Bayesian Approaches to Issues Arising in Spatial Modelling

A THESIS SUBMITTED TO
THE SCIENCE AND ENGINEERING FACULTY
OF QUEENSLAND UNIVERSITY OF TECHNOLOGY
IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Earl Duncan

Principal supervisor: Professor Kerrie Mengersen

Associate supervisor: Doctor Nicole White

School of Mathematical Sciences

Science and Engineering Faculty

Queensland University of Technology

2017

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: QUT Verified Signature

Date: 25 Aug 2017

Abstract

The collection of spatial and spatiotemporal data is ever-growing, as is the demand for methodologically advanced and computationally efficient statistical models to analyse them. Spatial data pose some unique and challenging problems to statisticians. One of the most prominent issues in spatial data analysis is spatial autocorrelation, a term which implies that the underlying process(es) which generates the data results in observations which are similar if they are nearby geographically. Neglecting this phenomenon can lead to inaccurate conclusions. However, accounting for spatial autocorrelation is not straightforward. This problem is exacerbated when spatial data is collected over multiple time points, known as spatiotemporal data. One particular solution to this problem, which reoccurs throughout this thesis, can be viewed as an advantage of the statistical methodology.

Another problem that frequently arises in the analysis of spatial data is the task of identifying and/or classifying subsets of the data. Examples include identifying areas associated with a high risk, identifying points or areas that deviate from some norm, and classifying the data into groups according to some common attribute. The purpose of such analyses is to provide insight into the nature of the processes driving the data and possibly offer advice on how the spatial processes may be altered or interrupted in order to produce desirable outcomes.

This thesis aims to address these issues, with a particular focus on spatio-temporal data. The thesis begins with an application of Bayesian hierarchical presented in the epidemiological context of analysing mammography screening rates. The models are designed to identify areas which are associated with aberrant temporal trends, and reveal any spatial patterns according to the nature of the temporal trends. Uncovering these patterns is an important step in understanding the reasons for service under-utilisation, and it aids managers to improve these health services. This initial analysis revealed potential methodological improvements which prompted a second analysis of the same data to exemplify the extensions to these models.

The experience from performing these analyses and greater familiarity with the literature raised some new research questions. These questions relate to two fundamental aspects of the methodology used in the analysis of the mammography screening data, and have been commonly used in other spatial and spatio-temporal analyses. These aspects are spatial smoothing and mixture models. Despite the vast literature on both concepts, there remain several important questions and opportunities for improving the methodology.

On the subject of spatial smoothing, this thesis highlights the positive effect that spatial smoothing has on model fit and predictive ability, even when spatial autocorrelation of the observed data is not particularly strong. Various definitions of the smoothing weights are reviewed, and new alternatives are proposed, including the notion of covariate-based weights.

Mixture models allow for a more flexible model specification by allowing the likelihood to be described as a mixture of component densities. However, in the context of Bayesian estimation, the posterior distribution can become invariant to the permutations of the labels, which makes statistical inference difficult. A common response to this problem is to attempt to prevent label switching from occurring, while at the same time the literature indicates that label switching is a necessity. This is one example of the misunderstandings surrounding the subject of label switching that this thesis seeks to clarify. More significantly, existing relabelling algorithms designed to undo the label switching are reviewed in a systematic manner for the first time, and a new relabelling algorithm is proposed.

Keywords

Spatial, Spatio-temporal, Mixture, Label switching, Relabelling Algorithm, Mammography Screening, Cancer, Epidemiology, Bayesian, Smoothing, Covariate Space

Acknowledgements

To my supervisors, Kerrie Mengersen and Nicole White, thank you for the supervision you have provided me while allowing me to explore interesting research topics. This thesis is a testament to your patience, guidance, and tireless efforts to provide me with helpful feedback and keep me on track.

To all my lecturers and mentors at QUT who have taught and inspired me, thank you. To you I owe my knowledge of statistics, and my programming and research skills. This thesis would not have been possible without your contribution to my academic prowess.

Special thanks to the members of the Bayesian Research and Application Group for sharing your ideas and knowledge. And to all my colleagues at QUT, I wish to express my thanks for your valuable support – your tips and tricks for Bayesian modelling, hacks for R, and assistance with those vexatious travel forms is much appreciated. It has also been a privilege to get acquainted with each of you and share some laughs together.

Last but not least, I wish to express my sincere gratitude to my fiancé, family, and friends. Your endless love, encouragement, and gentle words of wisdom will not be forgotten.

Table of Contents

List of Figures

List of Tables

List of Publications Arising from this Thesis

Chapter 3: Duncan, E. W., White, N. M., and Mengersen, K. 2016. Bayesian spatiotemporal modelling for identifying unusual and unstable trends in mammography utilisation. *BMJ Open*, **6** (5): e010253. doi: 10.1136/bmjopen-2015-010253.

Chapter 4: Duncan, E. W., White, N. M., and Mengersen, K. Improved Bayesian methods for identifying aberrant temporal trends in spatio-temporal data with application to mammography screening services, publication submitted to *PLOS ONE*.

Chapter 5: Duncan, E. W. and Mengersen, K. New and existing solutions to the label switching problem in Bayesian mixture models: a systematic review, publication submitted to *Journal of the American Statistical Association*.

Chapter 6: Duncan, E. W., White, N. M., and Mengersen, K. Spatial Smoothing in Bayesian Models: A Comparison of Weights Matrix Specifications and their Impact on Inference, publication submitted to *International Journal of Health Geographics*.

Chapter 1

Introduction

1.1 Research Aim and Objectives

The overall research aim of this thesis is to apply and develop Bayesian models for analysing spatio-temporal data exhibiting aberrant temporal trends, and to extend the related methodology.

To meet this aim, the following research objectives are considered:

1. To apply recently developed Bayesian models for analysing spatio-temporal data in order to understand the temporal patterns of mammography screening service utilisation in Brisbane.

2. To extend the models in objective 1 through the inclusion of covariates and a mixture model in order to address the limitations of these models and improve inferences.

3. To extend the methodology of spatial modelling by systematically reviewing existing relabelling algorithms and proposing a new relabelling algorithm to deal with the label switching problem affecting Bayesian mixture models.

4. To extend the methodology of spatial modelling by analysing the impact of the spatial weights matrix on inference.

These objectives cover a broad range of contemporary issues pertaining to spatial analysis and can be broadly classified into two groups: objectives 1 and 2 relate to the application and development of spatial models in the epidemiological context of cancer screening services;

objectives 3 and 4 relate to the improvement of specific methodological aspects frequently implemented in such models.

Objectives 1 and 2 are focused on applying contemporary Bayesian models to spatio-temporal data, in anticipation of identifying areas with aberrant temporal trends and consequently providing some understanding of the complex spatio-temporal processes driving the data. Specifically, the models are applied to mammography screening utilisation data in Brisbane with the overall aim of providing insight into the potential factors influencing mammography service utilisation rates to assist decision makers with the task of managing and improving these services. Aside from answering specific research questions relating to mammography screening, the initial analysis serves as a benchmark for the usefulness of these models and exploratory tool for identifying potential extensions to the methodology.

The statistical models considered in the analyses are sophisticated and purposefully designed to answer the research questions of interest. However, some limitations of these models have already been disclosed in the literature, and it is anticipated that an analysis using the models will provide further guidance on how the implementation and interpretation of the models may be enhanced. Extending these models is the focus of objective 2. The success of the proposed extensions is evaluated by comparing results of the initial and subsequent analyses.

Objectives 3 and 4 do not necessarily relate to mammography screening utilisation specifically or even epidemiology in general. However, these objectives arise in response to unanswered questions originating from two concepts which are fundamental to the methodology considered for the analysis of mammography screening data, and answers to these questions will most certainly be useful in many applications of spatial statistics and epidemiology. The first of these concepts is mixture models, or more precisely, the relabelling algorithms designed to reverse the effects of label switching – a phenomenon which can render statistical inference for mixture models meaningless. The second concept is spatial smoothing, which is a technique commonly used in spatial analysis to account for spatial autocorrelation and reduce the variability of spatial estimates. While these two research objectives relate to very particular methodological aspects, the implications are far-reaching.

1.2 Specific Research Questions

Evidence indicates that regular mammography screening helps detect early stages of breast cancer, and therefore irregular patterns in mammography screening may suggest a need for further investigation or intervention. At the very least, understanding the complex spatio-temporal processes behind the observed patterns assists in the management of mammography screening services. This is the rationale for modelling screening utilisation data with the intent of identifying areas with aberrant temporal trends. Specifically, the first research objective seeks to answer questions such as

- How does mammography screening utilisation in Brisbane vary over time?
- Which areas are associated with aberrant temporal trends?
- Do these areas have any characteristics in common? For example, are they rural rather than urban?
- What are the possible explanations for the aberrant nature of temporal trends?

Answering these questions is an important initial step in improving management of these services. A more detailed description of the scientific context and explanation of the reasons behind the focus on temporal trends is provided in Chapter 3.

The research aimed at answering the questions above is focused on the application of existing methods to estimate unknown parameters in a statistical model and make inferences based on those estimates. The second research objective is focused on reviewing and extending the methodology relating to the implementation and interpretation of the models in order to address some of the limitations of the models and improve statistical inferences. Specific details on these limitations and extensions require considerable background on the statistical framework, and are therefore deferred until Chapters 3 and 4. However, two concepts which are fundamental to the methodology considered in these analyses are mixture models and spatial smoothing. Each of these concepts gives rise to new research questions and objectives.

The third research objective is concerned with the phenomenon known as label switching and its impact on statistical inference in Bayesian mixture models. Mixture models allow a great deal of flexibility by allowing the likelihood to be described by a mixture of component distributions.

However, under certain conditions, the posterior distribution becomes invariant to permutations of the labels identifying the mixture components, which can be problematic for inference. Despite the substantial literature on mixture models and even the label switching phenomenon, the exact nature of label switching and its impact on inference is ambiguous and has been the subject of debates amongst academics. There are several facets of this research objective. The first is to review the literature on label switching and provide clarity on the subject. This includes providing answers to the following fundamental questions:

- Under what conditions will label switching occur?
- What is the extent of label switching as a problem for meaningful inference?
- If label switching makes inference difficult, why is it desirable?

Definitive answers to these questions will clear up the ambiguity and controversies surrounding this phenomenon. If label switching does occur and poses a problem for inference, then the only solution is to apply a relabelling algorithm to reverse the undesirable effects. The second facet of this research objective is to review and compare the existing relabelling algorithms from the literature. Although the literature does contain reviews of some methods, a systemic review with unified, consistent definitions and notation is currently lacking, as are comprehensive results on the accuracy and computational efficiency of each algorithm. This gives rise to two more facets of this research objective: to provide an extensive comparison of the algorithms by performing a simulation study under different scenarios designed to accentuate differences in accuracy and efficiency; and to answer the challenge of devising a new relabelling algorithm which outperforms existing algorithms.

The fourth research objective focuses on spatial smoothing refers to the idea of flattening the peaks and troughs of a spatial surface. In the context of disease mapping, the surface may represent incidence or mortality of a disease. If the aim is to identity spatial patterns of the underlying relative risk of that disease, then simply mapping the observed values can be misleading because the peaks and troughs may be exaggerated due to sampling variability, thereby obscuring genuine spatial patterns. Using statistical models to estimate the risk surface not only allows spatial smoothing, but also the inclusion of covariates which help to explain differences in the observed values. Smoothing is typically implemented by constructing a matrix of weights which captures the degree of spatial dependencies between neighbouring

areas, and thus controls the degree of smoothing. The literature seems to indicate that spatial smoothing is advantageous, but several questions require clarity. Some of the questions that objective 4 seeks to answer are:

- Is spatial smoothing necessary when the data are not spatially autocorrelated?
- What is the detriment of ignoring spatial smoothing?
- What are the recommendations for the number of neighbours?
- How should the smoothing weights be specified?

Regarding this last point, new strategies are proposed as alternatives to the common distance- or adjacency-based weight specifications. The answers to these questions are sought by comparing various weight specifications in the analysis of real and artificial data sets.

1.3 Thesis Outline

This thesis is intended to fulfil the requirements of thesis by publication. Chapters 3 through 6 are self-contained in the form of journal papers, and consequently there is some overlap in the content between these chapters. The references cited throughout the following chapters are listed in a bibliography at the end of the thesis. The structure of this thesis is as follows.

Chapter 2 provides a detailed literature review of the key ideas found in the subsequent chapters. There is some overlap between this literature review and the individual literature reviews in each of the chapters 3 through 6. The purpose of chapter 2 is to provide the reader with a broader statistical and contextual background than is possible within the confines of the chapters constituting journal papers.

Chapter 3 presents an application of Bayesian spatio-temporal modelling in the epidemiological context of cancer screening. Two state-of-the-art models presented in the literature are applied to 12 years of mammography screening utilisation data in Brisbane. These models are able to provide unique insight into the spatial and temporal patterns, and identify a number of potential extensions. The main results from the work presented in this chapter were disseminated at the Bayes on the Beach Conference, Gold Coast, Australia, 10-12 November 2014. This chapter in its entirety was published as a paper in *BMJ Open*.

Chapter 4 extends the methodology presented in Chapter 3. Several refinements are made to the models and the methods used to classify the nature of temporal trends. The refined models are applied to the same data to illustrate the significance of these improvements. Featured amongst the extensions to the methodology is the use of a mixture model which helped overcome one of the model limitations and simultaneously allowed more informative inferences. This chapter has been submitted for publication to the open access journal *PLOS ONE*. The results from the work presented in Chapters 3 and 4 were presented at the Social Computing Summit, Gold Coast, Australia, 14-15 December 2016.

Chapter 5 focuses on the theoretical and practical issues of label switching arising from the application of mixture models in the Bayesian framework. This chapter attempts to clarify some of the confusion and common misconceptions about label switching. Existing relabelling algorithms found in the literature are reviewed and a new relabelling algorithm is proposed. These algorithms are compared by testing their accuracy and efficiency in a simulation study. Some of the early findings were presented at the Bayes on the Beach Conference, Gold Coast, Australia, 07-09 December 2015, and the International Society for Bayesian Analysis, Sardinia, Italy, 13-17 June 2016. This chapter has been submitted for publication to the *Journal of the American Statistical Association*.

Chapter 6 address the issue of spatial smoothing. This chapter explains the rationale of spatial smoothing and reviews the methodologies for incorporating spatial smoothing in a statistical model with emphasis on generalised linear mixed models. Commonly used specifications of the weights matrix are also reviewed and alternative specifications are proposed. A comparison of the resulting models not only highlights the importance of spatial smoothing, but also reveals the effect that the weights matrix can have on spatial smoothing and ultimately the performance of the model. This chapter was submitted for publication to the *International Journal of Health Geographics*.

Chapter 7 summarises the main findings presented in this thesis, discusses the practical implications, and identifies potential extensions for future research.

Chapter 2

Literature Review

The following chapters in this thesis each contain an introductory section in which the relevant literature is reviewed in detail. The purpose of this chapter is to provide an overview of the literature on the key topics raised in this research. This chapter is organised as follows. Section 2.1 explains the theory of random fields, in particular Markov random fields, and describes several related models which feature throughout this thesis. Section 2.2 explores the notion of spatial modelling, with emphasis on spatio-temporal modelling and disease mapping. Statistical models are outlined and the concept of smoothing is also discussed. Section 2.3 describes and justifies the Bayesian approach to modelling, including specification of prior information and sampling techniques.

2.1 Random Fields

Consider an n-dimensional space where a generic location is denoted by the vector $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$. The elements $t_1, t_2, \ldots, t_n$ usually represent spatial coordinates or time, or a combination of both, but may also represent nominal coordinates. Collectively, these coordinates identify a unique location within the n-dimensional space (Shaddick and Zidek 2016; Vanmarcke 2010).

Consider the random variables $\{X_l = X(\boldsymbol{t}_l)\}$ for $l = 1, \ldots, L$ locations where realisations of these random variables are denoted by $x_1, \ldots, x_L$. If these random variables are independent, then

$$p(x_1, x_2, \ldots, x_L) = p(x_1)p(x_2)\ldots p(x_L)$$

for all values of $\{x_l\}$, where the left hand side represents the joint probability distribution of the random variables and the right hand side represents the product of the marginal distributions (here $p()$ is used laxly to denote a generic distribution). Conversely, the outcome of some variables may depend on the outcome of others. For example, suppose the locations represent a one-dimensional parameter space, such as time, and the random variables depend on the previous values. Then

$$p(x_1, x_2, \ldots, x_L) = p(x_1)p(x_2|x_1)\ldots p(x_L|x_1, x_2, \ldots, x_{L-1})$$

and the family of random variables $\{X_1, X_2, \ldots, X_L\}$ is called a stochastic (or random) process. If the dependence is limited to the previous p values, then this process is known as an autoregressive model of order p, often abbreviated to AR(p). More generally, if the parameter space is multidimensional, which permits far more complex dependencies, then the family of random variables is called a random field (Vanmarcke 2010).

> "An ideal random field model succeeds in capturing the essential features of a complex random phenomenon in terms of a minimum number of physically meaningful and experimentally accessible parameters" (Vanmarcke 2010, pp. 1).

The nature of the parameter space and the outcome of the random variables define the type of random field. For example, if the parameters are discrete and the observations are continuous, then the resulting random field is said to be a continuous-state, discrete-space random field. Some specific examples of random fields can be found in (Vanmarcke 2010).

Returning to the case of a one-dimensional parameter space, suppose that the process $\{X_1, \ldots, X_L\}$ is a first-order, discrete-space, stochastic process such that

$$p(x_l|x_{\setminus l}) = p(x_l|x_{l-1}) \tag{2.1}$$

for $l > 1$, where $x_{\setminus l}$ is shorthand for $\{x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_{l+1}\}$. Since the outcome of the random variable at location l depends on the outcome of the variable at $l-1$, then location $l-1$ is said to be a neighbour of l. The process described by Equation (2.1) indicates that the dependency between locations is unidirectional. For the bi-directional case, the process would

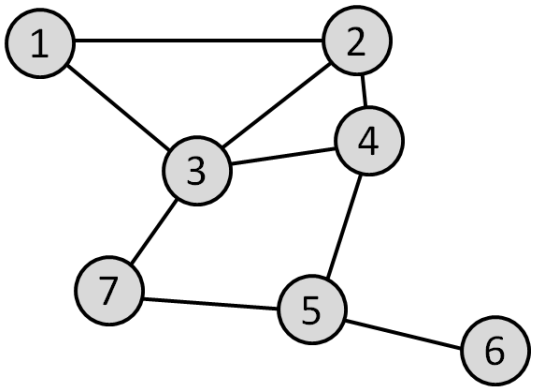
be expressed as

$$p(x_l|x_{\setminus l}) = p(x_l|x_{l-1}, x_{l+1}) \tag{2.2}$$

and each location l, for $2 \leq l \leq L-1$, would have 2 neighbours, namely $l-1$ and $l+1$, and the endpoints of the one-dimensional parameter space, 1 and L, would have only 1 neighbour each, namely locations 2 and $L-1$ respectively (Besag 1974; Vanmarcke 2010). Both of the processes described by Equations (2.1) and (2.2) are known as Markov processes because they possess the Markov property, which is conveyed by the conditional independence (Rue and Held 2005). For these two processes, the Markov property implies that the value of the random variable at one location depends only on the value of the random variable at the locations of immediately adjacent neighbour(s). This property is sometimes referred to as the 'memoryless' property, a synonym whose relevance may be more apparent when Equation (2.1) represents a first-order temporal process, where l denotes discrete points in time. Of course the Markov property and notion of conditional independence extend to n^{th}-order processes too. For example, if the values of a second-order temporal process depend only on the values at the previous two time periods, then such a process is still considered a Markov process.

For the example processes above, the neighbours were derived from the relationship between the random variables defining the stochastic process. In practice, the true nature of stochastic processes is not observable. To model a stochastic process, a neighbourhood structure is assumed, and this dictates the process (Besag 1974). This is true for higher dimensional parameter spaces too.

Now consider the case of a two-dimensional parameter space represented by a rectangular lattice, where locations are denoted by coordinate pairs (i, j). The first-order stochastic process can be generalised to a random field by defining a first-order neighbourhood for the locations. Each internal location has four neighbours, namely $(i-1, j)$, $(i+1, j)$, $(i, j-1)$, and $(i, j+1)$, while locations that lie on the boundaries will have two or three, depending on whether the location is at a corner. Of course more complex neighbourhood structures are possible, including extensions to higher-order processes, and each specification results in a different random field (Besag 1974). The parameter space is not restricted to regular lattices either; neighbourhoods can also be derived for irregular lattices and even point processes (Cressie 1993; Diggle 2013). A random field which possesses the Markov property is referred to as a Markov random field (MRF).



$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Figure 2.1: Relationship between an undirected graph and an adjacency matrix.

There are two common ways of representing the neighbourhood structure: as a finite graph, or as an adjacency matrix. Figure 2.1 shows an example of a simple undirected graph and the corresponding adjacency matrix where 1 denotes a connection between locations. Note that since the graph is undirected, the adjacency matrix is symmetric (Harary 1962).

The term 'Markov random field' is also given to the class of parametric models which encompass the theory of MRFs. MRFs have played an influential role in the development of statistical methodology for analysing spatial data (Clifford 1990), and their importance will be made evident throughout the remainder of this section and the following section. Note that a 'saturated' neighbourhood structure is also possible, in which each location includes every other location as neighbours. In this case, the random field is not a MRF, but alternative approaches to spatial modelling are available (Besag and Kooperberg 1995; Cressie 1993; Diggle 2013). However, what has not yet been discussed is the influence that each random variable has at neighbouring locations. Varying degrees of influence can be specified so that nearby neighbours have greater influence than neighbours further away, for example, by defining the weights as a decreasing function of distance (Congdon 2010; Earnest et al. 2007; Fahrmeir and Kneib 2011; Getis and Aldstadt 2008). If a threshold is applied such that neighbours with negligibly small weights are dropped, then the Markov property is restored.

In Chapters 3 and 4, a three-dimensional parameter space is considered; two parameters identifying location in Euclidean space[1], namely latitude and longitude, and one parameter for

[1] Although the earth's surface is not a Euclidean space, any map projection of a geographic coordinate system will result in a projected coordinate system on a plane which is Euclidean. However, calculating distances between the projected coordinates is not as accurate as calculating great-circle distances (unless distances are preserved by the projection) due to the curvature of the earth's surface. Therefore, in the subsequent chapters of this thesis, graphical illustrations are created using the projected coordinates, but all distances between coordinates represent great-circle distances. See Yang et al. (2000) for a comprehensive overview on the subject of map projections.

discrete time. In chapter 6, only spatial data are considered, and therefore only the two spatial coordinates are required to identify locations. For both the spatio-temporal data in Chapters 3 and 4 as well as the spatial data in Chapter 6, the spatial coordinates refer to an irregular lattice. Since the locations represented by the coordinates are not points, but rather areas which may vary greatly in size and shape and have no logical ordering, it is convenient to make the simplification of reducing the spatial coordinates to a single index which may be assigned to the areas arbitrarily. This index set together with a map convening the spatial dependencies comprises a spatial lattice (Cressie 1993). This is the approach used in Chapters 3, 4, and 6, and is typical in the literature when the spatial units are determined by administrative boundaries (for example, see Abellan et al. (2008), Breslow and Clayton (1993), citetBern95, Best et al. (2005), and Li et al. (2012)). When the locations are denoted by a single index set, the definition of neighbours will require a more elaborate scheme than the four nearest neighbours in the cardinal directions. That is not to say that a 'nearest neighbours' scheme is not feasible for irregular lattices, but the manner of defining the neighbours is less straightforward.

As aforementioned, the term 'Markov random field' may describe a generalisation of a Markov process to higher dimensions, or to describe a model which formalises the functional form of $p(x_l|x_{\setminus l})$, that is, which models the underlying random field. A model which models a Markov process is referred to as a Markov model, which can be considered a special case of a MRF model. Another special case of a MRF is a hidden MRF where the states are hidden, also referred to as 'latent' or 'unobserved'. A hidden Markov model (HMM) is therefore a special case of a hidden MRF model where the parameter space is one-dimensional, and a special case of a Markov model in which the states are hidden.

Two more special cases of a MRF are the mixture model and the conditional autoregressive (CAR) model. Both mixture models and a specific variety of the CAR model feature extensively in the following chapters, and therefore warrant a more detailed description, which is provided below. Figure 2.2 depicts the relationship between all these models which are generalised by the Markov random field (MRF) model.

2.1.1 CAR Model

The term 'CAR model' may actually refer to one of several models. In the Bayesian framework, the CAR model may refer to a specific prior distribution or a general class of CAR priors (Lee

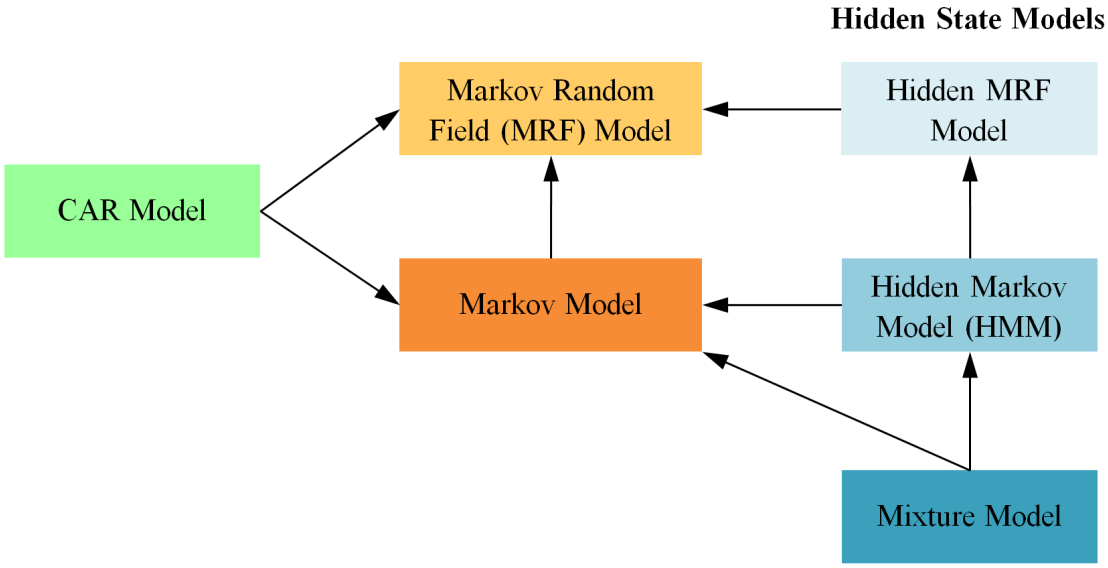


Figure 2.2: Relationships between a various models which can all be considered special cases of the Markov random field (MRF) model. The models at the head of the arrows are generalisations of those at the tail.

2011). To illustrate, consider the random field $\boldsymbol{S} = \{S_1, \ldots, S_N\}$. The first CAR prior which was introduced by Besag (1974) and later clarified by Besag et al. (1991) [2], called the intrinsic CAR (ICAR) prior, is defined by the set of conditional distributions

$$S_i|\boldsymbol{s}_{\setminus i} \sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}} \sum_j w_{ij} s_j, \frac{\sigma_s^2}{\sum_j w_{ij}}\right) \tag{2.3}$$

where w_{ij} is the element of a spatial weights matrix $\mathbf{W}$ corresponding to row i and column j (Assunção and Krainski 2009; Banerjee et al. 2014; Besag et al. 1991; Congdon 2010; Fahrmeir and Kneib 2011; Johnson 2004; Kandhasamy and Ghosh 2017; Lee 2011, 2013; Lee and Mitchell 2012; Lu et al. 2007; Mugglin et al. 1999; Rue and Held 2005; Shaddick and Zidek 2016; Wall 2004). $\mathbf{W}$ determines the spatial proximity between the random effects, and it is most commonly defined as a binary, first-order, adjacency matrix, whereby

$$w_{ij} = \begin{cases} 1 & \text{if areas } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases}. \tag{2.4}$$

Areas are considered adjacent if they share a common boundary (Banerjee et al. 2014; Bell and

[2] Besag et al. (1991) were the first to use the term 'intrinsic' to describe this specific CAR formulation. This is probably why Besag et al. (1991) are often credited with the introduction of the ICAR prior rather than Besag (1974).

Bockstael 2000; Bernardinelli et al. 1997; Best et al. 2001; Congdon 2010; Conlon and Waller 1999; Earnest et al. 2007; Fahrmeir and Kneib 2011; Lee 2011; Morrison et al. 2012; Wall 2004; Waller et al. 1997; Xia and Carlin 1998). By definition, this construction implies that $w_{ii} = 0$, which is a general principle that applies to any definition of weights (Assunção and Krainski 2009; Banerjee et al. 2014; Getis and Aldstadt 2008). This model implies that the conditional expectation of S_i is equal to the mean of the random effects at neighbouring locations. If the $\{w_{ij}\}$ are given by Equation (2.4), then the conditional variance is inversely proportional to the number of neighbours (Lee 2011, 2013).

When the ICAR prior is considered within the broader class of generalised linear models, described below in Section 2.2.3, the random field given by Equation (2.3) can be regarded as structured spatial random effects which account for the implied neighbourhood structure. In these statistical models, it is common to also include unstructured spatial random effects. This particular model is sometimes referred to as the convolution model, or the BYM model in honour of Besag et al. (1991) who proposed this variation (Abellan et al. 2008; Best et al. 2001; Breslow and Clayton 1993; Kandhasamy and Ghosh 2017; Lee 2011, 2013). Despite the subtle modification, the BYM model is regarded as distinct CAR formulation in its own right, with several studies aimed at comparing the differences resulting from use of the ICAR and BYM models, amongst others (for example, see Kandhasamy and Ghosh (2017), and Lee (2011)).

Brook's Lemma (Brook 1964) allows the joint probability distribution for $\boldsymbol{S}$ to be obtained from the full conditionals given by Equation (2.3). For the ICAR prior, this turns out to be a multivariate Gaussian distribution

$$\boldsymbol{S} \sim MVN\left(\mathbf{0}, \mathbf{B}\right) \tag{2.5}$$

where $\mathbf{B}$ is a precision matrix,

$$\mathbf{B} = \frac{1}{\sigma^2}\left(\mathbf{D} - \mathbf{W}\right) \tag{2.6}$$

and $\mathbf{D}$ is a diagonal matrix with elements

$$d_{ii} = \sum_{j}^{N} w_{ij}$$

(Banerjee et al. 2014; Besag et al. 1991; Conlon and Waller 1999; Kandhasamy and Ghosh 2017; Lee 2011; Lee and Mitchell 2012; Lu et al. 2007; Rue and Held 2005; Shaddick and

Zidek 2016; Wall 2004). The full conditionals for the ICAR prior given by Equation (2.3) are proper, but the joint distribution given by Equation (2.5) with precision matrix $\mathbf{B}$ given by (2.6) is improper since $\mathbf{B}$ is singular (Assunção and Krainski 2009; Banerjee et al. 2014). The impropriety of the ICAR prior can be overcome by redefining the precision matrix

$$\mathbf{B} = \frac{1}{\sigma^2}(\mathbf{D} - \rho\mathbf{W})$$

such that the full conditionals are

$$S_i|\boldsymbol{s}_{\setminus i} \sim \mathcal{N}\left(\frac{\rho}{\sum_j w_{ij}}\sum_j w_{ij}s_j, \frac{\sigma_s^2}{\sum_j w_{ij}}\right) \tag{2.7}$$

with the constraint $|\rho| < 1$. This ensures that the precision matrix $\mathbf{B}$ is positive definite and therefore $\boldsymbol{S}$ has a proper joint distribution (Banerjee et al. 2014; Kandhasamy and Ghosh 2017; Lekdee and Ingrisawang 2013; Lu et al. 2007). The prior specified by Equation (2.7) is referred to as the proper CAR prior (Assunção and Krainski 2009; Cressie 1993; Kandhasamy and Ghosh 2017; Lee 2011).

Another variation of the CAR model was proposed by Leroux et al. (1999),

$$S_i|\boldsymbol{s}_{\setminus i} \sim \mathcal{N}\left(\frac{\rho\sum_j w_{ij}s_j + (1-\rho)\mu}{\rho\sum_j w_{ij} + 1 - \rho}, \frac{\sigma_s^2}{\rho\sum_j w_{ij} + 1 - \rho}\right), \tag{2.8}$$

which only requires a single set of random effects (Kandhasamy and Ghosh 2017; Lee 2011; MacNab 2003). The ICAR prior is therefore a limiting case of both CAR priors given by Equations (2.7) and (2.8) when ρ is set to 1 (Assunção and Krainski 2009; Lee 2013; MacNab 2003; Wall 2004).

In practice, CAR models are often used in Bayesian analyses as prior distributions for spatial random effects as a means of accounting for spatial autocorrelation. In this case, $\boldsymbol{S}$ represents a spatial random field. (The topics of spatial autocorrelation and spatial random fields are discussed in more depth in Section 2.2.) However, the CAR models are not limited to spatial random fields. For example, the ICAR prior has been used to model temporal and spatio-temporal random fields (Abellan et al. 2008; Duncan et al. 2016; Li et al. 2012). Nor are the CAR formulations restricted to the Gaussian case – analogous formulations for the binomial, Poisson, and exponential cases are provided in Besag (1974).

The Spatial Weights Matrix

The spatial weights matrix $\mathbf{W}$ is an important component of many spatial models, not only CAR models (Banerjee et al. 2014; Florax and Rey 1995; Wall 2004; Zhang 2012). Equation (2.4) is the most common definition used for the weights, but it is not the only one (for example, see Banerjee et al. (2014); Bell and Bockstael (2000); Earnest et al. (2007); Getis and Aldstadt (2008); Griffith (1996) and Zhang (2012)). Although $\mathbf{W}$ formalises the spatial dependence between observations (Anselin 1988; Zhang 2012), it is misleading to refer to $\mathbf{W}$ as the 'spatial dependence matrix' because the connection between $\mathbf{W}$ and the resulting spatial correlations is not obvious (Assunção and Krainski 2009; Wall 2004). Consequently, the task of determining appropriate weights so that they represent the spatial correlation is not straightforward, and this remains an ongoing area of research (Johnson 2004). This is the focus of Chapter 6.

2.1.2 Mixture Models

Consider the stochastic process $Z_1, \ldots, Z_L$. Suppose that this process is not directly observable, but that each variable in this process gives rise to a second process, $Y_1, \ldots, Y_L$, which is observable. Since the states $\{Z_l\}$ are not observed, they are said to be hidden, or latent. If the process $Z_1, \ldots, Z_L$ also possesses the Markov property, then the two processes $Z_1, \ldots, Z_L$ and $Y_1, \ldots, Y_L$ are known collectively as a hidden Markov process. (It is usually assumed that the observations $\{Y_l\}$ are conditionally independent given the hidden states $\{Z_l\}$.)

A hidden Markov model (HMM) generates a sequence of states from a hidden Markov process, and each state emits a value according to emission probabilities. The sequence of states forms a Markov chain, and transition probabilities determine the probability of changing between states. The possible values that the observed random variables $Y_1, \ldots, Y_L$ may take are defined by emission probability distributions, which are potentially different for each state. Therefore, the probability of an observed sequence is the product of the state transition probabilities and the emission probabilities (Baum and Eagon 1967; Baum and Petrie 1966; Rabiner 1989; Robert et al. 1993; Rydén and Titterington 1998). An example of simple HMM consisting of two states is shown in Figure 2.3.

A mixture model is a special case of a HMM in which the hidden states are not related by a Markov process, but are generated independently. Consequently, there are no transition

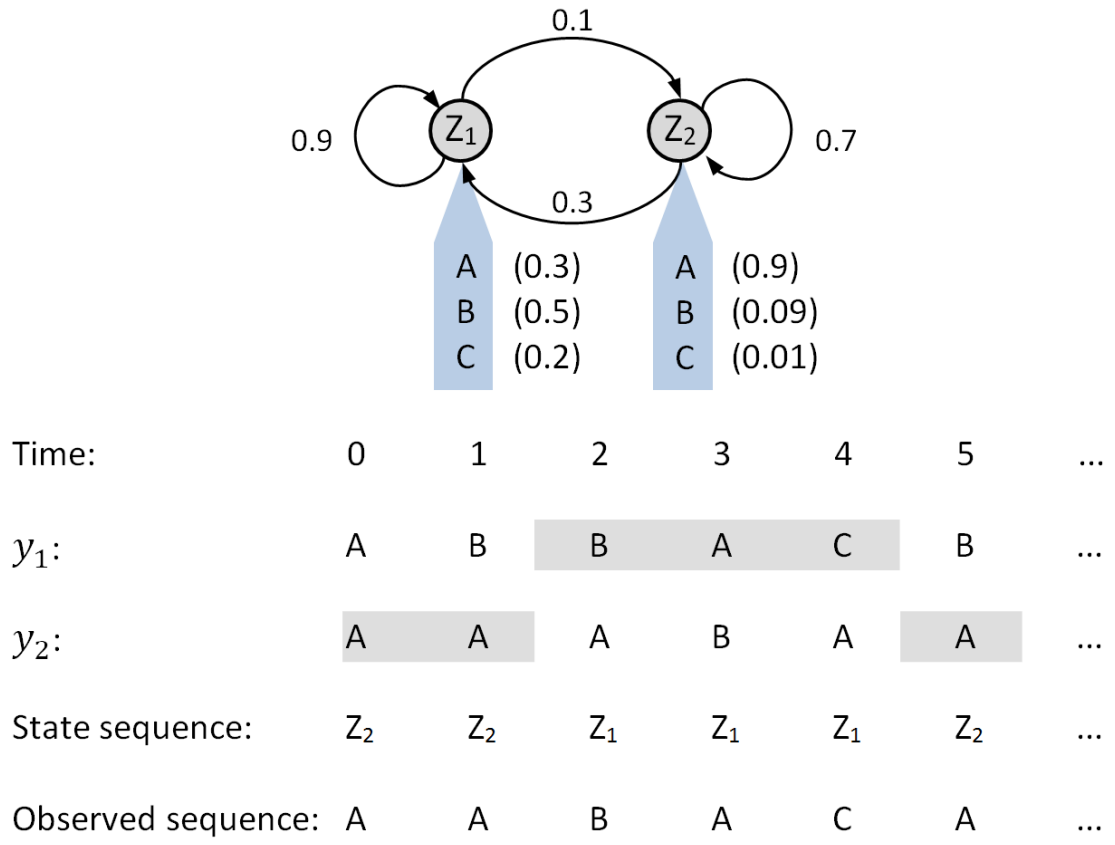


Figure 2.3: An example of a simple, discrete, two-state HMM. State transitions and their associated probabilities are indicated by arrows, and the emissions are shown below the states with the associated emission probabilities given in parenthesis.

probabilities; all states are considered to be 'active' simultaneously throughout the generating process. The output of interest is not one possible sequence, but rather all sequences generated for each state. In fact, multiple observations may be observed for a given state (Robert et al. 1993; Rydén and Titterington 1998).

The states in a mixture model are typically referred to as subpopulations, which collectively comprise a heterogeneous population, and the emission probabilities are usually summarised by a probability density function. Since the subpopulations are hidden, the membership of each observation to a given subpopulation, the parameters characterising the probability densities of each subpopulation, and even the number of subpopulations must be estimated (Marin and Robert 2007; Marin et al. 2005; Rasmussen 2000; Richardson and Green 1997). A mixture model attempts to determine this missing information by estimating the probability density functions such that the resulting subpopulations are consistent with the observed values. If the probability densities are similar, then it may be difficult to determine to which of the subpopulations an observation belongs. A mixture model is not compelled to classify the membership

of each observation to only one subpopulation, but rather it estimates the probability of each observation belonging to each subpopulation.

Suppose a total of N values are observed which are assumed to have been generated from L subpopulations, where $N > L$. Given observed data $\boldsymbol{y} = (y_1, \ldots, y_N)$, the mixture model represents a probability density function for the random variable Y, given by

$$\boldsymbol{Y} \sim p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\phi}) = \prod_{i=1}^{N} \sum_{l=1}^{L} w_l f_l(y_i|\boldsymbol{\phi}_l)$$

where $\boldsymbol{\phi}_l$ denotes the parameters defining the density for the l^{th} subpopulation, $f_l(\cdot)$, and $\boldsymbol{w} = (w_l, \ldots, w_L)$ are the mixture weights representing the probability of membership to the l^{th} subpopulation. The true number of subpopulations may be less than L, in which case some subpopulations may become 'empty' (Celeux et al. 2000; Marin and Robert 2007; Marin et al. 2005; Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014; Stephens 2000b). Although the hidden states $\{Z_l\}$ are not observed, it is customary to treat $\boldsymbol{Z} = (Z_1, \ldots, Z_L)^{\text{T}}$ as a random variable with realisations z_i such that

$$Y_i|z_i, \boldsymbol{\phi} \sim f_{z_i}\left(y_i|\boldsymbol{\phi}_{z_i}\right).$$

The flexibility of mixture models makes them particularly useful in Bayesian analyses of heterogeneous data. They permit a flexible prior specification, and provide a means of simultaneously estimating parameters and classifying data according to those estimates (Abellan et al. 2008; Richardson and Green 1997). The theory of mixture models is discussed in greater depth in Chapter 5.

2.2 Spatial Modelling

Spatial analysis is concerned with the analysis of data arising from a spatial random field or process. Using the terminology introduced in Section 2.1, if $\{X_l = X(t_l)\}$ for $l = 1, \ldots, L$ denotes a spatial random field, then the elements $t_l = \{t_{l,1}, t_{l,2}, \ldots, t_{l,n}\}$ represent spatial coordinates in an n-dimensional space. In many applications, especially in epidemiology and ecology, these spatial coordinates will be geographical coordinates, such as latitude and longitude, and may represent point locations or areas (Assunção and Krainski 2009; Wall 2004).

Observations arising from a spatial random field are usually autocorrelated and/or clustered, which are the result of structured spatial dependencies. In practice, the distinction between the two is not readily discernible, and the term 'autocorrelation' is used collectively to communicate the notion that observations at nearby locations tend to have similar values (Assunção and Krainski 2009; Bell and Bockstael 2000; Florax and Nijkamp 2003; Wall 2004). Ignoring autocorrelation amongst spatial data is analogous to ignoring the order of time-series data (Bell and Bockstael 2000). Consequently, fitting a standard regression model to spatial data will most likely result in the error terms being spatially autocorrelated, which violates one of the assumptions of such models. Practically, this means that the variance of the estimates will be inflated, which makes inference more difficult (Griffith 1996). Therefore, when constructing a statistical model for spatial data, it is necessary to account for spatial autocorrelation in some way (Wall 2004). The CAR model discussed in Section 2.1.1 is one example. Alternative models can be found in Congdon (2010); Diniz-Filho et al. (2009); Dormann et al. (2007).

There are two different ways to model the spatial random field. The first is the geostatistical approach which assumes a continuous-space random field where the observations are only available at discrete points $1, \ldots, L$ (Cressie 1993; Wall 2004). The second way assumes a discrete-space Markov random field, restricting the spatial dependency between locations by defining a neighbourhood structure based on the shape, size, or other features of the lattice (Wall 2004). In both cases, the neighbourhood structure can be summarised by a spatial weights matrix, as described in Section 2.1.1.

While it is clear that it is important to account for spatial autocorrelation, it may not always be apparent how best to capture it. A useful starting point is Tobler's first law of geography which states:

> "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970).

This may suggest measuring spatial dependency as a function of distance, perhaps using geostatistical approaches (Cressie 1993; Diggle 2013). However, geographical distance is not the only measure of distance between two points in a geographical space. This idea is explored in Chapter 6. The introduction of MRFs has led to some very useful models, such as the ICAR model, which offer advantages in simplicity an efficiency of parameter estimation over

geostatistical approaches (Allcroft and Glasbey 2003; Clifford 1990). Both the geostatistical and MRF approaches to incorporating spatial autocorrelation leads to the notion of spatial smoothing, which is discussed below in Section 2.2.4.

2.2.1 Spatio-temporal Modelling

Spatio-temporal data add another layer of complexity to the statistical analysis. Such data can be considered realisations from a random field where the parameter space consists of both spatial coordinates and time (Lekdee and Ingrisawang 2013; Vanmarcke 2010). Clearly the issue of autocorrelation is relevant to space and time, but how it should be accounted for in a spatio-temporal model is not straightforward (Waller et al. 1997). One approach is to account for the spatial autocorrelation across the areas, and the temporal autocorrelation within each area (Lekdee and Ingrisawang 2013). Another approach is to account for the overall spatial and temporal autocorrelation separately, for example, by applying separate CAR prior distributions to the random effects for space and time (Abellan et al. 2008; Duncan et al. 2016; Li et al. 2012). Another approach is to account for the spatio-temporal autocorrelation jointly, for example, by fitting a single CAR model to the data or spatio-temporal random effects (Allcroft and Glasbey 2003) . This last approach reflects the assumption about the underlying process being spatio-temporal most accurately, and might be considered the most appropriate if the bond between space and time is very strong.

2.2.2 Disease Mapping

As the name suggests, disease mapping is the practice of producing maps which convey spatial patterns of some variable relating to a particular disease, such as incidence, mortality, or relative risk. These maps usually take the form of a choropleth map or some other thematic map and serve to visualise and communicate spatial relationships. Traditionally, disease maps made poor analytical tools, as the maps simply connected observed data with geography, and made no attempt to account for varying population density, risk factors, confounding variables, etc. (Banerjee et al. 2014; Barrett 2000; Besag et al. 1991; Bithell 2000; Lee 2011; Rue and Held 2005). Nonetheless, early pioneers in mapping were able to make useful inferences from their maps. A famous example is that of physician John Snow, who used a map of the local area around Broad Street in London to plot the number of cholera-related deaths during the

Soho epidemic of 1854. By mapping the distribution of cases, Snow was able to connect the concentration of cholera cases with contaminated water supplied from the Broad Street pump and convince the council to disable it, thus ending the epidemic (Aguirre, J. C. 2014; Bithell 2000; Shaddick and Zidek 2016; Snow 1855).

In modern times, disease maps continue to play an important role in visualising and communicating patterns of disease-related variables. The development of sophisticated statistical models and advances in technology, particularly software which is capable of both fitting statistical models and rendering thematic maps efficiently, means that disease maps can be used to display more meaningful values, thereby providing another way of conveying information to a wide audience (Assunção and Krainski 2009; Bithell 2000; MacNab 2003). For example, rather than display observed values, disease maps can be used to show patterns in excess relative risk that have been obtained from a statistical model which accounts for population density, risk factors, etc. Because of all these confounding layers, a disease map of the observed values may show no spatial patterns, while a map of the estimated relative risks using the same data may show very clear spatial patterns. This point is reiterated in the discussion of spatial smoothing in Section 2.2.4.

Spatio-temporal data complicate the task of disease mapping through the introduction of a temporal dimension. This not only raises the possibility of space-time interactions, but also makes it difficult to present the data visually as there are now three dimensions – two for space and one for time (Bithell 2000). Examples of disease mapping in recent spatial and spatio-temporal analyses can be found in Abellan et al. (2008), Bernardinelli et al. (1995), Knorr-Held and Besag (1998), Lekdee and Ingrisawang (2013), Li et al. (2012), Pascutto et al. (2000), Waller et al. (1997), Waller et al. (2007), and Xia and Carlin (1998).

2.2.3 Statistical Models

Consider a spatio-temporal random field which generates the random variables $\{Y_{it}\}$ with realisations $\{y_{it}\}$, where $i = 1, \ldots, N$ denotes the area and $t = 1, \ldots, T$ denotes time. A common approach to modelling such data is to assume that the underlying distribution of each random variable Y_{it} belongs to the exponential family, and specify a generalised linear model for the data as follows:

$$g[\mathrm{E}(Y_{it})] = g(\mu_{it}) = \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta}$$

where g is a function linking the expectation of Y_{it} to the covariates $\boldsymbol{x}_{it}^{\mathrm{T}}$, and $\boldsymbol{\beta}$ is the effect of these covariates (Breslow and Clayton 1993; Lee 2013; Schall 1991; Waclawiw and Liang 1993; Zeger and Karim 1991). The inclusion of random effects for space, time, and the interaction between space and time in an additive manner leads to a generalised linear mixed model (GLMM),

$$g(\mu_{it}) = \boldsymbol{x}_{it}^{\mathrm{T}}\boldsymbol{\beta} + S_i + U_t + V_{it}, \tag{2.9}$$

where S_i, U_t, and V_{it} denote the spatial, temporal, and space-time interaction random effects respectively (Lee 2013; Lekdee and Ingrisawang 2013; Schall 1991). GLMMs generalise the methodology of standard linear models by allowing for non-Gaussian response variables, non-linear link functions between the mean of the response and the predictors, and the inclusion of both random and fixed effects simultaneously (Breslow and Clayton 1993; McCulloch 1999; Nelder and Wedderburn 1972; Schall 1991; Zeger and Karim 1991).

Spatial autocorrelation can be incorporated into such a model by specifying the spatial random effect S_i appropriately (Lekdee and Ingrisawang 2013). Likewise for the temporal and interaction effects. In the Bayesian framework, a common way of addressing heterogeneity is to specify an intrinsic CAR prior for each of the random effects, as outlined in Section 2.1.1 (Abellan et al. 2008; Knorr-Held and Besag 1998; Lekdee and Ingrisawang 2013; Li et al. 2012).

In the context of disease mapping, the observations y_{it} often take the form of discrete counts. If the rate of occurrence of events giving rise to y_{it} is rare, such as in the case of rare diseases, then it is appropriate and very common to specify a Poisson likelihood for the count data,

$$Y_{it} \sim \mathrm{Po}(\mu_{it})$$

where the link function $g(\cdot)$ is the natural logarithm (Banerjee et al. 2014; Besag et al. 1991; Pascutto et al. 2000). For less rare events, a binomial likelihood may be more appropriate (for example, see Abellan et al. (2008) and Knorr-Held and Besag (1998)) where $g(\cdot)$ is the logit function,

$$\mathrm{logit}\,(\mu_{it}) = \frac{\mu_{it}}{1-\mu_{it}}.$$

The Poisson distribution implies some strong assumptions about the data. The mean and variance of a Poisson random variable are assumed to be equal. If the variance is considerably larger than the expected value, known as overdispersion, then the negative binomial likelihood is

an appropriate alternative which accounts for overdispersion by including an additional model parameter (Gardner et al. 1995). In a Bayesian model, overdispersion can also be accounted for by specifying an appropriate prior distribution on one or more of the random effects. For example, to allow for spatial overdispersion, a heavy-tailed or otherwise vague prior distribution for S_i may account for areas with unusually large variances (Pascutto et al. 2000; Schall 1991). Additionally, in the case of very rare events, the observed data may contain a large number of zeros. Modified versions of the Poisson model regression model have been proposed to deal with this (Heilbron 1994; Lambert 1992). The zero-inflated Poisson model proposed by Lambert (1992) assumes that the data are generated from a mixture of two processes, where one process is responsible for the excess zero counts. The likelihood is a mixture of a Poisson distribution and a degenerate distribution at zero,

$$p(y_{it}) = \begin{cases} w + (1-w)e^{-\mu_{it}} & \text{if } y = 0 \\ (1-w)\frac{\mu_{it}^{y_{it}} e^{-\mu_{it}}}{y_{it}!} & \text{if } y = 1, 2, \ldots \end{cases}$$

where w is the probability of excess zeros (Agarwal et al. 2002; Lambert 1992). Zero-inflated binomial (Hall 2000; Vieira et al. 2000) and negative binomial (Greene 1994) models have also been developed.

If data related to the response variable is available, this may be included as a covariate in the linear predictor, as shown in Equation (2.9). Sometimes it is important to include variables to control for population size or individual characteristics such as age and gender, which would otherwise cause the model to yield misleading estimates. Pascutto et al. (2000) show how such variables can be used to calculate expected values which can then be included in the model as an offset variable. For example, suppose in a spatial setting, the observed values are stratified by area and age, denoted y_{ij} for $i = 1, \ldots, N$ areas and $j = 1, \ldots, J$ age groups. To account for differences in age groups without modelling the area-age variable, the observed data are simply summed over each age group $y_i = \sum_j y_{ij}$ and the expected counts for area i are computed as

$$E_i = \sum_j P_{ij} \frac{\sum_i y_{ij}}{\sum_i P_{ij}}$$

where P_{ij} is the population of area i for age group j. This process is known as internal standardisation (Pascutto et al. 2000). The offset variable E_i is then included in the model

as

$$Y_i \sim \text{Po}(E_i \mu_i)$$

which is equivalent to including $\log(E_i)$ as an additive fixed effect in the linear predictor (Banerjee et al. 2014; Bernardinelli et al. 1997; Best et al. 2001; Fahrmeir and Kneib 2011; Lekdee and Ingrisawang 2013).

The statistical modelling framework outlined above, especially the Poisson GLMM, can be found in many recent papers dealing with spatio-temporal disease mapping. Notable papers include Bernardinelli et al. (1995), Bernardinelli et al. (1997), Best et al. (2005),Knorr-Held and Besag (1998), Lekdee and Ingrisawang (2013), Pascutto et al. (2000), Waller et al. (1997), and Xia and Carlin (1998). In summary, the generic statistical model for spatial or spatio-temporal disease mapping studies in a Bayesian framework can be formulated as a three-tier hierarchical model. The first tier is the likelihood, for example

$$Y_{it} \sim \text{Po}(E_{it} \mu_{it}).$$

The second tier is the linear predictor component of the GLMM given by Equation (2.9). The third tier consists of the prior distributions for each of the unknown parameters,

$$\begin{aligned} S_i &\sim p(\cdot|\boldsymbol{\theta}_S) \\ U_t &\sim p(\cdot|\boldsymbol{\theta}_U) \\ V_{it} &\sim p(\cdot|\boldsymbol{\theta}_V) \\ \boldsymbol{\beta} &\sim p(\cdot|\boldsymbol{\theta}_\beta) \end{aligned}$$

which are usually weakly informative in the absence of prior information, such as Gaussian distributions with zero mean and some large variance, or in the case of the random effects, may represent CAR priors to account for smoothing (Banerjee et al. 2014; Bernardinelli et al. 1995; Best et al. 2005; Hooten and Wikle 2008; Pascutto et al. 2000; Richardson 2003). The BYM model in Section 2.1.1 is often formulated in this three-tier hierarchical model approach (see for example Banerjee et al. (2014); Best et al. (2001, 2005); Kelsall and Wakefield (2002), and Pascutto et al. (2000)).

