Lancaster University

Geostatistical Methods for Modelling Spatially Aggregated Data

by

Olatunji Olugoke Johnson

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

Faculty of Health and Medicine,

Lancaster University
Medical School

16 January 2020

Declaration

I, OLATUNJI OLUGOKE JOHNSON, declare that this thesis titled, "*Geostatistical Methods for Modelling Spatially Aggregated 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: ____________________________

Abstract

Spatially aggregated epidemiological data is nowadays increasingly common because of ethical concern of data use as well as preservation of patient confidentiality. They are typically presented either as the count of disease cases or as an average measurement from districts partitioning a study region. In most cases, the partitioning is based on administrative convenience rather than information about the aetiology of any disease or public health problem. While inference for spatially aggregated data commonly make use of model that assumes a spatially discrete variation, we argue that a spatially continuous model should be considered when there is a scientific justification for its use, especially when the underlying generating process of the disease outcome is hypothesised to behave in a spatially continuous manner. In this thesis, we consider geostatistical methods as a framework that can be used to analyse spatially aggregated data. This thesis is a series of papers, two methodological and one public health application. In the first methodological paper, we developed a computationally efficient discrete approximation to log-Gaussian Cox process (LGCP) models for the analysis of spatially aggregated disease count data. We compare the predictive performance of our modelling approach with LGCP through a simulation study and an application to primary biliary cirrhosis incidence data in Newcastle-Upon-Tyne, UK. Our results suggest that when disease risk is assumed to be a spatially continuous process, the proposed approximation to LGCP provides reliable estimates of disease risk both on spatially continuous and aggregated scales. In the second methodological paper, We developed a model-based geostatistical approach that allows us to model the relationship between the Life expectancy at birth (LEB) and the index of multiple deprivation (IMD), when these are available over different partitions of the study region. We found that the effect of IMD on LEB is higher for males than for females. We show that our proposed model-based geostatistical approach does not only provide solution to any form of misalignment problem but also allows for spatially continuous inferences. In the third application paper, we developed a spatio-temporal model for monthly Chronic Obstructive Pulmonary Disease (COPD) emergency admissions data in South Cumbria and North Lancashire, UK, 2012-2018. We assess the relative contribution of socio-economic and environmental variables for forecasting COPD emergency admissions. In addition, we develop an early warning system that triggers an alarm whenever COPD emergency admissions exceeds a predefined incidence thresholds. The result of our analysis can potentially help NHS Morecambe Bay Clinical Commissioning Group stakeholders to define areas to target early intervention as well as inform resource allocation for healthcare system so that its limited resources can be used to maximum effect.

Acknowledgements

Foremost, I acknowledge the most powerful God for giving me the strength, wisdom, courage and ability to complete this research work.

My deepest gratitude is to my supervisors, Dr Emanuele Giorgi, Prof Peter Diggle and Prof Jo Knight. Their patience, immense knowledge and guidance helped me throughout the writing of this thesis. Thank you for your daily support. Thank you for sharing with me out of your wealth of knowledge. I have certainly learnt how to become a good supervisor to my future students from them.

I equally appreciate Dr Tim Gatheral, my clinical supervisor. Thank you for helping me to understand the fundamentals of COPD. Also, special thanks to Prof Mike Pearson, another clinical mentor. Our regular meetings are of great help and well appreciated.

I appreciate the funding from the Lancaster University arm of Connected Health Cities (CHC) project; and members of the team, namely Sumi Helal, Nick King, Jo Kendall, Helena Tendedez, Victoria Neumann, Nicola Platt and my supervisors. It is nice working with you. Particularly, I am grateful to CHC, whose generous funding and support enabled me to pursue a PhD. Also, I acknowledge the travel grant from African Institute for Mathematical Sciences that allows me to travel from Nigeria to the UK to start my PhD.

I want to appreciate the best Mum in the world, Mrs Foluke Abiodun Johnson for her parental care, love bestowed on me, advice and encouragement. You will live to reap the fruit of your labour in Jesus name. Special thanks to my brothers and sisters; Ajibola, Olaniran, Abimbola, Babatunde and Promise as well as my Queen, Stephanie.

Last but not the least, I equally acknowledge my friends, colleagues, well-wishers, amazing church people (RCCG Latter Rain, Lancaster and Redeemed Christian Campus Society (RCCS)) and the CHICAS family (Irene, Rachel, James, Erick, Claudio, Lisa, Poppy, Fran, Dileepa, Ben, Farhad, Juan, Laura, Tobias, Anne, Maddy, Max, ...).

Contents

1 Introduction **2**

1.1 The Standard Geostatistical Model for Spatially Point Referenced Data 4

1.1.1 Linear Geostatistical Model (LGM) 4

1.1.2 Generalised Linear Geostatistical Model (GLGM) 5

1.1.3 Family of correlation function 6

1.1.4 Spatial Exploratory Analysis 8

1.1.5 Exceedance Probability 11

1.2 Thesis Structure 11

2 Paper 1: A Spatially Discrete Approximation to Log-Gaussian Cox Processes for Modelling Aggregated Disease Count Data **16**

2.1 Introduction 18

2.2 Existing methods for modelling spatially aggregated disease count data 20

2.2.1 Gaussian Markov random field models 20

2.2.2 Log-Gaussian Cox process models 21

2.3 A spatially discrete approximation to Log-Gaussian Cox processes . . 23

2.3.1 Monte Carlo maximum likelihood 27

2.3.2 Continuous spatial prediction 29

2.4 Simulation Study 29

2.5 Application: mapping of primary biliary cirrhosis risk 33

2.6 Discussion 36

3 Paper 2: Dealing with spatial misalignment to model the relationship between deprivation and life expectancy in Liverpool: A model-based geostatistical approach 52

3.1 Background 54

3.2 Existing methods for analysing spatially misaligned data 57

3.2.1 Geostatistical method for spatially misaligned point-referenced data 58

3.2.2 Conditional autoregressive models for spatially misaligned areal data 60

3.3 Methods 61

3.3.1 Data 61

3.3.2 Modelling framework 63

3.3.3 Inference: parameter estimation and spatially continuous prediction 64

3.3.4 Model validation: testing for residual spatial correlation 67

3.3.5 Assessment of the coverage probabilities for the regression parameters and the spatial predictions 68

3.4 Results 69

3.5 Discussion 79

3.6 Conclusion 81

4 Paper 3: Spatio-temporal modelling of incidence in COPD emergency admissions in an area of Northwest England from 2012 to

2018. **87**

4.1 Introduction 89

4.2 Methods 92

4.2.1 COPD admission data 92

4.2.2 Statistical modelling and assessment of residual spatio-temporal correlation 94

4.3 Result 98

4.3.1 Descriptive Analysis 98

4.3.2 Spatio-temporal Analysis 98

4.4 Discussion 101

5 General discussion, conclusions and future work **112**

5.1 Summary and future extensions of SDALGCP models 112

5.2 Summary and future extensions of geostatiscal methods for analysing spatially misaligned areal data 116

5.3 Summary and future extensions of spatial-temporal modelling of COPD emergency admission 117

5.4 Software development 118

Appendix **121**

A.1 Proof of the Monte Carlo approximation 121

A.2 Diagnostic Plots of SDA Model 123

A.3 Convergence, Mixing diagnostic and Plots of LGCP Model 124

A.4 Mixing diagnostic of BYM Model 127

A.5 Mixing diagnostic of EV Model 128

A.6 Simulation Plot 129

A.7 Derivation of the equations in Paper 2 130

A.8 Map of the Observed IMD and LEB 136

A.9 Map of the Liverpool, MSOA, LSOA and Ward 138

List of Figures

1.1 Examples of parametric correlation functions: Fig a: showing the the Matérn family with different values of κ; Fig b: visualises exponential, Gaussian and spherical correlation functions. 8

1.2 Schematic representation of a typical variogram, with structural parameters indicated. 10

2.1 Maps of the estimated primary biliary cirrhosis (PBC) incidence in each lower layer super output area (LSOA) of Newcastle-Upon-Tyne from the four fitted models in Section 2.5. 38

2.2 Maps of the standard error of the estimated primary biliary cirrhosis (PBC) incidence in each lower layer super output area (LSOA) of Newcastle-Upon-Tyne from the five fitted models in Section 2.5. . . . 39

2.3 The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the estimated primary biliary cirrhosis (PBC) incidence in the lower layer super output areas (LSOA) of Newcastle-Upon-Tyne for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the estimated PBC incidence from each model. 40

2.4 The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the standard errors of primary biliary cirrhosis (PBC) incidence in the lower layer super output areas (LSOA) of Newcastle-Upon-Tyne for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the standard errors from each model. 41

2.5 Maps of the predicted relative risk surface $\exp\{S(x)\}$ from the fitted spatially discrete approximation to log-Gaussian Cox Process (SDA) using a population-weighted log-intensity average (SDA I, upper panel) and a simple average (SDA II, middle panel), and the exact LGCP model (lower panel). 42

2.6 Maps of the standard error of the predicted relative risk surface $\exp\{S(x)\}$ from the fitted spatially discrete approximation to log-Gaussian Cox Process (SDA) using a population-weighted log-intensity average (SDA I, upper panel) and a simple average (SDA II, middle panel), and the exact LGCP model (lower panel). 43

2.7 The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the estimated spatially continuous relative risk $\exp\{S(x)\}$ for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the estimated relative risk from each model. 44

2.8 The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the standard errors for the estimated risk $\exp\{S(x)\}$ for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the standard errors from each model. . 45

3.1 Map of Liverpool district council, UK showing the 30 electoral wards 72

3.2 Maps of the estimated female (upper panel) and male (middle panel) life expectancy at birth (LEB) and index of multiple deprivation (IMD) (lower panel). Middle Super Output Area (MSOA) with boundaries coloured in green correspond to the lowest estimated LEB, whilst those in white to the highest. For males, the lowest estimated LEB is 70.2 years and the highest is 85.2 years; for females, the lowest is 73.5 years and the highest is 89.6 years. 73

3.3 Spatially continuous prediction maps of female (upper panel) and male (lower panel) life expectancy at birth (LEB) in Liverpool, UK. In the upper panel, the white contour lines are for a LEB of 89.6 years and the green contour lines for a LEB of 73.5 yers; in the lower panel, the white contour lines correspond to 70.2 years and the green contour lines to 85.2 years. 74

3.4 Maps of the non-exceedance probability of female (upper panel) and male (lower panel) life expectancy at birth (LEB), with threshold 82.9 and 79.2 (average LEB in England, UK), respectively in Liverpool, UK. 75

3.5 Plots of the observed variograms (points) and the 95% tolerance bandwidth (dashed lines) generated under the assumption of absence of residual spatial correlation. 76

3.6 Scatter plots of the actual against the nominal coverage for the confidence intervals generated for β_1 and β_2 (upper panels), and for the spatially continuous and MSOA-level predictions of LEB (lower panels). The red lines in each panel correspond to the identity line. . . . 77

4.1 Map of South Cumbria and North Lancashire containing 209 LSOAs. 93

4.2 (a) Count of COPD emergency admission, by age group and sex, in South Cumbria and North Lancashire, 2012-2018; and (b) Incidence rate of COPD emergency admission per 1000 population, by age group and sex, in South Cumbria and North Lancashire, 2012-2018. 99

4.3 Boxplot showing seasonal variation in the monthly count of COPD emergency admission in South Cumbria and North Lancashire, 2012-2018. 100

4.4 Spatio-temporal variogram of the residual from an intercept only model. This shows an evidence of spatio-temporal variation. 102

4.5 Spatio-temporal variogram of the residual from model including all the predictors. This shows that there is no evidence of spatio-temporal variation. 103

4.6 The receiver's characteristics curve (ROC) with area under the curve (AUC) =0.78. The red dot indicates the value of the sensitivity and specificity for which the optimal cut off p = 0.85 value was chosen. . 104

4.7 The monthly-predicted surveillance maps comparing the predicted and the true alarm for each LSOA. Colour blue indicates an LSOA that is correctly predicted to be below the threshold; orange indicates an LSOA that is correctly predicted to be above the threshold; purple indicates an LSOA that is incorrectly predicted to be below the threshold; and red indicates an LSOA that is incorrectly predicted to be above the threshold. The incidence threshold used is 12 per 100,000. 105

1 The plots show the Autocorrelation plot of the process, S. 123

2 Diagnosing convergence to a posterior mode: a plot of the log-target, $\log\{\pi(\beta, \eta, S|N)\} + c$ up to an additive constant, c 124

3 The maps show the autocorrrlation plot three different lags Fig a: Lag 1 ; Fig b: Lag 5; Fig c: Lag 15. 125

4 The plots show the trace plot of the posteriors of the parameters. . . 126

5 The plots show the trace plot of the posteriors of the fixed effect parameters, and random effect parameters, that is β_1, β_2, σ^2 and ϕ. . 127

6 The plots show the trace plot of the posteriors of the fixed effect parameters, and random effect parameters, that is β_1, β_2, σ^2 and ϕ. . 128

7 Histogram of the Observed log-incidence from the data (Observed, left-upper panel) and the first simulated log-incidence (Simulation 1, right-upper panel) and the second simulated log-incidence (Simulation 2, left-lower panel), and the third simulated log-incidence (Simulation 3, right-lower panel). 129

8 Map of the observed female life expectancy at birth (LEB) 136

9 Map of the observed male life expectancy at birth (LEB) 137

10 Map of the observed index of multiple deprivation (IMD). 137

11 Maps of Liverpool wards with MSOA boundaries overlayed. Red lines are the ward boundaries while black lines are the MSOA boundaries . 138

12 Maps of Liverpool wards with LSOA boundaries overlayed. Red lines are the ward boundaries while black lines are the LSOA boundaries . 139

List of Tables

2.1 Average bias, root-mean-square-error (RMSE), width of the 95% prediction interval (WPI) and the 95% coverage probability (CP) for the LSOA incidence, λ_i, from the simulation study of Section 2.4. 32

2.2 Average bias, root-mean-square-error (RMSE), width of the 95% prediction interval (WPI) and the 95% coverage probability (CP) for the spatially continuous relative risk, $\exp\{S(x)\}$, from the simulation study of Section 2.4. 32

2.3 Point estimates and 95% confidence/credible intervals (CI) for the model parameters of the spatially discrete approximation to log-Gaussian Cox Process (LGCP) using a population-weighted log-intensity average (SDA I) and a simple average (SDA II), the exponential variogram (EV) model, Besag-York-Mollié (BYM) model and the LGCP model. 37

3.1 Point estimates and 95% confidence intervals (CI) for the three model parameters. 78

4.1 The table showing the set of predictors available for this study. 95

4.2 The table showing the relative variance reduced by the predictors. . . 101

List of Papers

Paper 1 *A Spatially Discrete Approximation to Log-Gaussian Cox Processes for Modelling Aggregated Disease Count Data.*

Authors: Olatunji Johnson, Peter Diggle and Emanuele Giorgi.

Published in *Statistics in Medicine*;

Contribution: Formulation of the idea. Development and implementation of the methods in the paper. Writing of the paper.

Paper 2 *Dealing with spatial misalignment to model the relationship between deprivation and life expectancy in Liverpool: A model-based geostatistical approach.*

Authors: Olatunji Johnson, Peter Diggle and Emanuele Giorgi.

Accepted in *International Journal of Health Geographics.*

Contribution: Formulation of the idea. Development and implementation of the methods in the paper. Writing of the manuscript.

Paper 3 *Spatio-temporal modelling of incidence in COPD emergency admissions in an area of Northwest England from 2012 to 2018.*

Authors: Olatunji Johnson, Jo Knight, Mike Pearson, Tim Gatheral, Peter Diggle and Emanuele Giorgi.

Contribution: Statistical analysis and writing of the paper.

To my Queen,

Stephanie

Chapter 1

Introduction

Aggregated epidemiological data is nowadays increasingly common. This is usually because of ethical concern of data use as well as preserving patient confidentiality. They are typically presented either as the count of cases or as average measurement from administrative districts partitioning of the area of study. The partitioning is of course in most cases based on administrative convenience rather than information about the aetiology of any disease. For example, in England, various geographies are used in the production of (health) statistics. The Output Areas (OA) is the lowest geographical level at which census estimates are provided. Because of confidentiality purposes, Super Output Areas (SOAs) were created to report official statistics, which are an aggregation of adjacent OAs. There are two tiers of SOAs, the Lower Layer Super Output Area (LSOA) which typically contain 4 to 6 OAs with a population of around 1500 and the Middle Layer Super Output Areas (MSOAs) which contains on average 7200 people. The caveat is that these SOAs were not designed based on any epidemiological characteristics.

A key idea of formulating a typical spatial (or spatio-temporal) statistical model for most health outcomes is to assume that outcome depends on a range of factors, some of which are known and some unknown - sometimes termed explained and unexplained. Therefore modelling the outcome as a linear combination of explained and unexplained variables. The explained component are measured characteristics, whilst the unexplained component can be modelled as an unobserved spatially (or spatio-temporally) varying stochastic process. The unexplained component can be theoretically modelled either as a spatially continuous variation or spatially discrete variation - the former is used in geostatistics (Diggle et al., 2007, 2013) specified through a Gaussian Random Field (GRF) (Abrahamsen, 1997), while the latter is modelled through the Gaussian Markov Random Field (GMRF), such as conditional autoregressive (CAR) structure (Besag et al., 1991; Lee and Durbán, 2009; Leroux et al., 2000). The link between GRFs and GMRFs have been studied by Lindgren et al. (2011) and Simpson et al. (2012).

In this thesis we address some of the issues related to analysing spatially aggregated data. In particular, we focus on how the geostatistical methods which are well established in spatially point referenced data can be adapted to analyse spatially aggregated data, since most spatially aggregated data is an aggregation of spatially point referenced data. While inference for spatially aggregated data commonly make use of model that assumes a spatially discrete variation, we argue that a spatially continuous model should be considered when there is a scientific justification for its use, especially when the underlying generating process of the disease outcome is hypothesised to behave in a spatially continuous manner.

In the next section, we describe the standard geostatistical models, linear geostatistical model (LGM) and generalised linear geostatistical method (GLGM) for spatially point referenced data. These methods will be extended for spatially aggregated data in the following chapters.

This thesis is a series of papers, two methodological and one applied. In the first methodological paper, we develop a spatially discrete approximation to Log-Gaussian Cox process for analysing spatially aggregated count data. In the second methodological paper, we developed a model-based geostatistical approach that allows us to model the relationship between the Life expectancy at birth (LEB) and the index of multiple deprivation (IMD), when these are available over different partitions of the study region. In the third paper, we developed a spatio-temporal model for monthly Chronic Obstructive Pulmonary Disease (COPD) emergency admission in North Lancashire and South Cumbria, UK, 2012 - 2018.

1.1 The Standard Geostatistical Model for Spatially Point Referenced Data

Here we describe the geostatistical methods for modelling spatially point referenced data, LGM and GLGM.

1.1.1 Linear Geostatistical Model (LGM)

Consider a continuous response variable, Y_i, measured at a discrete set of locations, $\{x_i : i = 1, \ldots, n\}$, where each x_i lies within a geographical region of interest, A. The standard linear geostatistical model for Y_i can be written as

$$Y_i = d(x_i)^\top \beta + S(x_i) + Z_i \tag{1.1}$$

where $d(x_i)$ is a vector of explanatory variables including environmental and socio-economic variables with associated regression coefficients β, usually used to explain some of the variations in Y_i; $S(x_i)$ is a spatial stochastic process modelled as zero-mean Gaussian process, used to account for unmeasured spatially structured risk factors; and Z_i is a zero-mean Gaussian noise sometimes referred to as the *nugget effect* used to capture intrinsic random variation owing to measurement error.

By assuming that $S(x)$ is a zero-mean, stationary and isotropic Gaussian process, the joint distribution of $S = (S(x_1), \ldots, S(x_n))$ is multivariate Gaussian with zero mean and covariance $\varSigma$, with ij-th entry given as

$$\varSigma_{ij} = \mathrm{Cov}\{S(x_i), S(x_j)\} = \sigma^2 \rho(\|x_i - x_j\|; \theta), \tag{1.2}$$

where σ^2 is the variance, $\|x_i - x_j\|$ is the Euclidean distance between locations x_i and x_j and $\rho(\cdot; \theta)$ is the isotropic and stationary correlation function of $S(x)$ indexed by the parameter θ. We shall define $\rho(\cdot; \theta)$ and give examples in Section 1.1.3. Finally, the joint distribution of $Z = (Z_1, \ldots, Z_n)$ is multivariate Gaussian with zero mean and covariance $\tau^2 \mathbb{I}_n$, where $\mathbb{I}_n$ is an $n \times n$ identity matrix and τ^2 is the variance.

1.1.2 Generalised Linear Geostatistical Model (GLGM)

The class of generalised linear geostatistical model is an extension of LGM used for response variables that are not normally distributed. In epidemiological research, the most common of this class of model are the Binomial logit-linear and Poisson log-linear geostatistical model. This class of model have been well studied, see for example Diggle and Ribeiro (2007). Specifically, extensive research on model-based geostatistics for binomial data with application in low resource setting can be found in Diggle and Giorgi (2016).

We retain the meaning of all notations used in the previous section except that Y_i is now a discrete response. We shall assume that conditionally on the random effects $S(\cdot)$ and Z, Y_i are mutually independent random variables from a family of the exponential distribution. If Y_i is the number of positive response out of m_i individuals sampled in a region $\mathcal{R}_i$, the natural model is Binomial logit-linear model, and if Y_i is the number of disease cases with m_i population at risk, the natural model is Poisson log-linear model. The two following ingredients are then needed to fully characterize the probabilistic distribution of Y_i (McCullagh, 2019).

- The *linear predictor* is defined as

$$\eta_i = d(x_i)^\top \beta + S(x_i) + Z_i.$$

- The *link function* $g(\cdot)$ such that

$$E[Y_i|\eta_i] = m_i g^{-1}(\eta_i),$$

A common choice of $g(\cdot)$ for Poisson model is the logarithm function and for Binomial model is logit function. We shall discuss an extension to spatially aggregated response data in Chapter 2.

1.1.3 Family of correlation function

The main ingredient of defining a fully parametric geostatistical model is a positive-definite correlation function $\rho(u;\theta)$, where $u = \|x - x'\|$, x and x' are arbitrary locations and $\theta = (\kappa, \phi)$ or $\theta = \phi$. The following are the common functions that are usually used.

1.1.3.1 Matérn family

Matérn family is the most popular and most often used class of correlation function. This family is named after Matérn (1960) and it is characterised by two parameters,

$\theta = (\kappa, \phi)$:

- $\kappa > 0$, the shape (or smoothness) parameter, determines the differentiability of the process $S(x)$. More specifically, this will be $\lceil \kappa \rceil$ (i.e. the smallest integer greater than or equal to κ) minus one times differentiable; and
- $\phi > 0$ the range (or scale) parameter, regulates the rate at which the spatial correlation decays for increasing distance u.

More specifically, its expression is given by

$$\rho(u;\theta) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{u}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{u}{\phi}\right),$$

where $K_{\kappa}(\cdot)$ is the modified Bessel function of the second kind of order $\kappa > 0$.

1.1.3.2 Exponential correlation function

$$\rho(u;\theta) = e^{\frac{-u}{\phi}}.$$

This is a special case of the Matérn family with parameter $\kappa = 1/2$. The resulting process $S(x)$ has sample functions that are not differentiable but are mean-squared continuous, since $\rho(\cdot;\theta)$ is continuous at the origin. It drops asymptotically towards zero as $u \longrightarrow \infty$.

1.1.3.3 Gaussian correlation function

$$\rho(u;\theta) = e^{\frac{-u^2}{\phi}}.$$

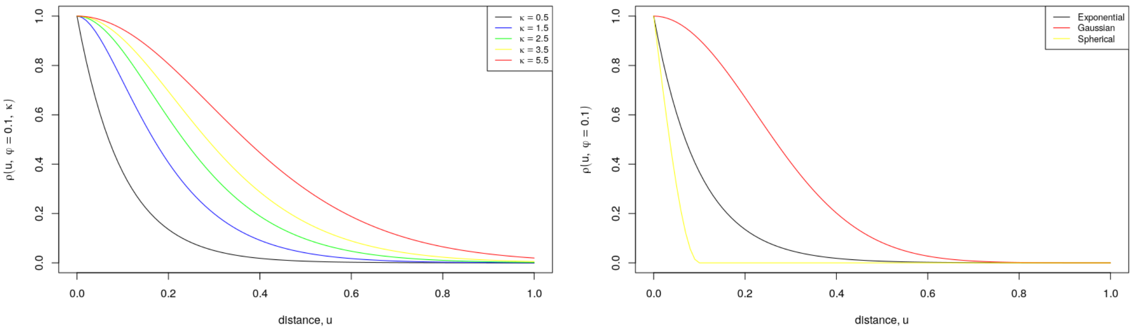
This is also a special case of the Matérn family with parameter $\kappa \longrightarrow \infty$. The resulting process $S(x)$ has sample functions that are infinitely times differentiable.

1.1.3.4 Spherical correlation function

$$\rho(u;\theta) = \begin{cases} 1 - \frac{3}{2}(u/\phi) + \frac{1}{2}(u/\phi)^3 & 0 \leq u \leq \phi \\ 0 & u > \phi \end{cases}.$$

The resulting process $S(x)$ has paths that are not differentiable but are continuous and it depends on a single unknown scale parameter, ϕ.

Figure 1.1a shows the Matérn family with different values of κ. Larger values of κ lead to correlation functions with a larger scale, and thus stronger correlations for larger distances. Figure 1.1b shows some examples of different correlation curves: exponential; Gaussian; and Spherical. More examples of correlation functions as well as its theoretical properties can be found in Wackernagel (2013).



(a)

(b)

Figure 1.1: Examples of parametric correlation functions: Fig a: showing the the Matérn family with different values of κ; Fig b: visualises exponential, Gaussian and spherical correlation functions.

1.1.4 Spatial Exploratory Analysis

The key starting point of every spatial analysis is exploratory spatial data analysis (ESDA) as we have in other types of statistical analysis. ESDA methods focus on assessing the spatial data for spatial autocorrelation and spatial heterogeneity.

A general approach used in geostatistics is the variogram. Variogram has been well utilised in classical geostatistical analysis to describe the degree of spatial dependence of a spatial random field or stochastic process $S(x)$. The theoretical semivariogram for process $W(x) = S(x) + Z$ is defined as

$$\begin{aligned}\gamma(x, x') &= \frac{1}{2}\text{var}\{W(x) - W(x')\} \\ &= \frac{1}{2}E[\{W(x) - W(x')\}^2] \\ &= \tau^2 + \sigma^2(1 - \rho(u; \theta)),\end{aligned}$$

for a stationary and isotropic spatial process $S(x)$. Clearly, since $\rho(u; \theta)$ is a monotonically decreasing function in u, the variogram is a monotonically increasing function in u.

