, future work could include applying each framework to different applications that exhibit similar features and data challenges compared with those presented in this thesis. Each of the proposed approaches could be extended to incorporate covariate information. Additionally, the framework for spatial compositional data could be adapted to use alternative spatial models, such as a spatial conditional autoregressive (CAR) model. Overall, these frameworks can be tailored to meet specific application and modelling requirements. As we noted that computational cost of our proposed methods could be improved upon, we could consider implementing faster inference using alternative samplers or methods. Specifically, we could explore applying the GDM to the application presented in Chapter 3, using artificial counts derived from the proportions of the compositional elements. Another aspect worth examining is when compositional data take the form of a spatio-temporal structure, integrating insights from Chapters 4 and 5.

In the broader context of compositional data, we hope that this thesis has demonstrated the value of expanding the traditional definition of compositional data, allowing more innovative approaches to be applied to the analysis of this unique type of data. We believe that the proposed methods offer significant benefits across a wide range of compositional data applications, extending beyond the specific cases examined in this thesis, making them valuable tools for real-world analysis of compositional data.

Future work could explore testing more intuitive or practical approaches to better capture the data while preserving both absolute and relative information. It would also be interesting to investigate more situations where the total is informative, highlighting the risk of discarding this information by focusing solely on log-ratios and relative proportions. Additionally, exploring the nature of compositional data where either the constraint or the component is defined first could provide valuable insights into the construction of compositional data analysis.

Appendices

A Initials & Acronyms

- `GDM` - Generalised-Dirichlet-Multinomial
- `HMM` - Hidden Markov Model
- `GD` - Generalised-Dirichlet
- `GAM` - Generalised Additive Model
- `RW` - Random Walk Model
- `DLM` - Dynamic Linear Model
- `MCMC` - Markov Chain Monte Carlo
- `PSRF` - Potential Scale Reduction Factor
- `ALR` - Additive Log-ratio
- `CLR` - Centered Log-ratio
- `ILR` - Isometric Log-ratio
- `WHO` - World Health Organisation
- `VOC` - Variant of Concern
- `VOI` - Variant of Interest
- `ECE` - Expected Calibration Error
- `MAE` - Mean Absolute Error
- `RMSE` - Root Mean Square Error

B Bayesian Inference

B.1 Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) methods can be used to simulate and draw samples from distributions. Monte Carlo methods are used to approximate integrals and closed-form expressions that are otherwise extremely difficult or impossible to evaluate. Therefore, the use of MCMC methods allows samples to be collected directly from the posterior. After reaching a state of equilibrium, this is thought of as sampling from the desired target distribution. This process occurs after the chain has been run for a sufficient number of iterations, allowing it to explore the parameter space and discard any initial bias from the starting point. The two most commonly used MCMC algorithms are the Metropolis-Hastings algorithm (Metropolis et al., 1953) and the Gibbs sampler (Casella et al., 1992). Depending on the model, only one of these methods may need to be implemented, but a mix of both can be employed also.

Since the sampling algorithm requires an initial number of iterations before converging to the target, a "burn-in" period can be specified, such that a chosen number of iterations can be discarded, with only the draws made after this period used in the analysis. This allows the chain to reach a stationary distribution before it keeps the samples. The initial part of the chain is discarded as it is thought that it will have been influenced by the starting point and may not be representative of the true distribution.

Autocorrelation in MCMC samples refers to the degree of dependence or correlation between consecutive samples in the chain. It measures how much the one sample is related to the previous sample in the chain. High autocorrelation indicates that the samples are highly correlated, resulting in slow mixing, meaning it takes more iterations for the chain to explore the different regions of the parameter space and converge to the target distribution. Reducing autocorrelation is desirable because it allows for more independent and representative samples.

We can "thin" our chains by storing every m-th draw of the sampler, with the rest of the draws discarded. Thinning an MCMC can be advantageous by reducing the computational practicality of the chain as it does not store as many samples. As only a subset of the samples are saved, it reduces the storage requirements and process power to run the full chain. However, Link et al. (2012) argue that keeping all samples leads to more accurate inference, where feasible.

We can run multiple chains to check the convergence of the model, which is the process of running the MCMC multiple times from different starting points. This can allow us to potentially get more samples from our model. As the MCMC chains are independent of one another, they can be run in parallel across multiple CPU cores to reduce computation times.

B.2 Checking convergence and quality of samples

To assess the convergence of the model, the following MCMC diagnostics are computed for each instance of model fit.