The ICAR model is overwhelmingly popular as a prior distribution for random effects in Bayesian

models, especially disease mapping studies (Abellan et al. 2008; Assunção and Krainski 2009; Banerjee et al. 2014; Cressie 1993; Li et al. 2012; Wall 2004). There are a few reasons for this. First, although the implied correlation structure is not obvious from examining the spatial weights matrix $\mathbf{W}$, the conditional neighbourhood structure is intuitive and the specification of the precision matrix is easily justified (Assunção and Krainski 2009). The ICAR prior for spatial random effects, for example, implies that the conditional expectation is the weighted average of the neighbouring values, where the spatial similarity between neighbours is governed by $\mathbf{W}$ (Conlon and Waller 1999). Second, ICAR priors are well suited for the task of smoothing the underlying risk in Bayesian disease mapping. If the neighbourhood structure defines a MRF, then the ICAR prior induces local smoothing by borrowing information from the neighbours (Assunção and Krainski 2009; MacNab 2003). Last, the convenient form of the full conditional distributions for any of the CAR priors facilitates their use in generalised linear models and MCMC samplers (Allcroft and Glasbey 2003; Banerjee et al. 2014).

2.2.4 Smoothing

Earlier in Section 2.2, autocorrelation was discussed in the context of random fields, and the CAR model was described in Section 2.1.1 which is commonly used to address the issue of autocorrelation, especially spatial autocorrelation. Smoothing is a by-product of using models such as the CAR model where the expectation of the variable at location i is smoothed towards the weighted mean of the values at the neighbouring locations (Banerjee et al. 2014; Best et al. 2005; Conlon and Waller 1999; Knorr-Held and Besag 1998; Waller et al. 1997). If Tobler's first law of geography is accurate, then smoothing should be desirable. Smoothing is particularly useful in the construction of maps where the data to be depicted would otherwise be noisy (Best et al. 2005; Johnson 2004; Knorr-Held and Besag 1998; Waller et al. 1997). The degree of spatial smoothing is determined, in part, by the spatial weights matrix $\mathbf{W}$, but the relationship is not obvious in the same way that the relationship between $\mathbf{W}$ and the spatial autocorrelation is unclear (Assunção and Krainski 2009; Banerjee et al. 2014; Lee 2011; Wall 2004). It is perhaps for this reason that $\mathbf{W}$, in the case of MRF models, is typically specified according to first-order adjacency, as defined in Equation (2.4).

The literature suggests that the correlation structure can be approximated by only a small number of neighbours (Allcroft and Glasbey 2003; Griffith 1996). This makes the MRF approach

an attractive option for parameter estimation, both in terms of parsimony and computational efficiency (Allcroft and Glasbey 2003). Empirical methods of smoothing are also possible (Boyle et al. 1989). However, these are not given any serious consideration in this thesis.

2.3 Bayesian Inference

For a generic vector of observations $\boldsymbol{y}$, the likelihood[3] can be expressed in general terms as

$$p(\boldsymbol{y}|\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ denotes all the parameters (and hyper-parameters from all tiers of the hierarchy) which are considered random variables in the Bayesian paradigm, and the goal is to estimate these parameters. Unlike classical statistical inference which summarises $\boldsymbol{\theta}$ with a point estimate, Bayesian inference provides a distribution of plausible values for the parameters given the observed data. This distribution is called the posterior distribution, and is related to the likelihood through Bayes' theorem (Banerjee et al. 2014; Bayes and Price 1763; Christensen et al. 2011):

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{2.10}$$

where $p(\boldsymbol{\theta})$ is the prior distribution of the parameters, and $p(\boldsymbol{y})$ is the marginal distribution of $\boldsymbol{y}$, given by

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{2.11}$$

Since $p(\boldsymbol{y})$ does not depend on the unobservable model parameters, it may be regarded as a constant in the computation of Equation (2.10). Therefore, it is usually sufficient to determine the posterior distribution up to a constant of proportionality,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.12}$$

The constant $p(\boldsymbol{y})$ is often referred to as the normalising constant, and thus Equation (2.12) is termed the unnormalised posterior distribution (Besag et al. 1991; Wikle et al. 2001). Almost

[3]More accurately, $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the sampling distribution of $\boldsymbol{y}$ given $\boldsymbol{\theta}$, and the likelihood $L(\boldsymbol{\theta}|\boldsymbol{y})$ is a function of $\boldsymbol{\theta}$ for fixed $\boldsymbol{y}$, but since $L(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})$, this distinction is unimportant in the derivation of the unnormalised posterior distribution. Consequently, $p(\boldsymbol{y}|\boldsymbol{\theta})$ is often referred to as the likelihood function.

always in practice, the posterior is intractable due to the complex, hierarchical construction of the model, meaning that the right-hand side of Equation (2.12) has no obvious recognisable form which facilitates the integration in Equation (2.11) to determine $p(\boldsymbol{y})$. Thus it is a common approach to estimate the posterior distribution using simulation techniques which are discussed in more detail in Section 2.3.2.

2.3.1 Prior distributions

The full specification of a Bayesian model requires a prior distribution for the parameters. Due to the hierarchical nature of the models, this prior distribution can be factorised into marginal prior distributions, typically one for each parameter, but may still include joint specification of some parameters, such as a CAR prior. This simplifies the task of deciding on an appropriate prior distribution and the process of simulating from the posterior. Nonetheless, choosing an appropriate prior distribution can be challenging.

One of the challenges is choosing prior distributions for variance parameters. Variance is by definition positive (and in most practical applications, strictly positive), which restricts the choice of distributions to those with a semi-infinite domain. Gelman (2006) offers some guidance on the choice of 'weakly-informative' priors for variance parameters and discusses associated caveats. Examples include the gamma and inverse gamma distributions, the uniform distribution on some positive interval, and truncated versions of the Gaussian and t-distributions.

The 'informativeness' of a prior distribution essentially refers to the variance of that distribution, and therefore how influential the distribution will be on the estimation of the posterior distribution. A distribution with a small variance means that the majority of the prior probability will be concentrated on a small domain. Such a prior is said to be informative as the posterior probability will be greatly influenced by the prior probability, even more so if the data consists of few observations. Prior distributions with large variances are sometimes referred to as 'vague', but even a flat distribution with infinite variance conveys some information about the researcher's prior beliefs regarding the plausibility of the parameter values. Hence Gelman (2006); Richardson and Green (1997) and others prefer to use the term 'weakly-informative'. In the case of mixture models, independent non-informative prior distributions are not possible (Richardson and Green 1997).

The subjective nature of specifying the prior distribution is sometimes criticised as anti-scientific.

However, formalising prior information in the form of a distribution is perfectly consistent with scientific practice. Both informative and weakly-informative prior distributions may be reasonable choices, and a reasonable prior ought to yield a sensible posterior (Christensen et al. 2011; Gelman 2006). What is important is that the researcher is able to justify the choice of the priors. Reasons for using the CAR prior, for example, will be obvious from the scientific context, but in this case justification should extend to the specification of the weight matrix defining the smoothing properties. This is the subject of Chapter 6.

2.3.2 Markov Chain Monte Carlo Sampling

Assuming conditional independence between the parameters, the posterior distribution in Equation (2.12) can be rewritten as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\theta_1|\boldsymbol{\theta}_{\setminus 1}) \dots p(\theta_H|\boldsymbol{\theta}_{\setminus H})$$

where H is the total number of parameters in the model (including the hyper-parameters). To make inferences about each of the parameters, it is often easier to consider the marginal posterior distributions. The marginal posterior distribution for the h^{th} parameter is found by integrating out the remaining parameters:

$$\begin{aligned} p(\theta_h|\boldsymbol{y}) &= \int \cdots \int p(\theta_1, \dots, \theta_h, \dots, \theta_H|\boldsymbol{y}) d\boldsymbol{\theta}_{\setminus h} \\ &= \int \cdots \int p(\theta_h|\boldsymbol{\theta}_{\setminus h}, \boldsymbol{y}) p(\boldsymbol{\theta}_{\setminus h}|\boldsymbol{y}) d\boldsymbol{\theta}_{\setminus h}. \end{aligned}$$

In practice, the integration necessary to compute $p(\theta_h|\boldsymbol{y})$ is often impossible (Banerjee et al. 2014; Christensen et al. 2011; Rossi et al. 2005). However, the theory of Markov processes from Section 2.1 offers a solution which avoids the need for any integration. By the careful construction of a Markov chain starting from initial values for each parameter, the posterior distribution can be estimated using Markov Chain Monte Carlo (MCMC) simulation (Christensen et al. 2011; Gilks et al. 1996). The exact definition of a Markov chain varies (Asmussen 2003), but in the context here, a Markov chain is taken to be a discrete-space Markov process.

The terms $p(\theta_h|\boldsymbol{\theta}_{\setminus h}, \boldsymbol{y})$ for $h = 1, \dots, H$ are referred to as the full conditional posterior distributions, or simply full conditionals. Gibbs sampling is an MCMC method which involves

sampling from each of the H full conditionals using the most recent estimates of the parameters Banerjee et al. (2014); Besag et al. (1991); Christensen et al. (2011); Rasmussen (2004). Because the samples constitute a Markov chain, the dependence on the initial values are forgotten, and under mild assumptions (Besag 1974; Cressie 1993; Hammersley and Clifford 1971), the chain converges to a stationary distribution which is the unnormalised posterior distribution as desired (Christensen et al. 2011; Geman and Geman 1984). The Gibbs sampler algorithm was proposed by Geman and Geman (1984) and later generalised by Gelfand and Smith (1990). The algorithm is provided below.

Algorithm 2.1: Gibbs sampler

Step 1: Initialise parameters $\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_H^{(0)}$.

Step 2: For $m = 1, \ldots, M$, sample from each of the full conditionals:

$$\begin{aligned}
\theta_1^{(m)} &\sim p\left(\theta_1^{(m-1)} | \boldsymbol{\theta}_{\setminus 1}^{(m-1)}, \boldsymbol{y}\right) \\
\theta_2^{(m)} &\sim p\left(\theta_2^{(m-1)} | \boldsymbol{\theta}_{\setminus 2}^{(m-1)}, \boldsymbol{y}\right) \\
&\vdots \\
\theta_H^{(m)} &\sim p\left(\theta_H^{(m-1)} | \boldsymbol{\theta}_{\setminus H}^{(m-1)}, \boldsymbol{y}\right)
\end{aligned}$$

If the full conditional distributions are recognisable, then sampling from them is trivial. In the case where one or more of the full conditionals do not have a standard, recognisable kernel density, then alternative MCMC techniques such as Metropolis-Hastings (Hastings 1970; Metropolis et al. 1953) or slice sampling (Damien et al. 1999) may be used. MCMC sampling is routinely performed with software such as R (R Core Team 2015) and WinBUGS (Lunn et al. 2000). This software is available on the Internet free of charge and both have features which make it very convenient to perform all aspects of a fully Bayesian analysis (Christensen et al. 2011; Hoeting 2009; Rue and Held 2005). As a result, MCMC methods have become more accessible to the wider research community. Examples of MCMC estimation in recent spatio-temporal analyses can be found in Abellan et al. (2008); Best et al. (2005); Lekdee and Ingrisawang (2013); Li et al. (2012); Wikle et al. (2001), and Duncan et al. (2016).

2.3.3 Advantages and Disadvantages of Bayesian Methods

From the complex nature of hierarchical models to the knowledge required to simulate from the posterior distribution, Bayesian methods can appear daunting. However, Bayesian methods also offer many advantages over classical inference. This is especially true for spatial and spatio-temporal models.

First, spatial data are, by nature, autocorrelated. In order to account for autocorrelation, it is necessary to use some sort of hierarchical structure. Bayesian models are inherently hierarchical and thus accounting for autocorrelation within the Bayesian framework is straightforward (Best et al. 2005). This includes the use of the CAR models (Besag 1974; Besag et al. 1991; Cressie 1993; Leroux et al. 1999) as MRF prior distributions for Bayesian inference. The complicated structure of hierarchical models often results in the posterior being intractable, but MCMC sampling techniques overcome this problem (Banerjee et al. 2014; Christensen et al. 2011).

Second, Bayesian hierarchical models are very flexible, accommodating for the uncertainty in estimated random and fixed effects, a priori knowledge through the specification of priors and hyper-priors, and even the sampling design (Banerjee et al. 2014; Breslow and Clayton 1993; Gelman 2006; Hoeting 2009; Legler et al. 2002). Despite accounting for heterogeneity in spatial data, non-Bayesian methods can lead to underestimation of uncertainty in the model parameters (Hoeting 2009).

Last, Bayesian analysis provides better estimates in a number of ways. Bayesian estimation provides a distribution rather than a point estimate which is more informative. Bayesian estimation avoids 'over-fitting' of the model by integrating over the model parameters (Rasmussen 2004). And estimating parameters in mixture models, especially when the number of mixture components is unknown, is more convenient and accurate in a Bayesian framework (Richardson and Green 1997).

The main disadvantage of Bayesian methods is the lengthy computation time required for the generated Markov chain to converge to the posterior (Banerjee et al. 2014; Hoeting 2009). There are also several difficulties associated with Bayesian modelling, such as assessing convergence and selecting appropriate prior distributions. However, the literature contains much guidance on these issues (for example, see Brooks and Gelman (1998) and Gelman (2006)), and such challenges are arguably less onerous than those associated with non-Bayesian methods.

Statement of Contribution of Co-authors for Chapter 3

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

This paper was published in *BMJ Open* in April 2016. The reference for this publication is: Duncan, E. W., White, N. M., and Mengersen, K. 2016. Bayesian spatiotemporal modelling for identifying unusual and unstable trends in mammography utilisation. *BMJ Open*, **6** (5): e010253. doi: 10.1136/bmjopen-2015-010253.

Contributor	**Statement of contribution**
E. W. Duncan Signature and date:	Wrote the code for data analysis, developed the statistical models, wrote the manuscript, revised the manuscript as suggested by co-authors and reviewers.
N. M. White	Supervised research, provided comments on and helped revise manuscript, aided in development of statistical models and interpretation of results.
K. L. Mengersen	Supervised research, provided comments on and helped revise manuscript, aided in interpretation of results.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

Chapter 3

Bayesian spatio-temporal modeling for identifying unusual and unstable trends in mammography utilisation

---

Preamble

This chapter relates to the first research objective. The chapter commences with a rationale for analysing spatio-temporal patterns in the utilisation of mammography screening services. A literature review ensues, highlighting the main results and shortcomings of past research. The aim of the methodology presented in this chapter is to build upon previous research in this field by applying two spatio-temporal models to the data. These models are specifically designed to identify areas with 'unusual' or 'unstable' temporal trends.

The analysis demonstrates the usefulness of the two models as tools for analysing spatio-temporal data with the objective of highlighting areas based on the nature of their corresponding temporal trends. The synergy of applying both models to the same data set is also demonstrated, strengthening the statistical inferences. The results reveal several important findings, and also identify a number of ways in which the methodology may be improved. Extensions to these models are the focus of Chapter 4.

This chapter was published as a paper in *BMJ Open* in April 2016. The online supplementary material provided in this paper includes schematic diagrams and WinBUGS code for both models. These figures and code are reproduced in Appendix A.1 through A.4.

3.1 Introduction

Breast cancer is one of the most common types of cancer faced by women in Australia, and the second most common cause of cancer-related deaths, accounting for 2914 deaths in 2011 alone (Australian Institute of Health and Welfare (AIHW) 2014). Evidence suggests that regular mammography screening to be effective at detecting early breast cancer thereby increasing the chance of survival (Alexander et al. 1999; Australian Government Department of Health (AGDH) 2014; Australian Institute of Health and Welfare (AIHW) 2012; Hyndman et al. 2000; Shen et al. 2005). However, recent studies have suggested that mammography screening services may be underutilised due to geographical factors such as the travel distance to a mammography screening facility (Hyndman et al. 2000; Jackson et al. 2009; Zenk et al. 2006).

Hyndman et al. (2000) conducted a study to investigate the effects of distance to a mammography screening facility and social disadvantage on service utilisation. This study found that women with a lower socioeconomic status had more difficulty travelling to a mammography screening facility, and suggested that facilities should be located closer to disadvantaged communities to increase utilisation. However, the influence of the location of mammography screening facility was less clear when socioeconomic factors play no significant role in service utilisation. In another work, Legler et al. (2002) modelled how state mammography screening rates depend on service user demographics, county-level socioeconomic factors, and previous mammography intervention research projects. The study found that states with one or more published intervention studies and states with higher levels of education tended to have higher rates of mammography screening service utilisation. Zenk et al. (2006) assessed the equitability of spatial accessibility to low-cost or no-fee mammography screening services in Chicago by modeling distance and time to travel to a facility as a function of geographic and sociodemographic variables including race and poverty. The study concluded that travel time and distance was generally less for poorer neighbourhoods, except for neighbourhoods with a higher proportion of African-American residents. It is unclear from this study, however, if these barriers to access translated into lower screening utilisation since mammography screening utilisation rates were not considered.

The methodologies used in the above studies vary greatly. Hyndman et al. (2000) used geographic information system (GIS) techniques and screening data from six mammography screening facilities in Perth, Australia. Legler et al. (2002) opted for a hierarchical model

to estimate the effects of education, occupation, and demographic group on mammography screening rates for each state. The model was fit to data collected at two time points, 1987 and 1993-1994, which mark a period during which numerous intervention studies were published. Although the model was applied to data at two time points, it only included spatial covariates, thus limiting inferences to differences between the two fitted models for each time point and spatial effects. Zenk et al. (2006) used ordinary least squares regression models to estimate the effects of the covariates on the accessibility measures. However, the authors found that the residuals exhibited spatial autocorrelation, even when endogenous spatial lag regression was used. Furthermore, travel times and distances were estimated using GIS and other software rather than observed, and required numerous assumptions, adding to the uncertainty and the potential bias of the estimates.

The literature contains many other studies that analyse low-cost or no-fee mammography screening utilisation rates and related data. The focus of these studies has typically been aimed at estimating the effect of one or more variables on screening rates. These variables are usually spatially dependent and include service user demographics, socioeconomic factors, accessibility factors, and variables relating to the spatial units of the study such as the degree of urbanisation (Engelman et al. 2002; Jackson et al. 2009; Makuc et al. 1999; Maxwell 2000; Selvin and Brett 2003). Whilst estimation of the covariate effects on mammography screening utilisation is useful, little attention has been given to identifying trends in screening utilisation rates, especially trends which vary over both space and time. Analysis of such trends permits a wider variety of statistical inferences, with implications for service management. Moreover, ignoring spatial and temporal correlation when present can lead to errors in prediction and inference (Hoeting 2009).

This paper aims to build upon previous research in this field by presenting two spatio-temporal models, applied to no-fee mammography screening facility attendance data in Brisbane, Australia. In short, these models are designed to identify ‘unusual’ or ‘unstable’ temporal patterns. The use of the terms ‘unusual’ and ‘unstable’ are model specific, and their meanings are discussed in further detail in Methods.

By nature, spatio-temporal data can be clustered and/or autocorrelated, and are sometimes sparse whereby some regions exhibit relatively low numbers of observed and expected counts. Spatial models typically account for these issues by encoding neighbourhood information as

part of the wider model. This has the added advantage of reducing estimated risks with high uncertainty towards the mean risk. Bayesian methods naturally incorporate this information using prior distributions and hierarchical model structures, and allow for estimation of a full probability model for the unknown parameters Bernardinelli et al. (1995); Best et al. (2005); Lekdee and Ingrisawang (2013); Zeger and Karim (1991); Zhang (2002).

Both models considered in this paper have in common the specification of spatial and/or temporal random effects, albeit in different forms, and a model indicator as a means of differentiating SLAs that exhibit a common/stable temporal trend as opposed to an unusual/unstable temporal trend. Both models are estimated using Bayesian techniques and overcome the difficulties associated with autocorrelated data by explicitly including the spatial and temporal dependencies in the models.

3.2 Methods

3.2.1 Data

The data used in this study consisted of the number of visits made to mammography screening facilities operated by BreastScreen Queensland in the Brisbane region per year, from 1997 to 2008 inclusive. For each year, the number of visits was recorded by statistical local area (SLA), with 158 SLAs included in the Brisbane region. The eligible population was defined as women aged 40 years or over at the time of screening, in line with the BreastScreen Australia Programme eligibility criteria (Australian Government Department of Health (AGDH) 2014).

The physical location and opening and closing dates of each mammography screening facility were also recorded. Some of the mammography screening facilities were mobile and therefore only available at a specific location for a shorter time period. Using these data, a covariate x_{it} was created, which represents the relative availability of services in a catchment area, defined as

$$x_{it} = \frac{d_{it}}{\max\limits_{i} d_{it}} \tag{3.1}$$

where d_{it} is the cumulative number of days that each mammography screening facility was operating in SLA i or any SLA that shares a border with SLA i, $\{i = 1, \ldots, 157\}$ during year t, $\{t = 1, \ldots, 12\}$. This catchment area service availability for odd years only is depicted

graphically in Figure 3.1.

Socioeconomic status was also considered as a covariate. However, a preliminary analysis of socioeconomic data did not indicate evidence of an effect. For this reason, it was excluded from the final models.

3.2.2 Model Formulation

Two models are proposed for spatio-temporal analysis of data. Both models are examples of Bayesian spatial generalised linear mixed models (GLMMs) that fall within the wider class of linear models (Schall 1991; Waclawiw and Liang 1993; Zeger and Karim 1991).

Let y_{it} denote the observed count of visits to mammography screening facilities in SLA i during year t. Given a population at risk R_{it}, the corresponding expected number of visits is given by

$$E_{it} = r_t R_{it}$$
$$r_t = \frac{\sum_i y_{it}}{\sum_i R_{it}}$$

where r_t is the reference screening rate in year t (Pascutto et al. 2000).

The first model considered was the BaySTDetect model proposed by Li et al. (2012). This model consists of two competing models: a common trend model where the temporal trend is the same for each SLA, and an area-specific model where the temporal trends are allowed to depart from the common trend. The two competing models are hierarchical in structure and are related to the likelihood via a model selection step, given by Equation 3.2. The BaySTDetect model assumes that the Y_{it} counts are a Poisson random variable, for example,

$$Y_{it} \sim \text{Po}\left(E_{it}\mu_{it}\right)$$

where

$$\log\left(\mu_{it}\right) = \begin{cases} \alpha + \eta_i + \gamma_t + \beta x_{it} & \text{if } p_i = 1 \\ u_i + \xi_{it} + \beta' x_{it} & \text{if } p_i = 0 \end{cases} \tag{3.2}$$

$$p_i \sim \text{Bern}\left(\delta_i\right). \tag{3.3}$$

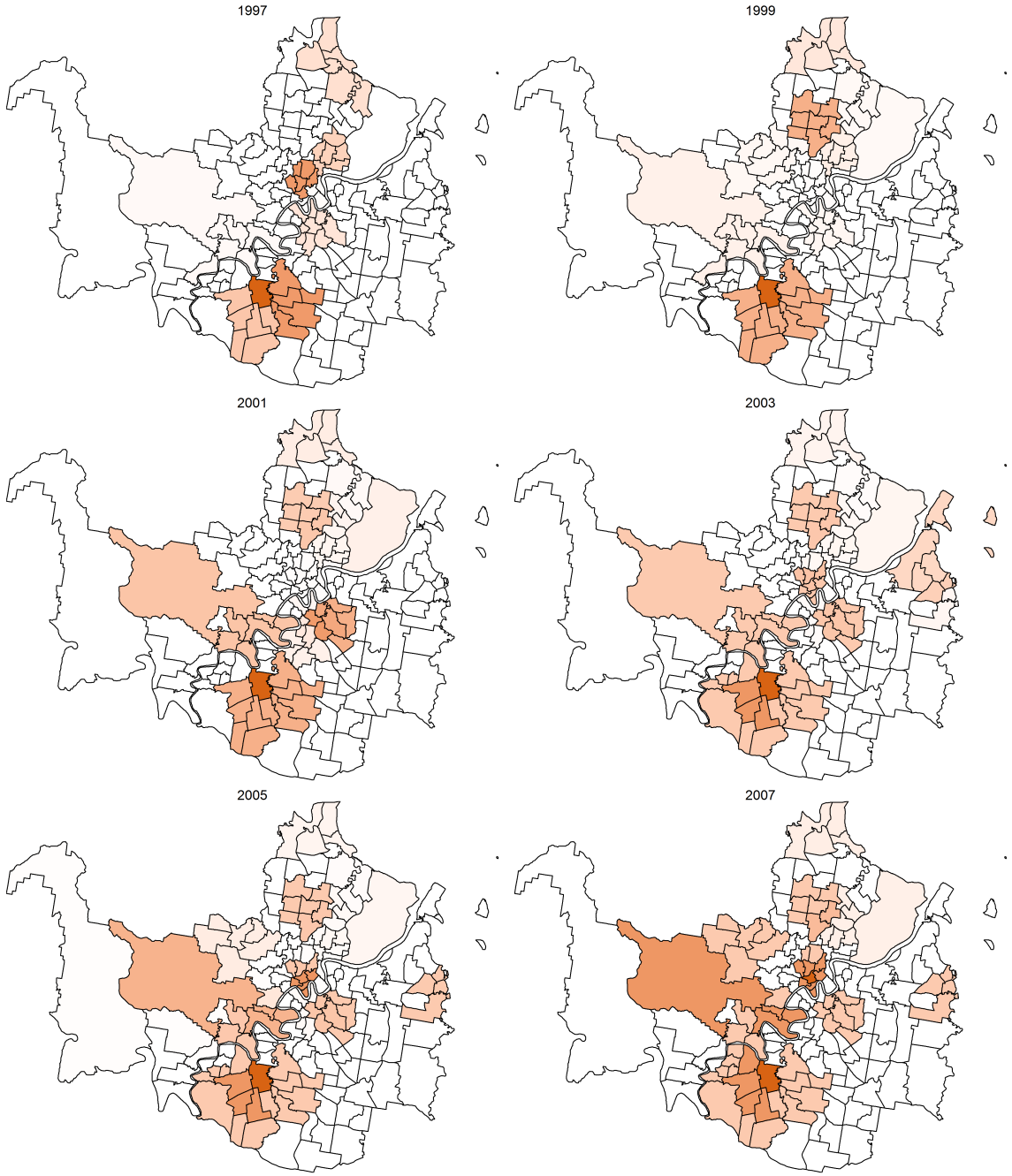


Figure 3.1: Map of the SLAs in the Brisbane region (Moreton Island not shown) depicting the relative availability of mammography screening services based on the operating duration and location of mammography screening facilities in each SLA and neighbouring SLAs over time (odd years shown only), as defined by Equation 3.1.

The components of 3.2 are as follows: α is the common intercept; η_i and γ_t are random effects of space and time respectively; u_i is the area-specific intercept, ξ_{it} is the area-specific random effect and x_{it} is the covariate defined by 3.1. Regarding the prior for p_i, it is expected that temporal trends are fairly homogenous for most SLAs, and thus the hyper-parameter δ_i is chosen to be 0.95. To incorporate spatial and temporal smoothing, intrinsic conditional autoregressive (ICAR) priors (Besag and Kooperberg 1995) are assigned to the random effects η_i, γ_t and ξ_{it}

$$\eta_i|\boldsymbol{\eta}_{\setminus i} \sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}}\sum_{j=1}^{N} w_{ij}\eta_j, \frac{\sigma_\eta^2}{\sum_j w_{ij}}\right), \tag{3.4}$$

$$\gamma_t|\boldsymbol{\gamma}_{\setminus t} \sim \mathcal{N}\left(\frac{1}{\sum_s w'_{st}}\sum_{s=1}^{T} w'_{st}\gamma_s, \frac{\sigma_\gamma^2}{\sum_s w'_{st}}\right), \tag{3.5}$$

$$\xi_{it}|\boldsymbol{\xi}_{i\setminus t} \sim \mathcal{N}\left(\frac{1}{\sum_s w'_{st}}\sum_{s=1}^{T} w'_{st}\xi_{is}, \frac{\sigma_{\xi,i}^2}{\sum_s w'_{st}}\right) \tag{3.6}$$

where $\setminus i$ denotes all areas excluding i, and w_{ij} is the i-j^{th} element of a symmetric spatial adjacency weight matrix $\mathbf{W}$ with elements $w_{ij} = 1$ if the i^{th} and j^{th} areas are neighbours, and zero otherwise. Similarly, $\setminus t$ denotes all years excluding t, and w'_{st} is the s-t^{th} element of a symmetric temporal adjacency weight matrix $\mathbf{W}'$ with elements $w'_{st} = 1$ if the s^{th} and t^{th} years are neighbours, and zero otherwise. Note that the temporal adjacency information is the same for each of the i terms of ξ_{it}.

The parameters α, β, u_i, and β' are assumed to be Normally distributed with mean 0 and variance 1000. The hyper-parameters σ_η and σ_γ were assigned weakly informative half-Normal priors to reflect a lack of prior knowledge about these parameters but restrict their values to be strictly positive and yet not too large (Gelman 2006). The prior for the hyper-parameter $\sigma_{\xi,i}^2$ is log-Normal

$$\log\left(\sigma_{\xi,i}^2\right) \sim \mathcal{N}\left(a, c^2\right)$$

where the variance is given an informative prior relative to the data,

$$c \sim \mathcal{N}\left(0, 2.5^2\right)\mathbb{I}_{(0,+\infty)}$$

to reflect prior expectations about the temporal variability (Li et al. 2012).

The second model considered was based on the mixture model approach proposed by Abellan et al. (2008). This model estimates the common spatial and temporal trends based on the data

and identifies SLAs the residual temporal patterns of which show volatility, that is, are unstable. In this hierarchical model, the counts Y_{it} are modelled as Poisson random variables with mean $E_{it}\pi_{it}$, for example,

$$\begin{aligned} Y_i t &\sim \text{Po}\left(E_{it}\pi_{it}\right), \\ \log\left(\pi_{it}\right) &= \tau + \lambda_i + \psi_t + \nu_{it} + bx_{it}. \end{aligned} \tag{3.7}$$

Here the term τ is the common intercept, λ_i and ψ_t represent random effects for space and time respectively, and ν_{it} represents space-time interaction. Like the BaySTDetect model, the spatial and temporal random effects are modelled jointly using ICAR priors,

$$\lambda_i|\boldsymbol{\lambda}_{\setminus i} \sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}}\sum_{j=1}^{N} w_{ij}\lambda_j, \frac{\sigma_\lambda^2}{\sum_j w_{ij}}\right), \tag{3.8}$$

$$\psi_t|\boldsymbol{\psi}_{\setminus t} \sim \mathcal{N}\left(\frac{1}{\sum_s w'_{st}}\sum_{s=1}^{T} w'_{st}\psi_s, \frac{\sigma_\psi^2}{\sum_s w'_{st}}\right) \tag{3.9}$$

where w_{ij} and w'_{st} are as defined earlier (see Equations 3.4-3.6). Normal priors are defined for the intercept and covariate effect terms,

$$\begin{aligned} \tau &\sim \mathcal{N}\left(0, 1000\right) \\ b &\sim \mathcal{N}\left(0, 1000\right) \end{aligned}$$

while the prior distribution for ν_{it} is described by a mixture of two Normal distributions with different variances, one representing stable patterns and the other unstable patterns:

$$\nu_{it} = q\mathcal{N}\left(0, \sigma_{\nu_1}^2\right) + (1-q)\mathcal{N}\left(0, \sigma_{\nu_2}^2\right)$$

The variance is determined by a latent model indicator variable z_{it}, specified in the model by the multinomial distribution consisting of a single draw,

$$z_{it} \sim \text{Mult}(1, \boldsymbol{q})$$

where the prior for the mixture weights $\boldsymbol{q} = (q, 1-q)$ is a Dirichlet distribution

$$\boldsymbol{q} \sim \mathcal{D}(1,1)$$

The latent indicators take the value 1 if $\nu_i t$ is modelled by $\mathcal{N}\left(0, \sigma^2_{\nu_1}\right)$ or 2 if $\nu_i t$ is modelled by $\mathcal{N}\left(0, \sigma^2_{\nu_2}\right)$, with $\sigma^2_{\nu_1} < \sigma^2_{\nu_2}$. To avoid the issue of label-switching (Redner and Walker 1984; Stephens 2000b), and in line with the model specification, the priors for the two variances are specified as

$$\begin{aligned}
\sigma_{\nu_1} &\sim \mathcal{N}(0, 0.01)\, \mathbb{I}_{(0,+\infty)} \\
\sigma_{\nu_2} &= \sigma_{\nu_1} + \kappa \\
\kappa &\sim \mathcal{N}(0, 100)\, \mathbb{I}_{(0,+\infty)}
\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function $\mathbb{I}_{a,b} = 1$ if $a < \kappa < b$.

By analysing the posterior frequencies of the latent indicator variables z_{it}, this mixture model can be used to identify SLAs with unstable temporal trends. For example, let

$$P_{it} = \Pr(z_{it} = 2 | y_{it}) \tag{3.10}$$

represent the posterior probability that nu_{it} follows $\mathcal{N}\left(0, \sigma^2_{\nu_2}\right)$, that is, the posterior probability that the space-time interaction has a large variance. Thus the closer the P_{it} values are to 1, and the more P_{it} values that are close to 1 for $t = 1, \ldots, 12$, the more unstable the temporal patterns are for the i^{th} SLA. Abellan et al. (2008) propose two rules for classifying SLAs as unstable. The first rule considers the i^{th} SLA to be unstable if $P_{it} > P_{cut}$ for at least one t, where P_{cut} is some specified threshold. The second rule classifies the i^{th} SLA as unstable if the average of the three largest P_{it} values $> P_{cut}$. Rule 2 is slightly more conservative since it averages the P_{it} values over three years. Both of these rules were used.

3.2.3 Comparing the Two Models

While the distinction between unusual and unstable temporal trends may seem trivial, these two models aim to address two very different questions relating to spatio-temporal patterns, and hence each model may provide unique insights.

The BaySTDetect model assumes one common temporal trend, γ_t, across all areas and uses a model choice step to fit a competing model with independent random temporal effects for each area if there is considerable departure from the common trend. This allows identification of SLAs which have an unusual temporal trend. For example, assuming a constant mammography screening utilisation rate on average (the common trend), then SLAs which exhibit a high screening rate one year followed by a low rate the next year would be considered to have an unusual temporal trend and would therefore most likely be modelled by the area-specific model.

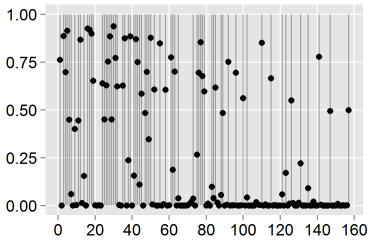
In contrast, the space-time mixture tries to estimate the overall spatio-temporal trend. If the annual screening counts for a given SLA are quite different from that which is predicted by the model, then this apparent departure from the overall spatio-temporal trend suggests that the screening rate for this SLA is unstable.

3.2.4 Implementation

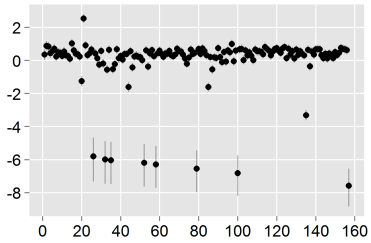
Both models were estimated using Markov Chain Monte Carlo (MCMC) techniques, implemented in WinBUGS through R using the `R2WinBUGS` package (Lunn et al. 2000; R Core Team 2012; Sturtz et al. 2005). The results are based on 25 000 iterations after discarding an initial 100 000 iterations as burn-in. Convergence was assessed informally via visual checks of trace and density plots, as well as formally using the Geweke convergence diagnostic (Geweke 1992).

Initially, both models were implemented with the hierarchical structure and priors as specified by the respective authors as described above. These models were then adapted to our scientific problem of interest through a number of modifications. The main changes to the models involved modelling the spatial and temporal random effects η_i, λ_i, and ψ_t using ICAR priors directly, rather than modelling their respective means, due to a lack of strong autocorrelation between parameters and issues with identifiability of parameters (results not shown). Both models were also extended to include the covariate given by Equation 3.1.

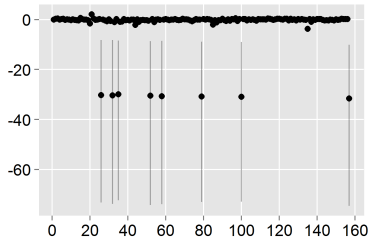
Schematic diagrams of the BaySTDetect and mixture models, after taking into account the changes outlined above, are provided in online supplementary figure S1 and online supplementary figure S2 respectively, available online. The WinBUGS code is also provided, in online supplementary Codes 1 and 2. The posterior distributions of the key model parameters for each model are summarised in Figures 3.2 and 3.3.



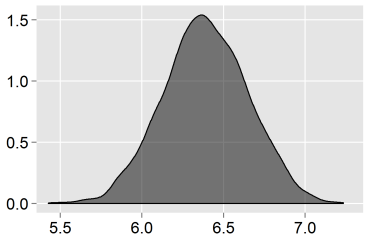
(a) Posterior mean and 95% CI for p_i



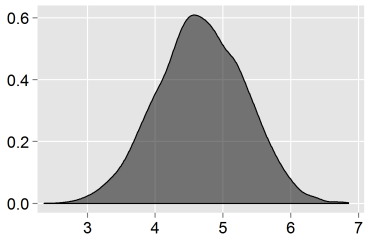
(b) Posterior mean and 95% CI for η_i



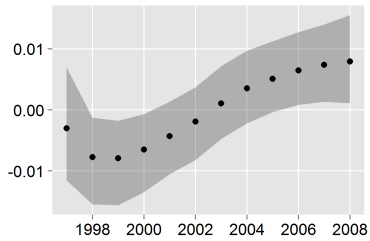
(c) Posterior mean and 95% CI for u_i



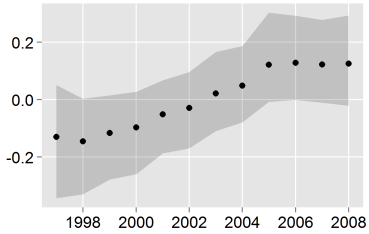
(d) Posterior density for β



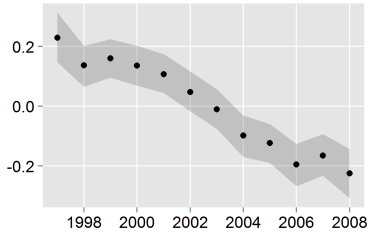
(e) Posterior density for β'



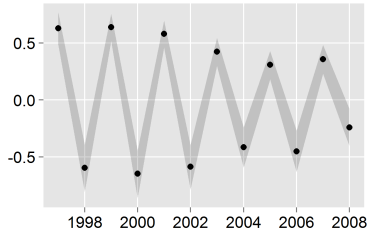
(f) Posterior mean and 95% CI for γ_t



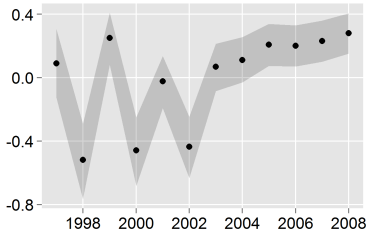
(g) Posterior mean and 95% CI for ξ_{10t}



(h) Posterior mean and 95% CI for ξ_{22t}

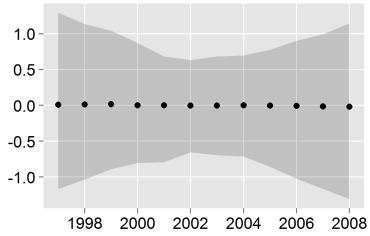


(i) Posterior mean and 95% CI for ξ_{57t}



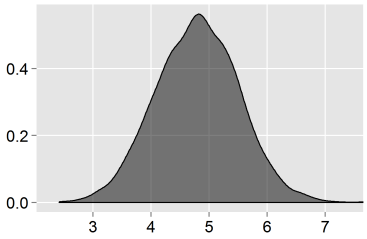
(j) Posterior mean and 95% CI for ξ_{68t}

(k) Posterior mean and 95% CI for ξ_{92t}

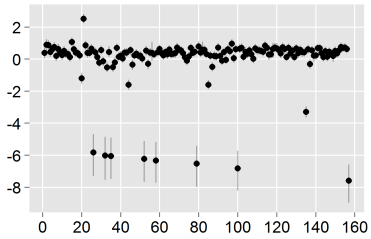


(l) Posterior mean and 95% CI for ξ_{158t}

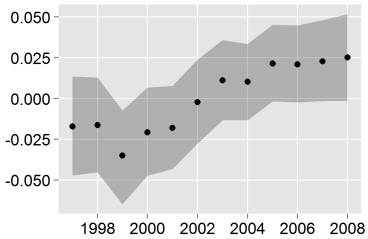
Figure 3.2: Posterior summary of the main model parameters for 1 chain of the final (modified) BaySTDetect model.



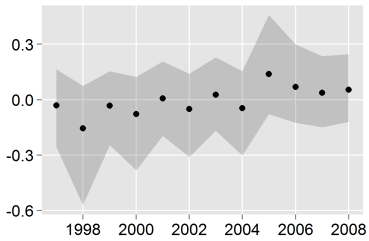
(a) Posterior density for b



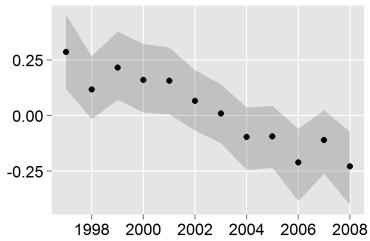
(b) Posterior mean and 95% CI for λ_i



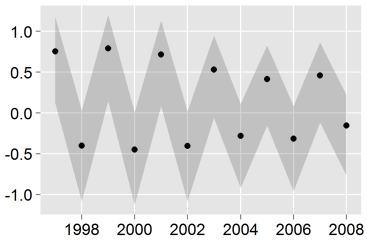
(c) Posterior mean and 95% CI for ψ_t



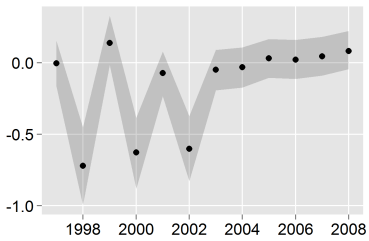
(d) Posterior mean and 95% CI for ν_{10t}



(e) Posterior mean and 95% CI for ν_{22t}

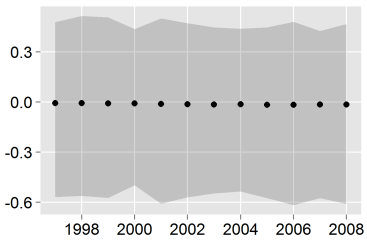


(f) Posterior mean and 95% CI for ν_{57t}

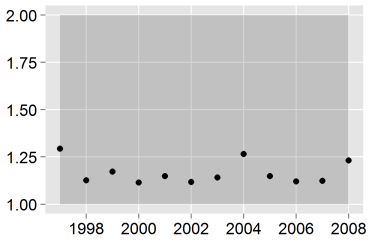


(g) Posterior mean and 95% CI for ν_{68t}

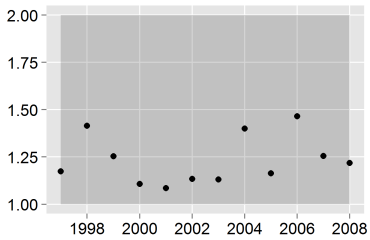
(h) Posterior mean and 95% CI for ν_{92t}



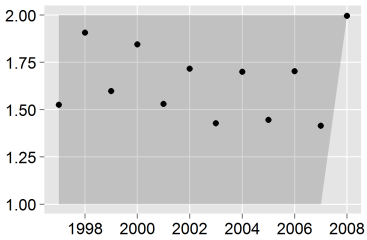
(i) Posterior mean and 95% CI for ν_{158t}



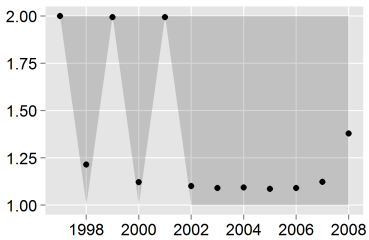
(j) Posterior mean and 95% CI for z_{10t}



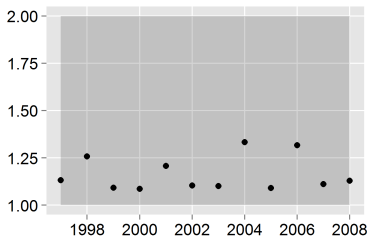
(k) Posterior mean and 95% CI for z_{22t}



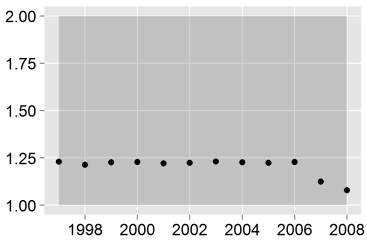
(l) Posterior mean and 95% CI for z_{57t}



(m) Posterior mean and 95% CI for z_{68t}



(n) Posterior mean and 95% CI for z_{92t}



(o) Posterior mean and 95% CI for z_{158t}

Figure 3.3: Posterior summary of the main model parameters for 1 chain of the final (modified) space-time mixture model.

3.2.5 Assessment of Model Fit and Predictive Performance

Posterior predictive checks (PPCs) were performed to assess the goodness-of-fit and predictive performance of the models given by Equations 3.2 and 3.7. In brief, PPCs aim to assess the consistency between predictions from the model and the observed data (Gelman et al. 2013; Ibrahim and Laud 1994; Ntzoufras 2009). If y_{it}^{pred} is a prediction of y_{it} from the specified model, PPCs involve draws from the posterior predictive distribution:

$$p\left(Y_{it}^{\text{pred}}|y_{it}\right) = \int p\left(Y_{it}^{\text{pred}}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|y_{it}\right) d\boldsymbol{\theta}$$

where $\boldsymbol{\theta}$ denotes all the parameters in the model (Ntzoufras 2009). These predictions were formed by sampling 200 times from the joint posterior distribution and using each posterior sample to generate $y_{it}^{\text{pred}(k)}$, $k = 1, \ldots, 200$. The consistency between the predicted and observed counts for each SLA-year was evaluated using the L-criterion (Ibrahim and Laud 1994), defined as the square root of the mean squared prediction error,

$$L_{it} = \sqrt{\text{E}\left(\sum_{k=1}^{200}\left(Y_{it}^{\text{pred}(k)} - y_{it}\right)^2 |y_{it}\right)} \tag{3.11}$$

The estimate $\hat{L}_{it}$ of this quantity is easily computed from the MCMC estimate of the posterior predictive distribution. The results of these PPC are discussed below.

3.3 Results

3.3.1 Predictive Performance of the Models

As a summary of the differences between the predicted and observed counts for each SLA, Figure 3.4 shows the L-criterion estimates $\hat{L}_{it}$ averaged over time for the BaySTDetect model. (The spatial composition of average $\hat{L}_{it}$ values for the mixture model is almost identical and thus omitted). The $\hat{L}_{it}$ values were about 19.21 and 18.70 counts on average for the BaySTDetect and mixture models, respectively, suggesting acceptable and comparable predictive performance. While there were two regions of SLAs with predominantly larger $\hat{L}_{it}$ values (the north and south east), there did not appear to be any correlation between SLAs with larger $\hat{L}_{it}$ values and service availability (compare Figures 3.1 and 3.4).

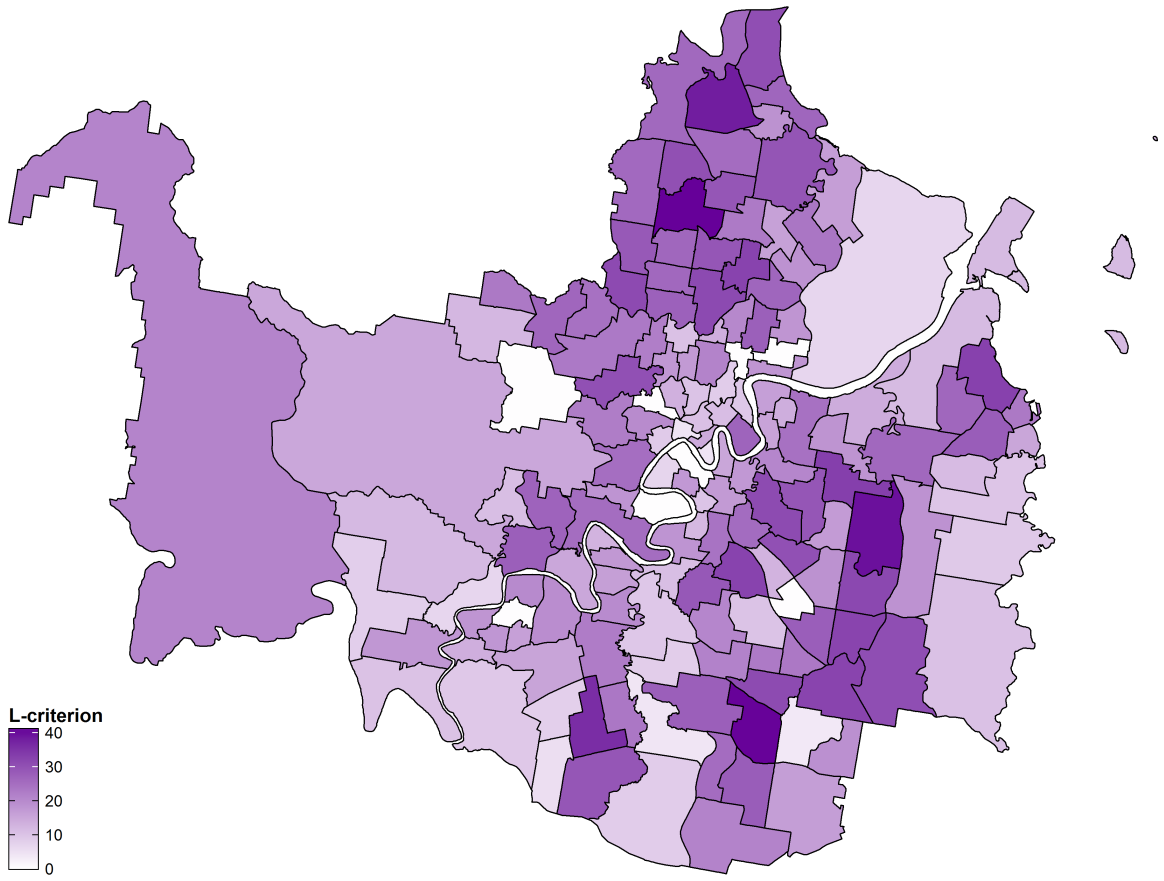


Figure 3.4: Map of SLAs in the Brisbane region (Moreton Island not shown) depicting the closeness between y_{it} and $\mathrm{E}\left(Y_{it}^{\mathrm{pred}}|y_{it}\right)$ for each SLA for the final (modified) BaySTDetect model, as specified by the L-criterion defined in Equation 3.11. Lighter regions represent SLAs with smaller aggregated L-criterion estimates.

3.3.2 BaySTDetect Model

Figure 3.2a-3.2c show the posterior means of the three spatially indexed parameters, p_i, η_i, and u_i, ordered by the average population at risk, $\sum_t R_{it}/12$ from smallest to largest, left to right.

Figure 3.2a shows the posterior means for the model indicator parameter p_i which represent the posterior probabilities of selecting the common trend model for the ith SLA (refer to 3.3). For those SLAs whose posterior mean of p_i was close to zero, the visits to mammography screening facilities Y_{it} were better modelled by the area-specific model because these SLAs had temporal trends that differed considerably from the common trend γ_t. This was the case for most SLAs, with 91 (58%) SLAs having a posterior mean $p_i \leq 0.05$, where 61 (39%) SLAs actually had a posterior mean p_i equal to zero. The p_i values for SLAs with a larger average population at risk tended to have a smaller posterior mean, indicating that the temporal trend for SLAs with a larger population at risk tended to be less similar to the common temporal trend. The spatial

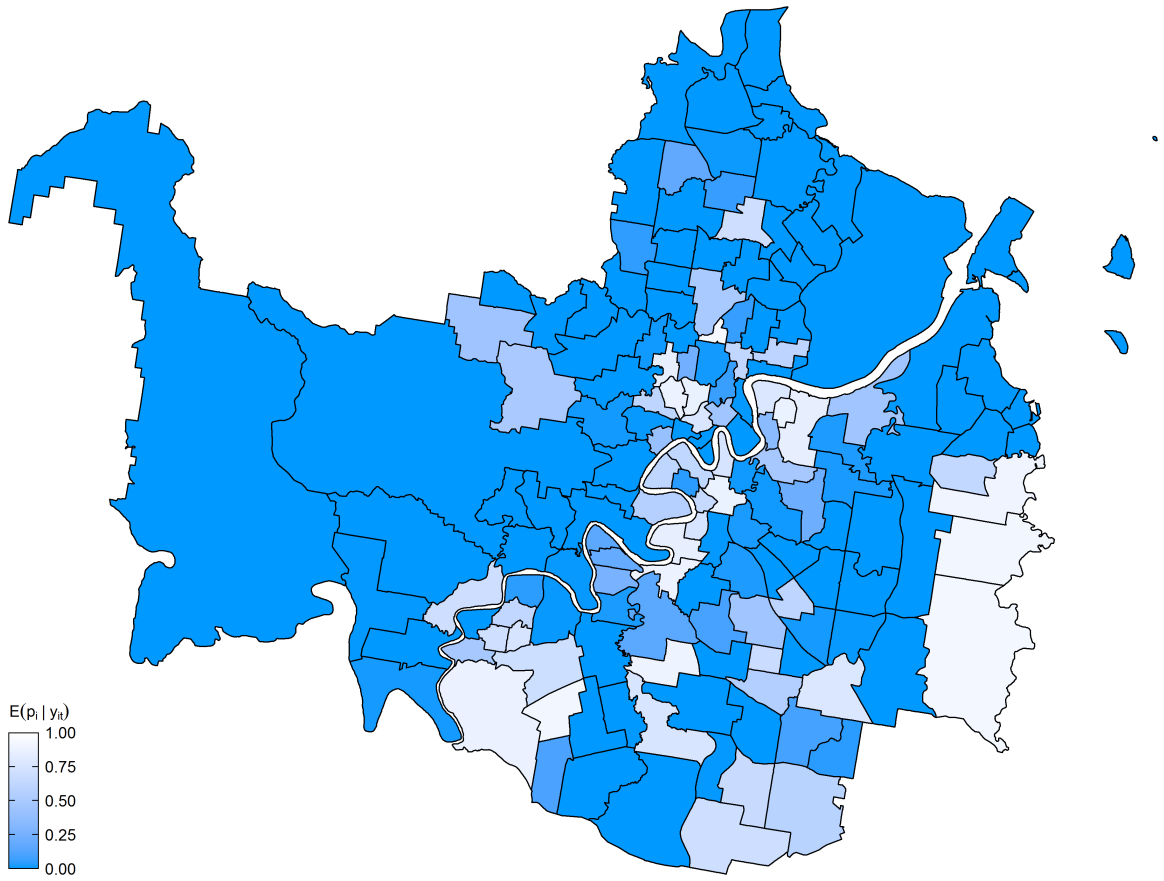


Figure 3.5: Map of SLAs in the Brisbane region (Moreton Island not shown) representing the degree to which SLAs follow the common temporal trend (lighter regions) or exhibit unusual temporal trends (darker regions).

formation of these posterior means is provided in Figure 3.5.