In practice, the empirical variogram is used to test for the presence of residual spatial correlation in the residuals after fitting a non-spatial model to the data. Empirical variogram helps to describe the spatial dependence in the data and to estimate the autocorrelation structure of the underlying stochastic process. Let $\hat{W}_i$ denote the predicted residual. Metheron (1963) defined semivariance function, $\hat{\gamma}(u)$ as the half of the average square difference between residuals at points that are separated by an Euclidean distance u, written algebraically as,

$$\hat{\gamma}(u) = \frac{1}{2|n(u)|} \sum_{(p,q) \in n(u)} \left(\hat{W}_p - \hat{W}_q\right)^2,$$

where $n(u)$ is the set that contains all the neighbouring pairs at distance u, $|n(u)|$ is the number of distinct pairs of $n(u)$. A schematic example of a typical variogram is shown in Figure 1.2 (Johnson, 2016). A rising trend up across u divulges a presence of spatial variation. To highlight the features: nugget variance τ^2 corresponds to $\hat{\gamma}(u)$ at $u = 0$; sill is the total variance, sum of the nugget variance τ^2 and the signal variance σ^2 obtained as $u \longrightarrow \infty$; and practical range is the distance at which the semivariance value achieves 95% of the sill, that is the value of u when $\hat{\rho}(u;\theta) = 0.05$.

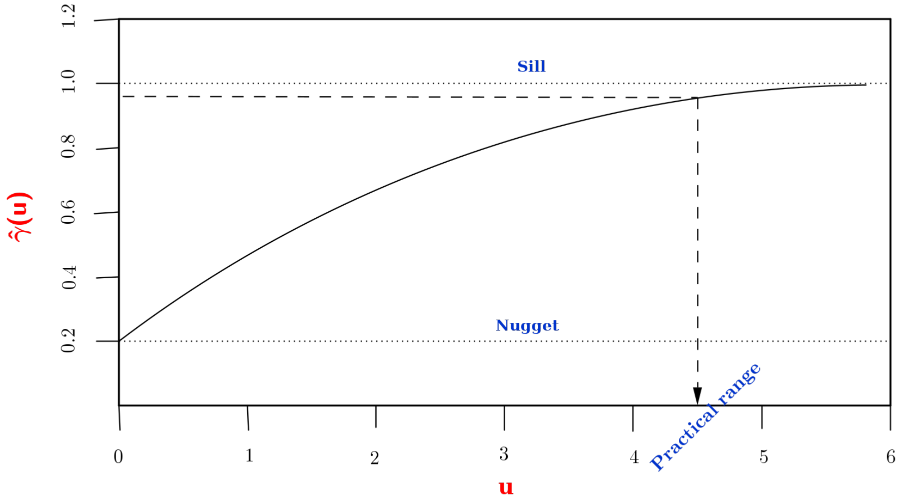


Figure 1.2: Schematic representation of a typical variogram, with structural parameters indicated.

The next step after constructing a variogram is to establish whether the observed patterns are or are not compatible with random fluctuations. A simple Monte Carlo test is used to test for the presence of residual spatial correlation via the following steps:

1. Randomly permutes the labelling of $\hat{W}_i$ by holding the regions fixed.
2. Compute $\hat{\gamma}(u)$ in Equation 1.1.4 using the permuted $\hat{W}_i$.
3. Repeat the steps in 1 and 2 for large samples, say B.
4. Use the resulting B of $\hat{\gamma}(u)$ to obtain 95% tolerance interval for each distance bin, under the hypothesis that $\hat{W}_i$ is spatially independent.

After completing step 4, if the initial $\hat{\gamma}(u)$ falls outside the 95% tolerance band, then we conclude that there is evidence against the assumption of spatial independence.

1.1.5 Exceedance Probability

A more natural way to quantify uncertainty is to provide a standard error map but they usually do not convey much information (Giorgi et al., 2018), especially when the interest is on providing information on the degree of uncertainty. For example, in health decision making, when the interest is to reliably identify areas where the disease risk exceeds or go below a policy-relevant threshold. A more useful way to convey the meaningful uncertainty in this setting is to use the exceedance probability (EP) map. Let $\hat{\lambda}(x)$ be the predicted disease risk at location x, the expression for the EP is

$$Pr(\hat{\lambda}(x) > l|data),$$

where l is a predefined threshold. In general, values of EP close to 1 indicate that disease risk is highly likely to be above l, while the values of EP close to zero indicate that disease risk is highly likely to be below l. Finally, values of EP around 0.5 indicate that disease risk is equally likely to be above or below l, thus implying a scenario with the highest uncertainty.

1.2 Thesis Structure

In Chapter 2, we proposed an alternative method to analyse spatially aggregated count data. In this work, we developed spatially discrete approximation to log-Gaussian Cox process (LGCP) for the analysis of spatially aggregated data. The methodology extends the LGCP method for analysing spatially aggregated case-count data proposed by Li et al. (2012) and Diggle et al. (2013). We consider Y_i as a spatially aggregated case-count data derived from a point process data. We give an overview of the existing method for modelling spatially aggregated case-count data including hierarchical Poisson-Gaussian Markov random field model (Besag et al., 1991) and LGCP models (Diggle et al., 2013; Li et al., 2012). We define and develop our spatially discrete approximation to the LGCP models, by approximating the conditional log-intensity of an LGCP as piecewise constant by taking its weighted or simple average over $\mathcal{R}_i$. We carry out parameter estimation for the model using the Monte Carlo maximum likelihood (MCML) method (Christensen, 2012). We conducted a simulation study to assess the predictive performance of the proposed approximation in (2.3) when the underlying process is an LGCP model. We consider the prediction of the incidence λ_i and the spatially continuous relative risk, $\exp\{S(x)\}$. We applied our method to analyse the incidence data on primary biliary cirrhosis (PBC) in Newcastle-Upon-Tyne, UK.

In Chapter 3, we proposed a novel joint geostatistical approach to model the relationship between two spatially misaligned dataset. We considered an application to life expectancy at birth and the index of multiple deprivation in Liverpool, UK. We estimate the parameters of the model using the Maximum Likelihood (ML) method. We carry out a spatially continuous prediction of male and female LEB in Liverpool. We used Non-exceedance probability (NEP) map to identify areas in the Liverpool council district whose LEB is highly likely to fall below a threshold l, by setting l to be England-wide average years for males ($l = 79.2$ years) and females ($l = 82.9$ years). Finally, we developed an online web application that allows the user to dynamically change the threshold.

In Chapter 4, we analyse the monthly COPD emergency admission dataset. Predict the incidence of monthly COPD emergency admission for 12 months ranging from April 2017- March 2018. We develop an early warning system that triggers an alarm whenever COPD emergency admissions signal the likely exceedance of predefined incidence thresholds.

Chapter 5 is a concluding general discussion where we present a summary of the main contributions, the implications of our results on the analysis of COPD emergency admission and explore possible future extensions of the developed methodologies in the previous chapters.

Bibliography

Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Christensen, O. F. (2012). Monte carlo maximum likelihood in model-based geostatistics. *Journal of computational and graphical statistics.*

Diggle, P., Ribeiro, P., and Geostatistics, M.-b. (2007). *Springer Series in Statistics.* Springer.

Diggle, P. J. and Giorgi, E. (2016). Model-based geostatistics for prevalence mapping in low-resource settings. *Journal of the American Statistical Association*, 111(515):1096–1120.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Diggle, P. J. and Ribeiro, P. J. (2007). Generalized linear models for geostatistical data. *Model-based Geostatistics*, pages 79–98.

Giorgi, E., Osman, A. A., Hassan, A. H., Ali, A. A., Ibrahim, F., Amran, J. G., Noor, A. M., and Snow, R. W. (2018). Using non-exceedance probabilities of policy-relevant malaria prevalence thresholds to identify areas of low transmission in somalia. *Malaria journal*, 17(1):88.

Johnson, O. (2016). Model-based geostatistical mapping of river blindness prevalence in cameroon. `https://docs.google.com/a/aims.ac.tz/viewer?a=v&pid=sites&srcid=YWltcy5hYy50enxhaW1zLXRhbnphbmlhLWFyY2hpdmV8Z3g6Mjk2ZGQ0YjRhOTdhYTUzZA`. [Master's thesis].

Lee, D.-J. and Durbán, M. (2009). Smooth-car mixed models for spatial count data. *Computational Statistics & Data Analysis*, 53(8):2968–2979.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

Li, Y., Brown, P., Gesink, D. C., and Rue, H. (2012). Log gaussian cox processes and spatially aggregated disease incidence data. *Statistical methods in medical research*, 21(5):479–507.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Matérn, B. (1960). Spatial variation. *Lecture Notes in Statistics.*

McCullagh, P. (2019). *Generalized linear models.* Routledge.

Metheron, G. (1963). Principles of geostatistics, economic geology. *Economic Geology*, 58(8):1246–1266.

Simpson, D., Lindgren, F., and Rue, H. (2012). Think continuous: Markovian gaussian models in spatial statistics. *Spatial Statistics*, 1:16–29.

Wackernagel, H. (2013). *Multivariate geostatistics: an introduction with applications.* Springer Science & Business Media.

Chapter 2

A Spatially Discrete Approximation to Log-Gaussian Cox Processes for Modelling Aggregated Disease Count Data

Olatunji Johnson, Peter Diggle, Emanuele Giorgi

CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK

Summary

In this paper, we develop a computationally efficient discrete approximation to log-Gaussian Cox process (LGCP) models for the analysis of spatially aggregated disease count data. Our approach overcomes an inherent limitation of spatial models based on Markov structures, namely that each such model is tied to a specific partition of the study area, and allows for spatially continuous prediction. We compare the predictive performance of our modelling approach with LGCP through a simulation study and an application to primary biliary cirrhosis incidence data in Newcastle-Upon-Tyne, UK. Our results suggest that when disease risk is assumed to be a spatially continuous process, the proposed approximation to LGCP provides reliable estimates of disease risk both on spatially continuous and aggregated scales. The proposed methodology is implemented in the open-source R package `SDALGCP`.

Keywords: disease mapping; geostatistics; log-Gaussian Cox process; Monte Carlo maximum likelihood.

2.1 Introduction

In this paper our concern is to make inference on a spatially continuous disease risk surface using aggregated counts of reported disease cases, say y_i, over regions $\mathcal{R}_i$ forming a partition of a geographical area of interest A. In this context, information on risk factors and on the population at risk may also be available, possibly at different spatial scales. We shall denote these by $d(x)$ and $m(x)$, respectively, when available on a spatially continuous scale, and by d_i and m_i when they are spatially aggregated.

Existing methods from small area estimation (SAE) only allow spatial prediction at the aggregated level of the regions $\mathcal{R}_i$ and are usually based on a Gaussian Markov random field (GMRF) structure. (Besag, 1974; Rue and Held, 2005) Typically, non-zero elements of the precision matrix of a GMRF are restricted to contiguous pairs of the $\mathcal{R}_i$. Hence, the formulation and interpretation of a GMRF is tied to the specific partition of A, which will usually have been drawn up for administrative, historical, or other reasons unrelated to the disease aetiology. The use of such models also becomes impractical when the spatial units $\mathcal{R}_i$ change over time. Wall (2004) points out that the use of GMRFs is especially problematic when dealing with irregular geometries, which can induce counter-intuitive forms for the correlation structure between variables associated with the $\mathcal{R}_i$.

The geostatistical paradigm, unlike SAE, treats disease risk as a spatially continuous phenomenon irrespective of the data-format. Diggle et al. (2013) argue that the analysis of spatially aggregated counts can be regarded as a special case of the class of geostatistical problems and propose to model the y_i as an aggregated realisation of a Log-Gaussian Cox process (LGCP). Unlike GMRFs, LGCPs allow for prediction of disease risk at any spatial scale, while avoiding the ecological fallacy (Wakefield and Shaddick, 2006). However, fitting of LGCP models using the aggregated counts y_i is computationally demanding due to the iterative imputation of the unobserved locations for each reported case within a region $\mathcal{R}_i$ (Li et al., 2012).

In this paper, our objective is to develop a computationally efficient approximation to LGCPs in order to predict disease risk at any desired spatial scale. We argue that this provides a more realistic alternative to GMRF models when LGCPs are not computationally feasible, and can also be used as an exploratory tool in order to inform more complex modelling approaches based on LGCPs.

In Section 2.2 of the paper, we review existing methods for modelling spatially aggregated disease counts. In Section 2.3, we develop a computationally efficient spatially discrete approximation to LGCP models. In Section 2.4 we carry out a simulation study to investigate the predictive performance of the proposed approximation and compare this with an exact fitting algorithm for LGCP models. In Section 2.5 we show an application of the method to a data-set on primary biliary cirrhosis (PBC) incidence in Newcastle, UK. Section 2.6 is a concluding discussion on the advantages and limitations of the proposed approach.

The method has been implemented in the open-source R package `SDALGCP` (Johnson et al., 2018), available from the Comprehensive R Network Archive. The R code for reproducing the results of Section 2.5 is available as supplementary material.

2.2 Existing methods for modelling spatially aggregated disease count data

2.2.1 Gaussian Markov random field models

Let Y_i denote the reported disease count in region $\mathcal{R}_i$. Conditionally on a zero-mean Gaussian process $S = (S_1, \ldots, S_n)$, assume that the Y_i are mutually independent Poisson random variables with expectations

$$\lambda_i = m_i \exp\{d_i^\top \beta + S_i\}, i = 1, \ldots, n \tag{2.1}$$

where β is a vector of regression coefficients and m_i is the population count or a standardised expectation of the number of cases, taking into account the demographics of the population in subregion $\mathcal{R}_i$ but assuming that risk is otherwise spatially homogeneous. Spatially discrete models are then developed by specifying the precision matrix for the Gaussian process S. Here, we focus on the two most commonly used formulations for S, namely the conditional autoregressive (CAR) (Leroux et al., 2000) and intrinsic conditional autoregressive (ICAR) (Besag et al., 1991) models.

Let $i \sim j$ be a shorthand notation for "$\mathcal{R}_i$ and $\mathcal{R}_j$ are neighbours". A CAR model then assumes that

$$S_i | S_{-i} \sim N\left(\rho_c \sum_{j \sim i} c_{ij} S_j, \tau_i^2\right), \tag{2.2}$$

where $S_{-i} = \{S_j : j \neq i\}$, ρ_c is the spatial dependence parameter and c_{ij} are known quantities such that $c_{ij} \neq 0$ if and only if $j \sim i$ and $j \neq i$. It follows from Brook's Lemma (Brood, 1964) and the Hammersley-Clifford Theorem (Besag, 1974) that the joint distribution of S is a multivariate zero-mean Gaussian distribution with covariance matrix

$$(I - \rho_c C)^{-1}\tilde{D}, \tag{2.3}$$

where $\tilde{D} = \{\tau_1^2, \ldots, \tau_n^2\}$, while the specification of C is generally tied to the specific arrangement of the partition of the region of interest. The most common approach is to set $c_{ij} = 1$ if $j \sim i$ and 0 otherwise. The matrix in (2.3) is then a valid covariance matrix if $\xi_{max}^{-1} < \rho_c < \xi_{min}^{-1}$ (Cressie, 1993, pg. 472), where ξ_{min} and ξ_{max} are the minimum and maximum eigenvalues of C, respectively. Scaling of the matrix C so as to obtain a weighted average of the S_j in (2.2) also implies that $-1 < \rho_c < 1$.

The ICAR model is a special case of the CAR model when $\rho_c = 1$ in (2.2). Although this leads to an improper distribution for S because of the singularity of its covariance matrix, the associated conditional distribution of S given Y is proper.

2.2.2 Log-Gaussian Cox process models

A spatial point process is a stochastic mechanism that generates a countable set of events $x_i \in \mathbb{R}^2$. The class of inhomogeneous Poisson processes with intensity $\lambda(x)$ is defined by the following postulates.

1. The number of events, $N(\mathcal{A})$, in any planar region $\mathcal{A} \subset \mathbb{R}^2$ follows a Poisson distribution with mean $\int_{\mathcal{A}} \lambda(x)dx$.

2. Conditionally on $N(\mathcal{A})$, each event in $\mathcal{A}$ is an independent random sample from a distribution on $\mathcal{A}$ with probability density function proportional to $\lambda(x)$.

A Cox process (Cox, 1955) is defined by a non-negative valued stochastic process $\Lambda(x)$ such that, conditional on a realisation of $\Lambda(x)$, the process is an inhomogenous Poisson process with intensity $\Lambda(x)$. If we assume that $\log\{\Lambda(x)\} = S(x)$ is a

Gaussian process, we obtain the log-Gaussian Cox process (LGCP); for more details on the theoretical properties of LGCPs, see Møller et al. (1998).

Diggle et al. (2013) develop a modelling framework for aggregated disease count data using LGCPs. They assume that, conditionally on $S(x)$, the Y_i are mutually independent Poisson variables with means

$$\int_{\mathcal{R}_i} m(x) \exp\{d(x)^\top \beta + S(x)\}\, dx, \tag{2.4}$$

where $d(x)$ is a vector of covariates at location x with associated regression coefficients β.

A first notable difference between (2.1) and (2.4) is that the latter uses spatially continuous information on the distribution of the expected cases, $m(x)$, hence, unlike (2.1), avoids the questionable assumption of a homogeneous distribution of the population at risk within $\mathcal{R}_i$. However, population density is often only available in the form of small-area population counts, implying a piece-wise constant surface $m(x)$. Note, however, that modelled spatially continuous maps for population density have been made freely available; see, for example, `sedac.ciesin.columbia.edu/data/collection/`

Furthermore, unlike the spatially discrete models described in the previous section, LGCP is not tied to any particular partition of the area of interest and therefore provides a route to a solution to the problem of combining information at multiple spatial scales. However, this is offset by a substantial increase in the computational burden arising from the need to impute the unobserved locations for each of the reported cases within each of the $\mathcal{R}_i$, $i = 1, \ldots, n$ (Li et al., 2012). In the next section, we circumvent this issue by proposing a spatially discrete approximation to $S(x)$ which allows to model the counts y_i as the realisation of a Poisson log-linear mixed model.

2.3 A spatially discrete approximation to Log-Gaussian Cox processes

Let $w_i(x)$ be a positive function with domain $\mathcal{R}_i$, such that $\int_{\mathcal{R}_i} w_i(x)\,dx = 1$. Using the same notation as in Section 2.2.2, we approximate the conditional log-intensity of an LGCP as piecewise constant by taking its weighted average over $\mathcal{R}_i$ to give

$$\begin{aligned}\log\{\Lambda(x)\} &\approx \int_{\mathcal{R}_i} w_i(x)\left[d(x)^\top\beta^* + S^*(x)\right]\,dx \\ &= \int_{\mathcal{R}_i} w_i(x)\,d(x)^\top\beta^*\,dx + \int_{\mathcal{R}_i} w_i(x)\,S^*(x)\,dx \\ &= d_i^\top\beta^* + S_i^*,\ x \in \mathcal{R}_i, \qquad (2.5)\end{aligned}$$

where β^* is a vector of regression coefficients for the aggregate explanatory variables d_i and $S^*(x)$ is a Gaussian process. The rationale for using the weighting function $w_i(x)$ is to account for the potential non-homogeneous distribution of disease cases within a region $\mathcal{R}_i$. For example, a larger number of cases may concentrate in more densely populated areas, thus a natural choice for $w_i(x)$ would be to set this equal to $m(x)/m_i$ with $m_i = \int_{\mathcal{R}_i} m(x)dx$, if $m(x)$ is available. If $m(x)$ is instead unavailable, a pragmatic approach would be to set $w_i(x) = 1/|\mathcal{R}_i|$.

Following from (2.5), we obtain the following approximation for the conditional mean of the counts Y_i

$$\begin{aligned}\lambda_i = \int_{\mathcal{R}_i} m(x)\Lambda(x)\,dx &\approx \int_{\mathcal{R}_i} m(x)\exp\left\{d_i^\top\beta^* + S_i^*\right\}\,dx \\ &= m_i\exp\{d_i^\top\beta^* + S_i^*\} \\ &= m_i\exp\{\eta_i\} \\ &= \mu_i. \qquad (2.6)\end{aligned}$$

The joint distribution of $S^* = (S_1^*, \ldots, S_n^*)$ is multivariate Gaussian with zero mean and covariance function

$$\mathrm{Cov}\{S_i^*, S_j^*\} = \sigma^2 \int_{\mathcal{R}_i} \int_{\mathcal{R}_j} w_i(x) w_j(x') \, \rho(\|x - x'\|; \phi) \, dx \, dx', \tag{2.7}$$

where $\| \cdot \|$ is the Euclidean distance and $\rho(\cdot; \phi)$ is the isotropic and stationary covariance function of $S^*(x)$ indexed by the parameter ϕ. Hence, the resulting model (2.6) falls under the class of generalized linear mixed models. Also, note that the variance of S_i^* depends on the size and shape of $\mathcal{R}_i$, with larger regions leading to smaller variances.

We now provide further details on the computation of the covariance function in (2.7). Among the class of isotropic and stationary covariance functions for $S^*(x)$ in (2.6), one of the most commonly used is the Matérn covariance function,(Stein, 2012) which has expression

$$\mathrm{Cov}\{S^*(x), S^*(x')\} = \frac{\sigma^2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{u}{\phi}\right)^{\kappa} \mathcal{K}_\kappa \left(\frac{u}{\phi}\right), \tag{2.8}$$

where $u = \|x - x'\|$ is the Euclidean distance between any two locations x and x', σ^2 is the variance, ϕ is a scale parameter that regulates the rate at which the spatial correlation decays for increasing distance u, κ is the shape parameter that determines the differentiability of the process S and $\mathcal{K}_\kappa(\cdot)$ is the modified Bessel function of the second kind of order $\kappa > 0$. Estimating κ reliably requires a large amount of densely sampled data, which in this context is not available. As shown by Zhang (2004), not all of the three parameters σ^2, ϕ and κ can be consistently estimated under in-fill asymptotics, and in practice this translates to κ often being poorly identified. This issue is likely to be further exacerbated in this context. As a pragmatic approach, we then set $\kappa = 0.5$ which reduces (2.8) to

$$\mathrm{Cov}\{S^*(x), S^*(x')\} = \sigma^2 \exp\{-u/\phi\}$$

corresponding to a mean-square continuous process. However, in our application, we tried $\kappa = 1.5$ and $\kappa = 2.5$ and it gives similar prediction.

We approximate (2.7) as a discrete sum over L_i and L_j randomly chosen points in $\mathcal{R}_i$ and $\mathcal{R}_j$ to give

$$\int_{\mathcal{R}_i} \int_{\mathcal{R}_j} w_i(x) w_j(x')\, \rho(\|x - x'\|; \phi)\, dx\, dx' \approx \frac{\sum_{k=1}^{L_i} \sum_{k'=1}^{L_j} w_i(x_k) w_j(x_{k'})\, \rho(\|x_k - x_{k'}\|; \phi)}{\sum_{k=1}^{L_i} \sum_{k'=1}^{L_j} w_i(x_k) w_j(x_{k'})}, \tag{2.9}$$

To attain a good spatial coverage of $\mathcal{R}_i$ and $\mathcal{R}_j$, we propose to draw each of the x_k and $x_{k'}$ in the above equation using a class of inhibition processes (Diggle, 2013, pp. 110-116) which combine simple sequential inhibition with rejection sampling. More specifically, we proceed through the following steps.

1. Compute $w_{max} = \max_{x \in \mathcal{R}_i} w_i(x)$.
2. Generate x_{prop} over $\mathcal{R}_i$ from a homogeneous Poisson process with intensity w_{max}.
3. Compute $p(x_{prop}) = w_i(x_{prop})/w_{max}$.
4. Generate a sample u from the uniform distribution on $(0, 1)$.
5. If $k = 1$, set $x_1 = x_{prop}$ if $u \leq p(x_{prop})$; for $k > 1$ and given $\{x_j : j = 1, \ldots, k-1\}$, set $x_k = x_{prop}$ if $u \leq p(x_{prop})$ and x_{prop} falls at the intersection of $\mathcal{R}_i$ and $\{x \in \mathcal{R}_i : \|x - x_j\| > \delta(1 - w(x_j)/w_{max})\}$. Otherwise, reject x_{prop}.
6. Repeat 2 to 5, until $k = L_i$.

To identify a suitable value for L_i (the total number of generated points within $\mathcal{R}_i$), a possible solution is to use the packing density for a sequential inhibitory point process given by

$$\gamma = \frac{L_i \pi \delta^2}{4|\mathcal{R}_i|}, \tag{2.10}$$

where δ is the minimum permissible distance between points. The maximum possible value for γ is obtained by close-packed discs whose centres form an equilateral triangular lattice with sides of length $\delta = \pi/\sqrt{12}$. Through a simulation study, Tanemura (1979) suggested to set $\gamma = 0.55$ in order to achieve good spatial coverage in a relatively small number of iterations. Once γ and δ are fixed, we can then obtain L_i through equation (2.10).

An alternative solution is to leave choose γ as a function of ϕ using the following adaptive algorithm.

1. For a given ϕ, initialize a batch size k and a relative tolerance ϵ;
2. Locate k quadrature points with packing intensity $\gamma(k) = k\pi\delta^2/4|\mathcal{R}_i|$, evaluate the integral in (2.9) and denote its value as I_{old};
3. Add k points using a packing intensity $\gamma(k)/2$, re-evaluate the integral and denote its value as I_{new};
4. If $I_{new} = I_{old}$, stop the algorithm. Otherwise, set $I_{new} = I_{old}$, add k points with $\gamma(k)/3$ and repeat until $|I_{old} - I_{new}| < \epsilon|I_{new}|$.

Since ϕ is almost always unknown, the adaptive algorithm becomes more computationally demanding, especially in the case of a large number of regions in the study domain and for small values of ϕ which require a finer grid for a satisfactory approximation of (2.7). When fitting the model in (2.6) (see next section for more details), our recommendation is to use the non-adaptive algorithm first, in order to locate the likely value of ϕ, and then to run a final estimation using the adaptive algorithm. In the application in Section 2.5, the adaptive algorithm increases the elapsed time by about 10 minutes (592 seconds) on a laptop with 7.6GiB memory and 2.40GHz × 4 processor. Furthermore, in order to reduce the computational burden, we propose to discretise ϕ over a finite set of values and pre-compute the covariance matrix as defined by (2.9) for each of the pre-defined values. To obtain a 95% confidence interval for ϕ, we then compute the profile likelihood over the discrete set and interpolate it using a natural cubic spline. In our experience, the fineness of the discretisation does not have tangible effects on the spatial predictions but, instead, directly affects the goodness of the numerical approximation of the 95% confidence interval based on the profile likelihood.

2.3.1 Monte Carlo maximum likelihood

We carry out parameter estimation for the model in (2.6) using the Monte Carlo maximum likelihood (MCML) method (Christensen, 2012).

Let $f(\cdot)$ be a shorthand notation for "the density function of $\cdot$". Let $y^\top = (y_1, \ldots, y_n)$ and linear predictor $\eta^\top = (\eta_1, \ldots, \eta_n)$; it then follows that conditionally on $S^* = (S_1^*, \ldots, S_n^*)^\top$, the joint distribution of Y is

$$f(y|\eta) = \prod_{i=1}^{n} f(y_i|\eta_i),$$

where

$$f(y_i|\eta_i) \propto \exp\{y_i \log \mu_i - \mu_i\}.$$

Let $\psi = (\beta, \sigma^2, \phi)$ denote the vector of the model parameters, then the likelihood function for ψ is obtained by integrating out S^*, i.e.

$$L(\psi) \;\; = \;\; \int_{\mathbb{R}^n} f(y|\eta) \; f(\eta;\psi) \, d\eta. \tag{2.11}$$

In (2.11) $f(\eta;\psi)$ is a multivariate Gaussian distribution function with mean $D\beta$, where D denotes a matrix of explanatory variables, and covariance matrix $\varSigma$, whose (i,j)-th entry is given by (2.7). To reduce the computational burden accrued from the numerical approximation (2.9), we restrict the maximization of (2.11) to a finite set of predefined values for ϕ and, for each of these, pre-compute the covariance matrix Σ together with its inverse, determinant and Cholesky decomposition.