MCMC convergence is usually assessed by running multiple MCMC chains from different, sometimes randomly generated initial values. We can compute the Gelman Diagnostic (Gelman et al., 1992) from the resulting samples, which compares the variance between the chains to the variance within the chains. If the two variances are similar then this typically results in a Gelman value less than 1.05. A Gelman value of around 1.05 or less indicates that the chains have converged to the same distribution, which is usually the posterior but in some cases, all chains may have converged to another distribution such as the prior or the MCMC proposal distribution. In `R`, we can compute this using the *coda* package (Plummer et al., 2006), the `gelman.diag` function computes the Gelman diagnostic for each of the parameters.

In addition, traceplots can be produced for the sampled parameters from the model. Traceplots are a visualisation tool used in Bayesian inference to assess the convergence and mixing properties of the chains. A traceplot displays the values of the parameter of interest against the number of iterations. The evolution of the parameter values over time can be visually examined. If the traceplot shows a random pattern with no apparent trends or patterns where the points appear scattered throughout the plot. This suggests that the chain has explored the different regions of the parameter space and hence indicates good mixing has been reached. The points in the traceplot should also exhibit stable and consistent variability across the iterations. The spread of points should not change dramatically over time. Having a stable variability indicates that the chain is exploring the target distribution consistently and is not getting stuck in particular areas. As multiple chains are employed in the MCMC process, it is expected that the chains exhibit overlap and demonstrate similar mixing patterns. This behaviour suggests that all chains are effectively exploring the parameter space, including both local and global maxima. If the chains converge to similar regions of the parameter space and display consistent sampling from these regions, it indicates that the algorithm is adequately exploring the distribution and is not getting stuck in local maxima. The presence of such overlap and mixing suggests that the chains are collectively converging to the global maximum (or a good approximation thereof) of the target distribution.

C Further Classification Checking for Chapter 3

C.1 Classification Performance

We can consider a number of classification performance metrics to further assess the classification of the glass items into one of the five glass use types, as described in Chapter 3. Here we will compare each approach using seven performance metrics.

Goodman and Kruskal's τ (Agresti et al., 1990) is a measure of the reduction in the expected conditional variability in comparison to the marginal variability. Theil's U (Agresti et al., 1990) provides another measure of variation. These metrics measure association and examine the proportional reduction in prediction error. They examine how much the predicted glass use types differ from the actual type. Higher values of τ and U indicate a stronger association between predicted and actual classes. Cohen's κ (Agresti et al., 1990) is a measure of agreement that takes into account any agreement that can occur by chance. This measures the agreement between the classified and actual glass type. A value of $\kappa = 0$ is equivalent to that of agreement by chance, whereas perfect agreement would have a value of $\kappa = 1$. Matthews Correlation Coefficient (MCC) (Baldi et al., 2000) is a correlation coefficient between the observed and predicted classifications which takes all possible prediction outcomes into account. Lastly, the F1-score combines the precision and recall of a classifier using a confusion matrix associated with that classifier. Measures of both association and agreement are explored as it would be possible to have an association without agreement. For both the MCC and F1-score, higher values indicate better classification performance. In addition, the accuracy and the percentage of misclassification are evaluated.

Tables C1 and C2 report the classification performance metrics for each of the classification results from the different approaches. Overall, the results within Tables C1 and C2 align with those outlined in Section 3.4.7.

Table C1: Classification performance metrics for the classification results presented in Chapter 3, Section 3.4.7 of the classification of glass items into one of the the five glass use types.

Approach	**Performance Measure**				
	Accuracy	% miss-classified	τ	U	κ
No spilt: untransformed	0.331	66.9%	0.05	0.11	0.09
No spilt: square root	0.656	34.4%	0.39	0.53	0.45
Manual: configurations	0.756	24.4%	0.51	0.54	0.67
Pre-clustering: hierarchical	0.763	23.8%	0.53	0.56	0.68
Pre-clustering: k-means	0.753	24.7%	0.51	0.54	0.67
Integrated clustering	0.763	23.8%	0.54	0.54	0.68

Table C2: Classification performance measures for the classification results, presented in Chapter 3, Section 3.4.7 of the classification of glass items into one of the the five glass use types, split by each compositional elemental. The highest F1-score or Matthew's Correlation Coefficient (MCC) are highlighted in green, with a 2% tolerance has been applied in cases of near ties.