Figures 3.2b, 3.2c shows the posterior means of the parameters for the effects of space (on the logarithm scale) for the common-trend and area-specific models respectively. While the majority of the posterior means of η_i were close to zero, zero was included in only 5 of these 95% credible intervals (CIs), and a small quantity of these means were quite far from zero. In particular, eight SLAs had a posterior mean of < -5 which corresponds to SLAs with zero observed counts. The posterior distributions for u_i in the area-specific model were similar to those for η_i in that the majority of posterior means were close to zero, and it is the same eight SLAs which had a large negative posterior mean. Incidentally, these eight SLAs have the eight smallest aggregated L-criterion estimates, and can easily be identified in Figure 3.4 as the white regions.

The posterior densities of the parameters for the effects of the covariate xit in the two competing models are shown in Figure 3.2d, 3.2e. The densities of these parameters indicate a positive

marginal effect of the catchment covariate on service utilisation, that is, a tendency for service utilisation to be higher in SLAs that fall within the catchment area of a mammography screening facility, as would be expected.

Figure 3.2f shows the posterior means of the parameters for the effects of time for the common-trend model. The posterior means generally decrease with time, indicating a fairly consistent downward trend. The observed counts of visits to mammography screening facilities, however, generally increase over time. While this result is surprising, the temporal effect is very small. More interestingly, there are few SLAs for which the temporal trends agree with this common trend, as indicated by the posterior means of p_i. This is partly explained by the variety of space-time trends in the area-specific model. The posterior means of these space-time trends ξ_{it} for six selected SLAs are shown in Figure 3.2g, 3.2l.

3.3.3 Space-Time Mixture Model

While the BaySTDetect model aims to determine SLAs with unusual temporal trends, the space-time mixture model is designed to identify SLAs whose residual temporal trend exhibits volatility. Figure 3.3a shows the posterior density of the parameter for the covariate effect, whose estimation and interpretation is comparable to that of β' in the BaySTDetect model. Figure 3.3b shows the posterior means of λ_i, which are almost identical to those of the spatial effect term in the BaySTDetect model. (The eight smallest values of λ_i correspond to the SLAs with zero observed counts.)

The temporal effect ψ_t shown in Figure 3.3c indicates a slight, decreasing trend overall. While this differs to the common temporal trend in the BaySTDetect model, the effect size in both cases is small.

Figure 3.3d-3.3i show the posterior means and 95% CIs for the space-time interaction effects ν_{it} for the same six selected SLAs in Figure 3.2g-3.2l, respectively. They exhibit a variety of SLA specific temporal trends to ξ_{it} in the BaySTDetect model. Figure 3.3j-3.3o show the posterior means of the latent indicator variables z_{it} associated with the space-time interaction parameters (equal to $P_{it} + 1$). By analysing the posterior probabilities P_{it} given by Equation 3.10, SLAs with unstable residual temporal trends can be identified. The two rules for classifying unstable SLAs proposed by Abellan et al. (2008) were used using a variety of different values for P_{cut}; the results are summarised graphically in Figure 3.6.

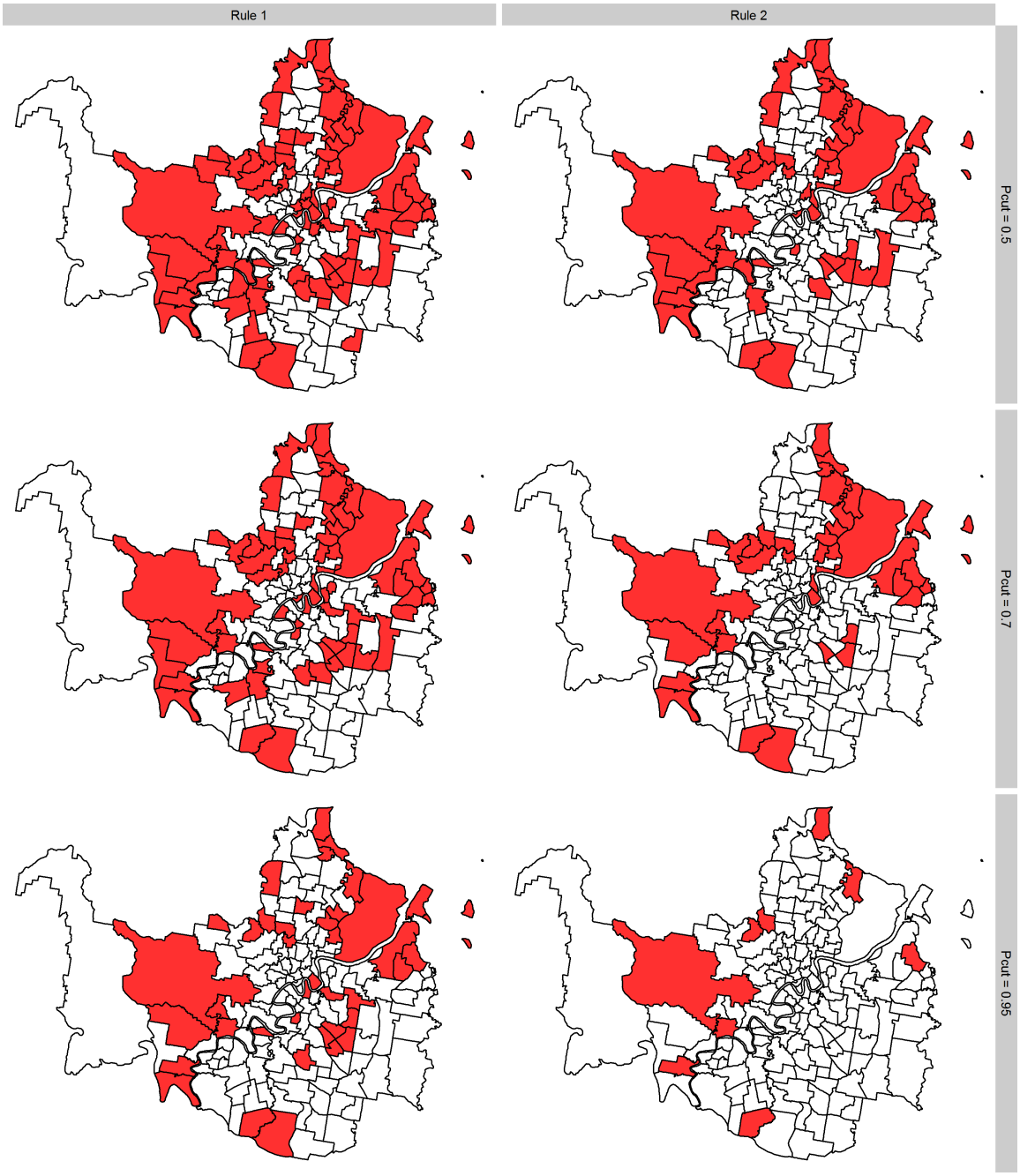


Figure 3.6: Map of SLAs in the Brisbane region (Moreton Island not shown) with unstable trends (shaded areas) determined by Rule 1 and Rule 2 using different values for the threshold, P_{cut}.

3.4 Discussion

This paper has presented two Bayesian hierarchical spatio-temporal models that were used to analyse the utilisation patterns of no-fee mammography screening services in Brisbane over 12 years. In contrast to previous studies, the models sought to identify SLAs with unusual or unstable temporal patterns as an initial step in improving management of these services. The results from both the BaySTDetect and space-time mixture models provide useful insight into the spatial and temporal patterns in mammography screening service utilisation.

First, the BaySTDetect model highlighted a large number of SLAs which had unusual temporal trends relative to the common trend. Although a covariate for the relative availability of services was included in the model to account for mobile facility relocations and facility operating times, service utilisation rates for these SLAs changed from year to year in a way that differs from the common trend.

Second, although the BaySTDetect model estimates a common temporal trend γ_t, it is not common in the sense that very few SLAs exhibit this particular temporal trend. To understand why this is the case, consider the space-time trend ξ_{it} in the area-specific model. Figure 3.2g, 3.2l show the area-specific temporal trend residuals for six selected SLAs. For SLA 10, ξ_{it} exhibits a fairly stable upward trend; SLA 22 shows a downward trend; SLA 57 has a distinctive oscillating pattern; SLA 68 exhibits an oscillating pattern for the first 7 years followed by a relatively stable upward trend; SLA 92 shows no discernable pattern and SLA 158 has a constant trend. This variety of trends suggests that there is not one but several common temporal trends, and explains why the area-specific model is favoured by the Bayesian model choice. Given the large proportion of SLAs that have a temporal trend which departs from the common trend, departures from this trend should be interpreted with care.

Third, there is an apparent correlation between SLAs which follow the common temporal trends (lighter regions in Figure 3.5) and SLAs with stable temporal trends (white regions in Figure 3.6). This is most noticeable for smaller values of P_cut, especially when Rule 1 is used to classify unstable SLAs. This is unsurprising since the common temporal trend is itself fairly flat (only ranges between -0.02 and 0.02 approximately), that is, stable. However, there also exist SLAs the temporal trends of which are unusual but stable, and SLAs the temporal trends of which are usual but unstable.

Fourth, the analysis of the space-time mixture model identified a number of SLAs as unstable, depending on the value of P_cut and the classification rule used. Figure 3.6 indicates that unstable SLAs tend to be situated on the outskirts of the Brisbane region, that is, unstable SLAs tend to be more rural than urban, particularly for larger values of P_cut. Comparing Figures 3.1 and 3.6, these unstable SLAs also tend to be in regions outside of catchment areas. This suggests that distance to a screening facility has an impact on the consistency of clients accessing mammography screening services.

Last, the predicted values for eight SLAs with zero observed counts were the most consistent with the data in both the BaySTDetect and mixture models, as evidenced by the smallest L-criterion estimates. This implies that the existing model components and covariates are adequate in accounting for the lack of service utilisation from these SLAs. Figure 3.1 shows that these SLAs tend to fall outside the catchment areas, which reinforces the notion that service utilisation is influenced by the distance to the nearest screening facility.

The two models presented in this paper are not without limitations. Li et al. (2012) raised a concern about the number of time periods over which the BaySTDetect model detects changes in the temporal trend. The authors advise that a single model indicator pi for each SLA may be 'too restrictive' when the number of time periods is > 10 because the current design assumes only one common temporal trend for the whole period, which is less likely to be the case for longitudinal data collected over many time points. This could be addressed by changing the model indicator to apply to SLAs and years, say p_{it}. Another potential issue with the BaySTDetect model is the *a priori* specification of the prior for the model indicator, p_i. Li et al. (2012) use 0.95 for the Bernoulli probability in Equation 3.3 to reflect their belief that only a small proportion of areas are actually unusual. This rather informative prior may have been adequate for the chronic disease mortality data analysis performed by Li et al. (2012), but based on the results that indicate a large proportion of unusual SLAs, it may be more appropriate to specify a hyperprior for the Bernoulli probability, perhaps using additional spatial covariate information if available.

Similarly, the Dirichlet prior for $\boldsymbol{q}$ in the mixture model could be extended to include additional effects of space and/or time. The change in the dimensionality of $\boldsymbol{q}$ to $\boldsymbol{q}_i$, $\boldsymbol{q}_t$, or $\boldsymbol{q}_{it}$ should be straightforward since the space-time effect ν_{it} is already indexed by space and time. However, this may increase the computational burden significantly.

In both models, there are a large number of parameters to be estimated, some of which have posterior means close to zero (in particular some of the space-time trends ξ_{it} and ν_{it}). It may be beneficial to zero out such parameters using appropriate spike and slab priors.

In this study, the spatial autocorrelation between the observed data for any given year appears to be weak, as indicated by the posterior means of η_i and λ_i, shown in Figure 3.2b and 3.3b. However, measures of spatial autocorrelation such as Moran's I and Geary's C indicate the contrary (results not shown). Although Geary's C is more sensitive to local spatial autocorrelation, such statistics imply that spatial autocorrelation in this dataset is global rather than local, and thus perhaps not easily captured through ICAR priors. Results may be improved by changing the adjacency weight elements w_{ij} to be non-zero for second- and third-order neighbors, for example.

A possible extension to the work of Abellan et al. (2008) concerns the rules used to classify SLAs as unusual or not. The methodology proposed by Abellan et al. (2008) allows SLAs with unstable temporal trends to be identified, but methods to identify the degree to which an SLA is unstable would undoubtedly be more informative and comparable to the results from the BaySTDetect model. Another avenue for future research is the inclusion of additional covariates that vary in space and/or time, such as accessibility to public transport, which may improve inferences relating to the spatial and/or temporal trends.

The L-criterion values also provide insight into the observed trends. It is speculated that the larger $\hat{L}_{it}$ values may be attributed to some unknown factor such as the influence of services offered by private mammography screening facilities. For both models, the predictive performance tended to decrease with time, with the annual average $\hat{L}_{it}$ values increasing by about 4 between 1997 and 2008. Inclusion of other temporal factors may improve predictive performance in later years.

Overall, this paper has shown that the BaySTDetect and space-time mixture models are useful in analysing mammography screening service utilisation data. In particular, the BaySTDetect model was able to identify SLAs which had temporal trends that differed from the overall temporal trend, and the space-time mixture model identified SLAs with unstable temporal trends. Analysis of these models has shown insight into patterns of the observed trends, and showed potentially important factors not yet considered.

Statement of Contribution of Co-authors for Chapter 4

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

This paper was submitted to *PLOS ONE* for publication and is currently under review. The title of this paper is *Improved Bayesian methods for identifying aberrant temporal trends in spatio-temporal data with application to mammography screening services*.

Contributor	**Statement of contribution**
E. W. Duncan Signature and date:	Wrote the code for data analysis, developed the statistical models, wrote the manuscript, revised the manuscript as suggested by co-authors.
N. M. White	Supervised research, provided comments on and helped revise manuscript, aided in interpretation of results.
K. L. Mengersen	Supervised research, provided comments on and helped revise manuscript, aided in interpretation of results.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

Chapter 4

Improved Bayesian methods for identifying aberrant temporal trends in spatio-temporal data with application to mammography screening services

---

Preamble

This chapter relates to the second research objective. This chapter introduces several extensions to the methodology presented in Chapter 3. These extensions aim to address some of the limitations of the models and improve the inferences that can be drawn from the model estimates. Aside from the model limitations, the ideas for some of these extensions originated from a closer examination of the literature on mammography screening studies. This chapter begins with a literature review, focusing on potential factors that may explain variation in screening rates. This leads to the definition of three covariates to be used in the models as predictors of the relative risk. The first covariate, denote $x_{1,it}$, is the same as x_{it} in the previous chapter. The subscript '1' has been added to distinguish it from the other two new covariates. Likewise, subscripts have been added to each of the covariate effects for clarity.

As the result of another extension, a subscript is also added to each of the parameters in the linear predictor of the BaySTDetect model to denote membership to a particular component of a mixture model. This mixture model is introduced to deal with one of the limitations of the BaySTDetect model, namely the ability to estimate only a single common temporal trend.

The data analysed in this chapter is the same data presented in Chapter 3, allowing direct comparisons of the results between the two studies. In particular, the following figures can be compared: Figure 4.2 and Figure 3.2; Figure 4.5 and Figure 3.3; Figure 4.6 and Figure 3.5; and Figure 4.7 and Figure 3.6.

Another key change to the methodology from the previous chapter is that the variances defining the area-specific temporal random effects in the space-time mixture model are generated independently. In Chapter 3, these variance parameters were constrained in an attempt to prevent label switching. This methodology is consistent with the specification of the original models and with common practice. However, further research into the label switching problem revealed that this methodology may be flawed. The problem of label switching is the topic of Chapter 5.

This chapter, and a brief summary of the work in Chapter 3, was presented at the Social Computing Summit, Gold Coast, Australia, 14-15 December 2016.

4.1 Introduction

Regular mammography screening services play a vital role in the early detection of breast cancer. However, evidence suggests that participation in mammography screening programs is irregular, and participation rates are suboptimal (Brown et al. 2009; Engelman et al. 2002; Jackson et al. 2009; Peek and Han 2004; Zenk et al. 2006). The reasons for screening service underutilisation and differences in screening rates between subpopulations of different demographics have been the focus of several studies. These studies investigated the impact of geographic proximity to mammography screening facilities, availability of transport, socio-economic status, and other factors on mammography screening rates. The aim of this paper is to consider the key factors affecting screening rates identified in the literature as covariates in statistical models designed to identify aberrant temporal trends in mammography screening utilisation data, and extend the existing methodology to better understand the reasons for these aberrant patterns. Ultimately, this will allow decision makers to manage and improve screening services, and hopefully increase participation in mammography screening programs.

Duncan et al. (2016) applied two Bayesian hierarchical models for analysing spatio-temporal data: the BaySTDetect model proposed by Li et al. (2012), and another model proposed by Abellan et al. (2008). Both models fall within the broad class of linear models and more

specifically constitute spatial generalised linear mixed models (Schall 1991; Zeger and Karim 1991). The goal of both models is to identify areas with aberrant temporal trends, where the definition of an aberrant temporal trend is specific to the model. The BaySTDetect model considers aberrant temporal trends as temporal trends which did not adhere to an overall common trend, while the model of Abellan et al. (2008) considers aberrant temporal trends to be temporal trends which exhibit more excessive volatility, or instability.

In this paper, a number of extensions to both of these models are presented. These extensions aim to address some of the shortcomings of the original models, as identified by Li et al. (2012) and Abellan et al. (2008), and the concluding remarks discussed in Duncan et al. (2016). Specifically, we include covariates to account for differences in spatial accessibility of screening services and spatial heterogeneity in socioeconomic status, relax the assumption that there is a single common temporal trend by introducing a mixture model, and prescribe more informative methods for classifying temporal trends. The idea behind this first extension is well founded in the wider literature, which we subsequently review.

Previous studies of mammography screening have analysed the effect of spatial accessibility on service utilisation rates (Engelman et al. 2002; Hyndman et al. 2000; Kreher et al. 1995), or health-related outcomes such as stage at cancer diagnosis (Huang et al. 2009). Three common measures of accessibility include travel distance to the nearest facility (Engelman et al. 2002; Huang et al. 2009; Hyndman et al. 2000; Jackson et al. 2009; Kreher et al. 1995; Zenk et al. 2006), travel time, either by public or private transportation, to the nearest facility (Jackson et al. 2009; Kreher et al. 1995; Zenk et al. 2006), and the availability of transportation (Kreher et al. 1995). Screening utilisation rates were often reported to be negatively correlated with travel distance (Engelman et al. 2002; Hyndman et al. 2000; Maxwell 2000). Maxwell (2000) found that the effect of such variables on screening rates was relatively small compared to the effect of socioeconomic status. Kreher et al. (1995) found travel distance, travel time, and a lack of available transportation did not have a significant impact on utilisation rates, However, this conclusion was based on questionnaire responses from women already seeking mammography screening services, and as Kreher et al. (1995) cautiously note, such spatial accessibility indicators may be a far greater impediment to women who do not access screening services.

Hyndman et al. (2000) make the important observation that screening rates reported by facilities

offering no- to low-fee mammography screening services may be lower than expected due to some women favouring private, opportunistic screening services. The main advantage of organised screening to participants is the cost, as organised screening services are usually offered free of charge or are heavily subsidised (Bulliard et al. 2009; Miles et al. 2004; Peek and Han 2004). In terms of follow-up procedures, quality assurance, and protection from harm such as over-screening and over-diagnosis, the evidence of superiority of one mode of screening over the other is controversial (Bihrmann et al. 2008; Bulliard et al. 2009; Chamot et al. 2007; Jørgensen and Gøtzsche 2009; Miles et al. 2004; Peek and Han 2004). Thus preference for a particular mode of screening depends on the individual's perception of efficacy, safety, and cost-effectiveness of that mode, and their individual needs (Aro et al. 2001; Chamot et al. 2007). Possible reasons for seeking opportunistic screening include a desire for more frequent screening than that offered by an organised screening program, ineligibility for organised mammography services, and the perception that opportunistic screening is more effective or of a higher standard (Bulliard et al. 2009; Chamot et al. 2007; Miles et al. 2004).

Socioeconomic status and spatial accessibility have been widely cited as important factors in explaining variation in screening service utilisation, even when these services are offered free of charge (Bulliard et al. 2009; Chamot et al. 2007; Jensen et al. 2005; Miles et al. 2004; Peek and Han 2004). Socioeconomic factors that have been commonly used included income, health insurance, having a usual source of health care, age, level of education, occupation, employment status, marital status, English proficiency, and smoking habits (Brown et al. 2009; Jackson et al. 2009; Legler et al. 2002; Makuc et al. 1999; Mandelblatt et al. 1999; Selvin and Brett 2003). The relevance and importance of these factors varied between studies since they are inextricably linked to the population being studied. Nonetheless, studies have consistently reported that women with a higher socioeconomic status have better access to health care, and are more likely to use mammography screening services (Jackson et al. 2009; Mandelblatt et al. 1999; Maxwell 2000; Selvin and Brett 2003).

The urbanity and physical size of the spatial units may also be important spatial accessibility indicators (Zenk et al. 2006). It is often reported that the effect of spatial accessibility variables is greater for women living in rural areas and/or socioeconomically disadvantaged areas (Engelman et al. 2002; Huang et al. 2009; Jackson et al. 2009; Mandelblatt et al. 1999; Zenk

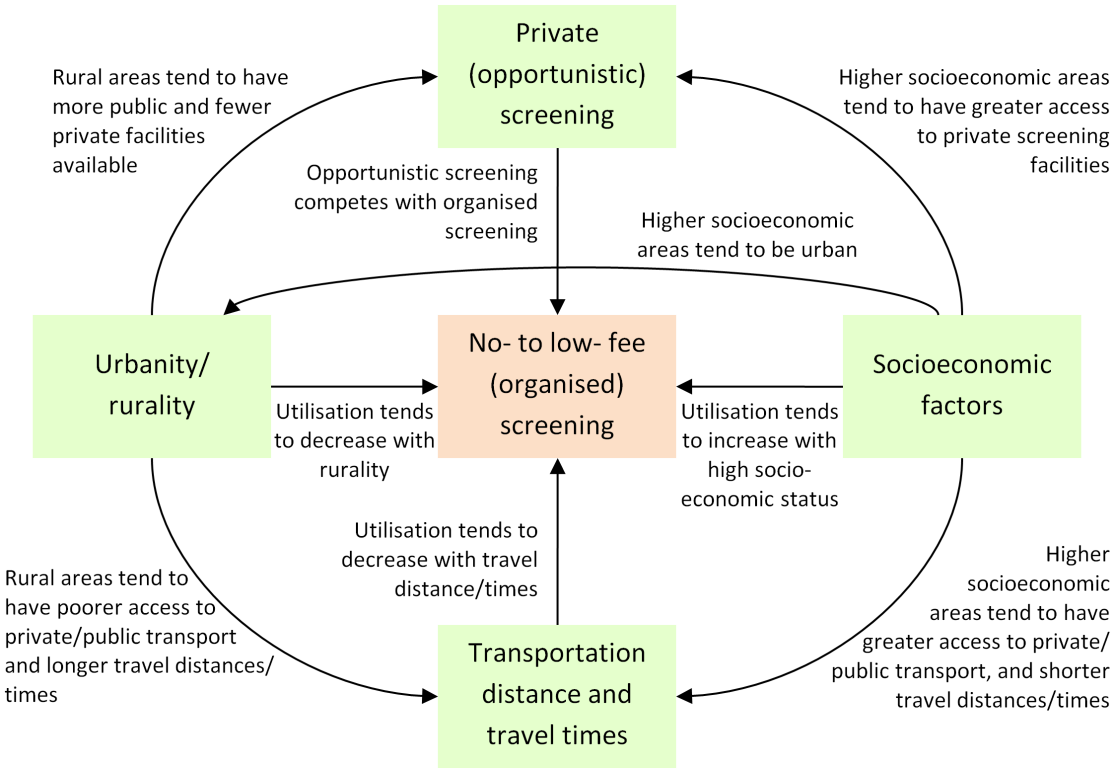


Figure 4.1: Relationship between potential factors affecting no- to low-fee mammography screening.

et al. 2006). Interventions aimed at increasing accessibility, such as the strategic location or relocation of mammography screening facilities and the use of mobile facilities, have been suggested, particularly interventions targeted at women living in socioeconomically disadvantaged and/or remote areas (Chamot et al. 2007; Hyndman et al. 2000; Legler et al. 2002; Maxwell 2000; Miles et al. 2004; Peek and Han 2004; Selvin and Brett 2003). While not definitive, the notion that women who have a higher socioeconomic status and/or live in an urban area tend to prefer opportunistic screening over organised screening is plausible (Chamot et al. 2007; Miles et al. 2004).

The complex relationship among socioeconomic status, the degree of urbanity or remoteness of living, distance and travel times to a screening facility, and the availability of private screening facilities are summarised in Figure 4.1.

The structure of the remainder of this paper is as follows. Section 4.2 begins by describing the data and the original model specification as presented by Duncan et al. (2016). A number of extensions for each model are then proposed, followed by a description of how these models were implemented. Section 4.3 provides the results from the different model extensions. Section 4.4 contains a discussion on the results, the methodology presented in this paper, and the

limitations.

4.2 Methods

4.2.1 Data

The models presented in this study are applied to the same dataset analysed in Duncan et al. (2016). These data consist of annual numbers of visits to no-fee mammography facilities in the Brisbane City Council region between 1997 and 2008 inclusive, by statistical local area (SLA). The availability of opportunistic and organised screening in Brisbane, the use of mobile screening services, and the geography of Brisbane, especially the variation in the size of SLAs, make this an interesting example to study. The eligible population was defined as women aged 40 years or more at the time of screening, in accordance with the BreastScreen Australia Program eligibility criteria (Australian Government Department of Health (AGDH) 2014).

Three covariates are considered for inclusion in the models. A covariate reflecting the relative availability of services in a catchment area was derived previously by Duncan et al. (2016), defined as

$$x_{1,it} = \frac{d_{it}}{\max_i d_{it}} \tag{4.1}$$

where d_{it} is the cumulative number of days that each mammography screening facility was operating in SLA i or any SLA that shares a border with SLA i, $\{i = 1, \ldots, 158\}$ during year t, $\{t = 1, \ldots, 12\}$. This covariate was included to account for the fact that some mammography screening facilities did not operate for the full 12-year period, for example mobile facilities.

The index of relative socio-economic advantage and disadvantage (IRSAD), published by the Australian Bureau of Statistics (ABS), was selected as an area-level measure of socioeconomic status (Australian Bureau of Statistics (ABS) 2013a). For the 158 Brisbane SLAs, the IRSAD scores range from 798 to 1202. For the analysis presented in Section 4.2.3, a scaled version of IRSAD $x_{2,i}$ was used such that the largest value of $x_{2,i}$ is 1.

Using the geocodes for each screening facility, combined with the geocodes for each SLA centroid, obtained from a shapefile (Australian Bureau of Statistics (ABS) 2011), the travel time from each SLA centroid to each available screening facility was calculated for each year using the Google Maps Distance Matrix API (Google 2015) via R using the packages `RCurl`

(Temple Lang, D. and the CRAN team 2015a) and `XML` (Temple Lang, D. and the CRAN team 2015b). The shortest travel time to an available screening facility was then determined for each year, where an available screening facility is taken to be any facility, mobile or permanent, that is operating for at least one day in a year. Standardising these values, a covariate for travel time was defined as:

$$x_{3,it} = \frac{t_{it}}{\max_i t_{it}} \tag{4.2}$$

where t_{it} is the shortest travel time in seconds from the centroid of SLA i to a screening facility in year t. This method required numerous assumptions which are discussed in Section 4.4.

Additional covariate information to account for the effects of opportunistic screening and rurality of SLAs was also considered. However, these covariates were ultimately not used in the final models for a variety of reasons. An explanation of the acquisition of such data and the reasons for their exclusion is discussed in section 4.4.

4.2.2 Original Model Specification

For each of the ensuing models and extensions, a Poisson random variable Y_{it} is assumed to model the count data, y_{it}. For the BaySTDetect model (Li et al. 2012),

$$Y_{it} \sim \text{Po}\left(E_{it}\mu_{it}\right)$$

where

$$E_{it} = \frac{\sum_i E_{it}}{\sum_i R_{it}} R_{it} \tag{4.3}$$

is the expected number of visits calculated using internal standardisation (Pascutto et al. 2000), R_{it} is the population at risk, and

$$\log\left(\mu_{it}\right) = \begin{cases} \alpha + \eta_i + \gamma_t + \beta_1 x_{1,it} & \text{if} p_i = 1 \\ u_i + \xi_{it} + \beta_1' x_{1,it} & \text{if} p_i = 0 \end{cases}. \tag{4.4}$$

Under Equation (4.4), the log-relative risk in SLA i during year t is given by one of two competing models, depending on the value of the unknown model indicator, p_i. For $p_i = 1$, α denotes a common intercept, and η_i and γ_t denote random effects of space and time respectively. For $p_i = 0$, u_i is interpreted as an area-specific intercept and ξ_{it} is an area-specific random effect.

The covariate $x_{1,it}$ is given in Equation (4.1). When $p_i = 1$, the model fitted to the data includes a random effect for time, γ_t; conversely when $p_i = 0$, the effect for time is encapsulated by the area-specific trend, ξ_{it}. Hence these two competing models are referred to as the common trend and area-specific trend models respectively. The prior distributions for unknown parameters are as follows:

$$\alpha \sim \mathcal{N}\left(0, \sigma_\alpha^2\right)$$

$$\eta_i | \boldsymbol{\eta}_{\setminus i} \sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}} \sum_{j=1}^{N} w_{ij}\eta_j, \frac{\sigma_\eta^2}{\sum_j w_{ij}}\right) \tag{4.5}$$

$$\gamma_t | \boldsymbol{\gamma}_{\setminus t} \sim \mathcal{N}\left(\frac{1}{\sum_s \tilde{w}_{st}} \sum_{s=1}^{T} \tilde{w}_{st}\gamma_s, \frac{\sigma_\gamma^2}{\sum_s \tilde{w}_{st}}\right) \tag{4.6}$$

$$\beta_1 \sim \mathcal{N}\left(0, \sigma_{\beta_1}^2\right)$$

$$u_i \sim \mathcal{N}\left(0, \sigma_{u_i}^2\right)$$

$$\xi_{it} | \boldsymbol{\xi}_{i\setminus t} \sim \mathcal{N}\left(\frac{1}{\sum_s \tilde{w}_{st}} \sum_{s=1}^{T} \tilde{w}_{st}\xi_{is}, \frac{\sigma_{\xi,i}^2}{\sum_s \tilde{w}_{st}}\right) \tag{4.7}$$

$$\beta_1' \sim \mathcal{N}\left(0, \sigma_{\beta_1'}^2\right)$$

$$p_i \sim \text{Bern}\left(\delta_i\right). \tag{4.8}$$

The notation $\setminus i$ denotes all areas excluding i, and $\setminus t$ denotes all years excluding t. The values w_{ij} and $\tilde{w}_{st}$ are the i-j^{th} and s-t^{th} elements of the symmetric spatial and temporal weight matrices $\mathbf{W}$ and $\tilde{\mathbf{W}}$ respectively, whose diagonal elements w_{ii} and $\tilde{w}_{tt}$ are zero. The priors in Equations (4.5), (4.6) and (4.7) are intrinsic conditional autoregressive (ICAR) priors (Besag and Kooperberg 1995) which allow for spatial and temporal smoothing. The variances σ_α^2, $\sigma_{\beta_1}^2$, $\sigma_{u_i}^2$, and $\sigma_{\beta_1'}^2$ are set to 1000 to instil vague priors for the respective parameters. The probability of belonging to the common trend model, δ_i, was originally fixed at 0.95 to reflect the assumption that trends were unlikely to be area-specific.

For the space-time mixture model, Abellan et al. (2008), defined a space-time mixture model for describing the log-relative risks. This model takes the following general form:

$$Y_{it} \sim \text{Po}\left(E_{it}\pi_{it}\right),$$

$$\log\left(\pi_{it}\right) = \tau + \lambda_i + \psi_t + \nu_{it} + b_1 x_{1,it} \tag{4.9}$$

where E_{it} is given by Equation (4.3), and

$$\begin{aligned}
\tau &\sim \mathcal{N}\left(0, \sigma_{\tau}^2\right) \\
\lambda_i|\boldsymbol{\lambda}_{\setminus i} &\sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}} \sum_{j=1}^{N} w_{ij}\lambda_j, \frac{\sigma_{\lambda}^2}{\sum_j w_{ij}}\right) \\
\psi_t|\boldsymbol{\psi}_{\setminus t} &\sim \mathcal{N}\left(\frac{1}{\sum_s \tilde{w}_{st}} \sum_{s=1}^{T} \tilde{w}_{st}\psi_s, \frac{\sigma_{\psi}^2}{\sum_s \tilde{w}_{st}}\right) \\
\nu_{it} &= q\mathcal{N}\left(0, \sigma_{\nu_1}^2\right) + (1-q)\mathcal{N}\left(0, \sigma_{\nu_2}^2\right) \\
b_1 &\sim \mathcal{N}\left(0, \sigma_{b_1}^2\right).
\end{aligned}$$

The variances σ_{τ}^2 and $\sigma_{b_1}^2$ are set to 1000. The area-specific temporal random effects ν_{it} are described by a two-component Gaussian mixture model, characterised by component-specific variances,

$$\begin{aligned}
\sigma_{nu_1} &\sim \mathcal{N}\left(0, 0.01\right) \mathbb{I}_{(0,+\infty)} \\
\sigma_{nu_2} &\sim \mathcal{N}\left(0, 100\right) \mathbb{I}_{(0,+\infty)},
\end{aligned}$$

where $\mathbb{I}$ denotes the indicator function $\mathbb{I}_{(a,b)} = 1$ if the random variate lies between a and b, and 0 otherwise. As per the standard approach, the latent indicator variables are assumed to be distributed according to a multinomial distribution for a single event,

$$z_{it} \sim \text{Mult}(1, \boldsymbol{q})$$

defined by the mixing proportions $\boldsymbol{q} = (q, 1-q)$ are given a Dirichlet prior,

$$\boldsymbol{q} \sim \mathcal{D}(1, 1) \tag{4.10}$$

If $z_{it} = 1$, the variance of ν_{it} is small, signifying a stable temporal pattern for SLA i. Conversely, for $z_{it} = 2$, the variance of ν_{it} is large, signifying an unstable temporal pattern. Hence the posterior probabilities of the latent indicator variables can be used to identify SLAs with unstable temporal trends.

4.2.3 Model Extensions

Extensions to the BaySTDetect Model

We provide three extensions to the BaySTDetect Model. The first extension is the inclusion of covariate $x_{2,i}$ to reflect spatial heterogeneity in socioeconomic status. However, a review of the literature revealed studies which found socioeconomic status to be an important predictor of service utilisation. Including this covariate as an additional term into the linear predictor of the regression equations, given by Equation (4.4), results in issues with multicollinearity between regression parameters, and its effect is not readily identifiable. An alternative is to introduce this covariate information through the model indicator p_i, which overcomes the identifiability problem and simultaneously avoids the need to assume a value for δ_i *a priori* which necessitates a strong prior belief about the proportion of areas with unusual temporal trends.

Thus rather than specifying $\delta_i = 0.95$ in Equation (4.8), this parameter is estimated as follows:

$$\delta_i \sim \text{Beta}\left(\alpha_{\delta,i}, 2\right) \tag{4.11}$$

$$log\left(\alpha_{\delta,i}\right) \sim \mathcal{N}\left(\beta_2 x_{2i}, \sigma^2_{\alpha_\delta}\right) \tag{4.12}$$

$$\beta_2 \sim \mathcal{N}\left(0, \sigma^2_{\beta_2}\right) \tag{4.13}$$

A beta prior is used for δ_i, with the second shape parameter fixed at 2. This reduces the model complexity but still provides adequate flexibility in the estimation of δ_i. If the effect of $x_{2,i}$ is negative, the model choice step will favour the area-specific model, and if the effect of $x_{2,i}$ is positive, the model choice step will favour the common trend model. Thus β_2 is given a vague prior centred around zero, and the variances $\sigma^2_{\alpha_\delta}$ and $\sigma^2_{\beta_2}$ are set to 100.

A review of the literature revealed that some women may find it too difficult or inconvenient to access mammography screening due to a lack of transportation, or long travel distances and/or travel times. The covariate $x_{1,it}$ is partly designed to account for such spatial accessibility barriers. However, the assumptions used to define the catchment areas do not necessarily make this covariate a realistic surrogate for travel distance or time. For example, it is possible that one may live outside a catchment area and yet be only a short distance from a screening facility, and one may live within a catchment area but take a long time to reach a screening facility due

to the nature of the transport system.

The second extension to this model is designed to measure spatial accessibility more accurately. We take a similar approach to recent studies, for example Zenk et al. (2006), Engelman et al. (2002), and Huang et al. (2009), and estimate the shortest time taken to travel from the centroid of each SLA to a screening facility, for each year. The additive effects of this covariate are estimated by replacing Equation (4.4) with

$$\log(\mu_{it}) = \begin{cases} \alpha + \eta_i + \gamma_t + \beta_1 x_{1,it} + \beta_3 x_{3,it} & \text{if } p_i = 1 \\ u_i + \xi_{it} + \beta_1' x_{1,it} + \beta_3' x_{3,it} & \text{if } p_i = 0 \end{cases} \tag{4.14}$$

where β_3 and β_3' are given suitably vague priors,

$$\beta_3 \sim \mathcal{N}\left(0, \sigma^2_{\beta_3}\right)$$
$$\beta_3' \sim \mathcal{N}\left(0, \sigma^2_{\beta_3'}\right).$$

The variances $\sigma^2_{\beta_3}$ and $\sigma^2_{\beta_3'}$ are set to 1000. Another limitation of the BaySTDetect model is that it only accommodated for a single common trend. The analysis in Duncan et al. (2016) revealed that there were very few areas with a temporal trend similar to the common temporal trend, and speculated that there may there may exist not one, but several clusters of common temporal trends. As a third extension, we address this limitation by allowing for multiple common trends using a Gaussian mixture model. To achieve this, the common trend component of the log-relative risk in Equation (4.14) is represented as a finite mixture,

$$\log(\mu_{it}) = \sum_{k=1}^{K} \rho_k \left(\alpha_k + \eta_{ik} + \gamma_{tk} + \beta_{1,k} x_{1,it} + \beta_{3,k} x_{3,it}\right) \text{if } p_i = 1.$$

A latent allocation variable

$$\zeta_i \sim \text{Mult}(1, \boldsymbol{\rho})$$

is used to identify to which of the K components each $\{\mu_{it} | p_i = 1\}$ belongs. The prior for the mixture weights is a Dirichlet distribution,

$$\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)^{\mathsf{T}} \sim \mathcal{D}\left(\alpha_{\rho,1}, \ldots, \alpha_{\rho,K}\right)$$

where the concentration parameters $\{\alpha_{\rho,1}, \ldots, \alpha_{\rho,K}\}$ are assigned independent gamma prior distributions,

$$\alpha_{\rho,k} \sim \text{Gam}\,(0.5, 0.005) \text{for } k = 1, \ldots, K.$$

The priors for α_k, η_{ik}, γ_{tk}, $\beta_{1,k}$, and $\beta_{3,k}$ retain their same form but are estimated for each component independently,

$$\begin{aligned}
\alpha_k &\sim \mathcal{N}\left(0, \sigma_\alpha^2\right) \\
\eta_{ik}|\boldsymbol{\eta}_{\setminus i,k} &\sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}} \sum_{j=1}^{N} w_{ij}\eta_{jk}, \frac{\sigma_{\eta,k}^2}{\sum_j w_{ij}}\right) \\
\gamma_{tk}|\boldsymbol{\gamma}_{\setminus t,k} &\sim \mathcal{N}\left(\frac{1}{\sum_s \tilde{w}_{st}} \sum_{s=1}^{T} \tilde{w}_{st}\gamma_{sk}, \frac{\sigma_{\gamma,k}^2}{\sum_s \tilde{w}_{st}}\right) \\
\beta_{1,k} &\sim \mathcal{N}\left(0, \sigma_{\beta_1}^2\right) \\
\beta_{3,k} &\sim \mathcal{N}\left(0, \sigma_{\beta_3}^2\right).
\end{aligned}$$

The impact of these extensions on inference is as follows. For each Markov Chain Monte Carlo (MCMC) iteration, both the area-specific and common trend models are fit to the data, and then the model choice step is performed. However, for the purpose of summarising the posterior estimates of particular parameters in this model, it is useful to apply a dichotomous classification rule based on p_i so that each SLA is considered to belong to one of two groups: SLAs with a common temporal trend, or SLAs with an unusual temporal trend. To this end, SLAs satisfying

$$\text{E}\,(p_i|y_{it}) \leq 0.05 \tag{4.15}$$

are classified into the latter group. Similarly, to facilitate meaningful inferences about parameters pertaining to the common trend mixture model, membership to a particular component is determined by

$$\arg\max_k \left(\frac{1}{k}\sum_{m=1}^{M} \zeta_i^{(m)} = k \middle| y_{it}\right) \tag{4.16}$$

where M is the number of MCMC iterations and $\zeta_i^{(m)}$ is the value of the latent model indicator for SLA i in the m^{th} iteration. Although cluster membership is not generally as unanimous in reality, for most SLAs, cluster membership in this analysis is fairly homogenous. The

implication of combining these simplifications is that some mixture components may appear empty or near-empty, depending on which SLAs satisfy Equation (4.15).

Extensions to the Space-time Mixture Model

The space-time mixture model, described in Equation (4.9), is extended in three ways. First, a new covariate is added to better account for spatial heterogeneity and estimate the effect of socioeconomic status on the log-relative risk. The motivation for including this covariate is the same as for the BaySTDetect model. This covariate, denoted $x_{2,i}$, is introduced by replacing the Dirichlet prior on the mixture weights in Equation (4.10) with

$$\begin{aligned} \text{logit}\left(q_i\right) &= b_2 x_{2,i} \\ b_2 &\sim \mathcal{N}\left(0, 100\right). \end{aligned}$$

The mixture weights $\boldsymbol{q}_i = (q_i, 1 - q_i)$ are now area-specific and the prior for the latent model indicator is easily adapted:

$$z_{it} \sim \text{Mult}(1, \boldsymbol{q}_i).$$

The effect of travel time to a screening facility is estimated by extending the linear predictor in Equation (4.9) to include $x_{3,it}$:

$$\log\left(\pi_{it}\right) = \tau + \lambda_i + \psi_t + \nu_{it} + b_1 x_{1,it} + b_3 x_{3,it}.$$

In addition to these two methodological extensions, we revised the analytical approach of Abellan et al. (2008) to classifying temporal trends as unstable. The authors proposed two rules for classifying the stability of temporal trends. These rules compare the posterior probabilities $P_{it} = \Pr\left(z_{it} = 2 | y_{it}\right)$ that ν_{it} has a large variance to some cut-off threshold P_{cut}. Rule 1 classifies the temporal trend of SLA i to be unstable if $P_{it} > P_{cut}$ for at least one t, and Rule 2 classifies the temporal trend of SLA i to be unstable if the average of the three largest P_{it} values are greater than P_{cut}. However, the aberrant nature of temporal trends is not necessarily as unequivocal as this. For the BaySTDetect model, for example, the degree to which SLAs are unusual is considered. We take a similar approach to Abellan et al. (2008) by replacing these

two rules with two new rules; the instability of the temporal trends in the i^{th} SLA is measured as the proportion of P_{it} values greater than P_{cut} (Rule 3), and the difference between the average of the $l < T$ largest P_{it} values and P_{cut}, with a lower bound at zero (Rule 4).

4.2.4 Implementation

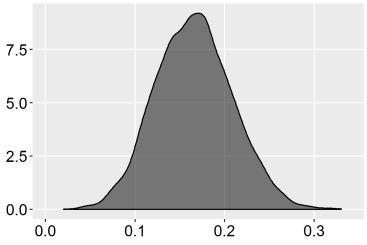
The two models with the aforementioned extensions were implemented in WinBUGS (Lunn et al. 2000) using the `R2WinBUGS` package in R (R Core Team 2015; Sturtz et al. 2005). The posterior distributions for the BaySTDetect and space-time mixture models were estimated using 25000 MCMC iterations (with a thinning factor of 5), after discarding 25000 burn-in iterations.

Prior to implementing the extended BaySTDetect model, a cluster analysis was performed on the posterior means of the area-specific trends, $\text{E}(\xi_{it}|y_{it})$, obtained from fitting the original BaySTDetect model (Duncan et al. 2016). A k-means clustering analysis (MacQueen 1967) was performed for various numbers of clusters using the Hartigan-Wong algorithm (Hartigan and Wong 1979), implemented in R via the `kmeans` function in the `stats` package R Core Team (2015). The temporal trends ξ_{it} in the area-specific model may be interpreted as temporal noise for SLA i, and therefore no spatial autocorrelation between these values is expected. The purpose of the k-means analysis is to help determine what spatial autocorrelation was unexplained, and evaluate the effectiveness of extending the BaySTDetect model to allow for multiple common trends. Comparisons are also made between the cluster memberships determined by the k-means analysis and the mixture component allocations from the extended BaySTDetect model.

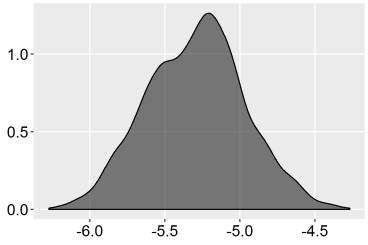
4.3 Introduction

4.3.1 BaySTDetect Model

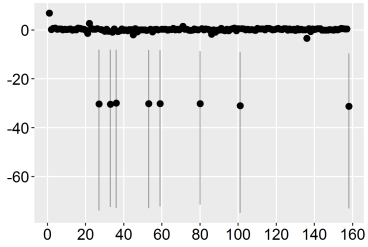
Figure 4.2 summarises the marginal posterior distributions of the main model parameters in the extended BaySTDetect model. Based on Equations (4.15) and (4.16), components 3 and 7 are comprised entirely of SLAs with an unusual temporal trend, and thus effectively empty, while component 5 is comprised of only 5 SLAs which are not considered unusual.



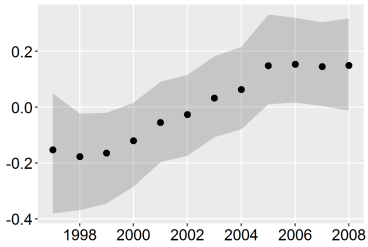
(a) Posterior density for β_1'



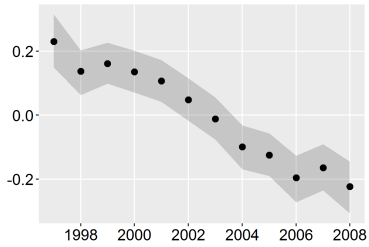
(b) Posterior density for β_3'



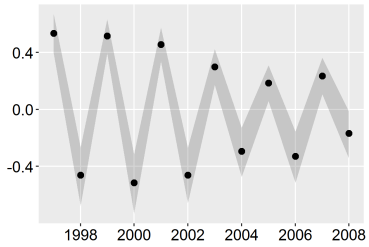
(c) Posterior mean and 95% CI for u_i



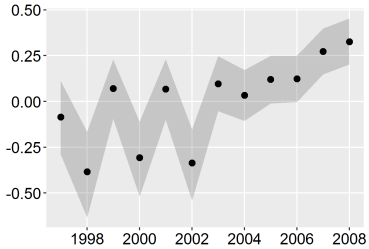
(d) Posterior mean and 95% CI for ξ_{10t}



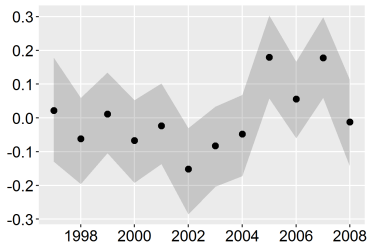
(e) Posterior mean and 95% CI for ξ_{22t}



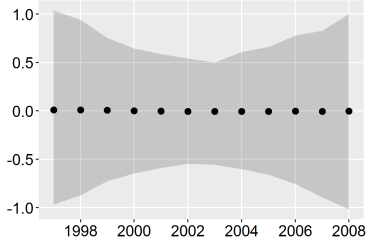
(f) Posterior mean and 95% CI for ξ_{57t}



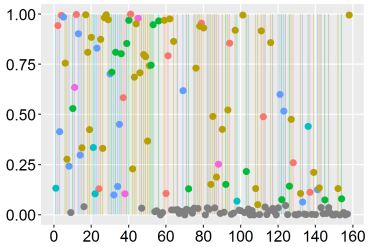
(g) Posterior mean and 95% CI for ξ_{68t}



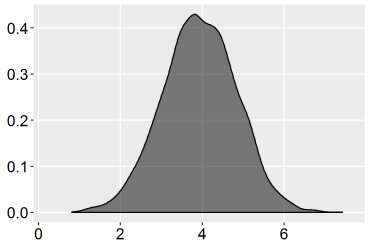
(h) Posterior mean and 95% CI for ξ_{92t}



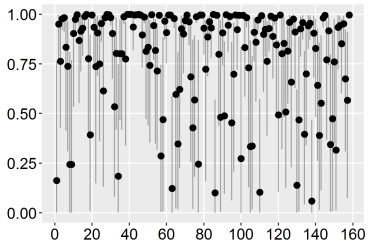
(i) Posterior mean and 95% CI for ξ_{158t}



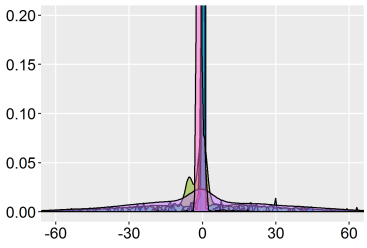
(j) Posterior mean and 95% CI for p_i



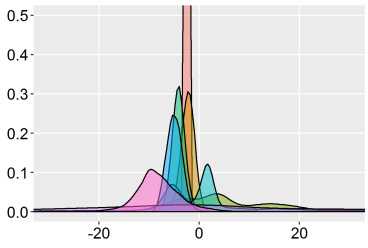
(k) Posterior density for β_2



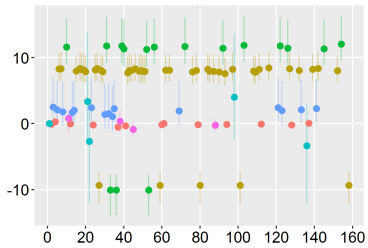
(l) Posterior mean and 95% CI for δ_i



(m) Posterior density for $\beta_{1,k}$

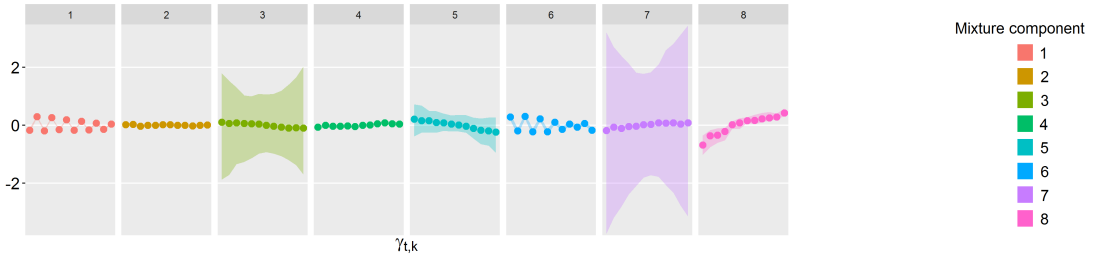


(n) Posterior density for $\beta_{3,k}$



(o) Posterior mean and 95% CI for $\eta_{i,k}$

Figure 4.2: Posterior summary of the main model parameters for 1 chain of the BaySTDetect model.



(p) Posterior mean and 95% CI for $\gamma_{t,k}$

Figure 4.2: Posterior summary of the main model parameters for 1 chain of the BaySTDetect model (cont.).

Figures 4.2a-4.2i summarise the marginal posterior distributions of the main parameters relating to the area-specific model, namely those in Equation (4.4). Figures 4.2a-4.2b show the posterior densities for the effects of relative service availability $x_{1,it}$ and travel time $x_{3,it}$ on service utilisation respectively. As expected, the marginal effect of $x_{1,it}$ is positive while the marginal effect of $x_{3,it}$ is negative. These estimates both indicate that service utilisation tends to be higher when screening services are more accessible, in particular, when time required to travel to a screening facility is shorter. Figure 4.2c shows the posterior means of the spatial random effects in the area specific model, and Figures 4.2d-4.2i show the posterior means of the temporal random effects for six selected SLAs. While the estimated effect of $x_{3,it}$ (Figure 4.2b) is quite strong, the values are relatively small, and consequently a wide variety of temporal trends are estimated by the area-specific model (Figure 4.2d-4.2i). However, the inclusion of a finite mixture in the common trend model allows SLAs with similar temporal trends to be clustered. The mixture applies not only to the random temporal effect, but also to the other parameters in the linear predictor, including the coefficients of the covariates $x_{1,it}$ and $x_{3,it}$, which provides a better model fit.

As a result of the finite mixture, only 51 (32%) SLAs satisfy Equation (4.15), compared to 91 (58%) when the unmodified BaySTDetect model is used (Duncan et al. 2016). The posterior means $\mathrm{E}(p_i|y_{it})$ are shown in Figure 4.2j, where the colours denote the majority component membership as determined by Equation (4.16), and the 51 unusual SLAs are shown in grey. The SLAs are in ascending order of the average population at risk, $\sum_t R_{it}$ from left to right. Figure 4.2j shows that SLAs with a smaller population at risk tend to have a common temporal trend, which is consistent with the results from the BaySTDetect model with a single common temporal trend (Duncan et al. 2016). The marginal effect of socioeconomic status $x_{2,i}$ is shown in Figure 4.2k, which has a posterior mean close to 4, and hence the model choice step tends to

favour the common trend model, as indicated by the posterior means of the Bernoulli probability parameter δ_i shown in Figure 4.2l.

The effects of $x_{1,it}$ and $x_{3,it}$ for each mixture component in the common trend model are shown in Figures 4.2m-4.2n respectively. For components 3, 5, and 7, which appear empty or near-empty, the effects of $x_{1,it}$ are estimated with greater uncertainty. In any case, the marginal effect appears to be close to zero. This is not surprising given the assumptions in constructing the covariate $x_{1,it}$, as aforementioned in section 4.2.3, and the presence of covariates $x_{2,i}$ and $x_{3,it}$ which have noticeably stronger effects. Ignoring components 3, 5, and 7, the posterior densities of $\beta_{3,k}$, $k = \{1, \ldots, 8\}$ shown in Figure 4.2n indicate a strong inverse relationship between time required to travel to a screening facility and service utilisation, as in the area-specific model.