Since the high-dimensional integral in (2.11) cannot be solved analytically, we use Monte Carlo methods for the approximation of the likelihood. Let ψ_0 denote our best guess of ψ. We re-write (2.11) as

$$L(\psi) \quad \propto \quad E_{\eta|y}\left[\frac{f(\eta;\psi)}{f(\eta;\psi_0)}\right], \tag{2.12}$$

where the expectation E is taken with respect to the conditional distribution of η given y with parameters vector ψ_0. We provide the proof of this in Appendix A.1 of the supplementary material.

To generate N samples, say $\eta_{(j)}$, from the conditional distribution of η given y, we use a Monte Carlo Markov chain (MCMC) algorithm implemented in the `Laplace.sampling.MCML` function in the PrevMap package(Giorgi and Diggle, 2017). This function uses a Metropolis-adjusted Langevin MCMC algorithm to update the standardised vector of random effects, $\tilde{\eta} = \hat{\Sigma}^{-\frac{1}{2}}(\eta - \hat{\eta})$, where $\hat{\eta}$ and $\hat{\Sigma}$ are the mode and the inverse of the negative Hessian of $f(\eta;\psi_0)$ at $\hat{\eta}$. We can then approximate the likelihood function in (2.12) as

$$L(\psi) \approx L_N(\psi) \quad = \quad \frac{1}{N}\sum_{j=1}^{N}\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}. \tag{2.13}$$

As $N \to \infty$, in the above equation, $L_N(\psi)$ converges to $L(\psi)$. Geyer (1994, 1996); Geyer and Thompson (1992)

Finally, we maximize (5) using a constrained quasi-Newton optimization algorithm, implemented in the `nlminb` function in the R software environment, by providing analytical expressions for the first and second derivatives of (5) with respect to ψ.

If $\hat{\psi}_N$ denote the resulting MCML estimate, we then set $\psi_0 = \hat{\psi}_N$ and repeat the previous steps until convergence.

2.3.2 Continuous spatial prediction

We now consider the problem of carrying out spatial prediction of $S^*(x)$ at a pre-defined location x within the study area A. Using the same notation as in the previous section, we first note that

$$\begin{aligned} f(S^*(x)|y) &= \int_{\mathbb{R}^n} f(\eta, S^*(x)|y)\, d\eta \\ &= \int_{\mathbb{R}^n} f(\eta|y) f(S^*(x)|\eta, y)\, d\eta \\ &= \int_{\mathbb{R}^n} f(\eta|y) f(S^*(x)|\eta)\, d\eta. \end{aligned} \tag{2.14}$$

Hence, we sample from $f(S^*(x)|y)$ using the following two-step procedure: (1) draw samples $\eta_{(j)}$, for $j = 1, \dots, N$ from $f(\eta|y)$ using the MCMC algorithm described in the previous section; (2) for each $\eta_{(j)}$, for $j = 1, \dots, N$ simulate from $f(S^*(x)|\eta_{(j)})$, a Gaussian distribution with mean $\mu^*(x) = c(x)^\top \Sigma^{-1}(\eta_{(j)} - D\beta)$ and variance $v^2(x) = \sigma^2 - c(x)^\top \Sigma^{-1} c(x)$, where $c(x)^\top = (c_1(x), \dots, c_n(x))$, $c_i(x) = \sigma^2 \int_{\mathcal{R}_i} w(x)\rho(\|x - x'\|)\, dx'$, and we use (2.9) to approximate the integral. The resulting samples from $f(\eta|y)$ can then be used to compute non-linear properties of $S^*(x)$ and to summarise these using, for example, predictive means and standard errors.

2.4 Simulation Study

We now conduct a simulation study to assess the predictive performance of the proposed approximation in (2.3) when the underlying process is an LGCP model.

We simulate $B = 1,000$ data-set of counts using the administrative boundaries of the lower layer super output areas (LSOAs) in Newcastle-Upon-Tyne, UK, as in the application of Section 2.5. We specify the offsets $m(x)$ using population density estimates from the OpenPopGrid database (Murdock et al., 2015) and simulate the locations of the events using an inhomogeneous Poisson process with intensity $m(x)\exp\{S(x)\}$. We define three scenarios by setting the standard deviation of the Gaussian random field $S(x)$ to $\sigma = 0.706$ and let ϕ (whose unit of measure is metres) vary over the set $\{100, 800, 1500\}$, which correspond to a case of small, medium and large spatial correlation, respectively. The value of the standard deviation corresponds to the posterior mean obtained from the fitted LGCP in the application to primary biliary cirrhosis data described in the next section. Finally, for each of the $1,000$ simulated data-sets of aggregated counts at LSOA level, we fit the following models.

- *LGCP.* We use a Bayesian data augmentation technique implemented in the `lgcp` package (Taylor et al., 2015). We overlay a computational grid at a spacing of of 300×300 metres onto the area of interest and fit the model in (2.4). We run 3,100,000 iterations of the MCMC algorithm with a burn-in of 100,000 samples and then retain every 300-th sample.

- *Spatially discrete approximation (SDA) to LGCP.* We fit the approximation in (2.3) using a population weighted average (SDA I, with $w_i(x) = m(x)/m_i$) and simple average (SDA II, with $w_i(x) = 1/|\mathcal{R}_i|$) of the log-intensity. For both, we use the MCML method described in Section 2.3.1 and run 110,000 iterations of the MCMC algorithm with a burn-in of 10,000 samples and then retain every 10-th sample.

We summarise the results from the simulation study through the bias, root-mean-square-error (RMSE), width of the predictive interval (WPI) and the 95% coverage probability (CP) for the incidence at LSOA level, λ_i, and for the spatially continuous relative risk, $\exp\{S(x)\}$. Let $\lambda_i^{(j)}$ denote the true simulated incidence for $\mathcal{R}_i$ at the j-th simulation; hence

$$BIAS = \frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}(\hat{\lambda}_i^{(j)} - \lambda_i^{(j)}),$$

$$RMSE = \sqrt{\frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}(\hat{\lambda}_i^{(j)} - \lambda_i^{(j)})^2},$$

$$WPI = \frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}(PI_{0.95,U}^{(j)} - PI_{0.95,L}^{(j)}),$$

$$CP = \frac{1}{nB}\sum_{i=1}^{n}\sum_{j=1}^{B}I(\lambda_i^{(j)} \in PI_{0.95}^{(j)}),$$

where $\hat{\lambda}_i^{(j)}$ is the mean of the predictive distribution for $\lambda_i^{(j)}$, $I(\lambda_i^{(j)} \in PI_{0.95}^{(j)})$ is an indicator function that takes value 1 if $\lambda_i^{(j)}$ falls inside the 95% prediction interval and 0 otherwise, and $PI_{0.95,U}^{(j)}$ and $PI_{0.95,L}^{(j)}$ are the upper and lower limits of the 95% prediction interval, respectively. Similarly, we compute the three indices for the relative risk $\exp\{S(x)\}$ by averaging each of these over the regular grid at a spacing of 300 metres covering the whole of Newcastle-Upon-Tyne, UK.

Table 2.1 reports the results for the prediction of λ_i, the incidence at LSOA level. We observe that SDA I and II have a slightly lower bias and RMSE than LGCP in all three scenarios, with SDA I having the best performance. The coverage probability is close to the 95% nominal level and the WPI is comparable for all three models.

The results for the spatially continuous relative risk, $\exp\{S(x)\}$, are shown in Table 2.2. LGCP has the lowest bias and RMSE followed by SDA I in all three scenarios, with larger differences for $\phi = 800$ and $\phi = 1500$. Both SDA I and II are more conservative than LGCP and provide prediction intervals with a larger coverage than the nominal level, as the result of a large RMSE. We also observe that the use of the population weighted average in SDA I leads to a tangible reduction in RMSE and bias with respect to SDA II.

Table 2.1: Average bias, root-mean-square-error (RMSE), width of the 95% prediction interval (WPI) and the 95% coverage probability (CP) for the LSOA incidence, λ_i, from the simulation study of Section 2.4.

	$\phi = 100$			$\phi = 800$			$\phi = 1500$		
	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP
Bias	-0.006	-0.007	-0.009	-0.002	-0.003	-0.004	-0.008	-0.008	-0.011
RMSE	0.020	0.021	0.022	0.003	0.004	0.006	0.027	0.029	0.030
WPI	0.015	0.016	0.017	0.002	0.003	0.004	0.026	0.028	0.028
95%CP	0.940	0.942	0.948	0.942	0.943	0.952	0.943	0.944	0.945

Table 2.2: Average bias, root-mean-square-error (RMSE), width of the 95% prediction interval (WPI) and the 95% coverage probability (CP) for the spatially continuous relative risk, $\exp\{S(x)\}$, from the simulation study of Section 2.4.

	$\phi = 100$			$\phi = 825$			$\phi = 1500$		
	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP	SDA I	SDA II	LGCP
Bias	-0.575	-0.582	-0.566	0.842	0.965	-0.108	0.299	0.316	0.227
RMSE	2.590	2.800	0.045	0.439	0.531	0.005	2.070	2.260	0.002
WPI	2.525	2.739	0.564	0.719	0.806	0.108	2.048	2.238	0.227
95%CP	0.988	0.990	0.940	0.979	0.983	0.948	0.975	0.982	0.942

2.5 Application: mapping of primary biliary cirrhosis risk

We analyse incidence data on primary biliary cirrhosis (PBC) in Newcastle-Upon-Tyne, UK, obtained from the original study carried out by Prince et al. (2001); the data-set is freely available from the `lgcp` R package. The data consist of geo-referenced cases of definite or probable PBC between 1987 and 1994. The objective of this analysis is to quantify the difference in the predictive inferences between the gold-standard LGCP model and the proposed spatially discrete approximation (or SDA), on PBC incidence at LSOA level and the spatially continuous relative risk surface. In the case of SDA, we fit the population weighted (SDA I) and simple average (SDA II) versions described in the previous section. We also consider the exponential variogram (EV) model proposed by Wall (2004) consisting of a geostatistical Poisson model for the counts whose spatial structure is defined using the centroids of each LSOA. Finally, we fit the Besag et al. (1991) (BYM) model, one of most commonly used approaches in small area estimation, with linear predictor

$$\log \lambda_i = d_i^\top \beta^* + S_i + Z_i$$

where S_i is a zero-mean intrinsically autoregressive process with variance σ^2 and Z_i is Gaussian noise with variance τ^2.

In all five models, we use the index of multiple deprivation (IMD) as a covariate of the linear predictor. The IMD is publicly available from the UK Government online archives (`webarchive.nationalarchives.gov.uk`). The regression coefficients for the IMD are denoted by β_i in the LGCP model and by β_i^* in the BYM, EV and SDA models, with $i = 0$ corresponding to the intercept and $i = 1$ the effect of IMD.

For the SDA models, we run 110,000 iterations of the MCMC algorithm with a burn-in of 10,000 samples, and then retain every 10-th sample. We discretise ϕ using 100 equally spaced values between 50 and 2000 meters.

For the LGCP model, we specify independent priors as follows: $\log \sigma \sim N(\log 1, 0.15)$, $\log \phi \sim N(\log 500, 2)$ and $(\beta_0, \ldots, \beta_7) \sim MVN(0, 10^6 I)$. We run 3,100,000 iterations of the MCMC algorithm with a burn-in of 100,000 samples and retain every 3000-th sample so as to obtain a set of 1,000 weakly dependent samples.

Fitting of the BYM model using `CARBayes` (Lee, 2013) is carried out by iterating the MCMC algorithm 1,100,000 times with a burn-in of 100,000 samples and retaining every 100-th sample.

Finally, for the EV model which fit using the `spBayes` (Finley et al., 2007) R package, we specify an conjugate inverse-Gamma prior on the variance parameter σ^2 with shape parameter 1 and scale parameter 2. The spatial scale parameter ϕ is assigned a uniform prior in the interval $[50, 2500]$. For the regression coefficients β, we use a flat prior. We run 1,100,000 iterations of the MCMC algorithm with a burn-in of 100,000 samples and retain every 40-th sample.

A conjugate inverse-gamma prior was specified for the variance parameter σ. Generally, conjugate priors have appealing computational properties and for this reason it is widely used in practice. Also, we used a weakly informative prior for β because it allows the likelihood to dominate if there is a reasonably large information in the data. No formal sensitivity analysis was done on the prior, however, through the simulation study in section 2.4, we demonstrated that the models have a good coverage probability. Meaning that the posterior distribution is not strongly influenced by our choice of prior.

Trace-plots and correlograms are used to assess convergence of the MCMC algorithms in each of the fitted models. These are reported in the Appendix, from section A.2 to A.5, and all indicate a good mixing of the resulting MCMC samples. Note that we run a larger MCMC run for LGCP model because the target is high dimensional as there are 8196 parameters to estimate.

Tables 2.3 reports the point and interval estimates for the parameters of each of the fitted models. We observe that the differences amongst the point estimates of the regression coefficients from the five models are small.

Figure 2.1 shows a map of the estimated PBC incidence at LSOA level from the five models. The spatial spatial pattern estimated by each of these is comparable, as indicated by the scatter plots of Figure 2.3. The same consideration holds for the predictive standard errors (Figures 2.4 and 2.2). More specifically, the estimated incidence from the LGCP model has a correlation of about 0.7 with the other models, expect the BYM model for which the correlation is about 0.6. The good performance of the EV model can be explained by the fact that, in this scenario, the size of most of the LSOAs is small relative to the range of the spatial correlation, hence the use of the centroid becomes less problematic.

Figure 2.5 shows the map of the estimated continuous relative risk surface $\exp\{S(x)\}$ over a 300×300 meters regular grid covering the whole of the study area and Figure 2.6 shows its standard error. The scatter plots (Figures 2.7 and 2.8) indicate that the point estimates from the LGCP and the SDA approach are strongly similar, with a correlation of 0.862 between SDA I and LGCP and of 0.884 between SDA II and LGCP. However, we also observe that the standard errors from SDA, both I and II, are larger than those from LGCP. This is consistent with our results from the simulation study of the previous section.

2.6 Discussion

In this article we have developed a spatially discrete approximation (SDA) to log-Gaussian Cox process (LGCP) models in order to carry out spatial prediction of disease risk at any desired spatial scale using spatially aggregated disease count data.

As variation in disease risk occurs in a spatial continuum irrespective of the format in which the data are available, we consider the LGCP framework to be a natural statistical paradigm for modelling aggregated disease count data. However, when computational constraints make the fitting of an LGCP infeasible, we argue that SDA provides a computationally efficient solution while respecting the spatially continuous nature of disease risk. SDA also overcomes some of the limitations inherent to other spatially discrete models, such as CAR models. In addition to providing spatially continuous predictions, SDAs can also deal with the issue of changing administrative boundaries over time and allow incorporation of covariates available at any spatial scale.

Kelsall and Wakefield (2002) developed a similar approach to the proposed SDA for modelling count data available at areal level. Specifically, by assuming an intercept-only model, they approximate (2.4) using a multivariate log-Gaussian distribution with mean

$$E[\lambda_i] = \exp\{\beta_0 + \sigma^2/2\}$$

and covariance

$$\begin{aligned}\text{Cov}\{\lambda_i, \lambda_j\} &= \exp\{\beta_0 + \sigma^2/2\} \times \\ &\left[\int_{\mathcal{R}_i}\int_{\mathcal{R}_j} w_i(x)w_j(x')\exp\{\sigma^2\rho(\|x - x'\|;\phi)\}\,dx\,dx' - 1\right].\end{aligned}$$

Table 2.3: Point estimates and 95% confidence/credible intervals (CI) for the model parameters of the spatially discrete approximation to log-Gaussian Cox Process (LGCP) using a population-weighted log-intensity average (SDA I) and a simple average (SDA II), the exponential variogram (EV) model, Besag-York-Mollié (BYM) model and the LGCP model.

Model	**Parameter**	**Estimate**	**95% CI**
SDA I	σ^2	1.043	(0.907, 1.180)
	ϕ	742.857	(453.153, 1005.405)
	β_0^*	-8.080	(-8.248, -7.912)
	β_1^*	0.008	(0.004, 0.011)
SDA II	σ^2	1.020	(0.898, 1.142)
	ϕ	857.143	(489.590 1037.638)
	β_0^*	-7.876	(-8.029, -7.722)
	β_1^*	0.006	(0.002, 0.010)
EV	σ^2	0.316	(0.246, 0.369)
	ϕ	525.570	(367.719, 949.950)
	β_0^*	-8.069	(-8.177, -7.957)
	β_1^*	0.009	(0.006, 0.011)
BYM	τ^2	0.108	(0.012, 0.470)
	ν^2	0.023	(0.003, 0.173)
	β_0^*	-7.917	(-8.167, -7.694)
	β_1^*	0.007	(0.001, 0.014)
LGCP	σ^2	0.479	(0.237, 0.914)
	ϕ	1163.877	(528.618, 1967.756)
	β_0	-19.333	(-19.738, -19.013)
	β_1	0.008	(0.001, 0.015)

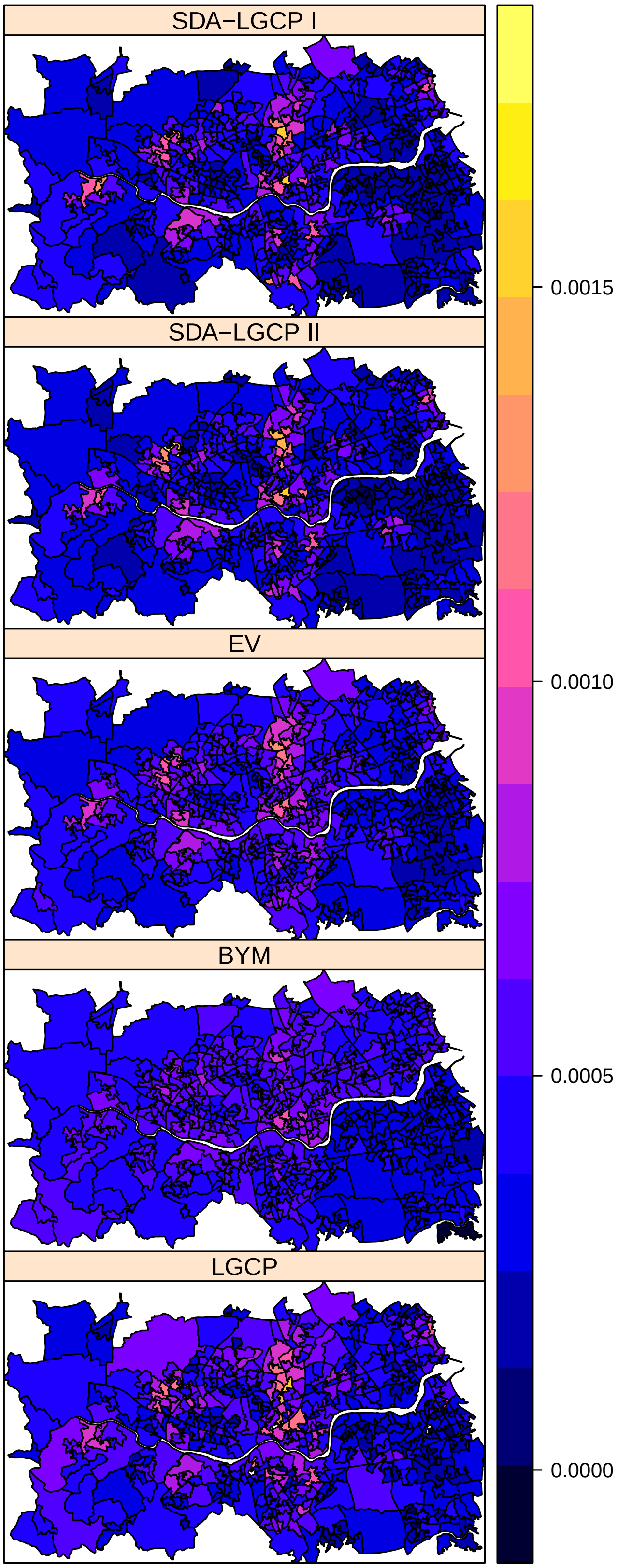


Figure 2.1: Maps of the estimated primary biliary cirrhosis (PBC) incidence in each lower layer super output area (LSOA) of Newcastle-Upon-Tyne from the four fitted models in Section 2.5.

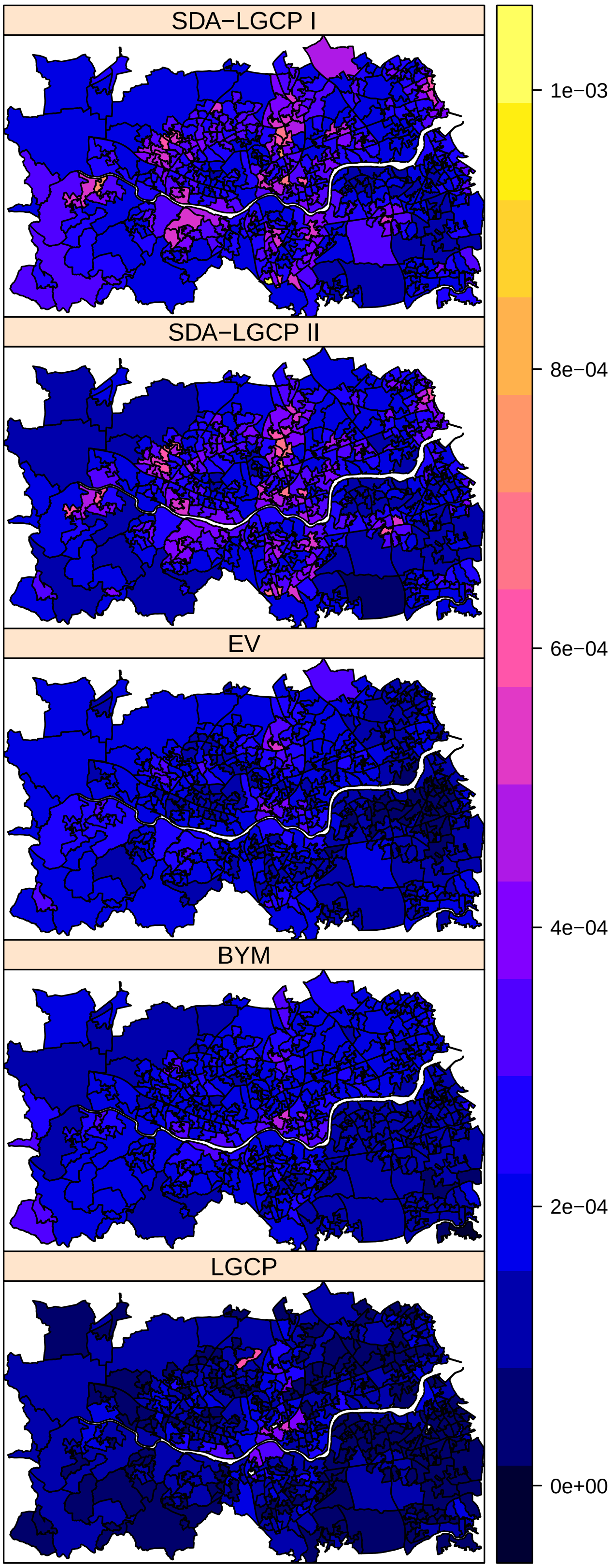


Figure 2.2: Maps of the standard error of the estimated primary biliary cirrhosis (PBC) incidence in each lower layer super output area (LSOA) of Newcastle-Upon-Tyne from the five fitted models in Section 2.5.

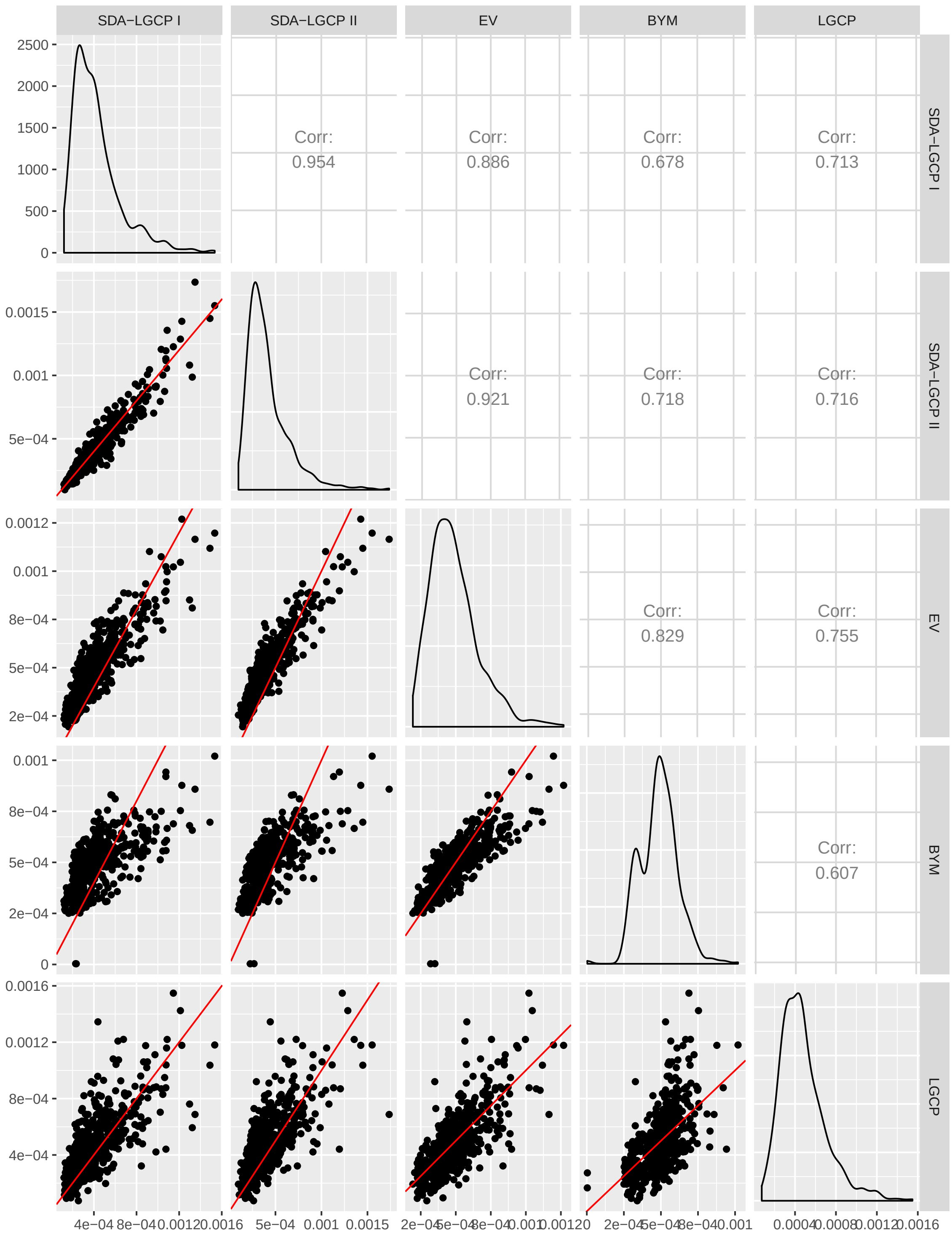


Figure 2.3: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the estimated primary biliary cirrhosis (PBC) incidence in the lower layer super output areas (LSOA) of Newcastle-Upon-Tyne for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the estimated PBC incidence from each model.

