Modelling techniques for areal spatial data

By:

Kevin Horan

Under the supervision of:

Prof. Chris Brunsdon   Dr. Katarina Domijan

A thesis submitted in fulfillment of the requirements
for the Ph.D. degree in Statistics

at the

Hamilton Institute
Maynooth University
Maynooth, Co. Kildare, Ireland

August 2025

Declaration

I, Kevin Horan, declare that this thesis titled, "**Modelling techniques for areal spatial data**" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 3 September 2025

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Chris Brunsdon and Dr. Katarina Domijan for their guidance and support over the past four years. Every meeting and discussion we had was a pleasure and I learned so much from you both.

I am grateful to the CRT programme directors, Prof. Ken Duffy, Prof. David Malone, Prof. Claire Gormley and Prof. James Gleeson, and to Janet, Joanna, Rosemary, Kate, Patsy, Peg and all at the CRT and in the Hamilton Institute for making everything run smoothly.

A particular thanks to Prof. Ken Duffy for his encouragement in the early days that this was something I could do.

Thanks to Research Ireland for funding my PhD and all the fantastic conferences, masterclasses and trips.

Special thanks to my 2021 cohort at Maynooth: Conor, Dara, Gabriel, Jack, Niloufar, Oluwayomi, Solmaz and Victor. Thanks also go to all the others at UCD and UL, and to those ahead of me in the CRT program who were always helpful with tips and suggestions, especially Chang, Darshana, Jonathan, Nathan and Shauna.

To Jina, Skittles and lots of short-term feline guests, thank you for being wonderful housemates.

A big thank you as always to my wonderful supportive family.

Funding

This work was supported by Taighde Éireann – Research Ireland under grant number 18/CRT/6049.

Taighde Éireann
Research Ireland

Abstract

Accounting for spatial processes is an important aspect of modelling data which are associated with a geographical location. Failure to do so can compromise a model's performance. These processes can operate in different ways depending on the underlying mechanisms at play.

It is reasonable, for example, to expect that the characteristics of a single agricultural field should be closely related to those of other fields within the same farm, tended to by the same farmer. Similarly, various such farms in a single area of governance may be subject to one set of regulations leading them to more closely resemble each other than farms in another jurisdiction. By capturing this spatial hierarchy of field within farm within jurisdiction in a model, we would expect the model to perform better.

In another sense, it also seems reasonable to expect that fields which are geographically close to each other should be more similar than distant fields. They may be subject to similar climate, soil composition and local traditions of land use, so even if they are operated by different farmers on either side of a fence, their characteristics are not independent. Here we would expect that a spatially autoregressive model which accounts for this should perform better.

This thesis develops models which combine both of these types of spatial processes. Rather than looking at fields and farms, we instead focus on voter behaviour in individual constituencies across the UK, all of which are nested within counties and regions.

We begin by applying such a modelling structure, accounting for both of these types of spatial effect, to the 2019 UK General Election in England and Wales. Using this methodology, we can examine the proportion of variation in behaviour which is attributable to different levels of grouping, and estimate spatially varying coefficients.

A key component of such modelling is the construction of neighbourhood matrices which encode whether or not spatial units are to be considered as neighbours and thus more likely to share similarities than other units. We present an R package, `sfislands`, which reduces the workload in creating such matrices when complications occur due to the presence of islands or other geographic sources of discontiguity.

We conclude by applying a novel methodology to the 2024 UK General Election, which seeks to capture both of the above spatial effects in a different way. The proposed model allows the degree to which neighbouring constituencies are expected to be similar to vary according to hierarchical position. By comparing the plausibility of this framework to other candidate combinations of spatial structure, we find that this model represents the more plausible explanation of the underlying spatial processes of party support in this election.

Collaborations

Chris Brunsdon: As my supervisor, Professor Brunsdon (Maynooth University) supervised and collaborated on the work of all chapters. This includes reviewing and editing all chapters.

Katarina Domijan: As my supervisor, Dr. Domijan (Maynooth University) supervised and collaborated on the work of all chapters. This includes reviewing and editing all chapters.

Publications

The chapters contained in this thesis have been published by the peer-reviewed journals listed below.

Peer-reviewed published chapters:

- K. Horan, C. Brunsdon, and K. Domijan. A multilevel spatial model to investigate voting behaviour in the 2019 UK General Election. *Applied Spatial Analysis and Policy*, 17(2):703–727, Jan. 2024. URL `http://dx.doi.org/10.1007/s12061-023-09563-6`.

- K. Horan, K. Domijan, and C. Brunsdon. sfislands: Streamlines the process of fitting areal spatial models, 2024. *The R Journal*, 17(2):84-108, Jun. 2025. URL `http://dx.doi.org/10.32614/RJ-2025-015`.

- K. Horan, K. Domijan, and C. Brunsdon. Incorporating varying degrees of spatial cohesion in models of voter behaviour in the UK General Election 2024. *Journal of the Royal Statistical Society Series C: Applied Statistics*, Oct. 2025. URL `http://dx.doi.org/10.1093/jrsssc/qlaf055`.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

When modelling data which have spatial characteristics, failure to account for location can break the assumption of independence inherent in many modelling frameworks and lead to unreliable outcomes. How we go about accounting for geography depends not only on the type of spatial data (points, lines, polygons, networks etc.), but on the theories we might have to explain the role that location plays in each context. This thesis is concerned with the modelling of areal spatial datasets (where the spatial units are polygons) which exhibit two specific structural characteristics.

The first is that they are hierarchical in nature, whereby it should be possible to cluster the individual data units into groups and sub-groups, each of which is expected to display some degree of internal cohesion, distinct from other groups. In the geographical context, this could be national or sub-national divisions (e.g. countries, states, provinces) where, despite all other conditions being equal, we might still expect a different outcome from one group compared to another. This type of structure is often modelled using hierarchical modelling [28, 44].

The second characteristic is that nearby data units are expected to be more similar to each other than distant ones, regardless of whether or not two units lie on different sides of a grouping boundary [70]. In the context of areal data, conditional autoregressive (CAR) models are often used to capture this characteristic [5, 7, 6].

The combination of these two processes fits into the category of hierarchical spatial autoregressive (HSAR) modelling [19]. In this context, the hierarchical components are referred to as *vertical*, analogous to a branching tree. A spatial unit is expected to share characteristics with other units according to the similarity of their *vertical* branching paths. The autoregressive components are termed *horizontal*, where it is the values of nearby spatial units, as seen horizontally across the terminal nodes of a tree regardless of the path of their branching structures, which are expected to be similar.

While such data structures are prevalent across many domains, including environmental and ecological research, crime analysis, public health investigations, and housing market research, we predominantly focus on the case of voter behaviour in the UK. The aim of the research outlined in this thesis is both to develop novel methodologies in this area and to present an R package which streamlines the implementation of such models.

The chapters fit together as a story of adding increasing degrees of complexity or subtlety to the process of creating models of this type. We begin in chapter 2 by showing how these *vertical* and *horizontal* processes can be combined using the long-established and easily interpretable `spdep` and `mgcv` R packages. These provide, respectively, functions to construct neighbourhood matrices by contiguity, and modelling structures which use these matrices to combine both hierarchical and autoregressive processes. Chapter 3 focuses on developments to the neighbourhood structure component of the model, and considers ways of adjusting this structure to be consistent with additional spatial assumptions which the modeller may feel are appropriate. As such modifications are not easily implementable using `spdep`, a new package, `sfislands`, is introduced to streamline this process. Chapter 4 then focuses on developments to the modelling component, previously fitted using `mgcv`, which is shifted to a Bayesian framework where it is possible to implement further subtleties which `mgcv` and other packages for fitting CAR models do not allow.

1.2 Thesis outline

The thesis is presented as a collection of articles which have been published in peer-reviewed journals. Due to the interconnected nature of the research topics, some repetition of background material, methodology descriptions, and key findings may occur across chapters. This repetition has been retained to preserve the stand-alone nature and readability of each individual article.

Chapter 2: A multilevel spatial model to investigate voting behaviour in the 2019 UK General Election

A model of voter behaviour which incorporates a hierarchical structure and autoregressive component is implemented. The aforementioned *vertical* and *horizontal* effects are combined in a single model of the change in allegiance among voters from Labour to the Conservative Party, as measured by the Butler swing, using data from the 2019 UK General Election in England and Wales. This model is fit in the frequentist paradigm using the `mgcv` package in R, which conveniently offers the capabilities of including both hierarchical and autoregressive components within one model. Constituencies are nested according to county and region. Using this methodology, we can examine the proportion of variation in swing which is attributable to each level, and estimate spatially varying coefficients.

Chapter 3: `sfislands`: An R package for accommodating islands and disjoint zones in areal spatial modelling

We introduce the R package `sfislands`. For reasons of identifiability, all spatial units in a study must be contiguous to at least one other unit if they are to be included in autoregressive models. This condition is not always met in practice. For example, there are a small number of constituencies in England and Wales which are islands, albeit not very distant from the coast, and a decision must be made as to how to handle their inclusion or otherwise. This package streamlines the process of including these islands as neighbours to their nearest k constituencies, eliminating the need for manual geographic assignments. We describe the package's functions and consider two case studies of use: one in Indonesia and another in London.

In the first case, we are confronted with the large number of islands which constitute Indonesia and a scenario where the modeller may not be familiar with all of their names. We examine the incidence of strong earthquakes between 1985 and 2023. This is a process where nearby places might be expected to have similar risks, and discontiguity between two nearby islands does not imply an interruption in the seismic process.

In the second case study, while all of the wards of London are contiguous to at least one other, there is a geographical feature (in this case, a river) which may be an important consideration in the development of a neighbourhood structure. We demonstrate the use of `sfislands` to define a structure where wards which are separated by the River Thames are considered to be neighbours if they lie within 1 kilometre of a crossing such as a bridge or tunnel.

Chapter 4: Incorporating varying degrees of spatial cohesion in models of voter behaviour in the 2024 UK General Election

Instead of combining *vertical* and *horizontal* processes in the manner described in chapter 2, this chapter proposes a novel method which incorporates potential *vertical* effects as a parameter within the *horizontal* autoregressive component. In other words, the degree to which the *horizontal* effects occur, whereby neighbouring spatial units have a tendency to be similar to each other, can vary according to their position within a hierarchical *vertical* structure. To illustrate this, we examine the number of votes per constituency obtained by each of four parties in the 2024 UK General Election: Labour, Conservative, Liberal Democrat, and Reform UK. In this case, the response variable is modelled as a Poisson distribution using a Bayesian framework. We seek to control for common explanations of voter behaviour using census data and the *status quo ante* from the prior election of 2019. By comparing a selection of candidate models which capture spatial variation in different ways, this novel method emerges as the most plausible.

2

A multilevel spatial model to investigate voting behaviour in the 2019 UK General Election

2.1 Abstract

This paper presents a modelling framework which can detect the simultaneous presence of two different types of spatial process. The first is the variation from a global mean resulting from a geographical unit's *vertical* position within a nested hierarchical structure such as the county and region where it is situated. The second is the variation at the smaller scale of individual units due to the *horizontal* influence of nearby locations. The former is captured using a multi-level modelling structure while the latter is accounted for by an autoregressive component at the lowest level of the hierarchy. Such a model not only estimates spatially-varying parameters according to geographical scale, but also the relative contribution of each process to the overall spatial variation. As a demonstration, the study considers the association of a selection of socio-economic attributes with voting behaviour in the 2019 UK General Election. It finds evidence of the presence of both types of spatial effects, and describes how they suggest different associations between census profile and voting behaviour in different parts of England and Wales.

2.2 Introduction

While it is common to capture spatially varying phenomena using models based on a multi-level framework or using an autocorrelation component, the objective of this study is to build a model to test for the simultaneous presence of both of these processes. This is done by combining a nested tree-structure of administrative boundaries with an additional spatially autocorrelated process at the lowest level of the geographical hierarchy. This framework makes it possible to allocate spatial variation according to process and geographical scale. The output of such a model is a set of spatially-varying coefficients at different hierarchical levels, a spatially autocorrelated random component at the lowest level of the hierarchy, and an estimation of the relative contribution of each to overall variance. The autocorrelated component can take the form of a collection of random intercepts, random coefficients, or both, according to specification.

It thus falls into the category of hierarchical spatial autoregressive (HSAR) modelling introduced by Dong and Harris [19], except that it uses a frequentist rather than a Bayesian approach.

The model is applied to a case study of voter behaviour, examining the association of census variables which have been used by Beecham et al. [4] to study spatial variation in recent voting patterns, with voters' tendency to change allegiance from Labour to the Conservative Party in the 2019 UK General Election. By combining hierarchical and autocorrelated processes, it finds that at different levels of geographical hierarchy, the association of these variables with voting outcomes varies in magnitude and direction across the study area. A constituency with a greater ethnic diversity, for example, has markedly different associations with voting behaviour in London and parts of the East Midlands than it does in the North East. It also finds evidence of spatial effects reflecting a similarity among neighbouring locations, which are free to operate across administrative boundaries. Certain places have an additional tendency to vote a certain way which cannot be explained by census profile or nested location. Overall, it estimates that approximately 27% of the variation between constituencies is accounted for by the spatial hierarchy, while a further 41% can be attributed to constituency level '*spillover effects*' from adjoining constituencies.

This multilevel spatial approach to analysis is suitable not only for the study of elections, but can be easily adapted to any context or distributional family where a combination of spatial processes are hypothesised to be at play simultaneously, such as disease mapping [73] and survival models for businesses [9].

2.2.1 Spatial processes

The incorporation of spatial processes into models is a generic issue across quantitative human geography [25]. Firstly, human geography is not like physical geography in that it does not necessarily obey universal laws. While it may be possible to identify associations of certain covariates with an outcome, it is not always the case that these associations will be the same in different places. Furthermore, the observations within a dataset which has a geographic component can not be seen as independent. As Tobler's First Law of Geography [70] states, "everything is related to everything else, but near things are more related than distant things". These additional realities should be reflected in the structure of a spatial model.

Hierarchical process

One way to account for location is by using hierarchical or multi-level models, which model different spatial units at different levels. Much of the work done by Goldstein [28] in the development of multi-level models was focused on education research. In such a context, pupil outcomes could be seen as depending not only on the various decisions of local education boards (highest level), but within each of those, on the policies of different head-teachers in each school, and subsequently on the skills of or decisions made by each teacher within each school. The lowest level would then be the individual pupil. These different levels, however, could also be nested geographical divisions, such as regions, counties and electoral constituencies. The introduction of geographical levels was developed by Jones [44], and such a framework has previously been applied to voting behaviour [45]. This nested process can be characterised as *vertical* in the sense of correlations extending up and down through a branching tree.

This structure, however, requires us to know a priori what the appropriate scales are and to introduce hard boundaries accordingly. Standard mixed modelling assumes that beyond the random effects at these scales, no further correlation exists. In a geographical context, while this may explain a certain amount of the process, there could be a further spatial process which is better described using a spatially-autocorrelated framework.

Spatial autocorrelation process

Spatial processes can manifest themselves in a manner which is not consistent with discrete hard boundaries but is instead a continuous process where the value of each unit is related to that of its neighbours. Such modelling is long-established and commonplace in human geography, beginning with Geary [27]'s discussion of issues of spatial autocorrelation, or *contagion*, when examining Irish agricultural data, and the introduction of *kriging* [48], where point data is used to predict values of other unknown points based on proximity as measured by distance. Metrics such as Geary's C and Moran's I use different approaches to quantify this phenomenon.

For aggregated areal data, a similar principle to *kriging* can lead to the construction of contiguity matrices to capture proximity not in terms of distance, but whether or not areal units are adjacent. Unlike the hierarchical framework, this process can be seen more as a moving focal point. In any particular location, it is the places immediately adjoining it which influence it the most. As the moving-window of focus shifts to the next location, it will in turn share many of the same influences but will gain some new ones. In this sense, it can be seen as capturing *horizontal* correlations between adjoining units at the same level. When this process is extended across the study area, a different type of spatial effect is captured. In this framework, a priori groupings are irrelevant and correlation between spatial units is based only on their proximity.

ICAR models

This process in areal data can be captured by conditional autoregressive (CAR) structures, first introduced by Besag [5], of which intrinsic conditional autoregressive (ICAR) models are one type [6]. A CAR model captures spatial relationships using a contiguity matrix where all pairs of spatial units are classified as either neighbours or not neighbours. The conditional expected value for each unit then depends only on the values of adjacent units. In this way, it is an example of a Markov random field. An ICAR model assumes complete correlation between all units, the strength of which is based on their degree of contiguity. This broader dependence allows it to capture more extensive spatial relationships than the CAR structure.

Combination of both processes

But it can certainly be the case that both hierarchical and autoregressive processes are operating simultaneously.

In the hierarchical education context discussed above, a policy enacted by one education board could lead to an increase in school funding for a certain sport within their zone of governance. Such a policy would cease immediately upon crossing over into the neighbouring authority. However, the effects would not necessarily be so rigid. Should this sport prove popular, it is likely that the children will begin playing it more with their friends, regardless of what school their friends may attend. The same could then be true for friends of friends and this process would follow a contagion-like autocorrelated spatial pattern consistent with the distribution of children's friendship groups. Thus the process can propagate in ways which are not defined by the arbitrary borders of school administration policy.

2.2.2 Contribution

Our contribution is to outline an easily implementable methodology which incorporates both types of process simultaneously. It combines a hierarchical approach with a spatially autocorrelated component at the lowest level, which can not only model the process more accurately, but also allocates estimated variance according the level and type of spatial process. This aspect can be seen as analogous to an *analysis of variance* for different causes of spatial variability. Unlike a similar model applied to travel satisfaction in Beijing by Dong et al. [20], this model is fitted within a frequentist framework, using the well-established `mgcv` package in R (Wood 2011). It can be easily adapted to a range of different outcome distribution types beyond the Gaussian structure of the following example. While lme4 [3] and nlme [63] are popular R packages designed specifically for multi-level modelling, they are limited in their ability to take geography into account. The mgcv [77] package is primarily used for constructing a range of generalised additive models (GAMs), but it also contains functionality to create multilevel models equivalent to those in more specialised packages by using random effects splines. It has the further capability to combine these with Gaussian processes and Markov random fields, which are suitable for the type of spatial autocorrelation proposed above.

We demonstrate the implementation of such a model using a case study based on the UK General Election of 2019.

2.3 Data

The data used for this example are drawn from the parlitools R package [55], which includes a number of useful resources for analysing UK politics. It provides, among other features, convenient access to the British Election Study's record of recent published election results, and information from the 2011 census aggregated to the constituency level.

The 2019 UK election saw a large gain in support for the Conservative Party, often at the expense of the Labour Party. Many of their seat gains occurred in constituencies which they had not recently won, despite substantial improvements over the previous two election cycles. The Labour Party, the second largest party by a substantial margin and their principal competitor, received its lowest number of MPs since 1935. This contrasts with the Conservatives gaining a majority of 80 seats, their largest since 1987. Often referred to as the collapse of Labour's '*red wall*' [46], many of its losses followed a geographical pattern, notably in a collection of constituencies in the North and Midlands [68] which had been traditional Labour strongholds for many decades, albeit with declining majorities in recent elections.

In the context of UK voter behaviour, as with any data which has a geographical component, it is reasonable to hypothesise that there would be spatial processes at play in addition to differences which might be associated with socio-economic factors alone. Constituencies with similar types of census profile but in different locations do not necessarily produce similar election results. We further suspect that these spatial processes might occur in the form of both a *vertical* hierarchical process and lowest level *horizontal* neighbourhood effect. Dorling [21] has described how UK society has become ever more geographically fragmented since the 1970s. If the symptoms of this were a factor in voting allegiance, we could expect a spatial hierarchical structure to pick up on this. It is also well-established that voter behaviour is often subject to a neighbourhood effect, which Pattie and Johnston [60] summarised with the phrase "people who talk together vote together", referencing work done by Miller [52]. The presence of such an effect would take the form of positive spatial autocorrelation among neighbouring constituencies.

The specific change in voting behaviour which we seek to model in this example is the shift in allegiance from one party to another, referred to as '*swing*'.

2.3.1 Geographical context

Before discussing the dependent and explanatory variables of the proposed model, the geographical context is set. Data is restricted to England and Wales in this study. The reason is that, while Scotland and Northern Ireland saw interesting dynamics of their own in the 2019 election, the Conservatives and Labour were not the two primary competing parties in these parts of the UK. Neither party features to any extent in Northern Ireland, while the Scottish National Party (SNP) has dominated recent elections in Scotland.

Boundaries

The multilevel component of this model consists of three levels. The lowest level is composed of 571 individual *constituencies* in England and Wales[a]. This is the level at which election results are officially reported and, due to the secrecy of the ballot, is the lowest available unit of published voting data. Each of these is nested within one of 53 *counties*. London is modelled as both a region and county. The highest level considered is the *region*, although a still higher level of *nation* has been implicitly accounted for given the exclusion of Scotland and Northern Ireland, and the classification of all of Wales as a single *region*. The *region* level has 11 components. These are Wales, Merseyside, and the nine regions of England. With the exception of Merseyside, these are essentially the NUTS level 2 administrative divisions, and coincide more or less with the former European Parliament constituencies. These regions are shown in a guide map in Figure 2.1.

Merseyside has been extracted from the North West region and treated as a region in its own right for the purposes of this study. As it is well-known that Merseyside has consistently shown distinctive voting patterns in the past [46, 43], particularly in respect of the Conservative party, it is here given the opportunity to show variance in its own right. Such separation has become common in recent analyses of the Brexit vote (see Gordon (2018) [31]).

[a] There were actually 573 constituencies in England and Wales at this time but two have been excluded as they were held by the respective Speakers of the house in 2017 and 2019, and are traditionally not contested. In all constituency-level maps in this paper, these seats are coloured black.

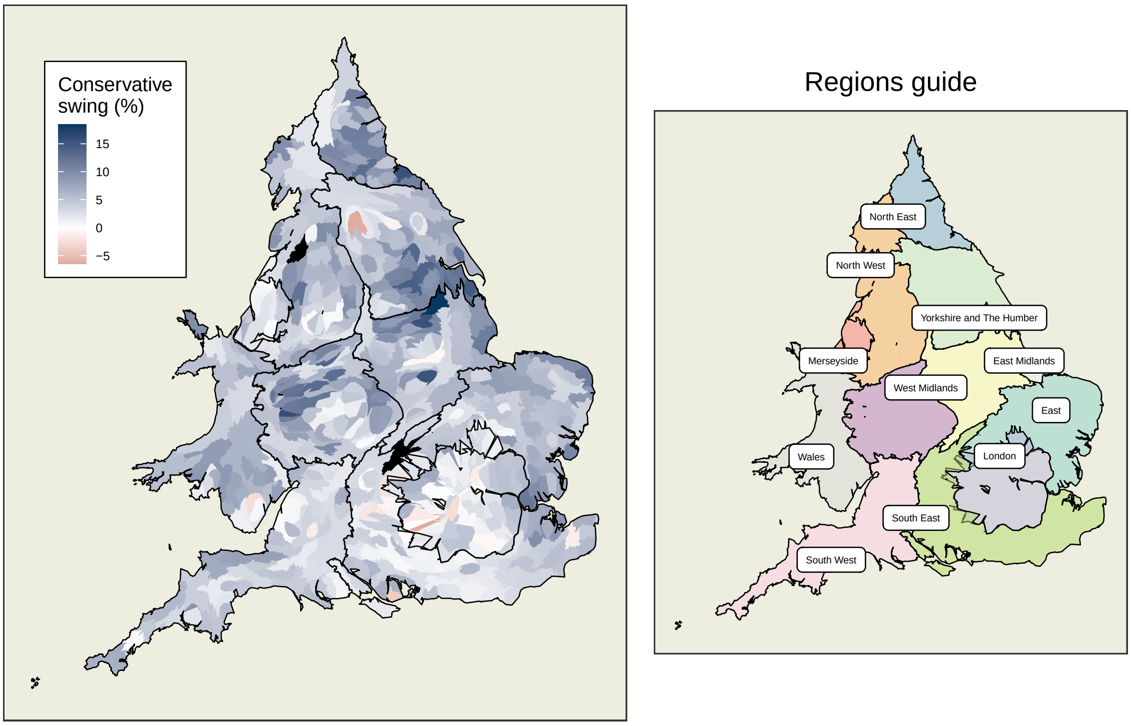


Figure 2.1: Butler swing in England and Wales and regions guide.

(L) Values of dependent variable, Butler swing to the Conservatives, mapped across constituencies of England and Wales. The vast majority of constituencies recorded a positive swing. Figures projected as Dougenik cartograms such that equal populations occupy equal area while maintaining constituency contiguities. (R) Guide map of the regions of England and Wales under a similar projection.

Contiguities

The spatial autocorrelation process at the lowest (constituency) level is based on whether or not constituencies are neighbours. Here, this structure is represented by first order queen contiguity (see Figure 2.2), where a constituency is considered a neighbour of another if they share at least one common point of boundary. Some additional contiguities have been added to account for invisible connectivity due to bridges and ferry crossings.

2.3.2 Dependent variable

Election swing is typically expressed as a positive or negative percentage point change. In the context of this analysis, the phenomenon under examination is the apparent change in voter preference in favour of the Conservatives and at the expense of Labour from the 2017 to the 2019 elections. The measurement of swing used here to represent this process is the '*conventional*', '*uniform*', or '*Butler*' swing [15], which is commonly utilised

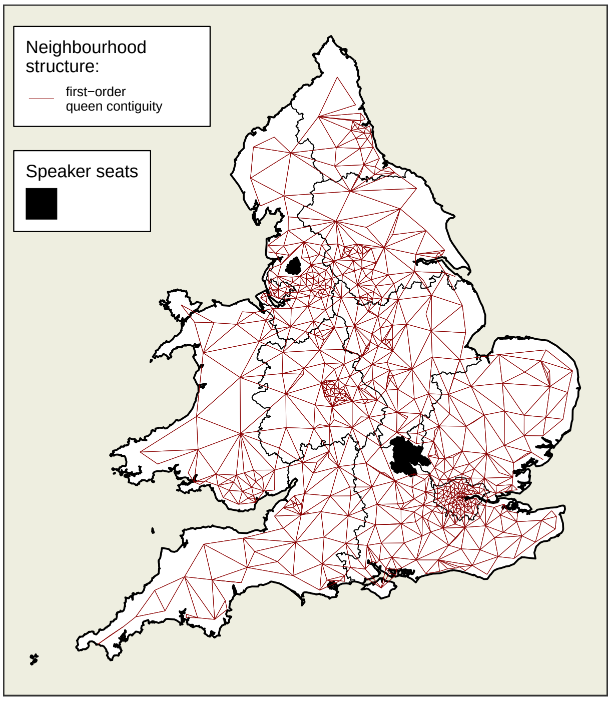


Figure 2.2: Modified queen contiguity structure of constituencies in England and Wales.

First order queen contiguity structure of constituencies in England and Wales, shown as edges radiating from nodes at the centroid of constituencies. Non-contiguous constituencies with bridge, ferry or tunnel services are also considered neighbours. Contiguity occurs regardless of region or county boundaries.

in popular discourse concerning election results. It is well-known to the public as it has been used in national television coverage of election results in the UK for many decades. The Butler swing is defined as the average of the percentage point gain of party A and the percentage point loss of party B. Thus the swing to the Conservatives in the context of this election can be represented as follows:

$$\text{Butler Swing} = \frac{(Con2019 - Con2017) - (Lab2019 - Lab2017)}{2}$$

where $Con2019$ and $Con2017$ represent the percentage of votes which were cast for the Conservative party in 2019 and 2017 respectively, while $Lab2019$ and $Lab2017$ correspond to the equivalent for the Labour party. It is calculated on the basis of total number of votes cast, including those cast for candidates other than Conservative or Labour. For example, an increase of Conservative vote share by 4.9%, combined with a decrease in Labour vote share of 7.9% would lead to a swing from Labour to the Conservatives of

$$\frac{4.9\% - (-7.9\%)}{2} = 6.4\%$$

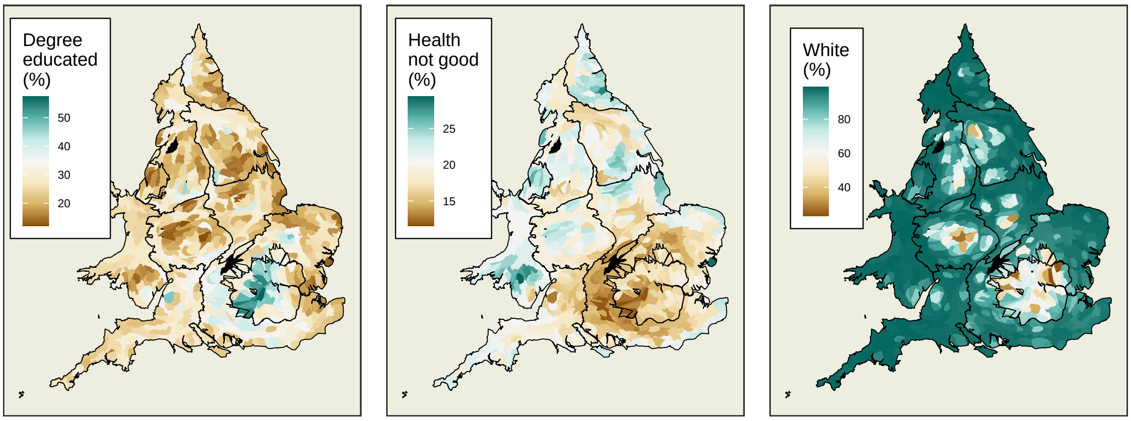


Figure 2.3: Independent variables from model.