Figures 4.2o-4.2p show the posterior means of the spatial and temporal random effects for each component in the common trend model respectively. For clarity, components 3 and 7 have been removed from Figure 4.2o. SLAs belonging to components with stronger temporal effects (e.g. components 1, 6, and 8) tend to have spatial random effects closer to zero, and vice versa. The variance of the estimates of the temporal effects for components 3 and 7 are noticeably larger since members of these clusters tend to correspond to unusual SLAs. Consequently, the extended BaySTDetect model is capable of identifying six common temporal trends. These six temporal trends are reproduced in Figure 4.3 to emphasise their differences.

k-means Analysis

Figure 4.4 shows a side-by-side comparison of the spatial composition of clusters determined

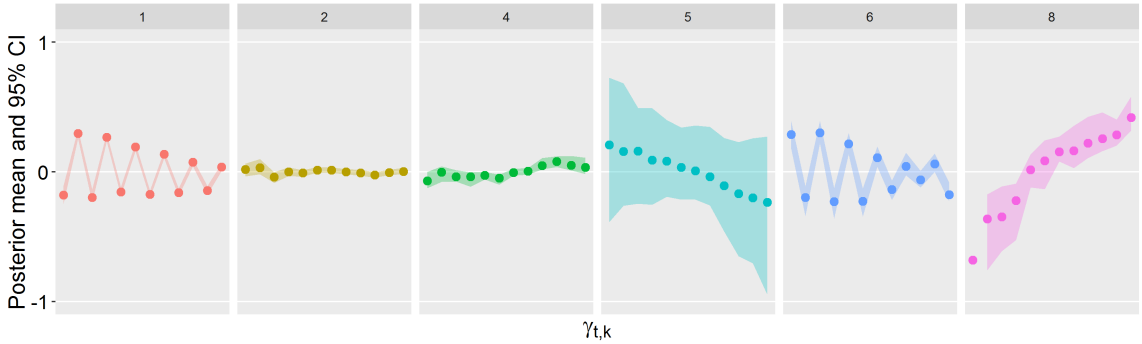


Figure 4.3: Posterior summary of the six common temporal trends estimated by the BaySTDetect model.

by the k-means analysis and the mixture components from the extended BaySTDetect model defined by Equation (4.16), for 2, 4, and 8 clusters/components. The k-means analyses cluster SLAs based on the posterior means of the space-time residuals, $\mathrm{E}(xi_{it}|y_{it})$, while the clustering resulting from the common trend mixture model is determined by similarities of random effects and covariates. The spatial clustering exhibited by the k-means clustering suggests that the area-specific, unmodified BaySTDetect model was lacking spatial information, while the spatial clustering exhibited by the mixture allocations indicate similarities amongst SLAs which can be attributed to a combination of random effects estimates and covariate information. While there exist some similarities between the two types of clustering, there are substantive differences, especially when the number of mixture components is 2 or 4, which validate the usefulness of the finite mixture.

4.3.2 Space-Time Mixture Model

Figure 4.5 summarises the marginal posterior distributions of the main parameters in the space-time model. The posterior densities for the effects of $x_{1,it}$ and $x_{3,it}$ on service utilisation are shown in Figure 4.5a and Figure 4.5b respectively. The marginal effect of $x_{1,it}$ is comparable to that in the BaySTDetect model. Surprisingly, the marginal effect of $x_{3,it}$ is also positive, which suggests that women who have a shorter travel time to a screening facility are less likely to use the mammography screening services offered by BreastScreen. We return to this in the Discussion section.

Figures 4.5c-4.5d show the posterior means of the spatial and temporal random effects for each SLA respectively. In Figure 4.5c, the SLAs are in ascending order of the average population at risk. The spatial random effects are very similar to those estimated by the unmodified BaySTDetect model (Duncan et al. 2016). The temporal random effects are also similar, but the differences in effect size between earlier and later years are larger, which may partly explain the instability of the area-specific trends.

Figures 4.5e-4.5f show the posterior density for the effect of $x_{2,i}$ and the posterior means of the multinomial probability q_i respectively. The positive marginal effect of $x_{2,i}$ results in the posterior means being around 0.75 for each SLA. While the covariate information on socioeconomic status contributes to the model, it seems that differences between the 158 SLAs are too minor to warrant area-specific estimates.

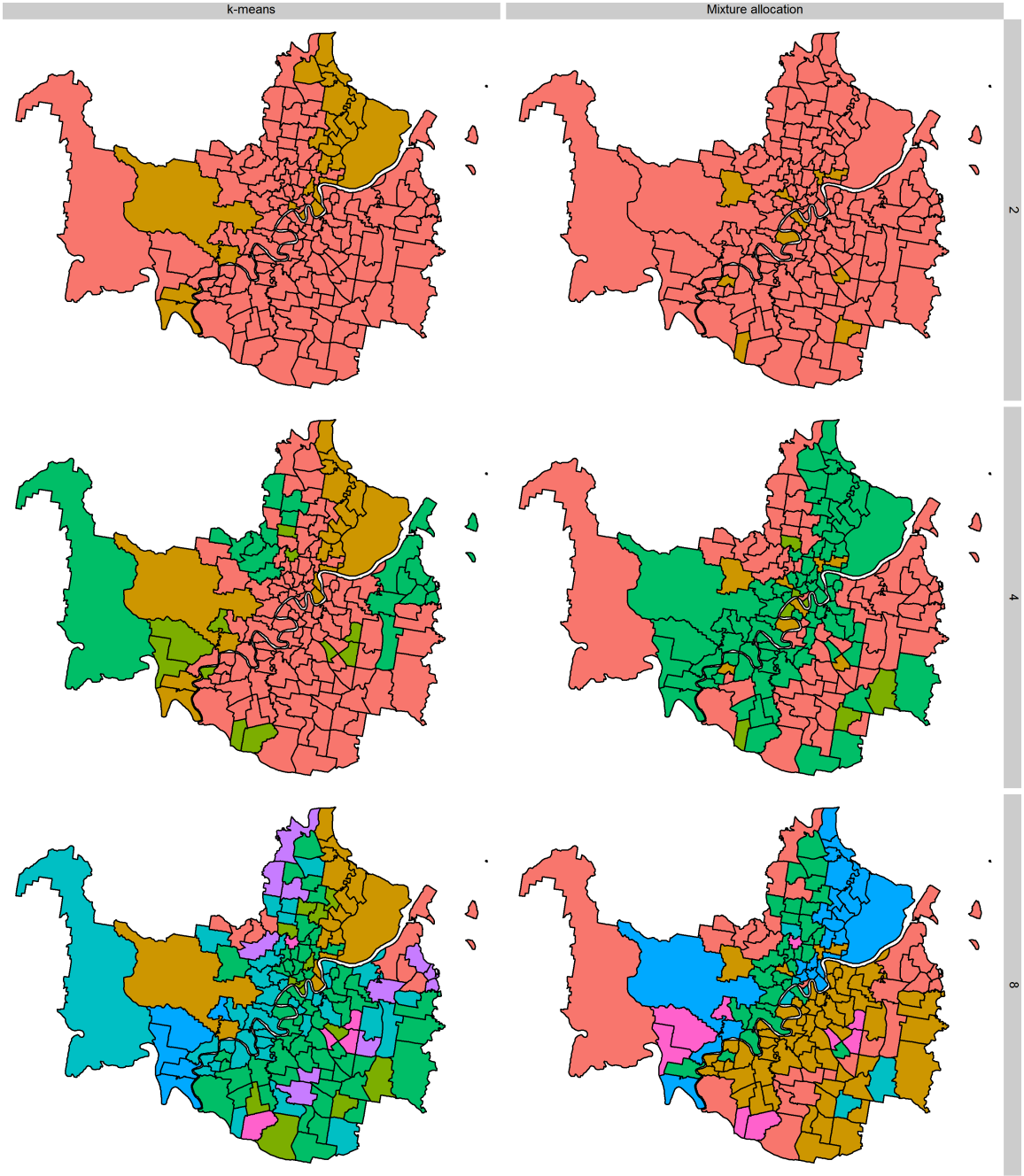


Figure 4.4: The spatial composition of clusters as determined by the k-means analysis and the extended BaySTDetect model for 2, 4, and 8 clusters/components.

The posterior means of the space-time interaction effects ν_{it}, for six selected SLAs, are shown in Figures 4.5g-4.5l. These estimates are similar to those obtained from the unmodified BaySTDetect model. The corresponding mixture allocations, however, are in many cases very different. The posterior means of the mixture allocations, shown in Figures 4.5m-4.5r, were larger on average, indicating that greater temporal instability is being detected by this model.

4.3.3 Aberrant Temporal Trends

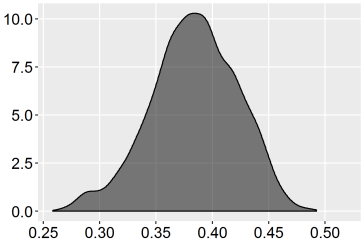
The aims of the BaySTDetect and space-time mixture models are to identify unusual and unstable temporal trends respectively. Figure 4.6 shows the degree to which the estimated temporal trends are unusual, based on $\mathrm{E}(p_i|y_{it})$. Figure 4.7 shows the degree to which temporal trends are unstable, using Rules 3 and 4 described in Section 4.2.

Based on the original BaySTDetect model, the analysis of Duncan et al. (2016) found that there appeared to be some correlation between SLAs with unusual trends and SLAs with unstable trends. This still largely appears to be the case, with notable exceptions in the northeast and southeast regions of Brisbane. The inclusion of additional covariate information and the mixture model to account for multiple common trends has decreased the number of SLAs with unusual temporal trends, but increased the number of SLAs with unstable temporal trends. Thus SLAs which are allocated to one of the eight common temporal trends, which may previously have been considered to be unusual, may still have unstable trends according to the space-time model.

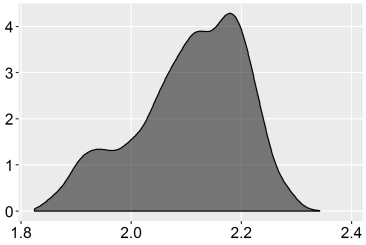
4.4 Discussion

The methodology presented in this paper relies on several assumptions and posed several difficulties. These are now discussed.

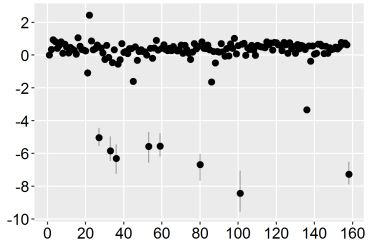
The covariate for area-level socioeconomic status was computed as the SLA IRSAD score. IRSAD is one of four socioeconomic indexes released by the ABS. Other indexes may produce different results. The scores are based on the 2011 national census, three years after the study period ends. While IRSAD scores from 2006 census data were also available, the 2011 scores were preferred since they were computed with more up-to-date methods. Both sets of scores are similar in value and were recorded near the end of the study period, so the effect of this choice is likely to be negligible. The Australian Bureau of Statistics (ABS) (2013b) suggests



(a) Posterior density for b_1

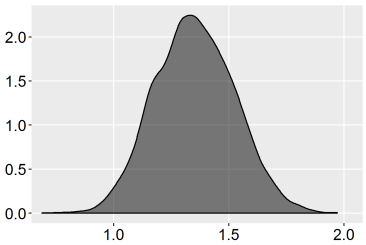


(b) Posterior density for b_3

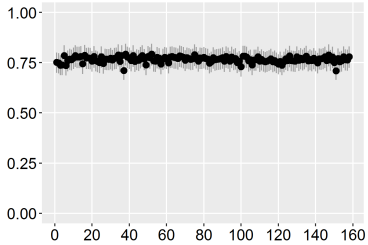


(c) Posterior mean and 95% CI for λ_i

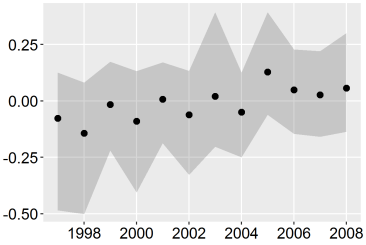
(d) Posterior mean and 95% CI for ψ_t



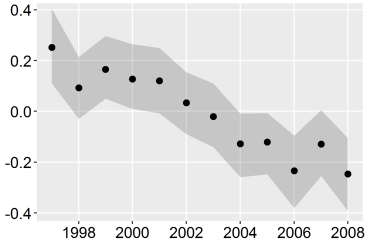
(e) Posterior density for b_2



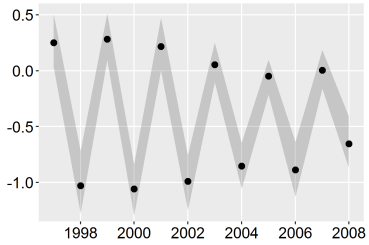
(f) Posterior mean and 95% CI for q_i



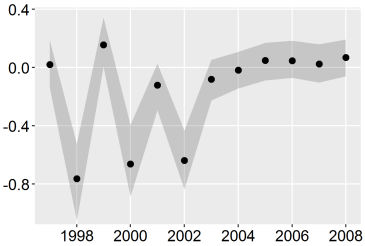
(g) Posterior mean and 95% CI for ν_{10t}



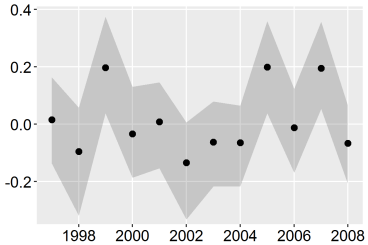
(h) Posterior mean and 95% CI for ν_{22t}



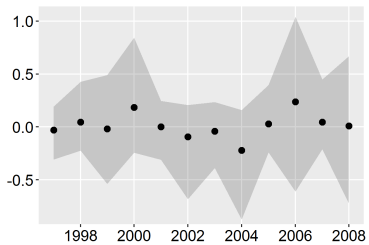
(i) Posterior mean and 95% CI for ν_{57t}



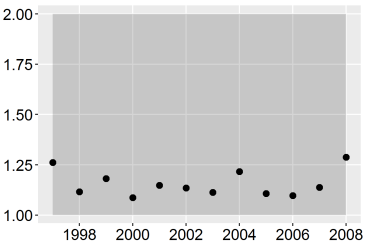
(j) Posterior mean and 95% CI for ν_{68t}



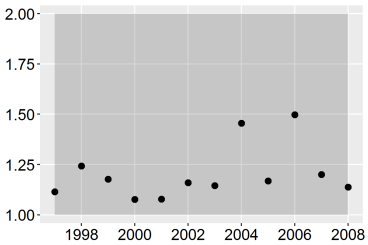
(k) Posterior mean and 95% CI for ν_{92t}



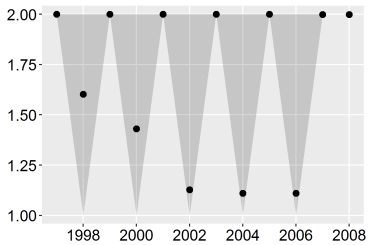
(l) Posterior mean and 95% CI for ν_{158t}



(m) Posterior mean and 95% CI for z_{10t}

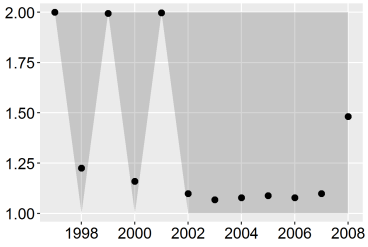


(n) Posterior mean and 95% CI for z_{22t}

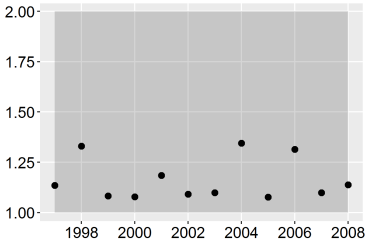


(o) Posterior mean and 95% CI for z_{57t}

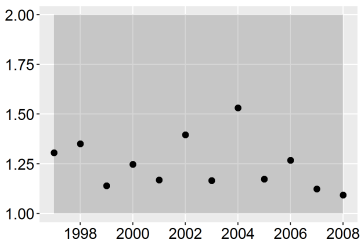
Figure 4.5: Posterior summary of the main model parameters for 1 chain of the space-time mixture model.



(p) Posterior mean and 95% CI for z_{68t}



(q) Posterior mean and 95% CI for z_{92t}



(r) Posterior mean and 95% CI for z_{158t}

Figure 4.5: Posterior summary of the main model parameters for 1 chain of the space-time mixture model (cont.).

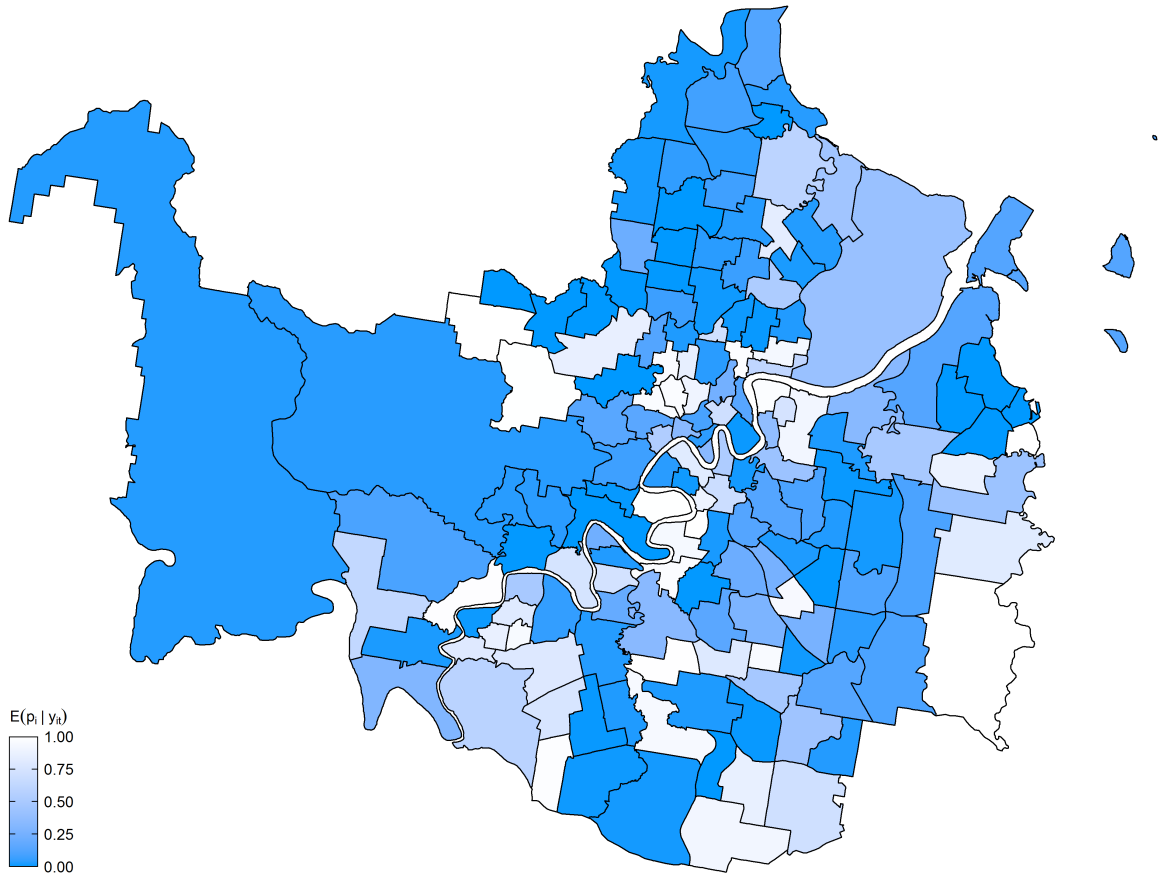


Figure 4.6: Map of SLAs in the Brisbane region (Moreton Island not shown) representing the degree to which SLAs follow the common temporal trend (lighter regions) or exhibit unusual temporal trends (darker regions).

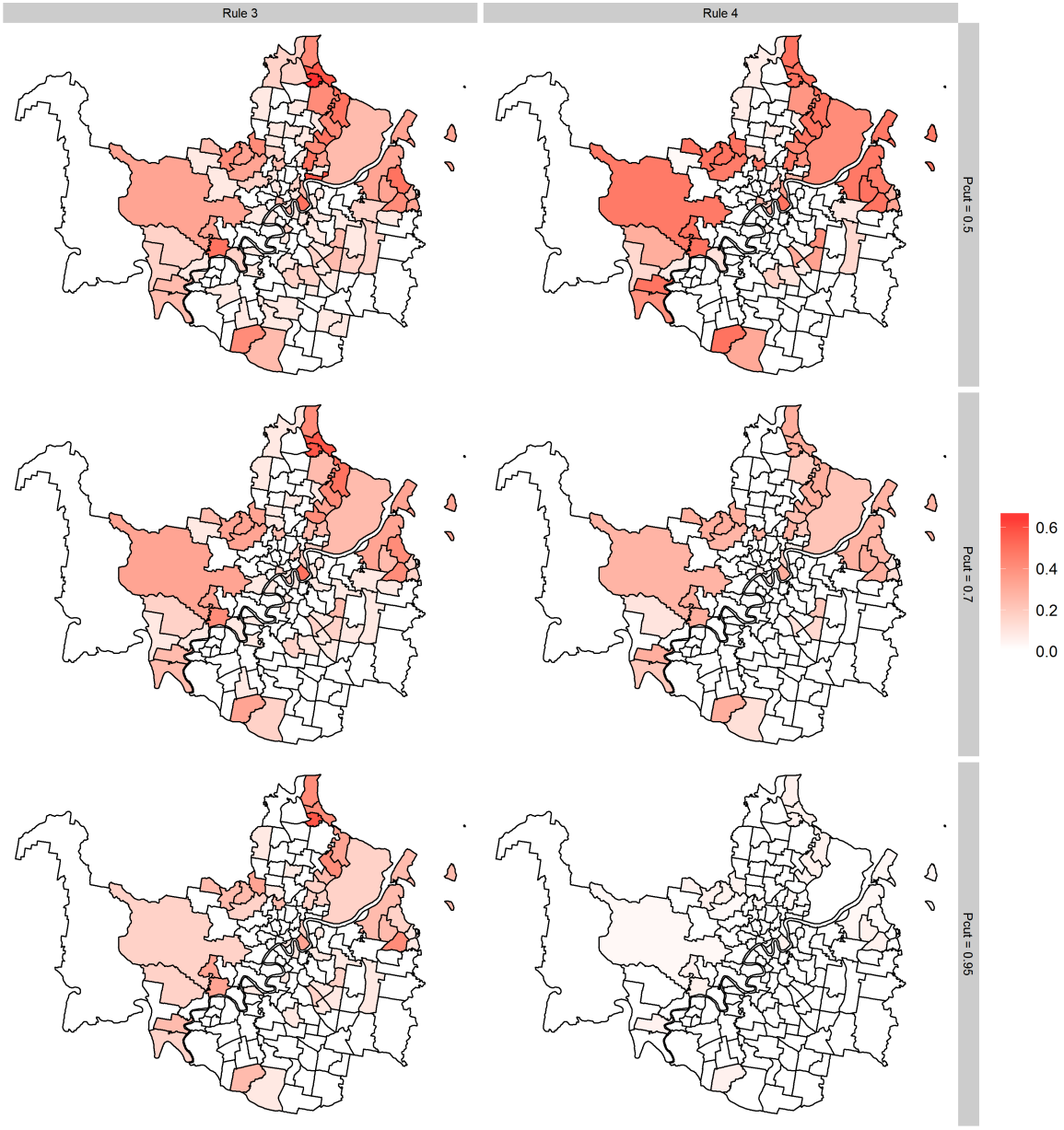


Figure 4.7: Map of SLAs in the Brisbane region (Moreton Island not shown) representing the degree to which SLAs exhibit temporal trends which are unstable, as determined by Rules 3 and 4 described in Section 4.2.

treating IRSAD as an ordinal measure by ranking the indexes and using quantiles to group areas of different socioeconomic status. However, we found that using standardised raw scores produced better estimates.

Calculation of the covariate for the shortest travel time to a screening facility required numerous assumptions. First, it was assumed that women elected to travel to the mammography screening facility nearest to the centroid of the SLA in which they lived as determined by the travel duration. In reality, women may have chosen to travel to a facility which was closer in terms of distance, or a facility with which they were familiar albeit much further, for example. Furthermore, the nearest facility in a given year may have been mobile and only available for a fraction of that year. A woman who missed the opportunity to access such a facility may have had to travel much further to find the next nearest facility. Travel time was chosen over travel distance as it seemed the more realistic decision factor that would be considered by women. Second, the mode of travel specified in the API was 'driving', which uses standard road networks to calculate the travel distance and duration (Google 2015). This assumes that women both have and use private transportation to get to and from a mammography appointment. Travel distances and times for public transport are also possible. However, these estimates rely on specification of time of departure during the day to be consistent, which then effects service availability, as some bus routes, for example, decrease frequency of services during off-peak times. Thus private transport times were chosen for simplicity and consistency. Third, SLA centroids were calculated as the geometric centre of the 'bounding box' of each SLA. For some irregularly shaped SLAs, this may not be a good representation of the centre, and women may not live near the SLA centroid anyway. This can lead to inaccurate estimates of travel time. However, this approach was deemed to be the least biased. The main limitation in calculating this covariate was due to the missing geocode information for some screening facilities. Our response was to use approximate street addresses to determine a geocode, or for mobile screening facilities which were typically named according to the suburb in which they were located, the geocode of the suburb centroid was used.

Many developed countries offer no- or low-fee mammography screening services to the public by inviting the population at risk to attend regular screening, administered by an organised screening program. Opportunistic screening may also be available to individuals by request or a doctor's referral. In order to determine if opportunistic screening could account for some of the variation in the number of observed visits to the publicly funded mammography screening

facilities, attempts were made to obtain information on the location and operating timeframe of private screening facilities in the Brisbane City Council region. Internet searches revealed a total of 61 private screening facilities, of which only 15 could be confirmed as offering mammography screening services. These 15 facilities were operated by 5 different companies. Each company was telephoned to solicit information about the dates of operation and confirm the location of each facility during the study period. This procedure revealed one additional facility not identified from the internet search. Out of these 16 facilities, four began operation post 2008, four were operating for the full study period, although accurate details of their operation times and locations could not be ascertained, and the outcome of eight facilities could not be ascertained despite several attempts to solicit this information. Based on this limited information, a covariate for the relative availability of services for the four operating facilities was computed using the same procedure used to construct $x_{1,it}$. However, due to the poor quality of this data, it was not possible to estimate the effect of this covariate reliably, especially in the presence of the other three covariates. However, inclusion of such data is recommended if available.

Covariate information on the rurality and remoteness of each SLA was obtainable, but not considered to be useful in this study because every SLA in the Brisbane City Council region is comparably urban compared to the remainder of Queensland. Furthermore, whatever spatial heterogeneity may exist due to rurality was most likely accounted for by the other three covariates.

Regarding the first extension to the BaySTDetect model, choosing a large value for δ_i *a priori* like 0.95 essentially instilled a penalty for choosing the area-specific model. When relaxing this assumption, care must be taken to avoid the area-specific model dominating the model choice step. This can happen if the prior specification is too restrictive, e.g.

$$\begin{aligned} \text{logit}\left(\delta_i\right) &= \beta_2 x_{2,i} \\ \beta_2 &\sim \mathcal{N}\left(0, \sigma^2_{\beta_2}\right). \end{aligned}$$

Treating δ_i, $\alpha_{\delta,i}$, and β_2 stochastically by assigning them appropriate prior distributions creates enough flexibility to overcome this problem. (Compare with Equations (4.11)-(4.13).)

Modifying the BaySTDetect model to allow for multiple common trends alleviates a concern

raised by Li et al. (2012), namely that indexing the model indicator p_i by space only, and not time, may be inappropriate when the number of time periods is large since each time point provides an opportunity for each estimated temporal trend to change, and therefore temporal trends are unlikely to remain consistent for the whole period. By allowing the model to choose from a number of common temporal trends, this concern is lessened, as such inconsistencies may instead become characteristics of a particular common trend.

In Li et al. (2012), Abellan et al. (2008), and Duncan et al. (2016), the elements w_{ij} of the spatial weights matrix $\mathbf{W}$ were defined as $w_{ij} = 1$ if the i^{th} and j^{th} areas are neighbours, and zero otherwise. However, the analysis of Duncan et al. (2016) indicated that the spatial autocorrelation between observations in this data set is global rather than local. In an attempt to improve estimation of whatever autocorrelation may exist, weights for higher order neighbours were considered, e.g.

$$w_{ij} = \begin{cases} \omega_h & \text{if the } i^{\text{th}} \text{ and } j^{\text{th}} \text{ areas are } h^{\text{th}}\text{-order neighbours} \\ 0 & \text{otherwise} \end{cases}.$$

where $\omega_1 > \cdots > \omega_H$ for $H \geq 2$. Defining the weights by radial proximity of SLA centroids rather than adjacency was also trialled in our analysis. However, neither alternate approach to definition of the weights had any noticeable effect on estimation of the posterior or predictive performance of the models.

The surprising result from section 4.3.2 warrants further discussion. The marginal effect for travel time was found to be positive in the space-time mixture model, suggesting that participation in mammography screening services offered by BreastScreen decreases as the travel time to the nearest screening facility decreases. This conflicts with the results obtained from the BaySTDetect model and highlights the importance of model specification and resultant inferences. There is a range of potential reasons why these results might be observed. For example, the parameters are estimated in the context of the other terms in the model, so the negative trend may be compensating for more positive contributions in the temporal trend or the complex interactions between time and space, in order to provide a good fit. A possible solution to this is to impose constraints in the prior, which we have not done here. The result also highlights the need to obtain better travel time data and information about relevant covariates, as well as gain a better understanding about why women make choices about if and where to go

for screening.

4.5 Conclusion

This paper has demonstrated several extensions to the models originally proposed by Li et al. (2012) and Abellan et al. (2008), and subsequently adapted by Duncan et al. (2016)). These extensions bring new insights into the nature of temporal patterns in mammography screening utilisation in Brisbane, and more generally, demonstrate the usefulness of these Bayesian spatio-temporal models as effective analytical tools in identifying "hot spots" and areas possibly in need of intervention.

Statement of Contribution of Co-authors for Chapter 5

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

This paper was submitted to the *Journal of the American Statistical Association* for publication and is currently under review. The title of this paper is *New and existing solutions to the label switching problem in Bayesian mixture models: a systematic review*.

Contributor	**Statement of contribution**
E. W. Duncan	Proposed and conducted the research, wrote the code for simulation study, developed the proposed algorithm, wrote the manuscript, revised the manuscript as suggested by co-author.
Signature and date:	
K. L. Mengersen	Supervised research, provided comments on and helped revise manuscript.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

Chapter 5

New and existing solutions to the label switching problem in Bayesian mixture models: a systematic review

---

Preamble

This chapter relates to the third research objective. The research presented in this chapter deviates from the application of spatio-temporal models to focus on the problem of label switching. Label switching is a problem which can affect inference from Bayesian mixture models, and since mixture models are a vital aspect of the methodology considered in this thesis, and commonly used in the wider literature, it is a research problem worth investigating.

This chapter describes label switching, and explains how it occurs, why it is considered a problem, and how it can be solved. Several misconceptions about label switching are also addressed. In brief, the solution to the label switching problem involves applying a relabelling algorithm designed to reverse the effects of the labels that have been switched. Relabelling algorithms are not always perfect, however, and their accuracy as well as computational efficiency can vary quite substantially. Twelve relabelling algorithms from the literature are reviewed in a systematic manner. A new algorithm is also proposed which ambitiously seeks to solve the label switching problem in an accurate and efficient manner.

The chapter concludes with a simulation study which compares all 13 algorithms, assessing their accuracy, computational efficiency, and robustness.

Aside from using y to denote a generic observation and general conventions, such as using $p(\cdot)$ to denote a probability distribution, the notation in this chapter bears no connection to the notation of the previous chapters. This chapter should be treated as a standalone chapter.

Some of the main results from this chapter were presented at two international conferences: the Bayes on the Beach Conference, Gold Coast, Australia, 07-09 December 2015, and the International Society for Bayesian Analysis, Sardinia, Italy, 13-17 June 2016.

5.1 Introduction

Mixture models comprise a number of component densities which can be used to approximate non-standard densities or model heterogeneous data, that is data with underlying subpopulations (Jasra et al. 2005; Marin et al. 2005; Papastamoulis 2013; Richardson and Green 1997; Rodriguez and Walker 2014; Sperrin et al. 2010; van Havre et al. 2015).

Given observed data $\boldsymbol{y} = (y_1, \ldots, y_N)$, the likelihood for a K-component mixture model is typically expressed as:

$$\boldsymbol{Y} \sim p(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\lambda}) = \prod_{i=1}^{N} \sum_{k=1}^{K} w_k f_k(y_i|\boldsymbol{\phi}_k, \boldsymbol{\lambda}) \tag{5.1}$$

where $\boldsymbol{\phi}_k$ and $\boldsymbol{\lambda}$ denote the unknown component-specific and common parameter(s) relating to the k^{th} mixture component respectively, and $f_k(\cdot)$ is the k^{th} component density with corresponding mixture weight w_k which satisfies the conditions $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0$ for $k = 1, \ldots, K$ (Celeux et al. 2000; Marin and Robert 2007; Marin et al. 2005; Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014; Stephens 2000a,b; van Havre et al. 2015). Note that the component densities may be of different distributional types (Marin et al. 2005). Denote the set of R component-specific parameters, which include the mixture weights, by $\boldsymbol{\theta} = \{\boldsymbol{\phi}_{1,(1,\ldots,K)}, \ldots, \boldsymbol{\phi}_{R-1,(1,\ldots,K)}, \boldsymbol{w}\}$.

It is convenient to associate with the random variable Y_i a latent allocation variable Z_i with realisation z_i such that

$$Y_i|z_i, \boldsymbol{\phi}, \boldsymbol{\lambda} \sim f_{z_i}\left(y_i|\boldsymbol{\phi}_{z_i}, \boldsymbol{\lambda}\right)$$
$$Z_i|\boldsymbol{w} \sim \text{Cat}\left(w_1, \ldots, w_K\right),$$

thereby identifying to which component each observation y_i belongs. Here $\text{Cat}(\cdot)$ denotes the categorical distribution, equivalent to the multinomial distribution with a single observation (Marin and Robert 2007; Marin et al. 2005; Pan et al. 2015; Papastamoulis and Iliopoulos 2010; Richardson and Green 1997; Rodriguez and Walker 2014; Sperrin et al. 2010; van Havre et al. 2015). The allocation variables $\boldsymbol{Z} = \{Z_1, \ldots, Z_N\}$ are unobserved and are therefore considered missing data (Papastamoulis and Iliopoulos 2010, 2013).

The likelihood given by Equation 5.1 is exchangeable, meaning that it is invariant to permutations of the labels identifying the mixture components. That is,

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\lambda}) = p(\boldsymbol{y}|\tau(\boldsymbol{\theta}), \boldsymbol{\lambda}) \tag{5.2}$$

for any permutation τ (Marin et al. 2005; Papastamoulis and Iliopoulos 2010; Sperrin et al. 2010; Yao 2013). If the prior is also exchangeable, then the posterior distribution will be invariant to permutations of the labels a phenomenon known as label switching (Celeux et al. 2000; Marin and Robert 2007; Marin et al. 2005; Pan et al. 2015; Papastamoulis and Iliopoulos 2010, 2013; Puolamäki and Kaski 2009; Rodriguez and Walker 2014; Rossi et al. 2005; Stephens 2000b; van Havre et al. 2015; Yao 2013). Geometrically, the impact of label switching is a posterior distribution which is symmetric in K dimensions, resulting in $K!$ symmetric modes. For example, if $\left(\hat{\theta}_1, \ldots, \hat{\theta}_K\right)$ is a local maximum, then $\left(\hat{\theta}_{\tau(1)}, \ldots, \hat{\theta}_{\tau(K)}\right)$ is as well for any permutation τ (Marin et al. 2005; Rossi et al. 2005; Stephens 2000b; Yao 2013).

The practical implications of label switching depend on the purpose of the mixture model. If the purpose is to approximate a non-standard density, then the role of $\boldsymbol{Z}$ is simply to facilitate estimation of the other parameters of interest, and the occurrence of label switching is inconsequential (Pauli and Torelli 2015; Rossi et al. 2005). Conversely, if estimation of the mixture components or component membership of the observations is important, then the estimation of both $\boldsymbol{\theta}$ and $\boldsymbol{Z}$ is of interest (Marin and Robert 2007; Marin et al. 2005; Puolamäki and Kaski 2009; Sperrin et al. 2010). In this case, label switching makes inference untenable – the marginal posterior distributions will look near identical, and thus attempting to summarise the parameters for each component through statistics like the posterior mean will be meaningless (Celeux et al. 2000; Jasra et al. 2005; Marin et al. 2005; Papastamoulis 2013; Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014; Sperrin et al. 2010; Stephens 2000b; van Havre et al. 2015; Yao 2013). Consequently, in order to make inferences about the components, it is

necessary to first reverse the label switching by applying a relabelling algorithm (Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014; Sperrin et al. 2010; Yao 2013). For the remainder of this paper, only the latter case is considered.

The aims of this paper are threefold. First, we review existing relabelling algorithms from the literature in a systematic manner. To our knowledge, this is the first time that these methods have been presented using unified, consistent definitions and notation, which facilitates straightforward comparison. Second, we introduce a new relabelling algorithm which combines different methodology to offer an algorithm that is both effective and computationally attractive. Third, we present a simulation study in which all the algorithms are tested for computational efficiency, their ability to relabel the parameters accurately, and their robustness to misspecification of the number of components.

The remainder of this paper is structured as follows. Section 5.2 discusses label switching in more depth, in particular the necessary conditions in order for label switching to occur, and the extent to which label switching is considered a problem for inference. Section 5.3 presents a general framework for solutions to the label switching problem. In section 5.4, a systematic review of existing relabelling algorithms is presented, followed by the introduction of a new algorithm. Section 5.5 describes the methodology of the simulation study, and the results are subsequently presented in section 5.6. This paper concludes with a discussion in section 5.7.

5.2 Label Switching

5.2.1 The Occurrence of Label Switching

For a mixture with K components there are $K!$ permutations which satisfy Equation (5.2). Each of these permutations corresponds to one of the symmetric modes of the posterior. It is often stated in the literature that a K-component mixture induces $K!$ symmetric modes (Celeux et al. 2000; Jasra et al. 2005; Stephens 2000b), but a mixture model may induce fewer than $K!$ modes, and even fewer still may be observed (Jasra et al. 2005; Rodriguez and Walker 2014; Rossi et al. 2005).

The occurrence of label switching depends on the degree of exchangeability. Informally, two

distributions are exchangeable if they are identical, partially exchangeable if they are not identical but intersect each other, and not exchangeable if they are disjoint (Diaconis and Freedman 1984). If the prior is not exchangeable, then label switching will not occur, and poses no concern. If the prior is fully exchangeable (i.e. the marginal prior distributions are identical for all components), then label switching should occur with asymptotically equal frequencies of each permutation. That is,

$$\Pr\left(\tau = \{S\}_{j\in\{1,\ldots,K!\}}\right) \to \frac{1}{K!} \tag{5.3}$$

where S is the set of $K!$ permutations and consequently

$$\Pr\left(Z_i = k|\boldsymbol{y}\right) \to \frac{1}{K} \tag{5.4}$$

as the number of Markov Chain Monte Carlo (MCMC) iterations tends to infinity (Pan et al. 2015; Papastamoulis and Iliopoulos 2010; Pauli and Torelli 2015; Puolamäki and Kaski 2009; Sperrin et al. 2010; Stephens 2000b). If the prior is partially exchangeable, then this may affect the frequency of the permutations and may restrict label switching to a subset of the K components (Celeux et al. 2000).

If $K!$ symmetric modes exist, the number of symmetric modes actually observed in a posterior sample may be fewer. This may occur because the sampler is unable to traverse the posterior surface easily due to the wide separation between the symmetric modes (Celeux et al. 2000; Jasra et al. 2005; Marin and Robert 2007; Papastamoulis and Iliopoulos 2010; Pauli and Torelli 2015; Robert, C. 2014; Rodriguez and Walker 2014; Rossi et al. 2005; van Havre et al. 2015).

5.2.2 The Extent of the Problem of Label Switching

As aforementioned, label switching poses a problem for meaningful inference. Despite this difficulty, the presence of label switching is actually desirable, even necessitated, as it indicates that the sampler has fully explored the posterior surface (Celeux et al. 2000; Jasra et al. 2005; Marin and Robert 2007; Marin et al. 2005; Papastamoulis and Iliopoulos 2010, 2013; Rodriguez and Walker 2014; van Havre et al. 2015).

A common deduction then is that an absence of label switching implies a lack of convergence (Jasra et al. 2005). However, this is not necessarily so (Rossi et al. 2005). A sampler which is

inefficient with respect to the labelling but otherwise efficient, such as the Gibbs sampler, will typically converge to one of the symmetric modes, thereby avoiding any problems associated with label switching (Celeux et al. 2000; Pauli and Torelli 2015; Rodriguez and Walker 2014). Hence the Gibbs sampler can be regarded as a solution to the label switching problem (Pauli and Torelli 2015; Robert, C. 2014). The real concern with an absence of label switching is that the sampler may have missed genuine, non-symmetrical modes, and not just the symmetric copies (Robert, C. 2014; Rodriguez and Walker 2014; van Havre et al. 2015).

To this end, difficulties in traversing the posterior surface might be alleviated by 'flattening' the peaks through the use of tempering (Celeux et al. 2000; Jasra et al. 2005; Marin and Robert 2007; van Havre et al. 2015), or by using a specialised MCMC sampler (Rodriguez and Walker 2014), or even a hybrid of samplers (Marin and Robert 2007). Modifying a sampler to simply force permutations of the labels, however, is pointless – it does not increase the confidence one may have about convergence, and worse, it may obscure far more serious problems with the sampler that are unrelated to label switching (Celeux et al. 2000; Robert, C. 2014). Additionally, it may have the adverse effect of analytically constraining the posterior space.

When the posterior is estimated using a MCMC sampler, label switching may occur at each iteration, according to the degree of exchangeability of the prior discussed above (Sperrin et al. 2010). However, a change in the labels between iterations might not necessarily indicate label switching. The labels are unknown quantities which must be estimated, and thus the labels for some observations may change between iterations due to overlap of the component densities, i.e. when one or more component densities have similar parameters. Moreover, it is not enough to identify the occurrence of label switching, but rather the permutation that caused it, such that the inverse permutation will successfully undo the label switching (Sperrin et al. 2010). This is a difficulty which increases with the number of components (Stephens 2000b). If a relabelling algorithm can correctly identify the inverse permutations necessary to undo the label switching, however, then the problem that label switching bears on inference is dispelled.

5.3 General Solutions to the Label Switching Problem

The earliest proposal to address the label switching problem involved imposing an artificial identifiability constraint (IC) on the component-specific parameters in an attempt to break the

symmetry of the prior, and thus of the posterior. For example, if the component densities in Equation (5.1) were Gaussian,

$$f_k\left(y_i|\boldsymbol{\phi}_k\right) = \mathcal{N}\left(y_i|\mu_k, \sigma_k^2\right) \qquad \text{for } k = 1, \ldots, K$$

then $\mu_1 < \mu_2 < \cdots < \mu_K$ or $\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_K^2$ are examples of such a constraint (Marin and Robert 2007; Papastamoulis 2016; Papastamoulis and Iliopoulos 2010, 2013; Rodriguez and Walker 2014; Sperrin et al. 2010; Stephens 2000b; Yao 2013). These ICs are equivalent to truncating the prior distribution, $p(\boldsymbol{w}, \boldsymbol{\phi})$, resulting in $p(\boldsymbol{w}, \boldsymbol{\phi})\mathbb{I}\left(\mu_1 < \cdots < \mu_K\right)$ or $p(\boldsymbol{w}, \boldsymbol{\phi})\mathbb{I}\left(\sigma_1^2 < \cdots < \sigma_K^2\right)$ where $\mathbb{I}$ is the indicator function, such that only one permutation of the parameters satisfies the chosen IC (Marin and Robert 2007; Sperrin et al. 2010). However, the logic of this method is flawed for two main reasons. First, choosing a suitable IC is not straightforward, and many choices will not guarantee removal of the symmetry in the posterior distribution. This is especially true for multivariate mixture models, models with many parameters, and when components are poorly separated (Celeux et al. 2000; Jasra et al. 2005; Pan et al. 2015; Richardson and Green 1997; Rodriguez and Walker 2014; Rossi et al. 2005; Stephens 2000b). Second, imposing an IC can have a large influence on the shape of the posterior distribution. If the basis of using an exchangeable prior is one of non-informativeness, then the impact of the IC on the posterior is counterproductive and indicates a contradiction of prior beliefs Celeux et al. (2000); Jasra et al. (2005); Marin and Robert (2007); Marin et al. (2005); Richardson and Green (1997); Sperrin et al. (2010); Stephens (2000b). It has been suggested that this latter concern can be alleviated by sampling from the unconstrained posterior and imposing the IC ex-post, that is, after the posterior sample has been obtained (Celeux et al. 2000; Marin et al. 2005; Richardson and Green 1997; Rossi et al. 2005). However, the former issue remains.

Other approaches are based on decision theory. These approaches make no attempt to prohibit label switching, but rather focus on identifying the permutations necessary to reverse it. Many of these approaches can be viewed as particular examples of a more general framework in which the goal is to minimise the posterior expectation of some loss function (Rodriguez and Walker 2014; Stephens 2000b). Formally, the objective is to choose an action a from a set of actions $\mathcal{A}$, where the loss incurred for choosing action a is expressed by the loss function $\mathcal{L}(a; \boldsymbol{\theta}, \boldsymbol{z})$, such

that the risk

$$\mathcal{R}(a) = \mathrm{E}\left[\mathcal{L}(a; \boldsymbol{\theta}, \boldsymbol{z})|\boldsymbol{y}\right]$$

is minimised (Stephens 2000b). Both $\boldsymbol{\theta}$ and $\boldsymbol{z}$ have been included in the expression for $\mathcal{L}$ to indicate that it may be a function of the parameters and/or latent labels, or some other quantity of interest derived from them. If $\boldsymbol{\theta}^{(m)} = \left\{\boldsymbol{\theta}_1^{(m)}, \ldots, \boldsymbol{\theta}_K^{(m)}\right\}$ and $\boldsymbol{z}^{(m)} = \left\{\boldsymbol{z}_1^{(m)}, \ldots, \boldsymbol{z}_N^{(m)}\right\}$ denote the estimates of the parameters and latent allocation vector for the m^{th} iteration of an MCMC sample, for $m = 1, \ldots, M$, then the risk $\mathcal{R}(a)$ can be approximated by the Monte Carlo risk

$$\hat{\mathcal{R}}(a) = \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}\left(a; \boldsymbol{\theta}^{(m)}, \boldsymbol{z}^{(m)}\right) \tag{5.5}$$

Since the likelihood is invariant to permutations of the parameters, the loss function should also be permutation invariant, that is

$$\mathcal{L}(a; \boldsymbol{\theta}, \boldsymbol{z}) = \mathcal{L}\left(a; \tau(\boldsymbol{\theta}), \tau^{-1}(\boldsymbol{z})\right) \tag{5.6}$$

for any $\tau \in S$ (Jasra et al. 2005; Stephens 2000b).

Here the function $\tau(\cdot)$ can be regarded as a generic permutation function which either permutes or relabels, depending on the supplied argument. To clarify, let $\tau = (\tau_1, \ldots, \tau_K)$ be a permutation of the index set $\{1, \ldots, K\}$, let $\boldsymbol{v} = (v_1, \ldots, v_K)$ be an arbitrary K-length vector, and let $\boldsymbol{z} = (z_1, z_2, z_3, \ldots)$ be an arbitrary length vector (or possibly scalar) containing only the values in S. Then

$$\tau(v_1, \ldots, v_K) = (v_{\tau_1}, \ldots, v_{\tau_K})$$

reorders the values of the vector v according to the permutation τ, and

$$\tau(z_1, z_2, z_3, \ldots) = (\tau_{z_1}, \tau_{z_2}, \tau_{z_3}, \ldots)$$

relabels the values of $\boldsymbol{z}$ (Jasra et al. 2005; Papastamoulis and Iliopoulos 2010; Puolamäki and Kaski 2009). However, note that if τ is the permutation required to reverse the effect of label switching on the parameters, then the corresponding latent allocation vectors must be relabelled according to the inverse permutation τ^{-1}.

5.3.1 Deterministic Relabelling Algorithms

If $\mathcal{L}_0(a;\boldsymbol{\theta},\boldsymbol{z})$ denotes a loss function which is not permutation invariant, then a permutation invariant loss function $\mathcal{L}$ can be obtained by defining

$$\mathcal{L}(a;\boldsymbol{\theta},\boldsymbol{z}) = \min_{\tau}\mathcal{L}_0\left(a;\tau(\boldsymbol{\theta}),\tau^{-1}(\boldsymbol{z})\right)$$

and thus the Monte Carlo risk (5.5) becomes

$$\hat{\mathcal{R}}(a) = \frac{1}{M}\sum_{m=1}^{M}\min_{\tau^{(m)}\in S}\mathcal{L}_0\Big(a;\tau^{(m)}\left(\boldsymbol{\theta}^{(m)}\right),(\tau^{-1})^{(m)}\left(\boldsymbol{z}^{(m)}\right)\Big) \tag{5.7}$$

(Pauli and Torelli 2015; Stephens 2000b). If the loss function $\mathcal{L}_0$ is of the form

$$\mathcal{L}_0(a;\boldsymbol{\theta},\boldsymbol{z}) = \sum_{k=1}^{K}\mathcal{L}_0\left(a;\boldsymbol{\theta}_k,\boldsymbol{z}(k)\right) \tag{5.8}$$

where $z(k)$ represents the vector z as some function of k, then the problem of minimising

$$\mathcal{L}_0\left(a;\tau(\boldsymbol{\theta}),\tau^{-1}(\boldsymbol{z})\right)$$

over all $\tau \in S$ is equivalent to minimising

$$\sum_{k=1}^{K} c_{\tau(k),k}$$

where

$$c_{j,k} = \mathcal{L}_0\left(a;\boldsymbol{\theta}_j,\boldsymbol{z}(j)\right)$$

is the cost of assigning the k^{th} element of τ the value j, i.e. $\tau_k = \tau(k) = j$. That is, the minimisation problem

$$\min_{\tau^{(m)}\in S}\mathcal{L}_0\Big(a;\tau^{(m)}\left(\boldsymbol{\theta}^{(m)}\right),(\tau^{-1})^{(m)}\left(\boldsymbol{z}^{(m)}\right)\Big)$$

is equivalent to the linear sum assignment problem (LSAP):

$$\min_{\tau^{(m)}\in S}\sum_{k=1}^{K} c_{\tau_k^{(m)},k} = \min_{b}\sum_{j=1}^{K}\sum_{k=1}^{K} b_{j,k}c_{j,k}^{(m)}$$

subject to

$$\sum_{j=1}^{K} b_{j,k} = \sum_{k=1}^{K} b_{j,k} = 1 \qquad \text{and} \qquad b_{j,k} \in \{0, 1\}.$$

(Bukard et al. 2009; Rodriguez and Walker 2014; Stephens 2000b). To clarify, if $\hat{b}$ is the optimal matrix which solves the LSAP, then this is the matrix representation of the optimal permutation under $\mathcal{L}_0$, that is:

$$\tau_k^{(m)} = \underset{j=1,\ldots,K}{\arg} \hat{b}_{j,k} \quad \text{where} \quad \hat{b}_{j,k} = \min_b \sum_{j=1}^{K} \sum_{k=1}^{K} b_{j,k} c_{j,k}^{(m)}$$

(Rodriguez and Walker 2014). Note that $c_{j,k}$ may also represent gains rather than costs, where the goal is to maximise the LSAP over b. This is equivalent to multiplying $c_{j,k}$ by -1 and solving as a minimisation problem.

Relabelling algorithms of this form are guaranteed to reach a fixed point (Jasra et al. 2005; Stephens 2000b), and are therefore referred to as deterministic relabelling algorithms (Sperrin et al. 2010). However, they have the potential to converge to a local rather than a global optimal solution, so it may be necessary to run the algorithm several times with different initialisations to reach a global optimum (Sperrin et al. 2010; Stephens 2000b).

Although there are some alternatives (see Pauli and Torelli (2015), for example), many deterministic relabelling algorithms follow this structure where the goal is to minimise the risk given by Equation (5.7), and many of these can be expressed as a LSAP to maximise efficiency.

5.3.2 Probabilistic Relabelling Algorithms

If the proposed loss function is inherently permutation invariant, satisfying Equation (5.6), then the resulting relabelling algorithm is said to be a probabilistic relabelling algorithm. The idea of probabilistic relabelling was first introduced by Jasra et al. (2005) and developed extensively by Sperrin et al. (2010). In contrast to deterministic relabelling algorithms which are guaranteed to reach a fixed point solution, probabilistic relabelling algorithms incorporate uncertainty of the permutations into the estimation (Rodriguez and Walker 2014; Sperrin et al. 2010). Probabilistic algorithms are solved using the expectation–maximisation (EM) algorithm (Dempster et al. 1977): the expected values of each permutation probability are computed for each MCMC iteration given initial estimates of the parameters, and then the parameter estimates are updated

by averaging over all permutations and iterations (Papastamoulis 2013, 2016; Sperrin et al. 2010; Yao 2013).

5.4 Specific Relabelling Algorithms

In this section, 12 different relabelling algorithms found in the literature are briefly described. These are presented in chronological order, except for Algorithms 5.6 and 5.7 which are extensions of Algorithm 5.5. This section concludes with the proposal of a new algorithm.

The Kullback-Leibler Divergence Algorithm

In addition to providing a general framework for deterministic relabelling algorithms, Stephens (2000b) presents a specific example of such an algorithm. The loss function is defined as the Kullback-Leibler distance from $\hat{\mathbf{P}}$ to $\mathbf{P}$ where $\mathbf{P}$ is the $N \times K$ matrix whose elements p_{ik} represent the probability of the i^{th} observation belonging to the k^{th} component, that is,

$$p_{i,k} = \frac{w_k f_k\left(y_i|\phi_k, \lambda\right)}{\sum_{j=1}^{K} w_j f_j\left(y_i|\phi_j, \lambda\right)}, \tag{5.9}$$

and $\hat{\mathbf{P}}$ is an estimate of $\mathbf{P}$. Thus the loss function associated with action a at the m^{th} iteration of an MCMC sample is

$$\mathcal{L}_0\Big(a; \tau^{(m)}(\mathbf{P}^{(m)})\Big) = \sum_{i=1}^{N}\sum_{k=1}^{K} p^{(m)}_{i,\tau^{(m)}(k)} \log\left(\frac{p^{(m)}_{i,\tau^{(m)}(k)}}{\hat{p}_{i,k}}\right)$$

where in this context a is a specified relabelling (Pauli and Torelli 2015; Stephens 2000b). The overall loss $\sum_{m=1}^{M} \mathcal{L}_0\left(a; \tau^{(m)}\left(\mathbf{P}^{(\mathbf{m})}\right)\right)$ is minimised with respect to the action a when

$$\hat{p}_{i,k} = \frac{1}{M}\sum_{m=1}^{M} p^{(m)}_{i,\tau^{(m)}(k)}. \tag{5.10}$$

Algorithm 5.1: Kullback-Leibler (KL) divergence algorithm

Step 1: Initialise the $M \times K$ matrix of permutations $\boldsymbol{\mathcal{T}} = \left\{\tau^{(1)}, \ldots, \tau^{(m)}\right\}$. This is usually initialised so that $\tau^{(m)} = \{1, \ldots, K\}$ for all m.

