



University  
of Glasgow

Anderson, C., Lee, D., and Dean, N. (2016) Spatial clustering of average risks and risk trends in Bayesian disease mapping. *Biometrical Journal*.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/122797/>

Deposited on: 30 August 2016

Enlighten – Research publications by members of the University of Glasgow  
<http://eprints.gla.ac.uk>

# Spatial clustering of average risks and risk trends in Bayesian disease mapping

Craig Anderson<sup>1\*</sup>, Duncan Lee<sup>2</sup>, Nema Dean<sup>2</sup>

<sup>1</sup> School of Mathematical Sciences, University of Technology, Sydney

<sup>2</sup> School of Mathematics and Statistics, University of Glasgow.

\*Email: craig.anderson@uts.edu.au

## Abstract

Spatio-temporal disease mapping focuses on estimating the spatial pattern in disease risk across a set of non-overlapping areal units over a fixed period of time. The key aim of such research is to identify areas which have a high average level of disease risk or where disease risk is increasing over time, thus allowing public health interventions to be focused on these areas. Such aims are well suited to the statistical approach of clustering, and while much research has been done in this area in a purely spatial setting, only a handful of approaches have focused on spatio-temporal clustering of disease risk. Therefore, this paper outlines a new modelling approach for clustering spatio-temporal disease risk data, by clustering areas based on both their mean risk levels and the behaviour of their temporal trends. The efficacy of the methodology is established by a simulation study, and is illustrated by a study of respiratory disease risk in Glasgow, Scotland.

**Keywords:** Clustering, CAR, Disease mapping, Spatio-temporal.

## 1 Introduction

The risks of almost all diseases vary in space for a number of reasons, including spatial changes in physical geography (e.g. temperature, sunlight, altitude), environmental factors (e.g. air quality, water quality), or the prevalence of risk inducing behaviours (e.g. poor diet, lack of exercise, smoking, alcohol consumption). Similarly, disease risk also varies in time as a result of temporal changes in the above factors, as well as due to changes in public health legislation, such as the ban on smoking in public places that was introduced in Scotland in 2005. It is therefore of great interest to public health professionals to estimate these spatial and temporal differences in disease risk, as it provides evidence for future public health policy. Spatial variation in disease risk is known as a health inequality (<sup>1,2</sup>), and interest lies in estimating the magnitude and spatial locations of this inequality so that interventions can be targeted at areas at greatest risk. The Equality Act (2010, <http://www.legislation.gov.uk/ukpga/2010/15/contents>) was introduced in the United Kingdom to reduce such inequalities, and thus there is keen interest in identifying whether these inequalities are getting wider or narrower over time. This requires the temporal trends in disease risk to be examined, to see whether areas initially at the lowest and highest risk are exhibiting increasing or decreasing trends over time.

The spatio-temporal pattern in disease risk is typically estimated at the population level, by partitioning the geographical region into  $n$  non-overlapping areal units, such as local council areas or electoral wards, and recording disease incidence at regular, such as yearly, time intervals. This population level approach is necessary because using individual level data would breach patient confidentiality, and because it allows public health professionals to look at the population as a whole. Counts of the numbers of disease cases in each areal unit and time period are then collected, and can be standardised by population demographic data to estimate the spatio-temporal pattern in disease risk which

can then be mapped. However, such standardised incidence ratios are known to be unstable <sup>(3)</sup>, and Poisson log-linear models are typically used to estimate the spatio-temporal pattern in disease risk using covariates and/or a set of random effects. The latter are included to account for spatio-temporal autocorrelation in the risk not captured by the covariates, and are typically modelled by a conditional autoregressive (CAR) prior <sup>(45)</sup> or spatio-temporal extensions thereof <sup>(6,78)</sup>.

These priors are based on the assumption that all pairs of random effects in geographically adjacent areal units are autocorrelated, which leads to spatially smooth mean levels and temporal trends in disease risk. However, such smoothing is contrary to the goal of identifying clusters of areal units that exhibit vastly different risk behaviour compared to their surrounding areas. Such differences in risk behaviour could be in an area’s mean level over time or its temporal trend, and the ability to cluster (group together) areas that exhibit similar characteristics has been the subject of work by <sup>9,10,11,12,13</sup>. However, all of the above research has focused on clustering purely spatial data by their mean risk levels, with the aim of identifying areas that exhibit very high disease risks. A range of spatio-temporal disease models have now been produced <sup>6,7,14,15,16,8</sup> but only <sup>17,18</sup> and <sup>19</sup> have focused on clustering in the spatio-temporal domain, focusing on detecting changing mean risk levels, shared latent structures and areas exhibiting unusual temporal trends respectively.

Therefore, this paper outlines a new modelling approach for clustering spatio-temporal disease risk data, by clustering areas based on both their mean risk levels and the behaviour of their temporal trends. The model proposed is an extension of the spatially varying intercept and linear trends model <sup>(6)</sup>, because in common with the majority of the existing literature our motivating application has data for a relatively small number of time periods ( $T = 10$  here) making it unwise to use more complex temporal trends such as smoothing splines. Our model contains two separate clustering components, the first is based on the average risk (intercept) for each area over time, while the second is based on the temporal trend (slope) in disease risk. This clustering is achieved in the same model that estimates disease risk, and this all-at-once approach is compared against a simpler two-stage approach in this paper. Our methodology is motivated by a study of respiratory disease risk in Glasgow, Scotland between 2002 and 2011, with particular interest in identifying potential changes in inequality over this period. The remainder of this paper is organised as follows. Section 2 gives a brief introduction to Bayesian disease mapping, and discusses a range of existing spatio-temporal modelling approaches. Section 3 proposes our new methodology, while Section 4 establishes its efficacy via simulation. Section 5 presents the motivating application for our methodology, while, Section 6 discusses the implications of this paper and ideas for future work.

## 2 Bayesian spatio-temporal disease mapping

The study region  $\mathcal{A}$  is partitioned into  $n$  non-overlapping areal units  $\mathcal{A} = \{\mathcal{A}_1, \dots, \mathcal{A}_n\}$ , and data about these areal units are collected for  $t = 1, \dots, T$  consecutive time periods. The response data take the form  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$ , where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})$  and  $Y_{it}$  represents the number of disease cases (hospital admissions with a primary diagnosis of respiratory disease in our application) in areal unit  $i$  during time period  $t$ . Covariate information may also exist, and is given by  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ , where  $\mathbf{X}_i = (x_{i1}, \dots, x_{iT})$ . The numbers of observed cases will depend on the size and demographic structure of the population at risk in each areal unit and time period, and this is accounted for by computing the expected numbers of disease cases based on standardised age and sex specific disease rates. These expected values are denoted by  $\mathbf{E} = (\mathbf{E}_1, \dots, \mathbf{E}_n)$ , where  $E_{it}$  is the expected number of disease cases for area  $i$  and time period  $t$ . A Poisson model with the general form  $Y_{it} \sim \text{Poisson}(E_{it}R_{it})$  is commonly used to model these data because  $Y_{it}$  is a count, and here  $R_{it}$  is the risk of disease in areal unit  $i$  during time period  $t$ . Based on this simple model the maximum likelihood estimator for  $R_{it}$  is  $\hat{R}_{it} = Y_{it}/E_{it}$ , which is known as the standardised incidence ratio (SIR). However, this ratio can lead to unstable estimates of  $R_{it}$ , especially if the disease is rare or the population at risk is small ( $E_{it}$  is small), so log-linear random effects models are used to collectively estimate disease risk in space and

time.