Values of independent variables mapped across England and Wales. Figures projected as Dougenik cartograms such that equal populations occupy equal area while maintaining constituency contiguities.

Put another way, if the Conservatives benefited from a two-percentage point swing having initially had an equal vote share, they would now have a four-percentage point majority over Labour.

British politics has traditionally been dominated by two parties which has made this measure of swing particularly suitable. In other situations, an alternative known as the ‘Steed’ swing can be used [18]. This follows an identical formula except that the percentage point scores are calculated relative to the total votes received only by the two parties of interest, rather than the total number of votes cast. However, when the same two parties occupy the first two places at successive elections, as is the case in the vast majority of seats in England and Wales, the Butler swing is considered a meaningful measure of change in support [71].

Such was the strength of the Conservatives’ performance in 2019 that they only experienced negative swing in 25 constituencies, almost half of which were in London. This dominance is represented by the Dougenik cartogram in Figure 2.1, a style of map which is used throughout this study. It is a distorted map of England and Wales where constituency sizes are inflated or deflated, while maintaining contiguities, such that equal population is represented by equal space on the map [22], implemented by the `cartogram` R package [41]. Such a map is particularly informative in this case because most small constituencies are very densely populated and vice versa. A heavily populated part of London would be virtually invisible on a standard choropleth map projection. These maps overcome this problem in that every unit area contains exactly the same number of people.

Table 2.1: Candidate explanatory variables for England and Wales swing model.

Explanatory variables considered by Beecham et al. (2018), separated into three thematic groupings.

Post-Industrial / Knowledge-Economy	**Diversity / Values / Outcomes**	**Metropolitan / Big-City**
degree educated	english-speaking	EU born, not UK
professional occupations	single-ethnicity	own home
younger adults	health not good	don't own car
	white	private transport to work
	christian	

2.3.3 Explanatory variables

The explanatory variables chosen for this study come from the 2011 census, the most recent prior to the election. They are based on those proposed by Beecham et al. [4] in their examination of spatial variation of voter behaviour in the 2016 Brexit referendum. They considered covariates based on "the media discourse around the Leave vote: that of the '*left-behind*' and of the varying experiences of de-industrialisation" [17]. This is consistent with the aforementioned spatial fragmentation process [21]. Places described as '*left-behind*' are often characterised by "chronic low skills, socially conservative and nativist values", as opposed to other areas with "more affluent, highly-educated and diverse populations" [30]. The candidate variables and their thematic groupings considered by Beecham et al. are reproduced in Table 2.1.

Preliminary models showed that those variables in the *metropolitan / "big-city"* category were not significant in the context of swing in the 2019 election. For this reason, the have been omitted from this study.

Upon examination of the remaining variables, shown in Table 2.2, three groups can be discerned. Each contains highly correlated variables which also show similar associations with swing. In order to mitigate against multicollinearity, one variable was chosen from each of these groups for this model as representative of this category. Table 2.3 shows how these three explanatory variables are calculated.

The variable *degree educated* is the percentage of the population of a constituency with at least a level 4 qualification (such as undergraduate degrees or similar qualifications). Figure 2.3 shows that the highest levels of this measure are concentrated overwhelmingly in London and its environs. Small pockets of high values can also be seen in other core cities such as Manchester and Bristol. The lower scores are found in areas which might be considered more peripheral, in particular a strip from South Wales to the Humber

Table 2.2: Grouped candidate explanatory variables.

Reduced list of relevant explanatory variables grouped by justification, and the choice of one representative variable from each group for subsequent models.

Grouped Candidate Variables	**Justification / Theory**	**Representative Variable**
group 1		
degree educated professional occupations	post-industrial / knowledge-economy / peripherality	degree educated
group 2		
younger adults health not good	life outcomes / young people	health not good
group 3		
english-speaking single-ethnicity white christian	ethnic / cultural diversity / values	white

Table 2.3: Description of explanatory variables used in subsequent models.

Explanatory Variable	**Calculation from Census**
degree educated	percentage of population with level 4 qualification or higher
health not good	percentage of population self-reporting 'poor', 'bad', or 'very bad' health
white	percentage of population of white ethnicity

estuary.

The variable *health not good* is the percentage of the population in the census who self-report their health as '*poor*', '*bad*', or '*very bad*'. It is assumed that there is a strong association between health outcomes and overall quality of life. It is also biologically more likely that areas with a higher proportion of younger people will have lower levels of poor health, other things being equal. Looking again at Figure 2.3, this variable shows the strongest indication of a north-south divide in England and Wales. It suggests a stark difference between values either side of a line drawn from the Bristol Channel to the Lincolnshire coast. Areas with particularly poor health outcomes can be seen in parts of South Wales, Merseyside and the North East. Unlike many of the other variables, London does not score at either extreme of *health not good.* Instead, it is areas to the West of London, stretching across to Bristol which show the lowest levels of poor health.

The final independent variable is *white.* This is the percentage of a constituency's population who identify as being of exclusively white ethnicity. Figure 2.3 shows that this measure of ethnic diversity has a different pattern again. While the areas below 50% white are predominantly urban constituencies, it is notable that not all large cities fall into this category. Some cities are composed of a much more ethnically diverse population than others.

Prior to constructing models, these explanatory variables are scaled such that they exhibit mean of zero and standard deviation of one. This means that units of increase in a dependent variable refer to unit changes in standard deviation of percentage points of that variable, which makes comparability of effects more interpretable. Furthermore, the mean of zero allows intercepts in regression models to be viewed as the estimated level of a dependent variable with all explanatory variables held at their mean value.

2.4 Model

A simple linear regression using these three explanatory variables and no spatial component, that is without hierarchy or spatially autocorrelated effects, produces residuals as shown in Figure 2.4. It is clear that swing in certain regions is predominantly over or under-estimated. These can be observed as block patterns of red or blue respectively. For example, the South West are Merseyside are overwhelmingly 'red' indicating that Labour actually performed better than would be predicted by this model. It also suggests lower level county variations in places such as the North East where there are major blocks of both 'red' and 'blue', again indicative of systemic error in one direction or another. All of this is suggestive of the presence of unaccounted-for hierarchical spatial processes. There are also clusters of similarly coloured constituencies which cross over regional boundaries, most notably between Yorkshire and the Humber and the East Midlands. These suggest

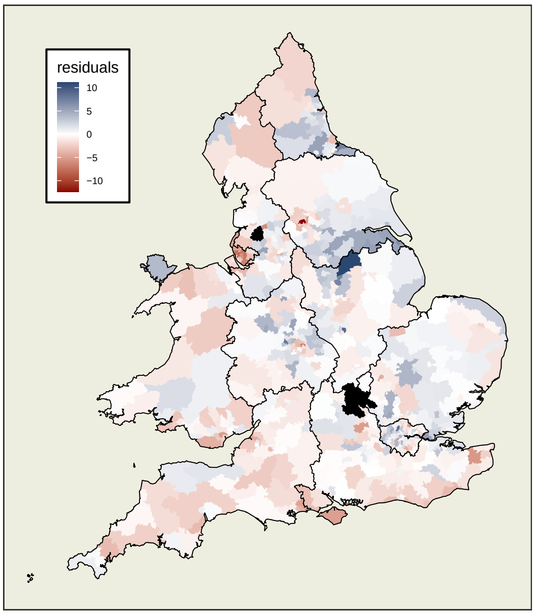


Figure 2.4: Residuals map from non-spatial model.

Map of residuals from a simple linear model which does not take geography into account. Regions such as the South West and Merseyside appear to be almost completely red (overprediction of swing), the North East show a block of red alongside a block of blue, while a blue pattern of underprediction spreads across the boundary between the East Midlands and Yorkshire and the Humber.

autocorrelated processes, operating beyond a hierarchical structure, which have not been captured.

Our proposed modelling framework allows for the inclusion of both hierarchical and autocorrelated effects, both of which are suspected to be present from a theoretical perspective and from examination of the residuals in Figure 2.4. The model described below, a hierarchical model with a spatially autocorrelated random component for each constituency, is calibrated to test for the presence of such processes and estimate them at different scales.

2.4.1 Hierarchical component

The form of the hierarchical model is as follows: Butler swing is taken as the response variable, and is assumed to have a linear relationship with each explanatory variable, controlling for the others. The three explanatory variables are *degree educated*, *health not good*, and *white*, as discussed earlier. The level of swing for any individual constituency is modelled as an intercept plus a linear combination of these three covariates. Each intercept and slope coefficient is composed of the overall mean slope and intercepts for

England and Wales, plus a differential of intercept and slopes for each region relative to the overall means, and another differential for each county relative to the means of the region in which it is nested.

The hierarchical structure means that, for example, a county intercept of zero would indicate that the county's mean is no different than the mean of the region in which it lies, controlling for the independent variables. Similarly, a regional level slope coefficient of zero for a particular explanatory variable would indicate that the association of that variable at the regional level is no different than the overall mean coefficient level, again holding other components constant.

In mixed models, the random intercepts and coefficients are not modelled directly. Instead, they are assumed to be Normally distributed with mean zero and their variance and covariances are estimated using, in this case, restricted maximum likelihood (REML). By examining whether confidence intervals around these variances include zero, hypotheses can be tested as to whether the slopes and coefficients vary significantly relative to the higher administrative level. The '*best linear unbiased predictors*' (BLUPs) of slope and intercept for individual regions and counties can then be calculated from the estimated variance components using the empirical Bayes method, which involves computing the posterior distribution of the random effects at each level of the hierarchy and taking the conditional mean as the BLUP for each unit at that level.

2.4.2 ICAR component

In addition to this, a spatially smoothing intrinsic conditional autoregressive (ICAR) component, as discussed earlier, is added at the lowest level to account for spatial dependence between adjacent constituencies which is not captured by the nested structure. The degree to which each neighbour of a given constituency influences it is proportional to the total number of neighbours of that constituency. Neighbours are defined by adjacency in this example but it could equally represent degree of connectivity by infrastructure, patterns of commuting, location of population concentrations, etc. A constituency is not considered to be its own neighbour, and each constituency is considered to be a neighbour of another if their boundaries share at least one common point. In the case of our model, the set of values generated by this component can be seen as random effects for each individual constituency.

Finally, there is an error term for each constituency which has mean zero and variance σ^2. This term accounts for differences in swing which can not be attributed to the three chosen independent variables, their nested location within county and region, or the constituency neighbourhood structure.

2.4.3 Model structure

The structure of the model is outlined below:

$$\begin{aligned} y_{ijk} &= \beta_0 + \beta_1 degree_{ijk} + \beta_2 health_{ijk} + \beta_3 white_{ijk} \\ &+ b_{0i} + b_{1i} degree_{ijk} + b_{2i} health_{ijk} + b_{3i} white_{ijk} \\ &+ b_{0ij} + b_{1ij} degree_{ijk} + b_{2ij} health_{ijk} + b_{3ij} white_{ijk} \\ &+ \gamma_l | \gamma_m, l \neq m \\ &+ \epsilon_{ijk} \end{aligned}$$

where y_{ijk} is the swing in constituency k in county j in region i for

- $i = 1, ..., 11$ regions,
- $j = 1, ..., J_i$ counties within region i,
- $k = 1, ..., K_{ij}$ constituencies within county j within region i, and
- $l = 1, ..., 571$ individual constituencies.
- β_0, β_1, β_2, β_3 are fixed effects.
- b_{0i}, b_{1i}, b_{2i}, b_{3i} are the random effects (intercept and three slopes) associated with region i,
- b_{0ij}, b_{1ij}, b_{2ij}, b_{3ij} are the random effects (intercept and three slopes) associated with county j in region i.
- ϵ_{ijk} are independent Normally distributed error terms.

Rather than estimate each of the random effect coefficients directly, the variance of each random effect is instead estimated. For the region and county level random effects, each is assumed to be independent of the others within its level, and to be Normally distributed with mean of zero. This independence is a key restriction in multi-level modelling with `mgcv` as opposed to other packages.

The γ_l's are constituency level random effects which model the spatial interactions at the lowest level of the model, based on an ICAR distribution. Let there be $m = 1, ..., M$ potential neighbouring constituencies, where $M = L = 571$. Each γ_l is conditional on the sum of the weighted values of its neighbouring γ_m's ($\mathrm{w}_{lm}\gamma_m$) and has unknown variance. As a constituency is not a neighbour to itself, the full conditional distribution can be written as follows:

$$\gamma_l | \gamma_m, l \neq m \sim \mathcal{N}\left(\frac{\sum_{l \neq m} \gamma_l}{d_l}, \frac{\sigma_l^2}{d_l} \right)$$

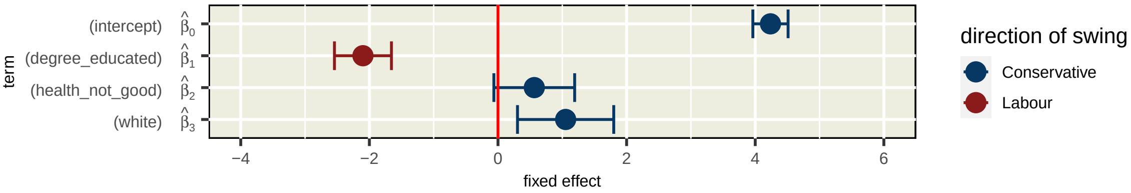


Figure 2.5: Fixed effects.

Plot of fixed or global intercept and coefficients from combined model, coloured according to direction of association with swing. A higher proportion of people of white ethnicity in an average constituency is associated with a swing to the Conservatives while the opposite is true for increases in the proportion of degree-educated.

where the term d_l represents the number of neighbours. Thus the mean of each γ_l is equal to the average of its neighbours, while its variance decreases as the number of neighbours increases.

The joint specification of the ICAR random vector γ when centred at 0 with common variance 1 rewrites to the pairwise difference formulation:

$$\gamma \propto \exp\Bigg(-\frac{1}{2}\Sigma_{l \neq m}(\gamma_l - \gamma_m)^2 \Bigg)$$

To overcome the problem of unidentifiability, the constraint $\Sigma_L \gamma_l = 0$ is added to centre the model.

2.5 Results

The aim of this modelling structure was to enable us

1. to test for the presence of spatial effects resulting in different associations between covariates and the dependent variable according to geography, taking into account both hierarchical and autoregressive spatial processes (*spatial heterogeneity*),
2. and also to estimate the relative variance associated with each type of process at different spatial scales (‘*analysis of variance*’ of spatial processes).

Firstly, the global intercept and coefficients, having controlled for region, county and constituency spatial effects, are shown in Figure 2.5. They suggest that on average, an increase in the *white* proportion of a constituency by one standard deviation, which can be interpreted as lower diversity, is significantly associated with a 1 percentage point swing to the Conservatives. In contrast, an increase of similar size in *degree educated* people among a constituency's population is associated with a 2.1 percentage point swing away from them. However, these mean global effects do not tell the full story.

2.5.1 Region level

Looking at the random effects at the region level in Figure 2.6, it is clear that divergences in association occur. This is not the case, however, for variable 1, *degree educated*. At this level, its association with swing does not vary significantly from the global mean ($\hat{\beta}_1$) of -2.1. An increase in the percentage of *degree educated* voters within a constituency is associated with a swing away from the Conservatives.

Looking at the various random coefficients for *health not good* ($\hat{\beta}_{2i}$'s for i regions), not only do they operate in different directions, the negative values are often sufficiently large to counteract the positive global effect ($\hat{\beta}_2$) of 0.564. Thus, while a higher proportion of a constituency's population with poor health is generally associated with an increase in swing to the Conservatives, this effect is much stronger in the Midlands regions and the North East, but is reversed in Merseyside, and neutralised in the East and the South East.

A similar process occurs for the third covariate, although the heterogeneities occur in different regions. The global association of a higher proportion of *white* ethnicity in a constituency and swing to the Conservatives ($\hat{\beta}_3$) is positive. This is much more strongly the case in the North East and Yorkshire and the Humber. However, in the East Midlands, London and Merseyside, the counter-acting negative random effect is enough to reverse the direction of association. It is also drawn closer to zero for Wales and the South West.

Table 2.4 shows that these region-level hierarchical effects account for 18.5% of total variance in the data.

2.5.2 County level

Looking at the county level random effects of the first covariate, *degree educated* (Figure 2.7), we can see within-region deviations. These are particularly pronounced in the East. However, while they do strengthen or weaken the negative association (-2.1) of this covariate with swing to the Conservatives across all regions, they are not sufficient at any location to alter its direction.

Table 2.4 shows that there is no further significant random effect divergence at county level for the second variable, *health not good*.

Similarly to *degree educated*, the *white* covariate (Figure 2.8) shows strong within-region county level variation, particularly in Nottinghamshire and Leicestershire in the East Midlands, which show a strengthening and a reversal respectively of the positive association between low ethnic diversity and swing to the Conservatives within this region.

Table 2.4: Model variance by geographical level.

Variance explained by the model, with associated measures of significance, at different levels and for different spatial processes. There is significant variation in the association with 'health' and 'white' at the region level. At county level, it is 'degree' and 'white' which show significant divergence.

Level	**Variance**	**Variance %**	**Cumulative Variance %**	**F-test p-val**	
region					
$\sigma^2_{region,intercept}$	<0.01	<0.01	<0.01	0.074	.
$\sigma^2_{region,degree}$	<0.01	<0.01	<0.01	0.344	
$\sigma^2_{region,health}$	0.63	7.2	7.2	0	***
$\sigma^2_{region,white}$	0.99	11.2	18.5	0.001	***
county					
$\sigma^2_{county,intercept}$	<0.01	<0.01	18.5	0.082	.
$\sigma^2_{county,degree}$	0.44	5	23.5	0	***
$\sigma^2_{county,health}$	<0.01	<0.01	23.5	0.54	
$\sigma^2_{county,white}$	0.34	3.9	27.4	0.045	*
constituency (ICAR)					
$\sigma^2_{constituency,ICAR}$	3.63	41.3	68.7	0	***
residuals					
σ^2	2.74	31.3	100		***

Note:

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

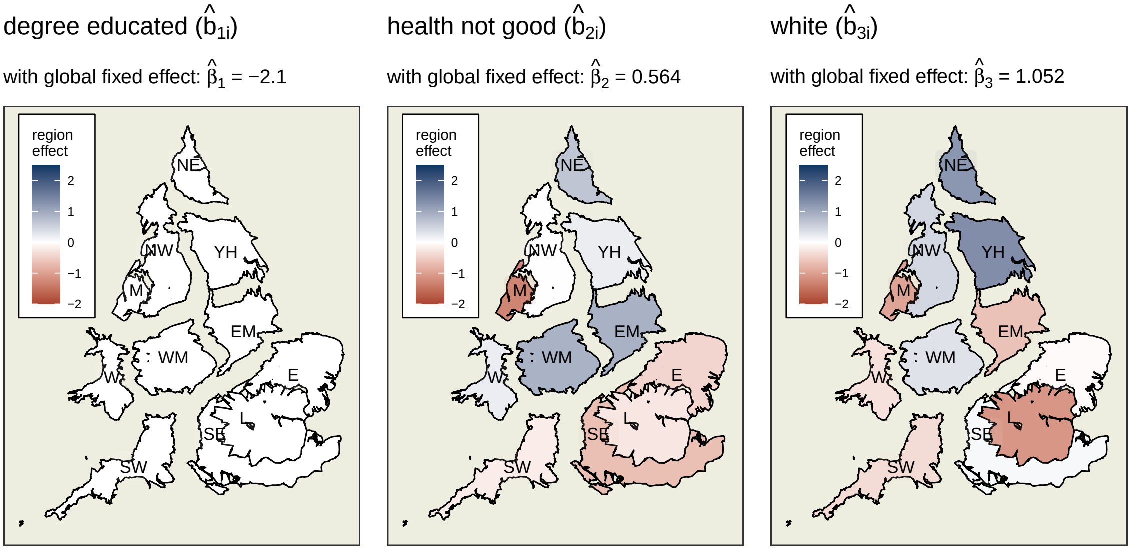

degree educated ($\hat{b}_{1i}$)
with global fixed effect: $\hat{\beta}_1$ = −2.1
health not good ($\hat{b}_{2i}$)
with global fixed effect: $\hat{\beta}_2$ = 0.564
white ($\hat{b}_{3i}$)
with global fixed effect: $\hat{\beta}_3$ = 1.052
region effect
2
1
0
−1
−2
NE
NW
YH
M
EM
WM
W
E
L
SE
SW

Figure 2.6: Region-level random effects.

Regions of England and Wales, coloured according to direction and magnitude of region-level random effects of each covariate with swing to the Conservatives. The global fixed effect from which these divergences occur is shown above each map. Unlike 'health not good' and 'white', 'degree educated' does not show significant divergence at this level from its global coefficient.

Referring again to Table 2.4, these county-level hierarchical effects account for 8.9% of total variance in the data and are much less impactful than those at the region level.

2.5.3 Net hierarchical effects

These hierarchical random effects at region and county level can be added to the global coefficients to show a picture of the net associations and their variability across England and Wales (Figure 2.9). In the case of *degree educated*, for instance, this is the sum of the global coefficient ($\hat{\beta}_1$) and the random effects for region ($\hat{b}_{1i}$) and county ($\hat{b}_{1ij}$).

As discussed above, the level of *degree educated* in a constituency is negatively associated with swing to the Conservatives in all locations, but less so in some counties. Looking at *health not good*, higher levels of poor health outcomes within a constituency are associated with a swing to the Conservatives to varying degrees, except for Merseyside and the East and South East where it is reversed or negligible. Finally, the *white* covariate suggests that lower levels of ethnic diversity in a constituency are associated with swing to the Conservatives in most of England and Wales. This is particularly strong in certain counties of the North East and Yorkshire and the Humber. However, the reverse is also observed in London, Merseyside and part of the East Midlands.

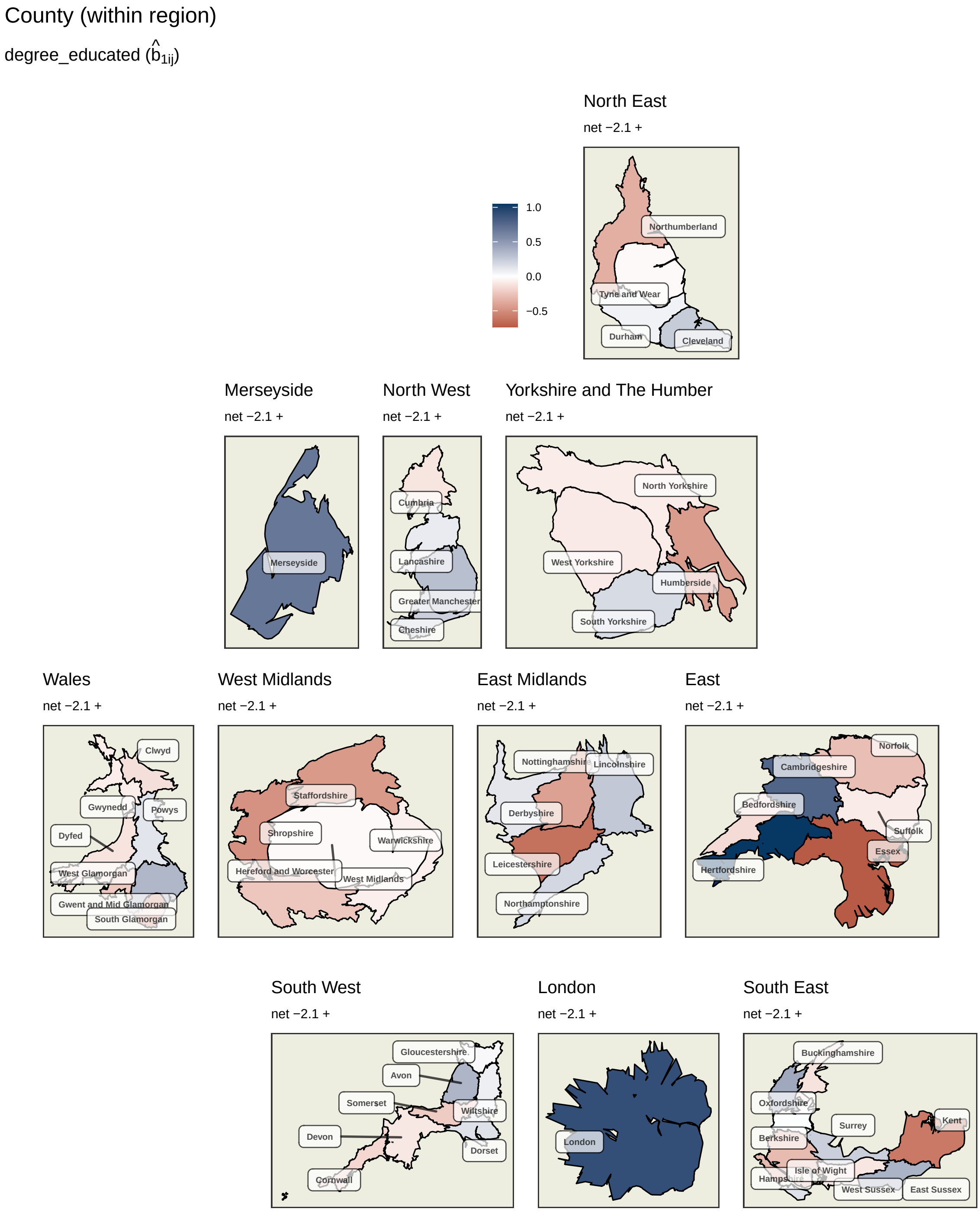


Figure 2.7: County-level 'degree educated' random effects in England and Wales. Particularly strong within-county variation can be observed in the East.

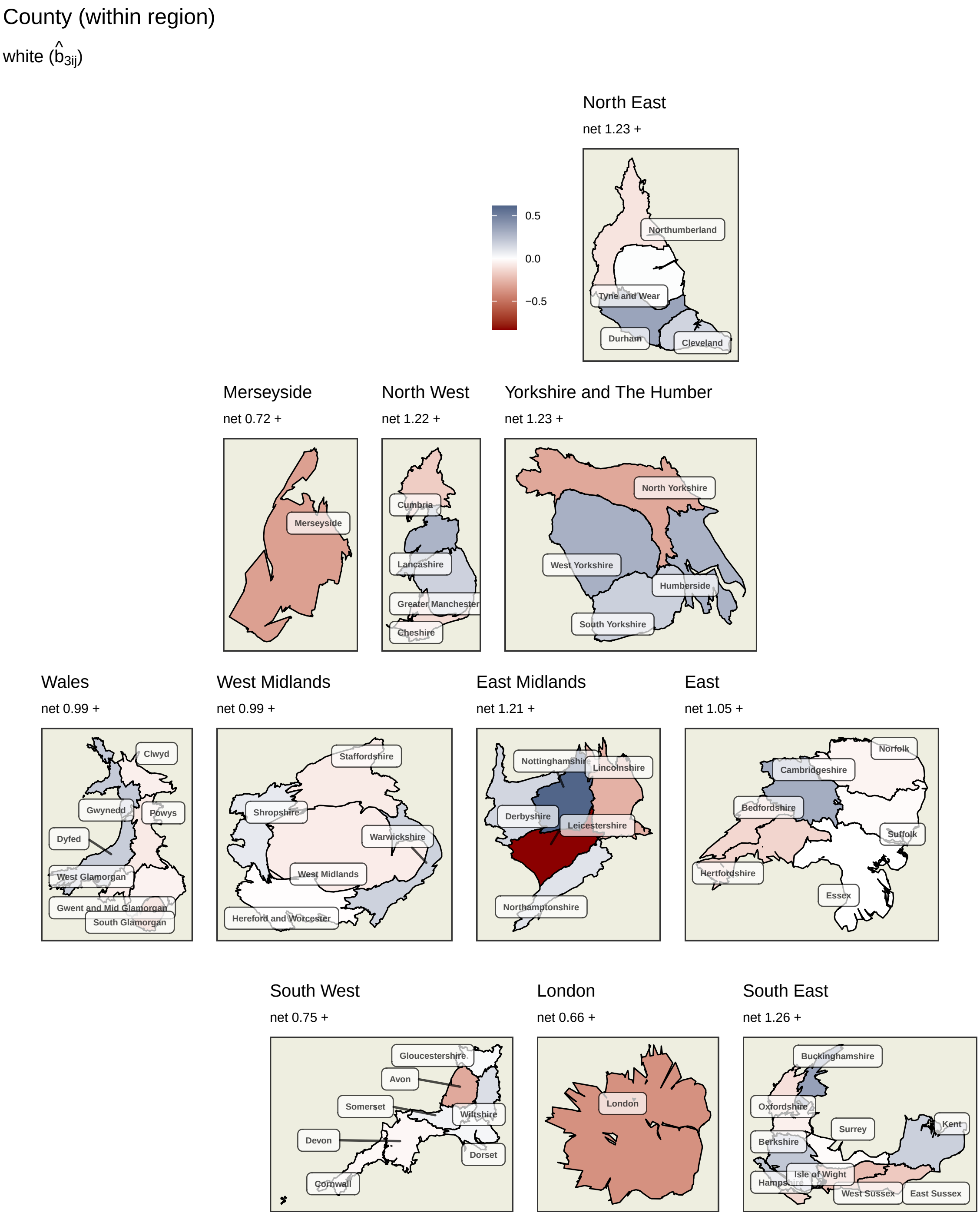


Figure 2.8: County-level 'white' random effects in England and Wales.