Step 2: For $i = 1, \ldots, N$ and $k = 1, \ldots, K$, calculate

$$\hat{p}_{i,k} = \frac{1}{M} \sum_{m=1}^{M} p_{i,\tau^{(m)}(k)}^{(m)}.$$

Step 3: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP using costs

$$c_{j,k}^{(m)} = \sum_{i=1}^{N} p_{i,j}^{(m)} \log\left(\frac{p_{i,j}^{(m)}}{\hat{p}_{i,k}}\right).$$

Step 4: If an improvement in $\sum_{m=1}^{M} \hat{\mathcal{L}}_0^{(m)}$ has been achieved, return to step 2 and repeat, otherwise stop.

The Pivotal Reordering Algorithm

The deterministic algorithm proposed by Marin et al. (2005) is geometric-based, data-driven alternative to an IC. The idea is to choose the parameter estimates from one of the M MCMC samples as a pivot, and then apply to each of the M samples the permutation which minimises the Euclidean distance between the pivot and permuted parameter vectors. If the permutations are chosen correctly, the parameters comprising all the symmetric modes will be permuted to a single mode, thereby removing the symmetry of the posterior (Marin et al. 2005; Papastamoulis 2016).

Algorithm 5.2: pivotal reordering algorithm (PRA)

Step 1: Define the pivot $\boldsymbol{\theta}^* = \boldsymbol{\theta}^{(m^*)}$ where m^* is the iteration which corresponds to the Monte Carlo approximation of the *maximum a posteriori* (MAP) estimate of $\boldsymbol{\theta}$.

Step 2: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by maximising the scalar product

$$\tau^{(m)} = \underset{\tau \in S}{\operatorname{argmax}} \sum_{r=1}^{R} \sum_{k=1}^{K} \theta_{r,\tau_k}^{(m)} \theta_{r,k}^*.$$

(Note that maximising the scalar product $\tau\left(\boldsymbol{\theta}^{(m)}\right) \cdot \boldsymbol{\theta}^{*}$ is equivalent to minimising the Euclidean distance between $\tau\left(\boldsymbol{\theta}^{(m)}\right)$ and $\boldsymbol{\theta}^{*}$.)

Contrary to the literature (see Papastamoulis (2013) and Papastamoulis (2016), for example), the PRA algorithm is not condemned to computational inefficiency when K is large. The scalar product represents a gain function which satisfies Equation (5.8), and therefore the problem may be formulated as a LSAP where the costs are given by

$$c_{j,k}^{(m)} = -\sum_{r=1}^{R} \theta_{r,j}^{(m)} \theta_{r,k}^{*}.$$

The Bernoulli Mixture Algorithms

The latent allocation variables $\left\{z_i^{(m)}\right\}$ can alternatively be expressed in terms of binary indicator variables

$$\hat{z}_{i,k}^{(m)} = \begin{cases} 1 & \text{if} z_i^{(m)} = k \\ 0 & \text{otherwise} \end{cases}.$$

Let $\hat{\boldsymbol{Z}}$ be the $M \times N \times K$ array containing the posterior sample of these values. An estimate of the latent allocation variables $\boldsymbol{Z}$ or equivalently $\hat{\boldsymbol{Z}}$ is provided by the posterior sample. However, in the presence of label switching, the empirical distributions of each Z_i will be symmetric and virtually identical just like the component-specific parameters in the model (and thus summaries of $\hat{\boldsymbol{Z}}$ will be misleading).

The approach of Puolamäki and Kaski (2009) is to create a second mixture model to estimate the distribution of $\hat{\boldsymbol{Z}}$. In the original mixture model, $\hat{\boldsymbol{Z}}$ is treated as an unknown parameter, while in this second mixture model, $\hat{\boldsymbol{Z}}$ is viewed as data, where the posterior estimates $\hat{z}_{i,k}^{(m)}$ are used as observations (Puolamäki and Kaski 2009). By Equation (5.4), the mixture weights will all be equal to $1/K$, and thus the likelihood for $\hat{\boldsymbol{Z}}$ is given by the Bernoulli mixture model

$$p(\hat{\boldsymbol{z}}|\mathbf{Q}) = \prod_{m=1}^{M} \prod_{k=1}^{K} \sum_{k=1}^{K} \frac{1}{K} \prod_{i=1}^{N} q_{i,k}^{\hat{z}_{i,k}^{(m)}} \left(1 - q_{i,k}\right)^{1-\hat{z}_{i,k}^{(m)}} \tag{5.11}$$

where the unknown parameters $\mathbf{Q} = \{q_{i,k}\}$ represent

$$q_{i,k} = \Pr\left(Z_i = k|\boldsymbol{y}, \boldsymbol{\theta}\right).$$

In the limit as $M \to \infty$, these posterior probabilities will correspond to one of the symmetric modes. The likelihood of Equation (5.11) can be maximised using the following EM algorithm (Puolamäki and Kaski 2009).

Algorithm 5.3: Bernoulli mixture (BM) algorithm

Step 1: Initialise the $N \times K$ matrix $\mathbf{Q}$ with elements $0 \leq q_{i,k} \leq 1$ at random, or

$$q_{i,k} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(z_i^{(m)} = k\right).$$

Also choose a tolerance threshold t for determining convergence, e.g. $t = 0.01$.

Step 2 (E-step): For $m = 1, \ldots, M$, $j = 1, \ldots, K$, and $k = 1, \ldots, K$, compute

$$\gamma_{m,j,k} = \frac{\prod_{i=1}^{N} \left(q_{i,k}\right)^{\mathbb{I}(z_i^{(m)}=j)} \left(1 - q_{i,k}\right)^{1-\mathbb{I}(z_i^{(m)}=j)}}{\sum_{k'=1}^{K} \prod_{i=1}^{N} \left(q_{i,k'}\right)^{\mathbb{I}(z_i^{(m)}=j)} \left(1 - q_{i,k'}\right)^{1-\mathbb{I}(z_i^{(m)}=j)}}.$$

Step 3 (M-step): For $i = 1, \ldots, N$, and $k = 1, \ldots, K$, compute

$$q_{i,k}^{*} = \frac{\sum_{m=1}^{M} \sum_{j=1}^{K} \gamma_{m,j,k} \mathbb{I}(z_i^{(m)} = k)}{\sum_{m=1}^{M} \sum_{j=1}^{K} \gamma_{m,j,k}}.$$

Step 4: Evaluate the change in $\mathbf{Q}$ between iterations and compare it to the tolerance threshold to assess convergence, e.g. if

$$\underset{i,k}{|} \, q_{i,k}^{*} - q_{i,k}| > t$$

then set $q_{i,k} = q_{i,k}^{*}$, return to step 2 and repeat, otherwise continue to step 5.

Step 5: For $m = 1, \ldots, M$ and $k = 1, \ldots, K$ set $\tau_k^{(m)}$ as

$$\tau_k^{(m)} = \underset{j=1,\ldots,K}{\gamma_{m,j,k}}.$$

Puolamäki and Kaski (2009) also suggest an alternative EM approach, called the Bernoulli mixture permutation algorithm, which minimises a cost function. This idea is re-iterated in Pauli and Torelli (2015). However, the exact specification of the cost function is unclear. Nor is it clear how the cost function might be minimised as an EM algorithm. The BMP algorithm, as

described by Puolamäki and Kaski (2009), was investigated but did not converge. As another alternative, we propose an iterative deterministic algorithm which solves the LSAP using the cost function

$$c_{j,k}^{(m)} = -\sum_{i=1}^{N} q_{i,k} \mathbb{I}\left(z_i^{(m)} = j\right)$$

where $q_{i,k}$ is updated using the latest estimate of the permutations. Further details of the cost functions proposed by Puolamäki and Kaski (2009) and Pauli and Torelli (2015) are provided in Appendix C.1.

Algorithm 5.4: Bernoulli mixture permutation (BMP) algorithm

Step 1: Initialise the $M \times K$ matrix of permutations $\boldsymbol{\mathcal{T}} = \left\{\tau^{(1)}, \ldots, \tau^{(m)}\right\}$. This is usually initialised so that $\tau^{(m)} = \{1, \ldots, K\}$ for all m.

Step 2: For $i = 1, \ldots, N$ and $k = 1, \ldots, K$, calculate

$$q_{i,k} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(z_i^{(m)} = \tau(k)\right).$$

Step 3: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP using costs

$$c_{j,k}^{(m)} = -\sum_{i=1}^{N} q_{i,k} \mathbb{I}\left(z_i^{(m)} = j\right).$$

Step 4: If an improvement in $\sum_{m=1}^{M} \hat{\mathcal{L}}_0^{(m)}$ has been achieved, return to step 2 and repeat, otherwise stop.

The Equivalence Classes Representatives Relabelling Algorithms

If $\boldsymbol{z}_1 = \tau(\boldsymbol{z}_2)$ for some permutation τ, then the two allocation vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are said to be equivalent. An equivalence class is defined as the unique set of allocation vectors which are all equivalent to each other (Papastamoulis and Iliopoulos 2010). The equivalence classes representatives (ECR) algorithm assigns each allocation vector $\boldsymbol{z}^{(m)}$ to an equivalence class, and then selects one allocation vector from each class as a representative (Papastamoulis 2016; Papastamoulis and Iliopoulos 2010). One way to determine a good representative for each class is to first determine a pivot $\boldsymbol{z}^*$ and then choose the permuted allocation $\tau(\boldsymbol{z})$ which is

most similar to $\boldsymbol{z}^*$. The permutations necessary to undo the label switching are then simply the permutations which match each allocation vector to its representative (Papastamoulis and Iliopoulos 2010). The similarity (or dissimilarity) between vectors can be measured in numerous ways. One simple and effective suggestion for measuring dissimilarity between two vectors $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ is the simple matching distance

$$SMD(\boldsymbol{z}_1, \boldsymbol{z}_2) = N - \sum_{i=1}^{N} \mathbb{I}\left(z_{i,1} = z_{i,2}\right) \tag{5.12}$$

(Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014). The task of determining the equivalence classes, the representatives for each class, and the permutations necessary to undo the label switching is rather computationally demanding. However, this approach can be greatly simplified. The permutations can be determined by finding the permutation τ which minimises some dissimilarity measure between $\tau^{-1}\left(\boldsymbol{z}^{(m)}\right)$ and the pivot $\boldsymbol{z}^*$ directly, for each allocation vector $\boldsymbol{z}^{(m)}$. For example, using Equation (5.12),

$$\tau^{(m)} = \underset{\tau \in S}{\operatorname{argmin}} \left\{ SMD(\tau^{-1}\left(\boldsymbol{z}^{(m)}, \boldsymbol{z}^*\right)\right\}$$

Further improvements to computational efficiency can be achieved by formulating the problem as a LSAP using the costs:

$$c_{j,k}^{(m)} = \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) - \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) \mathbb{I}\left(z_i^* = k\right).$$

(Rodriguez and Walker 2014). The results will depend somewhat on the choice of the pivot $\boldsymbol{z}^*$, so it is important to choose a good allocation vector as the pivot. The allocation vector corresponding to the Monte Carlo *maximum a posteriori* (MAP) estimate has often been suggested as a good choice for the pivot (Papastamoulis 2016; Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014). Alternatives include the allocation vector corresponding to the maximum of the complete likelihood function and the most probable allocation (Papastamoulis and Iliopoulos 2010; Rodriguez and Walker 2014).

Two variations to the ECR algorithm were provided by Rodriguez and Walker (2014) which reduce the dependence on the choice of the pivot by updating the estimate of the pivot iteratively. The first variation initialises the permutations $\left\{\tau^{(1)}, \ldots, \tau^{(M)}\right\}$ and calculates the pivot $\boldsymbol{z}^* =$

$(z_1^*, \ldots, z_N^*)$ by setting

$$z_i^* = \text{mode}\left\{(\tau^{-1})^{(1)}\left(z_i^{(1)}\right), \ldots, (\tau^{-1})^{(M)}\left(z_i^{(M)}\right)\right\}.$$

for $i = 1, \ldots, N$. If the solution leads to a reduction in the overall loss, the process is repeated until no further improvements can be made. Neither the original ECR algorithm nor this iterative variation guarantees that the final pivot will be 'good'. The second variation addresses this issue by using the classification probabilities, given by Equation (5.9), to estimate the pivot (Papastamoulis 2016; Rodriguez and Walker 2014). In both cases, the costs are the same as those used in the original ECR algorithm and the permutations are updated by solving the LSAP.

Algorithm 5.5: equivalence classes representatives (ECR) algorithm

Step 1: Choose one iteration m^* with corresponding allocation vector $\boldsymbol{z}^* = (z_1, \ldots, z_N)^{(m^*)}$ as the pivot. One suggestion is to choose m^* as the iteration which corresponds to the Monte Carlo approximation of the MAP estimate of $\boldsymbol{\theta}$.

Step 2: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP using costs

$$c_{j,k}^{(m)} = \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) - \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) \mathbb{I}\left(z_i^* = k\right).$$

Algorithm 5.6: iterative ECR algorithm version 1 (ECR 1)

Step 1: Initialise the $M \times K$ matrix of permutations $\boldsymbol{\mathcal{T}} = \left\{\tau^{(1)}, \ldots, \tau^{(m)}\right\}$. This is usually initialised so that $\tau^{(m)} = \{1, \ldots, K\}$ for all m.

Step 2: For $i = 1, \ldots, N$, calculate the pivot $\boldsymbol{z}^*$ with elements

$$z_i^* = \underset{m=1,\ldots,M}{\text{mode}}\left\{(\tau^{-1})^{(m)}\left(z_i^{(m)}\right)\right\}.$$

Step 3: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP as per Algorithm 5.5.

Step 4: If an improvement in $\sum_{m=1}^{M} \hat{\mathcal{L}}_0^{(m)}$ has been achieved, return to step 2 and repeat, otherwise stop.

Algorithm 5.7: iterative ECR algorithm version 2 (ECR 2)

Step 1: Initialise the $M \times K$ matrix of permutations $\boldsymbol{\mathcal{T}} = \left\{\tau^{(1)}, \ldots, \tau^{(m)}\right\}$. This is usually initialised so that $\tau^{(m)} = \{1, \ldots, K\}$ for all m.

Step 2: For $i = 1, \ldots, N$, calculate the pivot $\boldsymbol{z}^*$ with elements

$$z_i^* = \underset{k=1,\ldots,K}{\text{argmax}} \left\{\hat{p}_{i,k}\right\}.$$

where $\hat{p}_{i,k}$ is given by Equation (5.10).

Step 3: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP as per Algorithm 5.5.

Step 4: If an improvement in $\sum_{m=1}^{M} \hat{\mathcal{L}}_0^{(m)}$ has been achieved, return to step 2 and repeat, otherwise stop.

The SJW Probabilistic Relabelling Algorithm

Given the true values of the parameters, the permutations can be determined by maximising the discrete probability density over all possible permutations, i.e.

$$z_i^* = \max_{\text{ƙ}=1,\ldots,K!} g_{\text{ƙ}}^{(m)}.$$

for $m = 1, \ldots, M$ where $g_{\text{ƙ}}^{(m)} = \Pr\left(\tau^{(m)} = S_{\text{ƙ}} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{z}\right)$. Because the parameters are also unknown, Sperrin et al. (2010) propose an iterative EM algorithm: use the observed data and estimates of the model parameters, including $\boldsymbol{z}$, to compute the probabilities $g_{\text{ƙ}}^{(m)}$, then use these probabilities to update the estimates of the parameters, repeating the process until a fixed point is reached (Papastamoulis 2013, 2016; Sperrin et al. 2010). This methodology has two major drawbacks. First, like all EM-type algorithms, this method may produce local maxima, so it may be necessary to run the algorithm from multiple starting values (Papastamoulis 2013; Sperrin et al. 2010). Second, the method requires the probabilities of each permutation to be estimated, and not simply the most likely permutation, so the computational burden increases dramatically with K (Papastamoulis 2013, 2016).

Algorithm 5.8: SJW algorithm

Step 1: Initialise an estimate of the parameters $\hat{\boldsymbol{\theta}} = \left\{ \hat{\boldsymbol{\phi}}_{1,(1,\ldots,K)}, \ldots, \hat{\boldsymbol{\phi}}_{R-1,(1,\ldots,K)}, \hat{\boldsymbol{w}} \right\}$

Step 2 (E-step): For $m = 1, \ldots, M$ and $\kappa = 1, \ldots, K!$, use the complete likelihood to compute the permutation probabilities:

$$\begin{aligned} g_{\kappa}^{(m)} &= \Pr\left(\tau^{(m)} = S_{\kappa} | \boldsymbol{w}, \boldsymbol{\phi}, \boldsymbol{\lambda}, \boldsymbol{y}, \boldsymbol{z}\right) \\ &= \prod_{i=1}^{N} \tau\left(w_{z_i^{(m)}}^{(m)}\right) f_{z_i^{(m)}}\left(y_i | \tau\left(\boldsymbol{\phi}_{z_i^{(m)}}^{(m)}\right), \boldsymbol{\lambda}^{(m)}\right), \quad \tau = S_{\kappa}. \end{aligned}$$

Step 3 (M-step): Update the parameter estimates:

$$\hat{\boldsymbol{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \sum_{\kappa=1}^{K!} g_{\kappa}^{(m)} \boldsymbol{\theta}^{(m)}.$$

Step 4: Return to step 2 and repeat until the optimal solution is reached.

The Data-Based Relabelling Algorithm

The relabelling algorithm proposed by Rodriguez and Walker (2014) relies on the fact that similar observations will tend to be allocated to the same component each iteration, notwithstanding label switching. Therefore, the data can be used to relabel the parameters by comparing statistics, such as the mean and standard deviation, of the data and clusters derived for a given set of permutations. These estimates are then used to update the permutations by minimising a cost function which incorporates these estimates and the data through a k-means type divergence measure (Rodriguez and Walker 2014).

Rodriguez and Walker (2014) suggest how good initial estimates of the cluster means and standard deviations may be obtained, and discuss how the permutations may be found, either by a "two-step procedure" (given by Algorithms 5 and 6 in Rodriguez and Walker (2014, pp. 35)), or recursively (via Algorithm 7 in Rodriguez and Walker (2014, pp. 36)). The "two-step procedure" is really a special case of the recursive algorithm where the number of iterations is fixed at two, while the recursive algorithm updates the permutations until no further improvements to the overall loss can be achieved. We shall refer to these variants of the algorithm "DB" and

"DB iterative" respectively. These two approaches are summarised in Algorithms 5.9 and 5.10 respectively.

Algorithm 5.9 and 5.10: data-based algorithm (DB and DB iterative)

Step 1: Initialise the $M \times K$ matrix of permutations $\boldsymbol{\mathcal{T}} = \left\{\tau^{(1)}, \ldots, \tau^{(m)}\right\}$. This is usually initialised so that $\tau^{(m)} = \{1, \ldots, K\}$ for all m.

Step 2: For $k = 1, \ldots, K$, initialise/update estimates of the cluster means and standard deviations, $\hat{\boldsymbol{m}}_k$ and $\hat{s}_k$. If initialising, Rodriguez and Walker (2014, pp. 35) suggest using:

$$\hat{\boldsymbol{m}}_k = \min(y) + \Big((\max(y) - \min(y)\Big)\left(\frac{k}{K+1}\right)$$
$$\hat{s}_k = \frac{\max(y) - \min(y)}{K}$$

If updating:

$$\hat{\boldsymbol{m}}_k = \frac{\sum_{m=1}^{M} \bar{y}^{(m)}_{\tau^{(m)}} \mathbb{I}\left(n^{(m)}_{\tau^{(m)}(k)} > 0\right)}{\sum_{m'=1}^{M} \mathbb{I}\left(n^{(m')}_{\tau^{(m')}(k)} > 0\right)}$$
$$\hat{s}_k = \frac{\sum_{m=1}^{M} s^{(m)}_{\tau^{(m)}} \mathbb{I}\left(n^{(m)}_{\tau^{(m)}(k)} > 1\right)}{\sum_{m'=1}^{M} \mathbb{I}\left(n^{(m')}_{\tau^{(m')}(k)} > 1\right)}$$

where

$$n^{(m)}_k = \sum_{i=1}^{N} \mathbb{I}\left(z^{(m)}_i = k\right)$$
$$\bar{y}^{(m)}_k = \frac{1}{n^{(m)}_k} \sum_{i=1}^{N} y_i \mathbb{I}\left(z^{(m)}_i = k\right)$$
$$s^{(m)}_k = \left[\sum i = 1^N \left(y_i - \bar{y}^{(m)}_k\right)^2 \mathbb{I}\left(z^{(m)}_i = k\right)\right]^{1/2}$$

Step 3: For $m = 1, \ldots, M$, determine $\tau^{(m)}$ by solving the LSAP using costs

$$c^{(m)}_{j,k} = n^{(m)}_j \sum_{i=1}^{N} \left(\frac{y_i - \hat{(}\boldsymbol{m}_k}{\hat{(}s_k}\right)^2 \mathbb{I}\left(z^{(m)}_i = j\right).$$

Step 4:

a) Non-iterative version: repeat steps 2 and 3 once to obtain the final estimates of $\tau^{(m)}$.

b) Iterative version: if an improvement in $\sum_{m=1}^{M} \hat{\mathcal{L}}_0^{(m)}$ has been achieved, return to step 2 and repeat, otherwise stop.

The iterative version must perform a minimum of two iterations (to be able to assess changes in the overall loss), so the non-iterative version will always be at least as fast. However, the non-iterative version is more reliant on the initial estimates, and therefore may be less accurate (Rodriguez and Walker 2014).

Alternative measures of central tendency and dispersion are discussed in Rodriguez and Walker (2014), which may increase robustness of the algorithm. However, we found the mean and standard deviation to perform adequately in our simulations and did not consider these alternatives.

The Pivotal Unit Relabelling Algorithm

The approach proposed by Pauli and Torelli (2015) is to identify K perfectly separated observations, that is, observations which are assigned to different components for all MCMC iterations, referred to as pivotal units, and use these to relabel the chains. The ability to identify K perfectly separated observations may not be possible, in which case the goal is to identify the K observations which maximise the number of iterations for which this condition is satisfied. The iterations which do not satisfy this condition are then removed from the posterior sample (Pauli and Torelli 2015). The suggested method of identifying the pivots is explained in Algorithm 5.11.

Algorithm 5.11: pivotal unit (PU) algorithm

Step 1: For $k = 1, \ldots, K$, assign the N observations to one of K sets $N_k \subset \{1, \ldots, N\}$. One option is to either systematically or randomly choose a particular iteration m^* and use the allocation vector $\boldsymbol{z}^* = (z_1, \ldots, z_N)^{(m)}$ to determine the $\{N_k\}$, e.g.:

$$N_k = \{i : z_i^* = k\}.$$

Step 2: Compute the $N \times N$ similarity matrix $\tilde{\mathbf{S}}$ with elements

$$\tilde{s}_{i,j} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(z_i^{(m)} = z_j^{(m)}\right).$$

Step 3: For $k = 1, \ldots, K$, determine the pivotal unit u_k from each set N_k by maximising (or minimising) an appropriate criterion, $\tilde{c}\left(\tilde{\mathbf{S}}, N_k\right)$, e.g.

$$u_k = \underset{i \in N_k}{\text{argmax}} \left(\underset{j \in N_k}{\max} \tilde{s}_{i,j} \right).$$

(Refer to text for alternative criteria.)

Step 4: Define the$M \times K$ matrix $\boldsymbol{Z}^*$ as the subset of allocation vectors corresponding to the pivotal units:

$$\boldsymbol{Z}^* = \left[\boldsymbol{z}_{u_1}, \ldots, \boldsymbol{z}_{u_K}\right].$$

Step 5: For $m = 1, \ldots, M$, remove iteration m from the posterior sample if the number of unique elements in the m^{th} row of $\boldsymbol{Z}^*$ is not exactly K (and remove this row from $\boldsymbol{Z}^*$). Let M_s be the size of the posterior sample remaining.

Step 6: For $m = 1, \ldots, M_s$, set $\tau^{(m)}$ as the m^{th} row of $\boldsymbol{Z}^*$.

Pauli and Torelli (2015, pp. 10) suggest five other example criteria for step 3:

$$\underset{i \in N_k}{\text{argmax}} \sum_{j \in N_k} \tilde{s}_{i,j}, \qquad \underset{i \in N_k}{\text{argmax}} \left(\sum_{j \in N_k} \tilde{s}_{i,j} - \sum_{j \notin N_k} \tilde{s}_{i,j} \right), \qquad \underset{i \in N_k}{\text{argmin}} \sum_{j \notin N_k} \tilde{s}_{i,j},$$

$$\underset{i \in N_k}{\text{argmin}} \left(\underset{j \in N_k}{\min} \tilde{s}_{i,j} \right), \qquad \underset{i \in N_k}{\text{argmin}} \left(\underset{j \notin N_k}{\max} \tilde{s}_{i,j} \right).$$

This method is simple to implement and should be computationally fast even for large K. However, the method has two major drawbacks: the difficulty in choosing a suitable criterion, and the possibility of discarding some portion of the posterior sample when it is impossible to identify K perfectly separated observations to construct the pivots (Pauli and Torelli 2015).

The Zswitch Relabelling Algorithm

The algorithm proposed by van Havre et al. (2015) is a hybrid of allocation-based and parameter-based relabelling. The idea is to choose one sample of the allocation vector as a pivot, $\boldsymbol{z}^*$. Then for each iteration, the joint distribution of the current allocation vector and the pivot is computed, summarised as a $K \times K$ matrix $\mathbf{M}$ with elements $\mathbf{M}_{j,k}$ representing the proportion of observations belonging to group j in the current iteration and group k in the pivot. After ignoring proportions which are less than some threshold ω, the non-zero cells in the matrix $\mathbf{M}$ indicate the only valid permutations for relabelling the current allocation vector so that it matches the pivot. In many cases, there might be only one valid permutation. For other iterations, this step will identify a set of candidate permutations, $\hat{S} \subseteq S$. In this case, the best permutation is decided as the permutation which minimises a loss function similar to that used by the PRA algorithm (van Havre et al. 2015).

The number of elements in $\hat{S}$ will vary, depending on the overlap of the component densities, but it is usually much smaller than $K!$, which has the potential to make the task of finding the best permutation in the second stage of the algorithm computationally feasible even when K is large (van Havre et al. 2015).

Algorithm 5.12: Zswitch (ZS)

Step 1: Choose one iteration m^* to be the reference, with corresponding allocation vector $z^* = (z_1, \ldots, z_N)^{(m^*)}$ and parameter values $\boldsymbol{\theta}^* = \left\{\boldsymbol{\phi}_{1,(1,\ldots,K)}, \ldots, \boldsymbol{\phi}_{R-1,(1,\ldots,K)}, \boldsymbol{w}\right\}^{(m^*)}$. One suggestion is to choose m^* as the iteration which corresponds to the Monte Carlo approximation of the MAP estimate of $\boldsymbol{\theta}$.

Step 2: For $m = 1, \ldots, M$:

Phase 1: Allocation-based relabelling

a) Construct a $K \times K$ matrix $\mathbf{M}$ with elements

$$\mathbf{M}_{j,k} = \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) \mathbb{I}\left(z_i^* = k\right), \quad j, k \leq K.$$

b) For $j = 1, \ldots, K$, define the set I_j as

$$I_j = \left\{ k : \frac{\mathbf{M}_{j,k}}{\sum_{k'=1}^{K} \mathbf{M}_{j,k'}} > \omega \right\}$$

c) Define $\hat{S} \subseteq S$ as the set of permutations arising from the K-fold Cartesian product of each set $\{I_j\}$:

$$\hat{S} = I_1 \times \ldots \times I_K.$$

For each permutation in $\hat{S}$, discard any permutation from this set which does not contain every label in the index set $\{1, \ldots, K\}$ to ensure the validity of permutations.

d) If $|\hat{S}| = 1$, set $\tau^{(m)} = \hat{S}$, otherwise set

Phase 2: Parameter-based relabelling

$$\tau^{(m)} = \underset{\tau \in \hat{S}}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{r=1}^{R} \left| \frac{\theta^*_{r,k} - \theta^{(m)}_{r,\tau(k)}}{\theta^*_{r,k}} \right|.$$

Note that it is possible for I_j to contain empty elements if ω is too large. See text for solutions to this problem.

The code for implementing this algorithm in R was provided by the authors as supplementary material (van Havre et al. 2015). However, that implementation was limited to univariate Gaussian mixture models. We provide an updated version of this algorithm which has no restriction on the choice of the likelihood.

One of the drawbacks to this method is the choice of the tuning parameter ω. Choosing a larger value for ω should decrease the total computation time. However, if ω is set too large, this can result in I_j being an invalid set for one or more iterations, that is, set I_j may have empty elements, prohibiting the calculation of a valid permutation. When this occurs, there are two courses of action: simply remove the affected iterations from the posterior sample, or use $\omega = 0$ for the affected iterations, thereby imposing parameter-based relabelling. We consider the second of these options.

5.4.1 A New Relabelling Algorithm

In the simulation study that follows, it will become clear that no one particular algorithm shows superiority in every regard. However, after running many different simulations, ZS seemed to offer an attractive compromise between accuracy and computational efficiency. ZS is designed to avoid computational overload for large values of K through the careful construction of the set $\hat{S}$. However, when there is considerable overlap between components and/or the sample size N is small, $\hat{S}$ can be very large. The only way to avoid this problem is to increase the value of ω, but this can result in I_j being invalid, as stated above. Note that this difficulty of unique label identification when components overlap considerably is common to all relabelling algorithms. The tuning parameter ω in ZS provides some flexibility in mitigating this problem, and thus despite the difficulty of choosing a reasonable value, the tuning parameter is actually an advantage.

Most relabelling algorithms avoid computational burden by formulating the problem as a LSAP, which can be solved using very efficient algorithms. Therefore, to overcome the computational problems susceptible to ZS, we suggest using the more traditional LSAP approach where the cost function is derived from the parameter-based relabelling strategy step, namely

$$c_{j,k}^{(m)} = \sum_{r=1}^{R} \left| \frac{\theta_{r,k}^{*} - \theta_{r,j}^{(m)}}{\theta_{r,k}^{*}} \right|.$$

and subsequently modified so that costs corresponding to infeasible permutations, according to the matrix $\mathbf{M}$, are infinite. This encapsulates the main principles of ZS, where the subset $\hat{S}$ is replaced by a cost matrix, and the set of permissible permutations is still governed by the matrix $\mathbf{M}$. This modification circumvents the problems associated with ω, and should increase the computational efficiency for larger values of K, albeit at a small cost for smaller values of K, while retaining a similar degree of accuracy. To further improve relabelling accuracy, the costs are rescaled by the matrix $\mathbf{M}$, which is very effective when one or more parameters are estimated with high uncertainty. We shall refer to this modified Zswitch algorithm as "Zswitch 2".

Algorithm 5.13: Zswitch 2 (ZS2)

Step 1: Choose one iteration m^* to be the reference, with corresponding allocation vector

$z^* = (z_1, \ldots, z_N)^{(m^*)}$ and parameter values $\boldsymbol{\theta}^* = \left\{\boldsymbol{\phi}_{1,(1,\ldots,K)}, \ldots, \boldsymbol{\phi}_{R-1,(1,\ldots,K)}, \boldsymbol{w}\right\}^{(m^*)}$. One suggestion is to choose m^* as the iteration which corresponds to the Monte Carlo approximation of the MAP estimate of $\boldsymbol{\theta}$.

Step 2: For $m = 1, \ldots, M$:

a) Construct a $K \times K$ matrix $\mathbf{M}$ with elements

$$\mathbf{M}_{j,k} = \sum_{i=1}^{N} \mathbb{I}\left(z_i^{(m)} = j\right) \mathbb{I}\left(z_i^* = k\right), \quad j, k \leq K.$$

b) Determine $\tau^{(m)}$ by solving the LSAP using costs

$$c_{j,k}^{(m)} = \begin{cases} \dfrac{1}{\mathbf{M}_{j,k}} \displaystyle\sum_{r=1}^{R} \left| \frac{\theta_{r,k}^* - \theta_{r,j}^{(m)}}{\theta_{r,k}^*} \right| & \text{if } \frac{\mathbf{M}_{j,k}}{\sum_{k'=1}^{K} \mathbf{M}_{j,k'}} > \omega \\ \infty & \text{otherwise} \end{cases}.$$

Note that in practice, it may be necessary to replace the infinite costs by arbitrarily large values. This is true for the `lp.assign` function in the R package `lpSolve` (Berkelaar, M. et al. 2015).

5.5 Simulation Studies

5.5.1 Previous Simulation Studies

Numerous simulation studies aimed at comparing alternative relabelling algorithms have been published. However, all of the studies we reviewed exhibited at least one of the following four constraints.

First, the scope of these simulation studies is usually limited to just one or a few algorithms. This is not surprising since the purpose of the study is usually to demonstrate the superiority of a particular algorithm, and the inclusion of other algorithms as benchmarks, if any, is to further quantify such claims. The largest simulation study of which we are aware compared 8 algorithms, but this included the two variations of ECR and also IC (Papastamoulis 2016). A wide range of relabelling algorithms now exists, and an extensive comparison of these methods

is now possible and warranted.

Second, the number of components is usually very small (5 or less). This was also observed by Papastamoulis (2013), who noted a lack of simulation studies containing mixtures with more than 6 components. Since the number of components is a prominent factor in the computational efficiency of any algorithm, comparable evaluation of their performance for large K is also warranted.

Third, in addition to a small number of components, simulation studies often entail simple scenarios, such as well separated components, equal mixture weights, and/or treating some parameters as known all of which make it easier for relabelling algorithms to succeed, and therefore harder to distinguish between the strengths and weaknesses of each approach.

Fourth, almost all simulation studies consider Gaussian mixture models exclusively. Since none of the 13 algorithms considered here are limited to Gaussian mixture models, this seems like a fairly large oversight.

A summary of previous simulation studies from the reviewed literature are provided in Appendix C.2.

5.5.2 Simulation Study

All 13 algorithms, including ZS 2, presented in this paper were implemented in R R Core Team (2015) under three different scenarios. Seven of the algorithms discussed in this paper are available in the R package `label.switching` (Papastamoulis 2015, 2016): KL, PRA, ECR, ECR 1, ECR 2, SJW, and DB. However, this package only offers the non-iterative version of the DB algorithm, `dataBased`, and the computation of the cost function in `dataBased` seems to differ from the text of both Rodriguez and Walker (2014) and Papastamoulis (2016). Also, the implementation of the PRA algorithm, `pra`, is not formulated as a LSAP, and therefore underestimates the computational efficiency of this algorithm. Consequently, the `label.switching` package was used to implement only the KL, ECR, ECR 1, ECR 2, and SJW methods. The remaining algorithms were implemented in R directly. The algorithms which are formulated as LSAPs were solved using the R package `lpSolve` (Berkelaar, M. et al. 2015). The R code for implementing the 13 algorithms, including a description of the input parameters and optional parameter defaults, are provided in Appendix C.3.

The scenarios presented in this simulation study involved fitting a K-component mixture model to randomly generated data consisting of G subpopulations, using 'true' values of the parameters. The model parameters are then estimated using the Gibbs sampler to deliberately avoid label switching. Although this strategy is not guaranteed to prevent label switching, it is easy to verify whether label switching occurs by visually inspecting the sequence of component-specific parameter values for each component. Only in a few cases where the components were overlapping considerably did this strategy falter, but these cases were readily resolved by making small changes to the 'true' parameters, or making the prior distribution less vague. For the purpose of this simulation study, it is important that label switching does not occur, as the posterior sample is used to check the accuracy of the relabelling algorithms. Once obtained, the posterior sample is duplicated and label switching is induced in accordance with Equation (5.3).

The three scenarios are designed to compare three fundamental characteristics of each relabelling algorithm: the computational efficiency, the accuracy of determining the correct permutations to recover the true posterior sample, and the robustness to misspecification of K. Computational efficiency was assessed by recording the computation time of each algorithm using the `microbenchmark` package in R Mersmann, O. (2015), which is allegedly accurate to the nanosecond. A lower computation time indicates greater efficiency. Note that all relevant intermediary calculations, such as the classification probabilities and MAP estimate, are included in the computational times. To avoid sampling bias, three realisations of each data set are generated, and the median computation time is averaged over each realisation. Using the algorithms to validate each other, the solutions produced by deterministic and iterative algorithms appeared to be globally optimal, with the exception of the BM algorithm, which was run until perfect accuracy was obtained, up to a maximum of 10 runs. The computational cost expended on these additional runs is not included in the recorded computation times for this algorithm.

The accuracy for the m^{th} iteration, denoted $A^{(m)}$, is defined as the proportion of the permutation indices, out of K, identified correctly according to the inverse of the permutations used to induce the label switching. From this, a mislabel severity index (MSI) is defined as

$$MSI = 1 - \frac{1}{M} \sum_{m=1}^{M} A^{(m)}, \tag{5.13}$$

where the value of $A^{(m)}$ is treated as zero for missing values. The MSI is a useful means of comparing the accuracy of each algorithm as it takes into account the number of correct permutation indices at each iteration and missing values, and not simply the number of iterations with correct permutations. A MSI value of zero indicates that the algorithm is able to undo the label switching with perfect accuracy.

In the first scenario, a data set is generated from a mixture of G Poisson likelihoods, $G = \{2, \ldots, 9\}$. Each set of generated data is 'well-behaved', that is, the subpopulations are equally weighted, and the sample size is relatively large. For $G = \{2, \ldots, 5\}$, the subpopulations are well separated, while for $G = \{6, \ldots, 9\}$, a moderate amount over overlap between the components exists. A Poisson mixture model with $K = G$ components was fit to each data set. This scenario represents a best-case scenario, designed to illustrate the feasibility of each method in terms of accuracy.

In the second scenario, a data set is generated from a mixture of G Gaussian likelihoods, $G = \{2, 3, \ldots, 10, 15, 20, 50, 100\}$. In this scenario, some subpopulations exhibit a high degree of overlap. For $G = \{2, \ldots, 5\}$, the subpopulation weights are very unbalanced, while equal weights are used for the other values of G. As with scenario 1, the number of components is assumed known, so $K = G$ for each mixture model fit to the data. This scenario is designed to highlight the computational efficiency with respect to K and the accuracy of each method under a worst-case scenario.

In the third scenario, the data consists of $G = 5$ equally weighted subpopulations. K-component gamma mixture models are fit to the data, where $K = \{2, 3, \ldots, 8\}$. This scenario is designed to demonstrate robustness to model misspecification when $K \neq G$.

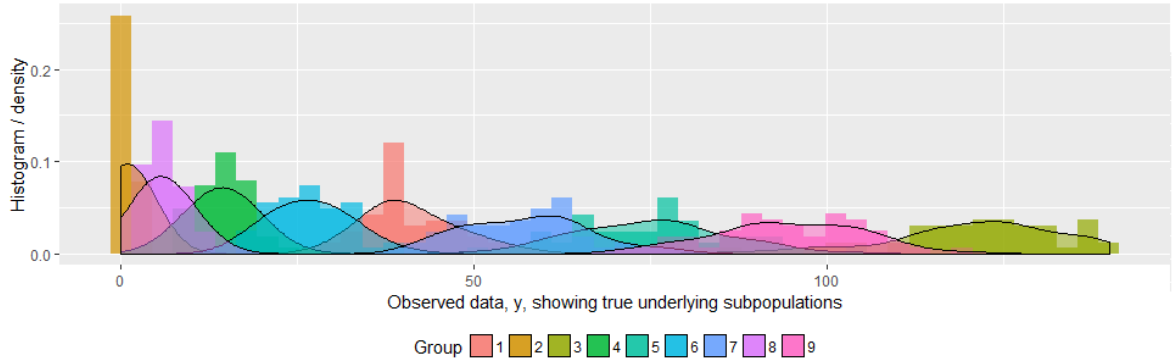
These models, including specification of the prior distributions, are summarised in Appendix C.4. To achieve reliable estimates when $K \geq 20$, it was necessary to order the generated labels, at least within blocks, and split the prior for the mixture means according to those blocks. Specifying a single, exchangeable, vague prior for the means conflicts with the observed data when they are far apart, and this can lead to poorly estimated parameters and missing components. These modifications are exemplified in Appendix C.4.

Table 5.1 summarises the sample sizes, the number of subpopulations and components, and the true parameters used to generate the data for each scenario.

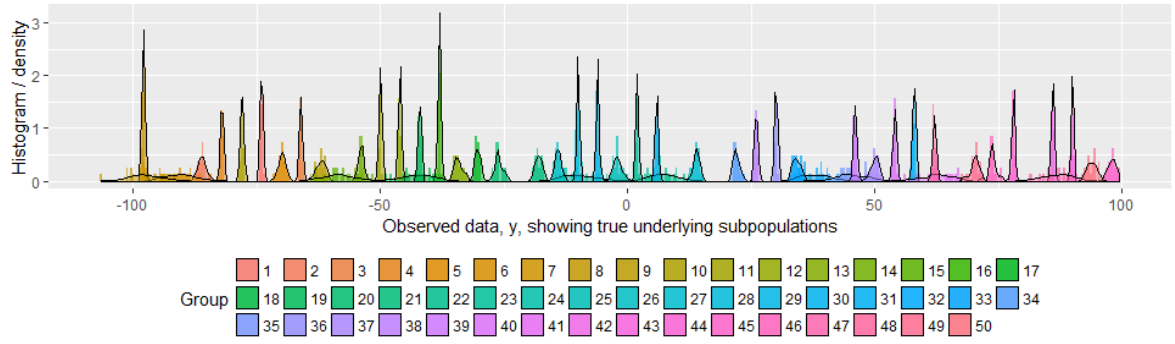
Sc.	N	K	G	True mixture weights	True parameter values used to generate the data
1	500	$K = G$	$2, \ldots, 9$	$\left(\frac{1}{G}, \ldots, \frac{1}{G}\right)$	The first G values in: $\boldsymbol{\eta} = (40, 1, 120, 16, 75, 27, 58, 6, 95)$
2	10^3	$K = G$	2	$(0.96, 0.04)$	The first G values in: $\boldsymbol{\mu} = (3, -4, 3, 5, -2)$ $\boldsymbol{\sigma}^2 = (4, 0.01, 0.1, 0.1, 1)$
			3	$(0.3, 0.05, 0.65)$	
			4	$(0.2, 0.01, 0.3, 0.49)$	
			5	$(0.5, 0.02, 0.09, 0.09, 0.3)$	
			$6, \ldots, 10, 15, 20, 50, 100$	$\left(\frac{1}{G}, \ldots, \frac{1}{G}\right)$	$\boldsymbol{\mu}$ randomly generated from finite interval whose length depends on G; $\boldsymbol{\sigma}^2$ selected from the set $\{0.05, 0.5, 8\}$ at random with replacement.
3	500	5	$2, \ldots, 8$	$\left(\frac{1}{G}, \ldots, \frac{1}{G}\right)$	$\boldsymbol{\delta} = (5, 14, 18, 40, 40)$ $\boldsymbol{\beta} = (3, 1, 3, 1.5, 0.9)$

Table 5.1: Summary of parameters used in each simulation study scenario.

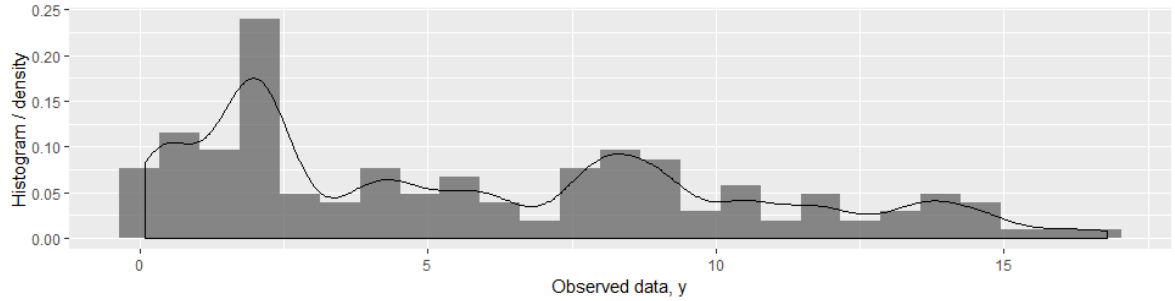
Figure 5.1 shows one realisation of the data generated in each scenario, for a selected value of G. Figure 5.1a shows some generated data for scenario 1 when $G = 9$. Note how the subpopulations are well separated. Figure 5.1b shows some generated data for scenario 2 when $G = 50$. Figures 5.1c and 5.1d are visualisations of the data in scenario 3 when the subpopulations are known and unknown respectively. In practice, only Figure 5.1c would be available. Without Figure 5.1d, the true number of subpopulations, and therefore how many components should be used in the mixture, is not obvious. Although several techniques are available to implement mixtures with an unknown number of components (see Richardson and Green (1997), Stephens (2000a) and van Havre et al. (2015) for example), the number of components is often presumed and treated as fixed. This gives rise to the problem of model misspecification with respect to the number of components. Thus the goal of this scenario is to assess the accuracy of each algorithm when $K = G = 5$, and contrast this to the accuracy when K is misspecified.



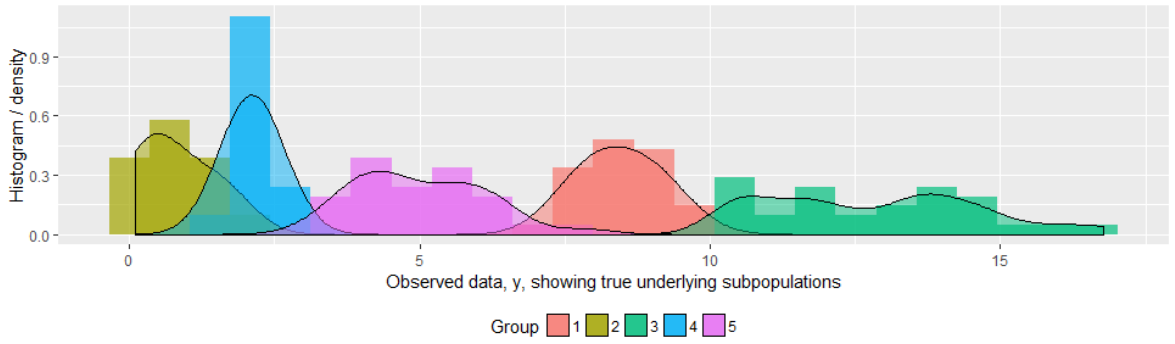
(a) Scenario 1



(b) Scenario 2



(c) Scenario 3 (subpopulations unknown)



(d) Scenario 3 (subpopulations known)

Figure 5.1: One realisation of the data, for a selected value of G, used in (a) scenario 1, (b) scenario 2, and (c-d) scenario 3. Note that (c) is how the data appear in practice, because the underlying subpopulations are unknown, while (d) is how the data would appear if the underlying subpopulations were known.

5.6 Results

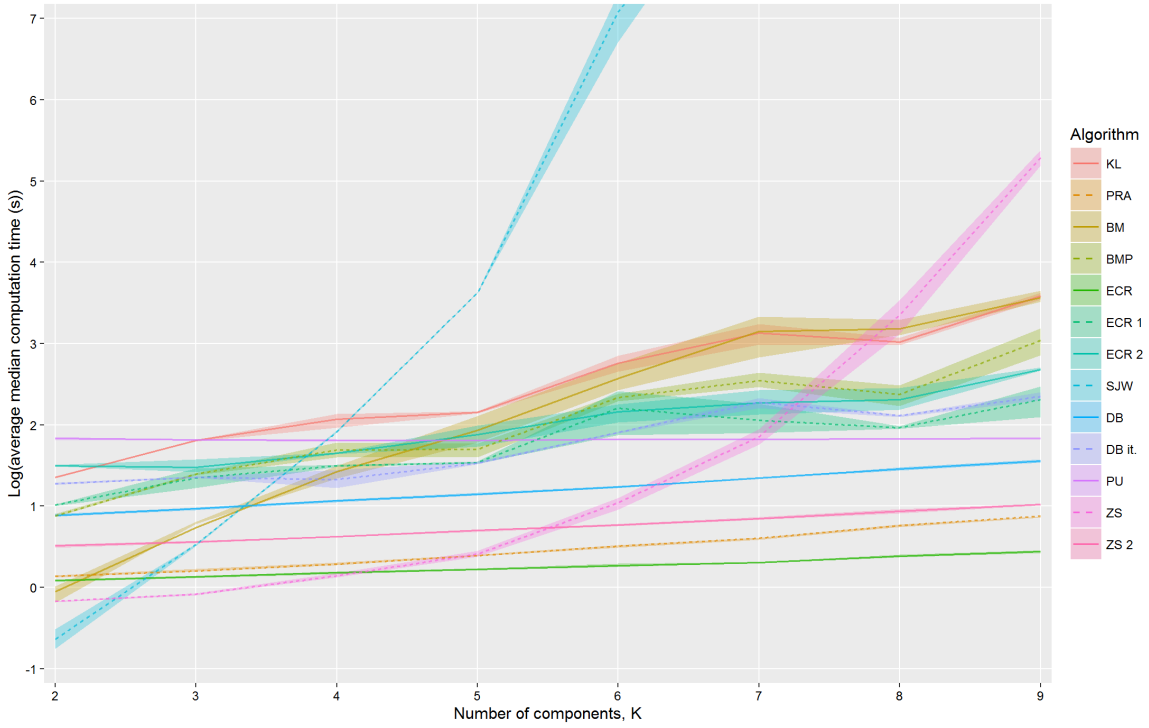
5.6.1 Computational Efficiency

Figure 5.2 shows the natural logarithm of the mean median computation time of each algorithm in scenarios 1 and 2, with bands representing the 30^{th} and 70th percentile computation times. The maximum value on the y-axis is approximately 1000 seconds, and once algorithms reach this threshold, they are not run for larger values of K as they are deemed to be too inefficient. Consequently, the MSI for these algorithms will be 1, as discussed below.

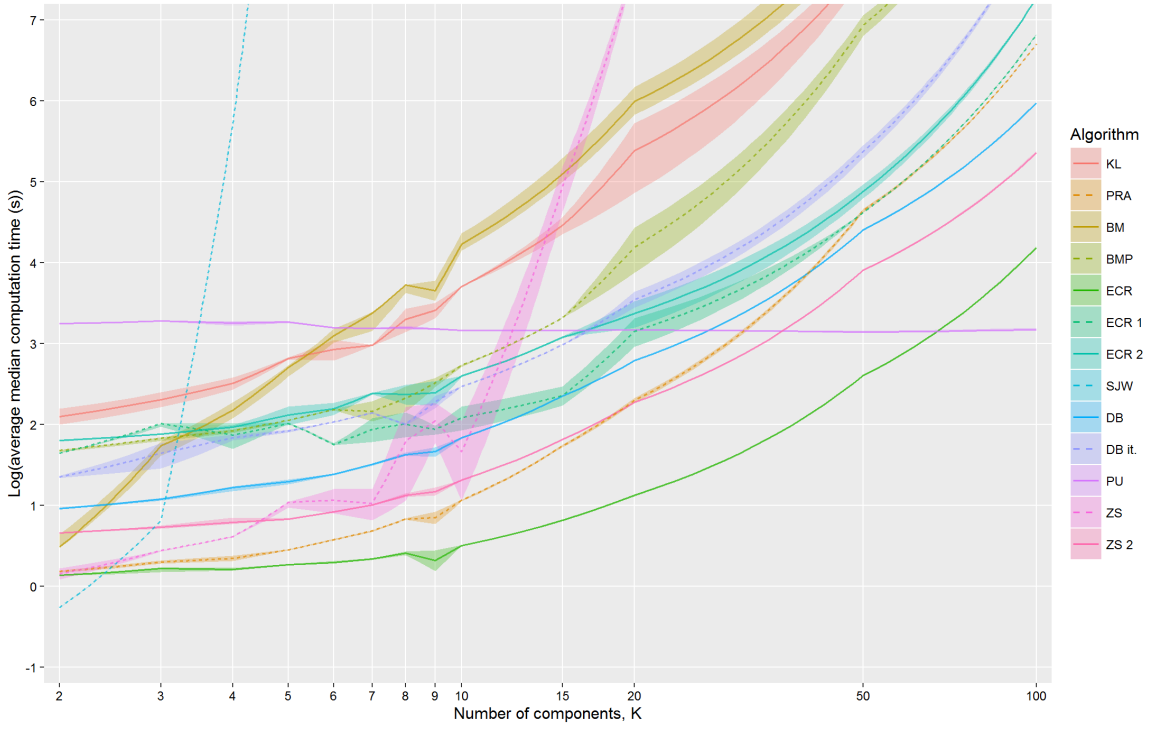
In both scenarios 1 and 2, when $K = 2$, the SJW algorithm was the most computationally efficient, but predictably, this efficiency was quickly lost as K increases, and this algorithm becomes prohibitively slow even for $K = 5$. Similarly, the KL and BM algorithms also scaled poorly with the number of components, although the BM algorithm did perform better for $K < 5$. Recall that the computation time shown for the BM algorithm is for one run only; the computational cost to achieve the accuracy discussed below may be much higher. The computational efficiency of ZS is an improvement on these three algorithms for smaller values of K, but the efficiency of ZS diminished rapidly after about $K = 7$. This shortcoming of ZS is adequately addressed by ZS 2 which displays a computational efficiency comparable to that attained by the PRA algorithm formulated as a LSAP, and was only marginally slower than ZS for $K < 4$ and $K < 5$ in scenarios 1 and 2 respectively. In scenario 2, ZS 2 even surpassed PRA in efficiency for $K > 20$.

The ECR algorithm was considerably faster than the ECR 1 and ECR 2 variants. Indeed, the computational efficiency of ECR was generally faster than any other algorithm, at least for $K > 5$, a finding which is consistent with the literature (Papastamoulis 2013). Unsurprisingly, the DB iterative algorithm was computationally slower than the DB algorithm, with the anticipation of increased accuracy. The performance of the BMP algorithm lies between that of the DB and KL algorithms.

This just leaves the incongruous PU algorithm, whose computation time appeared very consistent with respect to K. The computational burden of the PU algorithm is largely determined by the cost in computing the similarity matrix $\hat{\mathbf{S}}$, which increases with N rather than K. Note that the sample size between scenarios 1 and 2 doubles, while the computation time of the PU algorithm quadruples.