Approach	**Glass Use Type**	**Performance Measure**	
		F1-score	MCC
No split: untransformed	bulb	0.42	0.50
	car window	0.43	0.19
	headlamp	0	0
	container	0.12	0.01
	building window	0.36	0.08
No split: square root	bulb	0.78	0.77
	car window	0.50	0.29
	headlamp	0.71	0.71
	container	0.85	0.79
	building window	0.51	0.30
Manual: configurations	bulb	0.73	0.70
	car window	0.56	0.41
	headlamp	0.48	0.46
	container	0.85	0.80
	building window	0.53	0.37
Pre-clustering: hierarchical	bulb	0.98	0.98
	car window	0.68	0.55
	headlamp	0.79	0.79
	container	0.89	0.86
	building window	0.68	0.52
Pre-clustering: *k*-means	bulb	0.96	0.96
	car window	0.68	0.55
	headlamp	0.67	0.65
	container	0.88	0.84
	building window	0.68	0.52
Integrated clustering	bulb	0.93	0.92
	car window	0.75	0.64
	headlamp	0.19	0.20
	container	0.83	0.77
	building window	0.74	0.61

Bibliography

Agresti, Alan, Cyrus R Mehta and Nitin R Patel (1990). 'Exact inference for contingency tables with ordered categories'. In: *Journal of the American Statistical Association* 85.410, pp. 453–458.

Ahmed, Mohiuddin, Raihan Seraj and Syed Mohammed Shamsul Islam (2020). 'The k-means algorithm: A comprehensive survey and performance evaluation'. In: *Electronics* 9.8, p. 1295.

Aitchison, John (1982). 'The statistical analysis of compositional data'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 139–160.

– (1986). *The statistical analysis of compositional data*. Chapman and Hall.

– (1992). 'On criteria for measures of compositional difference'. In: *Mathematical Geology* 24, pp. 365–379.

Aitchison, John and Juan José Egozcue (2005). 'Compositional data analysis: where are we and where should we be heading?' In: *Mathematical Geology* 37.7, pp. 829–850.

Aitchison, John and Jim W. Kay (2003). 'Possible Solutions of Some Essential Zero Problems in Compositional Data Analysis'. In: *Compositional Data Analysis Workshop (CoDaWork)*. Girona, Spain.

Aitken, Colin GG and David Lucy (2004). 'Evaluation of trace evidence in the form of multivariate data'. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 53.1, pp. 109–122.

Al-Dhurafi, Nasr Ahmed, Nurulkamal Masseran and Zamira Hasanah Zamzuri (2018). 'Compositional time series analysis for air pollution index data'. In: *Stochastic Environmental Research and Risk Assessment* 32, pp. 2903–2911.

Ankam, Divya (2019). 'Distributions based Regression Techniques for Compositional Data'. PhD thesis. Concordia University.

Atzrodt, Cassandra L, Insha Maknojia, Robert DP McCarthy, Tiara M Oldfield, Jonathan Po, Kenny TL Ta, Hannah E Stepp and Thomas P Clements (2020). 'A Guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2'. In: *The FEBS Journal* 287.17, pp. 3633–3650.

Baldi, Pierre, Søren Brunak, Yves Chauvin, Claus AF Andersen and Henrik Nielsen (2000). 'Assessing the accuracy of prediction algorithms for classification: an overview'. In: *Bioinformatics* 16.5, pp. 412–424.

BBC News (2020). *New coronavirus variant: What do we know?* `https://www.bbc.co.uk/news/health-55388846`. Accessed: March 26, 2024.

Bennett, Alexander R, Jon Lundstrøm, Sayantani Chatterjee, Morten Thaysen-Andersen and Daniel Bojar (2025). 'Compositional data analysis enables statistical rigor in comparative glycomics'. In: *Nature Communications* 16.1, p. 795.

Bentahar, Jamal (2015). 'Modeling and Forecasting Time Series of Compositional Data: A Generalized Dirichlet Power Steady Model'. In: *International Conference on Scalable Uncertainty Management*. Springer, pp. 170–185.

Brier, Glenn W (1950). 'Verification of forecasts expressed in terms of probability'. In: *Monthly Weather Review* 78.1, pp. 1–3.

Brunsdon, Teresa M and TMF Smith (1998). 'The time series analysis of compositional data'. In: *Journal of Official Statistics* 14.3, p. 237.

Butler, Adam and Chris Glasbey (2008). 'A latent Gaussian model for compositional data with zeros'. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57.5, pp. 505–520.

Carreras, Miquel and Germà Coenders (2020). 'Principal component analysis of financial statements: a compositional approach'. In: *Revista de Métodos Cuantitativos para la Economía y la Empresa* 29, pp. 18–37.

Casella, George and Edward I George (1992). 'Explaining the Gibbs sampler'. In: *The American Statistician* 46.3, pp. 167–174.

Chayes, Felix (1960). 'On correlation between variables of constant sum'. In: *Journal of Geophysical Research* 65.12, pp. 4185–4193.

Clarotto, Lucia, Denis Allard and Alessandra Menafoglio (2022). 'A new class of α-transformations for the spatial analysis of compositional data'. In: *Spatial Statistics* 47, p. 100570.

Comas Cufí, Marc, Josep Antoni Martín-Fernández and Glòria Mateu-Figueras (2016). 'Log-ratio methods in mixture models for compositional data sets'. In: *SORT - Statistics and Operations Research Transactions* 40.2, pp. 349–374.