Particularly strong within-county variation can be observed in the East Midlands.

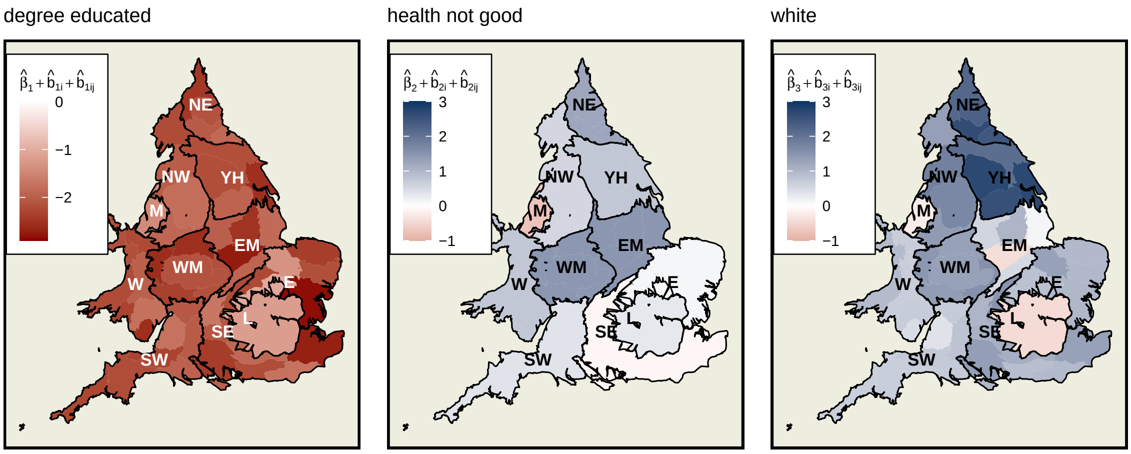

degree educated
health not good
white
$\hat{\beta}_1+\hat{b}_{1i}+\hat{b}_{1ij}$
0
−1
−2
$\hat{\beta}_2+\hat{b}_{2i}+\hat{b}_{2ij}$
3
2
1
0
−1
$\hat{\beta}_3+\hat{b}_{3i}+\hat{b}_{3ij}$
3
2
1
0
−1
NE
NW
YH
M
EM
WM
W
E
L
SE
SW

Figure 2.9: Net hierarchical effects of each variable.

These are the sum, for each coefficient, of its fixed effect and two random effects (at region and county level), showing the spatial heterogeneity accounted for by the hierarchical component of the model. 'Degree educated' is negatively associated with swing to the Conservatives across England and Wales, albeit to different extents. 'Health not good' and 'white' show not only different magnitudes but also different directions of association with swing in different regions and counties across the study area.

2.5.4 Constituency level

The final component of the model is the set of spatially autoregressive terms which, according to Table 2.4, account for 41.3% of the variance. They can be interpreted as random constituency-specific effects which account for further differences in voting behaviour which are not associated with the chosen census variables, nor with location within a region or county. Instead, they exhibit a pattern consistent with the defined neighbourhood contiguity structure and the resultant expected diffusion of political attitudes across nearby constituencies.

As can be seen in Figure 2.10, the pattern of the outcome does not align with the regions and county boundaries. Instead, we see a large and generally blue central area which is surrounded by a paler white section. This blue area, lying within parts of the East and West Midlands and Yorkshire and the Humberside, displays an increased tendency to swing to the Conservatives in 2019, over and above what the census variables and hierarchy would predict. There are patterns of particularly strong blue within this area which cross directly over regional boundaries (as highlighted by the circled areas of the map).

Conversely, across southern England, through Wales, and into parts of the far-north, the predominant trend is a lower level of swing to the Conservatives than would otherwise

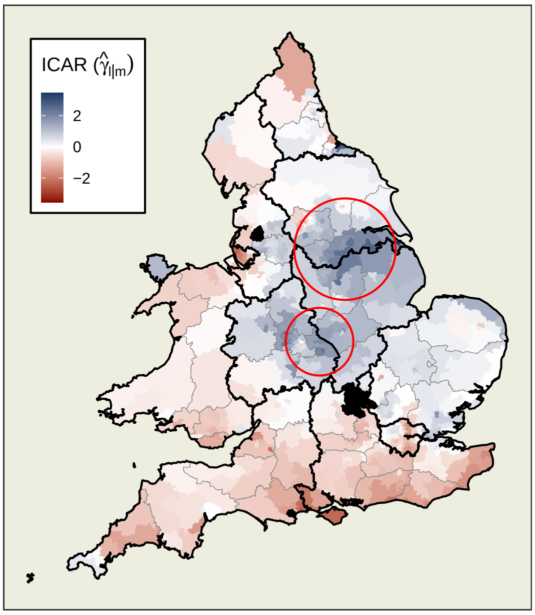


Figure 2.10: Constituency-level ICAR effect.

The spatially autoregressive (in this case, ICAR) component at the lowest level (constituency) of the model. It shows a blue area of increased tendency to swing to the Conservative party, surrounded by a paler band, and red areas to the south, west and parts of the north where the tendency is to swing to Labour, controlling for the covariates and hierarchical effects. Areas of clear cross-regional spillover of effects are highlighted with red circles.

be predicted.

While the model does not specify what the driving forces for these divergences above and below the expected levels of swing might be, it does detect their presence in the form of a hypothesised pattern of spatial diffusion. Such an insight could be useful to political scientists. It is, for example, consistent with how a set of shared attitudes or political culture in one part of the country would change gradually across neighbours. Similarly, support for a party due to local policy proposals, the benefits or drawbacks of which would not be restricted to that location alone, could be expected to show similar autoregressive patterns.

2.5.5 Alternative autoregressive components

The model presented above features a multilevel structure with an autocorrelated random effect at the lowest level for each constituency. This framework, however, also allows for more complex structures than this. For example, in addition to the random intercepts and slopes provided for in the hierarchical component, we have the option of using either

Table 2.5: Comparison of 2019 models.

Performances of a hierarchical structure with different combinations of spatial autocorrelation processes at lowest constituency level. Models 1 and 2 each incorporate one type of process while Model 3 includes both. Model 1 performs best on all metrics.

Model	Autoregressive Spatial Process(es)	AIC	RMSE	Adj.R^2	loglik
1	constituency component	2336	1.43	0.76	-1015
2	varying coefficients	2373	1.62	0.73	-1086
3	constituency component + varying coefficients	2381	1.64	0.73	-1094

1. a spatially autocorrelated random intercept at constituency level (as we did in this model),
2. spatially autocorrelated random slopes for each covariate in each constituency, or
3. both together.

To decide which of these three options was most suitable for this particular dataset, their performances can be compared. The fitting of such spatial models using the `mgcv` package requires the tuning of a parameter k which is the number of basis functions used to generate the autoregressive smoothing. Lower values of k lead to a smoother result. This is because k represents the number of components from the eigen decomposition of the variance-covariance structure which are to be used. Not all can be used because there are not enough data points for this to be computable. The k value has been optimised for each model such that the Akaike information criterion (AIC) is minimised, striking a balance between goodness of fit and model complexity.

Shown in Table 2.5 are performance metrics for each of these model combinations, named models **1-3**. Of these three potential structures, model **1**, which we have been discussing, has the best performance metrics and was deemed the most suitable structure for modelling this particular dataset. Such a process can be used to find the most suitable structure for any potential dataset.

2.5.6 Spatial diagnostics of model

Finally, unlike the pattern previously observed in Figure 2.4, the residuals from our model, mapped in Figure 2.11, show no evidence of any remaining unaccounted-for spatial processes. The spread of positive and negative residuals across the study area appear random by visual inspection, and a Moran's I test of randomness supports this observation.

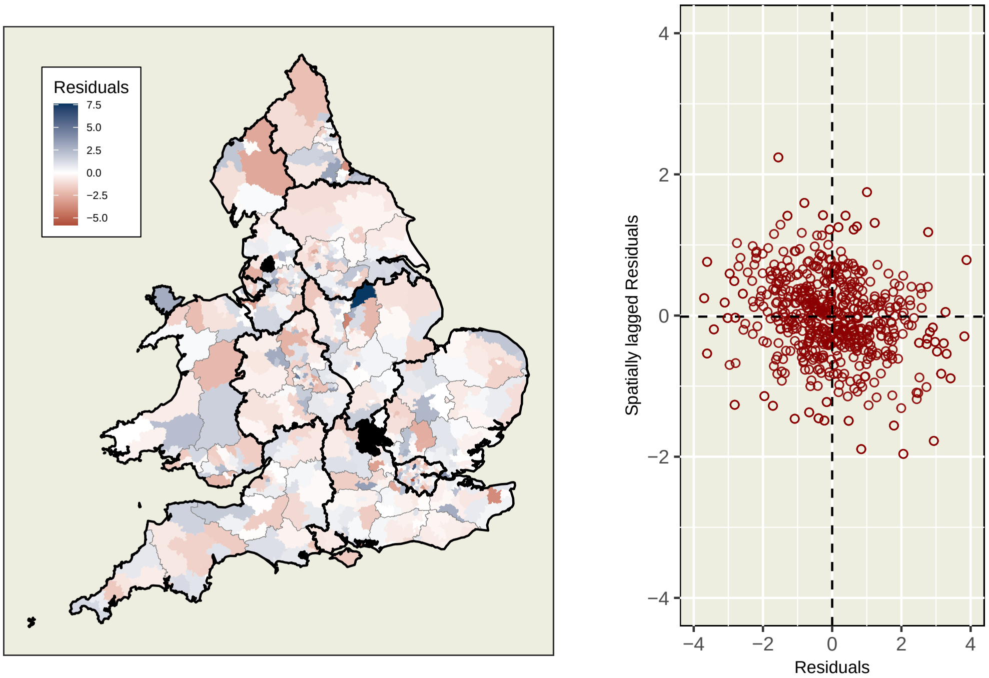


Figure 2.11: Spatial diagnostics of model.

(L) Residuals from model, mapped by location, which appear by visual inspection to be randomly distributed, and (R) dot-plot of residuals of constituencies against their spatially lagged neighbours which shows neither a positive nor negative association between a constituency's residuals and those of its neighbours.

2.6 Discussion and Conclusion

This study presents a framework for analysing spatial data which takes account of the different ways in which these spatial processes may be likely to occur. It simultaneously incorporates a *vertical* set of relationships between nested geographical areas with certain hypothesised shared characteristics, and a *horizontal* covariance among the lowest level units according to contiguity.

In the context of analysing elections, the model structure described above not only succeeds in modelling voting behaviour more accurately than less complex models, but also provides insight into the processes generating the results. It supports the hypothesis of different patterns of association of socio-demographic profile with voting behaviour in different parts of the study area.

It also identifies patterns of places which are more or less likely to swing to the Conservatives for reasons which can not be attributed to the census explanatory variables or hierarchical location, but which are consistent with a *neighbourhood* effect. These patterns are not immediately obvious from the raw data or from models which do not account for location in this way. Insights such as these could be valuable for political theorists.

The framework allows for the testing of different combinations of autocorrelation structure to find which types of spatial processes are most appropriate in a given context. Here, we chose what was essentially an autocorrelated random intercept for each constituency, but we could also fit a set of random slopes if the data suggested that such a structure was more appropriate. The model is fitted in a frequentist restricted maximum likelihood framework using the well-established `mgcv` package.

Another feature is the ability to estimate the relative contribution of each spatial process to overall variation. In this example, it apportions about 27% of variance to hierarchical processes, two-thirds of which occur at the region level, and a further 41% to *spillover* effects at the lowest constituency level. Furthermore, the spatial heterogeneity of the three covariates is shown to operate at different levels within the hierarchy. The variance in association of swing with *degree educated* occurs at the smaller-scale county level, while that of *health not good* and *white* is stronger at the region scale. Such patterns, which again are not otherwise immediately apparent, can contribute to the research of political scientists who are interested in understanding the geography of voting patterns.

3

sfislands: An R package for accommodating islands and disjoint zones in areal spatial modelling

3.1 Abstract

Fitting areal models which use a spatial weights matrix to represent relationships between geographical units can be a cumbersome task, particularly when these units are not well-behaved. The two chief aims of `sfislands` [35] are to simplify the process of creating an appropriate neighbourhood matrix, and to quickly visualise the predictions of subsequent models. The package uses visual aids in the form of easily-generated maps to help this process. This paper demonstrates how `sfislands` could be useful to researchers. It begins by describing the package's functions in the context of a proposed workflow. It then presents two worked examples showing a selection of potential use-cases. These range from earthquakes in Indonesia, to river crossings in London. We aim to show how the `sfislands` package streamlines much of the human workflow involved in creating and examining such models.

3.2 Introduction

A key feature which differentiates spatial statistics is the non-independence of observations and the expectation that neighbouring units will be more similar than non-neighbouring ones [70]. If this is not accounted for, the assumptions of many types of models will be violated. The relationships between all spatial units in a study can be represented numerically in a spatial weights matrix. In order to build this, we must first decide on what constitutes being a neighbour. We might see this as a continuous relationship where degree of neighbourliness is a function of connectivity, which could be represented as some measure of distance. Alternatively it could be a binary situation where each pair of units either are (1) or are not (0) neighbours. This can be based on a condition such as contiguity of some sort, or a distance constraint. It is the job of the modeller to formulate a hypothesis which justifies their choice of neighbourhood structure.

For R users, the `spdep` package [8] has long been popular for the creation of these matrices. More recently, in reference the increasing use of `sf` structures [62], the `sfdep` package [57] has presented generally similar functionality by wrapping `spdep` functions with functions that follow the `sf` naming convention (function names starting with `st_`), as well as a "use a data.frame for everything" attitude.

The most appropriate form of neighbourhood structure will depend on the specific context, which can be both spatial and thematic. Briz-Redón et al. [12] compared different structures in the context of COVID-19 data. They note that Earnest et al. [24] found that distance-based matrices were more appropriate when examining birth defects in Australia, whereas Duncan et al. [23] found that a first-order contiguity structure produced a better fit than others in the context of lip cancer incidence in Scotland.

The most commonly used neighbourhood structure is one based on first-order queen contiguity, where units are considered neighbours if they share at least a vortex of boundary. However, as the name suggests, this will lead to problems when non-contiguous units such as islands or exclaves are present. Less obviously, depending on how the geographic units are described, areas on either sides of rivers may be inappropriately classified as neighbours or not neighbours. Furthermore, the presence of infrastructure such as tunnels, bridges or ferry services might be satisfactory to meet our hypothesis of the required degree of connectivity to be considered neighbours. Again, such information may not be apparent from a basic set of polygons. In order to create what a researcher considers to be an appropriate neighbourhood structure, incorporating all of the domain knowledge that they might have about the system, it should be simple and intuitive to add and remove connections between spatial units. This might mean adding links to account for man-made infrastructure, or cutting links to incorporate natural barriers such as rivers or mountains.

The aim of `sfislands` is to deal with the situations described above in a convenient and open manner. It allows us to set up a structure, quickly map it, and then examine whether or not we are happy with how it represents our hypothesis of relationships between units. The structure can then be edited and the process repeated until we have described a spatial relationship structure with which we are satisfied.

It should be noted that while this package offers convenient tools for the examination, visualisation, addition and removal of neighbourhood linkages between units, such an approach to dealing with disconnected units is not always appropriate and other methodologies are available. These issues are discussed in more depth by Bivand and Portnov [10] and Freni-Sterrantino et al. [26].

The above can be considered as the *pre-functions* of the package. A second category of features, which we refer to as *post-functions*, are for use after the creation of a model.

Having fit a model with mgcv [77] in particular, the process of extracting estimates for certain types of effects can be somewhat awkward. These *post-functions* augment the original dataframe with these estimates and their standard errors in tidy format. They also allow for quick visualisation of the output in map form.

3.2.1 Typical use-cases

In this paper, we will look at two examples to show different use-cases for sfislands. The first example focuses on earthquakes in Indonesia. It shows a scenario where all of the functions are used, from setting up contiguities, to modelling and examining the predictions of the model. The second example looks at London and how, despite an absence of islands, the presence of a river means that some of the pre-functions of sfislands can be useful.

3.3 Why use sfislands?

Below, we outline some of the benefits of the package in the context of a proposed workflow for fitting areal spatial models.

Step 1: *Pre-functions* for setting up neighbourhood structure

1. It addresses an issue commonly seen in online help forums where an inexperienced user wishes to get started with a model but fails at the first hurdle because their neighbourhood structure contains empty records. sfislands will include a contiguity for all units.

2. It gives tools to immediately visualise this structure as a map.

3. These maps are created using ggplot2 [74], which allows users to apply additional styling and themes using ggplot2 syntax.

4. As the nodes can be labelled by index, it makes it very easy to add and remove connections as appropriate with confidence and without reference to the names of these areas.

5. Connections which have been induced by a function from the package but which are not based on geographical contiguity can be accessed to ensure openness in the process.

Table 3.1: Pre-functions: setting up a neighbourhood structure.

Function	Purpose
st_bridges()	create a neighbourhood contiguity structure, with a k-nearest neighbours condition for islands
st_quickmap_nb()	check structure visually on map
st_check_islands()	check the contiguities which have been assigned to islands
st_force_join_nb()	enforce changes by adding connections
st_force_cut_nb()	enforce changes by removing connections

Step 2: Modelling

These neighbourhood structures can be used in modelling packages such as `mgcv`, `brms` [13], `r-inla` [2] and more.

Step 3: *Post-functions* for models

1. It simplifies the process of extracting estimates from models, such as those with random effects and Markov random field structures created using `mgcv`. Compatibility with more packages can be added at a future date.

2. These effects can be quickly visualised as `ggplot2` maps.

3.4 Pre-functions

The first group of functions, shown in Table 3.1, deals with the creation of a neighbourhood structure in the presence of discontiguities. The resultant structure can be quickly mapped to check if it is satisfactory. Connections can be forcibly added or removed by name or index number. By an iterative process of changes and examination of a quickly-generated guide map, a satisfactory structure can be decided upon.

We will now go through each function in more detail using the set of rectangles shown in Figure 3.1 for demonstration purposes. Rectangles 1, 2 and 3 are contiguous while 4 and 5 can be viewed as "islands".

3.4.1 st_bridges()

This function requires at least two arguments: an `sf` dataframe and, from that, the name of one column of unique row identifiers, ideally names, of each spatial unit. It creates a neighbourhood structure where non-island units are joined by first-order queen

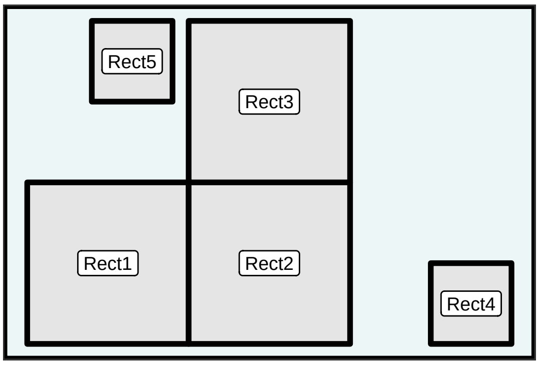


Figure 3.1: Simplified contiguity scenario with five rectangles.

contiguity, while island units are joined to their k-nearest neighbours. The output is a *named* neighbourhood structure in either list or matrix form as desired, which can be either a standalone object or included as an additional column in the original `sf` dataframe. While we have chosen to append the neighbourhood structure to the original data frame in this way by default, the user should be warned that any subsequent row sub-setting (filter) operation on this object will invalidate the list column involved. While it is not necessary in all modelling packages for the neighbourhood list or matrix to be *named*, it is good practice to do so and is mandatory when using, for example, `mgcv`.

One solution when confronted with islands in a dataset is to simply exclude them from the analysis. In the first two examples of using `st_bridges()`, we have chosen to ignore islands with the argument `remove_islands = TRUE` and to return a list and matrix structure respectively by specifying this in the `nb_structure` argument and choosing `add_to_dataframe = FALSE`:

```
# output a named list

st_bridges(rectangles,
           "name",
           remove_islands = TRUE,
           nb_structure = "list",
           add_to_dataframe = FALSE) |>
  head()

#> $Rect1
#> [1] 2 3
#>
#> $Rect2
#> [1] 1 3
```

```
#>
#> $Rect3
#> [1] 1 2
```

```
# output a named matrix

st_bridges(rectangles,
           "name",
           remove_islands = TRUE,
           nb_structure = "matrix",
           add_to_dataframe = FALSE) |>
  head()
```

```
#>       [,1] [,2] [,3]
#> Rect1    0    1    1
#> Rect2    1    0    1
#> Rect3    1    1    0
```

Alternatively, in the following examples, we choose to join islands to their 1 nearest neighbour, which is the default setting, and to return the output as a column called “nb” in the original sf dataframe (add_to_dataframe = "TRUE" is the default setting):

```
# output a named list as a column "nb" in original dataframe

st_bridges(rectangles,
           "name",
           link_islands_k = 1,
           nb_structure = "list") |>
  head()
```

```
#> Simple feature collection with 5 features and 2 fields
#> Geometry type: POLYGON
#> Dimension:     XY
#> Bounding box:  xmin: 0 ymin: 0 xmax: 6 ymax: 4
#> CRS:           NA
#>    name      nb                      geometry
#> 1 Rect1    2, 3 POLYGON ((0 0, 0 2, 2 2, 2 ...
#> 2 Rect2 1, 3, 4 POLYGON ((2 0, 2 2, 4 2, 4 ...
#> 3 Rect3 1, 2, 5 POLYGON ((2 2, 2 4, 4 4, 4 ...
#> 4 Rect4       2 POLYGON ((5 0, 5 1, 6 1, 6 ...
#> 5 Rect5       3 POLYGON ((0.8 3, 0.8 4, 1.8...
# output a named matrix as a column "nb" in original dataframe

st_bridges(rectangles,
           "name",
           link_islands_k = 1,
           nb_structure = "matrix") |>
  head()
```

```
#> Simple feature collection with 5 features and 2 fields
#> Geometry type: POLYGON
#> Dimension:     XY
#> Bounding box:  xmin: 0 ymin: 0 xmax: 6 ymax: 4
#> CRS:           NA
#>    name nb.1 nb.2 nb.3 nb.4 nb.5                       geometry
#> 1 Rect1    0    1    1    0    0 POLYGON ((0 0, 0 2, 2 2, 2 ...
#> 2 Rect2    1    0    1    1    0 POLYGON ((2 0, 2 2, 4 2, 4 ...
#> 3 Rect3    1    1    0    0    1 POLYGON ((2 2, 2 4, 4 4, 4 ...
#> 4 Rect4    0    1    0    0    0 POLYGON ((5 0, 5 1, 6 1, 6 ...
#> 5 Rect5    0    0    1    0    0 POLYGON ((0.8 3, 0.8 4, 1.8...
```

These structures can serve as the input to models in `brms`, `r-inla`, `rstan` [69] or `mgcv`. `brms` requires a matrix structure while `mgcv` models use a list. Rather than having a separate neighbours object, it is included in the original `sf` dataframe as a named list or matrix, in the spirit of the `sfdep` package.

3.4.2 st_quickmap_nb()

It is much more intuitive to examine these structures visually than in matrix or list format. This can be done with the `st_quickmap_nb()` function as shown in Figure 3.2.

```
# default is ’nodes = "point"’

st_bridges(rectangles,
           "name",
           link_islands_k = 1) |>
  st_quickmap_nb()
```

If we wish to make edits, it might be more useful to represent the nodes numerically rather than as points (Figure 3.3).

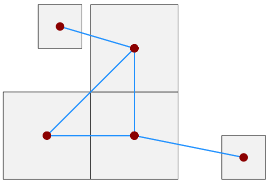


Figure 3.2: Output of `st_quickmap_nb()`.

Queen contiguity and islands connected to nearest neighbour, with nodes shown as points.

```
# with 'nodes = "numeric"'

st_bridges(rectangles,
           "name",
           link_islands_k = 1) |>
  st_quickmap_nb(nodes = "numeric")
```

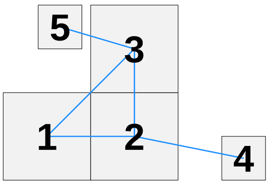


Figure 3.3: Output of `st_quickmap_nb(nodes = "numeric")`.

Queen contiguity and islands connected to nearest neighbour, with nodes shown as numeric indices.

3.4.3 st_check_islands()

This function will show us transparently what connections have been made which are not based on contiguity. It gives both the name and index number of each pair of added connections. In this example, two pairs have been added.

```
# show summary of non-contiguous connections in a dataframe

st_bridges(rectangles,
           "name",
           link_islands_k = 1) |>
  st_check_islands()

#>   island_names island_num nb_num nb_names
#> 1        Rect4          4      2    Rect2
#> 2        Rect5          5      3    Rect3
```

3.4.4 st_force_join_nb()

If we feel that 4 should also be connected to 3, this can be done by forcing a join (Figure 3.4).

```
# add an extra connection using numeric index

st_bridges(rectangles, "name",
           link_islands_k = 1) |>
  st_force_join_nb(3,4) |>
  st_quickmap_nb(nodes = "numeric")
```

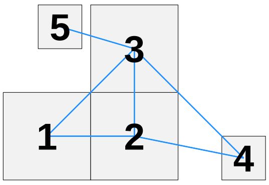


Figure 3.4: Addition of contiguities with st_force_join().

An additional connection beyond those imposed by st_bridges() is added between 3 and 4 using st_force_join(3,4).

3.4.5 st_force_cut_nb()

And perhaps there is a wide river between rectangles 1 and 2 which justifies removing the connection. We will edit it this time using names (Figure 3.5).

```
# remove an existing connection using unit name, not index

st_bridges(rectangles, "name",
           link_islands_k = 1) |>
  st_force_join_nb(3,4) |>
  st_force_cut_nb("Rect1","Rect2") |>
  st_quickmap_nb(nodes = "numeric")
```

Having decided upon an appropriate neighbourhood structure, the next step is to use this in the context of a model. The use of such structures is particularly associated

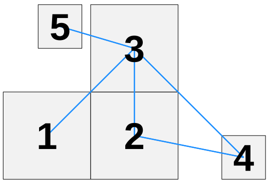


Figure 3.5: Removal of contiguities with `st_force_cut()`.

A connection previously imposed by `st_bridges()` is removed between 1 and 2 using `st_force_cut(1,2)`.

Table 3.2: Post-functions: tidy estimates from mgcv.

Function	**Purpose**
st_augment()	augment the original dataframe with model predictions
st_quickmap_preds()	generate quick maps of these predictions

with CAR (conditional autoregressive) or ICAR-type (intrinsic conditional autoregressive) models [5]. These are often implemented in a Bayesian framework using `brms`, `r-inla` or `rstan`. For example, the `brms` ICAR structure requires the neighbourhood relationships to be in matrix form. The pre-functions will output the neighbourhood structure in the desired format for use in any of these frameworks. A convenient frequentist alternative is to use the `mgcv` package which requires a named list of neighbours. It has the functionality to create such models using `bs="mrf"`. It also has the ability to combine these with a hierarchical structure using `bs="re"`. While the outputs from the Bayesian structures mentioned above can be extracted in the same way as any other component of the model, it can be somewhat awkward to get the estimates from `mgcv` models. `sfislands` has two post-functions to conveniently extract and visualise these.

3.5 Post-functions

Table 3.2 shows the second set of functions in the package and their purpose.