A number of spatio-temporal random effects models have been proposed to represent  $R_{it}$ , and our methodology in Section 3 extends the spatially varying intercept and linear trends model proposed by,<sup>6</sup> which is given by

$$\begin{aligned} Y_{it}|E_{it}, R_{it} &\sim \text{Poisson}(E_{it}R_{it}) & i = 1, \dots, n, \quad t = 1, \dots, T, \\ \ln(R_{it}) &= x_{it}^T \mathbf{B}_{it} + (\alpha + \phi_i) + (\beta + \delta_i)(t - \bar{t}), \\ \alpha, \delta &\sim \text{N}(0, 1000), \end{aligned} \quad (1)$$

where  $\bar{t} = (1/T) \sum_{t=1}^T t$ . Here the average risk over time for area  $i$  is  $\alpha + \phi_i$ , where  $\alpha$  is a global intercept term common to all areas while  $\phi_i$  is an area-specific adjustment to this global level. Similarly,  $\beta + \delta_i$  is the slope of the linear temporal trend for area  $i$ , which is decomposed into an average slope  $\beta$  and an area specific adjustment  $\delta_i$ . Prior distributions are specified for each of these components, and  $\alpha$  and  $\beta$  are each represented as fixed effects, with a diffuse Gaussian prior distribution with a large (on the log scale) variance. The two sets of spatial random effects  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$  and  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$  are assumed to be spatially autocorrelated, so that pairs of nearby areas have similar average risk levels and temporal trends. This was achieved in<sup>6</sup> by specifying conditional autoregressive priors for both  $(\boldsymbol{\phi}, \boldsymbol{\delta})$ , which induce spatial autocorrelation via a binary neighbourhood matrix  $\mathbf{W}$ . Elements of this matrix  $w_{ij} = 1$  if areal units  $(\mathcal{A}_i, \mathcal{A}_j)$  share a common border (denoted  $i \sim j$ ) and  $w_{ij} = 0$  otherwise, with  $w_{ii} = 0$  for all  $i$ . CAR priors can be specified as a set of  $n$  univariate conditional distributions, which for  $\boldsymbol{\phi}$  have the form  $f(\phi_i|\boldsymbol{\phi}_{-i})$ , where  $\boldsymbol{\phi}_{-i} = (\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n)$ . The CAR prior used in this context was the intrinsic CAR model<sup>(4)</sup>, which is given by

$$\phi_i|\boldsymbol{\phi}_{-i}, \mathbf{W} \sim \text{N}\left(\frac{\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{1}{\tau(\sum_{j=1}^n w_{ij})}\right) \quad i = 1, \dots, n, \quad (2)$$

where  $\tau$  is a conditional precision parameter. The same prior is used for  $\boldsymbol{\delta}$ , except that it has its own precision parameter  $\sigma$ . The conditional expectation of  $\phi_i$  is the mean of the random effects in neighbouring areal units, while the variance is inversely proportional to the number of neighbouring units. This set of conditional distributions correspond to a multivariate Gaussian distribution, with mean zero but an improper precision matrix given by  $\mathbf{Q} = \tau(\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W})$ , where  $\mathbf{W}\mathbf{1}$  is a vector containing the number of neighbours for each areal unit. One drawback of this model is the lack of a parameter to control the strength of the spatial autocorrelation. This means that the intrinsic model is only sensible in cases where the spatial autocorrelation in the data is strong; it is not sensible for cases where there is weak or moderate spatial autocorrelation across the study region because the model would tend to produce an overly smooth estimated risk surface in these cases. Therefore an extension to this intrinsic CAR model was proposed by,<sup>5</sup> and this Leroux CAR prior is given by

$$\phi_i|\boldsymbol{\phi}_{-i}, \mathbf{W} \sim \text{N}\left(\frac{\rho \sum_{j=1}^n w_{ij}\phi_j}{\rho \sum_{j=1}^n w_{ij} + 1 - \rho}, \frac{1}{\tau(\rho \sum_{j=1}^n w_{ij} + 1 - \rho)}\right) \quad i = 1, \dots, n, \quad (3)$$

where  $\rho$  controls the level of spatial autocorrelation. If  $\rho = 1$  then this corresponds to the intrinsic CAR model outlined above, while  $\rho = 0$  corresponds to complete spatial independence. This set of conditional distributions correspond to a multivariate Gaussian distribution, with mean zero and an improper precision matrix (assuming  $\rho \in [0, 1]$ ) given by  $\mathbf{Q} = \tau\{\rho[\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}] + (1 - \rho)\mathbf{I}\}$ , where  $\mathbf{I}$  is an  $n \times n$  identity matrix.

The linear time trends assumed in the Bernardinelli model were proposed for simplicity when the number of time periods  $T$  is small (in their original paper  $T = 4$ ), as more flexible time trends would

be better suited to a longer time series. A number of extensions to<sup>6</sup> (without any clustering) have been proposed in the literature to date, and a brief review is given below. The linearity of the time trends have been relaxed by replacing them with smooth functions<sup>(14)</sup> modelled by B-splines<sup>(20)</sup>, where a fixed effect is used to model the global temporal trend while random effects are used to model the localised trends for individual areal units. Inference for this model is carried out via penalised quasi-likelihood<sup>(21)</sup>, though the authors note that this approach is not ideal in terms of estimating model uncertainty. This is addressed by<sup>22</sup> who compares a number of estimators of prediction error to account for uncertainty, and a bootstrap adjusted Empirical Bayes variance estimator<sup>(23)</sup> is recommended. Smooth functions using P-splines<sup>(16)</sup> have been used instead of B-splines, while autoregressive time series models<sup>(15)</sup> have also been proposed to model the area specific temporal dynamics. The respiratory admissions data that motivated this paper have just  $T = 10$  time periods. We believe that while one could argue for the use of quadratics or other low level polynomials for modelling trends with  $T = 10$ , the increased number of parameters for such models would render the clustering aspect of our model difficult. As a result, the clustering model we propose in the next section will be an extension of the linear trends approach outlined by Bernardinelli et al.

Clustering within a spatio-temporal context has received little attention to date, and to our knowledge no research has attempted to cluster areas based on their temporal trends and mean risk levels as is proposed in the next section. The first attempt at spatio-temporal clustering was based on a mixture of Poisson distributions<sup>(17)</sup>, which aimed to identify areas that changed risk levels over time. More recently, work has focused on clustering areas based on the similarity of their latent structures<sup>(18)</sup>, as well as a modelling tool for outlier detection<sup>(19)</sup> that identifies areas that exhibit unusual temporal trends compared to the rest of the study region.