Connor, Robert J and James E Mosimann (1969). 'Concepts of independence for proportions with a generalization of the Dirichlet distribution'. In: *Journal of the American Statistical Association* 64.325, pp. 194–206.

Cressie, Noel (2015). *Statistics for spatial data*. John Wiley & Sons.

Day, William HE and Herbert Edelsbrunner (1984). 'Efficient algorithms for agglomerative hierarchical clustering methods'. In: *Journal of Classification* 1.1, pp. 7–24.

DeGroot, Morris H (2005). *Optimal Statistical Decisions*. John Wiley & Sons.

Dejean, Sébastien, Pascal GP Martin, Alain Baccini and Philippe Besse (2007). 'Clustering time-series gene expression data using smoothing spline derivatives'. In: *EURASIP Journal on Bioinformatics and Systems Biology* 2007, pp. 1–10.

Egozcue, Juan José, Caterina Gozzi, Antonella Buccianti and Vera Pawlowsky-Glahn (2024). 'Exploring geochemical data using compositional techniques: A practical guide'. In: *Journal of Geochemical Exploration* 258, p. 107385.

Egozcue, Juan José, Vera Pawlowsky-Glahn, Glòria Mateu-Figueras and Carles Barceló-Vidal (2003). 'Isometric logratio transformations for compositional data analysis'. In: *Mathematical Geology* 35.3, pp. 279–300.

Fahrmeir, Ludwig, Thomas Kneib and Stefan Lang (2004). 'Penalized structured additive regression for space-time data: a Bayesian perspective'. In: *Statistica Sinica*, pp. 731–761.

Feng, Xiaoping, Jun Zhu, Pei-Sheng Lin and Michelle M Steen-Adams (2017). 'Composite likelihood approach to the regression analysis of spatial multivariate ordinal data and spatial compositional data with exact zero values'. In: *Environmental and Ecological Statistics* 24, pp. 39–68.

Filzmoser, Peter and Karel Hron (2008). 'Outlier detection for compositional data using robust methods'. In: *Mathematical Geosciences* 40.3, pp. 233–248.

Firth, David and Fiona Sammut (2023). 'Analysis of composition on the original scale of measurement'. In: *Preprint*.

Fisher, Thomas J, Jing Zhang, Stephen P Colegate and Michael J Vanni (2022). 'Detecting and modeling changes in a time series of proportions'. In: *The Annals of Applied Statistics* 16.1, pp. 477–494.

Frantsuzova, Anastasia (2021). 'Prior distributions for stochastic matrices'. PhD thesis. University of Leeds.

Gelman, Andrew (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, Andrew, Donald B Rubin et al. (1992). 'Inference from iterative simulation using multiple sequences'. In: *Statistical Science* 7.4, pp. 457–472.

Gong, Wenping, Seppo Parkkila, Xueqiong Wu and Ashok Aspatwar (2023). 'SARS-CoV-2 variants and COVID-19 vaccines: Current challenges and future strategies'. In: *International Reviews of Immunology* 42.6, pp. 393–414.

Greenacre, Michael, Marina Martínez-Álvaro and Agustín Blasco (2021). 'Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation'. In: *Frontiers in Microbiology* 12, p. 727398.

Hartigan, John A, Manchek A Wong et al. (1979). 'A k-means clustering algorithm'. In: *Applied Statistics* 28.1, pp. 100–108.

Harvey, William T, Alessandro M Carabelli, Ben Jackson, Ravindra K Gupta, Emma C Thomson, Ewan M Harrison, Catherine Ludden, Richard Reeve, Andrew Rambaut, Sharon J Peacock et al. (2021). 'SARS-CoV-2 variants, spike mutations and immune escape'. In: *Nature Reviews Microbiology* 19.7, pp. 409–424.

Hastie, Trevor J (2017). 'Generalized additive models'. In: *Statistical Models in S*. Routledge, pp. 249–307.

Hastings, W. Keith (1970). 'Monte Carlo sampling methods using Markov chains and their applications'. In: *Biometrika* 57.1, pp. 97–109.

Humaira, Hestry and R Rasyidah (2020). 'Determining the appropiate cluster number using elbow method for k-means algorithm'. In: *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*.

Huston, Carolyn and Carl Schwarz (2012). 'Hierarchical Bayesian strategy for modeling correlated compositional data with observed zero counts'. In: *Environmental and Ecological Statistics* 19, pp. 327–344.

Iorio, Carmela, Gianluca Frasso, Antonio D'Ambrosio and Roberta Siciliano (2016). 'Parsimonious time series clustering using p-splines'. In: *Expert Systems with Applications* 52, pp. 26–38.