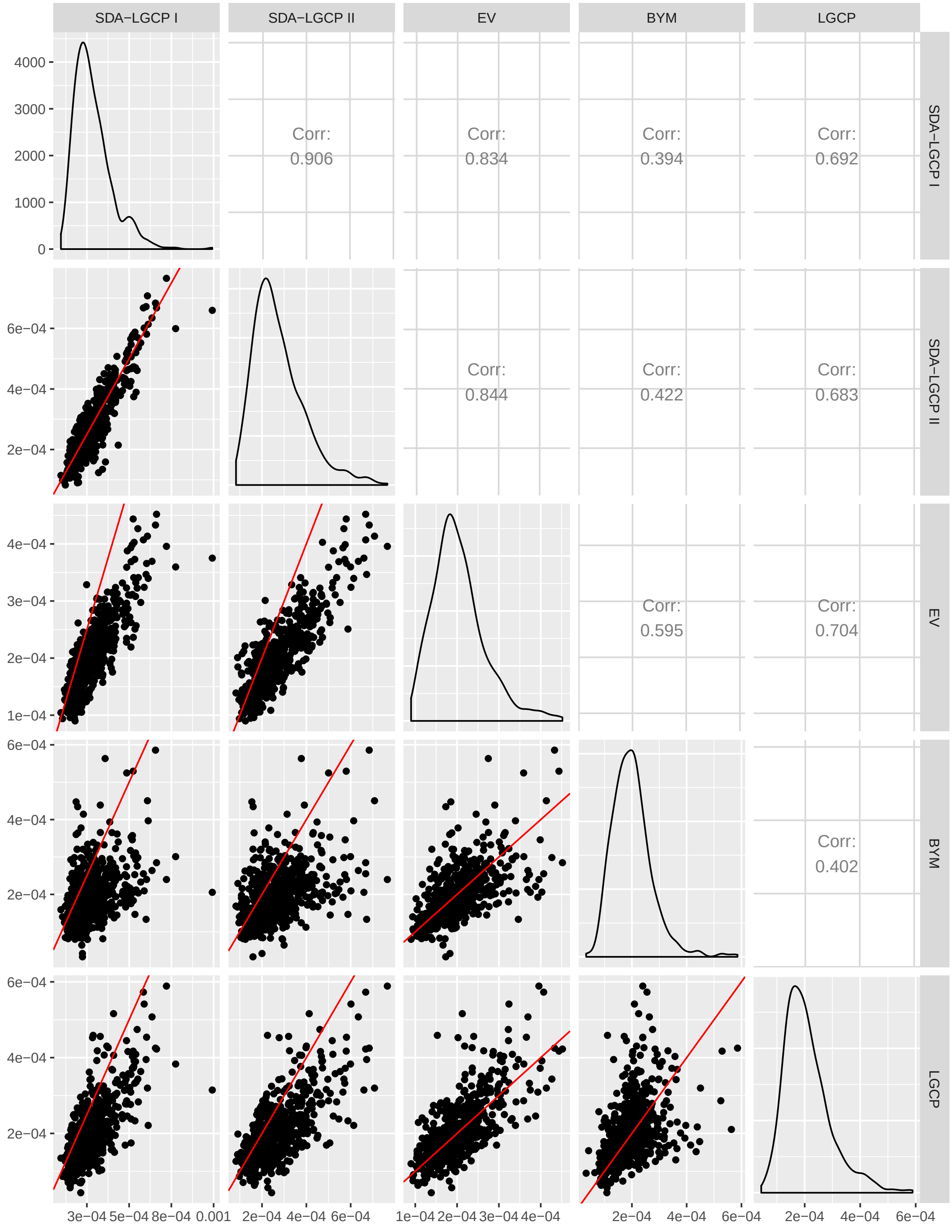


Figure 2.4: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the standard errors of primary biliary cirrhosis (PBC) incidence in the lower layer super output areas (LSOA) of Newcastle-Upon-Tyne for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the standard errors from each model.

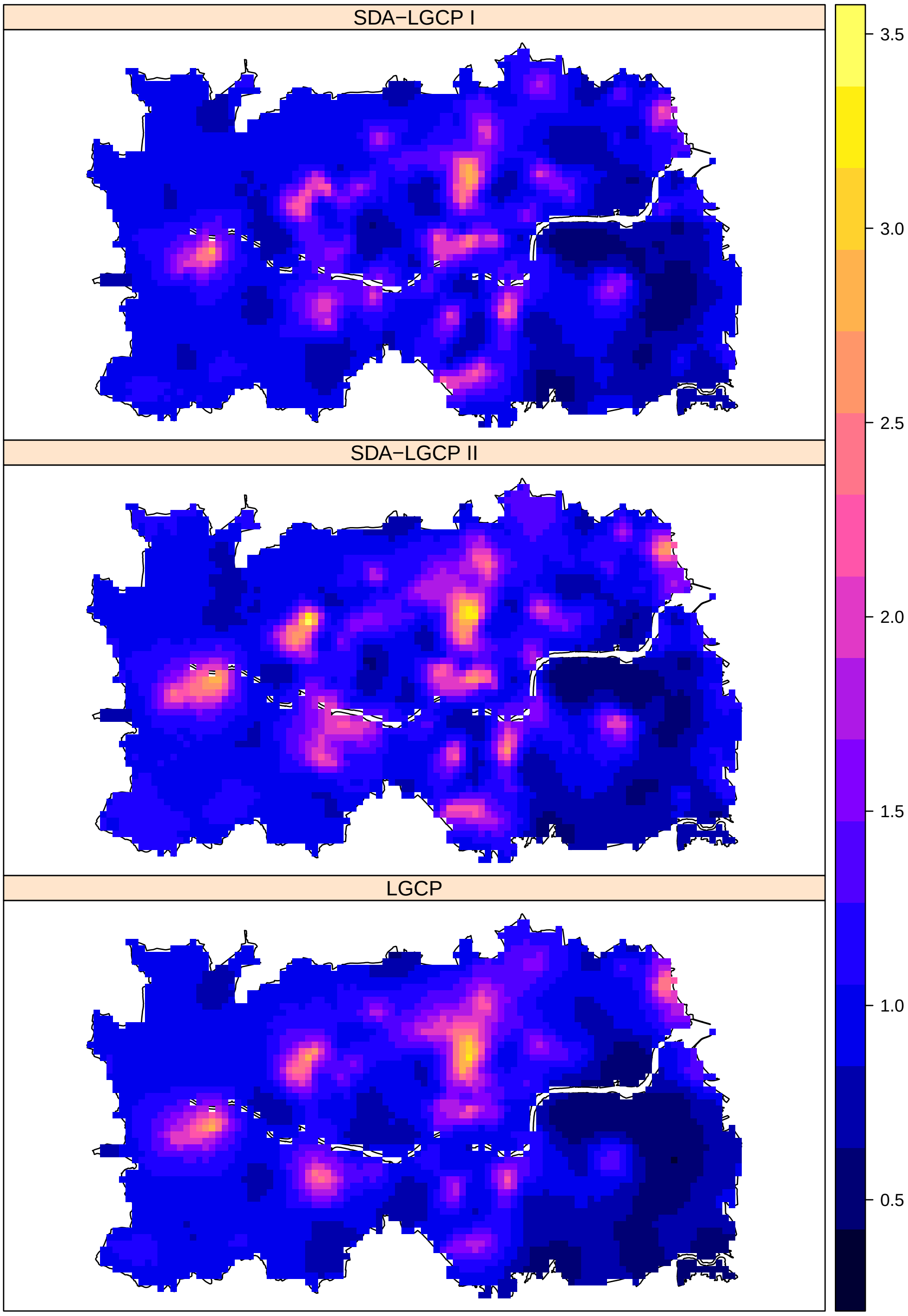


Figure 2.5: Maps of the predicted relative risk surface $\exp\{S(x)\}$ from the fitted spatially discrete approximation to log-Gaussian Cox Process (SDA) using a population-weighted log-intensity average (SDA I, upper panel) and a simple average (SDA II, middle panel), and the exact LGCP model (lower panel).

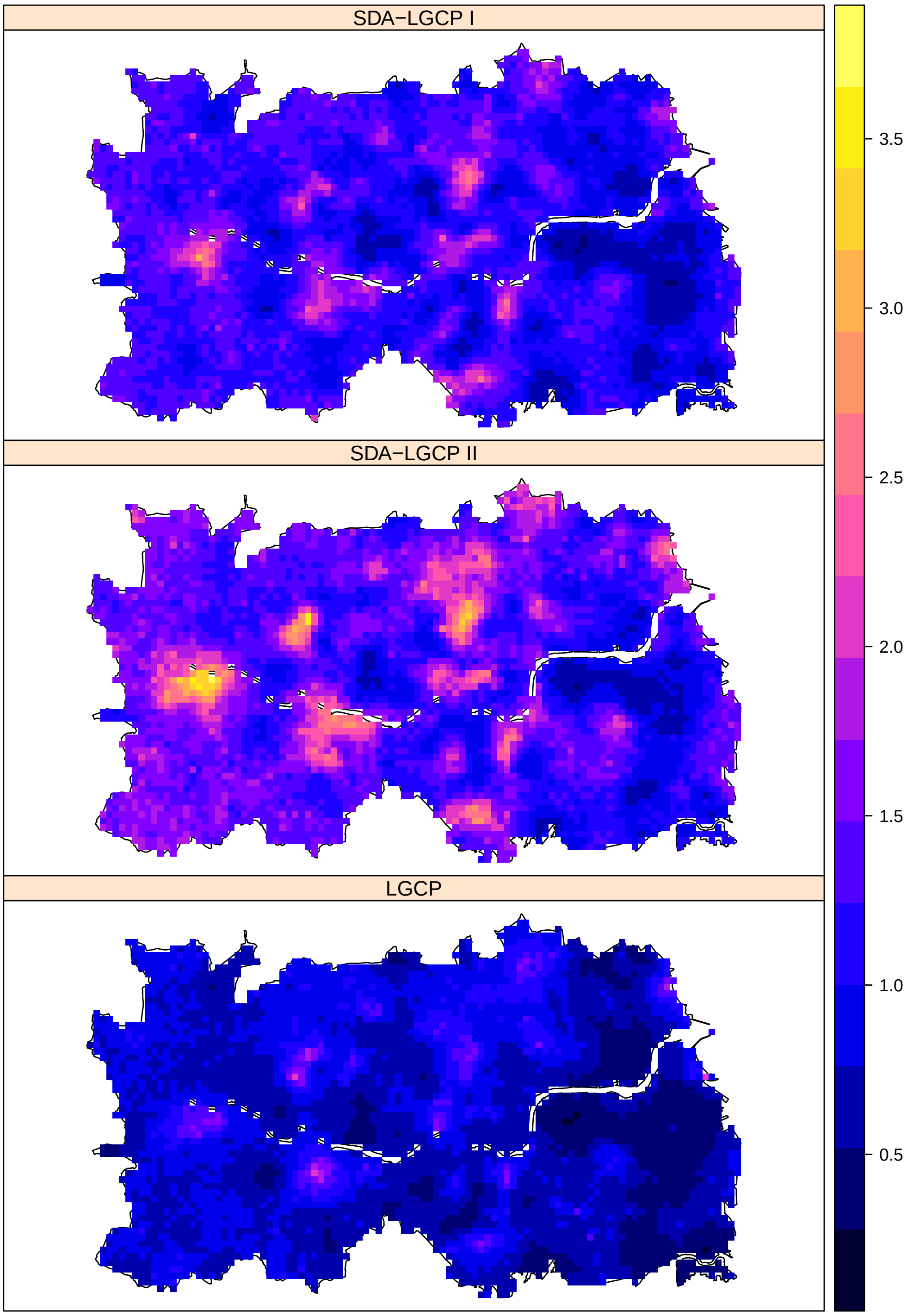


Figure 2.6: Maps of the standard error of the predicted relative risk surface $\exp\{S(x)\}$ from the fitted spatially discrete approximation to log-Gaussian Cox Process (SDA) using a population-weighted log-intensity average (SDA I, upper panel) and a simple average (SDA II, middle panel), and the exact LGCP model (lower panel).

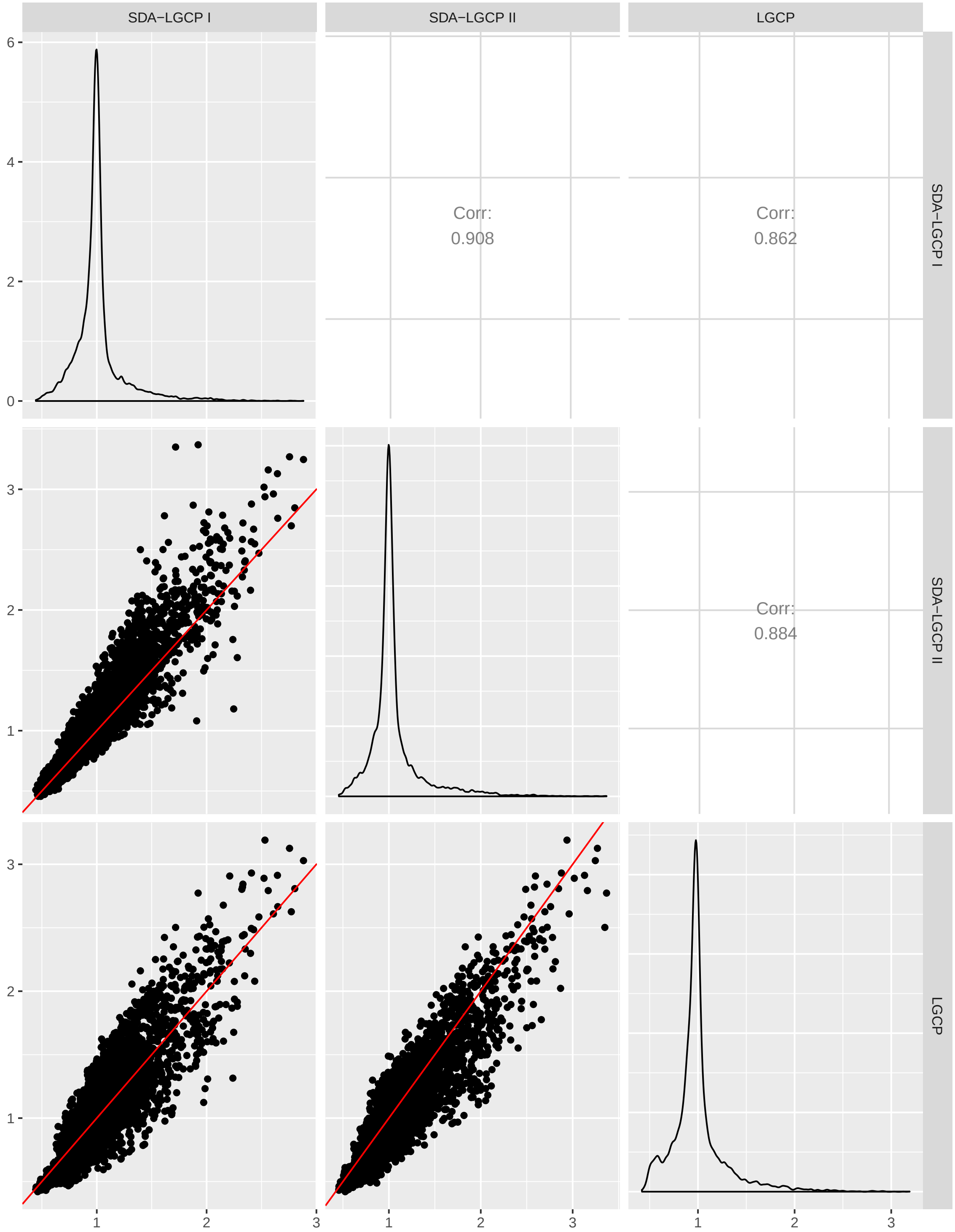


Figure 2.7: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the estimated spatially continuous relative risk $\exp\{S(x)\}$ for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the estimated relative risk from each model.

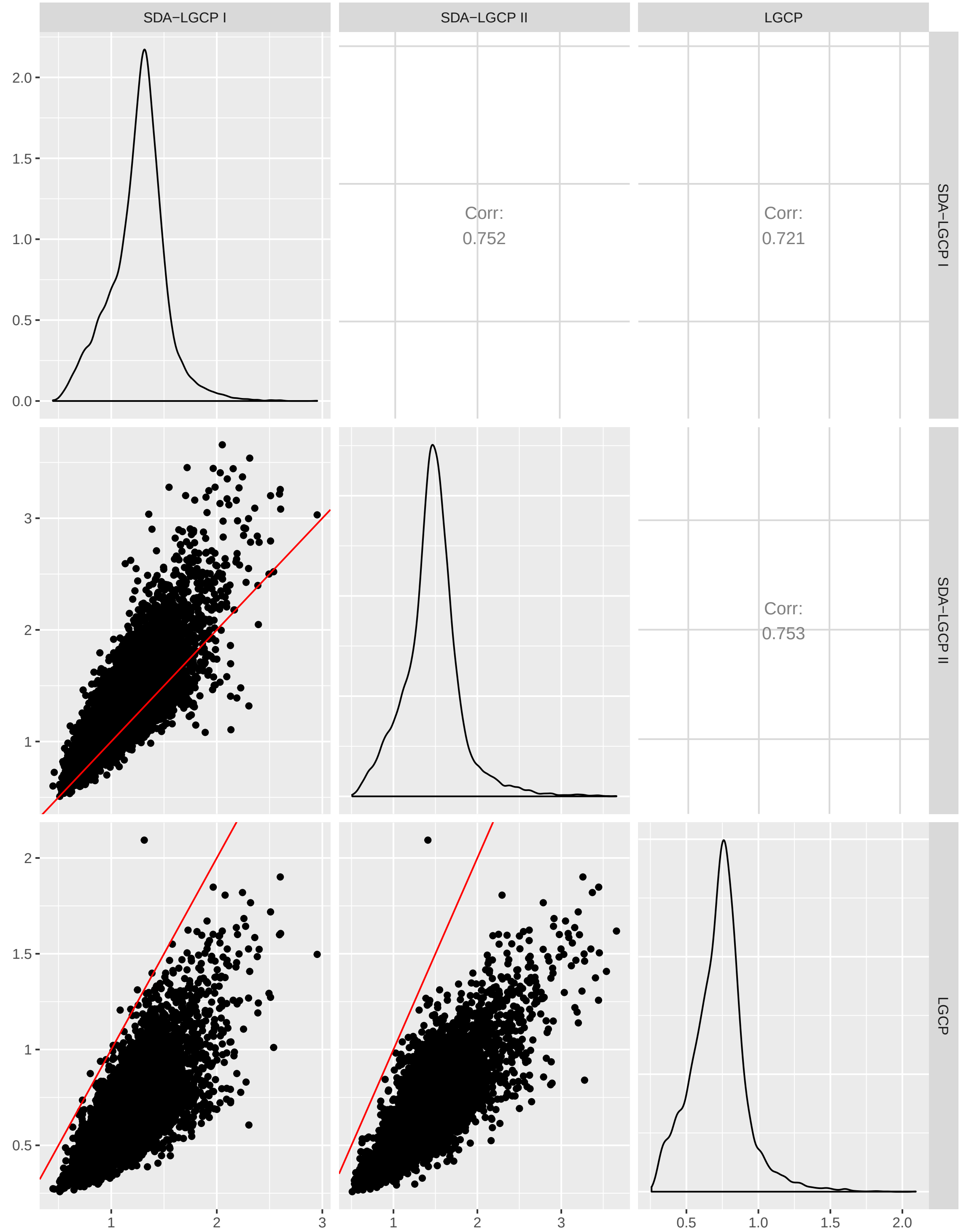


Figure 2.8: The lower and upper off-diagonal panels are scatter plots and correlation coefficients of the standard errors for the estimated risk $\exp\{S(x)\}$ for each pair of the fitted models in Section 2.5. The diagonal panels show smoothed histograms of the standard errors from each model.

Kelsall and Wakefield (2002) then advocate the use of the log-Gaussian approximation as a Bayesian prior for spatial smoothing but no reference is made to the LGCP framework. In this paper, instead, our objective was to develop a computationally efficient approximation to the LGCP model which, in Bayesian terms, is our chosen prior for modelling disease risk.

In fitting SDA models, most of the computational burden is due to the approximation of the integral in (2.7), which defines the area-level correlation between the spatial random effects. In our example, the SDA model is about 5 to 15 times faster to fit than the LGCP model, depending on the number of values used to discretise the scale of the spatial correlation ϕ. To make SDA even faster, efficient approximations to Gaussian processes should also be considered (see, for example, Lindgren et al. (2011)). These could be used to sample from the predictive distribution of $S^*(x)$ in (2.5) and avoid computation of the integral in (2.7).

We conclude that SDA is a reliable approximation to LGCP for carrying out predictions at areal-level, both in terms of point predictions and in the quantification of uncertainty. It also provides spatially continuous predictions in disease risk that are comparable to those from LGCP, but with larger standard errors and more conservative predictions intervals.

Finally, extension to the spatio-temporal case of the method discussed in this paper is possible and is work in progress. For example, let us consider counts y_{it} for the region $\mathcal{R}_i$ over the time interval $(t, t+1)$. Let $S(x,t)$ be a spatio-temporal Gaussian process with covariance function

$$\text{cov}\{S(x,t), S(x',t')\} = \sigma^2 \exp\{-|t-t'|/\psi\} \exp\{-\|x-x'\|/\phi\}.$$

By modelling the y_{it} as realisations of a spatio-temporal log-Gaussian Cox process with conditional intensity $\Lambda(x,t) = \exp\{\alpha + S(x,t)\}$, we can then approximate this with a spatio-temporally discrete Gaussian process $S^*_t = (S^*_{1t}, \ldots, S^*_{nt})$, such that

$$S^*_t = \varphi S^*_{t-1} + W_t, 0 < \varphi < 1,$$

where the temporal innovation W_t is modelled as a multivariate Gaussian distribution with covariance matrix given by (2.7). Preliminary results suggest that the reduction in computing time with respect to a spatio-temporal LGCP model is substantially larger than that observed for the purely spatial scenario presented in this paper.

Acknowledgements

We thank Dr Benjamin Taylor for his useful comment on the draft of this manuscript. We also acknowledge the travel grant from African Institute for Mathematical Sciences to OOJ.

Funding

OOJ holds a Connected Health Cities funded PhD studentship.

Bibliography

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Brood, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neihbor systems. *Biometrika*, 51:481–483.

Christensen, O. F. (2012). Monte carlo maximum likelihood in model-based geostatistics. *Journal of computational and graphical statistics.*

Cox, D. R. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164.

Cressie, N. A. (1993). Statistics for spatial data: Wiley series in probability and mathematical statistics. *Applied probability and statistics, rev. edn, New York: Wiley.*

Diggle, P. J. (2013). *Statistical analysis of spatial and spatio-temporal point patterns.* CRC Press.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Finley, A. O., Banerjee, S., and Carlin, B. P. (2007). spbayes: an r package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, 56:261–274.

Geyer, C. J. (1996). Estimation and optimization of functions. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice*, pages 241—258. London: Chapman and Hall.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54:657–699.

Giorgi, E. and Diggle, P. J. (2017). PrevMap: An R package for prevalence mapping. *Journal of Statistical Software*, 78(8):1–29.

Johnson, O., Giorgi, E., and Diggle, P. (2018). *SDALGCP: Spatially Discrete Approximation to Log-Gaussian Cox Processes for Aggregated Disease Count Data*. R package version 0.1.0.

Kelsall, J. and Wakefield, J. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 97(459):692–701.

Lee, D. (2013). Carbayes: an r package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24.

Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.

Li, Y., Brown, P., Gesink, D. C., and Rue, H. (2012). Log gaussian cox processes and spatially aggregated disease incidence data. *Statistical methods in medical research*, 21(5):479–507.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482.

Murdock, A., Harfoot, A., Martin, D., Cockings, S., and Hill, C. (2015). Openpopgrid: an open gridded population dataset for england and wales. openpopgrid.geodata.soton.ac.uk. GeoData, University of Southampton.

Prince, M. I., Chetwynd, A., Diggle, P., Jarner, M., Metcalf, J. V., and James, O. F. (2001). The geographical distribution of primary biliary cirrhosis in a well-defined cohort. *Hepatology*, 34(6):1083–1088.

Rue, H. and Held, L. (2005). *Gaussian random Markov fields.* Chapman & Hall/CRC.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging.* Springer Science & Business Media.

Tanemura, M. (1979). On random complete packing by discs. *Annals of the Institute of Statistical Mathematics*, 31(1):351–365.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. *Journal of Statistical Software*, 63:1–48.

Wakefield, J. and Shaddick, G. (2006). Health-exposure modeling and the ecological fallacy. *Biostatistics*, 7(3):438–455.

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of statistical planning and inference*, 121(2):311–324.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

Chapter 3

Dealing with spatial misalignment to model the relationship between deprivation and life expectancy in Liverpool: A model-based geostatistical approach

Olatunji Johnson, Peter Diggle, Emanuele Giorgi

CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK

Summary

Background

Life expectancy at birth (LEB), one of the main indicators of human longevity, has often been used to characterise the health status of a population. Understanding its relationships with the deprivation is key to develop policies and evaluate interventions that are aimed at reducing health inequalities. However, methodological challenges in the analysis of LEB data arise from the fact that different Government agencies often provide spatially aggregated information on LEB and the index of multiple deprivation (IMD) at different spatial scales. Our objective is to develop a geostatistical framework that, unlike existing methods of inference, allows to carry out spatially continuous prediction while dealing with spatial misalignment of the areal-level data.

Methods

We developed a model-based geostatistical approach for the joint analysis of LEB and IMD, when these are available over different partitions of the study region. We model the spatial correlation in LEB and IMD across the areal units using inter-point distances based on a regular grid covering the whole of the study area. The advantages and strengths of the new methodology are illustrated through an an analysis of LEB and IMD data from the Liverpool district council.

Results

We found that the effect of IMD on LEB is stronger in males than in females, explaining about 63.35% of the spatial variation in LEB in the former group and 38.92% in the latter. We also estimate that LEB is about 8.5 years lower between the most and least deprived area of Liverpool for men, and 7.1 years for women. Finally, we find that LEB, both in males and females, is at least 80% likely to be above the England wide average only in some areas falling in the electoral wards of Childwall, Woolton and Church.