### 3 Methodology

#### 3.1 Rationale

Model (1) represents the intercept and slope for the  $i$ th areal unit as  $(\alpha + \phi_i)$  and  $(\beta + \delta_i)$  respectively, and as the random effects  $(\phi, \delta)$  are spatially autocorrelated, then adjacent areal units are forced to have similar trends and average levels of disease risk. This specification is restrictive for two reasons, the first of which is that two adjacent areal units may have very different average risk levels or temporal trends, and wrongly forcing them to be similar will likely result in poorer estimation of the spatio-temporal pattern in disease risk. The second limitation with (1) is that it does not have an inbuilt clustering mechanism, that allows areal units to be grouped (or clustered) together based on the similarity of their mean risk levels and temporal trends. Therefore we propose an extension of (1) that clusters the  $n$  areal units in two ways, into  $N_C$  groups based on their average risk levels, and  $N_D$  groups based on their temporal trends, and the model specification is outlined below. Areal units are allocated to the intercept and slope clusters independently; it is possible for two areal units to lie in the same intercept cluster, but in different slope clusters, and vice versa. For example, consider areal unit  $\mathcal{A}_i$  where the level of disease risk is high on average and is increasing over time, and areal unit  $\mathcal{A}_j$  where the level of disease risk is high on average but is decreasing over time. Both areal units would be in the same intercept cluster on account of having high average disease risk, but would be in different slope clusters due to one having an increasing risk and the other having a decreasing risk. It would be possible to cluster intercept and slope together, for example via Dirichlet process mixtures<sup>(24)</sup>, but we believe that our simpler method is sufficient for the level of broad classification which we desire in our application.

#### 3.2 Proposed model

Our proposed model replaces the global risk level  $\alpha$  and temporal trend  $\beta$  in (1) with cluster-specific fixed effects,  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{N_C})$  and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{N_D})$ , where  $N_C$  and  $N_D$  are the number of clusters

for the intercept and slope respectively. These components are piecewise constant across the  $n$  areal units, and if two areas have the same  $(\alpha_j, \beta_j)$  values then they are in the same cluster in terms of their average risk level ( $\alpha$ ) and temporal trend ( $\beta$ ) respectively. This results in a model of the form

$$\begin{aligned} Y_{it}|E_{it}, R_{it} &\sim \text{Poisson}(E_{it}R_{it}) & i = 1, \dots, n, \quad t = 1, \dots, T, \\ \ln(R_{it}) &= x_{it}^T \mathbf{B}_{it} + \alpha_{C_i} + \phi_i + [\beta_{D_i} + \delta_i](t - \bar{t}), \end{aligned} \quad (4)$$

where the  $i$ th areal unit has an average risk level of  $\alpha_{C_i} + \phi_i$  and a slope of  $\beta_{D_i} + \delta_i$  for its linear time trend. Both sets of random effects  $(\phi, \delta)$  are modelled using (3), and their inclusion in the model allows two areal units in the same intercept and slope clusters, i.e. with the same  $(C_i, D_i)$  values, to have similar but not identical intercepts and slopes. It also means that adjacent areal units in the same groups  $(\alpha_j, \beta_j)$  have similar average risks and temporal trends, which respects the spatial structure of these data.

The piecewise constant fixed effects  $\alpha = (\alpha_1, \dots, \alpha_{N_C})$  and  $\beta = (\beta_1, \dots, \beta_{N_D})$  represent the mean intercept and slope for areal units in each of the  $N_C$  and  $N_D$  groups, and are assigned the following uniform priors:

$$\begin{aligned} \alpha_j &\sim \text{Uniform}(\alpha_{j-1}, \alpha_{j+1}) & j = 1, \dots, N_C, \\ \beta_j &\sim \text{Uniform}(\beta_{j-1}, \beta_{j+1}) & j = 1, \dots, N_D, \end{aligned} \quad (5)$$

where  $\alpha_0 = \beta_0 = -\infty$  and  $\alpha_{N_C+1} = \beta_{N_D+1} = \infty$ . These ordering constraints  $\alpha_{j-1} < \alpha_j < \alpha_{j+1}$  and  $\beta_{j-1} < \beta_j < \beta_{j+1}$  are enforced to mitigate against the label switching problem <sup>(25)</sup>, and a move from cluster  $\alpha_j$  to cluster  $\alpha_{j+1}$  thus represents an increase in average risk level. The variables  $\mathbf{C} = (C_1, \dots, C_n)$  and  $\mathbf{D} = (D_1, \dots, D_n)$  allocate each areal unit to an intercept and slope cluster, and  $C_i \in \{1, \dots, N_C\}$  while  $D_i \in \{1, \dots, N_D\}$ . Initially, equal prior probabilities,  $P(C_i = c) = \frac{1}{N_C}$  and  $P(D_i = d) = \frac{1}{N_D}$ , were considered for these indicators, but such an approach means that the choice of the number of clusters  $(N_C, N_D)$  will drastically affect the results. Therefore we propose a penalty prior for each of  $(\mathbf{C}, \mathbf{D})$ , which is similar in spirit to that used in penalised splines <sup>(26)</sup>. The priors we propose are given by

$$\begin{aligned} P(C_i = c) &= \frac{\exp(-\theta_C(c - \bar{C})^2)}{\sum_{j=1}^{N_C} \exp(-\theta_C(j - \bar{C})^2)} & c = 1, \dots, N_C, \\ P(D_i = d) &= \frac{\exp(-\theta_D(d - \bar{D})^2)}{\sum_{j=1}^{N_D} \exp(-\theta_D(j - \bar{D})^2)} & d = 1, \dots, N_D, \end{aligned} \quad (6)$$

where  $\bar{C} = \frac{N_C+1}{2}$  when  $N_C$  is odd, and  $\bar{C} = \frac{N_C}{2}$  when  $N_C$  is even, and likewise  $\bar{D} = \frac{N_D+1}{2}$  when  $N_D$  is odd and  $\bar{D} = \frac{N_D}{2}$  when  $N_D$  is even. Thus  $(\bar{C}, \bar{D})$  represent the middle cluster in terms of intercept and slope, and the prior (6) penalises the cluster indicators  $(C_i, D_i)$  towards these middle groups. As a result higher prior weight is given to the central clusters compared with the extreme ones, to ensure that areal units only move to extreme high or low clusters if the data support it. The amount of shrinkage undertaken by each prior is controlled by  $\theta_C$  and  $\theta_D$ , with larger values meaning higher weighting is assigned to the central clusters. It should be noted that  $N_C$  and  $N_D$  represent the maximum number of clusters allowed for the intercept and slope respectively, because it is possible for clusters to be empty (cluster  $c$  is empty if, for every areal unit  $i$ ,  $C_i \neq c$ ). Here we recommend setting  $(N_C, N_D)$  equal to a small positive value such as 3 or 5, as this allows areas to be split into a small number of groups which makes interpretation easier. For example, if three groups are chosen in each direction then this will likely classify areas as having low / average / high risks and decreasing / constant / increasing trends, which is simple to interpret epidemiologically. However, the simulation

study presented in the supplementary material assesses model sensitivity to this choice. Finally, the hyperparameters of this model are outlined as follows:

$$\begin{aligned}\tau, \sigma &\sim \text{Gamma}(\gamma, \psi), \\ \rho, \lambda &\sim \text{Uniform}(0, 1), \\ \theta_C, \theta_D &\sim \text{Uniform}(1, 100).\end{aligned}$$