Jain, Anil K (2010). 'Data clustering: 50 years beyond K-means'. In: *Pattern Recognition Letters* 31.8, pp. 651–666.

Janssen, Ian, Anna E Clarke, Valerie Carson, Jean-Philippe Chaput, Lora M Giangregorio, Michelle E Kho, Veronica J Poitras, Robert Ross, Travis J Saunders, Amanda Ross-White et al. (2020). 'A systematic review of compositional data analysis studies examining associations between sleep, sedentary behaviour, and physical activity with health outcomes in adults'. In: *Applied Physiology, Nutrition, and Metabolism* 45.10, S248–S257.

Karacan, C Özgen and Ricardo A Olea (2018). 'Mapping of compositional properties of coal using isometric log-ratio transformation and sequential Gaussian simulation–A comparative study for spatial ultimate analyses data'. In: *Journal of Geochemical Exploration* 186, pp. 36–49.

Khare, Shruti, Céline Gurry, Lucas Freitas, Mark B Schultz, Gunter Bach, Amadou Diallo, Nancy Akite, Joses Ho, Raphael TC Lee, Winston Yeo et al. (2021). 'GISAID's role in pandemic response'. In: *China CDC Weekly* 3.49, p. 1049.

Leininger, Thomas J, Alan E Gelfand, Jenica M Allen and John A Silander (2013). 'Spatial regression modeling for compositional data with many zeros'. In: *Journal of Agricultural, Biological, and Environmental Statistics* 18, pp. 314–334.

Link, William A and Mitchell J Eaton (2012). 'On thinning of chains in MCMC'. In: *Methods in Ecology and Evolution* 3.1, pp. 112–115.

Lobo, Manuel Duarte, Sérgio Miravent Tavares, Rui Pedro Pereira de Almeida and Manuel B Garcia (2025). 'Advancing Precision in Physical Education and Sports Science: A Review of Medical Imaging Methods for Assessing Body Composition'. In: *Global Innovations in Physical Education and Health*, pp. 293–326.

Lunn, David, David Spiegelhalter, Andrew Thomas and Nicky Best (2009). 'The BUGS project: Evolution, critique and future directions'. In: *Statistics in Medicine* 28.25, pp. 3049–3067.

Mardia, Kanti V and Peter E Jupp (2000). *Directional Statistics*. Vol. 2. Wiley Online Library.

Martín-Fernández, JA, Carles Barceló-Vidal and Vera Pawlowsky-Glahn (2000). 'Zero replacement in compositional data sets'. In: *Data Analysis, Classification, and Related Methods*. Springer, pp. 155–160.

Martín-Fernández, JA and S Thió-Henestrosa (2006). 'Rounded zeros: some practical aspects for compositional data'. In: *Geological Society, London, Special Publications* 264.1, pp. 191–201.

Martín-Fernández, Josep A, Carles Barceló-Vidal and Vera Pawlowsky-Glahn (2003). 'Dealing with zeros and missing values in compositional data sets using nonparametric imputation'. In: *Mathematical Geology* 35.3, pp. 253–278.

Martinez, Edson Z, Jorge A Achcar, Davi C Aragon and Marisa AA Brunherotti (2020). 'A Bayesian analysis for pseudo-compositional data with spatial structure'. In: *Statistical Methods in Medical Research* 29.5, pp. 1386–1402.

Matérn, Bertil (1960). 'Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations'. In: *Meddelanden fran Statens Skogsforskningsinstitut* 49.5.

Metropolis, Nicholas, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller and Edward Teller (1953). 'Equation of state calculations by fast computing machines'. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Murtagh, Fionn and Pedro Contreras (2012). 'Algorithms for hierarchical clustering: an overview'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1, pp. 86–97.

Naeini, Mahdi Pakdaman, Gregory Cooper and Milos Hauskrecht (2015). 'Obtaining well calibrated probabilities using Bayesian binning'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1.

Napier, Gary (2014). 'A Bayesian hierarchical model of compositional data with zeros: classification and evidence evaluation of forensic glass'. PhD thesis. University of Glasgow.

Napier, Gary, Agostino Nobile and Tereza Neocleous (2015). 'An online application for the classification and evidence evaluation of forensic glass fragments'. In: *Chemometrics and Intelligent Laboratory Systems* 146, pp. 418–425.

Neal, Radford M (2003). 'Slice sampling'. In: *The Annals of Statistics* 31.3, pp. 705–767.

Neocleous, Tereza, Colin Aitken and Grzegorz Zadora (2011). 'Transformations for compositional data with zeros with an application to forensic evidence evaluation'. In: *Chemometrics and Intelligent Laboratory Systems* 109.1, pp. 77–85.