Conclusion

The novel model-based geostatistical framework provides a feasible solution to the spatial misalignment problem. More importantly, the proposed methodology has the following advantages over existing methods: 1) it can deliver spatially continuous inferences using spatially aggregated data; 2) it can be applied to any form of misalignment with information provided at a range of spatial scales, from areal-level to pixel-level.

Keywords: deprivation; life expectancy; likelihood-based inference; model-based geostatistics; spatial misalignment; health inequality

3.1 Background

Over the last decades, access to better healthcare and education have led to a surge in human longevity, especially in high-income countries (Chetty et al., 2016; Kontis et al., 2017; Oeppen and Vaupel, 2002). Life expectancy at birth (LEB), one of the main indicators of human longevity, has often been used to characterise the health status of a population (OECD, 2017). Measuring deprivation is also important in order to describe health inequalities within a population and to better understand variation in health outcomes (Allik et al., 2016; Krieger et al., 2003). Previous studies have shown that the LEB is strongly affected by deprivation (Chetty et al., 2016; Tobias and Cheung, 2003; Woods et al., 2005) and that differences in LEB between most and least deprived individuals are larger among men than women (Auger et al., 2010; Tsimbos et al., 2014).

The main determinants of human longevity can be generally classified into social factors, genetic traits, life-style (e.g. consumption of tobacco, alcohol, dietary habits and physical activity) and environmental factors (e.g. overcrowded housing and pollution) (Christensen and Vaupel, 1996). As indices of deprivation are constructed by combining variables that are also likely determinants of human longevity, the reported associations with LEB are thus not surprising. However, linear regression models used to quantify the association between LEB and deprivation should also acknowledge the imperfect nature of the latter by making suitable distributional assumptions on the residuals of the model. Accounting for spatial correlation is especially important in this context so as to deliver reliable inferences on the regression relationship between LEB and deprivation. However, methodological challenges arise from the fact that different Government agencies often release spatially aggregated information on LEB and other socio-demographic variables, including deprivation, at different spatial scales. For example, in the UK, the Life Events and Population Sources Division of the Office for National Statistics releases information on LEB by Middle Super Output Area (MSOA) while the index of multiple deprivation (IMD), published by the Ministry of Housing, Communities and Local Government, is available at a higher spatial resolution by Lower Super Output Area (LSOA). An example of this is given by Figure 3.2 showing maps for male and female LEB and IMD in Liverpool, United Kingdom (UK). The rationale for calculating LEB at MSOA-level is that reliable estimates of LEB cannot be obtained from a population of less than 5000 individuals (Toson and Baker, 2003) and MSOAs satisfy this requirement, having 7200 inhabitants on average (Office for National Statistics, 2018).

In the recent paper by Buck et al. (2017), the authors investigate the association between LEB and IMD in England using a linear regression modelling framework. Their analysis is carried out at MSOA-level by taking the population-weighted average IMD based on the LSOAs falling in each of the corresponding MSOAs while assuming independent and identically distributed Gaussian residuals. This modelling approach ignores two important aspects: the within-MSOA variation which could result in a biased estimate for the regression coefficient associated with IMD; the residual spatial correlation in LEB, which affects the standard errors of the regression coefficient estimates (Thomson et al., 1999). Furthermore, the technique used by Buck et al. (2017) can only be reliably applied when spatial units at different scales are nested within each other.

The issue of spatial misalignment has been widely addressed in the statistical literature; see Gotway and Young (2002) and Banerjee et al. (2014) for an overview.

Our concern in this paper is with “areal-areal” misalignment, i.e. when data are available over misaligned, not necessarily nested, partitions of the same study area. A common approach used to address this problem is to predict the aggregated values of all the variables on a common set of spatial spatial units and use the resulting predictions to build a regression model; Buck et al. (2017) is an example of this. Madsen et al. (2008) refers to this strategy as “krige and regress”. They show that the estimator of the regression coefficient is consistent but the variance estimator can be biased. More rigorous approaches have been developed by joint modelling of the outcome variable and the covariates. For example, Agarwal et al. (2002) developed a joint model for outcomes observed at pixel-level and covariates at areal-level. The spatial correlation is modelled using conditionally autoregressive (CAR) models (Besag, 1974) for both pixel- and area-level spatial random effects. However, the use of CAR models for modelling outcomes aggregated over irregular spatial units (as in the case of LSOAs and MSOAs) is questionable because the adopted spatial structure is tied to the given partition of the study area, which is often drawn for administrative convenience. Also, Wall (2004) showed that when dealing with regions of varying size and shape, CAR models can induce counter-intuitive spatial correlation structure.

In this paper, our objectives are: 1) to develop a model-based geostatistical approach that allows the joint analysis of LEB and IMD data when these are available as spatially aggregated indices over misaligned partitions of the study area; 2) to carry out spatially continuous inference on LEB using spatially aggregated data. We illustrate our modelling approach through the analysis of LEB data from the Liverpool district council in the UK. Liverpool has been ranked as the most deprived local authority area in England in 2004, 2007 and 2010, and as the 4th most deprived in 2015 (Liverpool City Council, 2015). In 2018, LEB for both men and women was lower than the overall average in England (Public Health England, 2018). Understanding the relationship between deprivation and life expectancy within a single conurbation helps to develop policies and evaluate interventions that are aimed at reducing health inequalities (Bennett et al., 2018).

To address the aforementioned limitations of existing methods of inference, we develop a geostatistical framework that avoids the re-aggregation of IMD at MSOA-level. Instead, we jointly model LEB and IMD as aggregated outcomes of a spatially continuous stochastic process. More specifically, we model the spatial correlation across MSOAs for LEB and across LSOAs for IMD using inter-point distances based on a regular grid covering the whole of the study area. The methodology presented in this paper can also be used to model any spatially aggregated health outcome and estimate its association with risk factors that may be available at a range of spatial scales.

All the analyses presented in this paper have been developed R software environment (cran.r-project.org) and maps have been generate using the Q-GIS software (qgis.org). We provide the analysed data and the implemented R code in supplementary material.

3.2 Existing methods for analysing spatially misaligned data

Spatial misalignment has been well-studied in the literature, a good starting point the work done by Gotway and Young (2002) and Banerjee et al. (2014). Spatial misalignment can occur as point-point misalignment, point-areal misalignment areal-point misalignment and areal-areal misalignment. Geostatistical methods are popular solution to point-point misalignment, point-areal misalignment but its use for areal-areal and area-point misalignment is less studied. The common approach used to address areal-areal and area-point misalignment is conditional autoregressive (CAR) models (Besag, 1974). One of the limitations of this approach is inability to provide spatially continuous inference as they are tied to the data format.

In the sections that follow, we provide a review on model-based methods for spatially misaligned point-referenced data and spatially misaligned areal data using geostatistical methods and CAR models, respectively.

3.2.1 Geostatistical method for spatially misaligned point-referenced data

Let Y_i denote a continuous response variable, measured at a set of discrete locations, $\{x_i : i = 1, \ldots n\}$, where each x_i lies within a geographical region of interest, A and let D_k denote the predictor variable at measured at a set of discrete locations, $\{x_k : k = 1, \ldots m\}$, where each x_k lies within a geographical region of interest, A. We assume that the locations of the set of x_i are different the set of x_k implying that they are spatially misaligned. Also, let $U(x)$ and $U^*(x)$ denote spatially continuous Gaussian process defined over x_i and x_k, respectively, T_i is an independent and identically distributed Gaussian variable defined on a set of the x_i and let V_k is an independent and identically distributed Gaussian variable defined on a set of the x_k. Model-based approach for analysing such dataset proceeds by developing a joint model for the response and the predictor Madsen et al. (2008). Therefore, the joint model for Y_i and D_k takes the form

$$\begin{cases} Y_i = \alpha + \beta U(x_i) + T_i & \text{for } i = 1, \ldots, n \\ D_k = \gamma + U^*(x_k) + V_k & \text{for } k = 1, \ldots, m \end{cases}, \tag{3.1}$$

where the β parameters quantify the strength of the association between Y and D, whilst the α and γ are intercept parameters. $U(x)$ is defined as a spatially continuous Gaussian process, with stationary and isotropic covariance function such that

$$\text{Cov}\{U(x_k), U(x_{k'})\} = (\mathit{\Sigma}_m)_{(k,k')} = \tau^2 \rho(\|x_k - x_{k'}\|; \theta), \tag{3.2}$$

where τ^2 is the variance, $\|x_k - x_{k'}\|$ is the Euclidean distance between locations x_k and $x_{k'}$ and $\rho(\cdot; \theta)$ is the isotropic and stationary correlation function of $U(x)$ indexed by the parameter θ. T_i is Gaussian distributed with mean zero and variance ω^2 and V_k is Gaussian distributed with mean zero and variance ν^2. The likelihood function for the model define in Equation (3.1) is given as

$$\begin{aligned} L(\theta) &= [Y, D; \psi] \\ &= [Y \mid D; \psi][D; \psi], \end{aligned} \tag{3.3}$$

where ψ is the vector of the parameters, $[D; \psi]$ is multivariate Gaussian with mean $\gamma \mathbb{1}_{m\times 1}$ and covariance $\mathit{\Sigma}_m + \nu^2 \mathbb{I}_m$. Finally, $[Y \mid D; \psi]$ is a multivariate Gaussian with mean

$$\alpha \mathbb{1}_{n\times 1} + C^\top \mathit{\Sigma}_m^{-1}(D - \gamma \mathbb{1}_{m\times 1}), \tag{3.4}$$

and covariance

$$\mathit{\Sigma}_n - C^\top \mathit{\Sigma}_m^{-1} C, \tag{3.5}$$

where C is the cross-covariance between Y and D whose entries are given by

$$\text{Cov}\{Y_i, D_k\} = \beta^2 \tau^2 \rho(\|x_i - x_k\|; \theta),$$

where $\|x_i - x_k\|$ is the Euclidean distance between locations in set $\{x_i : i = 1, \ldots n\}$ and in set $\{x_k : k = 1, \ldots m\}$, (k, k') entry for $\varSigma_m$ is given in Equation 3.2 and $\varSigma_n = \beta^2 \varSigma_m + \omega^2 \mathbb{I}_n$.

3.2.2 Conditional autoregressive models for spatially misaligned areal data

The main referenced paper for areal misalignment are Mugglin et al. (2000) and Agarwal et al. (2002), they proposed a fully model-based approach implemented within a Bayesian framework. The advantage of working in Bayesian framework is that it allows estimation of model parameters and prediction jointly. Their approach is as follows: let Y_i denote a continuous response variable measured over a region $\mathcal{R}_i$, for $i = 1, \ldots n$, where Y_i's are considered as an aggregated measurement $\sum_k Y_{ij}$, where Y_{ij} is unobserved and the summation is over a regular grid indexed by k. Also, let D_k denote the predictor variable measured over a region $\mathcal{R}_k$, also considered as an aggregated measurement $\sum_l D_{kl}$, where D_{kl} is unobserved. The set of partitions Y_i and D_k are spatially misaligned. Random effect U_i was introduced to capture the spatial association among the Y_i's and random effect U_k^* was introduced to capture the spatial association among the D_k's. These random effects are given a CAR prior (Besag, 1974) specification with the assumption that the latent Y_{ij} inherit the effect, U_i and that the latent D_{kl} inherit the effect U_k^*. The joint distribution of Y and D can be expressed as

$$\prod_{i=1}^{n} [Y_{i1}, \ldots, Y_{ij_i} | Y_i] \prod_{i=1}^{n} [Y_i | U_i] \prod_{k=1}^{m} [D_{kl}, \ldots, D_{kl_k} | D_k] \prod_{k=1}^{m} [D_k | U_k^*]$$

The CAR prior on U_i assumes that

$$U_i | U_{-i} \sim N\left(\rho_c \sum_{j \sim i} c_{ij} U_j, \tau_i^2\right), \tag{3.6}$$

where $U_{-i} = \{U_j : j \neq i\}$, ρ_c is the spatial dependence parameter and c_{ij} are known quantities such that $c_{ij} \neq 0$ if and only if $j \sim i$ and $j \neq i$. It follows that the joint distribution of U is a multivariate zero-mean Gaussian distribution with covariance matrix

$$(I - \rho_c C)^{-1}\tilde{D}, \tag{3.7}$$

where $\tilde{D} = \{\tau_1^2, \ldots, \tau_n^2\}$, while the specification of C is generally tied to the specific arrangement of the partition of the region of interest. The most common approach is to set $c_{ij} = 1$ if $j \sim i$ and 0 otherwise.

In Section 3.3.2, we present how geostatistical framework can be used to solve this problem by assuming a spatially continuous process for U_i. An advantage of this is that it allows for spatially continuous prediction regardless of the format of the data.

3.3 Methods

3.3.1 Data

3.3.1.1 Index of Multiple Deprivation

IMD is a measure of relative deprivation and can thus be used to rank neighbourhoods. It combines seven distinct domains of deprivation: income; employment; education; skills and training; health deprivation and disabilit; crime, barriers to housing and services; and living environment. Weighted cumulative models are used to compute the IMD score, with weights obtained via the maximum likelihood method for factor analysis (Liu and Rubin, 1998; Smith et al., 2015). IMD data are made available either as a scores, deciles or ranks. In this study, we used the IMD score released in 2015, which was based on data collected between 2012 and 2013. Larger values of the IMD score can be interpreted as corresponding to a higher level of deprivation of an area relative to the others (UK Government, 2015).

3.3.1.2 Life expectancy at birth

Our outcome variable is the LEB released by the Office for National Statistics (2015) (ONS). The ONS estimates LEB using life tables that are constructed by applying the Chiang method (Chiang, 1984) to mortality data collected over five consecutive years, starting from 2009. This method assumes that the probability of dying is constant within a specified set of age intervals a_{t-1} and a_t. The resulting estimator is

$$LEB = \sum_{t=1}^{T} [(a_t - a_{t-1})p_t + m_t d_t]$$

where p_t is the fraction of the total population that has not died in the time interval (a_{t-1}, a_t), m_t is the average number of years lived in an interval by an individual who passes away in (a_{t-1}, a_t), d_t is the fraction of the total population that dies in (a_{t-1}, a_t) between ages a_{t-1} and a_t and T is the number of age intervals. In our case, we have $T = 19$, $(a_1, a_2) = (0, 1)$, $(a_2, a_3) = (1, 4)$ and for $t > 3$, $a_t - a_{t-1} = 5$.

Life tables are usually constructed separately for males and females because of their different mortality patterns (Gjonça et al., 1999). In the next section, we exploit the correlation between LEB for the two genders, and their associaton with IMD, in order to obtain more accurate estimates.

3.3.2 Modelling framework

Let LEB_{ij} denote the life expectancy at birth for males, if $i = 1$, and females, if $i = 2$, at the j-th MSOA, henceforth $MSOA_j$, for $j = 1, \ldots, n$. Similarly, we use IMD_k to denote the IMD score for the k-th LSOA, henceforth $LSOA_k$, for $k = 1, \ldots, m$.

Define $U(x)$ to be a spatially continuous Gaussian process, with stationary and isotropic exponential covariance function, i.e.

$$\text{Cov}\{U(x), U(x')\} = \tau^2 \exp\{-\|x - x'\|/\delta\},$$

where τ^2 is the variance and δ is a scale parameter regulating the rate of decay of the spatial correlation for increasing Euclidean distance $\|x - x'\|$ between any two locations x and x'.

We then model the cross-correlation in space between LEB and IMD through $U(x)$ as follows. Define the averaged spatial processes based on $U(x)$ over LSOAs and MSOAs as $U_j = \int_{MSOA_j} U(x)\, dx/|MSOA_j|$ and $U_k^* = \int_{LSOA_k} U(x)\, dx/|LSOA_k|$, where $|\mathcal{A}|$ corresponds to the area in m^2 of a spatial unit $\mathcal{A}$. The proposed joint model for LEB_{ij} and IMD_k takes the form

$$\begin{cases} LEB_{ij} = \alpha_i + \beta_i U_j + T_{ij} & \text{for } i = 1, 2; j = 1, \ldots, n \\ IMD_k = \gamma + U_k^* + V_k & \text{for } k = 1, \ldots, m \end{cases}, \tag{3.8}$$

where the β_i parameters quantify the strength of the association between LEB and IMD, whilst the α_i and γ are intercept parameters. Also in (3.8), the V_k are i.i.d. Gaussian variables with mean zero and variance ν^2, whilst (T_{1j}, T_{2j}) are i.i.d. bivariate Gaussian variables with mean zero and covariance matrix

$$\varOmega = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}.$$

It follows that the covariance between LEB_{ij} and IMD_k is

$$\mathrm{Cov}\{LEB_{ij}, IMD_k\} = \frac{\beta_i \tau^2}{|MSOA_j||LSOA_k|} f(MSOA_j, LSOA_k; \delta), \tag{3.9}$$

where

$$f(MSOA_j, LSOA_k; \delta) = \int_{MSOA_j} \int_{LSOA_k} \exp\left\{-\frac{\|x_j - x_k\|}{\delta}\right\} dx_j \, dx_k. \tag{3.10}$$

In order to understand how much of the spatial variation in LEB is explained by IMD, we compare the fitted model (3.8) with the special case of no association with IMD, i.e. $\beta_1 = \beta_2 = 0$.

An important feature of the spatial covariance structure defined by equation (3.9) is that it accounts for the different shapes and sizes of the various areal units involved.

3.3.3 Inference: parameter estimation and spatially continuous prediction

Let $LEB_i = (LEB_{i1}, \ldots, LEB_{in})$ and $IMD = (IMD_1, \ldots, IMD_m)$ and denote by θ the vector of model parameters. Also, let $\varSigma_{LSOA}$ and $\varSigma_{MSOA}$ be the spatial covariance matrices of the IMD at LSOA- and MSOA-level, respectively. The (k, k') entry for $\varSigma_{LSOA}$ is

$$(\varSigma_{LSOA})_{kk'} = \frac{\tau^2}{|LSOA_k||LSOA_{k'}|} f(LSOA_k, LSOA_{k'}; \delta) \tag{3.11}$$

where $f(LSOA_k, LSOA_{k'}; \delta)$ is as specified in equation (3.10). The elements of $\varSigma_{MSOA}$ are obtained similarly, replacing the domains of the integrals that define

(3.11) with those of the corresponding MSOAs. Using $[\cdot]$ as a shorthand notation for “the density function of the random variable $\cdot$,” the likelihood function for θ can now be expressed as

$$\begin{aligned} L(\theta) &= [LEB_1, LEB_2, IMD; \theta] \\ &= [LEB_1, LEB_2 \mid IMD; \theta][IMD; \theta], \end{aligned} \tag{3.12}$$

where $[IMD; \theta]$ is multivariate Gaussian with mean $\gamma \mathbb{1}_{m\times 1}$ and covariance $\Sigma_{LSOA} + \nu^2 \mathbb{I}_m$. Finally, $[LEB_1, LEB_2 \mid IMD; \theta]$ is a multivariate Gaussian with mean

$$\alpha \oplus \mathbb{1}_{n\times 1} + C^\top \Sigma_{LSOA}^{-1}(IMD - \gamma \mathbb{1}_{m\times 1}), \tag{3.13}$$

and covariance

$$\Sigma_{LEB} - C^\top \Sigma_{LSOA}^{-1} C, \tag{3.14}$$

where: $\alpha = (\alpha_1, \alpha_2)^\top$; $\oplus$ is the Kronecker product; $C = (C_1, C_2)^\top$ with C_i being the cross-covariance between LEB_i and IMD whose entries are given by Equation (3.9); finally,

$$\Sigma_{LEB} = \begin{pmatrix} \beta_1^2 \Sigma_{MSOA} + w_1^2 \mathbb{I}_n & \beta_1\beta_2 \Sigma_{MSOA} + w_{12} \mathbb{I}_n \\ \beta_1\beta_2 \Sigma_{MSOA} + w_{12} \mathbb{I}_n & \beta_2^2 \Sigma_{MSOA} + w_2^2 \mathbb{I}_n \end{pmatrix}.$$

We calculate each of the integrals in (3.9) and (3.11) using the numercial approximation described in Section 3 of Johnson et al. (2019). Finally, we estimate θ through maximization of the likelihood function in (3.12).

To quantify the contribution of IMD in explaining the spatial variation in LEB, we use the fraction of the total variance explained, given by

$$\frac{\text{Var}\{\beta_i U_j\}}{\text{Var}\{LEB_{ij}\}} = \frac{\beta_i^2 \tau^2}{\beta_i^2 \tau^2 + \omega_i^2}, \tag{3.15}$$

with $i = 1$ for the male population and $i = 2$ for the females, respectively.

We carry out spatial prediction over a regular grid at a spatial reslution of 250 by 250 meters, covering the whole of the Liverpool council area. Let $\{x_1, \ldots, x_q\}$ be the set of points forming the grid, with $q = 1787$, and let $LEB_i(x_h) = \alpha_i + \beta_i U(x_h)$ be the unobserved value of LEB at x_h, for $h = 1, \ldots, q$. Now, write $LEB^* = (LEB_1(x_1), \ldots, LEB_1(x_q), LEB_2(x_1), \ldots, LEB_2(x_q))^\top$; the predictive distribution for LEB^*, i.e. its conditional distribution given the data, is multivariate Gaussian with mean

$$\alpha \oplus \mathbb{1}_{q\times 1} + D^\top \Sigma_{LEB}^{-1}(LEB - \alpha \oplus \mathbb{1}_{n\times 1}), \tag{3.16}$$

and covariance matrix

$$\Sigma_{LEB^*} - D^\top \Sigma_{LEB}^{-1} D. \tag{3.17}$$

In (3.17), the (h, h')-th element of Σ_{LEB^*} is given by $(\Sigma_{LEB^*})_{hh'} = \tau^2 \exp\{-\|x_h - x_{h'}\|/\delta\}$. Also,

$$D = \begin{pmatrix} D_1 \\ \\ D_2 \end{pmatrix}$$

where D_i is the $n \times q$ matrix whose h-th column is $(d_1(x_h), \ldots, d_n(x_h))$, and $d_j(x_h) = \beta_i^2 \tau^2 \int_{MSOA_j} \exp\{-\|x_h - x\|/\delta\}\, dx$.

Using the above results, we can then draw samples for LEB^* and obtain any predictive summary of interest. For example, to identify areas in the Liverpool council district that are highly likely to fall below a threshold l, we map the non-exceedance probabilities (NEPs)

$$NEP_i(x) = Pr(LEB_i(x) < l \mid LEB_1, LEB_2, IMD). \tag{3.18}$$

In the results shown in the next section, we set l to be England-wide average years for males (l = 79.2 years) and females (l = 82.9 years). Values of NEP close to 1 indicate that LEB is highly likely to lie below l. Conversely, values close to 0 indicate locations whose LEB is highly likely to be above l. Finally, locations with values around 0.5 are equally likely to be below or above l, thus corresponding to the scenario with highest uncertainty.

Our results have been made publicly available at the following link

`http://fhm-chicas-apps.lancs.ac.uk/shiny/users/johnsono/LEBLiverpool/`

where interactive maps for NEPs can be generated from our model for any chosen threshold l.

We provide the derivation of all the equations in Appendix A.7 of the supplementary material.

3.3.4 Model validation: testing for residual spatial correlation

One of the main assumptions of the fitted bivariate model (3.8) is that all the spatial variation in LEB is captured by the IMD. To validate this assumption, we proceeds as follows. We first estimate the T_{ij} as

$$LEB_{ij} - \hat{\alpha}_i - \hat{\beta}_i \hat{U}_j \text{ for } i = 1, 2; j = 1, \ldots, n$$

where $\hat{\alpha}_i$ and $\hat{\beta}_i$ are the maximum likelihood estimates and $\hat{U}_j$ is the predictive mean of U_j. For each MSOA, we then extract the centroid associated with each of the $\hat{T}_{ij}$. For both males ($i = 1$) and females ($i = 2$), we then compute the empirical variogram given by

$$\hat{\gamma}_i(\mathcal{U}) = \frac{1}{2|\mathcal{U}|} \sum_{(j,k)\in\mathcal{U}} (\hat{T}_{ij} - \hat{T}_{i'j})^2, \tag{3.19}$$

where $\mathcal{U} = [u_0, u_1]$ is the set of all pairs of all pairs of centroids that no less than u_0 and no more than u_1 distant apart, and $|\mathcal{U}|$ is the number of pairs within the set. In the current analysis, we construct the empirical variogram by segmenting the interval $[0, 10]$ (km) into 12 equally spaced intervals.

In order to test whether the observed $\hat{\gamma}_i(\mathcal{U})$ is compatible with assumption of no residual spatial correlation, we use the following Monte Carlo approach to construct 95% tolerance intervals around $\hat{\gamma}_i(\mathcal{U})$:

1. permute the order of T_ij, while holding the centroid of the MSOAs fixed;
2. compute the empirical variogram $\hat{\gamma}_i(\mathcal{U})$ for the permuted T_{ij};
3. repeat step 1 and 2 for a large number of times, say B;
4. use the resulting B empirical variograms to generate 95% tolerance intervals at each of the predefined distance bins.

If $\hat{\gamma}_i(\mathcal{U})$ lies within the 95% tolerance intervals, we conclude that the assumption that the IMD fully captures the spatial variation in LEB is supported by the data. If, instead, $\hat{\gamma}_i(\mathcal{U})$ falls outside the 95% tolerance intervals, we conclude that the data show evidence against the fitted model in (3.8).

3.3.5 Assessment of the coverage probabilities for the regression parameters and the spatial predictions

In this section, we outline a simulation study which we carry out in order to assess the reliability of the confidence intervals generated for the regression coefficients β_i, the spatially continuous predictions and the MSOA-level predictions for LEB. This is especially important in our case as we carry out spatial predictions by plugging-in the maximum likelihood estimates, hence ignoring parameter uncertainty.

We then simulate $B = 10,000$ data sets under the bivariate the model in (3.8) using the administrative boundaries of Liverpool and proceed through the following iterative steps:

1. Simulate the spatially continuous process $U(x)$ over a 150 × 150 metres grid.
2. Simulate the spatially continuous surface for IMD and LEB on the same regular grid.
3. Average the LEB over the MSOAs boundaries and the IMD over the LSOAs boundaries.
4. Fit the model in (3.8) and compute confidence intervals of coverage α for β_1 and β_2.
5. Compute the prediction intervals of coverage α for the LEB at MSOA-level and over the 150 × 150 metres grid.

In this simulation we set the true value of the parameters to the point estimate reported for Model 1 in Table 3.1. We let the coverage probability α vary over the set $\{5i/100 : i = 1, 2, \ldots, 19\}$. Using the resulting 10,000 confidence intervals in step 4 and prediction intervals in step 5, we compute the fraction of times that the true values fall within those intervals in order to obtain the actual coverage.

3.4 Results

Table 3.1 shows the point and interval estimates for the model with (Model 1) and without (Model 2) IMD. The likelihood-ratio test for the null hypothesis $\beta_1 = \beta_2 = 0$ yields a p-value smaller than 0.001, hence indicating that Model 1 is a better fit to data. We find that the fraction of total variance explained (see equation 3.15)) is about 38.92% for females and 63.52% for males, respectively. We estimate that the range of the spatial correlation, defined as the distance beyond which the correlation is below 0.05, is approximately 4.6 km. The correlation in LEB between males and females, given by ratio $\omega_{12}/(\omega_1\omega_2)$, is 0.59 with associated 95% confidence interval (0.31, 0.90).