Here,  $\tau$  and  $\sigma$  are the precision hyperparameters for the intercept and slope random effects  $(\phi, \delta)$  respectively, and within this paper we set  $\gamma = 0.01$  and  $\psi = 0.01$ . A small sensitivity analysis relating to this choice of gamma prior is included in the supplementary material. The hyperparameters  $\rho$  and  $\lambda$  control the level of spatial autocorrelation within the intercept and slope random effects  $(\phi, \delta)$ , and are assigned uniform priors on the unit interval. As discussed above,  $\theta_C$  and  $\theta_D$  control the level of weighting towards the central clusters, and a uniform prior on a large range is specified. The lower bound of the Uniform distribution for these terms is chosen to be 1 rather than 0 since a value of  $\theta_C = 1$  corresponds to standard exponential decay. This is consistent with our prior belief that extra weight should be given to central clusters, and that areal units should only move to extreme clusters if the data provides substantial support for such a move.

Inference for this model was carried out using Markov-chain Monte Carlo sampling, based on a combination of Metropolis-Hastings and Gibbs sampling steps. All parameters were estimated using their posterior median, with the exception of the clustering parameters  $C_i$  and  $D_i$  which were based on the posterior mode. We can quantify the uncertainty surrounding each parameter estimate by constructing a credible interval based on the 2.5th and 97.5th percentiles of the posterior distribution. Software to run the model is available as part of the online supplementary material accompanying this paper. It can also be accessed at <https://bitbucket.org/craiganderson1989/model-code>

## 4 Simulation study

We present a simulation study to establish the efficacy of the Bayesian spatio-temporal clustering model outlined in the previous section. The template for this study is the set of 271 Intermediate Geographies comprising the Greater Glasgow and Clyde Health Board for a period of 10 years, which is the study region and time frame for the motivating application presented in Section 5. In this study we compare our clustering model proposed in Section 3 to the existing spatially varying intercept and linear trends model of,<sup>6</sup> given by (1). However, this latter model does not have an inbuilt clustering mechanism, so it is combined with a post-hoc clustering method based on a mixture model<sup>(27)</sup>. This clustering approach was carried out using the *mclust* R package<sup>(28)</sup>. This clustering method was applied with a maximum of  $N_C \times N_D = 25$  clusters, which matches that allowed in the model proposed in Section 3. Additionally, a sensitivity analysis was carried out to compare the performances of the methods with different choices of  $N_C$  and  $N_D$ , and this study is included in the supplementary material.

### 4.1 Data Generation

Each of the  $n = 271$  areal units in the Greater Glasgow and Clyde Health Board were assigned to an intercept cluster and a slope cluster separately, and in each case 3 groups were used to generate the data with  $\alpha = \beta = (-1, 0, 1)$  which are on the log-risk scale. These clusters have been designed so that every possible combination of the nine intercept and slope clusters is observed, and the template for this cluster structure is shown in Figure 1. In this figure the intercept levels are represented by the background colours of the areal units, where a darker shade of grey corresponds to a higher average disease risk over the study period. The slope levels are represented by the hatching, with the areal units with decreasing slopes represented by ‘downward’ hatches which go from top left to

bottom right, and the areal units with increasing slopes represented by ‘upward’ hatches which go from bottom left to top right. The base cluster means  $\boldsymbol{\alpha} = \boldsymbol{\beta} = (-1, 0, 1)$  are multiplied by a scalar  $Z$ , which varies the magnitude of the differences between the clusters. In this study three scenarios are considered. Scenario one sets  $Z = 1$  and corresponds to a case where there are large differences between the clusters, scenario two has  $Z = 0.5$  and corresponds to a more difficult case where there are smaller differences and  $Z = 0$  corresponds to a spatially smooth risk surface with no change over time where one would hope to identify a single cluster covering the entire study region.

Disease counts were generated for ten time points from a Poisson distribution with mean  $E_{it}R_{it}$ , where  $E_{it}$  was set to be equal to 100 for all  $i$  and  $t$ . The log risk surface was generated with separate linear time trends for each areal unit, where the set of  $n$  slopes and intercepts were generated from multivariate Gaussian distributions with a common spatially correlated precision matrix, given by  $\mathbf{Q} = (\text{diag}(\mathbf{W}\mathbf{1}) - \mathbf{W}) + \epsilon\mathbf{I}$ , which corresponds to the intrinsic CAR model with a small constant  $\epsilon = 0.001$  added to the diagonal to ensure  $\mathbf{Q}$  is diagonally dominant and hence invertible. Spatial clustering in the intercepts and slopes was obtained through a piecewise constant mean function for  $\phi$  and  $\delta$ , which follows the template shown in Figure 1.

Two hundred datasets were generated for each of the three data generation approaches ( $Z = 0, 0.5, 1$ ), which correspond to  $N_C = N_D = 1$  for  $Z = 0$  and  $N_C = N_D = 3$  for  $Z = 0.5, 1$ . However, in practice the true values of  $(N_C, N_D)$  are unknown, so we set  $N_C = N_D = 5$  to allow the possibility of the methods overestimating the number of clusters in the simulated data. As discussed in Section 3 our model may produce empty clusters, so selecting values of  $N_C$  and  $N_D$  which are larger than the true number of clusters does not mean that the correct cluster structure cannot be estimated. In the supplementary material accompanying this paper we tested our model under different values of  $N_C$  and  $N_D$  to investigate how reliant the model is on the user’s prior choice of these values, and the results obtained are consistent across all values tested. In all cases inference was based on 10,000 MCMC samples, of which 5,000 were discarded as burn-in.

## 4.2 Results

The results of the study are summarised in Figure 2, which displays a comparison of the relative performances of our approach and the Bernardinelli model with post-hoc clustering using three different metrics. The accuracy of the risk surfaces estimated by each approach is quantified by their root mean square error (RMSE), while the correctness of the estimated cluster structures is quantified by both the number of clusters identified and the Rand Index <sup>(29)</sup> between the true and estimated cluster structures. The latter is a measure of the similarity between two cluster structures and lies in the interval  $[0, 1]$ . It is computed as the proportion of pairs of areal units classified either in the same or in different clusters by both methods, that is the proportion of pairwise agreements between the two methods. A value of 1 indicates complete agreement between the two cluster configurations and a value of 0 indicates that no pair of areal units are classified in the same way under both configurations.