(a) Scenario 1



(b) Scenario 2

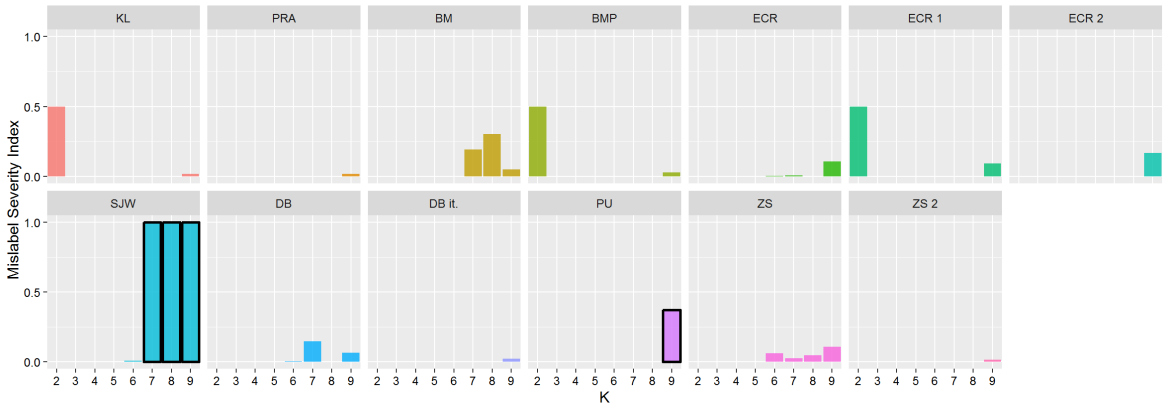
Figure 5.2: Natural logarithm of the median computation time, averaged over each realisation of the generated data, for each algorithm in (a) scenario 1, and (b) scenario 2. The bands represent the 30th and 70th percentiles for the median computation times pooled over all realisations of the data. The x-axis in (b) is also represented on the log scale to accentuate the differences for smaller values of K.

5.6.2 Accuracy and Robustness to Misspecification of K

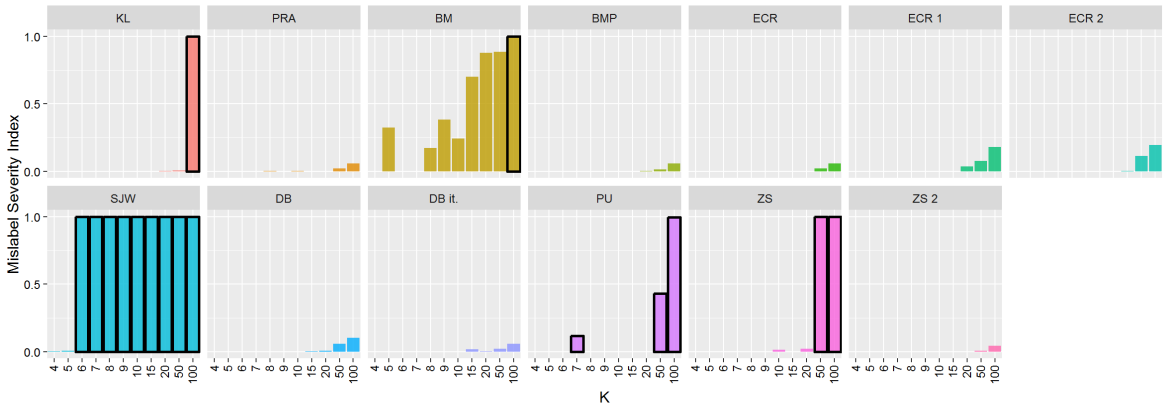
The MSI for each algorithm in each scenario is shown in Figure 5.3. As stated above, algorithms whose mean median computation exceeded 1000s for a given value of K are not considered for further assessment for larger values of K. The candidate permutations to reverse label switching are treated as missing, and consequently the overall accuracy for such algorithms will be zero and the MSI will be 1. The contribution that missing data has on the MSI is highlighted in Figure 5.3 by black outlines. For the PU and ZS algorithms, missing data may arise even when the algorithm is run, as described in Section 5.4.

In scenario 1, each algorithm achieved perfect accuracy for cases $K = \{3, 4, 5\}$, as indicated by the zero-valued MSIs in Figure 5.3a. For $K = 2$, the KL, BMP, and ECR 1 algorithms only achieved a 50% success rate. This surprising result is the consequence of a rather peculiar phenomenon, which is deferred to the Discussion. For $K > 5$, there is some overlap between the components, and this has a noticeable effect on accuracy. For $K = 6$, ZS was the most affected, while the remaining algorithms continued to perform with perfect or near-perfect accuracy. For $K = 7$, the BM algorithm needed to be run multiple times due to very poor accuracy on the first run, but even the highest accuracy achieved out of ten runs was relatively poor. The accuracy of the DB algorithm also wavered for $K = 7$. For $K = 8$, the BM algorithm suffered the same fate while the other algorithms performed exceptionally well. The most revealing results occur when $K = 9$. Here, the BM algorithm was able to produce fairly accurate results, albeit after 5 runs. The KL, PRA, BM, BMP, DB iterative, and ZS 2 algorithms performed exceptionally well, all achieving $MSI < 0.05$, while the other algorithms produced a large number of incorrect permutations or large number of missing values. This includes the SJW algorithm which was not run for $K > 6$ due to prohibitive computational costs. However, the MSI for the ECR, ECR 1, DB, and ZS algorithms were still moderate ($MSI < 0.11$), which indicates that the effect of the incorrect permutations were marginal.

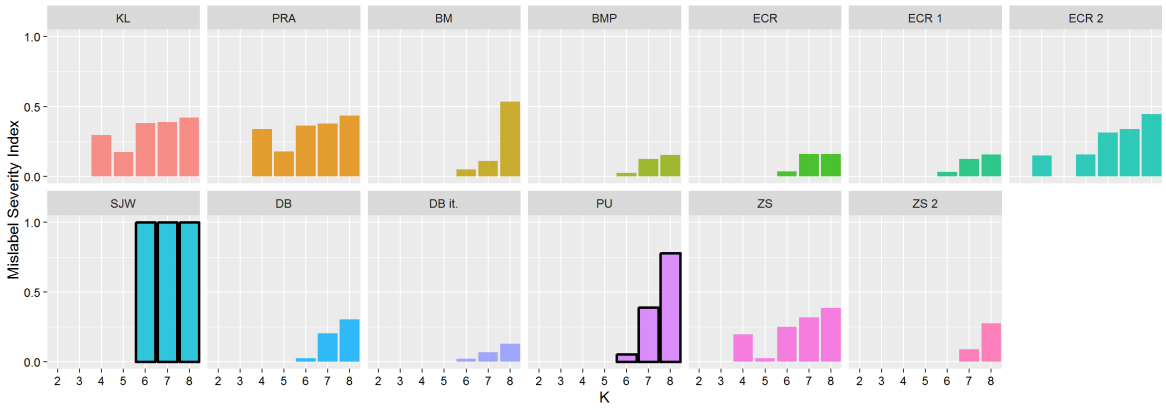
Figure 5.4 further illustrates the differences in accuracy between the algorithms in scenario 1, and the effect that incorrect permutations has on statistical inference. Figure 5.4a shows the accuracy $A^{(m)}$, $m = \{1, \ldots, 1000\}$ for the PU, ZS, and ZS 2 algorithms, while Figure 5.4b shows the effect of these relabelling algorithms on reconstructing the posterior estimate of the Poisson component-specific parameters $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$. The lower accuracy of the ZS algorithm is evident in the posterior estimate of $\boldsymbol{\eta}$. Some label switching between components



(a) Scenario 1



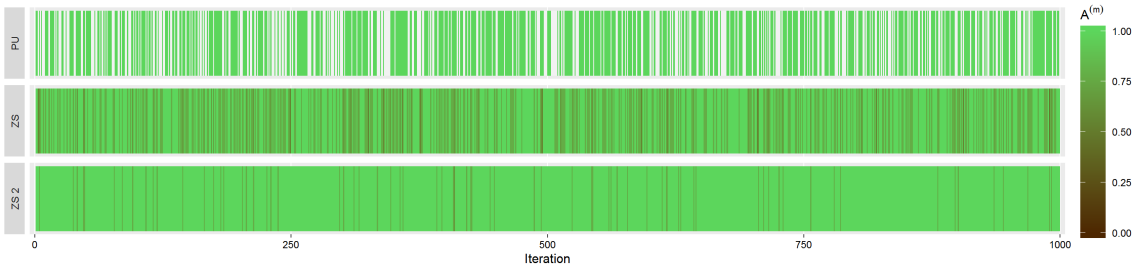
(b) Scenario 2



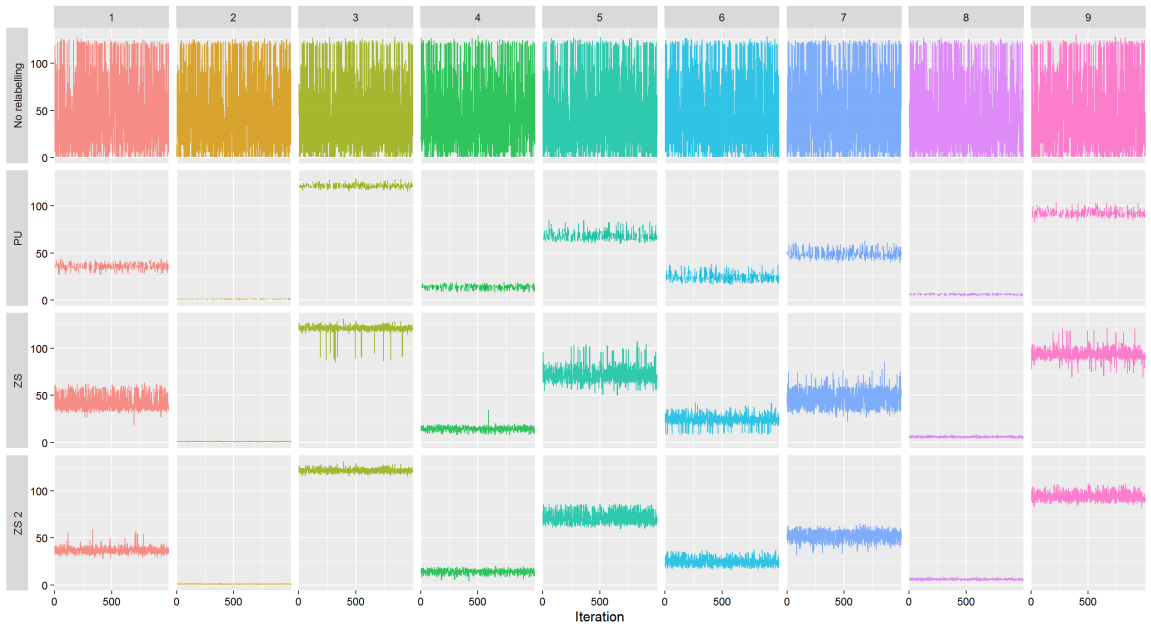
(c) Scenario 3

Figure 5.3: The MSI, given by Equation (5.13), for each algorithm for different values of K in (a) scenario 1, (b) scenario 2, and (c) scenario 3. The results for $K = 2$ and $K = 3$ are omitted in (b) since all algorithms achieved perfect accuracy. The bars with black outlines indicate the contribution of missing values to the MSI.

1 and 7, and components 3, 5, and 9 still appears to be present, increasing the uncertainty of this estimate. The missing values produced by the PU algorithm effectively reduced the posterior sample size by about one third. ZS 2 relabels with near-perfect accuracy, so the resulting estimate of $\boldsymbol{\eta}$ after relabelling is very close to how they estimate would appear in the absence of label switching. These three relabelling algorithms produce viable results, but a smaller MSI will lead to better posterior estimates.



(a) Accuracy, $A^{(m)}$, for the PU, ZS, and ZS 2 algorithms



(b) Posterior sample after relabelling

Figure 5.4: (a) The accuracy $A^{(m)}$ for $m = 1, \ldots, 1000$ for selected algorithms in scenario 1 when $K = 9$, and (b) visual representation of the posterior estimates of $\boldsymbol{\eta}_k$, $k = 1, \ldots, 9$, before and after relabelling.

In scenario 2, most algorithms achieved perfect or near-perfect accuracy for $K = \{2, 3, \ldots, 10, 15, 20\}$, as shown in Figure 5.3b. The only notable exceptions were the PU algorithm for $K = 7$ ($MSI = 0.116$), which was exclusively due to missing values, the BM algorithm, which generally performed worse as K increased, and the SJW algorithm which was not run for $K > 5$. For $K = 50$, inaccuracies were observed across the board, but these inaccuracies translated into large MSIs ($MSI > 0.1$) for only four algorithms: BM, SJW, PU, and ZS. The

KL and ZS 2 algorithms outperformed the other algorithms, each having an $MSI < 0.01$. The results for $K = 100$ indicated fewer iterations with perfect accuracy for all algorithms. In addition to SJW and ZS, the KL and BM algorithms were excluded due to high computation costs, and this time the PU algorithm only retained 4 iterations, 2 of which contained incorrect permutations. The best performers were ZS 2, PRA, DB iterative, ECR, and BMP. These results for scenario 2 are summarised in Figure 5.3b.

Scenario 3 is designed to test how well the algorithms determine the permutations when the number of mixture components differs to the true number of subpopulations in the data. The results for this scenario are summarised in Figure 5.3c. It is noteworthy that the KL, PRA, ECR 2, and ZS algorithms produced incorrect permutations even when $K = G = 5$. This is likely to be the result of some overlap between the subpopulations in combination with the model parameters being estimated with higher uncertainty than in scenarios 1 and 2. Like PRA and ZS, ZS 2 makes extensive use of the model parameters to determine the permutations, but even with a large degree of uncertainty in the parameter estimates, ZS 2 is able to identify the permutations correctly. When $K < G$, the algorithms generally perform well. In contrast, when $K > G$ the MSIs tend to be very large compared to the results from scenarios 1 and 2 when the same values of K are considered. This suggests that the larger MSIs are not simply due to the difficulty in untangling permutations with more indices, but are the result of models which have been misspecified. The algorithms which deal with this misspecification most robustly are BMP, ECR, ECR 1, DB iterative, and ZS 2.

5.7 Discussion

The results from the simulation study show that each algorithm considered in this paper has the potential to reverse label switching, but the accuracy of determining the permutations can vary quite substantially, and a higher computational cost does not always translate into higher accuracy. The results are fairly consistent across the three scenarios: most algorithms perform well in both accuracy and speed for small values of K, but as K increases, the same weaknesses are revealed and the less proficient algorithms fail in a predictable manner. The simplicity of the ECR algorithm makes it exceptionally fast in any scenario. ZS 2 is a close contender for the fastest algorithm, overtaking the PRA algorithm for large K, and performs only marginally slower than ECR. In general, ZS 2 was slightly more accurate than ECR, and more robust to

misspecification than PRA.

Scenario 1 produced some surprising results when $K = 2$, namely unexpectedly poor accuracy by KL, BMP, and ECR 1 – three algorithms which typically showed a high degree of accuracy. The phenomenon giving rise to these results will only occur under the following conditions: the component densities are so well separated that the estimated labels $\boldsymbol{z}^{(m)}$ are consistent for all $m = \{1, \ldots, M\}$, notwithstanding label switching; and the exchangeability of the prior causes the permutations to occur with exactly equal frequencies. In essence, the information that these algorithms rely on to reverse label switching provides no means of discriminating between an iteration which has switched labels and one which has not. Under these conditions, these algorithms deem that all iterations experienced label switching, and thus the resulting estimates are just as scrambled as before relabelling, albeit with the reversed labels. This phenomenon is unlikely to happen in practice, especially for $K > 2$, but in the event it does, it will not be apparent prior to relabelling since the permutations are unknown, and thus this phenomenon presents a genuine hazard for these three algorithms.

The solutions determined by KL, BM, BMP, ECR 1, ECR 2, SJW, and DB iterative algorithms may be locally optimal due to the dependence on initial values and iterative nature of finding the solution. Thus re-running these algorithms with different initialisations may lead to higher accuracy than that which was obtained in the simulation study. For a given scenario and value of K, it is apparent that whatever gains in accuracy were possible must be minor, judging by the consistency in results and comparability in accuracy of the remaining algorithms. These potential gains in accuracy do not warrant the additional computational costs that is required. The only exception is the BM algorithm which has a high tendency to reach locally optimal solutions which are far from globally optimal. The computational cost of running BM once, however, already makes this algorithm an unattractive choice.

There are several algorithms in the recent literature which were not reviewed in this paper. This includes the allocation variable-based probabilistic relabelling algorithm proposed by Pan et al. (2015) and two EM type algorithms proposed by Yao (2013). The algorithms selected for review in this paper were chosen based on several criteria, including their prominence in the literature, novelty, and ease of implementation. The importance of this last criterion should be emphasised – it is one thing for an algorithm to perform well, but if it cannot be easily implemented or understood, it is likely to be overlooked by practitioners in favour of a simpler

algorithm. For example, the SJW algorithm requires the practitioner to determine the full log-likelihood, and such effort may be a great deterrence.

Although no algorithm is flawless, some algorithms appear to be obsolete since the introduction of improved versions, and some variants appear to offer little or no advantage. Specifically, the DB iterative algorithm is slower than the two-stage DB algorithm, but offers only marginal improvements in accuracy. Likewise, the ECR 1 and ECR 2 algorithms were designed to improve accuracy of the ECR algorithm, yet the results from the simulation study indicated that the ECR algorithm was at least as accurate, and computationally faster. ZS 2 can also be viewed as an improvement on the PRA and ZS algorithms.

The simulation study presented in this paper was restricted to scenarios which highlight differences in accuracy, computational efficiency, and robustness to model misspecification. Additional simulation studies may provide further insight on the differences between these algorithms, for example, how the algorithms perform for multivariate models, or when K is treated as an additional unknown parameter.

No doubt new algorithms will continue to be developed, and it will be important to conduct further simulation studies to assess their usefulness as a solution to the label switching problem. The algorithms presented in this paper should serve as a useful benchmark.

5.8 Conclusion

This paper has presented an overview of the label switching problem, reiterated the general framework for relabelling algorithms proposed by Stephens (2000b), proposed a new relabelling algorithm, and through a simulation study, compared and contrasted 13 relabelling algorithms, highlighting differences in computational efficiency, accuracy in identifying the correct permutations, and robustness to model misspecification. Confusion surrounding the extent of the label switching problem and the necessity of label switching was addressed. The misconception that the PRA algorithm is necessarily inefficient was also laid to rest.

The results from the simulation study were able to distinguish the relabelling algorithms based on their computational efficiency, accuracy, and robustness to misspecification. Although no algorithm is supreme, ZS 2 clearly has the potential for a first-rate relabelling algorithm.

Statement of Contribution of Co-authors for Chapter 6

The authors listed below have certified that:

1. they meet the criteria for authorship in that they have participated in the conception, execution, or interpretation, of at least that part of the publication in their field of expertise;
2. they take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. there are no other authors of the publication according to these criteria;
4. potential conflicts of interest have been disclosed to (a) granting bodies, (b) the editor or publisher of journals or other publications, and (c) the head of the responsible academic unit, and
5. they agree to the use of the publication in the student's thesis and its publication on the QUT's ePrints site consistent with any limitations set by publisher requirements.

This paper was submitted to the *International Journal of Health Geographics* for publication and is currently under review. The title of this paper is *Spatial Smoothing in Bayesian Models: A Comparison of Weights Matrix Specifications and their Impact on Inference*.

Contributor	**Statement of contribution**
E. W. Duncan Signature and date:	Proposed and conducted the research, wrote the code for simulation study, performed statistical analysis, wrote the manuscript, revised the manuscript as suggested by co-authors.
N. M. White	Supervised research, provided comments on and helped revise manuscript.
K. L. Mengersen	Supervised research, provided comments on and helped revise manuscript.

Principal Supervisor Confirmation

I have sighted email or other correspondence from all co-authors confirming their certifying authorship.

Name: _____ Signature: _____ Date: _____

Chapter 6

Spatial Smoothing in Bayesian Models: A Comparison of Weights Matrix Specifications and their Impact on Inference

---

Preamble

This chapter relates to the fourth research objective. This chapter investigates the ways in which spatial smoothing is incorporated into a Bayesian model and what effect that this has on model fit and predictive performance. Spatial smoothing accounts for spatial autocorrelation by smoothing the estimates of a spatial random field towards the mean value of the respective 'neighbours'. The influence that each random variable in the spatial random field bears on its neighbours can be summarised in the form of a matrix of weights. The ideal definition of 'neighbours' and weights matrix specification is an open research problem.

This chapter begins with a literature review of spatial smoothing techniques followed by a description of a statistical framework commonly used in disease mapping. Various specifications of the weights matrix found in the literature together with some new proposals are described and subsequently applied to four data sets.

The statistical model used in this chapter can be viewed as a very simple, special case of the models discussed in Chapters 3 and 4. One of the simplifications is that in this chapter, the data are purely spatial rather than spatio-temporal, and therefore the data and model parameters are only indexed by i to denote their association to the i^{th} area. N is still used to denote the number of areas, and y_i and E_i denote the observed and expected counts for the i^{th} area. The

log-relative risk is expressed in the form of a regression model wich accounts for covariate, fixed, and random effects similar to the models in Chapters 3 and 4, but the notation for these parameters is different. Naturally, the prior distributions are also different to reflect prior beliefs about current data being analysed.

The extensions to the research presented in this chapter are not in relation to the statistical models per se, but rather the specification of the weights matrix. The weights matrix used to perform smoothing on the spatial random effects in the models in Chapters 3 and 4 is one of the specifications investigated in this chapter. This chapter provides 16 alternative specifications.

6.1 Introduction

Consider the problem of mapping disease incidence, prevalence, or mortality with the aim of identifying spatial patterns of the underlying relative risk. Such analyses may identify 'hot spots', provide insight into the causal processes, and guide researchers' efforts in further investigations (Bernardinelli et al. 1997; Huque et al. 2016; Langford et al. 1999; Xia and Carlin 1998). When analysing spatial data, it is important to account for spatial autocorrelation and sampling variability. Spatial autocorrelation refers to the idea that observations taken at locations near to each other tend to be similar (Assunção and Krainski 2009; Bell and Bockstael 2000; Florax and Nijkamp 2003; Wall 2004), while sampling variability refers to differences between areas due to small populations or heterogeneity of individuals within areas, for example (Bernardinelli et al. 1997; Conlon and Waller 1999; Earnest et al. 2007; Fahrmeir and Kneib 2011; Huque et al. 2016; Langford et al. 1999; Morrison et al. 2012; Shaddick and Zidek 2016; Wakefield and Elliott 1999; Xia and Carlin 1998). This is especially true for rare diseases, and when the areal units contain a small population (Langford et al. 1999; Xia and Carlin 1998).

Numerous statistical models have been developed to address these issues of spatial data. A general overview of some well-known models can be found in Congdon (2010); Cressie (1993); Diniz-Filho et al. (2009); Dormann et al. (2007); Fahrmeir and Kneib (2011), and Wheeler (2013). Each of these models has the common aim of accounting for spatial autocorrelation and sampling variability so as to satisfy the model assumptions and reduce uncertainty of the estimates. In many disease mapping studies, the modelling approach has been to model the observed data using a Bayesian generalised linear mixed model (GLMM) (Breslow and

Clayton 1993; Langford et al. 1999; McCulloch 1999; Rasmussen 2004), and account for spatial autocorrelation through spatial random effects in the linear predictor. A fairly standard approach is to use a three-stage random effects model: in the first stage, the likelihood for the data is specified by some distribution belonging to the exponential family; in the second stage, the expectation of the response variable is related to the linear predictor through a link function; and the parameters in the linear predictor are assigned prior distributions as the third stage. Examples of this framework can be found in recent papers such as Best et al. (2005); Johnson (2004); Morrison et al. (2012), and Pascutto et al. (2000), and is also described in Banerjee et al. (2014).

Common choices of prior distributions for random effects include the conditional autoregressive (CAR) model (Besag 1974; Besag et al. 1991) and the simultaneous autoregressive (SAR) model (Anselin 1988; Cressie 1993). Both models make use of a spatial weights matrix to quantify the relative influence that the random effects have on each other (Assunção and Krainski 2009; Banerjee et al. 2014; Best et al. 2001; Morrison et al. 2012). The effect that the weighting scheme has on the degree of smoothing and the analysis in general has received very little attention in the literature (Earnest et al. 2007). Some studies have considered multiple weighting schemes, for example Morrison et al. (2012) and Getis and Aldstadt (2008), but the motivation for doing so is usually to improve model fit and predictive ability. The results from these studies do, however, indicate that different weighting schemes can have a substantial impact on the analysis (Earnest et al. 2007; Morrison et al. 2012).

The aims of this paper are to 1) review the different specifications of the spatial weights matrix found in the literature; 2) choose a selection of weights matrices for comparison; and 3) use a GLMM with spatial smoothing to analyse a real and synthetic data set with the chosen weights matrices to compare and contrast the effect that they have on spatial smoothing, and the performance of the model.

Griffith (1996) provides several guidelines for defining the weights, two of which are particularly relevant. The first recommendation is that it is indeed better to apply smoothing than no smoothing at all, which reiterates previous statements about the necessity of smoothing. The second recommendation is that it is generally better to have a small number of neighbours, around 4 to 6. Getis and Aldstadt (2008) add that fewer neighbours is particularly appropriate if the data exhibits spatial heterogeneity. The number of neighbours and assigned weights are

often chosen arbitrarily, but more systematic approaches have been proposed. For example, it may be possible to infer a reasonable neighbourhood and weighting scheme from the scientific context (Dormann et al. 2007) or by filtering the spatial effects from the data (Griffith 1996). Other alternatives include examining the correlogram of relative risks over geographic distance Earnest et al. (2007) or even using trial and error with respect to the number of neighbours induced (Dormann et al. 2007).

The remainder of this paper is outlined as follows. Section 6.2 describes the proposed methods. This includes a detailed description of the CAR model specification including prior distributions, several alternate specifications of the spatial weights matrix, and a summary of the two data sets used in the analysis. Section 6.3 contains the results from the analysis. These results are discussed in Section 6.4.

6.2 Methods

6.2.1 CAR Model

For rare and non-infectious diseases, the true incidence or number of deaths in a given area is typically estimated by assuming a Poisson distribution,

$$y_i \sim \text{Po}(E_i \eta_i),$$

where y_i is the number of observed cases in area i, for $i = 1, \ldots, N$, E_i is the expected number of cases, and η_i is the area-specific relative risk. E_i can be regarded as on offset to account for differences in population, age, and/or risk factors between areas (Bernardinelli et al. 1997; Fahrmeir and Kneib 2011; Langford et al. 1999; Mugglin et al. 1999; Wakefield and Elliott 1999). If data on such characteristics are known, then an alternative is to include these data as additional covariates (Fahrmeir and Kneib 2011). Otherwise, E_i is generally computed as

$$E_i = \frac{\sum_i y_i}{\sum_i P_i} P_i \tag{6.1}$$

where P_i is the population at risk, which is known as internal standardisation (Fahrmeir and Kneib 2011; Mugglin et al. 1999; Pascutto et al. 2000; Xia and Carlin 1998). The area-specific

log-relative risk is then expressed as a regression model

$$\log(\eta_i) = \alpha + \beta x_i + \gamma_i + \varepsilon_i \tag{6.2}$$

where α is the overall fixed effect, β is the effect of the spatial covariate x_i, and γ_i and ε_i are structured and unstructured spatial random effects respectively. The unstructured spatial random effects are simply the errors which should be independent and identically distributed white noise with unknown variance σ_ε^2 if the spatial autocorrelation is adequately accounted for by the spatial covariate effect and structured spatial random effect (Fahrmeir and Kneib 2011; Huque et al. 2016). The structured spatial random effects are assumed to have arisen from a Gaussian Markov random field which is consistent with the belief that neighbouring areas have similar spatial effects. This spatial dependency is formalised by imposing the intrinsic conditional autoregressive (ICAR) prior distribution, proposed by Besag (1974) and Besag et al. (1991), on the structured spatial random effects

$$\gamma_i|\boldsymbol{\gamma}_{\setminus i} \sim \mathcal{N}\left(\frac{1}{\sum_j w_{ij}}\sum_j w_{ij}\gamma_j, \frac{\sigma_\gamma^2}{\sum_j w_{ij}}\right), \quad \text{for } i = 1, ..., N$$

where $\boldsymbol{\gamma}_{\setminus i}$ denotes the vector of structured spatial random effects for each area excluding area i, and w_{ij} is the element of a symmetric weights matrix $\mathbf{W}$ corresponding to row i and column j (Congdon 2010; Fahrmeir and Kneib 2011; Mugglin et al. 1999). To preserve the identifiability of these random effects, the ICAR prior is constrained by $\sum_{i=1}^N \gamma_i = 0$ (Xia and Carlin 1998). For the parameters α and β, weakly informative priors are chosen,

$$\alpha \sim \mathcal{N}(0, 100),$$
$$\beta \sim \mathcal{N}(0, 100).$$

In accordance with the assumption that the errors are uncorrelated, appropriate prior distributions for the unstructured spatial random effect and associated variance are

$$\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

$$\sigma_\varepsilon^2 \sim \mathcal{N}(0, 20)\mathbb{I}_{(0,\infty)}.$$

A suitable prior distribution for σ_γ^2 is less straightforward. Earnest et al. (2007) caution that the prior for the variance term of the CAR model can have a noticeable influence on the estimated spatial random effect. Commonly suggested priors for variances or standard deviations, include gamma, inverse gamma, half-Cauchy, and the uniform distribution (Earnest et al. 2007; Fahrmeir and Kneib 2011; Gelman 2006). If the neighbours are defined appropriately, then the structured spatial random effects for those neighbours should be similar, and therefore the distribution of the variance should have the bulk of the density close to zero. Therefore, a prior which seems consistent with this prior belief is the following gamma distribution:

$$\sigma_\gamma^2 \sim \mathcal{G}(0.5, 0.05).$$

The model presented so far is deliberately simplistic, as the focus of this paper is about the influence of the weights rather than model utility or complexity. However, this base model can be easily adapted to more complex situations. For example, if additional covariates are available, these can be incorporated as additional terms in Equation (6.2). Extensions to spatio-temporal data are also possible, where spatial and temporal smoothing can be employed separately or jointly, depending on the definition of neighbours (Fahrmeir and Kneib 2011).

6.2.2 Weights Matrix Specifications

As mentioned in the previous section, it is common for the weights to be defined as

$$w_{ij} = \begin{cases} 1 & \text{if areas } i \text{ and } j \text{ are neighbours} \\ 0 & \text{otherwise} \end{cases}. \tag{6.3}$$

The concept of 'neighbours' requires further clarification. Often neighbours are defined to be areas which share a common boundary, that is, are adjacent (Bernardinelli et al. 1997; Congdon 2010; Earnest et al. 2007; Fahrmeir and Kneib 2011; Morrison et al. 2012; Xia and Carlin 1998). In this case, the weights matrix may be referred to as the first-order adjacency matrix. Note that areas are not considered to be neighbours of themselves, and thus the elements on the diagonal of this matrix are zero by definition (Getis and Aldstadt 2008). If the areas comprise a regular grid, then the neighbourhood is comparable to restricted forms of the rook and queen chess moves, depending on whether areas which only share a common vertex are considered

neighbours (Earnest et al. 2007; Getis and Aldstadt 2008). The neighbourhood can be extended to include neighbours of neighbours, resulting in a second-order adjacency matrix, and so on. More generally, we can define an n^{th}-order adjacency matrix as

$$w_{ij} = \begin{cases} \omega_k & \text{if areas } i \text{ and } j \text{ are } k^{\text{th}}\text{-order neighbours}, k \leq n \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$ is the vector of weights corresponding to each order. Typically, larger weights are assigned to the closest neighbours, so $\boldsymbol{\omega}$ may be defined as a decreasing function of the order, e.g. $\omega_k = \exp\left((k-1)/(n-1)\right)$. The drawback to these adjacency-based approaches is that they do not account for areas of different sizes. One alternative is to define neighbours as a function of geographical distance. Distance between areas i and j is often measured as the Euclidean distance between their respective centroids (Dormann et al. 2007; Earnest et al. 2007; Fahrmeir and Kneib 2011; Johnson 2004; Morrison et al. 2012). Fahrmeir and Kneib (2011) point out that using the Euclidean distance implies the assumption of isotropy, that is, the influence between areas i and j is the same in both directions. For the purpose of defining a weights matrix to be used in the ICAR model, this is actually a desirable property since the weights matrix must be symmetric in order for the structured spatial random effects to yield a Markov random field (Fahrmeir and Kneib 2011). Let $\{d_{ij}\}$ denote these distances. A common definition for distance-based weights is the inverse distance power function:

$$w_{ij} = (1/d_{ij})^k \tag{6.5}$$

for positive integer k, with k often taken to be 1. The larger the exponent k, the greater the influence of those areas that are close relative to those further away (Earnest et al. 2007; Getis and Aldstadt 2008). Another definition is the exponential decay function

$$w_{ij} = \exp(-\lambda d_{ij}), \quad \lambda > 0 \tag{6.6}$$

where λ controls the rate of decay (Congdon 2010; Earnest et al. 2007; Fahrmeir and Kneib 2011). This decay parameter is often taken to be 1, as in Fahrmeir and Kneib (2011). Other authors have suggested more pragmatic approaches to determining this parameter. For example, Earnest et al. (2007) recommends setting $\lambda = 10$ based on the autocorrelation of the relative

risks. A similar definition is the Gaussian decay function

$$w_{ij} = \exp\left(-\frac{d_{ij}^2}{2b^2}\right), \quad b > 0 \tag{6.7}$$

where the inverse of the bandwidth parameter b determines the rate of decay (Congdon 2010; Wheeler 2013).

Other distance-based weights include the bi-square, bi-square nearest neighbour, tri-cube, and spherical kernel functions (Congdon 2010; Getis and Aldstadt 2008; Wheeler 2013), and further definitions are provided in Dormann et al. (2007); Earnest et al. (2007), and Mugglin et al. (1999). Distance-based weights can also be found in the literature on geospatial models, for example Cressie (1993) and Diggle (2013). All these adjacency and distance-based definitions may be collectively referred to as geometric weights. They share the assumption that closer areas have greater influence. However, it is conceivable that areas which are relatively far apart might have a greater influence on each other than areas which are simply nearby geographically. For example, areas with similar covariate values might be expected to have similar relative risk estimates. This yields alternative versions of the distance-based definitions of weights, where the distance d_{ij} is replaced by δ_{ij}, the absolute difference between the covariate values of areas i and j:

$$\delta_{ij} = |x_i - x_j|. \tag{6.8}$$

The smoothing that results from such weights matrices may be still regarded as spatial smoothing since the spatial random effects are smoothed towards the mean value of their neighbours, albeit neighbours which may be far apart geographically. The key difference is that the smoothing is conducted on the covariate space rather than the parameter space. This idea is not new, but it is seldom considered, and very rarely pursued in statistical analysis. For example, Dormann et al. (2007) mention that weights matrices can be defined in terms of 'environmental distance' as opposed to geographical distance, but only the latter is used in their analysis.

Two references which do actually use smoothing on the covariate space are Kuhnert (2003) and Earnest et al. (2007). However, the weights defined in the latter are a function of both the geographic distances and covariate distances, specifically

$$w_{ij} = \frac{1}{d_{ij}\delta_{ij}}.$$

Smoothing on the covariate space can be viewed as a more flexible alternative to smoothing on the parameter space for two reasons. First, it relaxes the assumption that the weights are (only) a function of geographic distance. Second, it relaxes the assumption that a large difference in the relative risk between adjacent areas is not possible. For example, if the covariate values corresponding to two neighbouring areas are dissimilar and this is reflected in the weights, then the relative risk estimates for these two areas should potentially be dissimilar too, depending on how accurate the covariate is as a predictor in the model. This second point can be viewed as a simpler alternative to adaptive Markov random fields (Brezger et al. 2007; Fahrmeir and Kneib 2011) and areal wombling Lu et al. (2007) where the weights are treated as a random variable and therefore allowed to vary.

Aside from a few general recommendations already mentioned, the results from previous studies are not particularly helpful in terms of providing advice on which weighting schemes should be considered for analysis. In fact, the results often suggest conflicting ideas. For example, Morrison et al. (2012) found that the weights based on first-order neighbours and four nearest neighbours produced the best model fit, while weights based on second-order neighbours performed worse. Conversely, of the geometric type weights explored by Getis and Aldstadt (2008), the 'rook' specification which contains at most four neighbours was the least effective, while the 'queen' specification improved the model fit. Similarly, Getis and Aldstadt (2008) found that the inverse distance power function specification, given by Equation (6.5) performs poorly, while Earnest et al. (2007) found that the inverse distances were considerably better than the 'rook' and 'queen' specifications. Both analyses agree, however, that the inverse distance power function appears to perform better when the exponent is 2 compared to the other values tested, namely 1, 3, or 5.

It should be pointed out that the weights matrix $\mathbf{W}$ is typically row-standardised such that each row sums to 1. This helps with interpretation of the parameters and seems to be preferred over global-standardisation (Getis and Aldstadt 2008). Note that the software used in our analysis automatically row-standardises the weights matrix; the following definitions are the unstandardised versions.

6.2.3 Study Design

Based on the aims of this paper and the recommendations from the literature, 17 definitions of the weights were chosen for the analysis. The first two sets of weights are based on neighbourhood adjacency, specifically first-order neighbours given by Equation (6.3) where neighbours are defined as areas which share a common boundary or vertex, and third-order neighbours given by Equation (6.4) where $k \leq 3$, and $\boldsymbol{\omega} = (e^0, e^{-0.5}, e^{-1})$. The corresponding models are denoted as A1 and A2.

The following distance-based weights are also considered: inverse distance power function given by Equation (6.5) for exponents $k = 1, 2,$ and 5, and the decay functions given by Equations (6.6) and (6.7). Rather than fix the decay and bandwidth parameters at some arbitrary value, these are computed as a function of the mean distance between spatial units, specifically

$$\lambda = \frac{10}{\underset{i,j}{\mathrm{E}}(d_{ij})}$$

$$b = \lambda^{-1}.$$

As noted by Getis and Aldstadt (2008), the scale characteristics of data are important. If the decay parameter value were fixed at 10 instead, this would result in a very different weights matrix if the areal units were large administrative regions spanning hundreds of kilometres compared to much smaller areas, including artificial rasters of areas which may be associated with relative distance only. This is also true if the units of geographical distance are changed from kilometres to metres, for example. The justification for the definitions used here is that it alleviates this dependency on the scale of the data and should be applicable to any spatial data set regardless of the scale. The value of 10 in the numerator was determined by trial and error such that for the four data sets analysed, the number of non-negligible neighbours appeared to be fairly consistent for a given model. These distance-based weights models will be denoted by D1 through D5. To compare the effect of smoothing on the covariate space, five new models are created by replacing the geographic distances, including those in the calculation of the decay and bandwidth parameters, with the covariate distances given by Equation (6.8). These models will be denoted C1 through C5.

The hybrid approach of Earnest et al. (2007) is simply the inverse distance specification, where the distance is the product of both the geographic and covariance distances. This approach

is certainly not limited to the inverse distance weighting scheme, however. In fact, we also include a hybrid version of each of the five distance-based weights mentioned above, where the distances are replaced by $d_{ij}\delta_{ij}$. The resulting models will be denoted H1 through H5.

The inverse distance power function will produce non-finite values if the distance between two areas is zero. This is theoretically possible for geographic distances, for example when one area is nested within another thus having the same centroid location, but is also applicable to covariate distances, and consequently the hybrid approach. To avoid this issue, Earnest et al. (2007) suggest adding a small correction to zero counts. However, the resulting weight will be highly dependent on that arbitrary value. The approach we adopt is to compute the weights without modification, then for each row of the weights matrix, replace the non-finite weights by the maximum finite value in that row. If graph representing the underlying spatial field is undirected, as is presumed here, then the weights matrix must be symmetric. To retain symmetry, the lower triangular portion of $\mathbf{W}$ is replaced by the upper triangular portion of $\mathbf{W}^{\mathrm{T}}$.

Figure 6.1 illustrates the differences and similarities of these weighting schemes for the Scottish lip cancer data set, described below. In general, the number of neighbours for each area for all variants of Models D, C, and H will be $N - 1$, albeit the weights will be very close to zero for some neighbours, effectively reducing the dependency to a subset of non-negligible neighbours. The average number of non-negligible neighbours for the Scottish lip cancer data set, as a function of the threshold defining which neighbours are negligible, is shown in Figure 6.2. If it is better to have a small number of neighbours as the literature suggests, then it might be expected that weighting schemes which result in a small number of non-negligible neighbours, for a given threshold, perform better.

As a benchmark for the usefulness of including spatial smoothing, a model which does not account for spatial autocorrelation is also included, bringing the total number of models to 18. This model, denoted B, has exactly the same specification except that γ_i is removed from Equation (6.2).

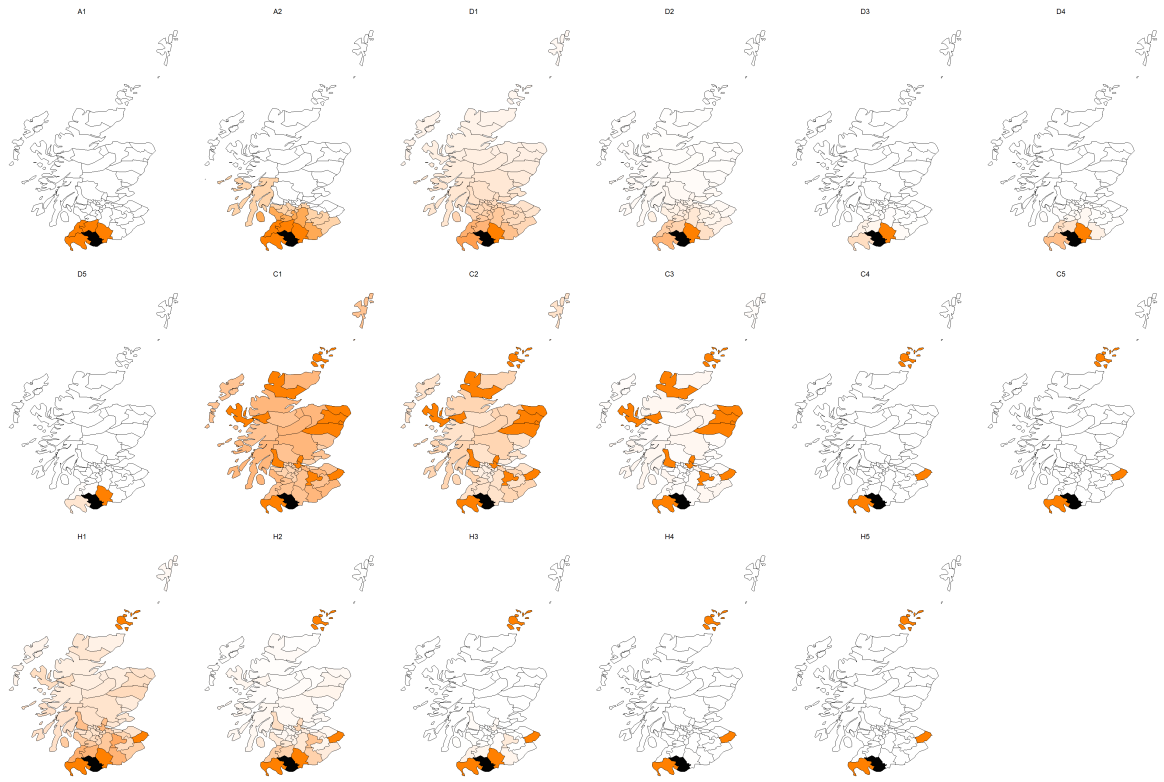


Figure 6.1: Unnormalised weights for 'area 43', shaded black, in the Scottish lip cancer data set. Darker orange areas are neighbours with large weights, signifying a high degree of correlation with area 43, as defined by the weights matrix of the respective models, while white areas have weights very close to zero, or in the case of models A1 and A2, equal to zero.

6.2.4 Data

Four data sets are analysed using the GLMM described above. The first data set is the well-known Scottish lip cancer data set, which has been analysed previously by Breslow and Clayton (1993); Rasmussen (2004), and Spiegelhalter et al. (2002) amongst others. The observed data represents incidence of lip cancer in 56 counties in Scotland. This data set also includes the expected cases which were computed using external standardisation, and a covariate which represents the percentage of the population working in industries that are typically related with high sun exposure, namely agriculture, fishing, and forestry. In both the analysis of Breslow and Clayton (1993) and Rasmussen (2004), the covariate was scaled by a factor of 10, as is done here.

The other three data sets comprise synthetic data where the spatial units are arranged as a 10 by 10 grid. The expected and covariate values are the same across these three data sets, while the observed vales are generated as if arising from 1) a spatial process with no autocorrelation; 2) a random field with strong positive spatial autocorrelation; or 3) a convolution of Gaussian

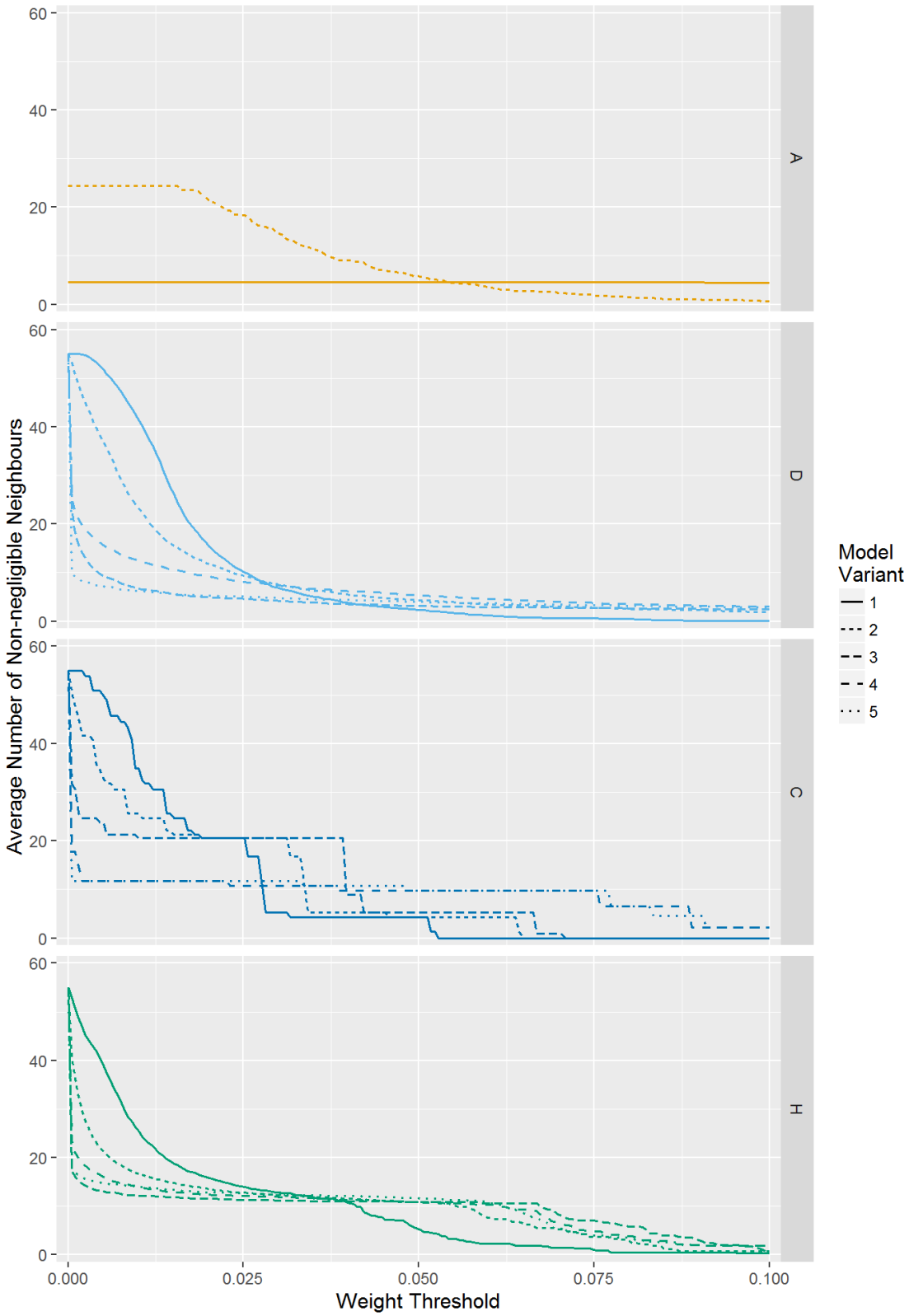


Figure 6.2: Average number of neighbours excluding areas with normalised weights less than the threshold for the Scottish lip cancer data set.

Markov random fields resulting in distinct clusters.

The log-relative risk, covariate, observed, and expected values were simulated as follows. First the expected values E_i in area i were simulated as

$$E_i \sim Gam(2, 2).$$

The covariate values were generated from a superimposition of six different clusters of values defined by Gaussian decay with added noise,

$$\tilde{x}_i \sim \mathcal{N}(0, 0.1)\mathbb{I}_{(0,\infty)}$$
$$\boldsymbol{x} = \tilde{\boldsymbol{x}} + \sum_{r=1}^{6} 2 \exp\left(-\frac{\boldsymbol{d}_{c_r}^2}{2b_r^2}\right)$$

where b_r is the bandwidth parameter for the r^{th} cluster, $b = \{0.7, 0.7, 0.3, 0.7, 1, 0.7\}$, and $\boldsymbol{d}_{c_r}$ is the ${c_r}^{\text{th}}$ row of the $N \times N$ matrix of geographic distances, where $\{c_1, \ldots, c_6\} = \{2, 21, 28, 39, 73, 96\}$. For example, when $r = 3$, $\boldsymbol{d}_{c_r}$ is the 28^{th} row of the distance matrix. Next, the log-relative risks were generated. For the first synthetic data set, the log-relative risks were simulated independently from a Gaussian distribution,

$$\log(\eta_i) \sim (-1, 2).$$

For the second synthetic data set, the log-relative risks were simulated from a Gaussian random field

$$R_j \sim \mathcal{N}(0.3, 0.09) \qquad \text{for } j = 1, \ldots, N$$
$$\log(\eta_i) = \sum_{j=1}^{N} \exp\left(-\frac{d_{ij}^2}{2}\right) R_j.$$

For the third synthetic data set, the log-relative risks were generated from a convolution of two Gaussian Markov random fields, similar to the approach of Devine et al. (1996) and Getis and Aldstadt (2008). The risks in the two clusters, centred at areas 85 and 12, exhibit strong spatial autocorrelation; the risks for all other areas are generated independently, simulating background

noise.

$$R_j \sim \mathcal{N}(0.035, 0.0009) \qquad \text{for } j = 1, \ldots, N$$

$$\log(\eta_i) = \begin{cases} \sum_{j=1}^{N} \exp\left(-\frac{d_{ij}^2}{50}\right) R_j & \text{if } d_{i,85} \leq 2.7 \\ \sum_{j=1}^{N} \exp\left(-\frac{d_{ij}^2}{128}\right) R_j & \text{if } d_{i,12} \leq 1.5 \\ \mathcal{N}(-0.1, 1) & \text{otherwise} \end{cases} .$$

The relative risks were then computed as the exponentiated log-relative risks after adding some effect for the covariate,

$$\eta_i \mapsto \exp\left(\log\left(\eta_i\right) + bx_i\right)$$

where $b = 0.7$ for the first synthetic data set, and $b = 0.4$ for the last two data sets.

Finally, the observed counts were simulated by drawing from the Poisson distribution given by Equation (6.1), using the simulated values for E_i and η_i above.

The R code used to generate these values is provided in Appendix D.1. The data for the Scottish lip cancer and synthetic data sets are summarised in Figure 6.3.

The spatial autocorrelation, or lack thereof, of the observed values for each data set is perhaps not as obvious as revealed by Figure 6.3. In each case, there is one particularly large observation which obscures the closeness between the values in other areas. Several statistics have been developed for the purpose of quantifying spatial correlation. Moran's I statistic (Moran 1950) is commonly used (for example, see Dormann et al. (2007); Getis and Aldstadt (2008); Morrison et al. (2012), and Wheeler (2013). This statistic is a function of the weights matrix, however, and different matrices will produce different results. Moran's I statistic was computed for the observed values in each data set using five different weights matrices, shown in Table 6.1. Based on the consensus of this statistic under various weight specifications, it would appear that the Scottish lip cancer data is not strongly spatially autocorrelated. As expected, the same conclusion holds for the first synthetic data set. The second and third synthetic data sets were generated to exhibit spatial autocorrelation, and this is generally reflected by Moran's I.

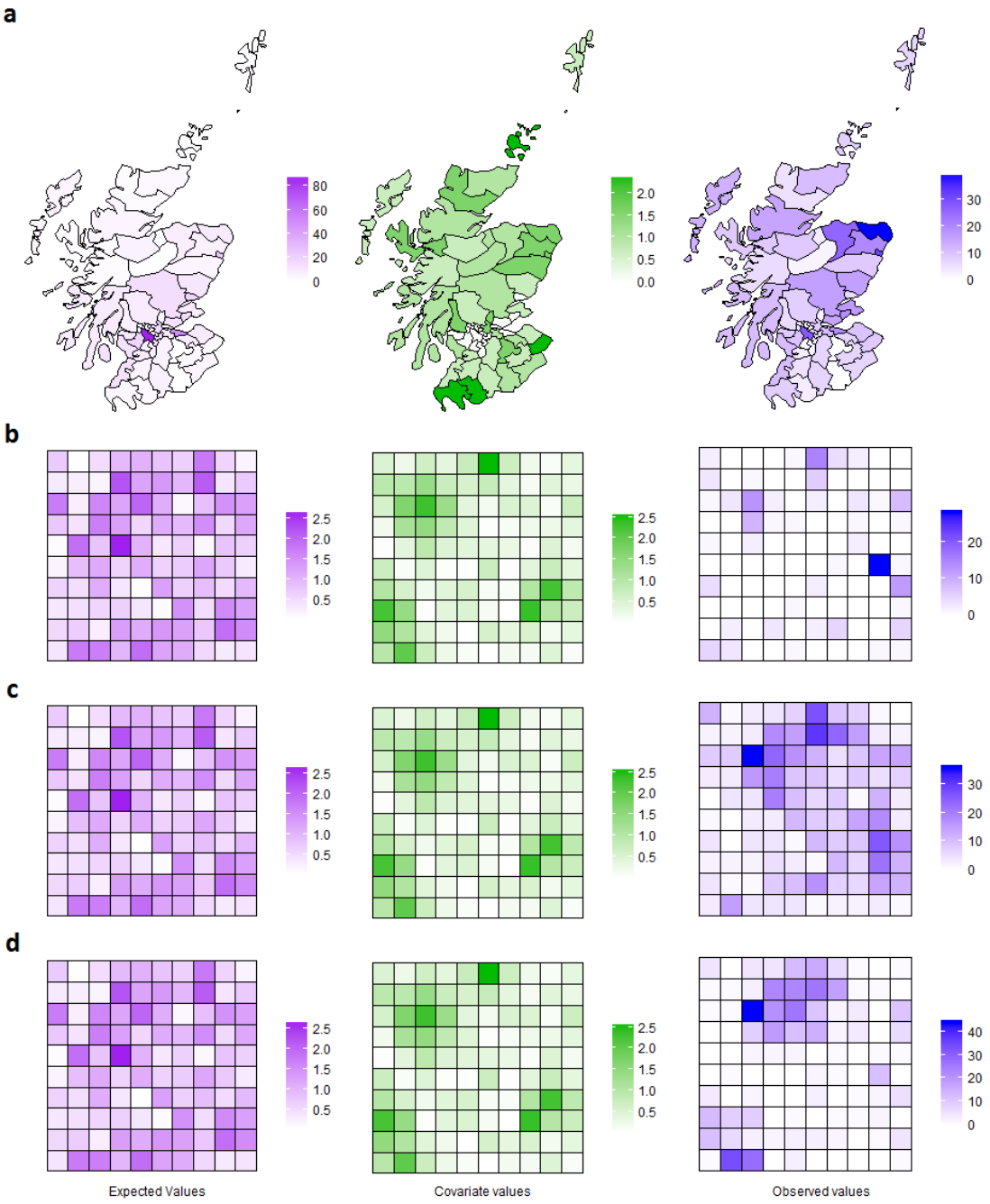


Figure 6.3: A spatial representation of the expected, covariate, and observed values for (a) the Scottish lip cancer data set, (b) the synthetic data set with weak spatial autocorrelation, (c) the synthetic data set with strong positive spatial autocorrelation, and (d) the synthetic data set with two clusters of elevated risk.