Figure 3.1 shows the boundaries of the electoral wards (EWs) in Liverpool district and their names. In commenting the results, we shall refer to the different areas of the Liverpool district council based on the EWs in Figure 3.1.

Figure 3.2 (upper and middle panel) shows the estimated surface of LEB at MSOA-level for females and males. As expected, female LEB is consistently higher than that for males, as also reflected in the spatially continuous predictions of Figure 3.3. In contrasting the maps of Figure 3.2 with those of Figure 3.3, we notice that spatially continuous predictions provide useful insights into the variation in LEB within MSOAs that is otherwise hidden by the aggregated estimates at MSOA-level. To demonstrate this, we selected the MSOA with the lowest and largest estimated value in LEB for both males and females; these MSOAs are identified identified by the white (largest LEB) and green (lowest LEB) boundaries in upper and middle panels of Figure 3.2. More specifically, for males, the lowest estimated value in LEB at MSOA-level is about 70.2 years and the largest is 85.2 years, whilst for females these are respectively 73.5 years and 89.6 years. In the maps of Figure 3.4, we then draw the contour lines for these same values in LEB. These reveal the actual extent of the areas where LEB reaches its highest and lowest values, that cannot be possibly discerned from Figure 3.2: the white contour lines encompass a relatively small at the intersection of Childwall, Woolton and Church; the green contour lines, instead, delineate a wide area consisting of three disjoint sub-regions in the north-west and north-east of Liverpool.

Figure 3.4 shows the non-exceedance probability maps of female and male LEB, with thresholds of 82.9 years and 79.2 years, respectively. These two values also correspond to the national average LEB in England for the two genders. For females, we find that LEB is at least 80% likely to be below 82.9 years in the areas of Kirkdale, Kensington and Fairfield and Princes Park; for males, a wider area is instead identified, comprising those same EWs with the addition of Fazakerley, Norris Green, Clubmoor, County, Anfield, Everton, Tuebrook and Stoneycroft, Picton, Central, St Michaels and Speke-Garston. On the other hand, areas that are at least 80% to be above the England-wide averages are are found in the EWs of Childwall, Woolton and Church for both males and females. In the EWs of West Derby and Mossley Hill the model is most uncertain as these are equally likely to have a LEB above or below the chosen thresholds for the both males and females.

Figure 3.5 the results for the variogram-based validation procedure. Since the observed variograms for both males and females lie within the 95% band, we interpret this as evidence that the data do not show any adittional residual spatial correlation. This leads us to conclude that the IMD was able to explain most of the spatial variation in LEB.

Figure 3.6 shows the scatter plots of the actual coverage, obtained from the simulation study, against the nominal coverage. For the spatial predictions, the actual coverage is averages over all the MSOAs and over the regular grid, respectively. The plots show a strong concordance between actual and nominal coverage levels. We then conclude that the interval estimates for the regression coefficients and the spatial predictions generated by the fitted model are in fact reliable when using plug-in estimates.

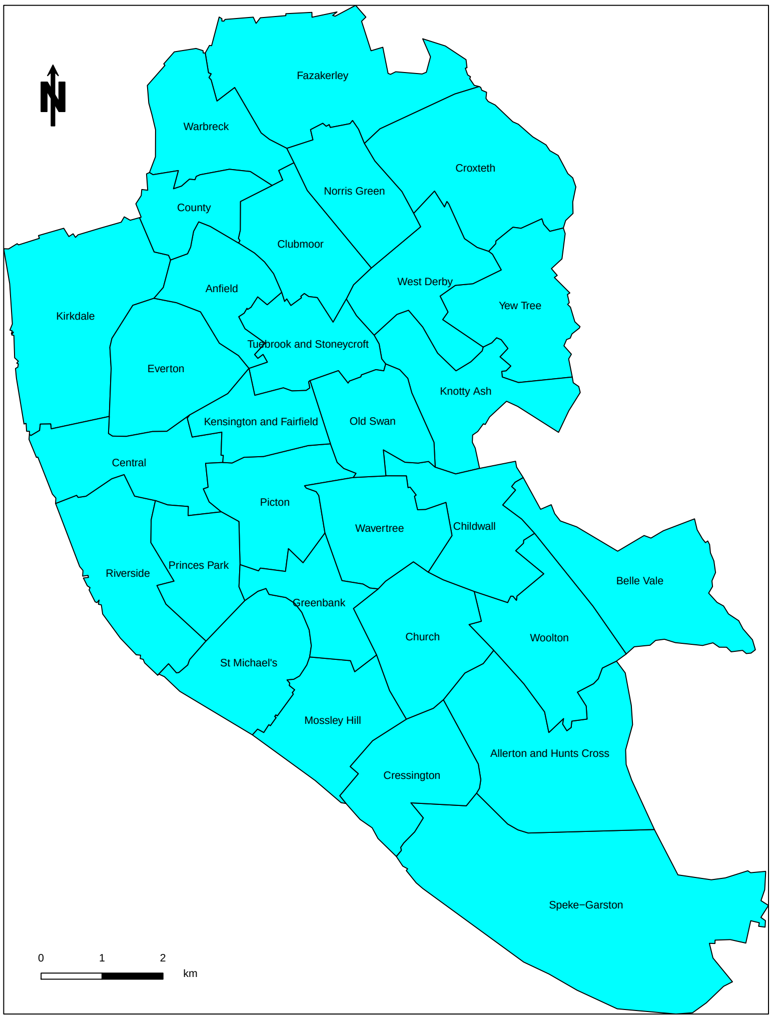


Figure 3.1: Map of Liverpool district council, UK showing the 30 electoral wards

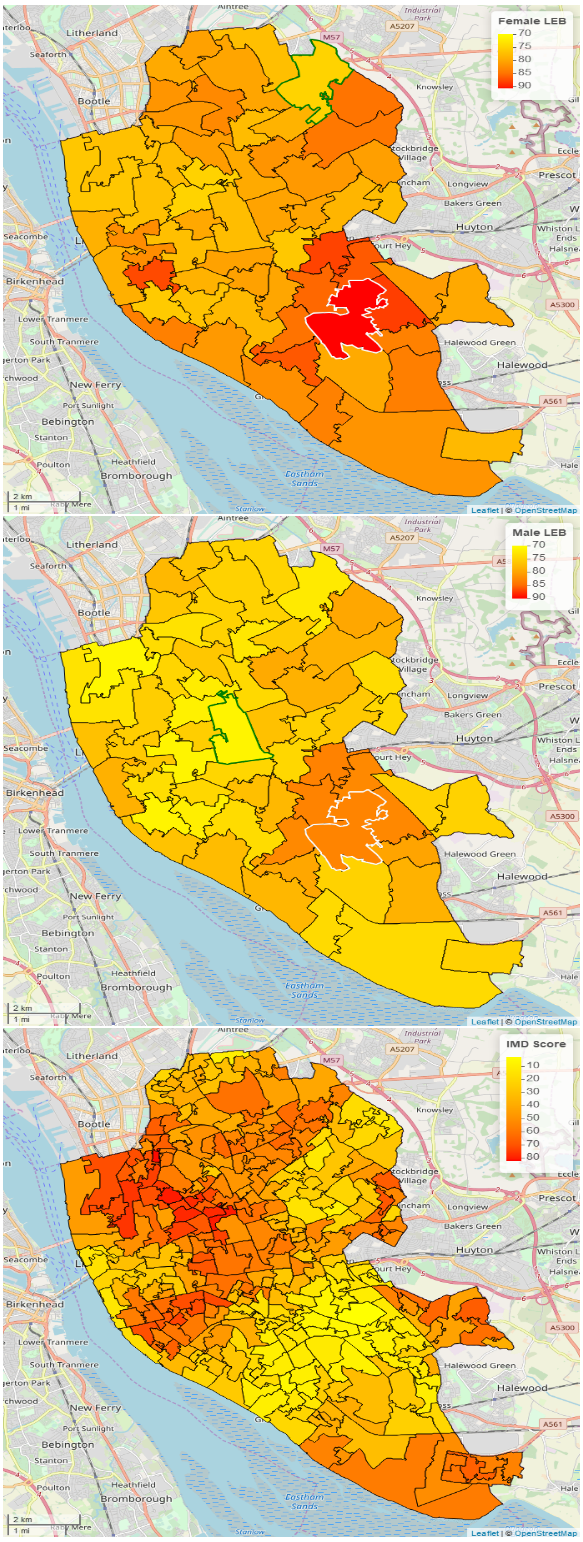


Figure 3.2: Maps of the estimated female (upper panel) and male (middle panel) life expectancy at birth (LEB) and index of multiple deprivation (IMD) (lower panel). Middle Super Output Area (MSOA) with boundaries coloured in green correspond to the lowest estimated LEB, whilst those in white to the highest. For males, the lowest estimated LEB is 70.2 years and the highest is 85.2 years; for females, the lowest is 73.5 years and the highest is 89.6 years.

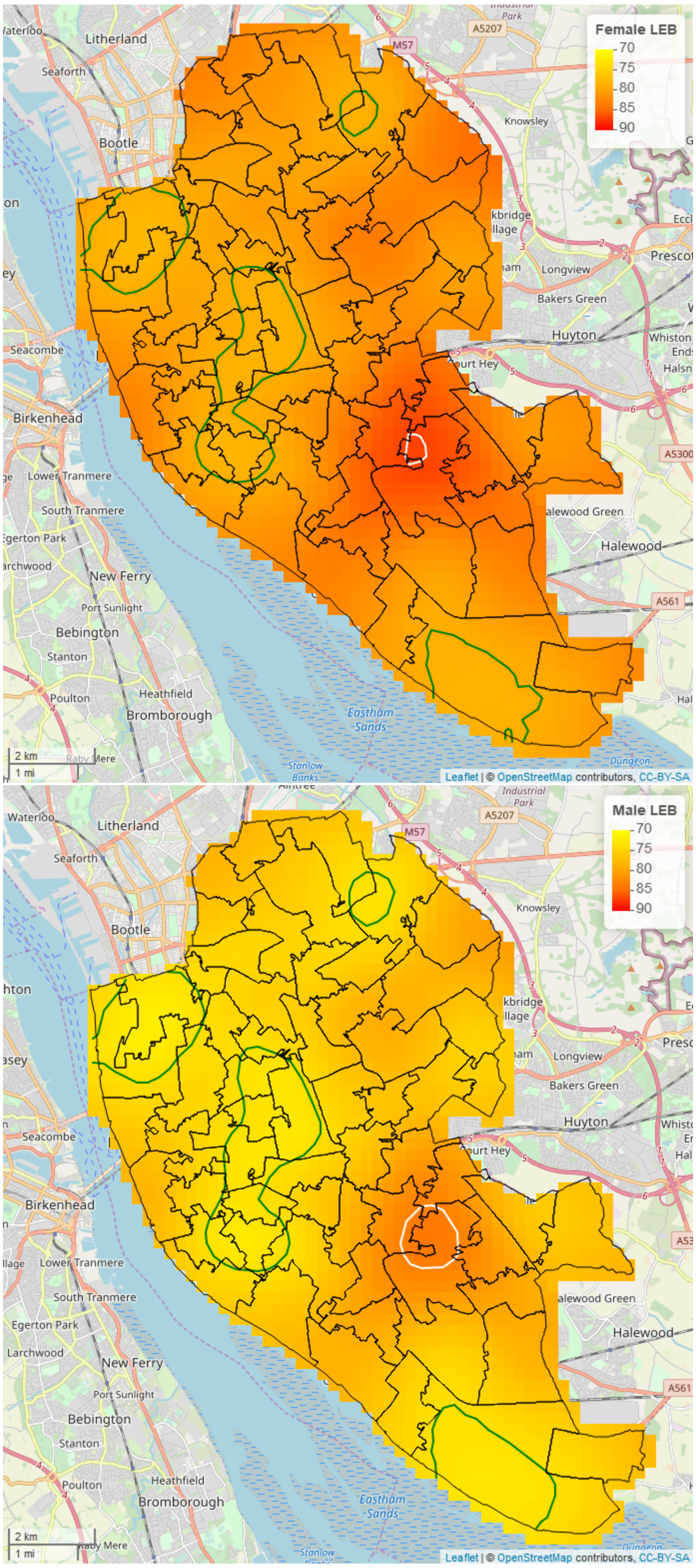


Figure 3.3: Spatially continuous prediction maps of female (upper panel) and male (lower panel) life expectancy at birth (LEB) in Liverpool, UK. In the upper panel, the white contour lines are for a LEB of 89.6 years and the green contour lines for a LEB of 73.5 yers; in the lower panel, the white contour lines correspond to 70.2 years and the green contour lines to 85.2 years.

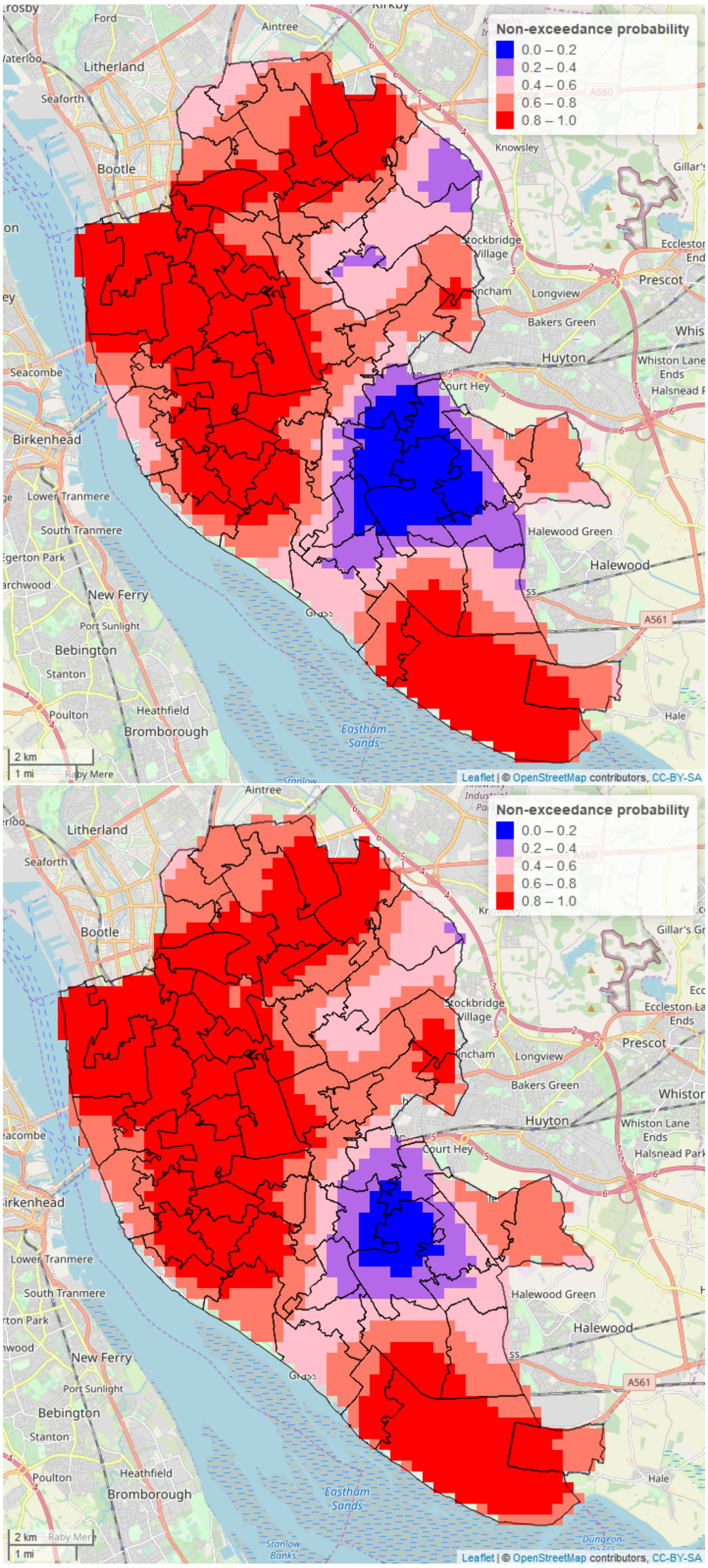
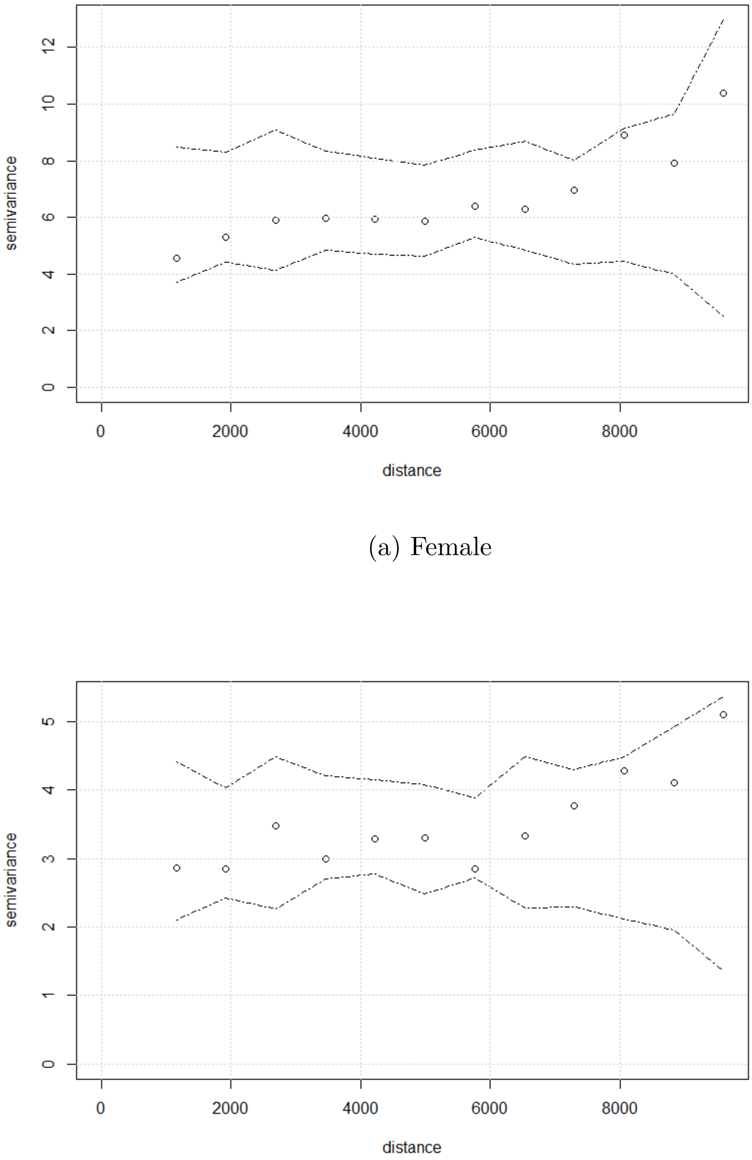


Figure 3.4: Maps of the non-exceedance probability of female (upper panel) and male (lower panel) life expectancy at birth (LEB), with threshold 82.9 and 79.2 (average LEB in England, UK), respectively in Liverpool, UK.



(a) Female

(b) Male

Figure 3.5: Plots of the observed variograms (points) and the 95% tolerance bandwidth (dashed lines) generated under the assumption of absence of residual spatial correlation.

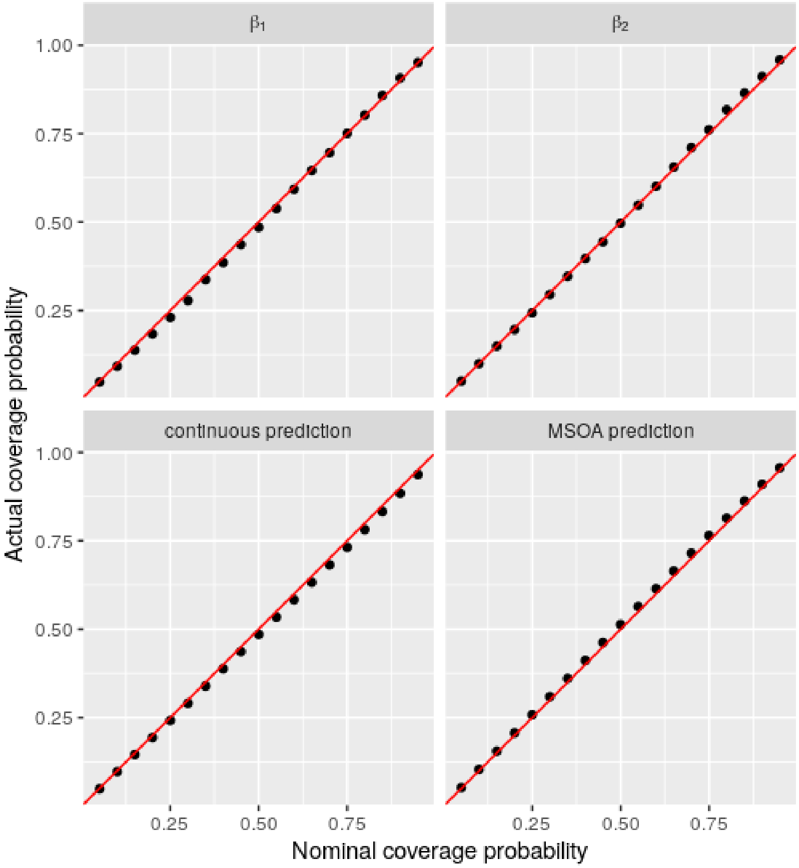


Figure 3.6: Scatter plots of the actual against the nominal coverage for the confidence intervals generated for β_1 and β_2 (upper panels), and for the spatially continuous and MSOA-level predictions of LEB (lower panels). The red lines in each panel correspond to the identity line.

Table 3.1: Point estimates and 95% confidence intervals (CI) for the three model parameters.

	Model 1		Model2	
Parameter	Estimate	CI 95%	Estimate	CI 95%
α_1	75.466	(75.596, 76.135)	75.131	(74.990, 75.272)
α_2	81.120	(80.883, 81.357)	81.375	(80.927, 81.823)
β_1	-0.154	(-0.180, -0.128)	-	-
β_2	-0.129	(-0.167, -0.091)	-	-
$\log \omega_1^2$	1.810	(1.494, 2.126)	3.036	(2.955, 3.117)
$\log \omega_2^2$	2.581	(2.272, 2.890)	3.160	(3.033, 3.287)
$\log \omega_{12}$	1.671	(1.257, 2.086)	2.871	(2.768, 2.974)
γ	39.221	(28.242, 50.200)	39.190	(28.073, 50.306)
$\log \tau^2$	6.226	(3.611, 8.841)	6.232	(5.678, 6.586)
$\log \delta$	7.336	(6.845, 7.827)	7.349	(6.318, 7.846)
$\log \nu^2$	2.586	(2.244, 2.927)	2.589	(2.064, 2.932)
Log-likelihood	-1429.491		-1465.432	

3.5 Discussion

We have developed a model-based geostatistical approach that allows to model the relationship between life expectancy and the index of multiple deprivation when these are provided over misaligned partitions of the study area. Unlike existing methods of analysis (e.g. Buck et al. (2017)), one of the main advantages of our approach is that it allows to combine information from multiple data sources without coarsening their resolution to a common spatial scale. The underpinning principle of our modelling framework is that spatially aggregated data should be treated as the realization of an aggregated spatially continuous stochastic process. This approach is strongly linked to that of Diggle et al. (2013) who propose the use of an integrated log-Gaussian Cox process to model disease counts at areal-level. As result of this, the proposed modelling paradigm allows to carry out spatially continuous inference which would be otherwise infeasible if the spatial models were tied to the specific data-format at which LEB and IMD are provided. Conditionally autoregressive models (Besag, 1974) are one of the most commonly used approaches to analyse areal-level data that suffer from this limitation (Agarwal et al., 2002; Mugglin et al., 2000).

Our novel methodology has highlighted the importance of dealing with variation in LEB occurring within areal units. In our application, the use of spatially continuous predictions was especially useful in order to visualize patterns in LEB that were hidden by the aggregated estimates. Furthermore, the use of non-exceedance probabilities also provides a way of measuring uncertainty in relation to a predefined threshold in LEB in order to identify areas that need urgent intervention.

One of the limitations of the model defined by equation (3.8), is that all the spatial variation in LEB and IMD is modelled through a single spatial process $U(x)$. The model could then be made more flexible through the introduction of a second spatial process, say $W(x)$, into the first line of equation (3.8), i.e.

$$LEB_{ij} = \alpha_i + \beta_i U_j + W_j + T_{ij}, \text{ for } i = 1, 2; j = 1, \ldots, n$$

where $W_j = |MSOA_j|^{-1} \int_{MSOA_j} W(x)\, dx$. In this model, the W_j would allow to account for unexplained spatial variation in LEB that is unrelated to IMD. However, in our attempt to fit such a model, we incurred in identifiability issues as the estimated spatial scale for the process $W(x)$ was well below the extent of the smallest MSOA. This also suggests that most of the large scale spatial variation in LEB is in fact well captured by the IMD and that unexplained variation occurring on a smaller spatial scale is instead accounted for by the unstructured component of the model T_{ij}.

Although our application to mapping LEB in Liverpool only dealt with areal misalignment, our methodology is more widely applicable to almost any scenarios of spatial misalignment. Consider, for example, the case where a second spatially varying factor associated with LEB is available in raster format over a regular grid, say $\{\tilde{x}_1, \ldots, \tilde{x}_q\}$, covering the whole of the Liverpool council area. Let $V(\tilde{x}_k)$ denote the value of such a variable at the grid location $\tilde{x}_k$, for $k = 1, \ldots, q$. Model (3.8) could then be extended by replacing the first line with

$$LEB_{ij} = \alpha_i + \beta_i U_j + \delta_i V_j + T_{ij},$$

where $V_j = |MSOA_j|^{-1} \int_{MSOA_j} V(x)\, dx$. Assuming a high enough spatial resolution of the raster file for $V(x)$, this integral could then be approximated by taking a sample average over the grid locations falling within $MSOA_j$. If, instead, the grid is too coarse, spatial variation in $V(x)$ within pixels can be accounted for by building a geostatistical model in a similar fashion as for the IMD in the second line of equation (3.8).

3.6 Conclusion

We have developed a novel joint geostatsitical approach to model the relationship between life expectancy at birth and the index of multiple deprivation while dealing with the issue of spatial misalignment. Unlike existing spatial methods based on conditional autoregressive models, one of the main strengths of the proposed modelling framework is the ability to carry out spatially continuous predictions regardless of the format of the data. Furthermore, it is also more widely applicable to more complex data scenarios where information is provided at a range of spatial scales, from pixel-level to areal-level.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

OJ, PD and EG conceived the idea. OJ and EG conducted the statistical analysis and developed the code. OJ wrote the first draft of the manuscript. OJ and EG reviewed the draft of the manuscript. All authors read and approved the final manuscript.

Abbreviations

LEB: Life Expectancy at Birth; IMD: Index of Multiple Deprivation; LSOA: Lower Super Output Area; MSOA: Middle Super Output Area; UK: United Kingdom.

Funding

OOJ holds a Connected Health Cities funded PhD studentship.