3.5.1 st_augment()

This function augments the original dataframe with the estimated means and standard errors of the spatially varying predictions from a fitted `mgcv` model in a similar manner to how the `broom` [67] package operates. The `geometry` column, as per convention, remains as the last column of the augmented dataframe, while the predictions are positioned immediately before it. [a] The spatially varying predictions which `st_augment()` extracts

[a] In a similar way, `st_augment()` can also be used to append the random effects from `lme4` and `nlme` models to an `sf` dataframe, which can then be easily mapped using `st_quickmap_preds()`. Compatability

from an mgcv model are

- random effects (which are called in mgcv with bs='re'), and
- ICAR components (bs='mrf').

Consider the model structure described in the code below using mgcv syntax. In this model y is the dependent variable which is being estimated with a fixed intercept, a fixed slope for some covariate, a set of random intercepts and slopes for the covariate at a *region* level, and a set of ICAR varying intercepts and slopes at a lower *sub-region* level.

```
# creating an mgcv model

mgcv::gam(
  y ~ covariate +               # fixed intercept and effect for covariate
    s(region, bs = "re") +              # random intercept at level region
    s(region, covariate, bs = "re") +      # random slopes at level region
    s(sub-region,
      bs = 'mrf',
      xt = list(nb = data$nb),
      k = k) +               # ICAR varying intercept at level sub-region
    s(sub-region, by = covariate,
      bs = 'mrf',
      xt = list(nb = data$nb),
      k = k),       # ICAR varying slope for covariate at level sub-region
  data = data,
  method = "REML")
```

When labelling the new prediction columns which are augmented to the original dataframe from such a model, st_augment() follows the formula syntax of the lme4 package, where the pipe symbol (|) indicates "*grouped by*". Table 3.3 shows how the augmented columns in this scenario would be named. Each column name begins with either random.effect. or mrf.smooth. as appropriate. An additional column is also added for the standard error of each prediction, as calculated by mgcv. These columns are named as above but with se. prepended (e.g. se.random.effect.region).

3.5.2 st_quickmap_preds()

These estimates can then be quickly mapped. As it is possible to include more than one spatially varying component, the output of this function is a list of plots. They

with models created using different packages can be introduced in the future.

Table 3.3: The naming procedure for augmented columns from different `mgcv` structures.

`mgcv` **Syntax**	**Column Name**
s(region, bs = 're')	random.effect.region
s(region, covariate, bs = 're')	random.effect.covariate|region
s(sub-region, bs = 'mrf', xt = list(nb = data$nb))	mrf.smooth.sub-region
s(sub-region, by = covariate, bs = 'mrf', xt = list(nb = data$nb))	mrf.smooth.covariate|sub-region

can be viewed individually by indexing, or all at once using, for example, the `plotlist` argument from the `ggarrange()` function which is part of the `ggpubr` [47] package. We will see this function in practice in the following example. The maps which it generates are automatically titled and subtitled according to the type of effect. For example, the map showing predictions for `random.effect.region` will have "*region*" as its title and "*random.effect*" as its subtitle.

3.6 Indonesia (example 1)

Modelling earthquakes in Indonesia serves as a good example to demonstrate this package. Firstly, Indonesia is composed of many islands. Secondly, earthquake activity is known to be associated with the presence of faults which exist below sea level and thus do not respect land boundaries. Therefore it is reasonable to expect similar behaviour in nearby provinces regardless of whether or not they are contiguous. We aim to model the incidence, or count per unit area, of earthquake activity by province across Indonesia, controlling for proximity to faults.

3.6.1 Data

The data for this section have been downloaded from the National Earthquake Information Center, USGS earthquake catalogue. The datasets with accompanying explanations are available at `https://github.com/horankev/quake_data`. They capture all recorded earthquakes in and close to Indonesia from the beginning of January 1985 to the end of December 2023. Figure 3.6 shows a map of Indonesia, divided into 33 provinces, with other neighbouring or bordering countries filled in grey. The many local faults which lie within 300km of the shore are shown in yellow with green outlines.

To get an interpretable measure of the concentration of faults in any area, these faults are transformed from linestrings to polygons by setting a buffer of 10km around them, which explains their green outline. Now both our faults and the sizes of provinces are in units of kilometres squared. This means we can generate a unitless metric of what

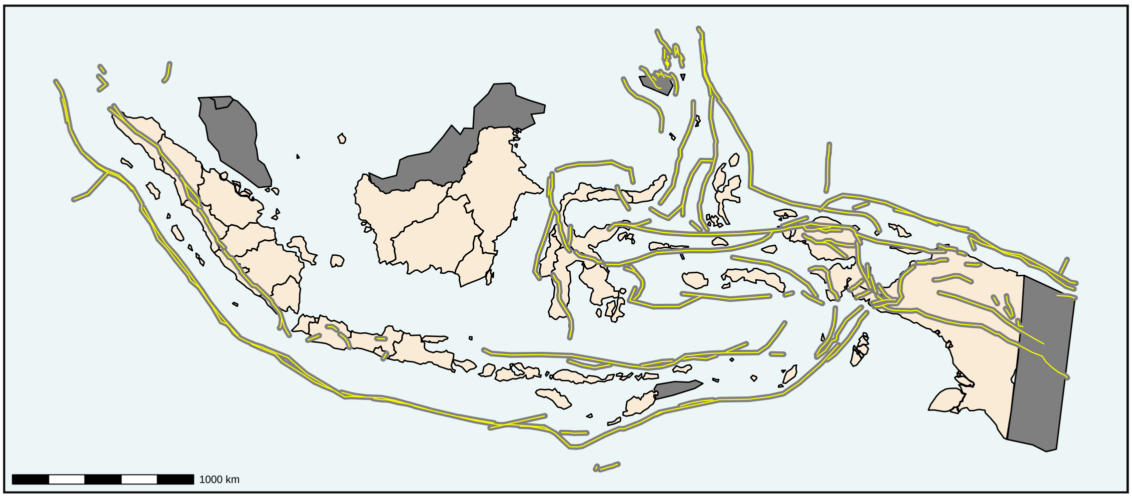


Figure 3.6: Indonesia faults surrounded by a 10 kilometre buffer.

proportion of any administrative unit is covered by these buffered faults. This measure across provinces is shown in Figure 3.7.

Earthquake incidence per province has been calculated as the total number of earthquakes with an epicentre within that province per unit area. We have restricted counts to earthquakes >5.5 on the moment magnitude scale, which is the point at which they are often labelled as potentially damaging.

The occurrences of these earthquakes are shown in Figure 3.8, their total per province in Figure 3.9, and finally, their incidence or count per square kilometre can be seen in Figure 3.10.

3.6.2 Model

As this is count data, we will model it as a Poisson distribution with λ as the mean count per province. For $i = 1, ..., n$ provinces, the dependent variable in this model is

$$y_i = \text{earthquake count}_i$$

while the explanatory variable is

$$x_i = \text{fault concentration}_i = \frac{\text{area of buffered faults in province}_i}{\text{province area}_i}.$$

Firstly, when excluding the incidence and just modelling counts, where

$$y_i = \text{earthquake count in province}_i,$$

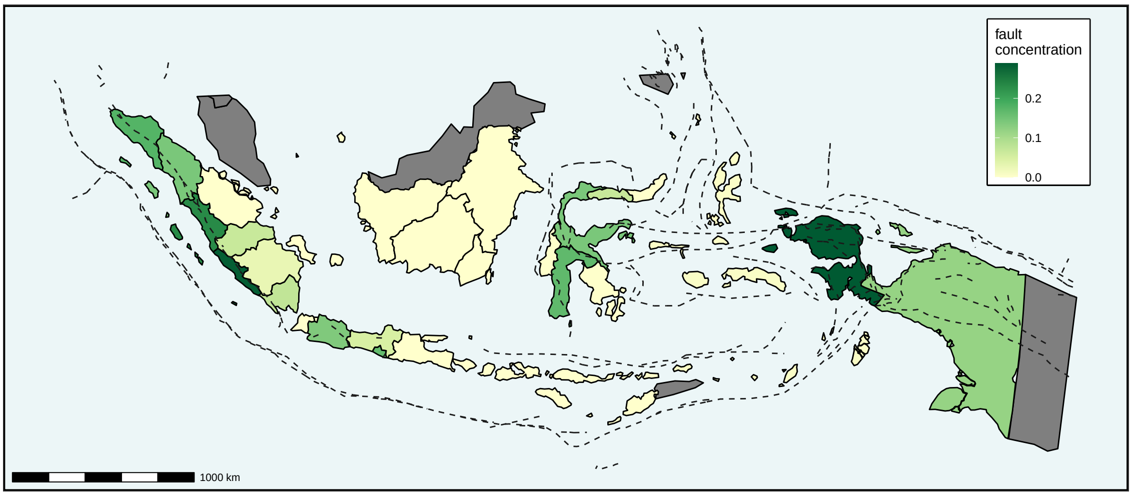


Figure 3.7: Concentration of faults by province of Indonesia.

Square kilometre of buffered fault per square kilometre of province area.

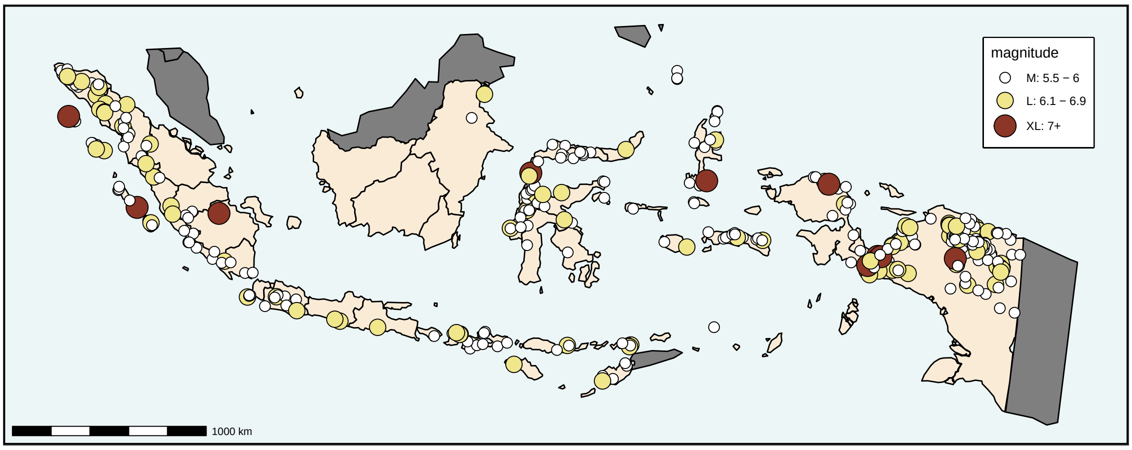


Figure 3.8: Earthquakes in Indonesia of magnitude > 5.5, 1985-2023.

Significant earthquakes categorised by magnitude as medium, large or extra-large.

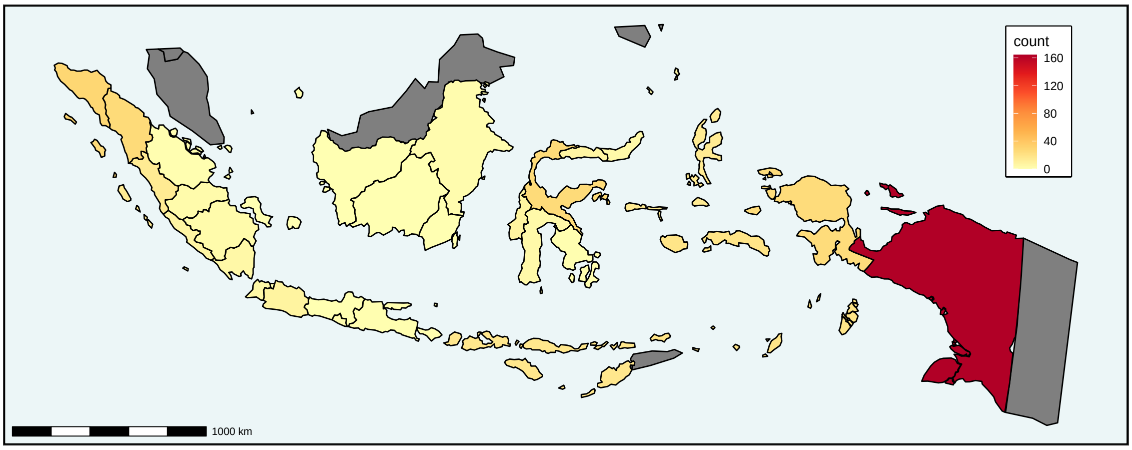


Figure 3.9: Earthquake count by province.

Earthquake count in Indonesia, 1985-2023, mag > 5.5: count by province.

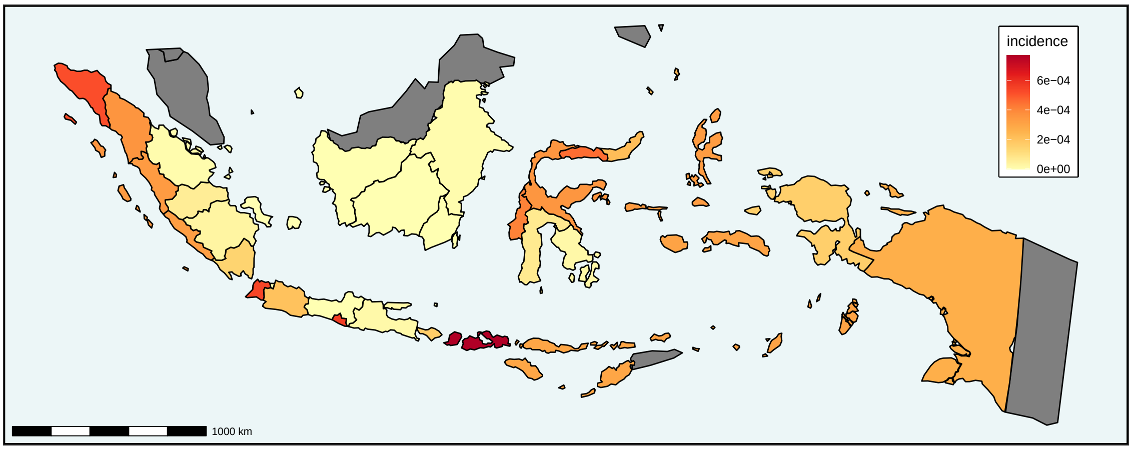


Figure 3.10: Earthquake incidence by province.

Earthquake incidence in Indonesia, 1985-2023, mag > 5.5: count per square kilometre by province.

the Poisson model is of the following form:

$$y_i|\lambda_i \sim \text{Pois}(\lambda_i)$$

with

$$E(y_i|\lambda_i) = \lambda_i.$$

We model

$$log(\lambda_i) = \beta_0 + \beta_1 x_i + \gamma_i.$$

where γ_i is a term with a correlation structure reflecting a province's location relative to other provinces.

We can describe these relationships by setting up a neighbourhood structure based on queen contiguity where a pair of provinces are considered neighbours if they share at least one point of boundary. This can be modelled as a Markov random field to generate an ICAR model with a spatially varying term. Each of these terms will be correlated with the others according to the neighbourhood structure we have defined.

The Markov random field here follows a multivariate Gaussian distribution. γ_i is a vector of province effects having a distribution with mean $\mathbf{0}$ and precision $\mathbf{P}$ where

$[\mathbf{p}]_{ij} = v_i$ if $i = j$ and v_i is the number of adjacent provinces to province i,

$[\mathbf{p}]_{ij} = -1$ if provinces i and j are adjacent, and

$[\mathbf{p}]_{ij} = 0$ otherwise.

A further constraint that $\Sigma_j \gamma_j = 0$ is applied so that the distribution is identifiable.

We now include an offset term (here, area) because we are more interested in modelling the incidence than in the actual count, such that

$$log(\frac{\lambda_i}{\text{area}_i}) = \beta_0 + \beta_1 x_i + \gamma_i$$

which is equivalent to

$$log(\lambda_i) = \beta_0 + \beta_1 x_i + \gamma_i + log(\text{area}_i).$$

We are still modelling $log(\lambda)$ rather than the incidence, but we are adding an offset to adjust for differing areas. Modelling $log(\lambda)$ and adding an offset is equivalent to modelling incidence, and coefficients can be interpreted that way.

When interpreting the estimated coefficients of the model, it can be useful to look at it in the following form:

$$\lambda_i = e^{\beta_0 + \beta_1 x_i + \gamma_i} \text{area}_i.$$

Expressing the model in this way clarifies that the β coefficients describe multiplicative effects on the rate of occurrence, independent of the size of the area.

Having described the type of model we wish to implement, we now show how sfislands can be used to streamline the process.

3.6.3 Pre-functions

Such models, however, can not incorporate locations which have no neighbours. In the case of Indonesia, this is quite problematic. It is composed of many islands. The estimated count of islands according to Andréfouët et al. [1] is 13,558. While it is not unusual for a country to have a number of often small offshore islands, Indonesia is entirely composed of (at least portions of) an archipelago of islands, so many of these islands or groups of islands are individual provinces in their own right. We might like to hypothesise that just because a province is a disconnected island, this should not mean that it is independent of other nearby provinces in terms of earthquake incidence. A standard first-order queen contiguity structure would mean the exclusion of disconnected units entirely from the model. An alternative strategy of assigning neighbour status based on a distance metric would overcome this, but the threshold size of distance necessary for such a structure might be inappropriately large for the non-island provinces. Many extra unwanted contiguities could be added when only those related to disconnected units were desired. We would like to use a compromise between these two strategies.

In this case, we use st_bridges() for setting up the queen contiguity structure as usual, but with the additional stipulation that unconnected units (provinces which are islands or collections of islands) are considered neighbours to their k nearest provinces. For this example, we have set the value of k to 2. The resulting neighbourhood structure is shown in Figure 3.11. Note how is can be styled with a combination of internal arguments (size, colour, fill etc.) and additional ggplot2 layers.

```
# join islands to k=2 nearest neighbours
# various arguments exist for altering colours and sizes
# additional ggplot themes and layers can be added

st_bridges(provinces_df, "province", link_islands_k = 2) |>
  st_quickmap_nb(fillcol = "antiquewhite1",
                 bordercol = "black", bordersize = 0.5,
                 linkcol = "darkblue", linksize = 0.8,
                 pointcol = "red", pointsize = 2) +
  theme(panel.background = element_rect(fill = "#ECF6F7",
                                        colour = "black",
                                        linewidth=1.5),
```

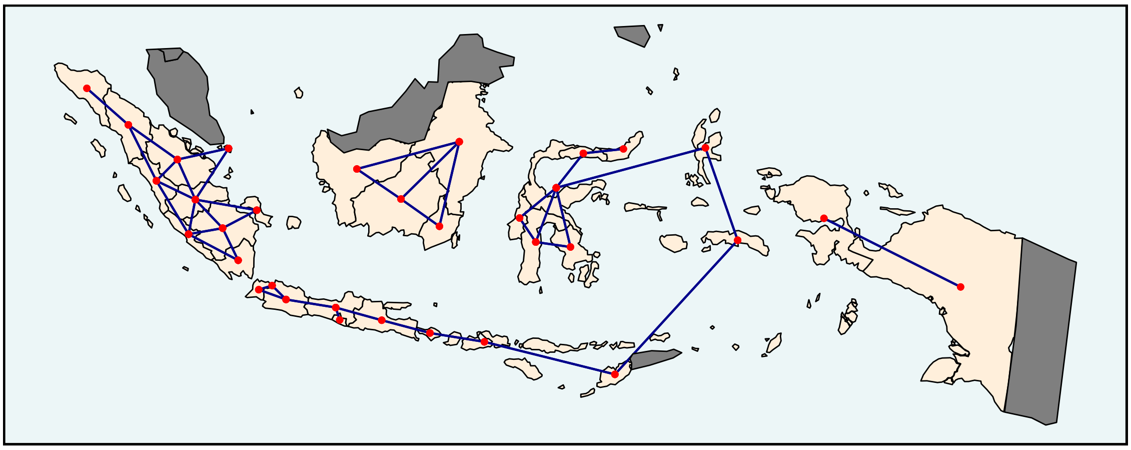


Figure 3.11: Basic neighbourhood strucutre for Indonesian provinces.

Neighbourhood structure for Indonesian provinces created by `st_bridges()` with `k=2`.

```
      axis.text = element_blank()) +
  geom_sf(data=nearby_countries_df,
          fill="gray50", linewidth=0.5, colour="black")
```

This neighbourhood structure now has no unconnected provinces so it is suitable for use in an ICAR model. However, if we are not entirely happy with this structure because of some domain knowledge about the inter-relationships between certain island provinces, we might wish to

- add some additional contiguities using `st_force_join_nb()`
- and remove one using `st_force_cut_nb()`.

To cater for the possibility that a modeller might not be familiar with the names of the various geographic units but still wishes to enforce alterations to their relationships, we can look at a map (Figure 3.12) where the nodes are shown by index number instead of as points (using the argument `nodes='numeric'`). This makes it easy to cut and join neighbour connectivities as desired. Furthermore, there is an option to show concave hulls drawn around each unit (using `concavehull = TRUE`). This is also shown in Figure 3.12. These shapes are not used in the assignment of contiguities but it can be useful to see them in a situation such as Indonesia where many individual provinces are actually multipolygons of more than one island. Without them, it is not clear whether an island is a province in its own right, or which group of islands together form one province.

```
# with 'concavehull = TRUE' and 'nodes = "numeric"'
```

```
st_bridges(provinces_df, "province", link_islands_k = 2) |>
  st_quickmap_nb(fillcol = "antiquewhite1",
                 bordercol = "black", bordersize = 0.5,
                 linkcol = "tomato", linksize = 0.5,
                 nodes = "numeric",
                 numericcol = "black", numericsize = 6,
                 concavehull = TRUE,
                 hullcol = "darkgreen", hullsize = 0.2) +
  theme(panel.background = element_rect(fill = "#ECF6F7",
                                         colour = "black",
                                         linewidth=1.5),
        axis.text = element_blank())
```

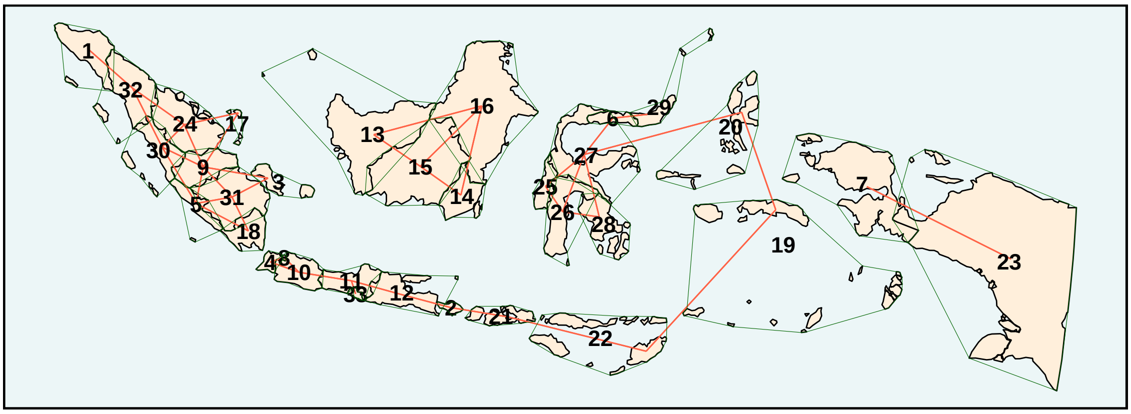


Figure 3.12: Annotated neighbourhood strucutre for Indonesian provinces.

Neighbourhood structure for Indonesian provinces viewed with `st_quickmap_nb()`, using the arguments `nodes = 'numeric'` and `concavehull = TRUE`.

Having enforced some adjustments to the neighbourhood structure, outlined in the code below, the new structure can be seen in Figure 3.13. Edge effects have also been mitigated by imposing additional connections on the two extreme provinces (1 and 23), which would otherwise have only one neighbour, so that they now also include their two next closest neighbours.

```
# a series of forced joins and cuts by index number

joins_df <- tribble(
  ~x, ~y,
  1, 24,
  1, 30,
  3, 13,
  13, 17,
  14, 25,
  20, 29,
  19, 23,
  16, 27,
  22, 23,
  7, 19,
  7, 20,
  19, 28,
  4, 18,
  21, 26,
  22, 28
)

st_bridges(provinces_df, "province", link_islands_k = 2) |>
  st_force_join_nb(xy_df = joins_df) |>
  st_force_cut_nb(19,22) |>
  st_quickmap_nb(fillcol = "antiquewhite1",
                 bordercol = "black", bordersize = 0.5,
                 linkcol = "darkblue", linksize = 0.8,
                 pointcol = "red", pointsize = 2) +
  theme(panel.background = element_rect(fill = "#ECF6F7",
                                        colour = "black",
                                        linewidth=1.5),
        axis.text = element_blank()) +
  geom_sf(data=nearby_countries_df,
          fill="gray50", linewidth=0.5, colour="black") +
  annotation_scale()
```

3.6.4 mgcv model

We now create the ICAR model using, in this case, the mgcv package. We will be able to use the output of st_bridges, which we have named prep_data, as both the data source for the model and the neighbourhood structure (by specifying the column nb which contains the neighbourhood list).

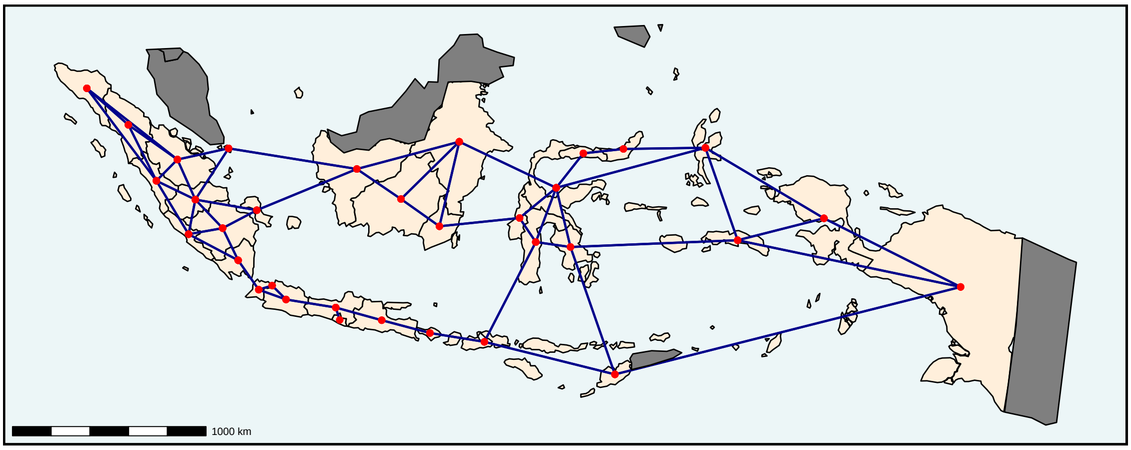


Figure 3.13: Modified neighbourhood strucutre for Indonesian provinces.

Neighbourhood structure for Indonesian provinces after alterations using `st_force_join()` and `st_force_cut()`. As many connections are being enforced at once, these have been fed to the function as a data frame rather than using consecutive function calls as before.

```
mod_pois_mrf <- gam(damaging_quakes_total ~
                      fault_concentration +
                      s(province,
                        bs='mrf', xt=list(nb=prep_data$nb), k=24) +
                      offset(log(area_province)),
                    data=prep_data, method="REML",family = "poisson")
```

We can see from the summary below that the adjusted R-squared is **0.983** and deviance explained is **93.3%**. The coefficient for `fault_concentration` confirms an expected positive mean global association between earthquake and fault incidence.

```
#>
#> Family: poisson
#> Link function: log
#>
#> Formula:
#> damaging_quakes_total ~ fault_concentration + s(province, bs = "mrf",
#>     xt = list(nb = prep_data$nb), k = 24) + offset(log(area_province))
#>
#> Parametric coefficients:
#>                       Estimate Std. Error z value Pr(>|z|)
#> (Intercept)            -9.5648     0.1744 -54.845  < 2e-16 ***
#> fault_concentration   5.9971     1.9245   3.116  0.00183 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Approximate significance of smooth terms:
#>               edf Ref.df Chi.sq p-value
#> s(province) 19.19     23  166.6  <2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> R-sq.(adj) =  0.983   Deviance explained = 93.3%
#> -REML = 104.81  Scale est. = 1         n = 33
```

Returning to the initial question, what is the additional risk level of earthquakes in a province, having controlled for the concentration of faults? This can be seen as a measure of the activity level of faults locally and it is spatially smoothed by the autoregressive process. It is represented in the model summary by the component s(province). However, the extraction of individual predictions for this component for each province from the mgcv model requires a number of steps. We now demonstrate how these are streamlined into a single function by the sfislands package.