Nguyen, Thi Huong An, Christine Thomas-Agnan, Thibault Laurent and Anne Ruiz-Gazen (2021). 'A simultaneous spatial autoregressive model for compositional data'. In: *Spatial Economic Analysis* 16.2, pp. 161–175.

Nielsen, Frank (2016). 'Hierarchical clustering'. In: *Introduction to HPC with MPI for Data Science*. Springer, pp. 195–211.

Nkemnole, EB and JO Oyewole (2023). 'An Analysis of the Hidden Markov Model for Surveilling the Transmission of Lassa Fever Epidemic Disease in Nigeria during Dry Season'. In: *International Journal of Tropical Disease & Health* 44.18, pp. 1–14.

Odeh, Inakwu OA, Alison J Todd and John Triantafilis (2003). 'Spatial prediction of soil particle-size fractions as compositional data'. In: *Soil Science* 168.7, pp. 501–515.

Oh, Junseop, Kyoung-Ho Kim, Ho-Rim Kim, Sunhwa Park and Seong-Taek Yun (2024). 'Using isometric log-ratio in compositional data analysis for developing a groundwater pollution index'. In: *Scientific Reports* 14.1, p. 12196.

Palarea-Albaladejo, J and JA Martin-Fernandez (2015). 'zCompositions – R package for multivariate imputation of left-censored data under a compositional approach'. In: *Chemometrics and Intelligent Laboratory Systems* 143, pp. 85–96. URL: http://dx.doi.org/10.1016/j.chemolab.2015.02.019.

Palarea-Albaladejo, Javier, Josep A Martín-Fernández and Juan Gómez-García (2007). 'A parametric approach for dealing with compositional rounded zeros'. In: *Mathematical Geology* 39.7, pp. 625–645.

Pawlowsky-Glahn, Vera and Juan José Egozcue (2016). 'Spatial analysis of compositional data: a historical review'. In: *Journal of Geochemical Exploration* 164, pp. 28–32.

Pawlowsky-Glahn, Vera, Juan José Egozcue and David Lovell (2015a). 'Tools for compositional data with a total'. In: *Statistical Modelling* 15.2, pp. 175–190.

Pawlowsky-Glahn, Vera, Juan José Egozcue, Ricardo A Olea and Eulogio Pardo-Igúzquiza (2015b). 'Cokriging of compositional balances including a dimension reduction and retrieval of original units'. In: *Journal of the Southern African Institute of Mining and Metallurgy* 115.1, pp. 59–72.

Pearson, Karl (1897). 'Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs'. In: *Proceedings of the Royal Society of London* 60.359-367, pp. 489–498.

– (1905). 'The problem of the random walk'. In: *Nature* 72.1867, pp. 342–342.

Petris, Giovanni, Sonia Petrone and Patrizia Campagnoli (2009). 'Dynamic linear models'. In: *Dynamic Linear Models with R*. New York, NY: Springer New York, pp. 31–84.

Pirzamanbein, Behnaz, Johan Lindström, Anneli Poska and Marie-José Gaillard (2018). 'Modelling spatial compositional data: Reconstructions of past land cover and uncertainties'. In: *Spatial Statistics* 24, pp. 14–31.

Plummer, Martyn (2003). 'JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling'. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Vienna, Austria, pp. 1–10.

Plummer, Martyn, Nicky Best, Kate Cowles and Karen Vines (2006). 'CODA: Convergence Diagnosis and Output Analysis for MCMC'. In: *R News* 6.1, pp. 7–11. URL: https://journal.r-project.org/archive/.

Qiu, Kun-Feng, Tong Zhou, David Chew, Zhao-Liang Hou, Axel Müller, Hao-Cheng Yu, Robert G Lee, Huan Chen and Jun Deng (2024). 'Apatite trace element composition as an indicator of ore deposit types: A machine learning approach'. In: *American Mineralogist* 109.2, pp. 303–314.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rabiner, Lawrence and Biinghwang Juang (1986). 'An introduction to hidden Markov models'. In: *IEEE ASSP Magazine* 3.1, pp. 4–16.

Ravishanker, Nalini, Dipak K Dey and Malini Iyengar (2001). 'Compositional time series analysis of mortality proportions'. In: *Communications in Statistics - Theory and Methods* 30.11, pp. 2281–2291.

Roux, Maurice (2018). 'A comparative study of divisive and agglomerative hierarchical clustering algorithms'. In: *Journal of Classification* 35, pp. 345–366.

Sangalli, Laura M, James O Ramsay and Timothy O Ramsay (2013). 'Spatial spline regression models'. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 75.4, pp. 681–703.

Scealy, JL and AH Welsh (2011). 'Regression for compositional data by using distributions defined on the hypersphere'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.3, pp. 351–375.