Model	Scottish Lip Cancer	Synthetic Data 1	Synthetic Data 2	Synthetic Data 3
A1	0.0509	0.2017	< 0.0001	< 0.0001
A2	< 0.0001	0.4278	< 0.0001	< 0.0001
D2	0.1465	0.4756	< 0.0001	< 0.0001
C3	0.7391	0.9248	0.9955	0.7898
H5	0.7592	0.0009	0.0024	< 0.0001

Table 6.1: Moran's I two-sided p-values for each data set using selected weight specifications defined in the previous section. A p-value close to zero suggests the presence of spatial autocorrelation in the observed data.

6.2.5 Implementation

Each weighting scheme results in a different CAR prior distribution, effectively yielding 18 different models. The model parameters were estimated using Markov chain Monte Carlo (MCMC) sampling, implemented in WinBUGS (Lunn et al. 2000). The remainder of the analysis was performed using the software R (R Core Team 2015). Two parallel MCMC chains were run for 10000 iterations following a burn-in period of 10000 iterations. Convergence of the chains was assessed by visual inspection of the posterior distributions and computation of the GelmanRubin statistic (Gelman and Rubin 1992). The WinBUGS code for the models with and without spatial smoothing are provided in Appendix D.2.

6.2.6 Model Evaluation

The Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002) is often used as a Bayesian measure of model fit and adequacy, compensating for overfitting by the inclusion of a penalty term. A smaller DIC indicates a better model fit. Following the suggestions of Burnham and Anderson (1998) and Spiegelhalter et al. (2002), the DIC is used to compare the models in the following way: models with a DIC within 2 of the 'best' model have a similar model fit, while models with a larger DIC have a decidedly worse model fit. However, there are some concerns about using DIC in a spatial context (see the discussion in Spiegelhalter et al. (2002, pp. 624)). As an alternative means of assessing model adequacy, the spatial autocorrelation in the residuals is measured using Moran's I statistic. If spatial autocorrelation exists within the data, and the model adequately adjusts the log-relative risks, then the residuals ought to be

spatially independent in accordance with the model assumptions. Both of these measures were used to assess and compare model fit and adequacy.

6.3 Results

6.3.1 Analysis of the Scottish Lip Cancer Data

The model evaluation measures for the Scottish lip cancer data set are summarised in Figure 6.4. On the left axis, the DIC for each model is shown. Note that the DIC is similar for each of the models that accounts for spatial autocorrelation. Except for models C3, C5, and H3, the difference between the smallest and largest DIC values was exactly 2, rounding to 3 decimal places. The DIC for model B, on the other hand is at least three times larger than the worst model which does account for spatial autocorrelation, emphasising a very poor model fit. The p-values for Moran's I statistic on the model residuals, shown on the right axis, provides further insight. Despite the comparable model fit for each model with a DIC within 2 of the smallest DIC, Moran's I statistic suggests that the models A2, C4, H4, and H5 account for spatial autocorrelation more satisfactorily in terms of the model assumptions.

The posterior estimates of the parameters for the two models with the smallest DIC and the two models with the largest DIC are summarised in Figure 6.5. The area-specific parameters are ordered in decreasing order of the standardised mortality ratio. Not surprisingly, models H4 and H5 have similar model parameters. The parameters for model B are quite different, not least of all the estimate for the covariate effect β which is negative rather than positive. Without a structured spatial random effect, the model can only capture deviations from the covariate term βx_i through the residuals ε_i. Note that the uncertainty of the parameter estimates is larger than for the other models, as indicated by the wider densities and large credible intervals (CIs), as well as many of the posterior means, both indicative of a model which is not able to explain the spatial variation in the relative risk well. Model C5 accounts for spatial autocorrelation, but unlike models H4 and H5, estimates the spatial autocorrelation to be zero. However, the posterior distribution of the covariate effect is positive, and the model does provide a much better fit than model B.

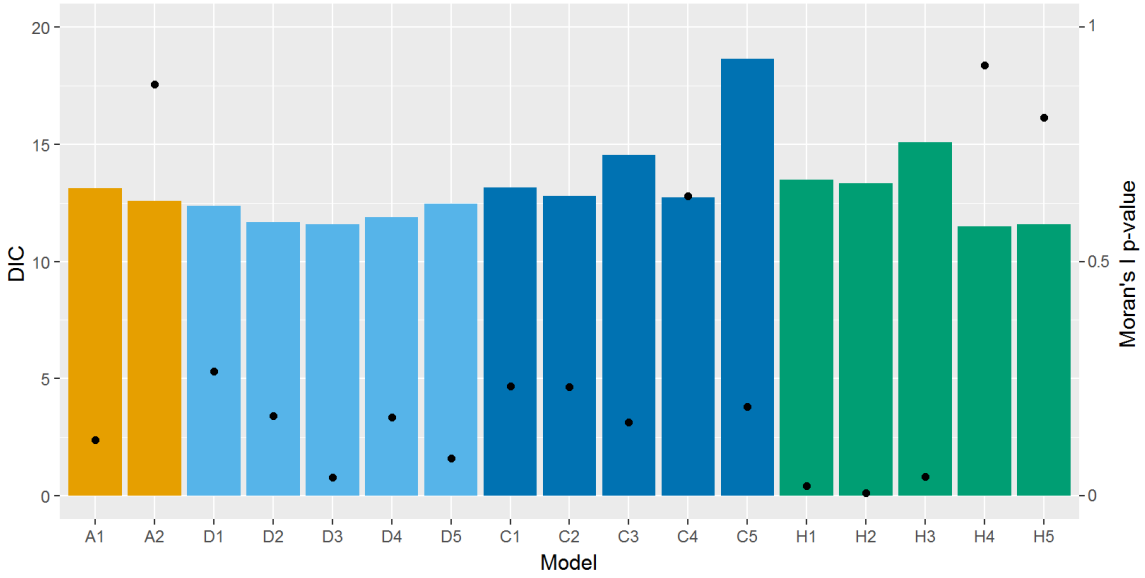


Figure 6.4: DIC for each model, except model B (left axis) overlayed by the two-sided p-values for Moran's I statistic on the posterior mean of the model residuals (right axis). The DIC for model B was 61.65.

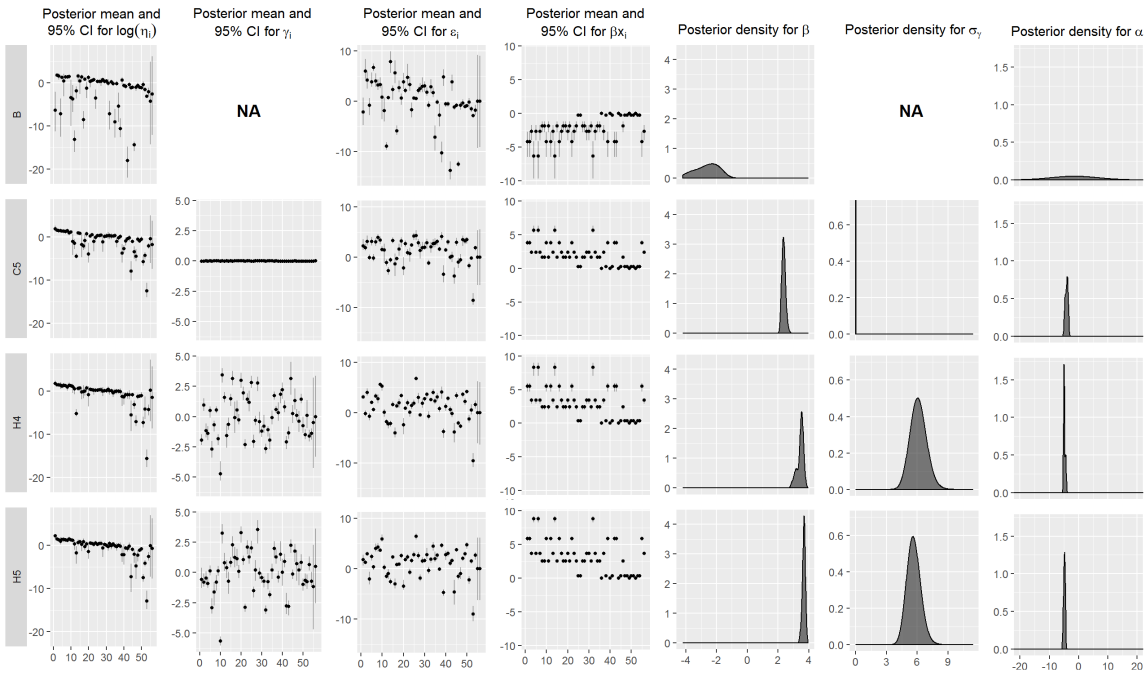


Figure 6.5: Summary of the parameter estimates based on the MCMC posterior sample for models B, C5, H4, and H5, for a single MCMC chain.

Spatial representations of the log-relative risk and model parameters for models B, C5, H4, and H5 are shown in Figure 6.6. This figure illustrates the estimates of Equation (6.2) and emphasises the contribution of each term towards the log-relative risk. The spatial random effects and covariate effect are shown on the same scale as the log-relative risk for comparison. The contribution of the structured spatial random effect is quite small compared to the covariate effect. The negative covariate effect for model B and the near-zero structured spatial random effect for model C5 are also emphasised in this figure. Note that model B tends to underestimate the relative risk

6.3.2 Analysis of the Synthetic Data Sets

The DIC and the p-values for Moran's I statistic on the model residuals ε_i are provided in Figure 6.7. Like the analysis of the Scottish lip cancer data, the DIC values are similar for most of the models, except for Model B, so the DIC does not offer much information on model adequacy except to assert that the models which incorporate spatial smoothing perform better than Model B. Despite the comparable model fit of the Model A and D variants, the two-sided p-values for Moran's I statistic indicate that the model residuals for some of these models, especially for the second synthetic data set, exhibit spatial autocorrelation. This may suggest that the Model C and H variants, in general, perform better.

The models with the smallest DIC for the analysis of the three synthetic data sets are D2, A2, and H5 respectively. The posterior estimates of the parameters for these three models are summarised in Figure 6.8. Since the expected values and covariate values are the same for these data sets, and since these three models each provide a good fit to the data, it is not surprising that the parameter estimates are also similar. However, there are slight differences. The posterior means of the log-relative risk, $\log(\eta_i)$, for Model A2 are more compact and closer to zero, and estimated with less certainty. This is most likely a result of the number of non-negligible neighbours used in this weighting scheme (refer to Figures 6.1 and 6.2). For the first synthetic data set, which exhibits no apparent spatial autocorrelation, the effect of the covariate is larger, which is offset by a smaller intercept. The minor posterior modes may indicate an identifiability issue between these two parameters, but not one of concern considering the good model fit.

Figure 6.6: Spatial representation of the relative risk, and a breakdown of the log-relative risk into the main components which comprise it: the structured and unstructured spatial random effects and the area-specific covariate effect, i.e. βx_i. The values are the posterior means.

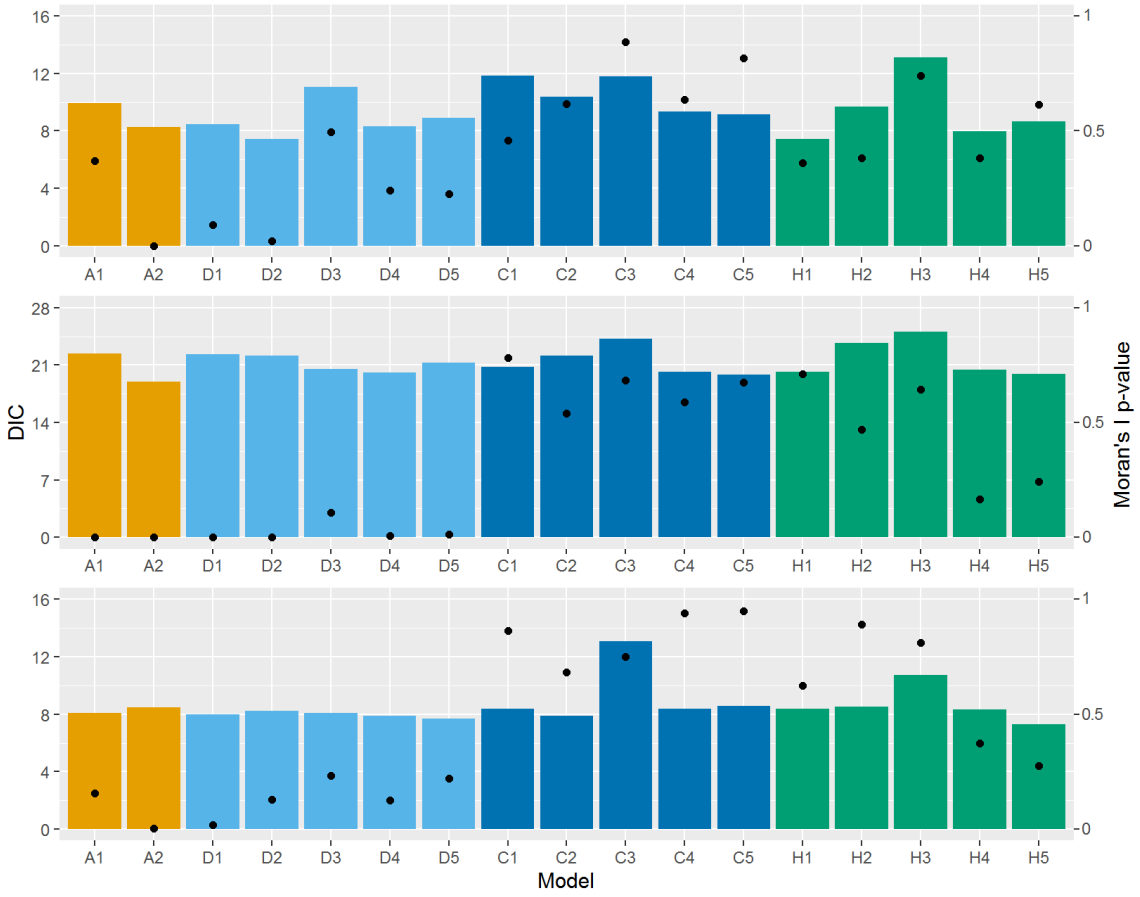


Figure 6.7: DIC for each model, except model B, across the three synthetic data sets (left axis) overlayed by the two-sided p-values for Moran's I statistic on the posterior mean of the model residuals (right axis). The DIC for model B was 37.00, 61.28, and 35.65 for the three data sets respectively.

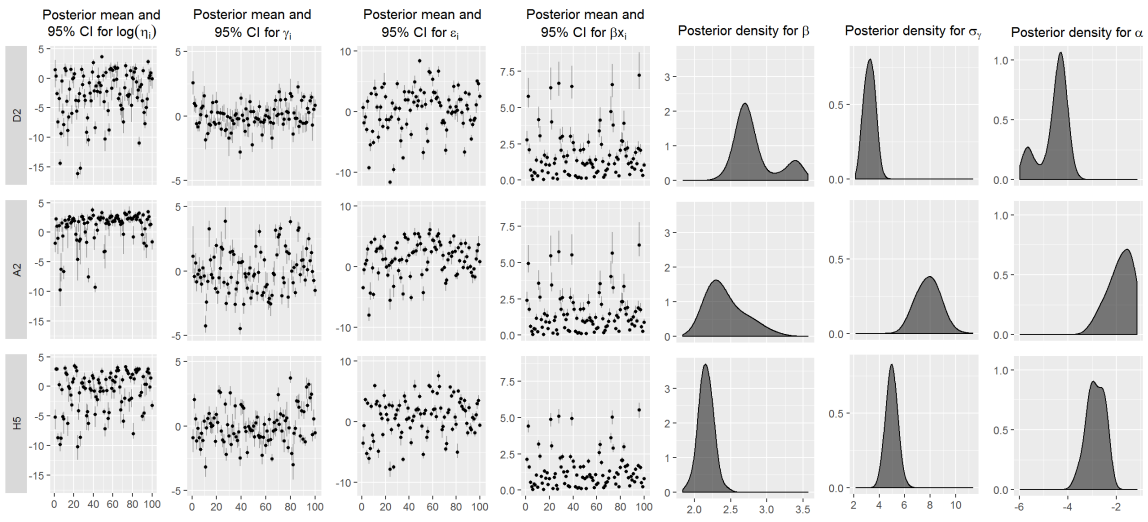


Figure 6.8: Summary of the parameter estimates based on the MCMC posterior sample for the best model in each of the synthetic data sets, for a single MCMC chain.

6.4 Discussion

It was noted by Earnest et al. (2007) that the prior distribution for the variance in the CAR prior distribution can have a very large impact on the model results. We reiterate and reinforce this point, and add that the prior for the variance of the unstructured spatial random effects can be highly influential as well. In fact, it was necessary to experiment with the priors for both variances before meaningful[1] results were obtained. Some of the prior specifications that we trialled for σ_γ^2 included

$$\sigma_\gamma^2 \sim \mathcal{N}(0, V)\mathbb{I}_{(0,\infty)}$$

for $V = 10$ and $V = 100$, and

$$\sigma_\gamma^2 \sim \text{Uniform}(0.0001, 100).$$

Similarly, the prior σ_ε^2 was initially a half-Normal distribution with variance 100, but the variance was decreased to 20 through trial and error. If these priors are too vague, the estimates may be very poor, leading to absurd predicted values and very poor model fit. Additionally, the sampler may have difficulty sampling from the full conditional distributions. At the other extreme, if the priors are too informative, or replaced by constants assumed known *a priori*, the parameter estimates can be very misleading and the impact of the weighting scheme becomes irrelevant.

There seems to be little evidence in support of the guidelines provided by Griffith (1996) regarding an appropriate number of neighbours in defining the weights matrix. Models which placed non-negligible weights on only a few neighbours, whether geographic neighbours or neighbours in the covariate space, did not necessarily perform any better according to measures such as DIC. In some instances, models C3, C5, and H3 stand out as attaining a somewhat worse fit to the data, but these models are not necessarily associated with having more non-negligible neighbours. It seems that each of the 17 weight specifications compared in these analyses are all viable. The task of choosing the 'best' weight specification without fitting several models and comparing their results is difficult, but also not very important. Figures 6.4, 6.6, and 6.7 in particular reveal that a model which incorporates at least some relevant spatial smoothing will

[1] In some of the early analyses, the predicted values differed from the observed values by a factor of 1000 or more.

result in a much better model fit than a model without spatial smoothing, regardless of how the weights are defined.

Figure 6.6 demonstrates the contribution of the estimated parameters on the log-relative risk for the Scottish lip cancer data. The structured spatial random effect might be smaller, that is, closer to zero than the effect of the covariate or residuals, yet the role of the structured spatial random effect in the model is clearly very important in order to attain good estimates of the log-relative risk. Notwithstanding this, as the results for model C5 show, while the inclusion of structured spatial random effects is necessary for good model fit, some weight specifications can have a noticeable impact.

Potential extensions to the work presented in this paper include consideration of multiple covariates. This could lead to new weight specifications and further insight into the impact that they have on spatial smoothing and statistical inference. Analysing larger data sets may also provide more opportunity to observe differences between these weight specifications.

Lastly, we wish to clarify the notion of smoothing and its effect on the relative risk. The term 'smoothed relative risk' is often used in the literature, but it should be apparent that the smoothing is applied to only one of the terms contributing to the log-relative risk, at least in the model used here. It would be quite erroneous to think that the estimated risk surface is smooth simply because the structured spatial random effects are smoothed. As Figure 6.6 indicates, the risk surface still permits adjacent areas with distinctly different values. The inclusion of spatial smoothing in the model clearly has a positive impact on estimation of the relative risk, but the extent of smoothing as it applies to the risk surface may be exaggerated. In other words, covariate effects and other fixed or random effects in the model may induce differences between areas in the estimated log-relative risk.

6.5 Conclusions

In summary, this paper compared 17 specifications of the weights matrix, including adjacency, distance-based, and covariate-based weights. Models using these weights were fit to both data based on simulated risk structures and real data for which the underlying spatial field is unobserved. In general, the effect of the weights matrix on model fit and estimation of the risk surface appears to be minimal. However, the analyses presented here indicate that it may

be beneficial to account for spatial smoothing, even if the observed data do not appear to be spatially autocorrelated.

Chapter 7

Overall Discussion and Conclusions

The overall aim of this thesis was to apply and develop Bayesian models for analysing spatio-temporal data exhibiting aberrant temporal trends, and to extend the related methodology. To achieve this aim, four specific research objectives were established:

1. To apply recently developed Bayesian models for analysing spatio-temporal data in order to understand the temporal patterns of mammography screening service utilisation in Brisbane.

2. To extend the models in objective 1 through the inclusion of covariates and a mixture model in order to address the limitations of existing models and improve inferences.

3. To extend the methodology of spatial modelling by systematically reviewing existing relabelling algorithms and proposing a new relabelling algorithm to deal with the label switching issue affecting Bayesian mixture models.

4. To extend the methodology of spatial modelling by analysing the impact of the spatial weights matrix on inference.

The first research objective was addressed in Chapter 3. The models presented in this chapter were able to identify 'hot-spots' and provide some useful insight into the spatial patterns in the excess relative risk. Aside from demonstrating the usefulness of each model and the synergy from applying them to the same data set, the analysis indicated some aspects of the methodology that could be improved to address the limitations of the models and improve statistical inferences. In particular, the BaySTDetect model estimated that most of the areas

had unusual temporal trends relative to an overall common trend, suggesting the existence of multiple common temporal trends rather than one. Both models indicated that distance to a mammography screening facility might be an important predictor. These ideas and others were presented as extensions in Chapter 4.

The second research objective was addressed in Chapter 4. This chapter provided a literature review of previous studies that considered various factors as potential predictors of relative risk. The covariates for the relative availability of services, socio-economic status, and shortest travel time to a screening facility were suggested multiple times in the literature and these were included one of the extensions to the BaySTDetect and space-time mixture models. The BaySTDetect model was also extended to include a mixture component through the linear predictor, which allowed the model to estimate multiple common trends. The methodology for classifying areas with unstable temporal trends identified by the space-time mixture model was also reviewed and extended. The resulting extended models were applied to the same data set as in Chapter 3. The results from this analysis demonstrated that the extensions provided improved inference.

The use of intrinsic conditional autoregressive prior distributions and the introduction of the K-component mixture model in Chapter 4 prompted new research questions regarding the issue of label switching and spatial smoothing, both of which present problems in spatial modelling. These questions gave rise to the third and fourth research objectives.

The third research objective was addressed in Chapter 5. A thorough literature review was conducted, which revealed a number of misconceptions regarding the occurrence of the label switching phenomenon and the practical implications. One facet of this research objective was to shed some light on these common misconceptions. Following this, an extensive, systematic review on the various relabelling algorithms to deal with label switching was provided. A new relabelling algorithm was also proposed. These algorithms were compared in a simulation study under three different scenarios, which revealed some interesting insights, highlighting the strengths and weaknesses of each algorithm in regards to accuracy, computational efficiency, and robustness to model misspecification. The newly proposed algorithm showed great promise in all three criteria.

The fourth research objective was addressed in Chapter 6. This chapter aimed to answer questions relating to the specification of the weights matrix which defines the nature and intensity

of smoothing. The primary goal was to determine what effect, if any, the specification of the weights has on parameter estimation and model inference. The analysis presented in this chapter provided some interesting and unexpected results regarding the effect of the weights matrix specification. The main findings were that differences between the effects of weight matrix specifications were marginal in most cases, and accounting for spatial autocorrelation is generally a good idea, even if autocorrelation among the observations is not obvious.

The application of the concepts considered in this thesis extend beyond the scope of each chapter. For example, the models presented in Chapters 3 and 4 will certainly be useful in other epidemiological problems and even other fields. Likewise, the relabelling algorithms of Chapter 5 may be of use to other latent variable models. Chapter 6 presented a wide range of weight matrix specifications, but these ideas could also be applied to other types of spatial smoothing models.

By meeting these research objectives, numerous gaps in the literature were addressed and the overall aim of this thesis was accomplished. The research presented in this thesis also identified new directions for future research.

In relation to the first two research objectives, additional covariates may be considered for inclusion in the models. Such an extension would likely be important if the data set was expanded to represent a wider geographic region, such as all of Australia rather than just Brisbane. Applying these models to new data sets may reveal new problems, such as the need to account for excess zero counts. Combining these models with zero-inflated models or similar constructs pose new extensions.

In relation to label switching, the systematic review could be extended to include additional relabelling algorithms, especially new algorithms as they are developed, using the results from this thesis as a benchmark. Likewise, the simulation study could be extended to include new measures of performance to help evaluate their differences. Another straightforward extension is to assess each of the relabelling algorithms on higher dimensional mixture models.

Regarding the investigation on the impact of the spatial weights matrix, the analysis presented in this thesis could be expanded to include other real-world data sets to validate the findings. Alternative specifications of the spatial weights matrix could also be considered, and these could be combined with the covariate and hybrid approaches to smoothing to produce new specifications.

Appendix A

Supplementary Material from Chapter 3

A.1 WinBUGS Code for the BaySTDetect Model (Supplementary Code S1)

```
model {
  for (i in 1:N) {
    for (t in 1:T) {
      y[i,t] ~ dpois(E.mu[i,t])
      E.mu[i,t] <- E[i,t] * mu[i,t]
      log(mu[i,t]) <- p[i] * crw.mix[i,t] + (1-p[i]) * rw.mix[i,t]
    }
    p[i] ~ dbern(0.95)
  }

# Common trend model
  for (i in 1:N) {
    for (t in 1:T) {
      y1[i,t] ~ dpois(mu1[i,t])
      log(mu1[i,t]) <- log(E[i,t]) + temp1[i,t]
      temp1[i,t] <- alpha + eta[i] + gamma[t] + beta * x[i,t]
      crw.mix[i,t] <- cut(temp1[i,t])
    }
  }
  alpha ~ dnorm(0,0.001)
```

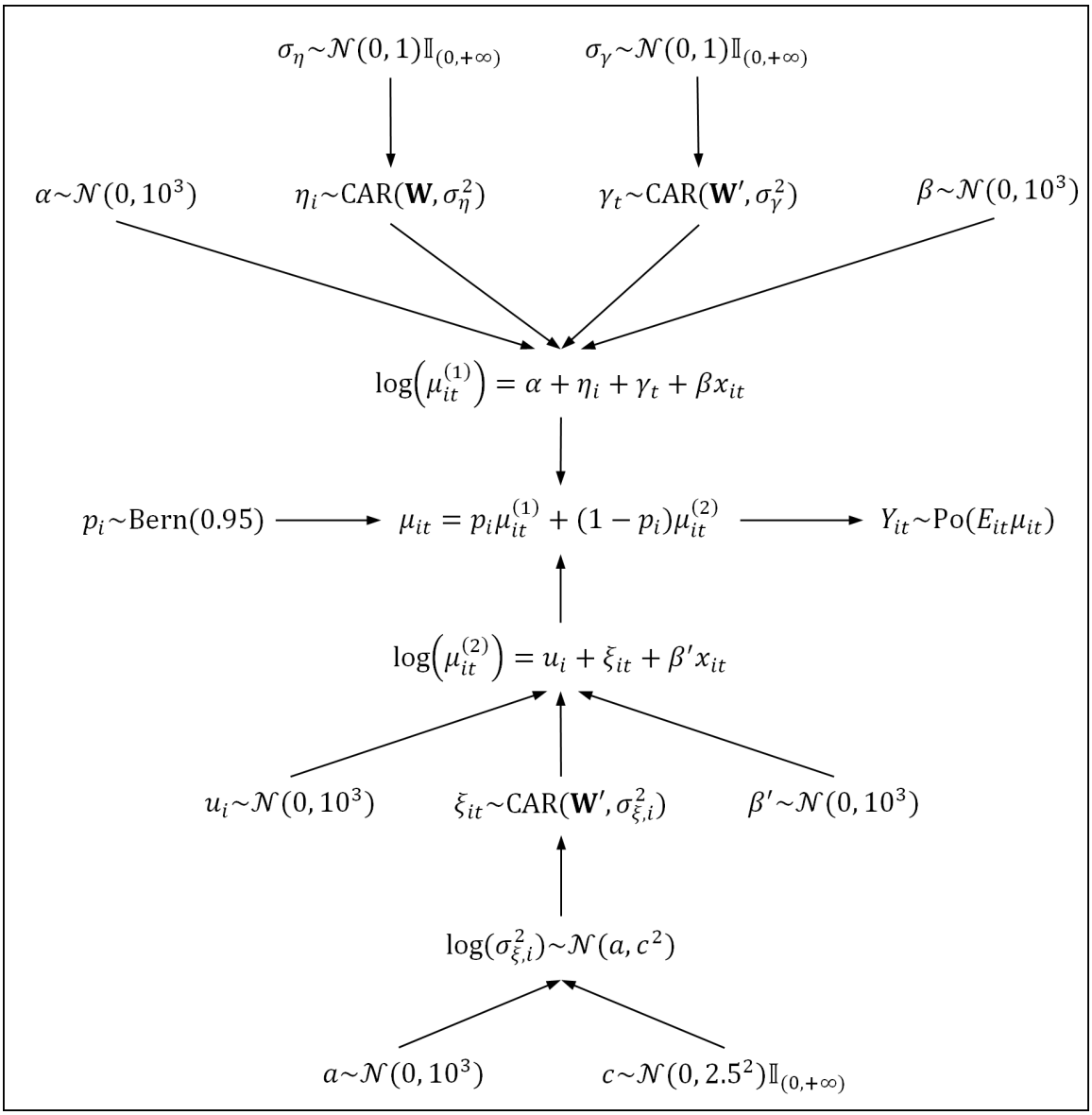
```
  gamma[1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.gamma)
  beta ~ dnorm(0, 0.001)
  eta[1:(N-1)] ~ car.normal(adj[],weights[],num[],tau.eta)
  eta[N] <- 0 # SLA with zero neighbours
  tau.gamma <- pow(sigma.gamma,-2)
  sigma.gamma ~ dnorm(0,1)I(0,)
  tau.eta <- pow(sigma.eta,-2)
  sigma.eta ~ dnorm(0,1)I(0,)

# Area-specific Model
  for (i in 1:N) {
    for (t in 1:T) {
      y2[i,t] ~ dpois(mu2[i,t])
      log(mu2[i,t]) <- log(E[i,t]) + temp2[i,t]
      temp2[i,t] <- u[i] + xi[i,t] + beta.dash * x[i,t]
      rw.mix[i,t] <- cut(temp2[i,t])
    }
    u[i] ~ dnorm(0,0.001)
    xi[i,1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.xi[i])
    tau.xi[i] <- pow(var.xi[i],-1)
    var.xi[i] <- exp(log.var.xi[i])
    log.var.xi[i] ~ dnorm(a, cc)
  }
  beta.dash ~ dnorm(0, 0.001)
  a ~ dnorm(0,0.001)
  cc <- pow(c,-2)
  c ~ dnorm(0,d)I(0,)
  d <- pow(2.5,-2)
}
```

A.2 WinBUGS Code for the Space-time Mixture Model (Supplementary Code S2)

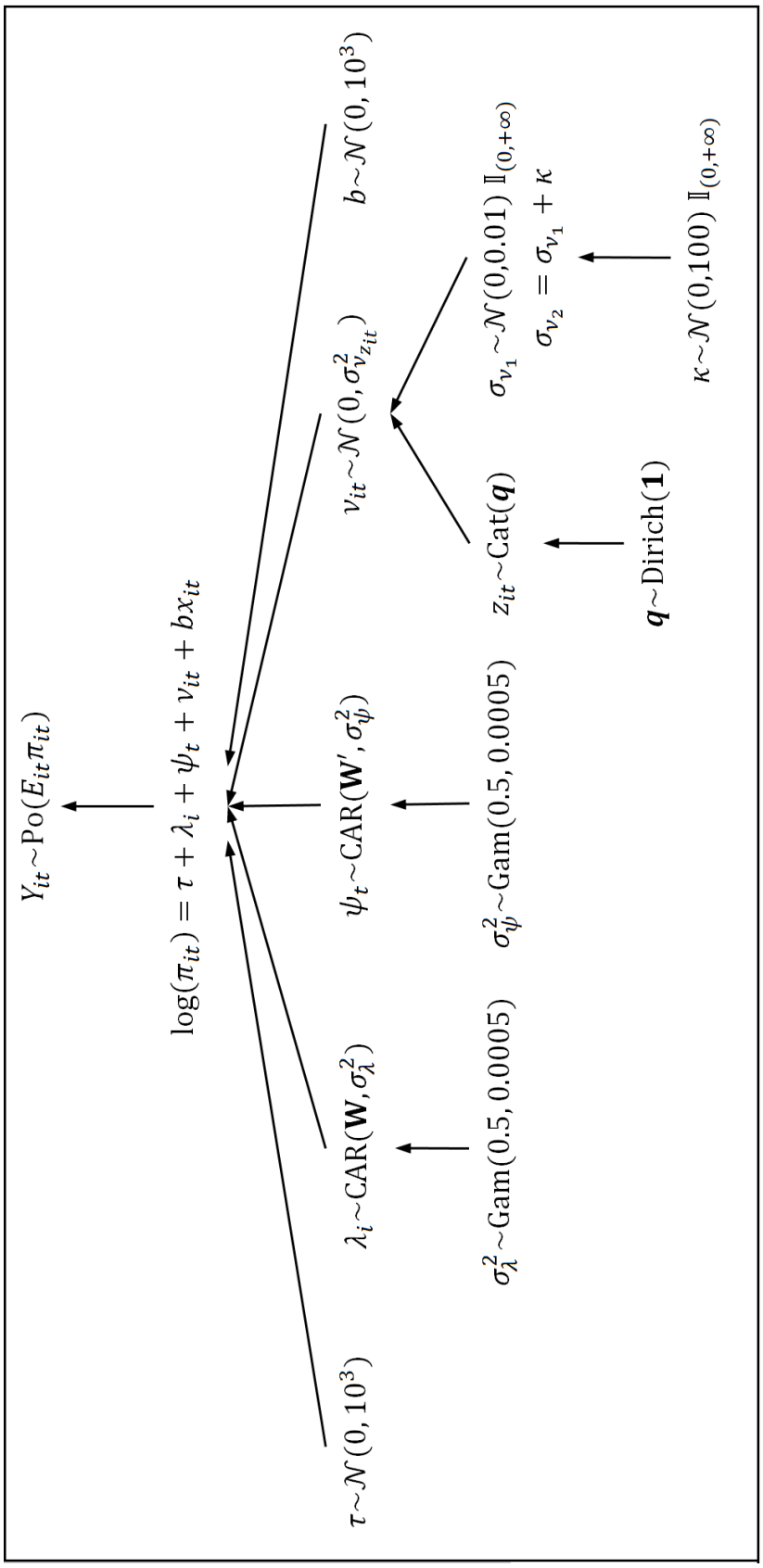
```
model {
  for (i in 1:N) {
    for (t in 1:T) {
      y[i,t] ~ dpois(E.pi[i,t])
      E.pi[i,t] <- E[i,t] * pi[i,t]
      log(pi[i,t]) <- tau + lambda[i] + psi[t] + nu[i,t] + b * x[i,t]
    }
  }

  # Priors
  tau ~ dnorm(0, 0.001)
  lambda[1:(N-1)] ~ car.normal(adj[],weights[],num[],tau.lambda)
  lambda[N] <- 0 # SLA with zero neighbours
  psi[1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.psi)
  for(i in 1:N){
    for(t in 1:T){
      nu[i,t] ~ dnorm(0, tau.nu[z[i,t]])
    }
  }
  b ~ dnorm(0, 0.001)

  # Hyperpriors
  for(i in 1:N){
    for(t in 1:T){
      z[i,t] ~ dcat(q[])
    }
  }
  sigma.lambda.sq <- pow(tau.lambda, -1)
  sigma.psi.sq <- pow(tau.psi, -1)
  tau.lambda ~ dgamma(0.5, 0.0005)
  tau.psi ~ dgamma(0.5, 0.0005)
```

```
  sigma.nu[1] ~ dnorm(0, 100)I(0.0,)
  kappa ~ dnorm(0, 0.01)I(0.0,)
  sigma.nu[2] <- sigma.nu[1] + kappa
  q[1:2] ~ ddirch(alpha.q[])
  for(i in 1:2){
    tau.nu[i] <- pow(sigma.nu[i], -2)
    alpha.q[i] <- 1
  }
}
```

A.3 Schematic of the BaySTDetect Model (Supplementary figure S3)



A.4 Schematic of the Space-time Mixture Model (Supplementary figure S4)



Appendix B

Supplementary Material from Chapter 4

B.1 WinBUGS Code for the Extended BaySTDetect Model

```
model {
  for (i in 1:N) {
    for (t in 1:T) {
      y[i,t] ~ dpois(E.mu[i,t])
      E.mu[i,t] <- E[i,t] * mu[i,t]
      log(mu[i,t]) <- p[i] * crw.mix[i,t] + (1-p[i]) * rw.mix[i,t]
    }
    p[i] ~ dbern(delta[i])
    delta[i] ~ dbeta(alpha.delta[i], 2)
    log(alpha.delta[i]) <- log.alpha.delta[i]
    log.alpha.delta[i] ~ dnorm(mean.alpha.delta[i], 0.01)
    mean.alpha.delta[i] <- beta.2 * x2[i]
  }
  beta.2 ~ dnorm(0, 0.01)

# Common trend model
  for (i in 1:N) {
    for (t in 1:T) {
      y1[i,t] ~ dpois(mu1[i,t])
      log(mu1[i,t]) <- log(E[i,t]) + temp[i,t]
```

```
    temp[i,t] <- alpha[z[i]] + eta[z[i],i] + gamma[z[i],t] +
       beta.1[z[i]] * x1[i,t] + beta.3[z[i]] * x3[i,t]
    crw.mix[i,t] <- cut(temp[i,t])
   }
   z[i] ~ dcat(rho[])
  }
  for (k in 1:K) {
   alpha[k] ~ dnorm(0,0.001)
   eta[k,1:(N-1)] ~ car.normal(adj[],weights[],num[],tau.eta[k])
   eta[k,N] <- 0 # Moreton Island; zero neighbours
   gamma[k,1:T] ~
      car.normal(adj.t[],weights.t[],num.t[],tau.gamma[k])
   beta.1[k] ~ dnorm(0, 0.001)
   beta.3[k] ~ dnorm(0, 0.001)
   tau.gamma[k] <- pow(sigma.gamma[k],-2)
   sigma.gamma[k] ~ dnorm(0,1)I(0,)
   tau.eta[k] <- pow(sigma.eta[k],-2)
   sigma.eta[k] ~ dnorm(0,1)I(0,)

   rho[k] <- rho.star[k] / sum(rho.star[])
   rho.star[k] ~ dgamma(alpha.rho[k], 1)
   # Equivalent to rho[1:K] ~ ddirch(alpha.rho[])
   alpha.rho[k] ~ dgamma(0.5, 0.005)
  }

# Area-specific Model
  for (i in 1:N) {
   for (t in 1:T) {
    y2[i,t] ~ dpois(mu2[i,t])
    log(mu2[i,t]) <- log(E[i,t]) + temp1[i,t]
    temp1[i,t] <- u[i] + xi[i,t] + beta.1.dash * x1[i,t] +
       beta.3.dash * x3[i,t]
    rw.mix[i,t] <- cut(temp1[i,t])
   }
```

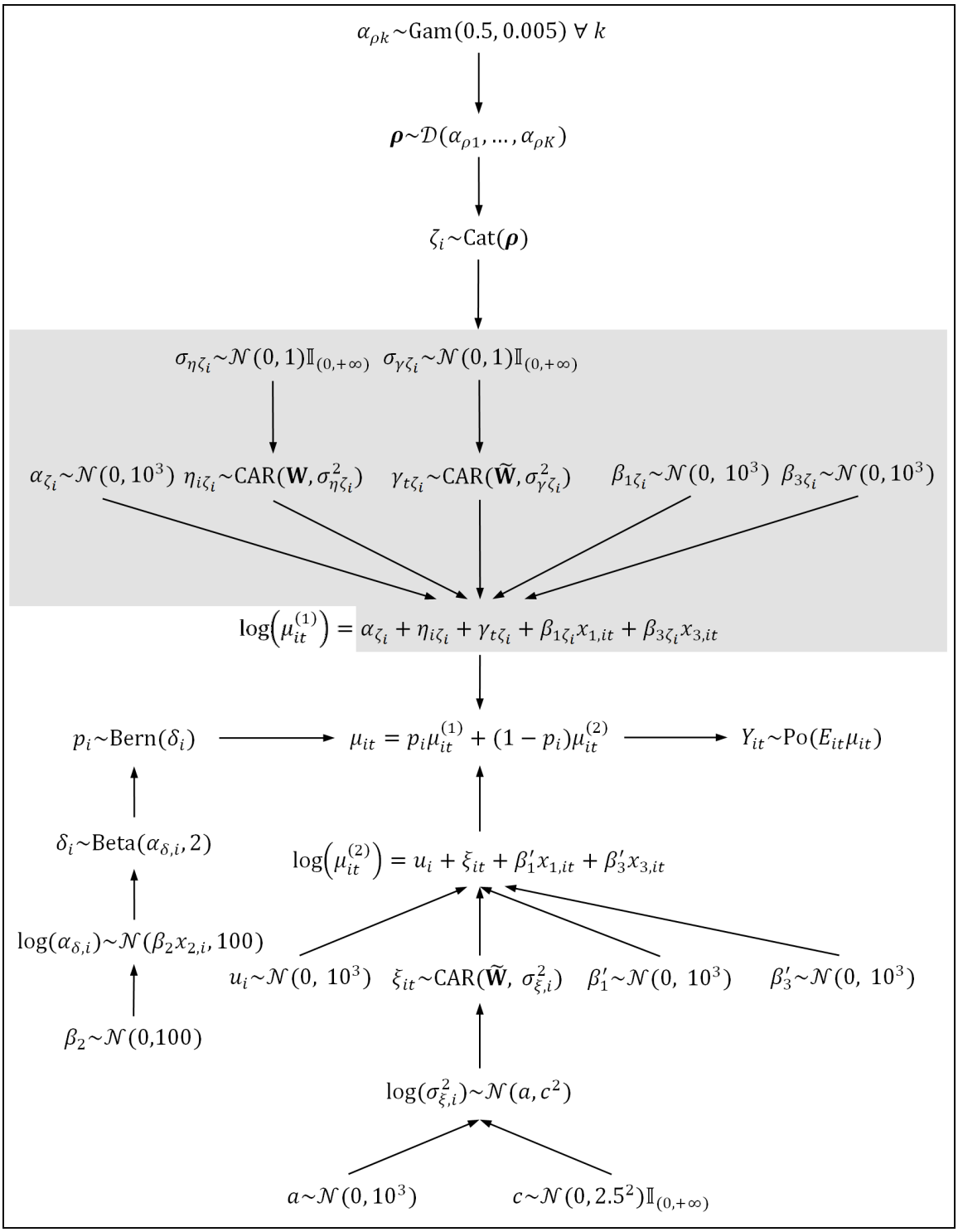
```
    u[i] ~ dnorm(0,0.001)
    xi[i,1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.xi[i])
    tau.xi[i] <- pow(var.xi[i],-1)
    var.xi[i] <- exp(log.var.xi[i])
    log.var.xi[i] ~ dnorm(a,cc)
  }
  beta.1.dash ~ dnorm(0, 0.001)
  beta.3.dash ~ dnorm(0, 0.001)
  a ~ dnorm(0,0.001)
  cc <- pow(c,-2)
  c ~ dnorm(0,d)I(0,)
  d <- pow(2.5,-2)
}
```

B.2 WinBUGS Code for the Extended Space-time Mixture Model

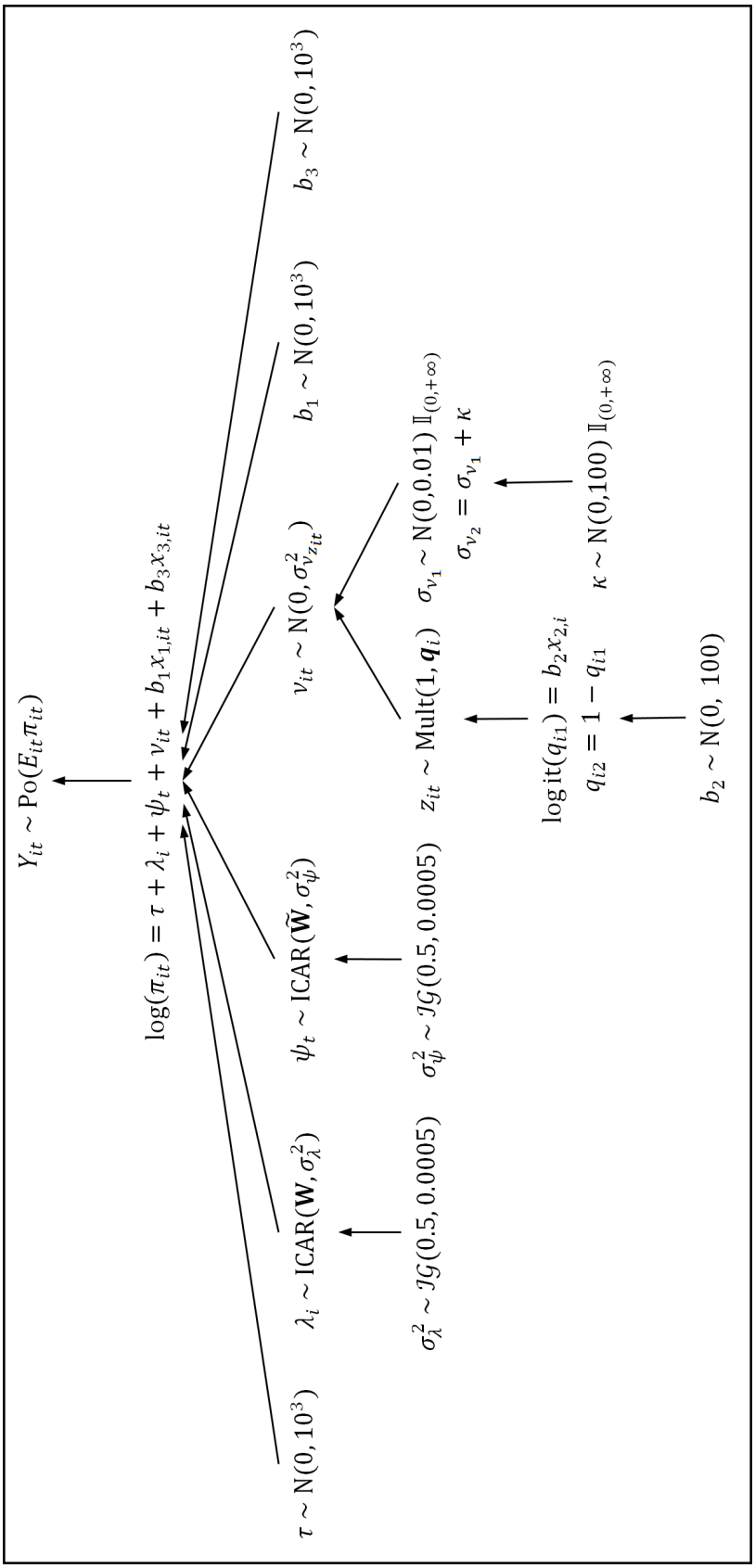
```
model {
  for (i in 1:N) {
    for (t in 1:T) {
      y[i,t] ~ dpois(E.pi[i,t])
      E.pi[i,t] <- E[i,t] * pi[i,t]
      log(pi[i,t]) <- tau + lambda[i] + psi[t] + nu[i,t] + b1 *
          x1[i,t] + b3 * x3[i,t]
    }
  }

  # Priors
  tau ~ dnorm(0, 0.001)
  lambda[1:(N-1)] ~ car.normal(adj[],weights[],num[],tau.lambda)
  lambda[N] <- 0 # Moreton Island; zero neighbours
  psi[1:T] ~ car.normal(adj.t[],weights.t[],num.t[],tau.psi)
  for(i in 1:N){
    for(t in 1:T){
```

```
    nu[i,t] ~ dnorm(0, tau.nu[z[i,t]])
   }
  }
  b1 ~ dnorm(0, 0.001)
  b3 ~ dnorm(0, 0.001)

  # Hyperpriors
  for(i in 1:N){
   for(t in 1:T){
    z[i,t] ~ dcat(q[i,])
   }
  }
  sigma.lambda.sq <- pow(tau.lambda, -1)
  sigma.psi.sq <- pow(tau.psi, -1)
  tau.lambda ~ dgamma(0.5, 0.0005)
  tau.psi ~ dgamma(0.5, 0.0005)

  sigma.nu[1] ~ dnorm(0, 100)I(0.0,)
  kappa ~ dnorm(0, 0.01)I(0.0,)
  sigma.nu[2] <- sigma.nu[1] + kappa
  for(k in 1:2){
   tau.nu[k] <- pow(sigma.nu[k], -2)
  }
  for(i in 1:N){
   logit(q[i,1]) <- b2 * x2[i]
   q[i,2] <- 1 - q[i,1]
  }
  b2 ~ dnorm(0, 0.01)
}
```

B.3 Schematic of the Extended BaySTDetect Model



B.4 Schematic of the Extended Space-time Mixture Model



Appendix C

Supplementary Material from Chapter 5

C.1 Notes on Algorithm 4

The notation used in the following quotations has been changed to be consistent with the notation presented in Chapter 5.

On the subject of the BM algorithm, Puolamäki and Kaski (2009) write:

"Another approach is to take explicitly into account the fact that there is a unique permutation of labels for each sample. The mixing matrix $\mathbf{Q}$ can then be found by optimizing the cost function given by

$$\prod_{m=1}^{M} \sum_{\tau \in S} \frac{1}{K!} \prod_{i=1}^{N} q_{i,\tau^{-1}} \left(z_i^{(m)} \right)$$

by using [an] EM algorithm, presented by Algorithm 4. We call this model [the] Bernoulli Mixture Permutation [BMP] model" [Notation mine].

Pauli and Torelli (2015) reiterate the BM algorithm and this alternative approach:

"Equivalently, the matrix $\boldsymbol{Q}$ can be found [by] minimizing the cost function

$$\prod_{m=1}^{M} \sum_{\tau \in S} \frac{1}{K!} q_{i,\tau^{-1}} \left(z_i^{(m)} \right).$$

" [Notation mine]

Note these cost functions are different. Neither Puolamäki and Kaski (2009) nor Pauli and Torelli (2015) make it clear how this cost function might be optimised in an EM algorithm. It

seems that the cost function might serve as the basis for an iterative deterministic algorithm. However, neither specification of the cost function appears to be correct. We believe that the cost function can be more clearly and accurately stated as

$$c_{j,k}^{(m)} = -\sum_{i=1}^{N} q_{i,k} \mathbb{I}\left(z_i^{(m)} = j\right)$$

Our version of the BMP algorithm is an iterative deterministic algorithm which solves the LSAP using these costs. This may differ to the algorithm that Puolamäki and Kaski (2009) and Pauli and Torelli (2015) had in mind, but this version of the BMP does produce reasonable results.

C.2 Summary of Previous Simulation Studies

Simulation	Algorithms	Density	K	N	M
Papastamoulis (2013)	IC, KL, ECR, PRA, SJW	Gaussian	3	80	50000
			10	100	10000
	IC, ECR, PRA	HMM	3	500	80000
Rodriguez and Walker (2014)	DB, ECR, KL	Gaussian	2	1000	30000
			4	200	30000
			5	600	30000
			4	245	100000
		Multivariate Gaussian	4	200	30000
van Havre et al. (2015)	ZS	Gaussian	3	100, 200	20000
			3	100, 200	20000
			2	100, 200	20000
			3	100, 200	20000
Yao (2013)	IC, KL, two others	Gaussian	2	400	20000
			8	400	5000
			3	155	20000
Sperrin et al. (2010)	IC, KL, SJW, two variants of SJW, two others	Gaussian?	3, 4, 5	82	100
		Gaussian	2	50, 100	100

Continued overleaf.

Simulation	Algorithms	Density	K	N	M
Stephens (2000b)	KL	Gaussian	6	82	20000
		t_4	3	82	10000
Papastamoulis (2016)	KL, PRA, ECR, ECR 1, ECR 2, SJW, IC, DB	Gaussian	5	256	10000
		Poisson	4	240	10000
		Multivariate Gaussian	4	100	10000
	KL, ECR, ECR 1, ECR 2, IC, DB		9	280	15000
Pan et al. (2015)	KL, ECR, SJW, two others	Poisson	2	10	100000
			2	100	100000
			2	100	100000
			2	100	100000
	ECR 2, DB, one other	Gaussian	3	160	100000
	DB, 1 other	Multivariate Gaussian	4	200	100000
Papastamoulis and Iliopoulos (2010)	ECR, PRA, KL	Gaussian	3, 5	160, 600	20000
	ECR, PRA	Gaussian	6	82	60000
	KL, ECR	Bivariate Gaussian	4	200	10000
			3	272	30000
Puolamäki and Kaski (2009)	IC, KL (STE), BM, BMP, 1 other	Gaussian	3	5	1000

C.3 R Code for Implementing the Relabelling Algorithms

Summary of relabelling algorithms used in our simulation study

Algorithm	In `label.switching`?	R Function Name	Input	Optional Parameter Defaults
KL	Yes	`stephens`	`p`	`threshold = 1e-6, maxiter = 100`
PRA	Yes, but not used	`PRA`	`theta, m.ref`	
BM	No	`BM`	`z`	`tol = 0.05, scale = 2`
BMP	No	`BMP`	`z`	`tol = 1e-6, maxiter = 50`
ECR	Yes	`ecr`	`z.ref, z, K`	
ECR 1	Yes	`ecr.iterative.1`	`z, K`	`threshold = 1e-6, maxiter = 100`
ECR 2	Yes	`ecr.iterative.2`	`z, K, p`	`threshold = 1e-6, maxiter = 100`
SJW	Yes	`sjw`	`theta, z, LLf, y`	`init = 0, threshold = 1e-6, maxiter = 100`
DB	Yes, but not used	`DB`	`y, z, iterative = FALSE`	
DB it.	No	`DB`	`y, z, iterative = TRUE`	`tol = 0.01, maxiter = 50`
PU	No	`pivotal.unit`	`z, m.ref`	`pivot.method = 3`

Continued overleaf.