3.6.5 Post-functions

The function st_augment() allow us to add the spatially varying predictions from the model as new columns to the original dataframe in a process similar to that of the broom package. For instance, we see from the output of the following code chunk that the original dataframe is now augmented with columns called mrf.smooth.province and se.mrf.smooth.province which show the predictions for the γ_i component and their standard errors. Note that this is how we would expect them to be named, based on the previous discussion surrounding Table 3.3. They are positioned immediately before the final geometry column of the sf dataframe, and after the neighbours list column, nb.

```
# column names of augmented dataframe

mod_pois_mrf |>
  st_augment(prep_data) |>
  names() |>
  dput()

#> c("province", "province_id", "S", "M", "L", "XL", "quake_total",
#> "quake_density", "damaging_quakes_total", "damaging_quakes_density",
#> "area_fault_within", "area_province", "fault_concentration",
#> "nb", "mrf.smooth.province", "se.mrf.smooth.province", "geometry"
#> )
```

This output can now be piped into the `st_quickmap_preds()` function to get a quick visualisation of these estimates for γ_i on a map, as shown in Figure 3.14. Again, note that the title and subtitle of the image are as previously discussed.

```
# st_quickmap_preds() outputs a list of ggplots

plot_mrf <- mod_pois_mrf |>
  st_augment(prep_data) |>
  st_quickmap_preds(scale_low = "darkgreen",
                    scale_mid = "ivory",
                    scale_high = "darkred",
                    scale_midpoint = 0)

# in this case, there is only one plot in the list
# so we call it by index
# it is then supplemented with additional ggplot functions

plot_mrf[[1]] +
  coord_sf(datum=NA, default = TRUE) +
  theme(panel.background = element_rect(fill = "#ECF6F7", colour = "black",
                                        linewidth=1.5),
        axis.text = element_blank()) +
  geom_sf(data=provinces_df, fill=NA, colour="black", linewidth=0.5) +
  geom_sf(data=nearby_countries_df, fill="gray50", colour="black",
          linewidth=0.5) +
  labs(fill="relative\nincidence") +
  annotation_scale() +
  coord_sf(datum=NA, default = TRUE) +
  theme(legend.position = "inside",
        legend.position.inside = c(0.92,0.77),
        legend.box.background = element_rect(colour = "black",
                                             linewidth = 1),
        legend.title = element_text())
```

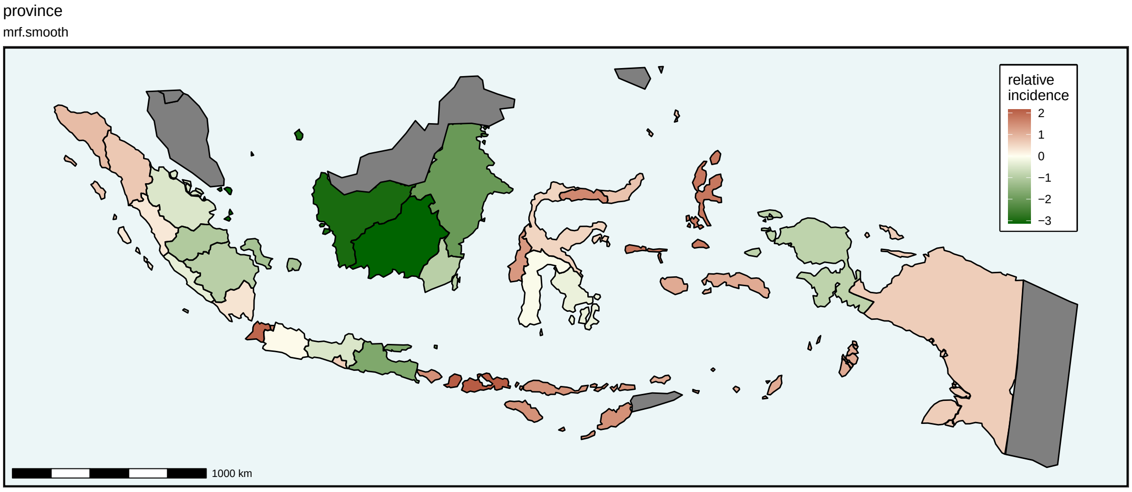


Figure 3.14: ICAR component from Indonesia model.

Estimates of γ_i shown as a map using `st_quickmap_preds()`.

If we wish to apply the inverse link function (the exponential function in the case of this Poisson model) to map these values to a more interpretable scale, this will not be generated by the function `st_quickmap_preds()`. Instead, we must use the augmented dataframe which is produced by `st_augment()` and create the appropriate extra column with the usual `tidyverse` [75] `mutate()` function. This allows us to produce the map in Figure 3.15. As these coefficients are multiplicatively related to the earthquake incidence, values below 1 imply an earthquake incidence which is lower than expected.

The provinces with the 3 most elevated incidences are labelled in red. We can see that, controlling for the effects of proximity to faults, the province of Nusa Tenggara Barat has 8.7 times the expected incidence, or number of major earthquakes per square kilometre. The two lowest-scoring provinces, labelled in green, have essentially no incidence of earthquake epicentres within their boundaries, controlling for what their proximity to faults alone would suggest.

3.6.6 Workflow summary

In this example, we have gone through a number of stages carefully, making changes to contiguities that we deemed appropriate as we went. However, in practice, at least in a first iteration, it might not be necessary to go through all of these steps. A rough and ready model, complete with spatially varying coefficients and visual output, can be generated with `sfislands` using nothing more than three or four lines of code, such as the following:

```
# workflow:
```

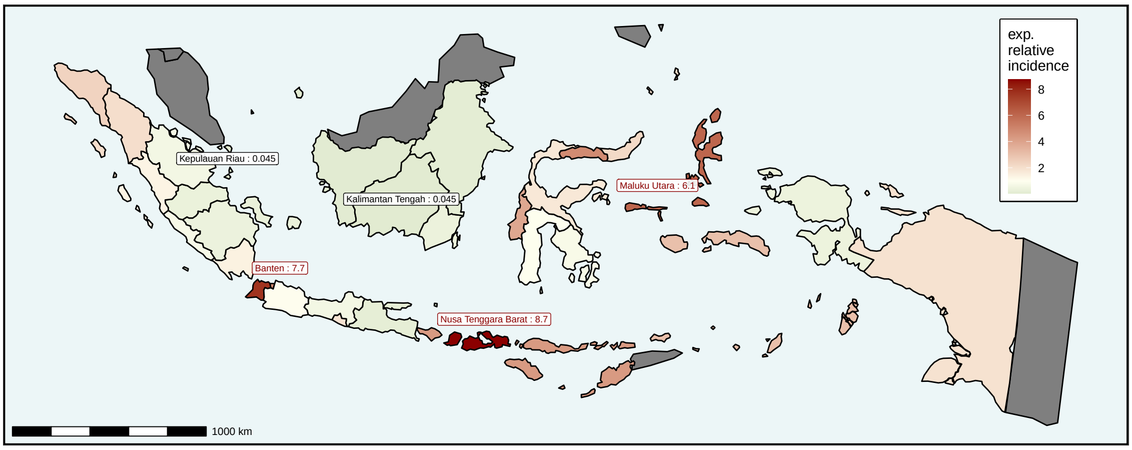


Figure 3.15: Exponential of ICAR component from Indonesia model.

Map showing estimates of $\exp(\gamma_i)$. This is produced by adding an additional column to the dataframe produced by `st_augment()`.

```
# 1. set up neighbourhood structure

prep_data <- st_bridges(provinces_df, "province")

# 2. define model

mod <- gam(quake_mlxl_total ˜
                  fault_concentration +
                  s(province, bs=’mrf’,
                    xt=list(nb=prep_data$nb), k=22) +
                  offset(log(area_province)),
        data=prep_data, method="REML",family = "poisson")

# 3. augment tidy estimates

tidy_ests <- st_augment(mod, prep_data)

# 4. visualise them

st_quickmap_preds(tidy_ests)
```

3.7 London (example 2)

The next example looks only at using the *pre-functions* of `sfislands`, but in a situation where the presence of actual *islands* is not the problem we seek to deal with. Consider the wards of London (sourced from the Greater London Authority's London Datastore) and available at `https://github.com/horankev/london_liverpool_data`. In Figure 3.16 the `st_bridges()` function is applied to them to construct a queen contiguity neighbourhood structure. As can be seen from the `st_check_islands()` function, this collection of London wards contains no isolated units.

```
st_bridges(london, "GSS_CODE") |>
  st_check_islands()

#> No disconnected units were found in original data

#> [1] 0
```

The `st_quickmap_nb()` function gives an immediate visual representation of the structure. [b] Because this map is created using `ggplot2`, it can be easily supplemented by adding a layer showing the course of the River Thames which is also visible in Figure 3.16.

```
# same as sfdep:st_contiguity() as there are no islands
# an extra layer for the River Thames

st_bridges(london, "GSS_CODE") |>
  st_quickmap_nb() +
  geom_sf(data=thames, colour="blue", linewidth=1.5) +
  theme(panel.background = element_rect(fill = "#F6F3E9",
                                        colour = "black",
                                        linewidth=1.5))
```

When a study area has a river running through it, problems can arise with constructing appropriate neighbourhood structures. Depending on how the geometries are defined, the presence of a river can cause problems in two ways. In one situation, the river could be expressed as a polygon in its own right meaning that, using the condition of queen contiguity, it severs any potential contiguity between units on either side of its banks. In this situation, no spatial units will be neighbours with the units directly across the river

[b] `st_quickmap_nb()` can also be used to visualise any contiguity structure created by `spdep` or `sfdep` as long as that structure is included in an `sf` dataframe as a column named `nb`.

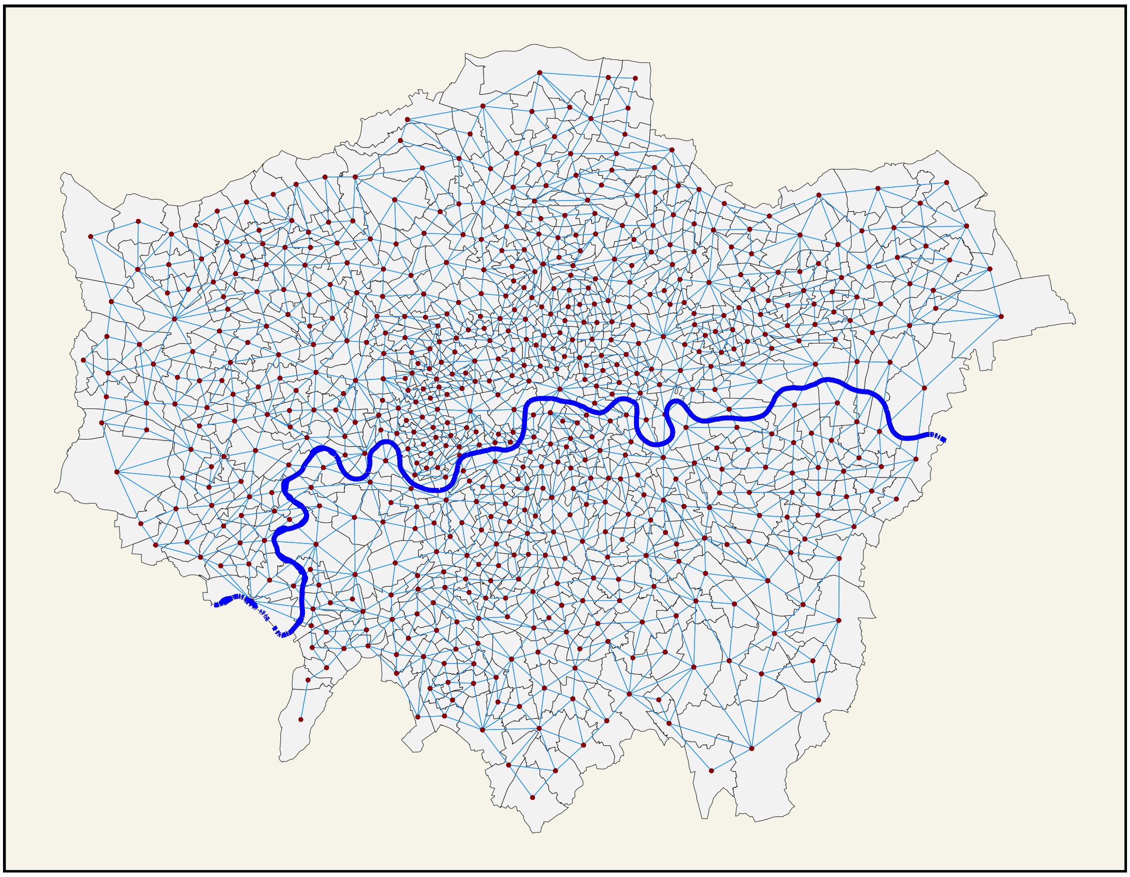


Figure 3.16: Neighbourhood structure of wards of Greater London by queen contiguity.

from them. At the other extreme, if the river is not included as a geometry (as is the case here) all units on opposing banks are automatically considered neighbours.

Depending on the presence of river crossings, two areas which are physically quite close but on opposing banks might be very distinct. If there is no means of crossing the river within a reasonable distance, somebody living on the banks of a river might be more likely to go about their life primarily on their side of the river, despite the short distance as the crow flies of facilities on the other side. This could be relevant in terms of, say, modelling of house prices where we might want to incorporate issues such as local amenities into a neighbourhood structure.

`sfislands` provides convenient functions for this sort of situation. Let us start by restricting our wards of interest to just those which are on either side of the River Thames. Figure 3.17 shows the resultant contiguities when the river is ignored.

```
# which wards are alongside the river

riverside <- thames |> st_intersects(london) |> unlist() |> unique()
```

```
# only map these wards

st_bridges(london[riverside,],"NAME") |>
  st_quickmap_nb(linksize = 0.5) +
  geom_sf(data=thames, colour="blue", linewidth=1.5) +
  annotation_scale(location="br") +
  coord_sf(datum=NA) +
  theme(panel.background = element_rect(fill = "#F6F3E9",
                                        colour = "black",
                                        linewidth=1.5))
```

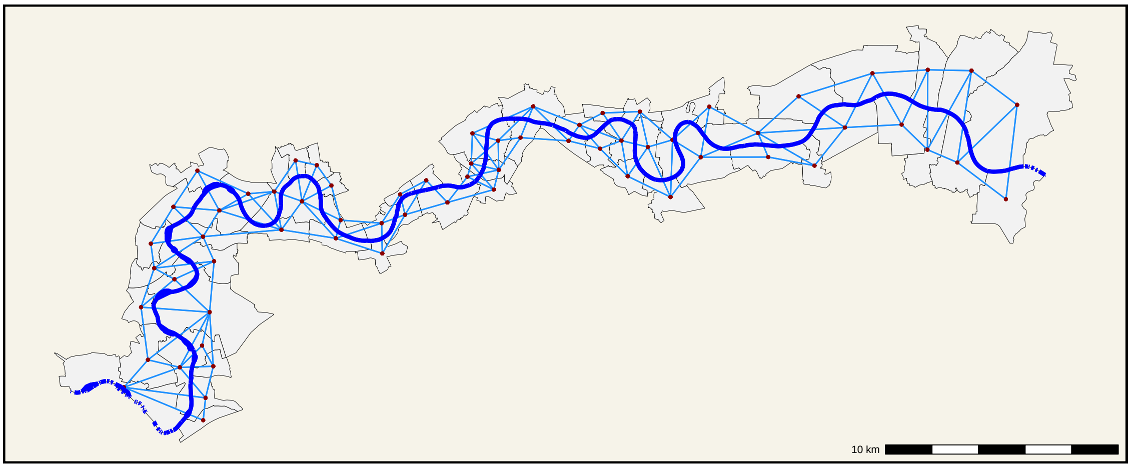


Figure 3.17: Neighbourhood structure of riverside wards of Greater London by queen contiguity.

> In this scenario, the presence of the River Thames has no effect on contiguity determination.

In order to take account of actual connectivity, we can add a layer showing the road and pedestrian bridges or tunnels. Details of these were sourced from the Wikipedia [76] article titled "*List of crossings of the River Thames*". In Figure 3.18, we have also drawn a 1 kilometre buffer around each crossing. This was chosen as an arbitrary measure of what might be considered a "reasonable" distance within which to consider opposing banks as being connected. The vast majority of units on opposing banks have access to a river crossing within this threshold and thus should be considered as neighbours. Only the extreme eastern units and one to the south west should not have a connection across the river according to this criterion.

In order the identify the changes we wish to make, we use the `nodes = "numeric"`

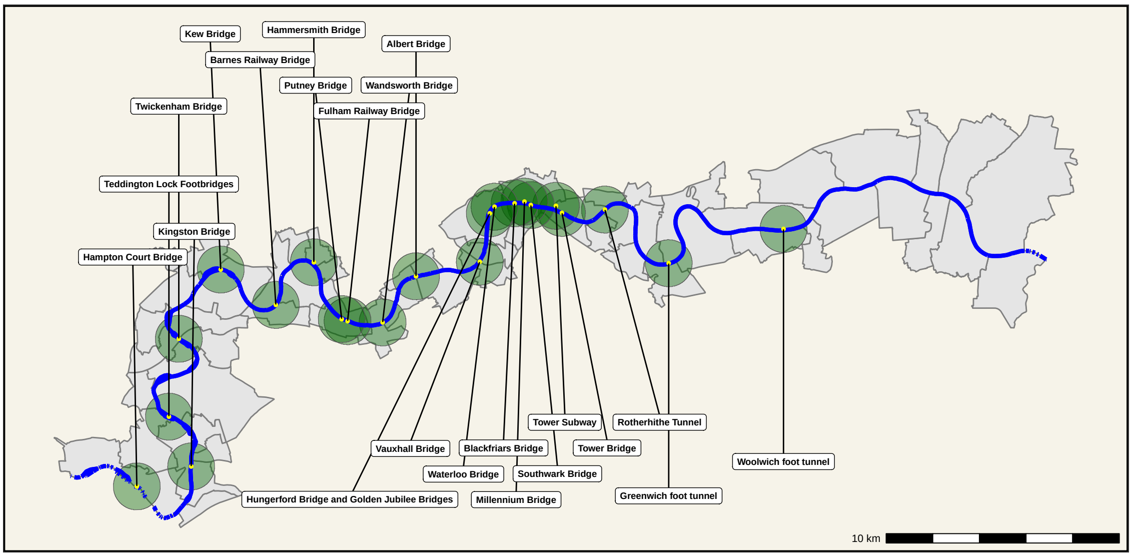


Figure 3.18: Thames crossings in Greater London.

Riverside wards of Greater London. Road and pedestrian crossings and tunnels are labelled and surrounded by 1 kilometre buffer shaded green.

argument in `st_quickmap_nb()`. Now we can identify each unit by its position in the contiguity structure. Here we have shaded in pink the units which are not within 1 kilometre of a river crossing (see Figure 3.19).

```
# with ’nodes = "numeric"’

st_bridges(london[riverside,],"NAME") |>
  st_quickmap_nb(nodes = "numeric",
                 numericsize = 4,
                 linksize = 0.5) +
  geom_sf(data=no_touch_buffer, fill="pink", alpha=0.3) +
  geom_sf(data=crossings_roadped |> st_buffer(1000),
          fill="darkgreen", alpha=0.3) +
  geom_sf(data=thames, colour="blue", linewidth=1.5) +
  geom_sf(data=crossings_roadped, size=1, colour="yellow") +
  annotation_scale(location="br") +
  coord_sf(datum=NA) +
  theme(panel.background = element_rect(fill = "#F6F3E9",
                                        colour = "black",
                                        linewidth=1.5))
```

This allows us to easily cut the ties across the river for these units by using the function

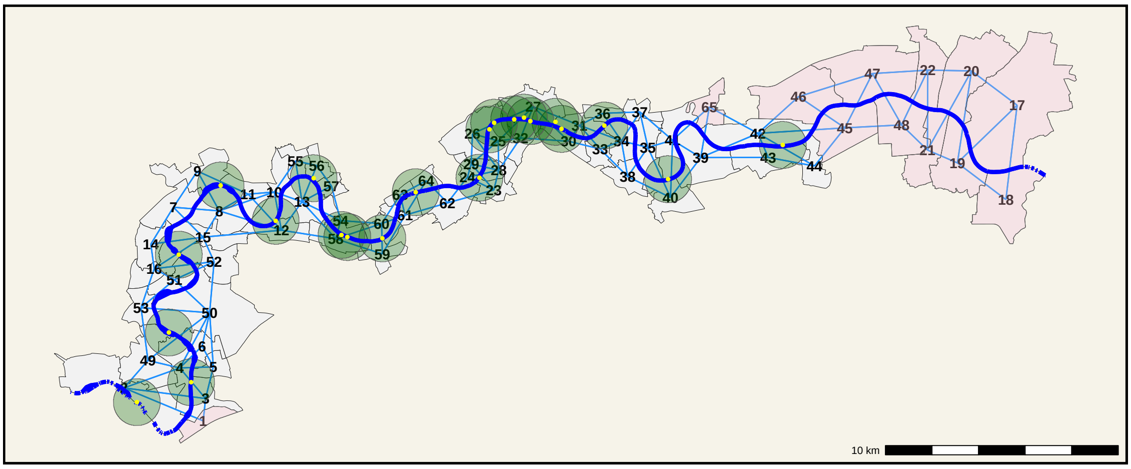


Figure 3.19: Determination of contiguity of riverside wards of Greater London across the River Thames.

> Riverside wards of Greater London. Index number for each ward shown at centroid. Wards which are not within 1 kilometre of a crossing are shaded pink.

`st_force_cut_nb()`. [c] Having made these adjustments, `st_quickmap_nb()` now shows a connectivity structure (Figure 3.20) which reflects our hypothesis of how influence should extend across the river in the presence or absence of crossings.

This example shows that the pre-functions of `sfislands` have uses for situations which do not involve islands. They can be used to apply domain knowledge to easily design the most appropriate neighbourhood structure.

```
# enforce cuts for the links where there is no crossing

cut_df <- tribble(
  ~x, ~y,
  18, 17,
  19, 17,
  19, 20,
  20, 21,
  21, 22,
  47, 48,
  45, 46,
  45, 47,
  39, 65,
```

[c] While we are using the index of the units in this example, the function also accepts names as arguments which may be more convenient in some circumstances.

```
  1, 2
)
st_bridges(london[riverside,], "NAME") |>
  st_force_cut_nb(xy_df = cut_df) |>
  st_quickmap_nb(bordercol = "black",
                 bordersize = 0.5,
                 linksize = 0.5) +
  geom_sf(data=no_touch_buffer, fill = "pink", alpha = 0.3) +
  geom_sf(data=crossings_roadped |> st_buffer(1000),
          fill= "darkgreen", alpha = 0.3) +
  geom_sf(data=thames, colour = "blue", linewidth = 1.5) +
  annotation_scale(location = "br") +
  coord_sf(datum=NA) +
  theme(panel.background = element_rect(fill = "#F6F3E9",
                                        colour = "black",
                                        linewidth = 1.5))
```

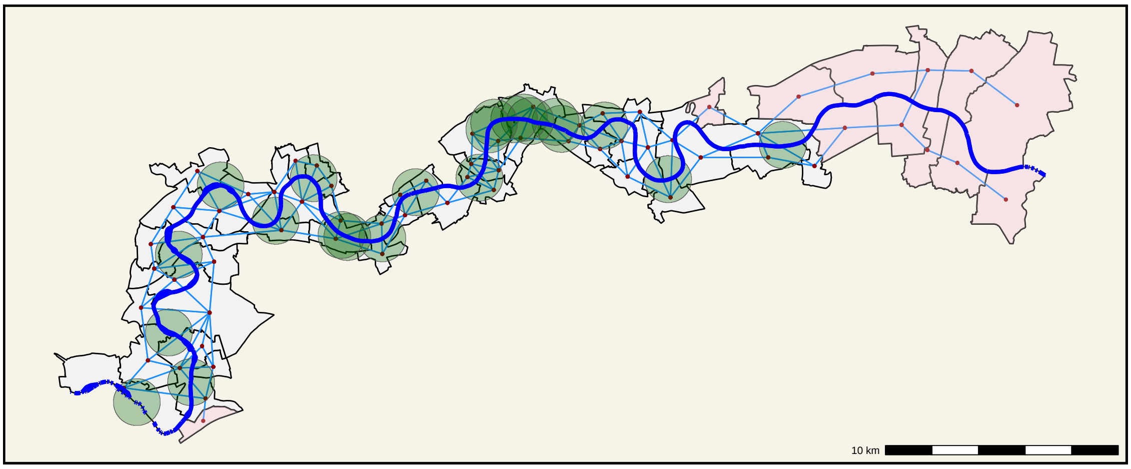


Figure 3.20: Modified contiguities of riverside wards of Greater London.

Riverside wards of Greater London. Contiguities across the river have been cut for the pink wards which do not lie within 1 kilometre of a crossing.

3.8 Summary

These examples have shown the varying scenarios in which `sfislands` can be useful. It aims to contribute to spatial modelling by making an awkward area less awkward. Rather than having a default attitude of ignoring islands when building neighbourhood structures based on contiguity, it offers a convenient system for enforcing linkages if that is deemed to be the most appropriate course of action. Even when no islands are present, it provides a simple procedure for tailoring a neighbourhood structure with bespoke contiguities to match a given hypothesis. It also provides helper functions to use these structures in spatial regression models, notably those built with `mgcv`, which streamline the human effort necessary to examine the estimates. In future, compatibility with other modelling packages can be added to broaden the package's capabilities.

4

Incorporating varying degrees of spatial cohesion in models of voter behaviour in the 2024 UK General Election

4.1 Abstract

Fitting models to data where multiple spatial processes are assumed to be operating simultaneously at different levels is often done by combining a hierarchical effects structure at the higher levels (region, county) with autoregressive structures at the lowest (individual data points). When dealing with areal data, an intrinsic conditional autoregressive (ICAR) component is often used to capture these lowest level spillover effects. One of the standard assumptions of ICAR models is a global spatial variance parameter, meaning all areas undergo the same degree of smoothing. We propose allowing this variance parameter to change, reflecting different degrees of spatial cohesion in different regions. We demonstrate an application using voter behaviour in the 2024 UK General Election. This technique has not enjoyed widespread exposure, particularly as regards election data. By fitting a series of models representing different hypothesised spatial processes for the vote count of each of the four largest parties and comparing the resultant marginal likelihoods, we find that this novel approach is the most plausible model, and reveals different patterns of spatial cohesion among the parties.

4.2 Introduction

When modelling spatial data, it is well understood that there is a tendency for nearby data points to be more similar than distant ones. It is also possible that groups of these data points can be seen as comprising regions, where a certain difference between and similarity within region is to be expected. Each of these phenomena describes a different type of spatial process, occurring at a different level of data aggregation. The practice of incorporating both of these processes simultaneously in a model has been termed hierarchical spatial autoregressive (HSAR) modelling by Dong and Harris [19]. They describe this as analogous to a combination of vertical and horizontal spatial processes.

The vertical processes can be captured using hierarchical modelling structures [29]. When applied in a geographical context, where the levels might refer to administrative units such as regions and counties [44], this top-down structure positions a spatial unit within a tree structure (e.g. a location in a town which lies within a province within a country). While this can do a good job at grouping potentially similar places together and sharing information across such groups, it can still lead to a situation where the potential relationship between neighbouring data points which happen to lie on either side of an administrative boundary is ignored.

The horizontal process can capture such neighbourhood relationships, regardless of group boundaries. This relies on an autoregressive structure to model spillover effects where each data point is dependent on nearby data points. Spatial units which are neighbours are expected to share similarities. One commonly-used such structure in areal models is an intrinsic conditional autoregressive (ICAR) component [5] which contributes a set of spatially smoothed random effects at the level of individual data points which are correlated according to contiguity, with each unit's effect being Normally distributed around the mean of that of its neighbours. The strength of this smoothing is controlled by a single variance parameter for the entire area.

Variations on this approach, particularly in the field of disease-modelling, include the BYM model [7] which combines this spatially structured component based on contiguity with an additional unstructured effect (i.i.d. Gaussian), and the related BYM2 model [66] which further parametrises the structure with a component to account for the relative proportion of each source of variation.

We propose a technique which combines group structures and an autoregressive component in a different way. Rather than capturing higher level group differences with a set of random effects where each group has a different mean divergence from the global mean, we instead incorporate group variation in the ICAR process itself, by allowing the variance parameter to vary by region, meaning that the strength of the spatial smoothing differs across groups. For units within each region, the random effect still depends on the average of the neighbours' random effects, but the amount of random variation around this average is determined by the group-specific variance parameter. Regions with lower variance exhibit stronger spatial smoothing (smaller variability among neighbours), while regions with higher variance allow for greater heterogeneity in the random effects within that region. This allows us to capture differences in spatial cohesion which may better account for the process than region random effects.