Schabenberger, Oliver and Carol A Gotway (2017). *Statistical methods for spatial data analysis*. Chapman and Hall/CRC.

Schielzeth, Holger and Shinichi Nakagawa (2013). 'Nested by design: model fitting and interpretation in a mixed model era'. In: *Methods in Ecology and Evolution* 4.1, pp. 14–24.

Scott, Steven L (2002). 'Bayesian methods for hidden Markov models: Recursive computing in the 21st century'. In: *Journal of the American Statistical Association* 97.457, pp. 337–351.

Shang, Han Lin, Steven Haberman and Ruofan Xu (2022). 'Multi-population modelling and forecasting life-table death counts'. In: *Insurance: Mathematics and Economics* 106, pp. 239–253.

Sharma, Shweta, Neha Batra et al. (2019). 'Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering'. In: *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE, pp. 568–573.

Sisk-Hackworth, Laura and Scott T Kelley (2020). 'An application of compositional data analysis to multiomic time-series data'. In: *NAR Genomics and Bioinformatics* 2.4.

Snyder, Ralph D, J Keith Ord, Anne B Koehler, Keith R McLaren and Adrian N Beaumont (2017). 'Forecasting compositional time series: A state space approach'. In: *International Journal of Forecasting* 33.2, pp. 502–512.

Stephens, Matthew (2000). 'Dealing with label switching in mixture models'. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.4, pp. 795–809.

Stephens, Michael A (1982). 'Use of the von Mises distribution to analyse continuous proportions'. In: *Biometrika* 69.1, pp. 197–203.

Stewart, Connie (2013). 'Zero-inflated beta distribution for modeling the proportions in quantitative fatty acid signature analysis'. In: *Journal of Applied Statistics* 40.5, pp. 985–992.

Stoner, Oliver and Theo Economou (2020a). 'An advanced hidden Markov model for hourly rainfall time series'. In: *Computational Statistics & Data Analysis* 152, p. 107045.

– (2020b). 'Multivariate hierarchical frameworks for modeling delayed reporting in count data'. In: *Biometrics* 76.3, pp. 789–798.

Stoner, Oliver, Gavin Shaddick, Theo Economou, Sophie Gumy, Jessica Lewis, Itzel Lucio, Giulia Ruggeri and Heather Adair-Rohani (2020c). 'Global household energy model: a multivariate hierarchical approach to estimating trends in the use of polluting and clean fuels for cooking'. In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 69.4, pp. 815–839.

Stroustrup, Bjarne (1986). 'An overview of C++'. In: *Proceedings of the 1986 SIGPLAN workshop on Object-oriented programming*, pp. 7–18.

Tahir, Hassam, Muhammad Shahbaz Khan, Fawad Ahmed, Abdullah M Albarrak, Sultan Noman Qasem and Jawad Ahmad (2023). 'Prediction of the SARS-CoV-2 Derived T-Cell Epitopes' Response Against COVID Variants'. In: *Computers, Materials & Continua* 75.2.

Tang, Zheng-Zheng and Guanhua Chen (2019). 'Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis'. In: *Biostatistics* 20.4, pp. 698–713.

Tanner, JM (1949). 'Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation'. In: *Journal of Applied Physiology* 2.1, pp. 1–15.

Taylor, Luke (2022). 'Covid-19: Omicron drives weekly record high in global infections'. In: *BMJ: British Medical Journal (Online)* 376.

Tibbits, Matthew M, Chris Groendyke, Murali Haran and John C Liechty (2014). 'Automated factor slice sampling'. In: *Journal of Computational and Graphical Statistics* 23.2, pp. 543–563.

Tjelmeland, Håkon and Kjetill Vassmo Lund (2003). 'Bayesian modelling of spatial compositional data'. In: *Journal of Applied Statistics* 30.1, pp. 87–100.

Tolosana-Delgado, Raimon, Ute Mueller and K Gerald van den Boogaart (2019). 'Geostatistics for compositional data: an overview'. In: *Mathematical Geosciences* 51.4, pp. 485–526.

Tolosana-Delgado, Raimon and KG Van Den Boogaart (2013). 'Joint consistent mapping of high-dimensional geochemical surveys'. In: *Mathematical Geosciences* 45, pp. 983–1004.

Torjesen, Ingrid (2021). *Covid-19: Omicron may be more transmissible than other variants and partly resistant to existing vaccines, scientists fear.*

Trobajo-Sanmartín, Camino, Iván Martínez-Baz, Ana Miqueleiz, Miguel Fernández-Huerta, Cristina Burgui, Itziar Casado, Fernando Baigorría, Ana Navascués, Jesús Castilla and Carmen Ezpeleta (2022). 'Differences in transmission between SARS-CoV-2 Alpha (B. 1.1. 7) and Delta (B. 1.617. 2) variants'. In: *Microbiology Spectrum* 10.2, e00008–22.