Algorithm	In `label.switching`?	R Function Name	Input	Optional Parameter Defaults
ZS	No	`Zswitch`	`z, m.ref, theta`	`omega = 0.01`
ZS 2	No	`Zswitch.2`	`z, m.ref, theta`	`omega = 0.01, inf.cost = 1e10`

For each algorithm not in `label.switching`, M, N, and K are inherited from R's global environment. For the optional parameters, only the defaults were used. Note that `pivot.method = 3` corresponds to the criterion shown in Algorithm 11.

Description of input parameters used in relabelling algorithms

Object	Type	Dimension	Details
`z`	Array (integer)	$M \times N$	Simulated allocation vectors
`y`	Numeric (real)	N	Observed data
`z.ref`	Numeric (integer)	N	Pivot allocation vector
`p`	Array (real)	$M \times N \times K$	Classification probabilities
`theta`	Array (real)	$M \times K \times R$	Simulated parameters
`m.ref`	Numeric (integer)	1	Reference iteration*
`LLf`	Function	–	Complete log-likelihood function

* This is usually taken to be the iteration which corresponds to the MAP estimate of $\boldsymbol{\theta}$.

Algorithm 2: pivotal reordering algorithm (PRA)

```
PRA <- function(theta, MAP){
  R <- dim(theta)[3]
  theta.ref <- t(theta[MAP,,]) # pivot
  perms <- matrix(NA, M, K)

  for(m in 1:M){
    # Construct cost matrix
    c <- matrix(0, K, K)
    for(j in 1:K){
      for (k in 1:K){
        for(r in 1:R){
          c[j, k] <- c[j, k] + theta[m, j, r] * theta.ref[r, k]
        }
      }
    }
    # Solve LSAP
    Solution <- lp.assign(c, direction = "max")$solution
    perms[m,] <- apply(Solution, 2, which.max)
  }
  return(list(permutations = perms))
}
```

Algorithm 3: Bernoulli mixture (BM) algorithm

```
BM <- function(z, tol = 0.05, scale = 2){
  z.hat <- array(NA, c(M, N, K))
  for(k in 1:K){
    z.hat[,,k] <- z
    z.hat[,,k][which(z == k, arr.ind = TRUE)] <- 1
    z.hat[,,k][which(z != k, arr.ind = TRUE)] <- 0
  }
  max.diff <- Inf
  Iterations <- 1
```

```
# STEP 1: Initialise the Bernoulli parameters (matrix Q)
q <- matrix(runif(N*K), N, K)

while(max.diff > tol){
  # STEP 2 (E-step):
  gamma <- array(1, c(M, K, K))
  for(m in 1:M){
    for(j in 1:K){
      scale.adj <- scale
      denominator <- 0
      max.gamma <- Inf
      while(max.gamma == Inf | denominator == 0){
        for(k in 1:K){
          gamma[m, j, k] <- prod(q[,k] ^ z.hat[m,,j] *
            (1 - q[,k]) ^ (1 - z.hat[m,,j]) * scale.adj)
        }
        max.gamma <- max(gamma[m, j, ])
        denominator <- sum(gamma[m, j, ])
        if(max.gamma == Inf){
          scale.adj <- scale.adj - 0.2
        }else if(denominator < 1e-290){
          scale.adj <- scale.adj + 0.2
        }
      }
      gamma[m, j, ] <- gamma[m, j, ] / denominator
    }
  }

  # STEP 3 (M-step):
  q.new <- matrix(0, N, K)
  for(i in 1:N){
    for(k in 1:K){
      q.new[i, k] <- sum(gamma[,,k] * z.hat[,i,]) /
        sum(gamma[,,k])
```

```
      }
    }

    # STEP 4: Assess convergence
    max.diff <- max(abs(q - q.new))
    message(paste0("Iteration ", Iterations, ": max change in q =
       ", max.diff, "."))
    if(max.diff <= tol){
      message(paste("BM algorithm converged in", Iterations,
         "iterations."))
      status <- paste0("Converged (", Iterations, " iterations)")
    }else{
      q <- q.new
      Iterations <- Iterations + 1
    }
    flush.console()
  }

  # STEP 5: Determine the most likely permutations
  perms <- matrix(NA, M, K)
  for(m in 1:M){
    perms[m, ] <- apply(gamma[m,,], 2, which.max)
  }
  return(list(permutations = perms, q = q, gamma = gamma, status =
     status))
}
```

Algorithm 4: Bernoulli mixture permutation (BMP) algorithm

```
BMP <- function(z, tol = 1e-6, maxiter = 50){
  L0 <- rep(-Inf, M)
  L0.new <- rep(NA, M)
  Repeat <- TRUE
  Iterations <- 1L
```

```
# STEP 1: Initialise permutations
perms <- matrix(1:K, M, K, byrow = TRUE)

while(Repeat == TRUE){
  # STEP 2: Calculate q
  q <- matrix(NA, N, K)
  for(k in 1:K){
    q[,k] <- apply(z, 2, function(x) length(which(x ==
       perms[,k]))) / M
  }

  # STEP 3: Determine permutations by solving LSAP
  for(m in 1:M){
    # Construct cost matrix
    c <- matrix(NA, K, K)
    for(j in 1:K){
      for(k in 1:K){
        c[j, k] <- sum(q[which(z[m,] == j), k])
      }
    }
    # Solve LSAP
    LSAP <- lp.assign(c, direction = "max")
    perms[m,] <- apply(LSAP$solution, 2, which.max)
    L0.new[m] <- LSAP$objval
  }

  # STEP 4: Assess convergence
  Improvement <- (1 - sum(L0) / sum(L0.new)) # Improvement in
     overall gain
  message(paste0("Iteration ", Iterations, ": overall gain = ",
     sum(L0.new),
    " (improvement = ", Improvement, ").","))
  if(Iterations >= maxiter){
    Repeat <- FALSE
```

```
      message("BMP algorithm stopped - maximum iterations reached.")
      status <- paste0("Max iterations reached (", Iterations, "
         iterations)")
    }else if(abs(Improvement) > tol){
      L0 <- L0.new
      Iterations <- Iterations + 1
    }else{
      Repeat <- FALSE
      message(paste("BMP algorithm converged in", Iterations,
         "iterations."))
      status <- paste0("Converged (", Iterations, " iterations)")
    }
    flush.console()
  }
  return(list(permutations = perms, status = status))
}
```

Algorithms 9 and 10: data-based algorithm (DB and DB iterative)

```
DB <- function (y, z, iterative = FALSE, tol = 0.01, maxiter = 50){
  # STEP 1: Initialisation
  perms <- matrix(1:K, M, K, byrow = TRUE)
  n <- matrix(NA, M, K)
  y.bar <- matrix(NA, M, K)
  s <- matrix(NA, M, K)
  for(k in 1:K){
    n[,k] <- apply(z, 1, function(x) length(which(x == k)))
    for(m in 1:M){
      y.bar[m,k] <- sum(y[which(z[m,] == k)]) / n[m,k]
      s[m,k] <- sqrt(sum(((y - y.bar[m,k])^2)[which(z[m,] == k)]) /
         (n[m,k] - 1))
    }
  }
  m.hat <- NULL
  s.hat <- NULL
```

```
L0 <- rep(Inf, M)
L0.new <- rep(NA, M)
Repeat <- 2L
Iterations <- 1L

while(Repeat > 0){
  # STEP 2: Initialise/update cluster means and standard
     deviations: m.hat, and s.hat
  if(is.null(m.hat)){
    # Initialise
    for(k in 1:K){
      m.hat[k] <- min(y) + (max(y) - min(y)) * k/(K + 1)
      s.hat[k] <- (max(y) - min(y))/K
    }
  }else{
    # Update
    for(k in 1:K){
      ind <- cbind(seq_along(perms[,k]), perms[,k]) #Indices to
         permute y.bar,s,n
      m.hat[k] <- sum(y.bar[ind][n[ind] > 0]) / length(n[ind] > 0)
      s.hat[k] <- sum(s[ind][n[ind] > 1]) / length(n[ind] > 1)
    }
  }

  # STEP 3: Determine permutations by solving LSAP
  for(m in 1:M){
    c <- matrix(0, K, K) # Cost matrix
    for(j in 1:K){
      ind <- which(z[m, ] == j)
      for(k in 1:K){
        c[j, k] <- length(ind) * sum((((y - m.hat[k]) /
           s.hat[k])^2)[ind])
      }
    }
```

```
      LSAP <- lp.assign(c)
      perms[m,] <- apply(LSAP$solution, 2, which.max)
      L0.new[m] <- LSAP$objval
    }

    # STEP 4: Decide when to stop algorithm
    if(iterative == FALSE){
      Repeat <- Repeat - 1
    }else{
      Improvement <- -(1 - sum(L0) / sum(L0.new)) # Improvement in
         overall loss
      message(paste0("Iteration ", Iterations, ": overall loss = ",
         sum(L0.new),
        " (improvement = ", Improvement, ").") )
      flush.console()
      if(Iterations >= maxiter){
        Repeat <- 0
        message("DB iterative algorithm stopped - maximum
           iterations reached.")
        status <- paste0("Max iterations reached (", Iterations, "
           iterations)")
      }else if(abs(Improvement) > tol){
        L0 <- L0.new
      }else{
        Repeat <- 0
        message(paste("DB iterative algorithm converged in",
          Iterations, "iterations.") )
        status <- paste0("Converged (", Iterations, " iterations)")
      }
      Iterations <- Iterations + 1
    }
  }
  if(iterative == FALSE){
    return(list(permutations = perms))
```

```
  }else{
    return(list(permutations = perms, status = status))
  }
}
```

Algorithm 11: pivotal unit (PU) algorithm

```
PU <- function(z, m.ref, pivot.method = 3){
  if(!(pivot.method %in% 1:6)){
    stop("Parameter 'pivot.method' must be an integer between 1 and
       6 inclusive.")
  }

  # STEP 1:
  groups <- z[m.ref,]
  N.k <- vector("list", K)
  for(k in 1:K){
    N.k[[k]] <- which(groups == k)
  }

  # STEP 2: Compute the similarity matrix
  s <- matrix(NA, N, N)
  for(i in 1:N){
    for(j in 1:N){
      s[i,j] <- length(which((z[, i] == z[, j]) == TRUE)) / M
    }
  }

  # STEP 3: Determine the pivotal units
  pivots <- rep(NA, K)
  for(k in 1:K){
    if(length(N.k[[k]]) == 1){
      pivots[k] <- N.k[[k]]
    }else{
      switch(pivot.method,
```

```
        pivots[k] <- N.k[[k]][which.max(apply(s[N.k[[k]],
           N.k[[k]]], 1, max))],
        pivots[k] <- N.k[[k]][which.max(apply(s[N.k[[k]],
           N.k[[k]]], 1, sum))],
        pivots[k] <- N.k[[k]][which.max(apply(s[N.k[[k]],
           N.k[[k]]], 1, sum) -
          apply(s[N.k[[k]], -N.k[[k]]], 1, sum))],
        pivots[k] <- N.k[[k]][which.min(apply(s[N.k[[k]],
           N.k[[k]]], 1, min))],
        pivots[k] <- N.k[[k]][which.min(apply(s[N.k[[k]],
           -N.k[[k]]], 1, max))],
        pivots[k] <- N.k[[k]][which.min(apply(s[N.k[[k]],
           -N.k[[k]]], 1, sum))],
     )
   }
}

# STEP 4: Subset the allocation vectors by the pivots
Z.star <- z[, pivots]

# STEP 5: Remove iterations with fewer/more than K unique elements
remove <- rep(NA, M)
for(m in 1:M){
  remove[m] <- length(unique(Z.star[m,])) != K
}
if(all(remove == FALSE)){
  remove <- NULL
}else{
  remove <- which(remove)
}

# STEP 6: Determine the permutations
perms <- Z.star
for(m in 1:M){
```

```
    if(m %in% remove){
      perms[m, ] <- rep(NA, K)
    }
  }

  # Return permutations
  return(list(permutations = perms, removed = remove))
}
```

Algorithm 12: Zswitch (ZS)

```
Zswitch <- function(z, m.ref, theta, omega = 0.01){
  Count.P2 <- 0L # Counter for parameter-based relabelling
  perms <- sapply(1:K, rep, M)

  # Define function for parmaeter based relabelling
  PBR <- function(x, S.hat, theta.ref, theta, m){
    perm <- unlist(S.hat[x, ])
    sum <- 0
    for(r in 1:nrow(theta.ref)){
      sum <- sum + sum(abs((theta.ref[r,] - theta[m, perm, r]) /
          theta.ref[r,]))
    }
    return(sum)
  }

  # STEP 1: Set reference allocation vector and corresponding
     parameters
  z.ref <- factor(z[m.ref,])
  R <- dim(theta)[3]
  theta.ref <- t(theta[m.ref,,])

  # STEP 2:
  for(m in 1:M){
    z.now <- factor(z[m,])  # Labels for this iteration
```

```
# a) Construct matrix M to identify potential candidate switches
M.jk <- table(z.now, z.ref)

# b) Determine the set I
Candidate.perms <- M.jk / apply(M.jk, 1, sum) > omega
Set.I <- vector("list", K) # Pre-allocate to ensure correct dim
for(k in 1:K){
  Set.I[[k]] <- which(Candidate.perms[,k] == "TRUE")
}

# c) Determine the set S.hat
S.hat <- expand.grid(Set.I)  # Set of permutations (as a matrix)
if(nrow(S.hat) == 1){   # Remove invalid rows
  Valid.rows <- which(length(apply(S.hat, 1, unique)) == K)
}else{
  Valid.rows <- which(sapply(apply(S.hat, 1, unique), length)
     == K)
}
S.hat <- S.hat[Valid.rows, ]

# d) Determine the permutations necessary to undo label
   switching
if(length(Valid.rows) == 0){
  warning(paste0("omega too large (m = ", m,
    ") resulting in empty set 'I'. Iteration omitted."))
  perms[m, ] <- rep(NA, K)
}
if(length(Valid.rows) == 1){
  perms[m, ] <- as.numeric(S.hat)
}else{
  # Parameter-based relabelling
  Sum <- sapply(1:nrow(S.hat), PBR, S.hat, theta.ref, theta, m)
  perms[m, ] <- as.numeric(S.hat[which.min(Sum), ])
```

```
      Count.P2 <- Count.P2 + 1
    }
  }
  # Print diagnostics
  message(paste("Parameter based relabelling used", Count.P2,
     "times."))

  # Return permutations
  return(list(permutations = perms, phase2 = Count.P2))
}
```

Algorithm 13: Zswitch 2 (ZS 2)

```
Zswitch.2 <- function(z, m.ref, theta, omega = 0.01, inf.cost =
   1e10){
  perms <- sapply(1:K, rep, M)

  # Set reference allocation vector and corresponding parameters
  z.ref <- factor(z[m.ref,])
  R <- dim(theta)[3]
  theta.ref <- t(theta[m.ref,,])

  # STEP 2:
  for(m in 1:M){
    z.now <- z[m,]  # Labels for this iteration

    # a) Construct matrix M to identify potential candidate switches
    M.jk <- table(z.now, z.ref)

    # b) Determine permutations by solving LSAP
    Candidate.perms <- M.jk / apply(M.jk, 1, sum) > omega
    c <- matrix(0, K, K) # Cost matrix
    for(j in 1:K){
      for (k in 1:K){
        for(r in 1:R){
```

```
            c[j, k] <- c[j, k] + abs((theta.ref[r, k] - theta[[r]][m,
               j]) /
                theta.ref[r, k])
          }
        }
      }
      # Scale cost matrix by M
      c <- c / M.jk

      # Set impossible values to arbitrary large cost
      c[which(Candidate.perms == FALSE)] <- inf.cost

      # Solve LSAP
      Solution <- lp.assign(c)$solution
      perms[m,] <- apply(Solution, 2, which.max)
    }
    # Return permutations
    return(list(permutations = perms))
}
```

C.4 Summary of Models used in Simulation Study

Scenario 1: Poisson mixture model

$$Y_i \sim \sum_{k=1}^{K} w_k \text{Po}(\eta_k)$$

$$\boldsymbol{w} \sim \mathcal{D}(\alpha, \ldots, \alpha)$$

$$\eta_k \sim \mathcal{N}\left(\mu_k, \sigma_k^2\right) \mathbb{I}\left(0, +\infty\right)$$

$$\mu_k \sim \text{Gam}(0.5, 0.05)$$

$$\sigma_k^2 \sim \text{Gam}(1, 0.5)$$

Scenario 2: univariate Gaussian mixture model

For $K < 20$:

$$
\begin{aligned}
Y_i &\sim \sum_{k=1}^{K} w_k \mathcal{N}\left(\mu_k, \sigma_k^2\right) \\
\boldsymbol{w} &\sim \mathcal{D}(\alpha, \ldots, \alpha) \\
\mu_k &\sim \mathcal{N}\left(\lambda, \xi^2\right) \\
\sigma_k^2 &\sim \text{Gam}\left(\delta, \beta\right) \\
\lambda &\sim \mathcal{N}(0, 12.5) \\
\xi^2 &\sim \text{Gam}(3, 1) \\
\delta &\sim \mathcal{N}(0, 4)\mathbb{I}(0, +\infty) \\
\beta &\sim \mathcal{N}(0, 3)\mathbb{I}(0, +\infty)
\end{aligned}
$$

For $K = 20$, the prior for μ_k and corresponding means is modified as

$$
\begin{aligned}
\mu_k &\sim \begin{cases} \mathcal{N}(\lambda_1, \xi^2) & \text{if } k \leq 10 \\ \mathcal{N}(\lambda_2, \xi^2) & \text{if } k > 10 \end{cases} \\
\lambda_1 &\sim \mathcal{N}(-20, 12.5) \\
\lambda_2 &\sim \mathcal{N}(20, 12.5)
\end{aligned}
$$

For $K = 50$ and $K = 100$ a similar modification is used.

Scenario 3: gamma mixture model

$$
\begin{aligned}
Y_i &\sim \sum_{k=1}^{K} w_k \text{Gam}\left(\delta_k, \beta_k\right) \\
\boldsymbol{w} &\sim \mathcal{D}(\alpha, \ldots, \alpha) \\
\delta_k &\sim \text{Uni}(0, 50) \\
\beta_k &\sim \text{Gam}(3, 1.5)
\end{aligned}
$$

Appendix D

Supplementary Material from Chapter 6

D.1 R Code for generating the synthetic data

```
# Gaussian decay function (Equation (6.7)): d is the N by N distance
# matrix, and str is a real number representing the relative
# strength of spatial autocorrelation (Gaussian bandwidth
# parameter).

Gaus.decay <- function(d, str){
  exp(-0.5 * (d/str)^2)
}

Set <- 1  # Choose data set 1, 2, or 3

set.seed(1)  # Set random seed to obtain same results

# Generate Expected Values
E <- rgamma(N, 2, 2)



# Generate Covariate Values
x <- abs(rnorm(N, 0, sqrt(0.1)))  # Global noise for all areas
hs <- c(2, 21, 39, 96, 28, 73)   # Hotspot centres
str <- c(rep(0.7, 4), 0.3, 1)   # Rel. autocorrelation strength
```

```
for(k in 1:length(hs)){
  x <- x + 2 * Gaus.decay(d[hs[k],], str[k])
}

# Generate Log-Relative Risks
if(Set == 1){     # No/weak auotcorrelation
  log.eta <- rnorm(N, -1, sqrt(2))
  beta <- 0.7
}else if(Set == 2){  # Strong positive auotcorrelation
  str <- 1          # Rel. autocorrelation strength
  Corr <- Gaus.decay(d, str)  # Correlation structure
  log.eta <- Corr %*%
    rnorm(N, 0.3, 0.3)   # Random field
  beta <- 0.4
}else{          # Clustered data
  log.eta <- rnorm(N, -0.1, 1)  # Start with uncorrelated risk
  centre <- c(85, 12)   # Centre of clusters
  radius <- c(2.7, 1.5)   # Radius of clusters
  str <- c(5, 8)       # Rel. autocorrelation strength
  for(k in 1:length(centre)){
    cluster <- which(d[centre[k],] < radius[k])
    Corr <- Gaus.decay(d, str[k])
    log.eta[cluster] <- (Corr %*%
      rnorm(N, 0.035, 0.0009))[cluster,1]
  }
  beta <- 0.4
}
log.eta <- log.eta + x * beta  # Add covariate effect
eta <- exp(log.eta)     # Relative risk

# Generate Observed Values
Y <- rpois(N, E * eta)
```

D.2 WinBUGS Models

Model without smoothing (Model B):

```
model {
  for (i in 1:N) {
    y[i] ~ dpois(E.eta[i])
    E.eta[i] <- E[i] * eta[i]
    log(eta[i]) <- epsilon[i] + beta * x[i]
    epsilon[i] ~ dnorm(0,tau.epsilon)
  }
  alpha ~ dnorm(0, 0.01)
  tau.epsilon <- pow(sigma.epsilon,-2)
  sigma.epsilon ~ dnorm(0, 0.01)I(0,)
  beta ~ dnorm(alpha, 0.01)
}
```

Model with smoothing (remaining 17 models):

```
model {
  for (i in 1:N) {
    y[i] ~ dpois(E.eta[i])
    E.eta[i] <- E[i] * eta[i]
    log(eta[i]) <- alpha + gamma[i] + epsilon[i] + beta * x[i]
    epsilon[i] ~ dnorm(0,tau.epsilon)
  }
  alpha ~ dnorm(0, 0.01)
  gamma[1:N] ~ car.normal(adj[], weights[], num[], tau.gamma)
  tau.gamma <- pow(sigma.gamma,-2)
  sigma.gamma ~ dgamma(0.5, 0.05)
  tau.epsilon <- pow(sigma.epsilon,-2)
  sigma.epsilon ~ dnorm(0, 0.05)I(0,)
  beta ~ dnorm(0, 0.01)
}
```

Bibliography

Abellan, J. J., Richardson, S., and Best, N. 2008. Use of space-time models to investigate the stability of patterns of disease. *Environmental Health Perspectives*, **116** (8): 1111–1119. doi: 10.1289/ehp.10814.

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. 2002. Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, **9** (4): 341–355. doi: 10.1023/A:1020910605990.

Aguirre, J. C. 2014. *The Unlikely History of the Origins of Modern Maps*. URL: `http://www.smithsonianmag.com/history/unlikely-history-origins-modern-maps-180951617/?no-ist`. Accessed 9 Apirl 2015.

Alexander, F. E., Anderson, T. J., Brown, H. K., Forrest, A. P. M., Hepburn, W., Kirkpatrick, A. E., Muir, B. B., Prescott, R. J., and Smith, A. 1999. 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. *Lancet*, **353**(9168):1903–1908. doi: 10.1016/S0140-6736(98)07413-3.

Allcroft, D. J. and Glasbey, C. A. 2003. A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **52** (4): 487–498. doi: 10.1111/1467-9876.00419.

Anselin, L. 1988. *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer Academic Publishers.

Aro, A. R., de Koning, H. J., Absetz, P., and Marjut, S. 2001. Two distinct groups of non-attenders in an organized mammography screening program. *Breast Cancer Research and Treatment*, **70** (2):145–153. doi: 10.1023/a:1012939228916.

Asmussen, S. 2003. *Applied probability and queues*. New York: Springer.

Assunção, R. and Krainski, E. 2009. Neighbourhood dependence in Bayesian spatial smoothing. *Biometrical Journal*, **51** (5): 851–869. doi: 10.1002/bimj.200900056.

Australian Bureau of Statistics (ABS). 2011. *Statistical Geography - Australian Standard Geographical Classification (ASGC), Digital Boundaries, 2006*, 'Statistical Local Area Digital Boundaries (ASGC 2006) in ESRI Shapefile Format', data cube: ESRI Shapefile, cat. no. 1259.0.30.002. URL: www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1259.0.30.0022006.

Australian Bureau of Statistics (ABS). 2013a. *Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia*, 'Statistical Local Area, Indexes, SEIFA 2011', data cube: Excel spreadsheet, cat. no. 2033.0.55.001. URL: www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2033.0.55.0012011.

Australian Bureau of Statistics (ABS). 2013b. *Socio-Economic Indexes for Areas (SEIFA) – Technical Paper, 2011*, cat. no. 2033.0.55.001. URL: www.abs.gov.au/ausstats/abs@.nsf/DetailsPage/2033.0.55.0012011. Accessed 3 August 2014.

Australian Government Department of Health (AGDH). 2014. *BreastScreen Australia Program*. URL: www.abs.gov.au/ausstats/abs@.nsf/DetailsPage/2033.0.55.0012011. Accessed 29 June 2015.

Australian Institute of Health and Welfare (AIHW). 2012. *Breast cancer in Australia: an overview. Cancer series no. 71*, cat. no. CAN 67. Canberra: AIHW. Accessed 27 November 2014.

Australian Institute of Health and Welfare (AIHW). 2014. *Australian Cancer Incidence and Mortality (ACIM) books: Breast cancer*. Canberra: AIHW. URL: http://www.aihw.gov.au/acim-books. Accessed 8 December 2014.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: CRC Press/Chapman & Hall, Monographs on Statistics and Applied Probability 135, 2nd edition.

Barrett, F. A. 2000. Finke's 1792 map of human diseases: the first world disease map? *Social Science & Medicine*, **50** (7–8): 915–921.

Baum, L. E. and Eagon, J. A. 1967. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, **73** (3): 360–363.

Baum, L. E. and Petrie, T. 1966. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37** (6): 1554–1563. doi: 10.1214/aoms/1177699147.

Bayes, T. and Price, R. 1763. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M., F. R. S. *Philosophical Transactions of the Royal Society of London*, **53**: 370–418. doi: 10.1098/rstl.1763.0053.

Bell, K. P. and Bockstael, N. E. 2000. Applying the generalized-moments estimation approach to spatial problems involving microlevel data. *The Review of Economics and Statistics*, **82** (1): 72–82. doi: 10.1162/003465300558641.

Berkelaar, M. et al. 2015. *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, R package version 5.6.13. URL: `http://CRAN.R-project.org/package=lpSolve`.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. 1995. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, **14** (21–22): 2433–2443. doi: 10.1002/sim.4780142112.

Bernardinelli, L., Pascutto, C., Best, N. G., and Gilks, W. R. 1997. Disease mapping with errors in covariates. *Statistics in Medicine*, **16** (7): 741–752. doi: 10.1002/(SICI)1097-0258(19970415)16:7⟨741::AID-SIM501⟩3.0.CO;2-1.

Besag, J. 1974. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **36** (2): 192–236.

Besag, J. and Kooperberg, C. 1995. On conditional and intrinsic autoregressions. *Biometrika*, **82** (4): 733–746. doi: doi:10.1093/biomet/82.4.733.

Besag, J., York, J., and Mollié, A. 1991. Bayesian image restoration with application in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43** (1): 1–20. doi: 10.1007/BF00116466.

Best, N., Cockings, S., Bennet, J., Wakefield, J., and Elliott, P. 2001. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, **164** (1): 155–174. doi: 10.1111/1467-985x.00194.

Best, N., Richardson, S., and Thomson, A. 2005. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14** (1): 35–59. doi: 10.1191/0962280205sm388oa.

Bihrmann, K., Jensen, A., Olsen, A. H., Njor, S., Schwartz, W., Vejborg, I., and Lynge, E. 2008. Performance of systematic and non-systematic ('opportunistic') screening mammography: a comparative study from Denmark. *Journal of Medical Screening*, **15** (1): 23–26. doi: 10.1258/jms.2008.007055.

Bithell, J. F. 2000. A classification of disease mapping methods. *Statistics in Medicine*, **19** (17–18): 2203–15. doi: 10.1002/1097-0258(20000915/30)19:17/18⟨2203::AID-SIM564⟩3.0.CO;2-U.

Boyle, P., Muir, C. S., and Grundmann, E. e. 1989. *Cancer Mapping*. Berlin: Springer-Verlag.

Breslow, N. E. and Clayton, D. G. 1993. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*, **88** (421): 9–25. doi: 10.2307/2290687.

Brezger, A., L., F., and Hennerfeind, A. 2007. Adaptive Gaussian Markov random fields with applications in human brain mapping. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **56** (3): 327–345. doi: 10.1111/j.1467-9876.2007.00580.x.

Brook, D. 1964. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, **51** (3/4): 481–483. doi: 10.2307/2334154.

Brooks, S. P. and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7** (4): 434–455. doi: 10.1080/10618600.1998.10474787.

Brown, K. C., Fitzhugh, E. C., Neutens, J. J., and Klein, D. A. 2009. Screening mammography utilization in Tennessee women: the association with residence. *Journal of Rural Health*, **25** (2): 167–173. doi: 10.1111/j.1748-0361.2009.00213.x.

Bukard, R., Dell'Amico, M., and Martello, S. 2009. *Assignment Problems*. Society for Industrial & Applied Mathematics. Accessed 7 September 2016. doi: 10.1137/1.9780898717754.

Bulliard, J.-L., Ducros, C., Jemelin, C., Arzel, B., Fioretta, G., and Levi, F. 2009. Effectiveness of organized versus opportunistic mammography screening. *Annals of Oncology*, **20** (7): 1199–1202. doi: 10.1093/annonc/mdn770.

Burnham, K. P. and Anderson, D. R. 1998. *Model Selection and Inference: a Practical Information-theoretic Approach*. New York: Springer.

Celeux, G., Hurn, M., and Robert, C. P. 2000. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, **95** (451): 957–970. doi: 10.2307/2669477.

Chamot, E., Charvet, A. I., and Perneger, T. V. 2007. Who gets screened, and where: A comparison of organized and opportunistic mammography screening in Geneva, Switzerland. *European Journal of Cancer*, **43** (3): 576–584. doi: 10.1016/j.ejca.2006.10.017.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. 2011. *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. Boca Raton: CRC Press.

Clifford, P. 1990. 'Markov random fields in statistics'. In Grimmett, G. R. and Welsh, D. J. A., editors, *Disorder in physical systems (A volume in honour of John M. Hammersley)*, pages 19–32. Oxford: Oxford University Press.

Congdon, P. D. 2010. *Applied Bayesian Hierarchical Methods*. Boca Raton: Chapman & Hall/CRC.

Conlon, E. M. and Waller, L. A. 1999. Flexible neighborhood structures in hierarchical models for disease mapping. Technical Report 1998-018, Division of Biostatistics, University of Minnesota.

Cressie, N. A. C. 1993. *Statistics for Spatial Data*. New York: John Wiley & Sons, Inc, rev. edition. doi: 10.1002/9781119115151.

Damien, P., Wakefield, J., and Walker, S. 1999. Gibbs sampling for Bayesian nonconjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **61** (1): 331–344. doi: 10.1111/1467-9868.00179.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **39** (1): 1–38.

Devine, O. J., Louis, T. A., and Halloran, M. E. 1996. Identifying areas with elevated disease incidence rates using empirical Bayes estimators. *Geographical Analysis*, **28** (3): 187–199. doi: 10.1111/j.1538-4632.1996.tb00930.x.

Diaconis, P. and Freedman, D. 1984. 'Partial exchangeability and sufficiency'. In Ghosh, J. K. and , J., editors, *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications, and New Directions*, pages 205–236. Calcutta: Indian Statistical Institute.

Diggle, P. J. 2013. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Boca Raton: Chapman & Hall/CRC Monographs on Statistics and Applied Probability, 3 edition.

Diniz-Filho, J. A. F., Nabout, J. C., Telles, M. P. d. C., Soares, T. N., and Rangel, T. F. L. V. B. 2009. A review of techniques for spatial modeling in geographical, conservation and landscape genetics. *Genetics and Molecular Biology*, **32** (2): 203–211. doi: 10.1590/s1415-47572009000200001.

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., and Wilson, R. 2007. Methods to account for spatial autocorrelation in the analysis of species distribution data: a review. *Ecography*, **30** (5): 609–628. doi: 10.1111/j.2007.0906-7590.05171.x.

Duncan, E. W., White, N. M., and Mengersen, K. 2016. Bayesian spatiotemporal modelling for identifying unusual and unstable trends in mammography utilisation. *BMJ Open*, **6** (5): e010253. doi: 10.1136/bmjopen-2015-010253.

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R., and Beard, J. 2007. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models. *International Journal of Health Geographics*, **6** (1): 54. doi: 10.1186/1476-072x-6-54.

Engelman, K. K., Hawley, D. B., Gazaway, R., Mosier, M. C., Ahluwalia, J. S., and Ellerbeck, E. F. 2002. Impact of geographic barriers on the utilization of mammograms by older rural women. *Journal of the American Geriatrics Society*, **50** (1): 62–68. doi: 10.1046/j.1532-5415.2002.50009.x.

Fahrmeir, L. and Kneib, T. 2011. 'Spatial smoothing, interactions and geoadditive regression'. In Fahrmeir, L. and Kneib, T., editors, *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*, pages 307–414. New York: Oxford University Press. doi: 10.1093/acprof:oso/9780199533022.001.0001.

Florax, R. J. G. M. and Nijkamp, P. 2003. Misspecification in linear spatial regression models. tinbergen institute discussion paper. Technical Report 2003-081/3.

Florax, R. J. G. M. and Rey, S. 1995. 'The impacts of misspecified spatial interaction in linear regression models'. In Anselin, L. and Florax, R. J. G. M., editors, *New Directions in Spatial Econometrics*, pages 111–135. Berlin: Springer. doi: 10.1007/978-3-642-79877-1_5.

Gardner, W., Mulvey, E. P., and Shaw, E. C. 1995. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. *Psychological Bulletin*, **118** (3): 392–404. doi: 10.1037/0033-2909.118.3.392.

Gelfand, A. E. and Smith, A. F. M. 1990. Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, **85** (410): 398–409.

Gelman, A. 2006. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, **1** (3): 515–533. doi: 10.1214/06-ba117a.

Gelman, A. and Rubin, R. D. B. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science*, **7** (4): 457–511. doi: 10.1214/ss/1177011136.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. 2013. *Bayesian Data Analysis*. Boca Raton: Chapman and Hall, 3 edition.

Geman, S. and Geman, D. 1984. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6** (6): 721–741. doi: 10.2307/2334940.

Getis, A. and Aldstadt, J. 2008. 'Constructing the spatial weights matrix using a local statistic'. In *Advances in Spatial Science: Perspectives on Spatial Data Analysis*, pages 147–163. Berlin: Spring-Verlag. doi: 10.1007/978-3-642-01976-0_11.

Geweke, J. 1992. 'Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments'. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics*, pages 169–193. Oxford: Clarendon Press.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. 1996. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

Google. 2015. *Google Maps Distance Matrix API*, last modified september 21, 2015. URL: `https://developers.google.com/maps/documentation/distance-matrix/intro`.

Greene, W. H. 1994. *Some accounting for excess zeros and sample selection in Poisson and negative binomial regression models*. Working Paper EC-94-10: Department of Economics, New York University.

Griffith, D. A. 1996. 'Some guidelines for specifying the geographic weights matrix contained in spatial statistical models'. In Arlinghaus, S. L., editor, *Practical Handbook of Spatial Statistics*, pages 65–82. Boca Raton: CRC Press.

Hall, D. B. 2000. Zero-inflated Poisson and binomial regression with random effects: a case study. *Biometrics*, **56** (4): 1030–1039. doi: 10.1111/j.0006-341X.2000.01030.x.

Hammersley, J. M. and Clifford, P. 1971. *Markov fields on finite graphs and lattices*. Unpublished.

Harary, F. 1962. The determinant of the adjacency matrix of a graph. *SIAM Review*, **4** (3): 202–210. doi: 10.1137/1004057.

Hartigan, J. A. and Wong, M. A. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **28** (1): 100–108. doi: 10.2307/2346830.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57** (1): 97. doi: 10.2307/2334940.

Heilbron, D. C. 1994. Zero-altered and other regression models for count data with added zeros. *Biometric Journal*, **36** (5): 531–547. doi: 10.1002/bimj.4710360505.

Hoeting, J. A. 2009. The importance of accounting for spatial and temporal correlation in analyses of ecological data. *Ecological Applications*, **19** (3): 574–577. doi: 10.1890/08-0836.1.

Hooten, M. and Wikle, C. 2008. A hierarchical Bayesian nonlinear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove. *Environmental and Ecological Statistics*, **15** (1): 59–70. doi: 10.1007/s10651-007-0040-1.

Huang, B., Dignan, M., Han, D., and Johnson, O. 2009. Does distance matter? Distance to mammography facilities and stage at diagnosis of breast cancer in Kentucky. *The Journal of Rural Health*, **25** (4): 366–371. doi: 10.1111/j.1748-0361.2009.00245.x.

Huque, M. H., Anderson, C., Walton, R., and Ryan, L. 2016. Individual level covariate adjusted conditional autoregressive (indiCAR) model for disease mapping. *International Journal of Health Geographics*, **15** (25): 1–13. doi: 10.1186/s12942-016-0055-7.

Hyndman, J. C. G., Holman, C. D. J., and Dawes, V. P. 2000. Effect of distance and social disadvantage on the response to invitations to attend mammography screening. *Journal of Medical Screening*, **7** (3): 141–145. doi: 10.1136/jms.7.3.141.

Ibrahim, J. G. and Laud, P. W. 1994. A predictive approach to the analysis of designed experiments. *Journal of the American Statistical Association*, **89** (425): 309–319. doi: 10.2307/2291227.

Jackson, M. C., Davis, W. W., Waldron, W., McNeel, T. S., Pfeiffer, R., and Breen, N. 2009. Impact of geography on mammography use in California. *Cancer Causes Control*, **20** (8): 1339–1353. doi: 10.1007/s10552-009-9355-6.

Jasra, A., Holmes, C. C., and Stephens, D. A. 2005. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science*, **20** (1): 50–67. doi: 10.1214/088342305000000016.

Jensen, A., Olsen, A. H., von Euler-Chelpin, M., Njor, S. H., Vejborg, I., and Lynge, E. 2005. Do Nonattenders in mammography screening programmes seek mammography elsewhere? . *International Journal of Cancer*, **113** (3): 464–470. doi: 10.10.1002/ijc.20604.

Johnson, G. D. 2004. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. *International Journal of Health Geographics*, **3** (1): 29. doi: 10.1186/1476-072X-3-29.

Jørgensen, K. J. and Gøtzsche, P. C. 2009. Overdiagnosis in publicly organized mammography screening programmes: systematic review of incidence trends. *British Medical Journal*, **339** (7714): 206–209. doi: 10.1136/bmj.b2587.

Kandhasamy, C. and Ghosh, K. 2017. Relative risk for HIV in India an estimate using conditional auto-regressive models with Bayesian approach. *Spatial and Spatio-temporal Epidemiology*, **20** : 27–34. doi: 10.1016/j.sste.2017.01.001.

Kelsall, J. and Wakefield, J. 2002. Modeling spatial variation in disease risk. *Journal of the American Statistical Association*, **97** (459): 692–701. doi: 10.1198/016214502388618438.

Knorr-Held, L. and Besag, J. 1998. Modelling Risk from a Disease in Time and Space. *Statistics in Medicine*, **17** (18): 2045–2060.

Kreher, N. E., Hickner, J. M., Ruffin, M. T., and Lin, C. S. 1995. Effect of distance and travel time on rural womens compliance with screening mammography: An UPRNet study. *Journal of Family Practice*, **40** (2): 143–147.

Kuhnert, P. M. 2003. *New methodology and comparisons for the analysis of binary data using Bayesian and tree based methods*. PhD thesis, Queensland University of Technology.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34** (1): 1–14. doi: 10.2307/1269547.

Langford, I. H., Leyland, A. H., Rasbash, J., and Goldstein, H. 1999. Multilevel modeling of the geographical distributions of diseases. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **48** (2): 253–268. doi: 10.1111/1467-9876.00153.

Lee, D. 2011. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, **2** (2): 79–89. doi: 10.1016/j.sste.2011.03.001.

Lee, D. 2013. CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, **55** (13): 1–24. doi: 10.18637/jss.v055.i13.

Lee, D. and Mitchell, R. 2012. Boundary detection in disease mapping studies. *Biostatistics*, **13** (3): 415–426. doi: 10.1093/biostatistics/kxr036.

Legler, J., Breen, N., Meissner, H., Malec, D., and C., C. 2002. Predicting patterns of mammography use: a geographic perspective on national needs for intervention research. *Health Services Research*, **37** (4): 929–947. doi: 10.1034/j.1600-0560.2002.59.x.

Lekdee, K. and Ingrisawang, L. 2013. Generalized linear mixed models with spatial random effects for spatio-temporal data: an application to dengue fever mapping. *Journal of Mathematics and Statistics*, **9** (2): 137–143. doi: 10.3844/jmssp.2013.137.143.

Leroux, B. G., Lei, X., and Breslow, N. 1999. Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, **116** : 179–191. doi: 10.1007/978-1-4612-1284-3_4.

Li, G., Best, N., Hansell, A., I., A., and Richardson, S. 2012. BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice. *Biostatistics*, **13** (4): 695–710. doi: 10.1093/biostatistics/kxs005.

Lu, H., Reilly, C., Banerjee, S., and Carlin, B. 2007. Bayesian areal wombling via adjacency modelling. *Environmental and Ecological Statistics*, **14** (4): 433–452. doi: 10.1007/s10651-007-0029-9.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. 2000. WinBUGS a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10** (4): 325–337. doi: 10.1023/a:1008929526011.

MacNab, Y. C. 2003. Hierarchical Bayesian spatial modelling of small area rates of non-rare disease. *Statistics in Medicine*, **22** (10): 1761–1773. doi: 10.1002/sim.1463.

MacQueen, J. 1967. 'Some methods for classification and analysis of multivariate observations'. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. Berkeley: University of California Press.

Makuc, D. M., Breen, N., and Freid, V. 1999. Low Income, Race, and the Use of Mammography. *Health Services Research*, **34** (1): 229–239.

Mandelblatt, J. S., Yabroff, K. R., and Kerner, J. F. 1999. Equitable access to cancer services: a review of barriers to quality care. *Cancer*, **86** (11): 2378–2390. doi: 10.1002/(SICI)1097-0142(19991201)86:11⟨2378::AID-CNCR28⟩3.0.CO;2-L.

Marin, J.-M. and Robert, C. P. 2007. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer-Verlag.

Marin, J.-M., Mengersen, K., and Robert, C. P. 2005. 'Bayesian modelling and inference on mixtures of distributions'. In Rao, C. and Dey, D., editors, *Handbook of Statistics*, pages 459–507. New York: Springer-Verlag. doi: 10.1016/s0169-7161(05)25016-2.

Maxwell, A. J. 2000. Relocation of a static breast screening unit: a study of factors affecting attendance. *Journal of Medical Screening*, **7** (2): 114–115. doi: 10.1136/jms.7.2.114.

McCulloch, C. E. 1999. 'An introduction to generalised linear mixed models'. In Friedl, H., Berghold, A., and Kauerman, G., editors, *Proceedings of the 11th International Workshop on Statistical Modelling*. Graz: Graz University of Technology.

Mersmann, O. 2015. *microbenchmark: accurate timing functions*, R package version 1.4-2.1. URL: `http://CRAN.R-project.org/package=microbenchmark`.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, **21** (6): 1087–1092.

Miles, A., Cockburn, J., Smith, R. A., and Wardle, J. 2004. A perspective from countries using organized screening programs. *Cancer*, **101** (S5): 1201–1213. doi: 10.1002/cncr.20505.

Moran, P. A. P. 1950. Notes on continuous stochastic phenomena. *Biometrika*, **37** (1/2): 17–23. doi: 10.2307/2332142.

Morrison, K. T., Nelson, T. A., Nathoo, F. S., and Ostry, A. S. 2012. Application of Bayesian spatial smoothing models to assess agricultural self-sufficiency. *International Journal of Geographical Information Science*, **2** (7): 1213–1229. doi: 10.1080/13658816.2011.633491.

Mugglin, B. P., A. S. amd Carlin, Zhu, L., and Conlon, E. 1999. Bayesian areal interpolation, estimation, and smoothing: an inferential approach for geographic information systems. *Environment and Planning A*, **31** (8): 1337–1352. doi: 10.1068/a311337.

Nelder, J. A. and Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **135** (3): 370–384.

Ntzoufras, I. 2009. *Bayesian Modeling Using WinBUGS*. Hoboken: Wiley. doi: 10.1002/9780470434567.

Pan, J.-C., Liu, C.-M., Hwu, H.-G., and Huang, G.-H. 2015. Allocation variable-based probabilistic algorithm to deal with label switching problem in Bayesian mixture models. *PLoS ONE*, **10** (10): e0138899. doi: 10.1371/journal.pone.0138899.

Papastamoulis, P. 2013. Handling the label switching problem in latent class models via the ECR algorithm. *Communications in Statistics Part B: Simulation and Computation*, **43** (4): 913–927. doi: 10.1080/03610918.2012.718840.

Papastamoulis, P. 2015. *label.switching: Relabelling MCMC Outputs of Mixture Models*, R package version 1.4. URL: `http://CRAN.R-project.org/package=label.switching`.

Papastamoulis, P. 2016. An R package for Dealing with the Label Switching Problem in MCMC Outputs. *Journal of Statistical Software*, **69** (1): 1–24. doi: 10.18637/jss.v069.c01.

Papastamoulis, P. and Iliopoulos, G. 2010. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, **19** (2): 313–331. doi: 10.1198/jcgs.2010.09008.

Papastamoulis, P. and Iliopoulos, G. 2013. On the convergence rate of Random Permutation Sampler and ECR algorithm in missing data models. *Methodology and Computing in Applied Probability*, **15** (2): 293–304. doi: 10.1007/s11009-011-9238-7.

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Staines, A., and Elliot, P. 2000. Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*, **19** (17-18): 2493–2519. doi: 10.1002/1097-0258(20000915/30)19:17/18⟨2493::aid-sim584⟩3.0.co;2-d.

Pauli, F. and Torelli, N., 2015. Relabelling in bayesian mixture models by pivotal units. arXiv:1501.05478v1.

Peek, M. E. and Han, J. H. 2004. Disparities in screening mammogrpahy. *Journal of General Internal Medicine*, **19** (2): 184–194. doi: 10.1111/j.1525-1497.2004.30254.x.

Puolamäki, K. and Kaski, S. 2009. 'Bayesian solutions to the label switching problem'. In Adams, N. M., Robardet, C., Siebes, A., and Boulicaut, J.-F., editors, *Advances in Intelligent Data Analysis VIII*, pages 381–392. Berlin: Spring-Verlag. doi: 10.1007/978-3-642-03915-7_33.

R Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL: `http://www.R-project.org/`.

R Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL: `http://www.R-project.org/`.

Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77** (2): 257–86. doi: 10.1109/5.18626.

Rasmussen, C. E. 2000. 'The Infinite Gaussian Mixture Model'. In Solla, S., Leen, T., and M$'$uller, K.-R., editors, *Advances in Neural Information Processing Systems 12*, pages 237–259. Oxford: Oxford University Press.

Rasmussen, S. 2004. Modeling of the discrete spatial variation in epidemiology with SAS using GLIMMIX. *Computer Methods and Programs in Biomedicine*, **76** (1): 83–89. doi: 10.1016/j.cmpb.2004.03.003.

Redner, R. A. and Walker, H. F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *Society for Industrial and Applied Mathematics*, **26** (2): 195–239. doi: 10.1137/1026034.

Richardson, S. 2003. 'Spatial models in epidemiological applications'. In Green, P. J., Hjort, N. L., and Richardson, S., editors, *Highly structured stochastic systems*, pages 237–259. Oxford: Oxford University Press.

Richardson, S. and Green, P. J. 1997. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59** (4): 731–792. doi: 10.1111/1467-9868.00095.

Robert, C. P., Celeux, G., and Diebolt, J. 1993. Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistical & Probability Letters*, **16** (1): 77–83. doi: 10.1016/0167-7152(93)90127-5.

Robert, C. 2014. *Xi'ans Og*, label switching in Bayesian mixture models, October 31. URL: `https://xianblog.wordpress.com/2014/10/31/label-switching-in-bayesian-mixture-models`. Accessed 24 September, 2015.

Rodriguez, C. E. and Walker, S. 2014. Label Switching in Bayesian Mixture Models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, **23** (1): 25–45. doi: 10.1080/10618600.2012.735624.

Rossi, P. E., Allenby, G. M., and McCulloch, R. 2005. 'Markov Chain Monte Carlo Methods'. In *Bayesian Statistics and Marketing*. Chichester: John Wiley & Sons. doi: 10.1002/0470863692.ch3.

Rue, H. and Held, L. 2005. *Gaussian Markov random fields: theory and applications*. Boca Raton: CRC Press.

Rydén, T. and Titterington, D. M. 1998. Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, **7** (2): 194–211. doi: 10.2307/1390813.

Schall, R. 1991. Estimation in generalized linear models with random effects. *Biometrika*, **78** (4): 719–727. doi: 10.2307/2336923.

Selvin, E. and Brett, K. M. 2003. Breast and cervical cancer screening: sociodemographic predictors among white, black, and Hispanic women. *American Journal of Public Health*, **93** (4): 618–623. doi: 10.2105/ajph.93.4.618.

Shaddick, G. and Zidek, J. V. 2016. 'Spatial patterns in disease'. In *Spatio-Temporal Methods in Environmental Epidemiology*. Boca Raton: CRC Press.

Shen, Y., Yang, Y., Inoue, L. Y. T., Munsell, M. F., Miller, A. B., and Berry, D. A. 2005. Role of detection method in predicting breast cancer survival: analysis of randomized screening trials. *Journal of the National Cancer Institute*, **97** (16): 1195–1203. doi: 10.1093/jnci/dji239.

Snow, J. 1855. *On the Mode of Communication of Cholera*. London: Churchill, 2 edition.

Sperrin, M., Jaki, T., and Wit, E. 2010. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, **20** (3): 357–366. doi: 10.1007/s11222-009-9129-8.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and der Linde, V. 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64** (4): 583–640. doi: 10.1111/1467-9868.00353.

Stephens, M. 2000a. Bayesian analysis of mixture models with an unknown number of components an alternative to reversible ump methods. *Annals of Statistics*, **28** (1): 40–74.

Stephens, M. 2000b. Dealing with label Switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **62** (4): 795–809. doi: 10.1111/1467-9868.00265.

Sturtz, S., Ligges, U., and Gelman, A. 2005. R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, **12** (3): 1–16. doi: 10.18637/jss.v012.i03.

Temple Lang, D. and the CRAN team. 2015a. *RCurl: General Network (HTTP/FTP/...) Client Interface for R*, R package version 1.95-4.7. URL: `http://CRAN.R-project.org/package=RCurl`.

Temple Lang, D. and the CRAN team. 2015b. *XML: Tools for Parsing and Generating XML Within R and S-Plus*, R package version 3.98-1.3. URL: `http://CRAN.R-project.org/package=XML`.

Tobler, W. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, **46** (2): 234–240. doi: 10.2307/143141.

van Havre, Z., White, N., Rousseau, J., and Mengersen, K. 2015. Overfitting Bayesian mixture models with an unknown number of components. *PLoS ONE*, **10** (7): e0131739. doi: 10.1371/journal.pone.0131739.

Vanmarcke, E. 2010. *Random Fields: Analysis and Synthesis*. Singapore: World Scientific Publishing Company. doi: 10.1142/5807.

Vieira, A. M. C., Hinde, J. P., and Demetrio, C. G. B. 2000. Zero-inflated proportion data models applied to a biological control assay. *Journal of Applied Statistics*, **27** (3): 373–389. doi: 10.1080/02664760021673.

Waclawiw, M. A. and Liang, K. 1993. Prediction of Random Effects in the Generalized Linear Model. *Journal of the American Statistical Association*, **88** (421): 171–178. doi: 10.2307/2290711.

Wakefield, J. and Elliott, P. 1999. Issues in the statistical analysis of small area health data. *Statistics in Medicine*, **18** (17-18): 2377–3299. doi: 10.1002/(SICI)1097-0258(19990915/30)18:17/18⟨2377::AID-SIM263⟩3.0.CO;2-G.

Wall, M. M. 2004. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **121** (2): 311–324. doi: 10.1016/s0378-3758(03)00111-3.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. 1997. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92** (438): 607–617. doi: 10.1080/01621459.1997.10474012.

Waller, L. A., Goodwin, B. J., Wilson, M. L., Ostfeld, R. S., Marshall, S., and Hayes, E. B. 2007. Spatio-temporal patterns in county-level incidence and reporting of Lyme disease in the northeastern United States, 1990–2000. *Environmental and Ecological Statistics*, **14** (1): 83–100.

Wheeler, D. C. 2013. 'Geographically weighted regression'. In M., F. and Nijkamp, P., editors, *Handbook of Regional Science*, pages 1435–1459. Berlin: Springer. doi: 10.1007/978-3-642-23430-9_77.

Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. 2001. Bayesian spatiotemporal modelling for identifying unusual and unstable trends in mammography utilisation. *BMJ Open*, **6** (5): e010253. doi: 10.1136/bmjopen-2015-010253.

Xia, H. and Carlin, B. P. 1998. Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Statistics in Medicine*, **17** (18): 2025–2043. doi: 10.1002/(SICI)1097-0258(19980930)17:18⟨2025::AID-SIM865⟩3.0.CO;2-M.

Yang, Q. H., Snyder, J. P., and Tobler, W. R. 2000. *Map projection transformation: Principles and applications*. London: Taylor & Francis.

Yao, W. 2013. A simple solution to Bayesian mixture labelling. *Communications in Statistics Part B: Simulation and Computation*, **42** (4): 800–813. doi: 10.1080/03610918.2012.655825.

Zeger, S. L. and Karim, M. R. 1991. Generalized linear models with random effects; a Gibbs sampling approach. *Journal of the American Statistical Association*, **86** (413): 79–86. doi: 10.1080/01621459.1991.10475006.

Zenk, S. N., Tarlov, E., and Sun, J. 2006. Spatial Equity in Facilities Providing Low- or No-Fee Screening Mammography in Chicago Neighborhoods. *Journal of Urban Health*, **83** (2): 195–210. doi: 10.1007/s11524-005-9023-4.

Zhang, C. 2012. 'Spatial prominence and spatial weights matrix in geospatial analysis'. In *Progress in Geospatial Analysis*, pages 73–84. doi: 10.1007/978-4-431-54000-7_5.

Zhang, H. 2002. On Estimation and Prediction for Spatial Generalized Linear Mixed Models. *Biometrics*, **58** (1): 129–136. doi: 10.1111/j.0006-341x.2002.00129.x.