There are other approaches to the idea of a non-stationary variance component in the literature. Brewer and Nolan [11] introduced edge-specific weights within the contiguity structure to enable locally varying smoothness between all pairs of units. An iterative algorithm was proposed by Lee et al. [49] to detect spatial discontinuities and update the adjacency structure itself. Jack et al. [38] used a spatio-temporal model where the degree of spatial cohesion could vary with time. However, these methods do not seek to incorporate a hypothesised hierarchical structure within the ICAR variance parameter.

In this article, we examine an application of this approach using data from the 2024 UK General Election to see if it proves more plausible than other hypothesised approaches. Before discussing the spatial processes which we will consider, we first mention some other factors which are at play in the modelling of voting behaviour.

As Griffiths et al. [32] point out, it is well known that socio-economic factors - predominantly social class [14], education, age, deprivation and ethnicity - have long been strongly associated with political preferences in the UK, although Jennings et al. [40] observe that both their strength and direction of association have been changing over time for different parties.

In the coverage of this particular election, there was also a strong narrative of potential tactical voting where a voter might choose which party to vote for not based on a preference for that party but with the aim of maximising the chances of unseating the incumbent. This is particularly encouraged by the UK's first-past-the-post electoral system. Although difficult to model explicitly, it is clearly connected to the marginality and incumbency of a seat [54]. The size of the lead held by an incumbent party, whether or not that party is currently in government, and the identity of its closest competitor could all influence the likelihood of tactical voting occurring. By controlling for the incumbent party in each constituency, and also the second placed party from the previous election and the size of the majority, we seek to capture trends which are consistent with this type of behaviour.

Having controlled for these socio-economic and *status quo ante* factors, we still theorise that there would be unaccounted for spatial processes at play in determining voter behaviour. There are a number of reasons why voters might behave in a way which would reflect a vertical regional structure. Different parts of the country have deep-rooted political cultures associated with place and may be reluctant to change from these. The regional scale has proven useful [59] and can, for some contextual effects, be more powerful than local scales [61]. Furthermore, policies which a party proposes in a manifesto can have a spatial aspect to them which can make them more or less appealing to voters in a particular region. In the UK context, this could apply to "levelling-up", issues regarding the green belt, energy policy etc.

We would also expect a horizontal neighbourhood component when modelling voter behaviour. People who live nearby, work together, or associate together in bars, clubs and other organisations have a tendency to align in their political views, as summarised by Pattie and Johnston [60] in the maxim “those who speak together vote together”, referencing work done by Miller [52].

These two types of processes have been combined together in a HSAR modelling framework in the context of voting behaviour by Horan et al. [34], more specifically, examining swing between Labour and Conservatives in the 2019 election.

In the following application, to avoid complications arising from parties which competed only in Scotland, Wales and Northern Ireland, we restrict the study area to England where this election saw increased party system fragmentation [64] with the emergence of four dominant parties (each receiving greater than 10% of the national vote and a combined total of almost 90%): Labour, Conservatives, Liberal Democrats and Reform UK (see Table 4.1). We aim to model the number of votes obtained by each party in each constituency using a Bayesian approach. We propose a selection of credible models and compare their plausibility using log marginal likelihoods. Before turning to the 2024 election data, exploratory simulations are carried out to examine whether this approach can correctly identify known underlying model structures as being the most plausible. For the election data, we first fit models with no explicitly spatial component and examine their residuals. The presence of residual spatial patterns justifies the subsequent modelling exercises. Seven further models are then fitted for each of the four parties. These models increase in complexity from those which contain only one spatial process to others using a HSAR modelling framework with combined processes. One such model is our novel approach allowing the variance of the ICAR structure to change by region. We are then in a position to be able to rank these models according to marginal likelihood and determine which is most plausible for each party, given the voting data. In our example, we find that the novel approach is the best model for all four parties, revealing different patterns of spatial cohesion for each.

The utility of this model is not only that it seems to provide a more plausible mechanism than other models in some circumstances. The set of variance parameters which it estimates are of interest in their own right as a measure of the degree to which nearby places tend to be similar. Their interpretation will vary depending on the field of study, but they offer an additional avenue of investigation for spatial processes.

4.3 Data

4.3.1 Election 2024

The 2024 UK General Election saw Labour defeat the incumbent Conservative government. In England, the focus of this model, Labour increased their seat count by 166,

Table 4.1: Summary of party performance in 2024 UK General Election.

Percentage vote share, seat count, and percentage of seat totals in England for parties in the 2024 General Election. Four parties achieved a vote share greater than 10%, totalling 89% of the total votes cast and over 98% of seats won.

Party	**Percentage of Votes**	**Seat Count**	**Percentage of Seats**
Labour	34.5	347	64.0
Conservative	25.9	116	21.4
Reform UK	15.4	5	0.9
Liberal Democrat	13.2	65	12.0
Green	7.3	4	0.7
Other	3.7	5	0.9

the Conservatives lost 229 seats, the Liberal Democrats increased their seat count by 59, while Reform UK won 5 seats. However, owing to the first-past-the-post electoral system, these numbers are far from reflective of the relative number of votes cast for each party, particularly so in this election [58]. Summaries of these seat and vote counts for each party are shown in Table 4.1. Despite winning only a fraction of the number of seats which the Liberal Democrats attained, Reform UK actually acquired a greater number of votes across England. This is a result of the voting system and its interplay with differences in concentration and distribution of votes by location, emphasising the particularly key role played by geography in UK elections. For the purposes of this study, we are not concerned with whether or not parties won a particular seat. Instead, we attempt the model the total number of votes gained by a party in each constituency.

The two data sources in this study are the 2021 census and the results of the 2024 and 2019 General Elections. Census results were sourced from the Office for National Statistics [56] and voting data from the House of Commons Library [37]. The census results are available at the level of the 2024 constituency boundaries so these were joined with the most recent election results. When including results from the 2019 election as covariates, some complications arose because of boundary changes. As a result, it has been necessary for the purposes of comparison to re-project the 2019 constituency vote counts to their 2024 equivalents. This was achieved by assigning 2019 constituency votes to 2024 constituencies according to the proportion of population living in areas of overlap [36].[a]

Of the 543 constituencies in England, 542 were considered. We omitted Chorley as this was the constituency of the Speaker of the House which, by convention, is not contested by the major parties. In subsequent maps, it can be identified as an uncoloured unit towards the north west. A Labour candidate competed in all 542 constituencies, a Conservative

[a] The conversion procedure is outlined in more detail at `https://github.com/horankev/voteReproject`.

or Liberal Democrat candidate in 541, and a Reform UK candidate in 521.

4.3.2 Dependent variable

The outcome we are seeking to model is vote count per constituency for each party. To give an idea of the spatial patterns and variation between parties, Figure 4.1 shows the percentage of votes which each party attained in 2024, diverging about the overall median. We show percentages in this figure to control for different constituencies having different populations. In these maps, the constituencies are shown as hexagons of equal size, while still approximating relative position. This overcomes the problem of invisibility arising from some small, densely populated urban constituencies. In later maps, where a more precise idea of relative position and contiguity is an important consideration (as is the case when exploring neighbourhood relationships) maps will be shown in regular projection.

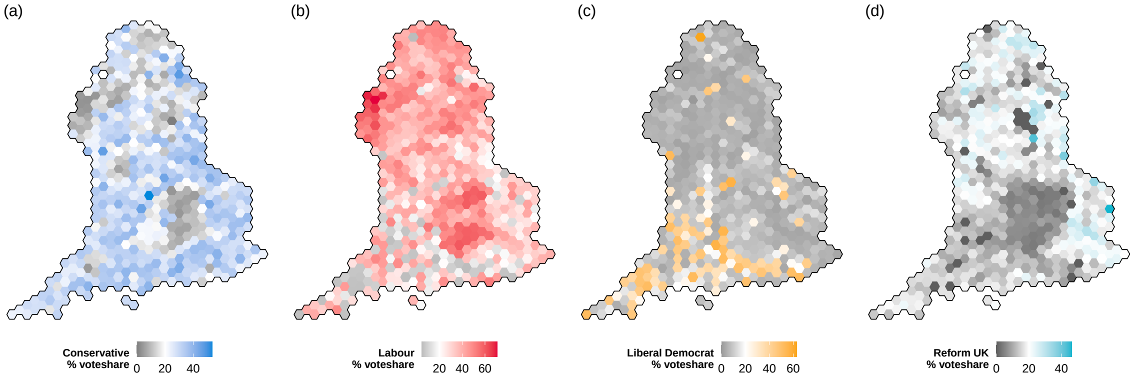

(a)
(b)
(c)
(d)
Conservative % voteshare 0 20 40
Labour % voteshare 20 40 60
Liberal Democrat % voteshare 0 20 40 60
Reform UK % voteshare 0 20 40

Figure 4.1: Party vote share in 2024 UK General Election.

(a-d) Percentage vote share by constituency for four largest parties in 2024 UK General Election. The white hexagon with black outline towards the north west in these and subsequent maps represents the Speaker's seat which, by custom, is not contested.

4.3.3 Explanatory variables

The three socio-economic variables included in the model from the 2021 census are shown in Figure 4.2, diverging about their median values. They represent, respectively, the proportion of the population of each constituency with a degree, the proportion self-reporting fair, bad or very bad health, and the proportion of white ethnicity. These three variables are a subset of those considered by Beecham et al. [4] in a study of the connection between the narrative of “left-behind” places and voting behaviour in the Brexit referendum. It seems reasonable to expect that cleavages which they capture are likely to remain relevant in subsequent General Elections. Other variables which they considered included professional occupations, younger adults, English as main language and home ownership. They were, however, interested in examining individual relationships between these variables and voting behaviour. As we are modelling these covariates together, issues of multicollinearity would arise without careful selection. These three variables were found to be highly correlated with (and thus good proxies for) many other socio-demographic variables (e.g. degree education with professional occupations, poor health with age, white with English-speaking, Christian and single-ethnicity households), while also not leading to problematic levels of multicollinearity, In the subsequent models, they are scaled to have a mean of 0 and standard deviation of 1 to improve sampling efficiency and stability in a Bayesian context.

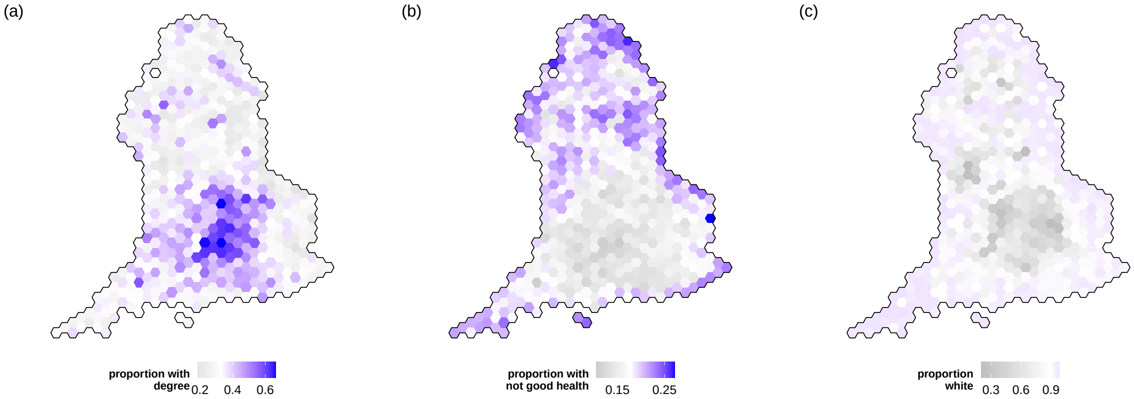

(a)
(b)
(c)
proportion with degree 0.2 0.4 0.6
proportion with not good health 0.15 0.25
proportion white 0.3 0.6 0.9

Figure 4.2: Explanatory variables from 2021 census.

Proportion of population by constituency (a) with a degree, (b) with fair, bad or very bad health, and (c) of white ethnicity.

The model also accounts for potential drivers of tactical voting - the winner and second placed party from the previous election, and the size of the majority of the winning party, as shown in Figure 4.3. This *marginality* is calculated as the difference between the total number of votes for the first and second placed parties. Because of varying constituency size, it is expressed as a proportion of the total overall number of votes cast in each constituency. There is evidence that tactical voting tends to be higher in marginal seats [16, 42]. For the same reasons as the socio-economic explanatory variables, these majorities were also scaled.

While Reform UK are shown separately in the second place map in Figure 4.3 (c), they were the second placed party in only one constituency in 2019. For this reason, they were included in the *other* category for this variable in the modelling process, along with other parties and independent candidates.

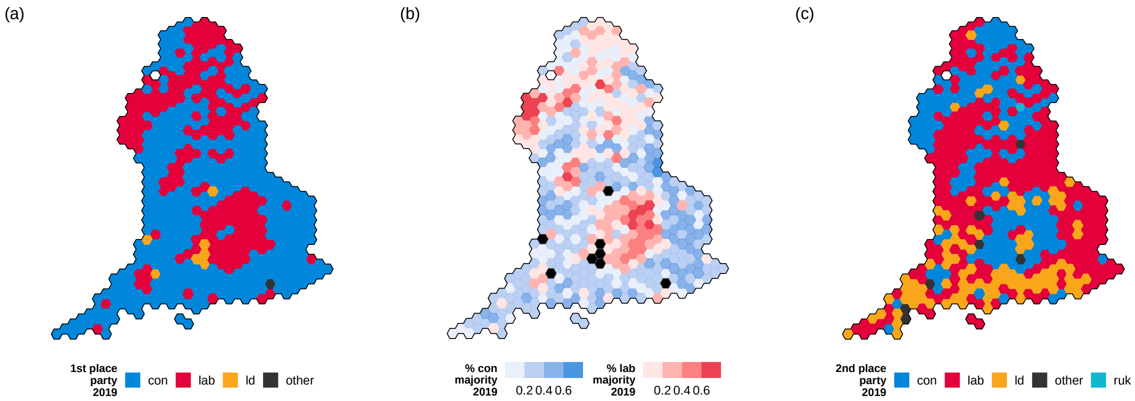


Figure 4.3: Status quo ante from 2019 UK General Election.

(a) First place parties by constituency in 2019, (b) majority as a proportion of total votes cast for Conservative and Labour seats, and (c) second place parties by constituency in 2019. All constituencies are reprojected to 2024 boundaries.

4.3.4 Spatial structures

When discussing space in these models, we will be concerned with vertical location as part of a tree with regions as its branches (Figure 4.4), and horizontal location as captured by contiguity (Figure 4.5). The nine regions of England which will form part of the modelling process and future discussion are mapped in Figure 4.4. They are the highest tier of sub-national division in England. They have a history of association as former constituencies for European Parliament elections, between 1994 and 2011 they had partly devolved functions, and they were previously the first level NUTS regions[b] within the European Union.

When describing connectivity between constituencies in a neighbourhood matrix, as is required for ICAR models, the presence of islands or otherwise disconnected units can lead to issues with computation. While there are no longer any island constituencies in England after the division of the Isle of Wight into two separate seats, we are still left with the situation where the constituencies of Isle of Wight East and West are entirely separate from the rest of the graph. The island is served by two regular ferry services which facilitate commuting for work, socialising and cultural exchange in a way that is consistent with the theory of how voter interaction can lead to neighbourhood effects. For this reason, using the package `sfislands` [35], additional neighbour connections have been added between Isle of Wight West and New Forest West, and between Isle of Wight East and Gosport. These correspond to the ferry connection routes. Even though Reform UK only competed in 521 out of 542 constituencies, this relative sparsity did not lead to

[b] Nomenclature of Territorial Units for Statistics (NUTS) is a Eurostat geocode standard for referencing the administrative divisions of countries for statistical purposes.

any disconnected units. The resultant neighbourhood structures for Labour and Reform UK can be seen in Figures 4.5 (a) and (b) respectively. The Conservative and Liberal Democrat contiguity maps, with 541 competing constituencies, closely resemble that of Labour.

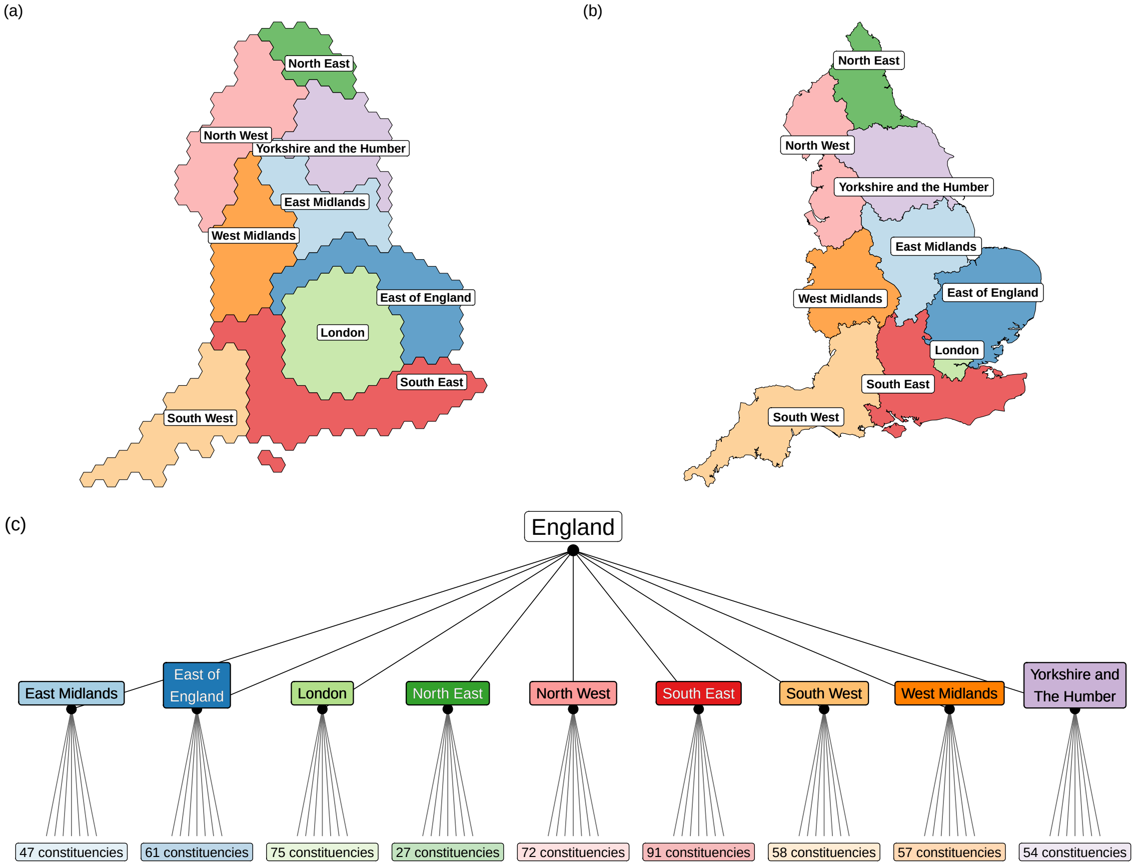


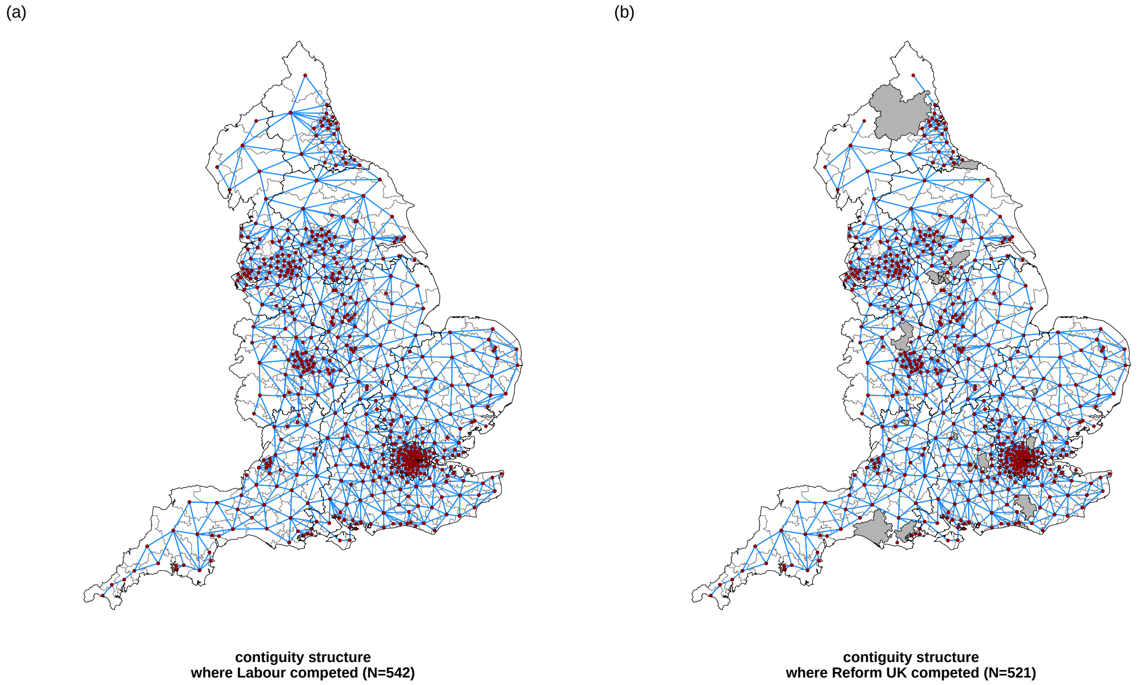
Figure 4.4: Regions of England and tree structure of constituencies.

Regions of England corresponding to (a) constituency hexagon maps and (b) regular projection. (c) Constituencies and regions of England visualised as a hierarchical tree structure. This can be seen as vertical modelling because only constituencies which branch from a common higher level region are assumed to share share similarities.

4.4 Methods

4.4.1 Non-spatial model

Prior to an examination of the relative merits of different spatial processes, we first seek to establish that there are indeed such processes which need to be accounted for. The



contiguity structure
where Labour competed (N=542)

contiguity structure
where Reform UK competed (N=521)

Figure 4.5: Neighbourhood structures for Labour and Reform UK.

Constituencies linked according to queen contiguity for (a) Labour and (b) Reform UK, with an additional connection made between each Isle of Wight constituency and the mainland constituency providing ferry connection. All contiguities are free to cross region boundaries. Conservative and Liberal Democrat contiguity structures (not mapped) only differ from Labour by one constituency each. In (b), constituencies where Reform UK did not compete are shown in grey.

proposed general structure for this model, without incorporating any spatial information, is a Poisson generalised linear model where the dependent variable Y_i is the count of the number of votes cast for a party in each constituency. We seek to model this as a linear combination of variables via a log link function. One such group of variables, as discussed above, is a set of covariates from the census. These have been used in previous models of electoral behaviour and serve as proxies for a broad range of socio-economic conditions. We also control for the political status quo ante by including as predictors the first and second placed parties from the previous election, the size of the majority, and its interaction with the second-placed party. An offset of the log of the total number of votes cast controls for different exposures due to the varying sizes of electorates by constituency. The model is fitted in a Bayesian framework in R using the `brms` package [13]. To assess convergence in these and subsequent models, we considered the convergence diagnostic R-hat [72], as well as the individual parameter trace plots.

Such a model structure has the following likelihood for the distribution of votes cast in each constituency i:

$$Y_i \sim \text{Poisson}(\mu_i),$$

where

$$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \text{offset}_i$$

The covariates in the $\mathbf{x}_i^\top \boldsymbol{\beta}$ component are arranged according to the structure in Table 4.2. We assign weakly informative but proper priors, Normal$(0, 10)$, to all β parameters, reflecting minimal prior knowledge or assumptions about their values.

When we fit the model described above, which contains no spatial components, a visual examination of its Pearson residuals, mapped in Figure 4.6, shows clear spatial patterns. We see, at constituency level, clusters of positive and negative residuals, suggesting that neighbours have a tendency to be similar to neighbours. At a regional level, we can see that some regions display similarities which are quite different to those of others.

4.4.2 Spatial models

The presence of these patterns suggests our model could be improved by accounting for spatial processes. We allow for the possibility that these could be at constituency level, region level, or both. With this in mind, seven further candidate models are fitted for each party containing different combinations of hypothesised processes, as summarised in Table 4.3.

For Models 1 and 2, we take the non-spatial model described above and additionally include region random effects only. In the case of Model 1, each effect is modelled as a

Table 4.2: Explanatory variables for 2024 election models.

Covariates with a brief description, classified as either coming from the 2021 census or the results of the previous election.

Explanatory Variables	Description
census 2021	
degree	scaled proportion of constituency population with a degree
not good health	scaled proportion of constituency population reporting fair, bad, or very bad health
white	scaled proportion of constituency population of white ethnicity
status quo ante	
first-placed party 2019	winner from the previous election
second-placed party 2019	closest party to winner from previous election
marginality	scaled difference between winning party and closest rival, as proportion of total votes
interaction of second-placed party and majority	influence of marginality can vary according to closest competitor

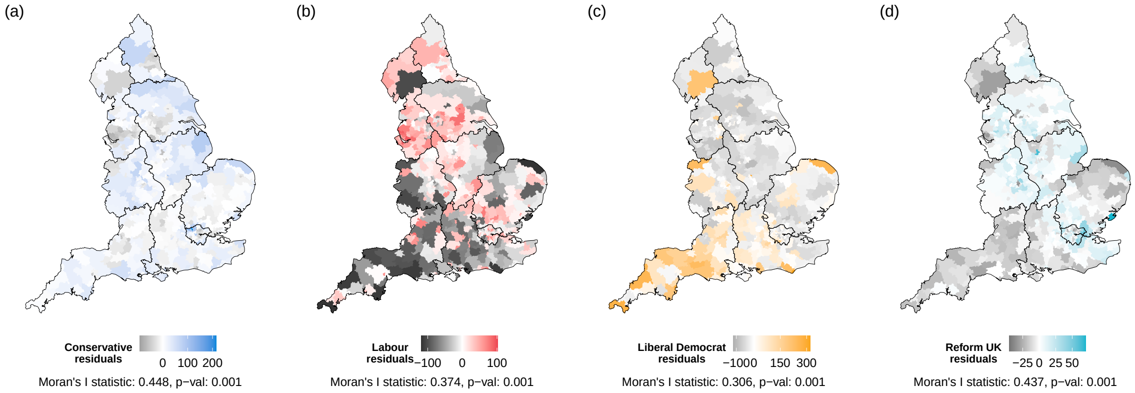

(a)
(b)
(c)
(d)
Conservative residuals
0 100 200
Moran's I statistic: 0.448, p−val: 0.001
Labour residuals
−100 0 100
Moran's I statistic: 0.374, p−val: 0.001
Liberal Democrat residuals
−1000 150 300
Moran's I statistic: 0.306, p−val: 0.001
Reform UK residuals
−25 0 25 50
Moran's I statistic: 0.437, p−val: 0.001

Figure 4.6: Residuals from non-spatial models.

(a-d) Pearson residuals from non-spatial models for each party. Clusters of positive and negative residuals are visible, in addition to regional patterns, suggesting the presence of spatial processes which have not been accounted for. Further evidence for this is provided by a Moran's I statistic of spatial autocorrelation for each set of residuals, and the p-values of hypothesis tests for random spatial distribution of values.

Table 4.3: Summary of candidate models.

Type of spatial process(es) incorporated in each model, and whether they occur at region level, constituency level, or both.

Model	Constituency Spatial Effect	Region Spatial Effect
non-spatial		
non-spatial	—	—
region only		
1	—	independent region random effects
2	—	region random effects from common distribution
constituency only		
3	ICAR process	—
4	BYM2 (ICAR + unstructured effect)	—
region and constituency		
5	ICAR process	independent region random effects
6	ICAR process	region random effects from common distribution
7	ICAR process	varying icar sd by region

dummy variable, considered to come from an independent sample. For Model 2, these effects are drawn from a common distribution of region effects as is the case in hierarchical modelling frameworks.

Model 3 includes only constituency level effects, where we account for the theory that the behaviour of voters in one constituency is likely to be similar to that of its neighbours using an ICAR process. This enforces a constant degree of spatial smoothing across the surface. Model 4 is the BYM2 variant of the ICAR model with a combination of spatially structured and unstructured effects at constituency level, whose relative proportion is also estimated.

Models 5 and 6 combine the region effects of Models 1 and 2 respectively with the ICAR component of Model 3. These fall into the category of hierarchical spatial autoregressive (HSAR) modelling, as previously discussed.