The top panel of Figure 2 shows boxplots of the number of combined slope-intercept clusters estimated under each approach in the 200 simulated data sets, where the true values of 1 (when  $Z = 0$ ) and 9 (when  $Z = 0.5, 1$ ) are represented by dashed lines. Our approach performs better than the Bernardinelli model for all three values of  $Z$ . When  $Z = 0$ , both models obtain a median of 1 cluster, but our approach has a standard deviation of 0.39 compared to 0.80 for the Bernardinelli model, while for  $Z = 0.5$ , our model obtains a median of 9 clusters and the Bernardinelli model underestimates the number of clusters, with a median of 4. For  $Z = 1$ , our model slightly overestimates the number of clusters for  $Z = 1$ , with a median of 11 clusters, while the Bernardinelli approach underestimates the number of clusters with a median of 6. The overestimation by our model is generally a result of the model partitioning a true high or low intercept or slope cluster into more than one group, or by the model placing a single outlier in a cluster on its own. This is likely to be less damaging than underestimating

the number of clusters, which will result in multiple true clusters being joined together and their estimated values smoothed towards each other. The effect of the overestimation and underestimation of clusters by the two approaches can be investigated by comparing the RMSE values as described below.

The middle panel of Figure 2 displays the Rand index values obtained under each approach. Again, our model performs better than the Bernardinelli approach in each case. For  $Z = 0$ , both models have a median Rand index of 1, but our approach has a much lower standard deviation of 0.005 compared to 0.20 for the Bernardinelli approach. The Bernardinelli model performs very poorly in some cases under  $Z = 0$ , with a Rand index as low as 0.32 obtained in one case; this makes a two-stage approach to risk estimation and clustering unreliable for clustering, because it identified a number of false positives clusters that do not exist. Thus, if this approach was applied to real data where the true cluster structure was unknown, then the possibility of a false positive would lead to doubt over the veracity of the results. For  $Z = 0.5$ , our model obtains a median Rand index of 0.75, while the Bernardinelli model obtains a median of 0.63. For  $Z = 1$ , both models have very high Rand index values; the median for our model is 0.91 compared to 0.92 for the Bernardinelli model. However, our model performs more consistently, with a standard deviation of 0.02 compared with 0.09 for the Bernardinelli model.

The bottom panel of Figure 2 displays boxplots of the root mean square error of the estimated risk surface obtained under each approach in the simulation study. Once more, our model provides more accurate risk estimates than the Bernardinelli approach. For  $Z = 0$ , both approaches have very similar results; our approach has a median of 0.025 compared to 0.024 for the Bernardinelli model, while our approach has a slightly higher standard deviation of 0.004 compared to 0.003 for Bernardinelli. When  $Z = 0.5$ , our approach has a median RMSE of 0.103 compared with 0.116 for the Bernardinelli model, and for  $Z = 1$  a median of 0.097 is obtained for our proposed model compared with 0.144 for the Bernardinelli model. We can see that both models produce similar risk estimates when the risk surface is smooth in space in terms of both intercept and slope ( $Z = 0$ ), but our model performs better in cases where clusters exist. This is unsurprising, since our model allows for different fixed effects for each cluster, while the Bernardinelli approach has fixed effects (one for intercept and one for slope) that are common to all areas.

## 5 Application to real data

### 5.1 Data description and study design

The spatio-temporal clustering model was motivated by a study of the changing nature of the respiratory hospitalisation risk in Glasgow, Scotland between 2002 and 2011. The study region is the Greater Glasgow and Clyde Health Board region, which contains the city of Glasgow in the east and the river Clyde estuary in the west. Glasgow is the largest city in Scotland, with a population of around 600,000 people. The health board is split into  $n = 271$  administrative units known as Intermediate Geographies (IGs), containing populations of between 2,244 and 10,877 people with a median value of 4,239. The disease data were obtained from the Scottish Neighbourhood Statistics database (<http://www.sns.gov.uk>), and consist of counts of the yearly numbers of respiratory admissions to hospital for each of the 271 areal units between 2002 and 2011.

The expected numbers of admissions were calculated for each areal unit and year, based on standardised age and sex specific disease rates. The average risks and temporal trends in the raw standard incidence ratios are summarised in Figure 3. The simple model  $Y_{it} \sim \text{Poisson}(E_{it}R_{it})$ ,  $\ln(R_{it}) = \alpha + \beta(t - \bar{t})$  was fitted separately for each areal unit  $i$ , and the top panel of the figure shows the estimated average  $\alpha$  while the bottom panel shows the estimated linear trend  $\beta$ . The top panel shows that the areas with the highest average risk can be found in the East End of the city, which is known to be an

area of high deprivation. The bottom panel shows that the risk of respiratory admissions is generally remaining stable or even decreasing in these areas, while increases can be identified in a number of areas to the outskirts of the city, including rural Dunbartonshire to the north-west and Eaglesham to the far south-east.

## 5.2 Results

The spatio-temporal clustering model outlined in Section 3 was then applied to these data with  $N_C = N_D = 5$ , which was chosen to allow for a possible distinction between high (and low) risk and extremely high (and low) risk intercept clusters, and also for different magnitudes of increasing and decreasing risk trends in each direction. No covariates were included in the model, because the goal of the analysis is to identify clusters in the disease risk surface, not in the residual surface after adjusting for covariate factors. Inference was based on 10,000 MCMC samples, of which 5,000 were discarded as burn-in. The result from fitting this model is that five intercept clusters and three slope clusters were identified, with the other two proposed slope clusters remaining empty. Figure 4 displays the combined clusters for respiratory disease risk in Glasgow, with the top panel displaying the clusters on the map while the bottom panel provides a visual representation of the slope and intercept for each cluster. The intercept clusters are represented by the background colours of the areal units in the top panel, with a darker shade of grey corresponding to a higher average disease risk over the study period. The slope clusters are represented by the hatching, with slope cluster 1 represented by ‘downward’ hatches which go from top left to bottom right, slope cluster 2 represented by no hatching, and slope cluster 3 represented by ‘upward’ hatches which go from bottom left to top right.

The bottom panel graphically displays the risk values estimated by the model for each areal unit within each possible cluster combination. Each column represents a different intercept cluster, with the intercept term increasing as you move from left to right. Each row represents a different slope cluster; the top row contains the cluster with increasing risk, the middle row contains the cluster with little or no change, and the bottom row contains the cluster with decreasing risk. Here we can see that the model ensures that areal units in the same cluster have similar risks over the study period, but does still allow for some variation in risk levels within a cluster. The figure shows that the five intercept clusters have median risks of 0.41, 0.57, 0.85, 1.14 and 1.43, while the three slope clusters have median slopes of -0.35, 0.00 and 0.40. The figure shows that 210 areas show little or no change in risk over time, where as 28 and 33 areas show decreasing and increasing risks respectively. Areas of high average risk appeared to be just as likely to experience a further increase as they did a decrease; of the areas in intercept clusters 4 and 5 (corresponding to high risks), 10% exhibited an increasing trend and a further 10% showed a decrease. In areas of low risk (intercept clusters 1 and 2), 10% showed an increase in disease risk over the study period, while 13% exhibited a decrease.