Bibliography

Agarwal, D. K., Gelfand, A. E., and Silander, J. A. (2002). Investigating tropical deforestation using two-stage spatially misaligned regression models. *Journal of agricultural, biological, and environmental statistics*, 7(3):420–439.

Allik, M., Brown, D., Dundas, R., and Leyland, A. H. (2016). Developing a new small-area measure of deprivation using 2001 and 2011 census data from scotland. *Health & place*, 39:122–130.

Auger, N., Alix, C., Zang, G., and Daniel, M. (2010). Sex, age, deprivation and patterns in life expectancy in quebec, canada: a population-based study. *BMC Public Health*, 10(1):161.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Bennett, J. E., Pearson-Stuttard, J., Kontis, V., Capewell, S., Wolfe, I., and Ezzati, M. (2018). Contributions of diseases and injuries to widening life expectancy inequalities in england from 2001 to 2016: a population-based analysis of vital registration data. *The Lancet Public Health*, 3(12):e586–e597.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 192–236.

Buck, D., Maguire, D., et al. (2017). Inequalities in life expectancy: changes over time and implications for policy. *Health*.

Chetty, R., Stepner, M., Abraham, S., Lin, S., Scuderi, B., Turner, N., Bergeron, A., and Cutler, D. (2016). The association between income and life expectancy in the united states, 2001-2014. *The Journal of the American Medical Association*, 315(16):1750–1766.

Chiang, C. L. (1984). *The life table and its applications*. Malabar Fla Robert E. Krieger Publishing 1984.

Christensen, K. and Vaupel, J. W. (1996). Determinants of longevity: genetic, environmental and medical factors. *Journal of internal medicine*, 240(6):333–341.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-gaussian cox processes: extending the geostatistical paradigm. *Statistical Science*, pages 542–563.

Gjonça, A., Tomassini, C., Vaupel, J. W., et al. (1999). *Male-female differences in mortality in the developed world*. Citeseer.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Johnson, O., Diggle, P., and Giorgi, E. (2019). A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. *Statistics in Medicine*, 38(24):4871–4887.

Kontis, V., Bennett, J. E., Mathers, C. D., Li, G., Foreman, K., and Ezzati, M. (2017). Future life expectancy in 35 industrialised countries: projections with a bayesian model ensemble. *The Lancet*, 389(10076):1323–1335.

Krieger, N., Chen, J. T., Waterman, P. D., Soobader, M.-J., Subramanian, S., and Carson, R. (2003). Choosing area based socioeconomic measures to monitor social inequalities in low birth weight and childhood lead poisoning: The public health disparities geocoding project (us). *Journal of Epidemiology & Community Health*, 57(3):186–199.

Liu, C. and Rubin, D. B. (1998). Maximum likelihood estimation of factor analysis using the ecme algorithm with complete and incomplete data. *Statistica Sinica*, pages 729–747.

Liverpool City Council (2015). The index of multiple deprivation 2015: A liverpool analysis. *Liverpool, United Kingdom: Liverpool City Council.*

Madsen, L., Ruppert, D., and Altman, N. (2008). Regression with spatially misaligned data. *Environmetrics*, 19(5):453–467.

Mugglin, A. S., Carlin, B. P., and Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, 95(451):877–887.

OECD (2017). Health at a glance 2017. *OECD Indicators, OECD Publishing, Paris. DOI: https://doi. org/10.1787/health_glance-2017-en. Accessed January, 2019.*

Oeppen, J. and Vaupel, J. W. (2002). Broken limits to life expectancy. *Science*, 296(5570):1029.

Office for National Statistics (2015). Health expectancies at birth for middle layer super output areas (msoas), england: 2009 to 2013. https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/healthandlifeexpectancies/articles/healthexpectanciesatbirthformiddlelayersuperoutputareasmsoasengland/2015-09-25. [Online; accessed 9-January-2019].

Office for National Statistics (2018). Middle super output area population estimates (supporting information). https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/middlesuperoutputareamidyearpopulationestimates. [Online; accessed 2-January-2019].

Public Health England (2018). Liverpool unitary authority health profile. http://fingertipsreports.phe.org.uk/health-profiles/2017/e08000012.pdf. [Online; accessed 30-April-2018].

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., and Plunkett, E. (2015). The english indices of deprivation 2015. *London: Department for Communities and Local Government*.

Thomson, M., Connor, S., D'Alessandro, U., Rowlingson, B., Diggle, P., Cresswell, M., and BM, G. (1999). Predicting malaria infection in gambian children from satellite data and bed net use surveys: The importance of spatial correlation in the interpretation of results. *The American journal of tropical medicine and hygiene*, 61:2–8.

Tobias, M. I. and Cheung, J. (2003). Monitoring health inequalities: life expectancy and small area deprivation in new zealand. *Population Health Metrics*, 1(1):2.

Toson, B. and Baker, A. (2003). Life expectancy at birth: methodological options for small populations. *National statistics methodological series*, 33.

Tsimbos, C., Kalogirou, S., and Verropoulou, G. (2014). Estimating spatial differentials in life expectancy in greece at local authority level. *Population, Space and Place*, 20(7):646–663.

UK Government (2015). English indices of deprivation 2015. `https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015`. [Online; accessed 9-January-2019].

Wall, M. M. (2004). A close look at the spatial structure implied by the car and sar models. *Journal of statistical planning and inference*, 121(2):311–324.

Woods, L. M., Rachet, B., Riga, M., Stone, N., Shah, A., and Coleman, M. P. (2005). Geographical variation in life expectancy at birth in england and wales is largely explained by deprivation. *Journal of Epidemiology & Community Health*, 59(2):115–120.

Chapter 4

Spatio-temporal modelling of incidence in COPD emergency admissions in an area of Northwest England from 2012 to 2018.

Olatunji Johnson[1], Peter Diggle[1], Michael Pearson[3], Tim Gatheral[2], Jo Knight[1] and Emanuele Giorgi[1]

[1] CHICAS, Lancaster Medical School, Lancaster University, Lancaster, UK

[2] Respiratory Medicine, Royal Lancaster Infirmary, Lancaster, UK

[3] Institute of Translational Medicine, Liverpool University, Liverpool, UK

Summary

Background

Chronic Obstructive Pulmonary Disease (COPD) is one of the leading causes of mortality worldwide with an estimated 3 million deaths in 2015, corresponding to 5% of all deaths globally. Acute exacerbations are a major contributor to the number of emergency admission in the UK. COPD is the second most common cause (after a heart attack) of admission to a medical ward in the UK - i.e., it's a huge cost burden and there is a belief that many cases could be prevented, hence the interest in predictions. In this study, we pursue two objectives: 1) to assess the relative contribution of socio-economic and environmental variables for forecasting COPD emergency admissions; 2) to develop a reliable surveillance system that triggers an alarm whenever COPD emergency admissions signal the likely exceedance of predefined incidence thresholds.

Methods

We developed a predictive model using a class of generalised linear mixed model. We select the best predictors using the root mean square error (RMSE). We developed an early warning system based on exceedance probabilities.

Results

The resulting predictors from our model selection are; minimum temperature; PM10; income deprivation; the proportion of males; and the proportion of the population aged above 75 years. We found that, overall, the selected predictor variables explain about 22% of the variability in the residual random effects. Among these variables, income deprivation attained the largest relative variance reduction of about 14%.

Conclusion

Our results demonstrate how to develop a predictive model as well as an early warning system for COPD emergency admission. Our model has the potential to predict correctly in most areas with high sensitivity and specificity. The early warning system can help to: identify and notify areas of a high incidence of COPD emergency admission; and inform resource allocation for the healthcare system.

Keywords: COPD; emergency admission; early warning system; variogram; geostatistics; generalised linear mixed model; predictive model

4.1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is one of the leading causes of mortality worldwide (Hasegawa et al., 2014; Mathers and Loncar, 2006) with an estimated 3 million deaths in 2015, corresponding to 5% of all deaths globally (World Health Organisation, 2016). Acute exacerbations are a major contributor to the number of emergency admission and hospitalization (Tian et al., 2012), especially during the winter months as a result of the increase in respiratory viral infections. The pathogenesis of COPD is still little understood while current research has been focused on understanding the risk factors associated with its exacerbation (Bahadori and FitzGerald, 2007; Chan et al., 2011; Osman et al., 2017).

While the majority of exacerbations are caused by infectious agents, especially rhinoviruses Wedzicha (2004), there has been evidence from previous studies that biological, environmental and socio-economic factors can also trigger COPD emergency hospitalisation (Hemming et al., 2009). Hemming et al. (2009) developed a Bayesian network approach in order to identify factors that can help predict COPD admissions in the UK and found a combination of environmental, socio-economic and health-related variables to be useful predictors. These included weather type (classified as sunny, cloudy, rainy, windy and snowy) temperature, outdoor air pollution, gas emissions, urbanisation, smoking, population age, environmental tobacco smoke, indoor air pollution (housing condition), income and education, infection load, number of previous admission and severity of the disease. However, most studies have examined these factors separately and only a few have assessed their joint contribution to COPD risk.

Predictive models have been developed in several studies to identify patients at high risk of COPD exacerbations (Billings et al., 2006; Samp et al., 2018; Urwyler et al., 2019; Yii et al., 2019) which add significant cost to the patients care. Hence, being able to accurately predict their occurrence can be especially useful in order to reduce avoidable COPD emergency admissions by targeting patients in most need. In order to develop a robust and scalable predictive models for COPD emergency admissions, the availability of comprehensive health records of patients is essential so as to ensure its reliability. Predictive power can also be further enhanced by incorporating risk factors concerning the lifestyle behaviour (e.g. smoking status), income, exposure to pollutants and other individual traits. However, such detailed information may not be readily available to researchers due to confidentiality issues or because it has not been collected. Notwithstanding, statistical modelling provides solutions that can be used to alleviate this issue. For example, generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993) are an extension of the classical generalized linear modelling framework that allows to account for the unavailability of risk factors through the use of so-called random effects. However, the full potential offered by this modelling framework has not been fully exploited in the analysis of COPD data and, in this paper, we aim to fill this gap.

While some analyses on COPD emergency admissions have focused on individual analysis where biological markers (e.g. forced expiratory volume in 1 seconds and blood level) were used to predict the risk of an emergency admission, here we focus our attention on studies that were concerned with understanding the geographical variation of COPD risk at population-level. Niyonsenga et al. (2018) model the prevalence of COPD and asthma over census units in the western area of Adelaide, South Australia, and assess the spatial clustering of cases using the local Getis-Ord's

Gi indices (Anselin, 1995). Kauhl et al. (2018) analyse how the prevalence of COPD varies across northeastern Germany and identify risk factors including proportions of insurants aged above 65, proportions of insurants with migration background, household size and area deprivation as statistically significant predictors for COPD. Holt et al. (2011) were the first to characterise geographic variations in COPD hospitalization across Health Service Areas (HSAs) and at state level across the United States. They found distinct geographical pattern in COPD hospitalisation rate in the HSA and state level, suggesting that different risk factors could be operating at different spatial scales. In another study conducted in Taiwan, Chan et al. (2014) analyse the spatio-temporal distribution of COPD mortality over a 9 year period, from 1999 to 2007. They found that smoking rate, the percentage of aborigines within a district, PM10, altitude and density of healthcare facilities were significantly associated with COPD mortality.

Most spatio-temporal analysis on COPD have used conditional autoregressive models (CAR) (Besag et al., 1991) to carry out spatial smoothing of COPD risk but did not attempt any forecasting. CAR models are formulated by defining a correlation structure between neighbouring areal units (e.g. districts or regions). In addition, all of these studies (Chan et al., 2014; Holt et al., 2011; Kauhl et al., 2018) have focused their efforts in predicting mean level of risks. In this paper, we argue that statistical modelling should, instead, aim to predict the exceedance of clinically relevant thresholds beyond which COPD risk is of public health concern.

In our analysis of COPD admissions, we pursue two specific objectives: 1) to assess the relative contribution of socio-economic and environmental variables for forecasting COPD emergency admissions; 2) to develop a reliable surveillance system that triggers an alarm whenever COPD emergency admissions signal the likely exceedance of predefined incidence thresholds. To the best of our knowledge, this is the first study that attempts to achieve these objectives using state-of-the-art spatio-temporal statistical methods for the analysis of data on COPD emergency admissions.

4.2 Methods

4.2.1 COPD admission data

Using the International Classification of Diseases (ICD) code (10th revision)49 , J44 for COPD, we extracted monthly counts of COPD emergency admissions for patients above 19 years living in the LA postcode area, covering parts of South Cumbria and North Lancashire in England (see Figure 4.1). The total population of the study region was 272,520 based on the 2011 census. The data cover the period from 1 April 2012 to 30 March 2018. To protect confidentiality and anonymity of the patients, spatial information on their place of residence was provided at the Lower Super Output Area (LSOA). From the same database, we also obtain the proportion of people older than 75 years and the proportion of male patients admitted, for each at LSOA-level.

4.2.1.1 Environmental variables

We obtained monthly weather data for 2012-2018 including monthly relative humidity, number of days of ground frost and temperature from the UK Met Office, freely available from the Centre for Environmental Data Analysis (`http://data.ceda.ac.uk/`). The spatial resolutions of the weather raster files is of $1 \times 1\text{km}^2$ across the

UK. We also obtained yearly pollution data including Particulate Matter less than 10 m in diameter (PM10), Sulphur Dioxide (SO2) and Nitrogen Dioxide (NO2), available from the Department of Environmental Food and Rural Affairs (DEFRA) (`https://uk-air.defra.gov.uk/data/pcm-data`). The estimate of the pollutants are provided at $1 \times 1\text{km}^2$resolution over the entire Great Britain. For our analysis, we computed the population weighted average of all the available raster data over the LSOAs shown in Figure 4.1.

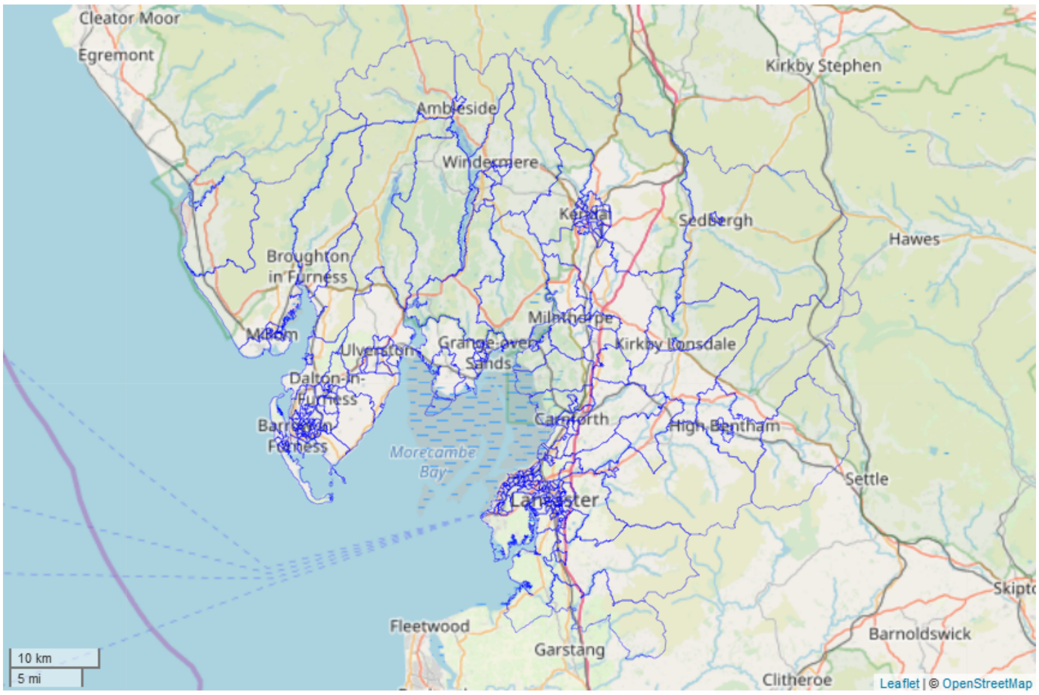


Figure 4.1: Map of South Cumbria and North Lancashire containing 209 LSOAs.

4.2.1.2 Socio-economic variables

We obtained the index of multiple deprivation (IMD) created by the Department for Communities and Local Government in order to account for socio-economic heterogeneities across LSOAs. The IMD combines seven domains which relate to income deprivation, employment deprivation, health deprivation and disability, education skills and training deprivation, barriers to housing and services, living environment deprivation, and crime. The IMD is available as either a score, decile or rank. In this study, we used the IMD score for 2015, the most recent release. Larger values of the score corresponds to a higher level of the domain deprivation 21.

4.2.1.3 Population data

We obtained the yearly population data per LSOA from the Office of National Statistics (ONS), UK. ONS usually updates their population estimates yearly based on migration data and any other physical adjustments (Office for National Statistics, 2018). The average population of LSOAs in England and Wales according to the census data in 2011 was 1,614 with 95% of LSOAs having a population of between 1,157 and 2,354.

4.2.2 Statistical modelling and assessment of residual spatio-temporal correlation

Let Y_{it} denote the monthly COPD emergency admission count at LSOA i and month t. We then assume that the Y_{it}, conditionally on a random effect Z_{it}, follow a Poisson distribution with mean $m_{it}\lambda_{it}$, where m_{it} denotes the population at LSOA i and month t and it represents the monthly incidence of COPD emergency admission at given LSOA.

We define the log-linear model for the incidence it as

$$\lambda_{it} = \exp\{d_{it}^{\top}\beta + Z_{it}\}, \tag{4.1}$$

where d_{it} is a vector of covariates with associated regression coefficients β. Finally, we assume that the Z_{it} are independent and identically distributed Gaussian variables with mean zero and variance σ^2. In order to build our regression model, we select predictors within three domains that are known to affect COPD admissions: weather, pollution and deprivation. The variables that we consider within each of these domains are listed in Table 4.1. As the variables within each group are highly collinear, our goal is to select the best predictor from each group. In addition to the variables of Table 4.1, we include proportion of males and proportion of the population aged above 75 years as background predictors at LSOA-level of the incidence of COPD emergency admission.

Table 4.1: The table showing the set of predictors available for this study.

Predictors	**Variables**
Weather	Minimum temperature; relative humidity; and number of days of ground frost.
Pollution	PM_{10} SO2; and NO2. All in micrograms per cubic metre (μgm^{-3})
Deprivation	Income deprivation; employment deprivation; health deprivation and disability; education skills and training deprivation; barriers to housing and services; living environment deprivation; and crime deprivation.

In order to carry out the selection of the best predictors, we split the dataset into training and test sets, with the former covering the months from April 2012 to March 2017 and the latter from April 2017 to March 2018. The rationale for the chosen test and training sets is that we aim to develop an early warning system that can better capture temporal features of the latest reported admissions. We then fit 63 models obtained by combining one predictor from each domain of Table 4.1 and, for each of those, we compute the root-mean-square-error (RMSE) for the predicted

COPD admissions incidence using the test set.

From the mixed model with the best set of predictors identified through the procedure outlined above, we assess whether the random effects Z_{it} show evidence of residual spatio-temporal correlation. To this end, we compute the empirical spatio-temporal variogram (ESTV) for the estimates of Z_{it}, using the centroid of each LSOA in order to quantify the proximity between LSOAs. Let $\hat{Z}(x_i, t_i)$ denote the estimate of Z_{it} from model (1) associated with the centroid x_i at time t_i. The expression of the ESTV is

$$\hat{\gamma}(u, v) = \frac{1}{2|n(u, v)|} \sum_{(i,j) \in n(u,v)} \{\hat{Z}(x_i, t_i) - \hat{Z}(x_j, t_j)\}^2,$$

where $|n(u, v)|$ is the number of pairs set.

We used Monte Carlo methods to construct a 95% tolerance interval around (u, v) in order to test the presence of residual spatio-temporal variation. We then proceed through the following iterative steps:

1. permute the order of $\hat{Z}(x_i, t_i)$, while holding (x_i, t_i) fixed;
2. compute the empirical variogram for $\hat{Z}(x_i, t_i)$;
3. repeat step 1 and 2 for a large number of times, say B times; and
4. use the resulting B empirical variogram to generate 95% tolerance interval at each of the predefined distance bins.

If $\hat{\gamma}(u, v)$ lies outside these intervals, then we conclude that the $Z(x_i, t_i)$ shows an evidence of residual spatio-temporal variation. To quantify the relative contribution of each predictor in the model, we compute the relative variance reduction (RVR) defined as

$$RVR = \frac{\sigma^2_{-j} - \sigma^2}{\sigma^2_{-j}}$$

where σ^2 the variance of the Z_{it} from the final model and σ^2_{-j} is the variance of the Z_{it} when the $j^{-\text{th}}$ predictor is excluded from the final model.

4.2.2.1 An early warning system based on exceedance probabilities

Using the best model identified in the previous stage of the analysis, we compute the exceedance probability (EP), i.e. the predictive probability that incidence exceeds a predefined threshold, say l, formally expressed as

$$EP_{it} = \Pr(\hat{\lambda}_{it} > l | y_{it}).$$

Values of EP close to one indicate that incidence is highly likely to be above l, while the values of EP close to zero indicate that incidence is highly likely to be below l. Finally, values of EP around 0.5 indicate that incidence are equally likely to be above or below l, thus implying a scenario with highest uncertainty.

For a given LSOA and month, an alarm is then triggered whenever the EP exceeds a value, say p. To identify an optimal value of p, we maximise the *sensitivity* (the ability of the early warning system to trigger alarms in districts where it exceeds l) and *specificity* (the ability of the early warning system not to trigger alarms in districts where it does not exceed l) of the early warning system using the test set from April 2017 to March 2018. Finally, we summarise the predictive power of the model using the area under the Receiver Operating Characteristic (ROC) (Bradley, 1997) curve (henceforth, AUC).

4.3 Result

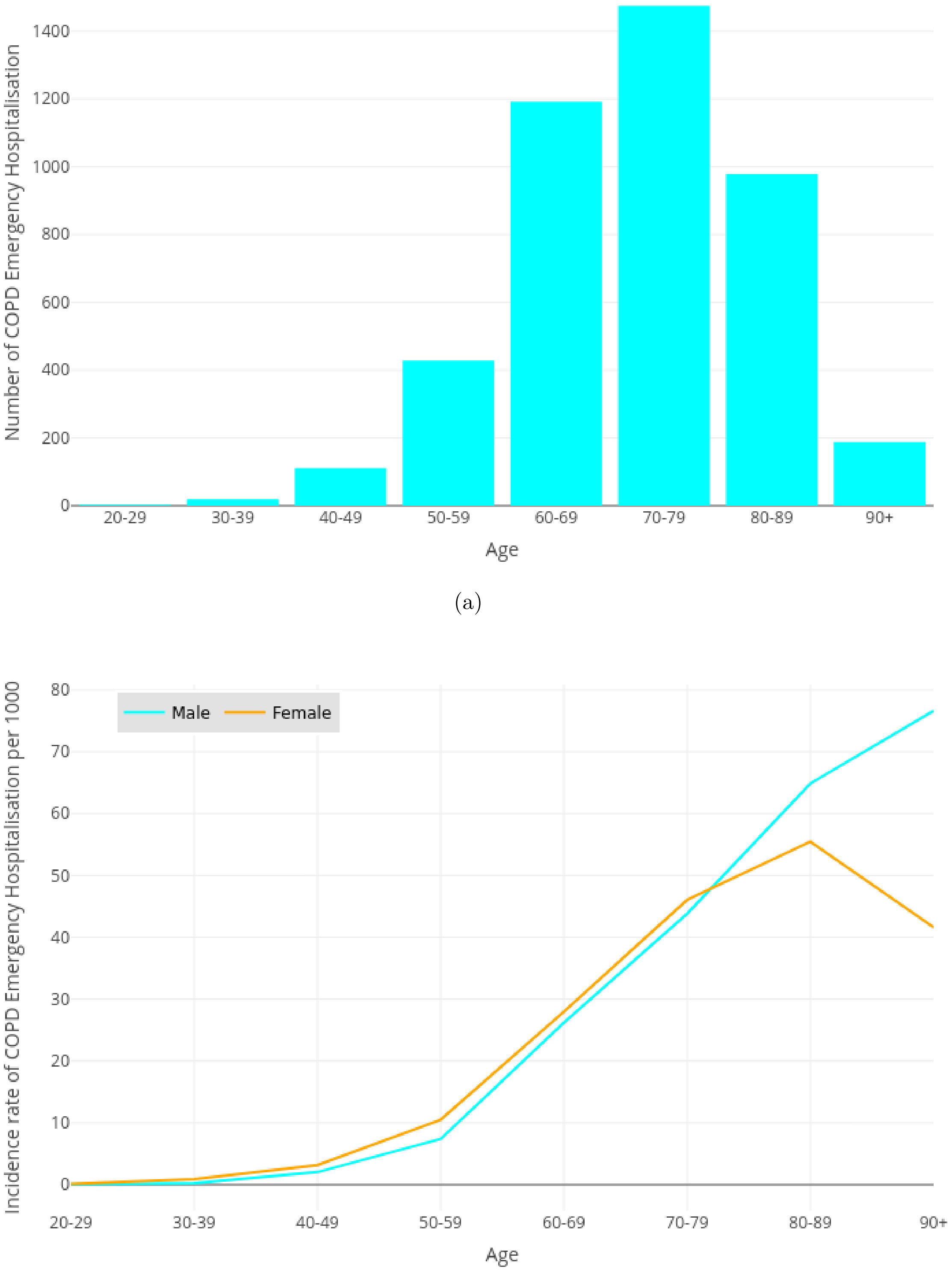
4.3.1 Descriptive Analysis

The age distribution of the COPD admissions is shown in Figure 4.2a. We observe the largest number of admissions for the age group 70-79. The COPD admissions incidence by sex show that females (Kilic et al., 2015). We also explore the incidence rate of COPD emergency admision by age group and sex in Figure 4.2b. These rates were calculated by dividing the number of admissions by the total number of male or female population in that age group. Clearly, incidence rate in males and females show a similar pattern, with slightly higher rate in females, up to the age group 70-79, beyond which incidence for female start to drop while incidence for males continue to increase. One reason for this could be that smoking (a major cause of COPD) was not common in females many years ago.

As expected, the empirical pattern of monthly counts of COPD emergency admission showed a seasonal pattern with the highest peaks found in the winter period each year, especially January and December (Figure 4.3). It is well established that COPD patients suffer from increased exacerbation and a decline in lung function during cold weather (Donaldson et al., 1999). The number of admissions is lowest in September.

4.3.2 Spatio-temporal Analysis

By applying the variables selection procedure described in Section 2.2, our final set of predictors consists of minimum temperature, PM10, income deprivation.



(a)

(b)

Figure 4.2: (a) Count of COPD emergency admission, by age group and sex, in South Cumbria and North Lancashire, 2012-2018; and (b) Incidence rate of COPD emergency admission per 1000 population, by age group and sex, in South Cumbria and North Lancashire, 2012-2018.