Finally, Model 7 contains a novel feature. Despite only featuring a constituency level ICAR component, in this scenario we account for regional variation by dropping the constraint that the standard deviation of the constituency level effects is constant, and allow it to vary by region. This allows the degree of spatial cohesion to be modelled and for this to be reflected in the resulting ICAR component. In some parts of the country, there may be a stronger or weaker tendency to behave like your neighbours than in other places. This feature allows a regional effect to operate in a more subtle way than by simply raising or lowering the mean region level through a random effect.

Models 1-6 are fitted in R using the `brms` package which relies on implementations by Morris et al. [53]. Model 7 uses the `rstan` package [69] with a modification of the Stan code of Model 3, as generated using the `brms` function `get_stancode()`.

A summary of the structural component of each model, $\log(\mu_i)$ or $\log(\mu_{ik})$ as appropriate for i constituencies and k regions, and a brief explanation are shown in Table 4.4.

Table 4.4: Structure of candidate models.

Structure of each model, showing which spatial processes are included.

Model	Description	Note
non-spatial		
non-spatial	$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \text{offset}_i$	- where $\mathbf{x}_i^\top \boldsymbol{\beta}$ contains explanatory variables from census and previous election
region only		
1	$\log(\mu_i) = \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}} + \text{offset}_i$	- where $\tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}}$ includes an additional region component
2	$\log(\mu_{ik}) = \mathbf{x}_i^\top \boldsymbol{\beta} + \gamma_k + \text{offset}_i$	- where γ_k is a region random effect from a common distribution, with $\gamma_k \sim \text{Normal}(0, \sigma_R^2)$
constituency only		
3	$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \phi_i + \text{offset}_i$	- where ϕ_i is a constituency random effect from an ICAR process
4	$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + \psi_i + \text{offset}_i$	where ψ_i is a mixture of spatial effects and unstructured noise according to a mixing parameter ρ
region and constituency		
5	$\log(\mu_i) = \tilde{\mathbf{x}}_i^\top \tilde{\boldsymbol{\beta}} + \phi_i + \text{offset}_i$	- region effects of model 1 and constituency effects of model 3
6	$\log(\mu_{ik}) = \mathbf{x}_i^\top \boldsymbol{\beta} + \gamma_k + \phi_i + \text{offset}_i$	- region effects of model 2 and constituency effects of model 3
7	$\log(\mu_{ik}) = \mathbf{x}_i^\top \boldsymbol{\beta} + \gamma_k + \phi_{ik} + \text{offset}_i$	- where ϕ_{ik} allows for varying standard deviation by region in the ICAR process

Priors

As in the non-spatial model, we assign weakly informative but proper priors, Normal$(0, 10)$, to all β parameters, reflecting minimal prior knowledge or assumptions about their values. In addition to the census variables and the previous election effects, such priors also apply to region effects where they form a part of the model.

The priors for the two different types of ICAR component, ϕ_i and ϕ_{ik}, are defined as follows. When incorporating an ICAR process, the conditional distribution of a constituency-level spatial random effect ϕ_i, given its neighbours, is:

$$p(\phi_i \mid \phi_j, j \neq i, \sigma) \sim \text{Normal}\left(\frac{\sum_{i\sim j}\phi_i}{d_i}, \frac{\sigma^2}{d_i}\right)$$

where d_i is the number of neighbours for constituency i, and $i \sim j$ refers to pairs of neighbouring constituencies. The individual spatial random variable ϕ_i for constituency i which has a set of neighbours $j \neq i$ whose cardinality is d_i, is Normally distributed with a mean equal to the average of its neighbours. Its variance decreases as the number of neighbours increases.

The joint distribution for the vector of all spatial random effects $\boldsymbol{\phi} = (\phi_1, ..., \phi_N)$ can be written as a pairwise difference:

$$p(\boldsymbol{\phi} \mid \sigma) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i\sim j}(\phi_i - \phi_j)^2\right)$$

This enforces similarity among neighbouring locations. If locations i and j are neighbours, their effects ϕ_i and ϕ_j are penalised for being different.

To ensure identifiability, a soft sum-to-zero constraint is applied:

$$\sum_{i=1}^{N}\phi_i \sim \text{Normal}(0, 0.001 \cdot N)$$

Instead of this ϕ_i ICAR effect, the BYM2 formulation in Model 4 includes a random effect ψ_i which is a combination of structured (ICAR) and unstructured effects at constituency level:

$$\psi_i = \sigma(\sqrt{1-\rho}\,\upsilon_i + \sqrt{\rho}\,\phi_i^*)$$

where σ is the overall standard deviation, ϕ_i^*, the spatially structured effect, is a scaled version of the ICAR prior with unit variance, υ_i represents unstructured noise, with $\upsilon_i \sim \text{Normal}(0,1)$, and ρ is the proportion of spatial variance in the mixing of these structured and unstructured components. A proper, uninformative prior of $\text{Beta}(1,1)$, uniform in the interval [0,1], is used for ρ. A value of 0 for ρ implies that all spatial variation is unstructured (independent random effects) while a value of 1 is equivalent to a pure ICAR model.

In Model 7, where the smoothness of the ICAR effect is allowed to vary by region k, the spatial prior changes to the following, where σ_k represents a different standard deviation for each region:

$$p(\phi_{ik} \mid \phi_j, j \neq i, \sigma_k) \sim \text{Normal}\left(\frac{\sum_{i\sim j} \phi_{ik}}{d_i}, \frac{\sigma_k^2}{d_i}\right)$$

The soft sum-to-zero constraint, this time for ϕ_{ik}, remains unchanged. The hyper-parameters σ or σ_k, representing the fixed or varying smoothness of the ICAR component respectively were initially assigned uniform priors. This, however, led to very inefficient sampling. Instead, a tighter prior was chosen with a probability distribution function derived from a generator of random values from a Student-t distribution with mean 0, standard deviation 2.5, and degrees of freedom 3, where the absolute value of each draw is used, implying an expectation that the value will be close to zero but not ruling out the possibility of much larger values:

$$\sigma \sim \text{Student-t}(3, 0, 2.5), \quad \sigma > 0$$

This achieved similar results to the uniform prior but with much less computation time. This prior is also used for the σ_R hyper-parameter of the γ_k region effect distribution in Models 2 and 5.

4.5 Exploratory simulations

Before fitting a sequence of competing models for each party, we first seek to verify that the proposed approach of using marginal likelihoods to determine relative plausibility is effective. If the underlying spatial structure is known, will the corresponding model emerge as the most plausible? This is done using some exploratory simulations.

To create a realistic but less computationally intensive geographical setting, we extract the 101 constituencies from the regions of Yorkshire and the Humber and the East of England. We then divide these into six pseudo-regions, with each containing only contiguous constituencies, by a process of hierarchical clustering about their centroid positions (see Figure 4.7).

Three distinct types of spatially structured data are randomly generated, each with five independent replicates: the first type follows an ICAR structure (as in Model 3); the second combines an ICAR structure with a region effect (Model 5); and the third uses an ICAR structure with region-specific variance σ_k (Model 7).

We then fit our eight competing models to each of the simulated structures and estimate the log marginal likelihood using the `bridgesampling` package [33], which provides a stable and accurate estimate via bridge sampling [51]. This method uses posterior samples to estimate the otherwise intractable marginal likelihood by iteratively computing the ratio between posterior and prior normalizing constants. For Models 3 and 5, we begin with a relatively tight ICAR variance (σ) of 0.7. In the Model 5 structure, a range

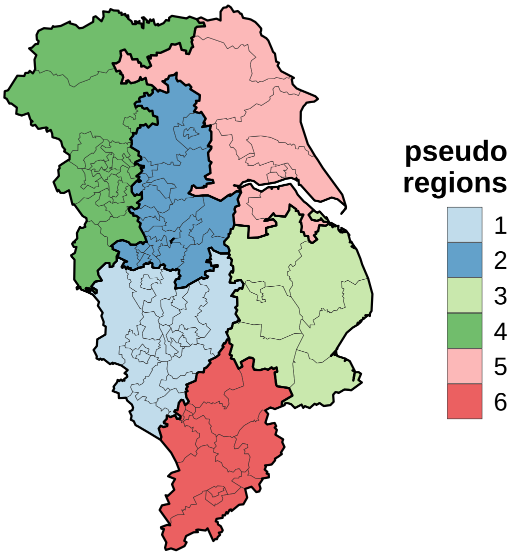


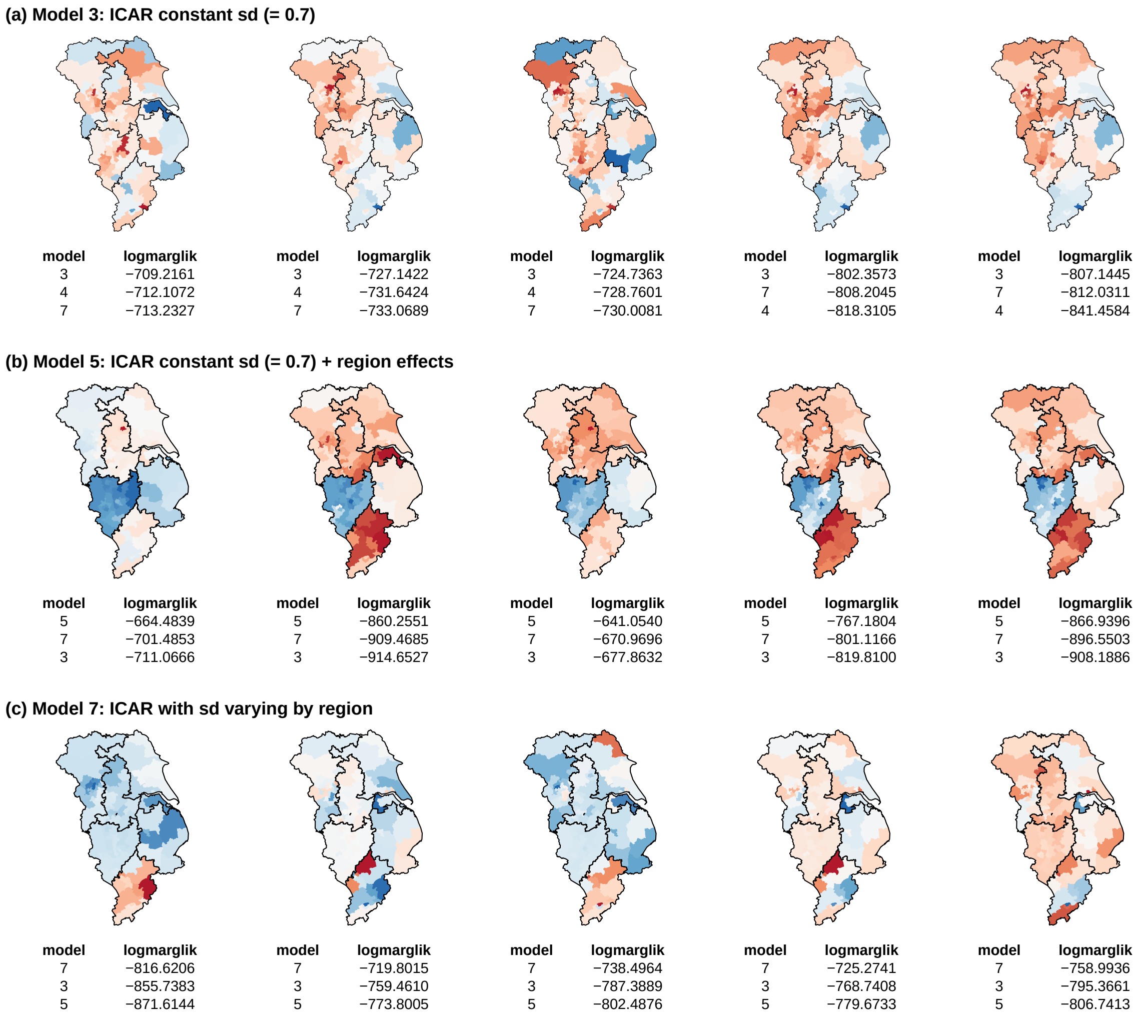
Figure 4.7: Pseudo-regions for simulations.

A spatial structure for performing simulations is created by extracting the boundaries of constituencies from the East of England and Yorkshire and The Humber and dividing them into 6 pseudo-regions, each containing only contiguous constituencies, by a process of hierarchical clustering about their centroid positions.

of six region effects is introduced: –1.5, 1, –0.5, 0.5, 1, and 1.2. For Model 7, samples are generated using six region-specific ICAR variances (σ_k): 0.2, 0.6, 1, 1.4, 1.8, and 2.2. Figure 4.8 displays each of these structures, with the three most plausible models (based on log marginal likelihood) shown below each map. In all cases, the correct model is preferred.

Additional noise is then introduced by increasing the ICAR σ value to 1.3 in Models 3 and 5 (see Figure 4.9). In one of the five replicates, the model with region-specific ICAR variance (σ_k) was marginally preferred over the true underlying ICAR structure. This outcome is expected: when spatial data are generated with high noise, it is not surprising that a more flexible structure (such as varying σ_k) may occasionally offer a better or equally good fit. In all other cases, the correct models were preferred.

While this exploratory simulation is not exhaustive, it provides preliminary evidence that the method can correctly identify the true underlying structure as the most plausible among competing alternatives, rather than favouring, for example, the model with the greatest complexity.



(a) Model 3: ICAR constant sd (= 0.7)

model	logmarglik
3	−709.2161
4	−712.1072
7	−713.2327

model	logmarglik
3	−727.1422
4	−731.6424
7	−733.0689

model	logmarglik
3	−724.7363
4	−728.7601
7	−730.0081

model	logmarglik
3	−802.3573
7	−808.2045
4	−818.3105

model	logmarglik
3	−807.1445
7	−812.0311
4	−841.4584

(b) Model 5: ICAR constant sd (= 0.7) + region effects

model	logmarglik
5	−664.4839
7	−701.4853
3	−711.0666

model	logmarglik
5	−860.2551
7	−909.4685
3	−914.6527

model	logmarglik
5	−641.0540
7	−670.9696
3	−677.8632

model	logmarglik
5	−767.1804
7	−801.1166
3	−819.8100

model	logmarglik
5	−866.9396
7	−896.5503
3	−908.1886

(c) Model 7: ICAR with sd varying by region

model	logmarglik
7	−816.6206
3	−855.7383
5	−871.6144

model	logmarglik
7	−719.8015
3	−759.4610
5	−773.8005

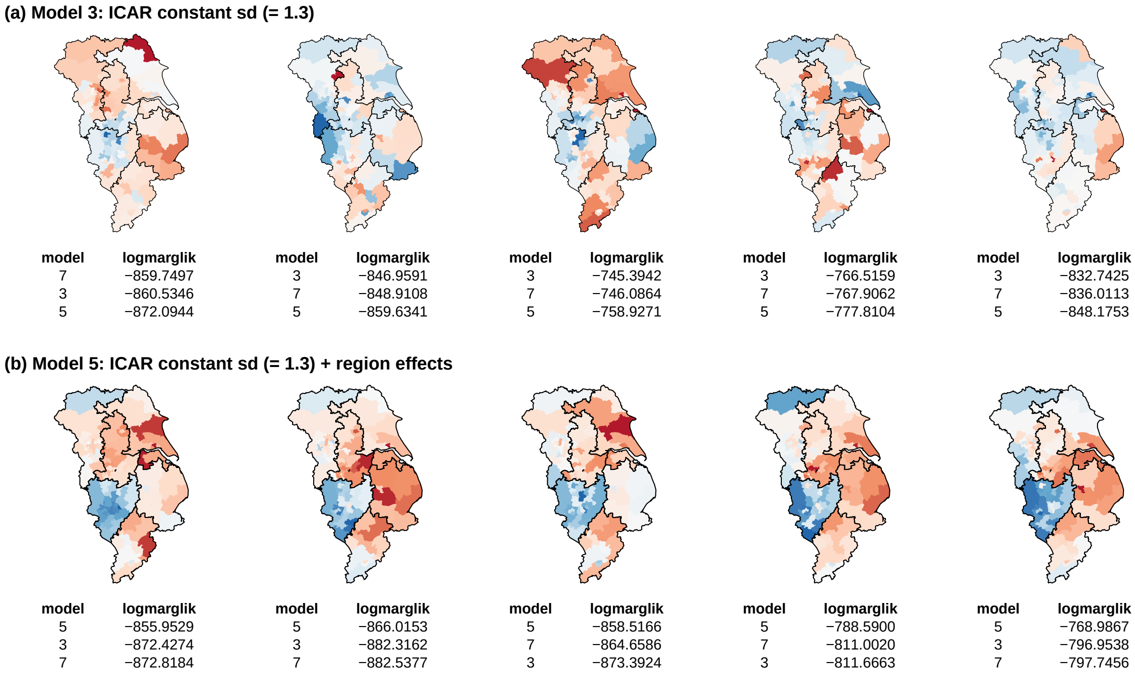
model	logmarglik
7	−738.4964
3	−787.3889
5	−802.4876

model	logmarglik
7	−725.2741
3	−768.7408
5	−779.6733

model	logmarglik
7	−758.9936
3	−795.3661
5	−806.7413

Figure 4.8: Simulation outputs from situation of moderate standard deviations.

Each row consists of five random samples of data with a known spatial association. The first row are ICAR, the second are ICAR with an additional region effect, the third are ICAR with standard deviation varying by region. In all cases, the use of marginal likelihoods identifies the correct underlying spatial structure as the most plausible.



(a) Model 3: ICAR constant sd (= 1.3)

model	logmarglik
7	−859.7497
3	−860.5346
5	−872.0944

model	logmarglik
3	−846.9591
7	−848.9108
5	−859.6341

model	logmarglik
3	−745.3942
7	−746.0864
5	−758.9271

model	logmarglik
3	−766.5159
7	−767.9062
5	−777.8104

model	logmarglik
3	−832.7425
7	−836.0113
5	−848.1753

(b) Model 5: ICAR constant sd (= 1.3) + region effects

model	logmarglik
5	−855.9529
3	−872.4274
7	−872.8184

model	logmarglik
5	−866.0153
3	−882.3162
7	−882.5377

model	logmarglik
5	−858.5166
7	−864.6586
3	−873.3924

model	logmarglik
5	−788.5900
7	−811.0020
3	−811.6663

model	logmarglik
5	−768.9867
3	−796.9538
7	−797.7456

Figure 4.9: Simulation outputs from situation of larger standard deviations.

Each row consists of five random samples of data with a known spatial association. The first row are ICAR, the second are ICAR with an additional region effect. These samples have a larger standard deviation of 1.3 meaning there is less spatial cohesion than in the previous example. In almost all cases, the use of marginal likelihoods identifies the correct underlying spatial structure as the most plausible.

4.6 Results

4.6.1 Model comparison

For each party, we evaluate eight competing models, each representing a different probabilistic structure that could have generated the observed voting data. We compute the log marginal likelihood of each model using bridge sampling. These marginal likelihoods can also be expressed as Bayes factors, enabling evidence-based comparison between models.

A Bayes factor compares the likelihood of the observed data under two competing models. Specifically, the Bayes factor for Model 1 relative to Model 2 quantifies how much more likely the data are under Model 1 than under Model 2. Values greater than 1 indicate evidence in favour of Model 1, while values less than 1 support Model 2. The magnitude of the Bayes factor reflects the strength of this evidence. We interpret these values using the scale proposed by Lee and Wagenmakers [50] (an update to the original guidelines of Jeffreys [39]) as shown in Table 4.5.

Table 4.5: Lee and Wagenmakers' interpretation of Bayes factors.

Quantifying how much more likely data are under one model compared to another according to different ranges of Bayes factor values.

Bayes Factor Range	**Model Comparison Evidence**
$> 0, < 1$	negative
$\geq 1, < 3$	anecdotal
$\geq 3, < 10$	moderate
$\geq 10, < 30$	strong
$\geq 30, < 100$	very strong
≥ 100	extreme

Table 4.6 shows the results of comparing the merits of the eight potential models for each of the four parties using this classification structure. The models are ranked for each party according to decreasing log marginal likelihood. Bayes factors are shown comparing each model to the model immediately below it. Because they have been ranked, Bayes factor values of less than 1 do not occur. For all sets of models in Table 4.6, the most plausible is Model 7, with an ICAR component whose standard deviation is allowed to vary by region. The evidence in favour of this model over its closest competitor is in each case classified as *extreme*. With some minor exceptions, the order of preference of models is consistent across parties.

4.6.2 Covariate estimates from optimal models

Having established which of our competing models is most appropriate for each of the parties, we can examine the posterior distributions of some parameters from the most favoured models for each party. All of these distributions are from Model 7. Looking at the socio-economic components (Figure 4.10), we can see differences in directions of association for each party. The proportion of people with a *degree* is negatively associated with Reform votes, while the association is positive for the Conservatives and Liberal Democrats. The opposite is the case for *health not good*, having a suggestion of positive association with Reform votes and negative for Conservatives and Liberal Democrats. The proportion of a constituency of *white* ethnicity is most strongly positively associated with Reform votes, but is also positive for Labour.

The associations of votes in 2024 with the winning party from 2019, and also with the second placed party, the majority, and their interaction are treated in this example as nuisance variables for control purposes. Interpretation of their posterior distributions is complicated because of their calibration with different relative baselines. For this reason, they are not examined here.

Table 4.6: Comparison of 2024 models across parties.

Comparison of models across parties, ranked by decreasing log marginal likelihood, alongside the degree of evidence for model improvement provided by Bayes factor comparison with the model immediately below it. The sequence is generally the same for each party with Model 7 being the most plausible.

Model	**Constituency Level Effects**	**Region Level Effects**	**Log Marginal Likelihood**	**Log Bayes Factor**	**Bayes Factor**	**Evidence**
Conservative						
7	ICAR	varying ICAR sd	-4809.537	49	>1e+6	extreme
3	ICAR	—	-4859.002	35	>1e+6	extreme
5	ICAR	independent distributions	-4894.030	14	>1e+6	extreme
6	ICAR	common distribution	-4907.638	42	>1e+6	extreme
4	BYM2	—	-4949.994	78000	>1e+6	extreme
2	—	common distribution	-83402.581	35	>1e+6	extreme
1	—	independent distributions	-83437.307	250000	>1e+6	extreme
non-spatial	—	—	-328526.802	—	—	—
Labour						
7	ICAR	varying ICAR sd	-5171.735	82	>1e+6	extreme
3	ICAR	—	-5253.968	32	>1e+6	extreme
5	ICAR	independent distributions	-5286.028	200	>1e+6	extreme
6	ICAR	common distribution	-5488.055	30	>1e+6	extreme
4	BYM2	—	-5518.061	170000	>1e+6	extreme
2	—	common distribution	-178712.220	30	>1e+6	extreme
1	—	independent distributions	-178742.703	480000	>1e+6	extreme
non-spatial	—	—	-659934.748	—	—	—
Reform UK						
7	ICAR	varying ICAR sd	-4253.464	53	>1e+6	extreme
3	ICAR	—	-4306.382	15	>1e+6	extreme
4	BYM2	—	-4321.256	16	>1e+6	extreme
6	ICAR	common distribution	-4337.488	4.7	110	extreme
5	ICAR	independent distributions	-4342.195	38000	>1e+6	extreme
2	—	common distribution	-42360.147	29	>1e+6	extreme
1	—	independent distributions	-42389.058	27000	>1e+6	extreme
non-spatial	—	—	-69501.304	—	—	—
Liberal Democrat						
7	ICAR	varying ICAR sd	-4713.785	25	>1e+6	extreme
3	ICAR	—	-4739.024	22	>1e+6	extreme
5	ICAR	independent distributions	-4760.673	230	>1e+6	extreme
6	ICAR	common distribution	-4991.153	84	>1e+6	extreme
4	BYM2	—	-5074.869	250000	>1e+6	extreme
2	—	common distribution	-258553.477	24	>1e+6	extreme
1	—	independent distributions	-258577.449	6e+05	>1e+6	extreme
non-spatial	—	—	-859667.026	—	—	—

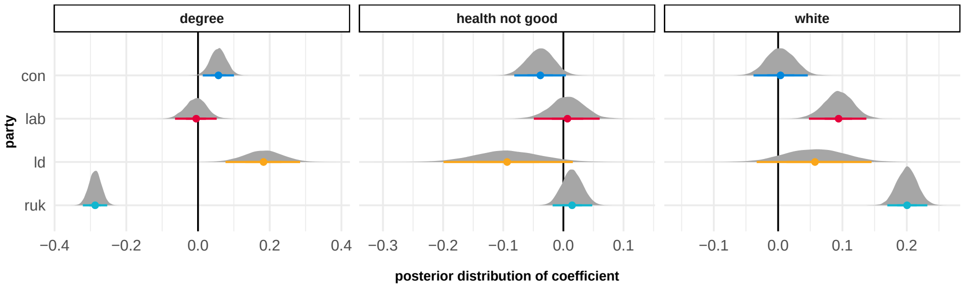


Figure 4.10: Covariate coefficient posterior distributions.

Posterior distributions of coefficients for socio-economic variables from Model 7 for Conservatives, Labour, Liberal Democrats and Reform.

4.6.3 Spatial estimates from optimal models

Moving to the spatial components, the mean posterior values of the ICAR component (which have regionally varying σ_k) are shown as maps (a-d) in row 1 of Figure 4.11. Differences in the degree of smoothness are clearly visible such as in the South West region for Labour, where strongly negative and positive values occur contiguously.

The second row (e-h) shows the mean posterior of the varying standard deviations mapped by region. These varying levels of spatial cohesion have been split into quintiles, ranging from those with high cohesion (where values in a given location are likely to be more similar to neighbouring values), to lower cohesion where the degree of similarity is reduced. These are quintiles of mean values of all parties in all regions. At the lower end of the scale, we have quintiles 1 and 2 in blue, which are 'most' and 'somewhat' cohesive respectively. Areas close to median spatial cohesion ('moderate') are shown in white, with more extreme values ('less' or 'least' cohesive) shown in orange.

Much of the Conservative vote pattern lies within the 'moderate' cohesion category. Exceptions to this are across southern regions where it is more cohesive and in Yorkshire and the Humber where neighbouring constituencies are not similar to the neighbours in terms of voting Conservative. Labour only shows 'moderate' levels of cohesion in London. Its cohesion pattern is split largely along a north-south divide with neighbouring constituencies showing similar behaviour in the north as opposed to the south.

The mean posterior ICAR σ_k values for the Liberal Democrats are all in the higher categories. Those of Reform UK, on the other hand, are all below the median level (with the exception of London), with most in fact falling into the category of 'most' cohesive. This is consistent with the notion of that party having a relatively consistent widespread level of support, although not quite to the extent where the first-past-the-post system would benefit them.

Plots (i)-(l) on row 3 focus on the 95% credible intervals rather than the means of

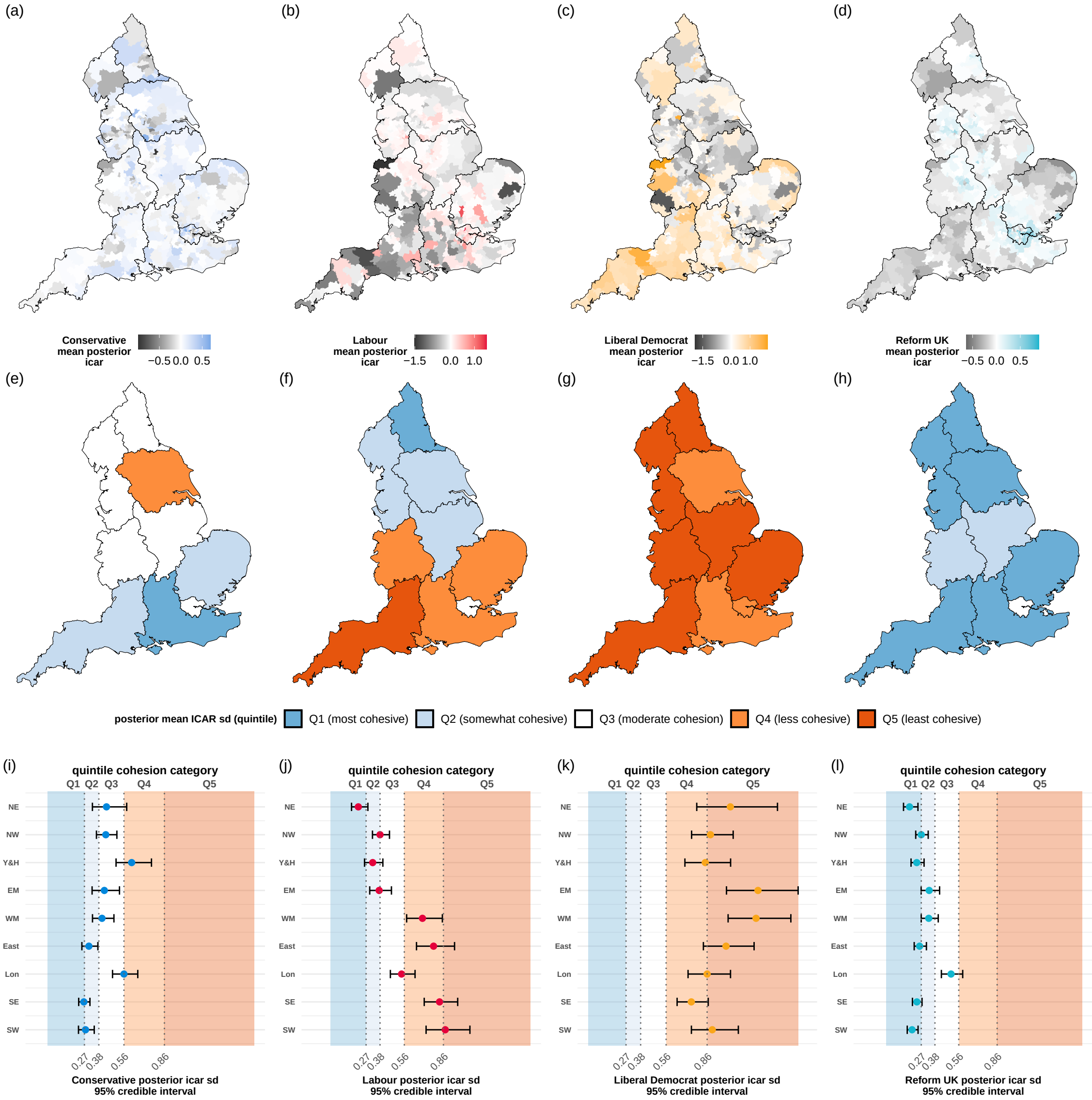


Figure 4.11: Spatial posterior components.