The most concerning cluster for health authorities will be that in the darkest grey with upward hatching. This cluster corresponds to intercept cluster 5 and slope cluster 3, and contains areal units which have a very high disease risk which is increasing over time. This cluster contains areas such as Drumry to the north of the city and Govan to the south which are known to have high levels of deprivation. An investigation into what these areal units have in common may lead to the identification of possible risk factors for respiratory disease. Conversely, areas that exhibit decreasing trends in risk are also of primary interest to health authorities, to identify factors that could be driving the improved health in these areas. Finally, the extent of health inequalities in Glasgow does not appear to have changed substantially over the ten years being investigated. The standard deviation of the fitted disease risks in 2001 was 0.304, while in 2010 a standard deviation of 0.303 was obtained. Likewise, the interquartile range of the fitted risks was 0.427 in 2001 and 0.461 in 2010.

## 6 Discussion

Here we have proposed a Bayesian spatio-temporal model which estimates the disease risk pattern over multiple time points and also identifies clusters of areas which have a similar disease risk characteristics over the study period. There are two separate clustering parameters within the model; the first is based on the average risk (intercept) and the second is based on the change in disease risk over time (slope). The model proposed here extends the<sup>6</sup> model by allowing different intercept and slope terms for each cluster. The model estimates disease risk via four parameters, a pair to estimate the intercept and a pair to estimate the slope. Each pair consists of a set of cluster-specific fixed effect terms and a set of spatially correlated random effects which follow a conditional autoregressive model. The fixed effects ensure that areal units within the same cluster will have similar intercept levels, but the random effects allow for some variation in intercept levels within a cluster.

The simulation study presented in Section 4 showed that our model outperforms the Bernardinelli model with post-hoc clustering across a variety of simulation scenarios. Our model was more accurate than the Bernardinelli model in terms of estimating the correct number of clusters, and also identified more accurate clusters as measured by the Rand index. The risk estimates from our model were also more accurate than those obtained from the Bernardinelli model. This improved estimation is a result of the additional fixed effect terms within our model; the Bernardinelli model has two fixed effects (one for intercept and one for slope) which are common to all areas across different clusters while our model allows for different fixed effects for each cluster. The simulation study in the supplementary material also showed that the performance of our model is not affected by the choice of  $N_C$  and  $N_D$ , the maximum number of clusters for intercept and slope respectively. Based on this result, it is our recommendation that the values of  $N_C$  and  $N_D$  are chosen to be a slightly larger than the number of clusters expected for intercept and slope respectively.

It is straightforward to combine model clusters to produce intercept-slope clusters containing areal units which have similar levels of average disease risk and similar changes in risk over time. This allows health authorities to identify groups of areal units with similar risk values across the entire study period, which has two important uses. Firstly, the clusters can be used to determine policy across the region; similar levels of resources can be allocated to areal units in the same cluster. Secondly, there may be interest in identifying factors which may be causing increased disease risk; for example the areal units in a cluster with increasing disease risk could be compared to identify possible common changes in these areas which could have caused the increase in risk.

As shown in the simulation study, this method represents an improvement on the<sup>6</sup> model and is more straightforward to implement than existing spatio-temporal clustering models such as,<sup>9</sup> which requires complex reversible-jump MCMC algorithms to identify the clusters. The existing approaches outlined in Section 2 all assume that the disease risk is constant within a cluster, but the approach proposed here allows disease risk to vary within a cluster via the random effects. Such variation is likely to exist in real datasets, and therefore the approach proposed here represents a more realistic alternative to the existing models.

This model currently has two separate clustering terms, one set for the intercept and the other for the slope, although it is straightforward to combine these. Nonetheless, it may be of interest to extend the model to include a single slope-intercept cluster term which can partition the areal units based on both characteristics rather than separately. This could be implemented within a similar modelling structure by allowing each intercept-slope cluster to have its own separate intercept and slope fixed effects. This would mean that the intercept-slope interactions were taken into account when estimating disease risk instead of forming these clusters by a post-hoc combination of intercept and slope clusters as is the case here. A possible challenge in such an approach would be avoiding overparameterisation as a result of the increased number of fixed effects. This model could also be extended by developing

a reversible-jump MCMC algorithm to allow the number of clusters to be shaped by the data, rather than relying on a user-defined maximum. This would allow the possibility of an additional cluster being formed, or two clusters being joined together, at each stage of the MCMC algorithm. It would also be of interest to extend this model to allow for a non-linear trend over time, which would enable the approach to be applied to data where there is a more complex temporal trend.

## Supplementary material

The supplementary material contains an additional simulation study which tests the sensitivity of the model to the user-defined choice of maximum cluster number. There is also software provided to allow users to implement the model proposed here.

## Acknowledgements

The authors gratefully acknowledge helpful suggestions made by the associate editor and referee, which have improved the motivation for and content of this paper.

The work of the first author was funded initially by the Carnegie Trust and then by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

## Conflict of Interest

The authors have declared no conflict of interest.

## References

- [1] Mackenbach J, Kunst A, Cavelaars A, Groenhouf F, Geurts J. Socioeconomic inequalities in morbidity and mortality in western Europe. *Lancet*. 1997;349:1655–1659.
- [2] Marmot M. Social determinants of health inequalities. *The Lancet*. 2005;365:1099–1104.
- [3] Elliott P, Wakefield J, Best N, Briggs D. *Spatial Epidemiology: Methods and Applications*. 1st ed. Oxford University Press; 2000.
- [4] Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991;43:1–20.
- [5] Leroux B, Lei X, Breslow N. Statistical Models in Epidemiology, the Environment and Clinical Trials, Halloran, M and Berry, D (eds). In: *Estimation of disease rates in small areas: A new mixed model for spatial dependence*. Springer-Verlag, New York; 1999. p. 135–178.
- [6] Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*. 1995;14:2433–2443.
- [7] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*. 2000;19:2555–2567.
- [8] Rushworth A, Lee D, Mitchell R. A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial Spatio-temporal Epidemiology*. 2014;10:29–38.
- [9] Knorr-Held L, Rasser G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*. 2000;56:13–21.