Tsagris, Michail (2018). 'Modelling Structural Zeros in Compositional Data'. In: Working Paper.

Tsagris, Michail, Simon Preston and Andrew TA Wood (2016). 'Improved classification for compositional data using the α-transformation'. In: *Journal of Classification* 33, pp. 243–261.

Tsagris, Michail and Connie Stewart (2018). 'A dirichlet regression model for compositional data with zeros'. In: *Lobachevskii Journal of Mathematics* 39.3, pp. 398–412.

Tsagris, Michail T, Simon Preston and Andrew TA Wood (2011). 'A data-based power transformation for compositional data'. In: *Proceedings of the 4th Compositional Data Analysis Workshop*.

Tsilimigras, Matthew CB and Anthony A Fodor (2016). 'Compositional data analysis of the microbiome: fundamentals, tools, and challenges'. In: *Annals of Epidemiology* 26.5, pp. 330–335.

Valpine, Perry de, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang and Rastislav Bodik (2017). 'Programming with models: writing statistical algorithms for general model structures with NIMBLE'. In: *Journal of Computational and Graphical Statistics* 26.2, pp. 403–413.

Vellingiri, Balachandar, Kaavya Jayaramayya, Mahalaxmi Iyer, Arul Narayanasamy, Vivekanandhan Govindasamy, Bupesh Giridharan, Singaravelu Ganesan, Anila Venugopal, Dhivya Venkatesan, Harsha Ganesan et al. (2020). 'COVID-19: A promising cure for the global panic'. In: *Science of the Total Environment* 725, p. 138277.

Walvoort, Dennis JJ and Jaap J de Gruijter (2001). 'Compositional kriging: a spatial interpolation method for compositional data'. In: *Mathematical Geology* 33.8, pp. 951–966.

Wang, Huiwen, Qiang Liu, Henry MK Mok, Linghui Fu and Wai Man Tse (2007). 'A hyperspherical transformation forecasting model for compositional data'. In: *European Journal of Operational Research* 179.2, pp. 459–468.

Ward Jr, Joe H (1963). 'Hierarchical grouping to optimize an objective function'. In: *Journal of the American Statistical Association* 58.301, pp. 236–244.

Watanabe, Sumio and Manfred Opper (2010). 'Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory.' In: *Journal of Machine Learning Research* 11.12.

Watkins, Rochelle E, Serryn Eagleson, Bert Veenendaal, Graeme Wright and Aileen J Plant (2009). 'Disease surveillance using a hidden Markov model'. In: *BMC Medical Informatics and Decision Making* 9.1, pp. 1–12.

Willmott, Cort J and Kenji Matsuura (2005). 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance'. In: *Climate Research* 30.1, pp. 79–82.

Wong, Tzu-Tsung (1998). 'Generalized Dirichlet distribution in Bayesian analysis'. In: *Applied Mathematics and Computation* 97.2-3, pp. 165–181.

Wood, S. N. (2003). 'Thin-plate regression splines'. In: *Journal of the Royal Statistical Society (B)* 65.1, pp. 95–114.

Wood, Simon N (2016). 'Just another gibbs additive modeller: interfacing JAGS and mgcv'. In: *Journal of Statistical Software*.

– (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Yoo, Jinkyung, Zequn Sun, Michael Greenacre, Qin Ma, Dongjun Chung and Young Min Kim (2022). 'A guideline for the statistical analysis of compositional data in immunology'. In: *Communications for Statistical Applications and Methods* 29, pp. 453–469.

Yoshida, Takahiro and Morito Tsutsumi (2018). 'On the effects of spatial relationships in spatial compositional multivariate models'. In: *Letters in Spatial and Resource Sciences* 11, pp. 57–70.

Zadora, G, T Neocleous and CGG Aitken (2010a). 'Recent developments in likelihood ratio models for multivariate compositional data'. In: *Science & Justice* 1.50, p. 30.

Zadora, Grzegorz, Tereza Neocleous and Colin Aitken (2010b). 'A two-level model for evidence evaluation in the presence of zeros'. In: *Journal of Forensic Sciences* 55.2, pp. 371–384.

Zhou, Shanglin, Paolo Braca, Stefano Marano, Peter Willett, Leonardo M Millefiori, Domenico Gaglione and Krishna R Pattipati (2021). 'Application of hidden Markov models to analyze, group and visualize spatio-temporal COVID-19 data'. In: *IEEE Access* 9, pp. 134384–134401.

Zuo, Renguang, Qinglin Xia and Haicheng Wang (2013). 'Compositional data analysis in the study of integrated geochemical anomalies associated with mineralization'. In: *Applied Geochemistry* 28, pp. 202–211.