(a-d) Mean posterior of ICAR component of Model 7 for Conservatives, Labour, Liberal Democrats and Reform UK respectively, (e-h) mean posterior value of standard deviation component from Model 7 for each party, varying by region, and (i-l) 95% credible intervals of the posterior distributions of these varying standard deviations.

the posterior σ_k values, and place their ranges within the context of the overall quintile levels. These plots tell a similar story, showing not only higher values of σ_k for Liberal Democrat voters, but a greater degree of uncertainty as to their value.

The results of this model comparison, which show strong evidence of differences in the level of spatial cohesion of voting patterns for different parties, clearly raises the question of why this would be the case. We could hypothesise that it is connected to unaccounted-for strategic voting, regional organisation, identity, urban versus rural constituencies, or other factors. There is, however, no evidence for any of this in the data. Further research would be required to rigorously investigate this.

4.6.4 Exceedance probabilities

Rather than comparing posterior means of spatial cohesion across all parties using a global average variance, we can instead calculate exceedance probabilities [65] separately for each party. These represent the posterior probability that a region's variance σ_k exceeds the party-specific mean, indicating lower spatial cohesion relative to that party's typical pattern across England. This within-party approach allows us to identify regions with unusually weak spatial structure for each party in its own context. The resulting maps are shown in Figure 4.12.

They show patterns broadly similar to those in Figure 4.11 for Conservative, Labour and Reform UK voters. In the case of Liberal Democrat voters, they highlight both the East and West Midlands as regions with particularly uncohesive voting patterns, even relative to the already high σ_k values typical of Liberal Democrat support.

4.7 Conclusions

In this paper, we proposed a novel structure for describing spatial processes occurring simultaneously at different levels. Rather than combining hierarchical models with a spatially autocorrelated process at the lowest level, we instead allowed the higher level differences to be reflected within the ICAR component itself, reflecting varying spatial cohesion. We constructed these models using a Bayesian framework and demonstrated how we could judge if our novel structure was more appropriate than others using log marginal likelihoods and Bayes factors.

This technique was applied to voting data from the UK, based on a theory that the tendency for voters in one constituency to behave similarly to their neighbours might not be the same in all parts of England. Some regions might demonstrate greater spatial cohesion in terms of electoral preferences than others. We found evidence that such differences did exist and that this new model was an improvement over standard ICAR models for all four of the largest parties. We found that, at a national scale, votes for Reform

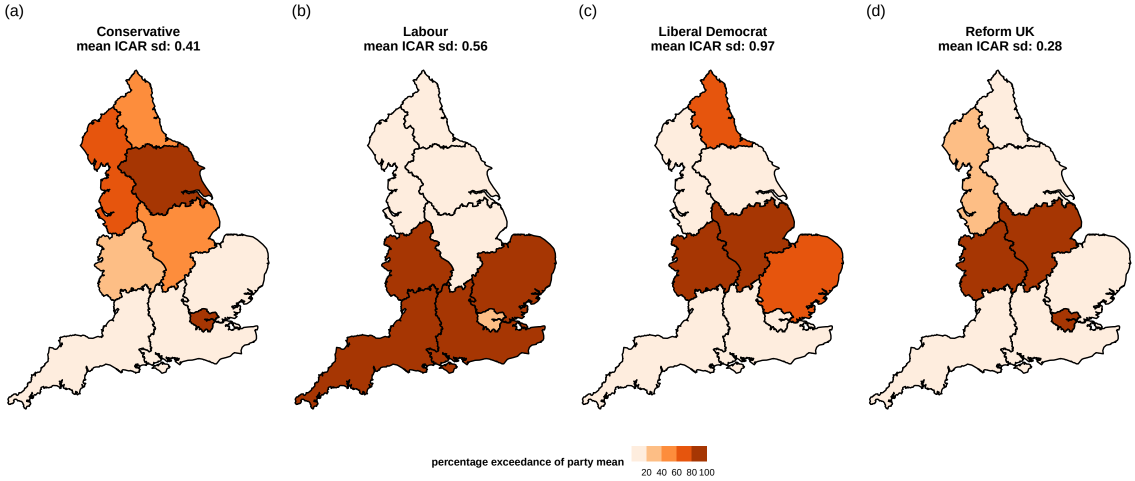


Figure 4.12: Exceedance probabilities by party and region.

Posterior exceedance by region, showing as percentages the posterior probability that a region's ICAR variance exceeds the party-specific mean (shown above map), indicating lower spatial cohesion relative to that party's typical pattern across England.

UK showed the highest degree of spatial cohesion between neighbouring constituencies while those for the Liberal Democrats showed the lowest. On a region by region basis, Conservative votes were more spatially cohesive in the south relative to the north, while the opposite was the case for Labour.

ICAR models with or without an additional hierarchical component are common in many areas of spatial analysis. This concept of allowing spatial cohesion to vary by region has wider application potential than just to voting data. In epidemiological modelling, for example, it might be desirable to allow the intensity of the spread of disease to be similarly variable by region. Rather than using region random effects, perhaps the process could be better captured by focussing only on the spatially smoothed ICAR component but allowing its degree of smoothness to vary. This idea could also be applied to studies of crime, social inequality, public health, environmental issues, ecology and housing markets. Within each of these areas, the meaning of the varying standard deviation of the ICAR structure would have its own interpretation.

It is this area of interpretation which may be one of the main limitations of this model. A combination of hierarchical effects and a spatially smooth process lends itself to a more standard and less complicated narrative than discussion of spatial cohesion. However, the explicit measure of spatial cohesion which this model provides may be of interest to test specific theories. Also, if the more complex model is deemed more plausible, this can raise questions about the simpler model's assumptions.

Another limitation which should be mentioned is the additional computational burden introduced by the varying σ_k parameter. Each model took several hours to fit on our hardware (MacBook Air, M1 chip, 8 GB RAM), with models incorporating varying σ_k requiring substantially longer runtimes. For this reason, we only sought to compare one instance of its use (in the form of Model 7) with a number of other candidate models, but there is scope for incorporating it within other similarly structured spatial frameworks as future work.

While our analysis of UK election data reveals distinct patterns of spatial cohesion among parties, understanding the underlying mechanisms driving these differences requires additional investigation beyond the scope of this work. The posterior estimates from our models do, however, provide a timely spatial analysis of this recent election, offering insights into geographic voting patterns that warrant further exploration.

Several extensions could strengthen this methodological framework. A more comprehensive simulation study, not conducted here due to computational constraints, would help establish the robustness of our approach across different scenarios and data structures, particularly examining performance under varying levels of spatial autocorrelation, strength of regional patterns, and sample sizes.

The regional variability framework introduced here could in the future, with access to more computational power, be extended to other spatial modelling components. For instance, the ρ parameter in BYM2 models, which governs the balance between structured and unstructured spatial processes, could itself vary spatially. This would allow different regions to exhibit distinct spatial dependence structures within that model's framework.

Another promising avenue for future work could involve imposing spatial structure on the varying ICAR precision parameters σ_k in our model. By smoothing these values across neighbouring regions, we could capture gradual transitions in spatial variance while preserving the model's ability to adapt to local patterns.

5 Conclusion

In this thesis, we have developed novel methods for modelling areal spatial data. More specifically, we have looked at data which can be seen as hierarchical in nature through aggregation of the component spatial units into blocks and sub-blocks of regions and subregions which are expected to share similar characteristics. We have sought to combine this approach to modelling with a concurrent autoregressive structure to capture expected similarities between neighbouring spatial units, regardless of which side of a hierarchical boundary they might lie on.

We have focused both on novel approaches to the specification of neighbourhood matrices for the autoregressive component, and on developments of the model structure itself to account for the influence of the spatial hierarchy within the autoregressive component itself.

In chapter 2, we presented an initial modelling structure which allowed for the combination of *vertical* hierarchical effects with *horizontal* autoregressive effects. This model was applied to voter behaviour in the 2019 UK General Election in England and Wales. It sought to model the degree of swing between the Conservative Party and Labour in each constituency, controlling for a number of socio-economic variables. The hierarchical structure saw constituencies nested within counties which were in turn nested within regions. This type of analysis allowed us to examine the proportion of variation in swing which was attributable to each level and to estimate spatially varying coefficients.

Chapter 3 examined the neighbourhood matrix portion of the model structure in more detail. It outlined the workings of an R package called `sfislands` which we developed to make it less complicated to accommodate islands (or other potential geographical issues) in autoregressive models. By examining the extreme case of Indonesia and its thousands of islands whose names were not familiar to the author, we showed how reasonable contiguity structures could be defined and subsequently examined with the help of easily-produced maps. We also examined a non-island situation where we might want to account for another type of potential spatial discontiguity between neighbouring units. The example chosen was London and the River Thames. We showed how the package could be used to create a contiguity matrix which satisfied the condition that spatial units separated by the river should be considered neighbours only if there was river-crossing infrastructure within a distance of 1 kilometre.

In chapter 4, we presented a novel model structure which combined *vertical* and *horizontal* effects but in a different way. Again, this model was demonstrated in the context of voter behaviour, this time from the 2024 UK General Election in England. The distinctive characteristic of this model was that it incorporated the *vertical* hierarchical component as a parameter within the *horizontal* autoregressive structure. This allowed the tendency of nearby spatial units to be similar to each other to vary according to a hierarchical structure. By comparison with a selection of candidate models which capture spatial variation in different ways, this novel method, allowing for variability of spatial cohesion, emerged as the most plausible.

5.1 Limitations

One of the key issues with this type of modelling is the nature of the geographical hierarchical structure. Specifically within the context of UK voter behaviour, but more broadly in other applications, it is not clear how meaningful the administrative divisions which serve as the definitions of the various levels necessarily are. Their delineation is complicated and based on historical associations which may no longer be especially relevant to the way that voters in any particular location behave. When used in conjunction with autoregressive effects, however, as is the case with these models, glaring misalignments can be identified in the form of strong clustering effects in the CAR component which straddle administrative boundaries.

The principal operational difficulty which we encountered was that, as additional complexities were added to the models discussed in chapter 4, convergence required a large and time-consuming number of iterations. While time issues and the size of resultant models are a well-known issue with CAR models in general, this made a more comprehensive simulation study unfeasible due to computational constraints.

5.2 Future work

Despite these difficulties with model size, a more thorough set of simulation studies in the future would be valuable and could seek to determine under what circumstances our novel model structure is best applied, and to examine how it performs under a wider range of scenarios.

There is also scope for further capabilities to be included in the `sfislands` package.

For example, a set of functions could be provided to automate a task such as in the London example in chapter 3 where a set of crossings (bridges, tunnels etc.) define whether or not a river or other such geographical feature should interrupt contiguity links between otherwise neighbouring spatial units. This process is demonstrated in a step-by-step fashion but it might be useful to simplify the procedure into one self-contained function.

The novel model structure in chapter 4, where a hierarchical component is allowed to feature as a parameter within the autoregressive structure, could also be developed further. Depending on context, it might be useful for these parameters themselves to be autocorrelated according to contiguity of the grouping structure. This would discourage sudden leaps in spatial cohesion across neighbouring regions and might be more reflective of a realistic spatial process.

To conclude, all proposed methods are freely available at `https://github.com/horankev` in the repositories named `swing_project`, `sfislands`, and `vary_cohesion`, for Chapters 2, 3 and 4, respectively. This ensures every analysis presented in this thesis is reproducible and methodologies are available to interested practitioners.

Bibliography

[1] S. Andréfouët, M. Paul, and A. R. Farhan. Indonesia's 13558 islands: A new census from space and a first step towards a one map for small islands policy. *Marine Policy*, 135:104848, 2022. URL `http://dx.doi.org/10.1016/j.marpol.2021.104848`.

[2] H. Bakka, H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, E. Krainski, D. Simpson, and F. Lindgren. Spatial modelling with r-inla: A review, 2018. URL `https://arxiv.org/abs/1802.06350`.

[3] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. URL `https://doi.org/10.18637/jss.v067.i01`.

[4] R. Beecham, A. Slingsby, and C. Brunsdon. Locally-varying explanations behind the United Kingdom's vote to leave the European Union. *Journal of Spatial Information Science*, 16:117–136, 2018. URL `https://doi.org/10.5311/JOSIS.2018.16.377`.

[5] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974. URL `http://dx.doi.org/10.1111/j.2517-6161.1974.tb00999.x`.

[6] J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746, 12 1995. URL `https://doi.org/10.1093/biomet/82.4.733`.

[7] J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–20, Mar. 1991. URL `http://dx.doi.org/10.1007/BF00116466`.

[8] R. S. Bivand. R packages for analyzing spatial data: A comparative case study with areal data. *Geographical Analysis*, 54(3):488–518, 2022. URL `https://doi.org/10.1111/gean.12319`.

[9] R. S. Bivand and V. Gómez-Rubio. Spatial survival modelling of business re-opening after Katrina: Survival modelling compared to spatial probit modelling of re-opening within 3, 6 or 12 months. *Statistical Modelling*, 21(1-2):137–160, 02 2021. URL `https://doi.org/10.1177/1471082X20967158`.

[10] R. S. Bivand and B. A. Portnov. Exploring spatial data analysis techniques using R: The case of observations with no neighbors. In L. Anselin, R. J. G. M. Florax, and S. J. Rey, editors, *Advances in Spatial Econometrics*, Advances in Spatial Science, chapter 6, pages 121–142. Springer, 2004. URL `https://ideas.repec.org/h/spr/adspcp/978-3-662-05617-2_6.html`.

[11] M. J. Brewer and A. J. Nolan. Variable smoothing in Bayesian intrinsic autoregressions. *Environmetrics*, 18(8):841–857, Mar. 2007. URL `http://dx.doi.org/10.1002/env.844`.

[12] Á. Briz-Redón, A. Iftimi, J. F. Correcher, J. De Andrés, M. Lozano, and C. Romero-García. A comparison of multiple neighborhood matrix specifications for spatio-temporal model fitting: A case study on COVID-19 data. *Stochastic Environmental Research and Risk Assessment*, 36(1):271–282, 2021. URL `http://dx.doi.org/10.1007/s00477-021-02077-y`.

[13] P.-C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. URL `http://dx.doi.org/10.18637/jss.v080.i01`.

[14] D. Butler and D. Stokes. *Political Change in Britain*. Palgrave Macmillan UK, 1974. URL `http://dx.doi.org/10.1007/978-1-349-02048-5`.

[15] D. Butler and S. D. Van Beek. Why not swing? Measuring electoral change. *PS: Political Science and Politics*, 23(2):178–184, 1990. URL `http://www.jstor.org/stable/420065`.

[16] B. E. Cain. Strategic voting in Britain. *American Journal of Political Science*, 22 (3):639, Aug. 1978. URL `http://dx.doi.org/10.2307/2110465`.

[17] E. Cox. Leaving the North behind led to Brexit. Here's what has to happen next. *The New Statesman*, 06 2016.

[18] J. Curtice and M. Steed. Proportionality and exaggeration in the British electoral system. *Electoral Studies*, 5(3):209–228, 12 1986. URL `https://www.sciencedirect.com/science/article/pii/0261379486900120`.

[19] G. Dong and R. Harris. Spatial autoregressive models for geographically hierarchical data structures. *Geographical Analysis*, 47(2):173–191, 2015. URL `https://doi.org/10.1111/gean.12049`.

[20] G. Dong, J. Ma, R. Harris, and G. Pryce. Spatial random slope multilevel modeling using multivariate conditional autoregressive models: A case study of subjective travel satisfaction in Beijing. *Annals of the American Association of Geographers*, 106(1):19–35, 01 2016. URL `https://doi.org/10.1080/00045608.2015.1094388`.

[21] D. Dorling. Persistent north-south divides. In N. M. Coe and A. Jones, editors, *The Economic Geography of the UK*, chapter 2, pages 12–28. Sage, London, 2010. URL `https://doi.org/10.4135/9781446269374.n2`.

[22] J. A. Dougenik, N. R. Chrisman, and D. R. Niemeyer. An algorithm to construct continuous area cartograms. *The Professional Geographer*, 37(1):75–81, 1985. URL `https://doi.org/10.1111/j.0033-0124.1985.00075.x`.

[23] E. W. Duncan, N. M. White, and K. Mengersen. Spatial smoothing in Bayesian models: A comparison of weights matrix specifications and their impact on inference. *International Journal of Health Geographics*, 16(1), 2017. URL `http://dx.doi.org/10.1186/s12942-017-0120-x`.

[24] A. Earnest, G. Morgan, K. Mengersen, L. Ryan, R. Summerhayes, and J. Beard. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models. *International Journal of Health Geographics*, 6(1):54, 2007. URL `http://dx.doi.org/10.1186/1476-072X-6-54`.

[25] A. S. Fotheringham and C. Brunsdon. Local forms of spatial analysis. *Geographical Analysis*, 31(4):340–358, 10 1999. URL `https://doi.org/10.1111/j.1538-4632.1999.tb00989.x`.

[26] A. Freni-Sterrantino, M. Ventrucci, and H. Rue. A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and Spatio-temporal Epidemiology*, 26:25–34, 2018. URL `http://dx.doi.org/10.1016/j.sste.2018.04.002`.

[27] R. C. Geary. The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5(3):115, 11 1954. URL `https://doi.org/10.2307/2986645`.

[28] H. Goldstein. *Multilevel models in education and social research.* Multilevel models in education and social research. Oxford University Press, New York, NY, US, 1987. Pages: viii, 98.

[29] H. Goldstein. Methods in school effectiveness research. *School Effectiveness and School Improvement*, 8(4):369–395, 12 1997. URL `https://doi.org/10.1080/0924345970080401`.

[30] M. Goodwin and O. Heath. The 2016 referendum, Brexit and the left behind: An aggregate-level analysis of the result. *The Political Quarterly*, (87):323–332, 2016.

[31] I. R. Gordon. In what sense left behind by globalisation? Looking for a less reductionist geography of the populist surge in Europe. *Cambridge Journal of Regions, Economy and Society*, 11(1):95–113, 03 2018. URL `https://doi.org/10.1093/cjres/rsx028`.

[32] J. Griffiths, S. Perrett, E. A. Fieldhouse, C. Prosser, J. Green, J. Mellon, J. Bailey, and G. Evans. The Brexit realignment amid electoral volatility: The role of party

blocs in the 2024 General Election. 2024. URL `https://dx.doi.org/10.2139/ssrn.5048763`.

[33] Q. F. Gronau and H. Singmann. bridgesampling: Bridge sampling for marginal likelihoods and Bayes factors, Mar. 2017. URL `http://dx.doi.org/10.32614/CRAN.package.bridgesampling`.

[34] K. Horan, C. Brunsdon, and K. Domijan. A multilevel spatial model to investigate voting behaviour in the 2019 UK General Election. *Applied Spatial Analysis and Policy*, 17(2):703–727, Jan. 2024. URL `http://dx.doi.org/10.1007/s12061-023-09563-6`.

[35] K. Horan, K. Domijan, and C. Brunsdon. *sfislands: Streamlines the process of fitting areal spatial models*, 2024. URL `http://dx.doi.org/10.32614/CRAN.package.sfislands`. R package version 1.1.2.

[36] House of Commons Library. Boundary changes: Current constituencies and new constituencies in the UK, 2023. URL `https://commonslibrary.parliament.uk/boundary-review-2023-which-seats-will-change/`. Accessed: 2025-01-28.

[37] House of Commons Library. Parliament elections data - data tools and resources, 2024. URL `https://commonslibrary.parliament.uk/data-tools-and-resources/parliament-elections-data/`. Accessed: 2025-01-28.

[38] E. Jack, D. Lee, and N. Dean. Estimating the changing nature of Scotland's health inequalities by using a multivariate spatiotemporal model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(3):1061–1080, Apr. 2019. URL `http://dx.doi.org/10.1111/rssa.12447`.

[39] H. Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

[40] W. Jennings, J. Furlong, G. Stoker, and L. Mckay. Fragmented and dealigned: The 2024 British General Election and the rise of place-based politics. *The Political Quarterly*, Dec. 2024. URL `http://dx.doi.org/10.1111/1467-923X.13483`.

[41] S. Jeworutzki. cartogram: Create cartograms with R. 2020. URL `https://CRAN.R-project.org/package=cartogram`.

[42] R. Johnston and C. Pattie. Using an entropy-maximizing procedure to estimate territorial social indicators: An introduction and illustration. *Social Indicators Research*, 27(3):235–256, Nov. 1992. URL `http://dx.doi.org/10.1007/BF00300463`.

[43] R. Johnston, D. Manley, C. Pattie, and K. Jones. Geographies of Brexit and its aftermath: Voting in England at the 2016 referendum and the 2017 General Election. *Space and Polity*, 22(2):162–187, 05 2018. URL `https://doi.org/10.1080/13562576.2018.1486349`.

[44] K. Jones. Specifying and estimating multi-level models for geographical research. *Transactions of the Institute of British Geographers*, 16(2):148–159, 1991. URL `https://doi.org/10.2307/622610`.

[45] K. Jones, M. I. Gould, and R. Watt. Multiple contexts as cross-classified models: The Labor vote in the British General Election of 1992. *Geographical Analysis*, 30 (1):65–93, 1998. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1998.tb00389.x`.

[46] J. Kanagasooriam and E. Simon. Red wall: The definitive description. *Political Insight*, 12(3):8–11, 09 2021. URL `https://doi.org/10.1177/20419058211045127`.

[47] A. Kassambara. *ggpubr: 'ggplot2'-based publication ready plots*, 2023. URL `https://CRAN.R-project.org/package=ggpubr`. R package version 0.6.0.

[48] D. G. Krige. Moving average surfaces for ore evaluation. *Joumd of the South Africun Institute of Mining and Metallurgy*, (66):13–38, 1966. URL `https://www.saimm.co.za/Conferences/DanieKrige/DGK10.pdf`.

[49] D. Lee, A. Rushworth, and S. K. Sahu. A Bayesian localized conditional autoregressive model for estimating the health effects of air pollution. *Biometrics*, 70(2): 419–429, Feb. 2014. URL `http://dx.doi.org/10.1111/biom.12156`.

[50] M. D. Lee and E.-J. Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, Apr. 2014. URL `http://dx.doi.org/10.1017/CBO9781139087759`.

[51] X.-L. Meng and W. H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.

[52] W. L. Miller. *Electoral Dynamics in Britain since 1918*. Palgrave Macmillan UK, 1977. URL `http://dx.doi.org/10.1007/978-1-349-15851-5`.

[53] M. Morris, K. Wheeler-Martin, D. Simpson, S. J. Mooney, A. Gelman, and C. DiMaggio. Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in Stan. *Spatial and Spatio-temporal Epidemiology*, 31:100301, Nov. 2019. URL `http://dx.doi.org/10.1016/j.sste.2019.100301`.

[54] D. Muller and L. Page. A new approach to measure tactical voting: evidence from the British elections. *Applied Economics*, 47(36):3839–3858, Mar. 2015. URL `http://dx.doi.org/10.1080/00036846.2015.1019037`.

[55] E. Odell. *parlitools: Tools for analysing UK politics in R*, 2017. URL `https://doi.org/10.5281/zenodo.591586`.

[56] Office for National Statistics. Census 2021 data - Nomis, 2021. URL `https://www.nomisweb.co.uk/sources/census_2021`. Accessed: 2025-01-28.

[57] J. Parry and D. H. Locke. *sfdep: Spatial dependence for simple features*, 2024. URL `https://sfdep.josiahparry.com`. R package version 0.2.4.

[58] C. Pattie and D. Cutts. Playing the system: Electoral bias in the 2024 UK General Election. *The Political Quarterly*, Oct. 2024. URL `http://dx.doi.org/10.1111/1467-923X.13471`.

[59] C. Pattie and R. Johnston. 'It's not like that round here': Region, economic evaluations and voting at the 1992 British General Election. *European Journal of Political Research*, 28(1):1–32, July 1995. URL `http://dx.doi.org/10.1111/j.1475-6765.1995.tb00485.x`.

[60] C. Pattie and R. Johnston. 'People who talk together vote together': An exploration of contextual effects in Great Britain. *Annals of the Association of American Geographers*, 90(1):41–66, 03 2000. URL `https://doi.org/10.1111/0004-5608.00183`.

[61] C. Pattie, R. Johnston, M. Schipper, and L. Potts. Are regions important in British elections? Valence politics and local economic contexts at the 2010 General Election. *Regional Studies*, 49(9):1561–1574, Nov. 2013. URL `http://dx.doi.org/10.1080/00343404.2013.847271`.

[62] E. Pebesma. Simple features for R: Standardized support for spatial vector data. *The R Journal*, 10(1):439–446, 2018. URL `https://doi.org/10.32614/RJ-2018-009`.

[63] J. Pinheiro, D. Bates, and R Core Team. *nlme: Linear and nonlinear mixed effects models*, 2023. URL `https://CRAN.R-project.org/package=nlme`. R package version 3.1-164.

[64] C. Prosser. Fragmentation revisited: The UK General Election of 2024. *West European Politics*, 48(6):1501–1513, Dec. 2024. URL `http://dx.doi.org/10.1080/01402382.2024.2430915`.

[65] S. Richardson, A. Thomson, N. Best, and P. Elliott. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112 (9):1016–1025, June 2004. URL http://dx.doi.org/10.1289/ehp.6740.

[66] A. Riebler, S. H. Sørbye, D. Simpson, and H. Rue. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165, Aug. 2016. URL http://dx.doi.org/10.1177/0962280216660421.

[67] D. Robinson, A. Hayes, and S. Couch. *broom: Convert Statistical Objects into Tidy Tibbles*, 2023. URL https://CRAN.R-project.org/package=broom.

[68] P. Rycroft. The December 2019 UK General Election: Reflections. *Revue Française de Civilisation Britannique. French Journal of British Studies*, XXV(3), 06 2020. URL https://journals.openedition.org/rfcb/5846.

[69] Stan Development Team. RStan: the R interface to Stan, 2024. URL https://mc-stan.org/. R package version 2.32.6.

[70] W. R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(sup1):234–240, 06 1970. URL https://doi.org/10.2307/143141.

[71] E. Uberoi and N. Baker. Electoral swing. 02 2023. URL https://commonslibrary.parliament.uk/research-briefings/sn02608/.

[72] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner. Rank-normalization, folding, and localization: An improved R-hat for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, 16(2), June 2021. URL http://dx.doi.org/10.1214/20-BA1221.

[73] M. Vranckx, T. Neyens, and C. Faes. Comparison of different software implementations for spatial disease mapping. *Spatial and Spatio-temporal Epidemiology*, 31: 100302, 11 2019. URL https://doi.org/10.1016/j.sste.2019.100302.

[74] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016. URL https://ggplot2.tidyverse.org.

[75] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani. Welcome to the tidyverse. *Journal of*

Open Source Software, 4(43):1686, 2019. URL https://doi.org/10.21105/joss.01686.

[76] Wikipedia. List of crossings of the River Thames — Wikipedia, the free encyclopedia, 2024. URL http://en.wikipedia.org/w/index.php?title=List%20of%20crossings%20of%20the%20River%20Thames&oldid=1184426738. Accessed 2024-03-15.

[77] S. N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. 73:3–36, 2011. URL https://CRAN.R-project.org/web/packages/mgcv/index.html.