- [10] Green P, Richardson S. Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*. 2002;97:1055–1070.
- [11] Wakefield J, Kim A. A Bayesian model for cluster detection. *Biostatistics*. 2013;14:752–765.
- [12] Forbes F, Charras-Garrido M, Azizi L, Doyle S, Abrial D. Spatial risk mapping for rare disease with hidden Markov fields and variational EM. *Annals of Applied Statistics*. 2013;7:1192–1216.
- [13] Anderson C, Lee D, Dean N. Identifying Clusters in Bayesian Disease Mapping. *Biostatistics*. 2014;15:457–469.
- [14] MacNab Y, Dean B C. Autoregressive Spatial Smoothing and Temporal Spline Smoothing for Mapping Rates. *Biometrics*. 2001;57:949–956.
- [15] Congdon P, Southall H. Trends in inequality in infant mortality in the north of England, 1921 to 1973, and their association with urban and social structure. *Journal of the Royal Statistical Society: Series A*. 2005;168:679–700.
- [16] Ugarte D M, Goicoa T, Militino F A. Spatio-temporal modelling of mortality risks using penalized splines. *Environmetrics*. 2010;21:270–289.
- [17] Bohning D. Empirical Bayes estimators and non-parametric mixture models for space and time-space disease mapping and surveillance. *Environmetrics*. 2003;14:431–451.
- [18] Choi J, Lawson AB. Evaluation of Bayesian spatial-temporal latent models in small area health data. *Environmetrics*. 2011;22:1008–1022.
- [19] Li G, Best N, Hansell A, Ahmed I, Richardson S. BaySTDetect: detecting unusual temporal patterns in small area data via Bayesian model choice. *Biostatistics*. 2012;13:695–710.
- [20] de Boor C. On Calculation With B-splines. *Journal of Approximation Theory*. 1972;6:50–62.
- [21] Breslow E N, Clayton G D. Approximate inference in generalised linear mixed models. *Journal of the American Statistical Association*. 1993;88:9–25.
- [22] Ugarte D M, Militino F A, Goicoa T. Prediction error estimators in Empirical Bayes disease mapping. *Environmetrics*. 2008;19:287–300.
- [23] MacNab Y, Farrell P, Gustafson P, Wen S. Estimation in Bayesian disease mapping. *Biometrics*. 2004;60:865–873.
- [24] Hossain M, Lawson A, Cai B, Jungsoon C, J L, Kirby R. Space-Time mixture models for small area disease risk and cluster estimation: model selection and optimum allocation. *Environmetrics*. 2014;25:84–96.
- [25] Stephens M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*. 2000;62:795–809.
- [26] Eilers P, Marx B. Flexible smoothing with B-splines and penalties. *Statistical Science*. 1996;11:89–121.
- [27] Fraley C, Raftery A. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*. 2002;97:611–631.
- [28] Fraley C, Raftery AE, Murphy TB, Scrucca L. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation [Technical Report]; 2012.
- [29] Rand W. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*. 1971;66:846–850.

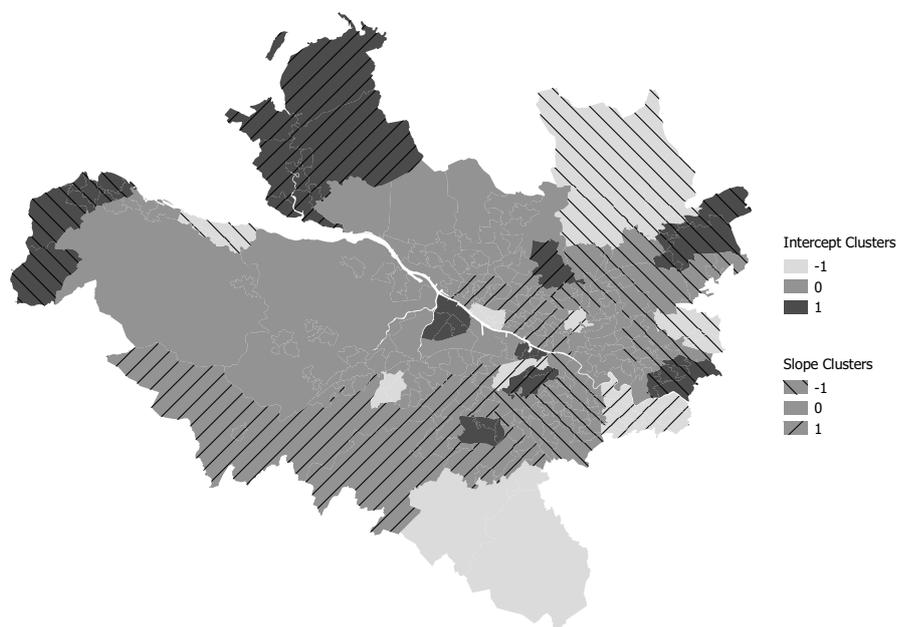


Figure 1: Plot of the set of combined intercept/slope clusters. The intercept clusters are represented by the background colours of the areal units, while the slope clusters are represented by the overlaid hatching.

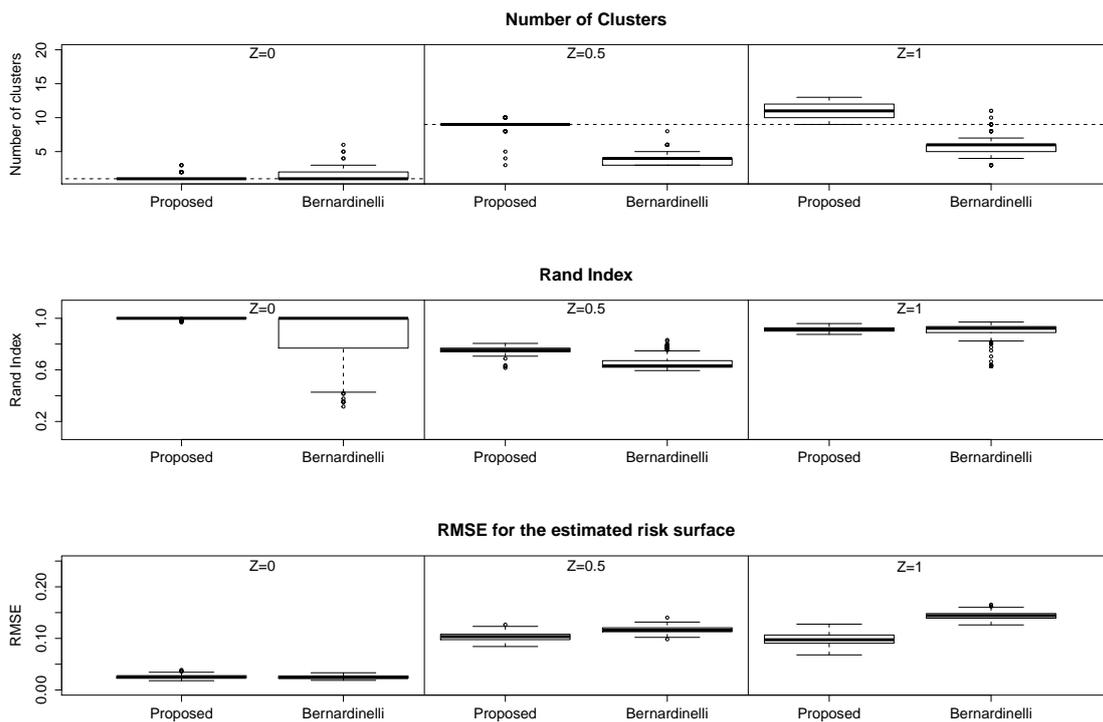


Figure 2: Summary of the results of the simulation study . The top, middle and bottom panels display boxplots for the number of clusters, Rand index and RMSE respectively. The results relate to  $Z = 0$  (left panels),  $Z = 0.5$  (middle panels) and  $Z = 1$  (right panels). In the top panel, the dashed lines represent the true number of clusters in each case.

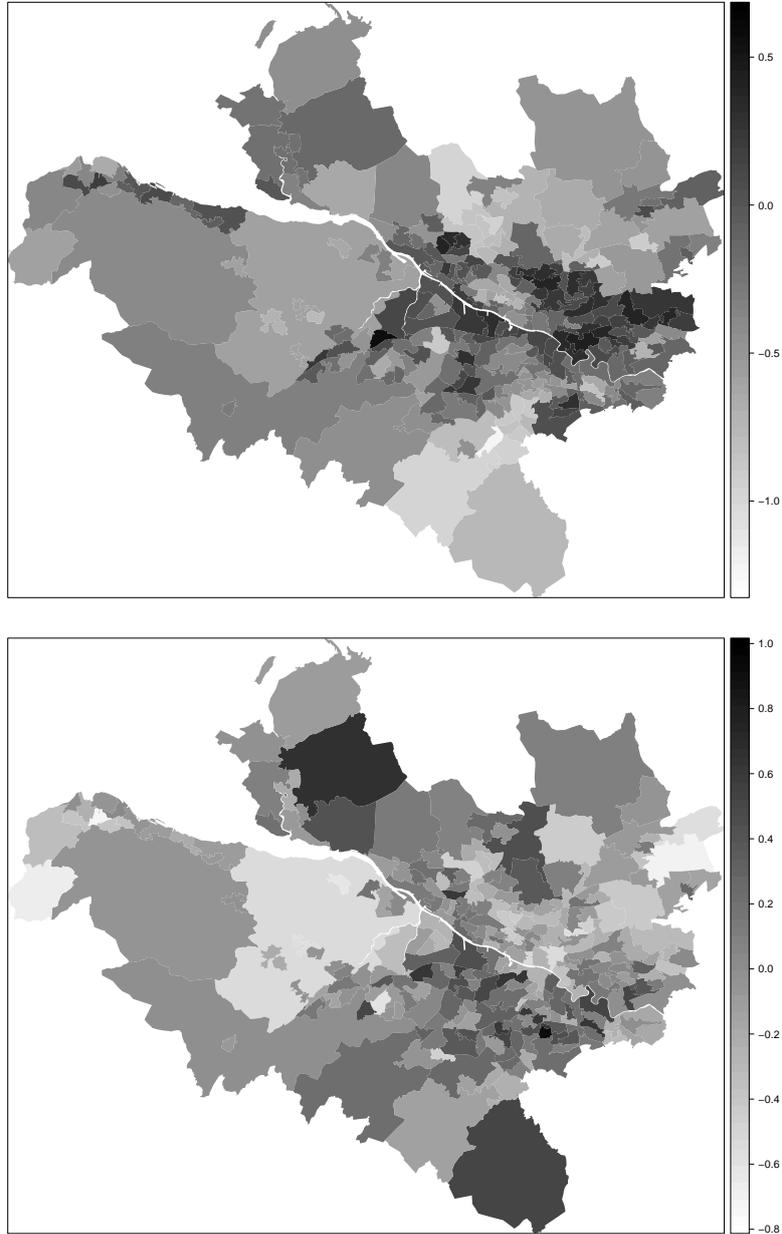


Figure 3: Plots of the intercepts (top panel) and slopes (bottom panel) obtained by fitting the simple model  $Y_{it} \sim \text{Poisson}(E_{it}R_{it})$ ,  $\ln(R_{it}) = \alpha + \beta(t - \bar{t})$  separately for each areal unit  $i$ .

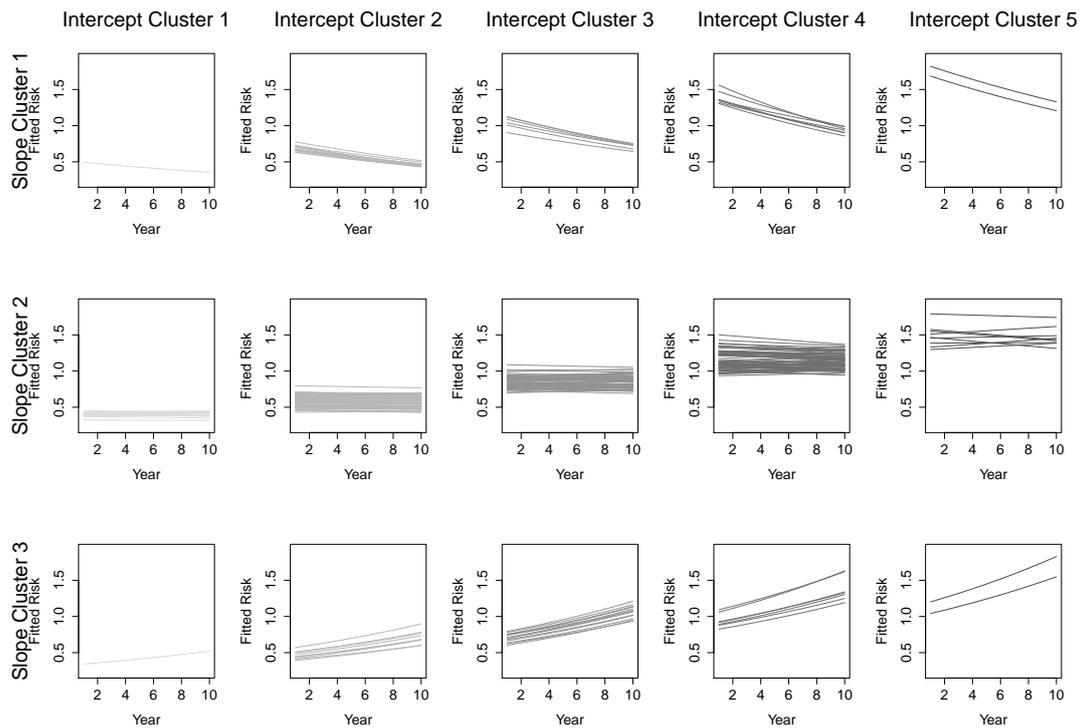


Figure 4: The top panel displays the combined intercept-slope clusters, while the bottom panel provides a visual representation of the characteristics of each cluster. The number of lines in each plot in the bottom panel corresponds to the number of areal units in that combination of intercept and slope clusters. The colours in the bottom panel correspond to the colours of the relevant intercept clusters in the top panel.