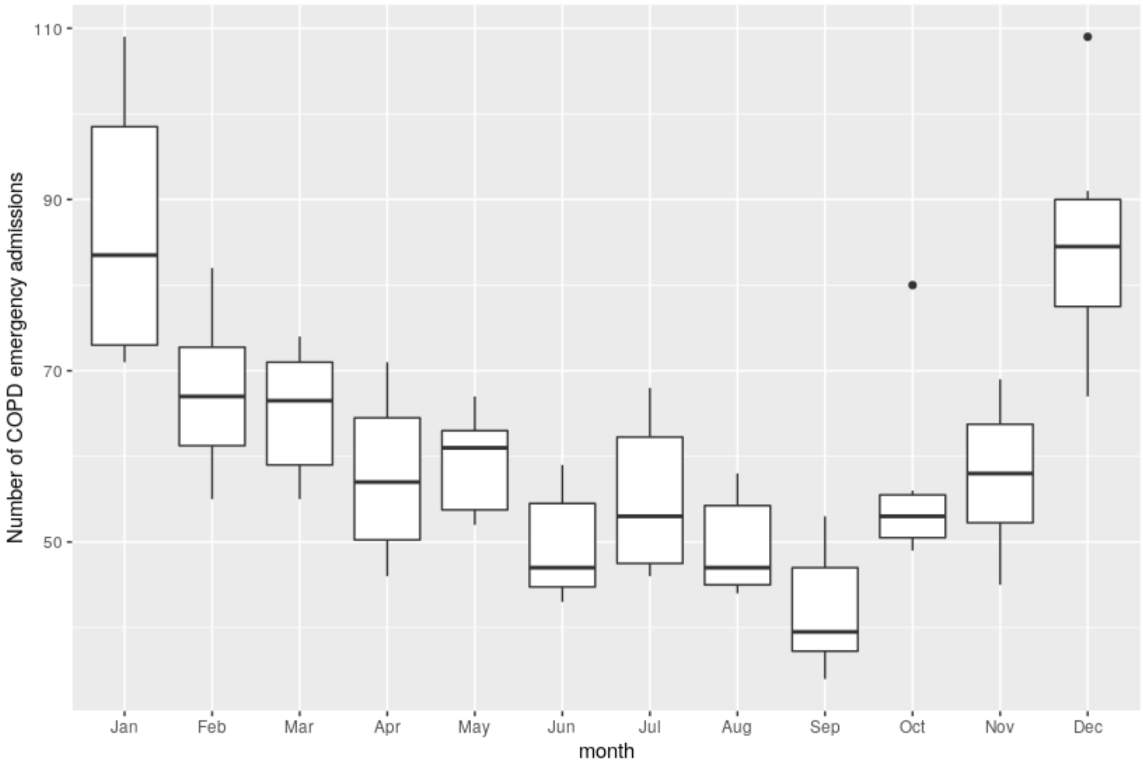


Figure 4.3: Boxplot showing seasonal variation in the monthly count of COPD emergency admission in South Cumbria and North Lancashire, 2012-2018.

Table 4.2 shows the relative variance reduction (RVR) of each predictor in the model. We find that, overall, the selected predictor variables explain about 22% of the variability in the residual random effects. Among these variables, income deprivation attained the largest RVR of about 14%.

In order to test whether the predictors included in this model can capture all the spatio-temporal correlation in the data, we applied the Monte Carlo procedure of Section 4.2.2 based on the spatio-temporal variogram for both an intercept-only model, that excludes all of predictors of the final model (Figure 4.4), and the final model (Figure 4.5). A comparison between Figures 4.4 and 4.5 indicates that the predictors used in the final model allowed us to capture most of the residual spatio-temporal correlation in COPD emergency admissions.

We then predict the incidence of COPD emergency admission for April 2017 – March

2018 and classify each LSOA as being above or below an incidence threshold l which we set to 12 per 100,000, a choice which is informed by the experts having tried different thresholds since there is currently no data or statistics to inform this value. For this threshold, we found that the value of EP that maximizes the sensitivity and specificity of the early warning system was $p = 0.85$, yielding a 72% sensitivity and a 70% specificity. We also found that the area under the curve of the final model was about 78% (Figure 4.6) which indicates a satisfactory predictive performance. Figure 4.7 shows the LSOA that were correctly and incorrectly classified based on our modelling approach. Whilst it is evident that our model can potentially predict correctly in most LSOAs, there exist a very few LSOAs with incorrect prediction.

Table 4.2: The table showing the relative variance reduced by the predictors.

Predictors	**RVR (%)**
Minimum temperature	1.23
PM_{10}	0.88
Income deprivation	14.33
Proportion over age 75	3.35
Proportion of male	0.05
All predictors	21.58

4.4 Discussion

We have developed a predictive statistical model for the incidence of COPD in South Cumbria and North Lancashire district (Northwest England). Our predictive model uses a combination of environmental and socio-economic variables as predictors.

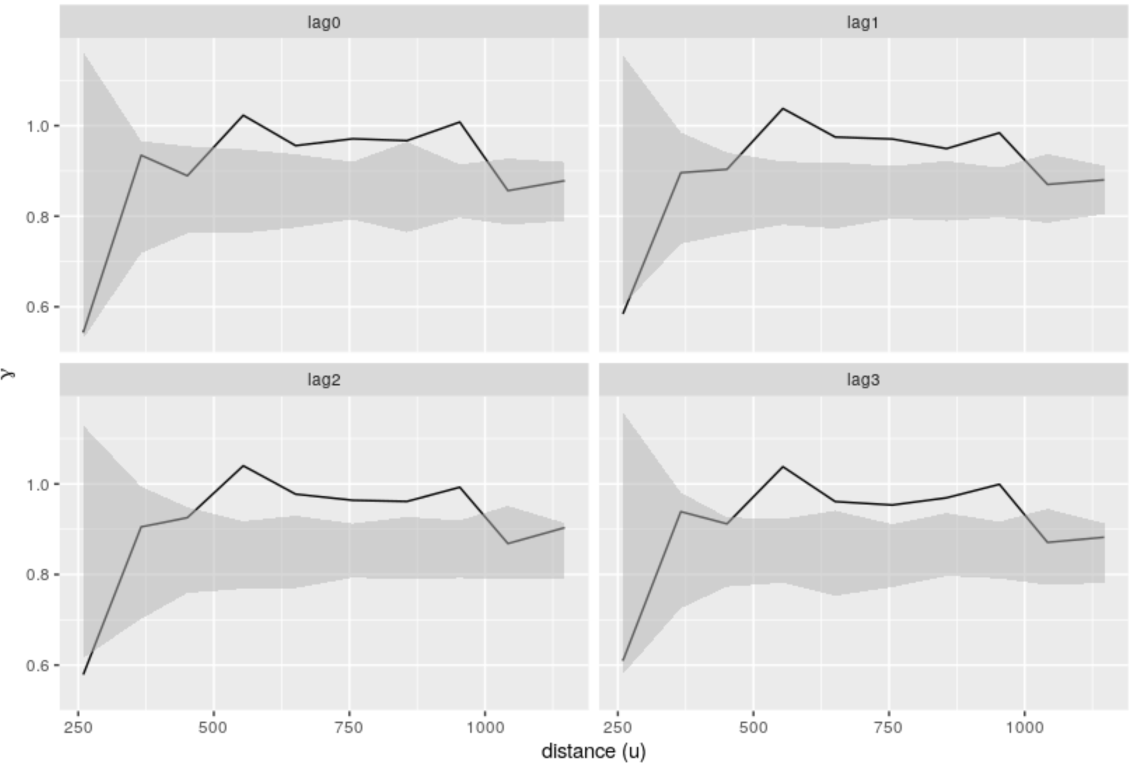


Figure 4.4: Spatio-temporal variogram of the residual from an intercept only model. This shows an evidence of spatio-temporal variation.

We also demonstrated that instead of predicting the incidence, a more meaningful prediction would be to predict the exceedance of clinically relevant threshold beyond which COPD risk is of public health concern.

Another major finding of this study is that after including the predictors into the model, we observed no presence of residual spatio-temporal variation meaning that the predictors have captured the spatio-temporal structure in the incidence. However, suppose a spatio-temporal structure is observed in the residual, we would have considered modelling Z_{it} as a spatio-temporal Gaussian process. Hence, consider y_{it} as a realisation of a spatio-temporal log-Gaussian Cox process, we refer the reader to Johnson et al. (2019).

Also, we found that income deprivation reduced the highest proportion of variance in the monthly incidence of COPD emergency admission. It has been shown in other

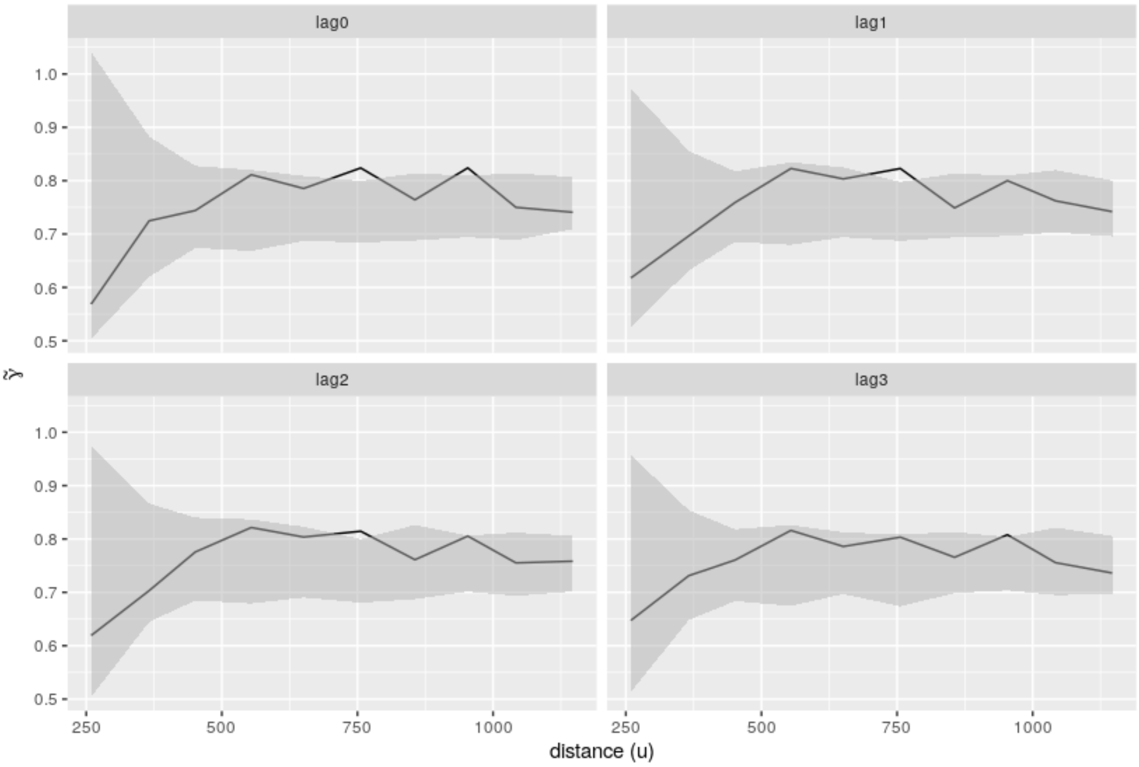


Figure 4.5: Spatio-temporal variogram of the residual from model including all the predictors. This shows that there is no evidence of spatio-temporal variation.

studies that people who live in deprivation are more likely to be admitted (Calderón-Larrañaga et al., 2011; McAllister et al., 2013). This suggests that a good predictive model for incidence of COPD should take into account socio-economic status.

A strong correlation exists among the set of potential predictors of COPD emergency admission. Liverani et al. (2016) have shown that IMD and air pollution are collinear; as well as some domains of deprivation. Income and employment deprivation is highly correlated, which is clearly due to the way income deprivation is measured. Income deprivation measures the proportion of the population experiencing deprivation due to low income, whereas people with low-income are those who are out of work or receive low earnings at work. Environmental deprivation is also correlated with PM10 - which makes the use of IMD and PM10 unfeasible in a single model. Furthermore, Income and education are also correlated. Therefore

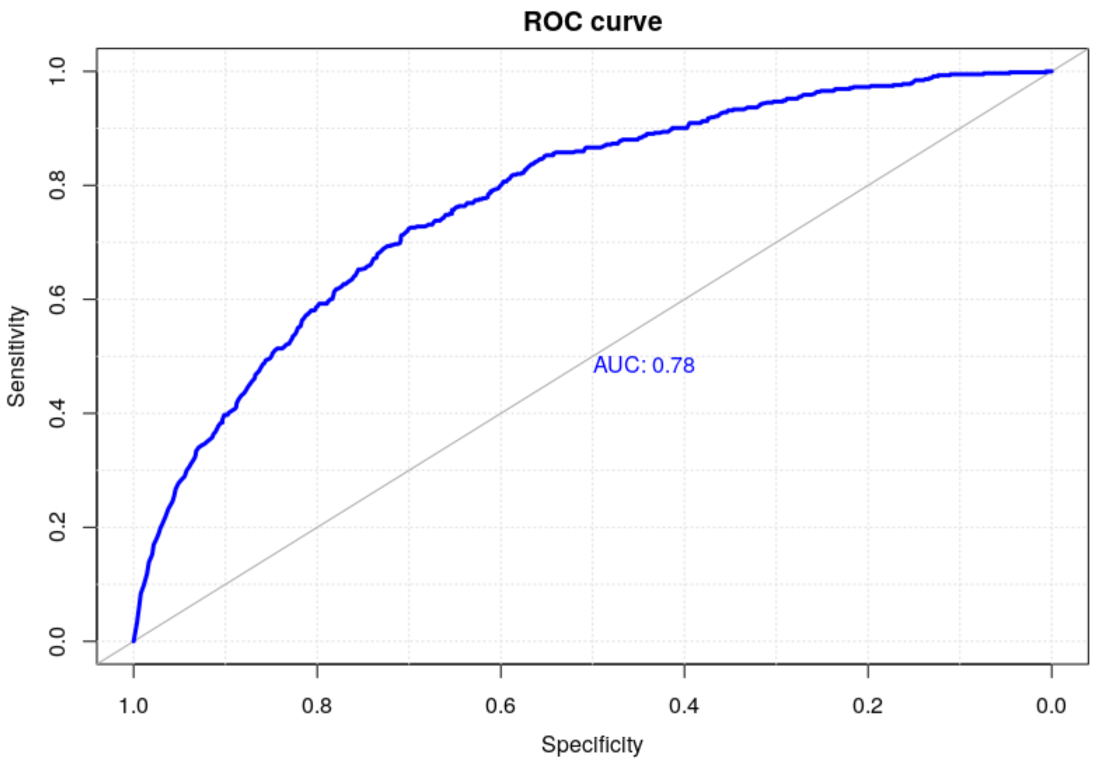


Figure 4.6: The receiver's characteristics curve (ROC) with area under the curve (AUC) =0.78. The red dot indicates the value of the sensitivity and specificity for which the optimal cut off p = 0.85 value was chosen.

including all the domains of IMD separately in a single model is not plausible.

Our developed warning system can potentially help to inform NHS Morecambe Bay CCG and policymakers where to target intervention/resources as well as reducing the need for hospital care or unplanned COPD emergency admission. The Met Office Health forecasting team has developed a similar approach in the past using an algorithm model, which predicts times when COPD patients are at elevated risk of having a flare-up (Bakerly et al., 2011; Hemming et al., 2009; Marno et al., 2010). The details of the model are no longer available on the Met Office webpage. However, we utilised a fully parametric statistical predictive model, which is well understood and can provide an estimate of the uncertainties. Our model can be updated in real-time, which in turn will lead to better sensitivity and specificity as one would

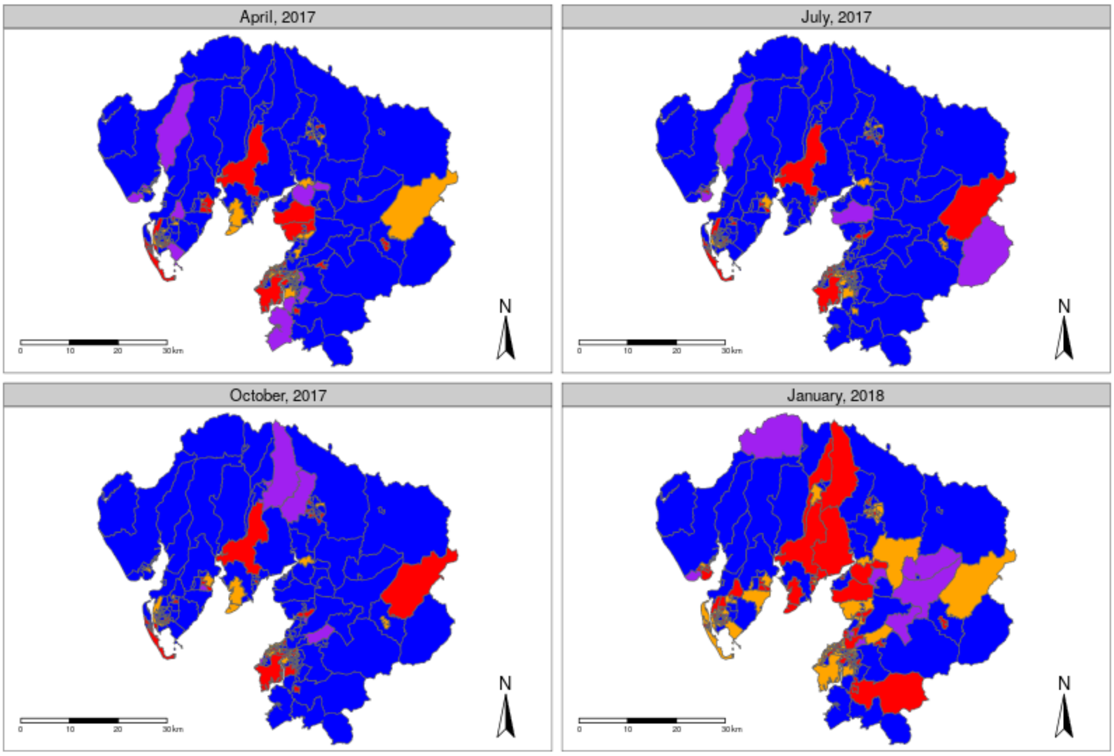


Figure 4.7: The monthly-predicted surveillance maps comparing the predicted and the true alarm for each LSOA. Colour blue indicates an LSOA that is correctly predicted to be below the threshold; orange indicates an LSOA that is correctly predicted to be above the threshold; purple indicates an LSOA that is incorrectly predicted to be below the threshold; and red indicates an LSOA that is incorrectly predicted to be above the threshold. The incidence threshold used is 12 per 100,000.

expect from short-range prediction.

The model performs fairly well at predicting LSOA-level incidence of COPD emergency admission in the test set, however, there is clear room for improving the predictive accuracy. The value of the true positive rate is quite interesting which suggest that our model can potentially identify 72% of the high incidence LSOAs. A good warning system model needs to achieve a balance between the sensitivity and specificity in order to avoid the waste of resources and identifying "real" high incidence LSOA. A warning system with high sensitivity is capable to detect LSOAs with “real” high incidence, but suffer losses from incurring additional resources due to low specificity. Similarly, a warning system with high specificity benefits from a significant reduction in the consumption of resources but has a decreased capacity to detect “real” high incidence LSOA due to low sensitivity. However, our model has high sensitivity and high specificity suggesting a good balance.

Also, note that choosing the optimal value of p by maximizing the sensitivity and sensitivity, implying sensitivity to be equal to specificity, is a pragmatic choice. There are other instances when sensitivity is preferred to be greater than specificity and vice versa depending on the risk and cost of the choice.

Limitations in the predictors and unavailability of other predictors affect the predictive accuracy of the model. Out of the predictors, the proportion of male and proportion of people over the age of 75 do not have any limitation as they were derived from the COPD health record provided by NHS Morecambe Bay. There is also a limitation in how a single year value of income deprivation was used for the entire years of study. The government published deprivation data at some specific time point and we used the one released in 2015 since that is the only one released during the period of study. PM10 data were only measured at few monitoring stations in the UK and the data was interpolated over the entire area. There are several limitations with this, one is that if there is a large variability between the monitoring sites it will increase the interpolation error and second is that aggregating the quantities over the LSOAs can further increase the error. The monthly minimum temperature data is available as a raster over a 1km grid and then aggregated over the LSOAs. The use of average temperature across the month is a limitation, as it does not allow us to account for the variation across the month.

The smoking rate would have been a very good predictor for our predictive model but getting such data is a challenge. We thought of using lung cancer rate as used in other studies as a proxy but it is also not readily available. Other variables that would have improved our predictive accuracy that is not available are influenza rate, and proportion of the population employed in mining or agriculture.

Our results demonstrate how to develop a predictive model and an early warning system for COPD emergency admission. Potential applications of the early warning system include identification and notification of high incidence areas of COPD emergency admission; and ability to support resource allocation for the healthcare system. Future studies will improve the model by accounting for more risk factors that are not captured in the study.

Bibliography

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, 27(2):93–115.

Bahadori, K. and FitzGerald, J. M. (2007). Risk factors of hospitalization and readmission of patients with copd exacerbation–systematic review. *International journal of chronic obstructive pulmonary disease*, 2(3):241.

Bakerly, N., Roberts, J. A., Thomson, A. R., and Dyer, M. (2011). The effect of copd health forecasting on hospitalisation and health care utilisation in patients with mild-to-moderate copd. *Chronic respiratory disease*, 8(1):5–9.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Billings, J., Dixon, J., Mijanovich, T., and Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *Bmj*, 333(7563):327.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Calderón-Larrañaga, A., Carney, L., Soljak, M., Bottle, A., Partridge, M., Bell, D., Abi-Aad, G., Aylin, P., and Majeed, A. (2011). Association of population and primary healthcare factors with hospital admission rates for chronic obstructive pulmonary disease in england: national cross-sectional study. *Thorax*, 66(3):191–196.

Chan, F. W., Wong, F. Y., Yam, C. H., Cheung, W.-l., Wong, E. L., Leung, M. C., Goggins, W. B., and Yeoh, E.-k. (2011). Risk factors of hospitalization and readmission of patients with copd in hong kong population: analysis of hospital admission records. *BMC health services research*, 11(1):186.

Chan, T.-C., Chiang, P.-H., Su, M.-D., Wang, H.-W., and Liu, M. S.-y. (2014). Geographic disparity in chronic obstructive pulmonary disease (copd) mortality rates among the taiwan population. *PloS one*, 9(5):e98170.

Donaldson, G., Seemungal, T., Jeffries, D., and Wedzicha, J. (1999). Effect of temperature on lung function and symptoms in chronic obstructive pulmonary disease. *European respiratory journal*, 13(4):844–849.

Hasegawa, W., Yamauchi, Y., Yasunaga, H., Sunohara, M., Jo, T., Matsui, H., Fushimi, K., Takami, K., and Nagase, T. (2014). Factors affecting mortality following emergency admission for chronic obstructive pulmonary disease. *BMC pulmonary medicine*, 14(1):151.

Hemming, D., Colman, A., James, P., Kaye, N., Marno, P., McNeall, D., McCarthy, R., Laing-Morton, T., Palin, E., Sachon, P., et al. (2009). Framework for copd forecasting in the uk using weather and climate change predictions. In *IOP Conference Series: Earth and Environmental Science*, volume 6, page 142021.

Holt, J. B., Zhang, X., Presley-Cantrell, L., and Croft, J. B. (2011). Geographic disparities in chronic obstructive pulmonary disease (copd) hospitalization among medicare beneficiaries in the united states. *International journal of chronic obstructive pulmonary disease*, 6:321.

Johnson, O., Diggle, P., and Giorgi, E. (2019). A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. *Statistics in Medicine*, 38(24):4871–4887.

Kauhl, B., Maier, W., Schweikart, J., Keste, A., and Moskwyn, M. (2018). Who is where at risk for chronic obstructive pulmonary disease? a spatial epidemiological analysis of health insurance claims for copd in northeastern germany. *PloS one*, 13(2):e0190865.

Kilic, H., Kokturk, N., Sari, G., and Cakır, M. (2015). Do females behave differently in copd exacerbation? *International journal of chronic obstructive pulmonary disease*, 10:823.

Liverani, S., Lavigne, A., and Blangiardo, M. (2016). Modelling collinear and spatially correlated data. *Spatial and spatio-temporal epidemiology*, 18:63–73.

Marno, P., Chalder, M., Laing-Morton, T., Levy, M., Sachon, P., and Halpin, D.

(2010). Can a health forecasting service offer copd patients a novel way to manage their condition? *Journal of health services research & policy*, 15(3):150–155.

Mathers, C. D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, 3(11):e442.

McAllister, D. A., Morling, J. R., Fischbacher, C. M., MacNee, W., and Wild, S. H. (2013). Socioeconomic deprivation increases the effect of winter on admissions to hospital with copd: retrospective analysis of 10 years of national hospitalisation data. *Primary Care Respiratory Journal*, 22(3):296.

Niyonsenga, T., Coffee, N., Del Fante, P., Høj, S., and Daniel, M. (2018). Practical utility of general practice data capture and spatial analysis for understanding copd and asthma. *BMC health services research*, 18(1):897.

Office for National Statistics (2018). *Mid-2017 estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. 2018.* Accessed 1 October 2019.

Osman, S., Ziegler, C., Gibson, R., Mahmood, R., and Moraros, J. (2017). The association between risk factors and chronic obstructive pulmonary disease in canada: A cross-sectional study using the 2014 canadian community health survey. *International journal of preventive medicine*, 8.

Samp, J. C., Joo, M. J., Schumock, G. T., Calip, G. S., Pickard, A. S., and Lee, T. A. (2018). Predicting acute exacerbations in chronic obstructive pulmonary disease. *Journal of managed care & specialty pharmacy*, 24(3):265–279.

Tian, Y., Dixon, A., and Gao, H. (2012). Data briefing. *King's Fund, London.*

Urwyler, P., Hussein, N. A., Bridevaux, P. O., Chhajed, P. N., Geiser, T., Grendelmeier, P., Zellweger, L. J., Kohler, M., Maier, S., Miedinger, D., et al. (2019). Predictive factors for exacerbation and re-exacerbation in chronic obstructive pulmonary disease: an extension of the cox model to analyze data from the swiss copd cohort. *Multidisciplinary respiratory medicine*, 14(1):7.

Wedzicha, J. A. (2004). Role of viruses in exacerbations of chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*, 1(2):115–120.

World Health Organisation (2016). *Chronic Obstructive Pulmonary Disease (COPD). World Health Organisation, fact sheet.* Accessed February 6, 2018.

Yii, A. C., Loh, C., Tiew, P., Xu, H., Taha, A. A., Koh, J., Tan, J., Lapperre, T. S., Anzueto, A., and Tee, A. K. (2019). a clinical prediction model for hospitalized copd exacerbations based on "treatable traits". *International journal of chronic obstructive pulmonary disease*, 14:719.

Chapter 5

General discussion, conclusions and future work

Whilst each paper contains its discussion, we further give an extensive discussion of each paper in this chapter. Also, we outline some future extensions of the problems tackled and how we can improve and widen the applicability of the methods developed.

5.1 Summary and future extensions of SDALGCP models

In Chapter 2 of this thesis, we developed a spatially discrete approximation (SDA) to LGCP models in order to carry out a spatial prediction of disease risk at any desired spatial scale using spatially aggregated disease count data. As variation in disease risk occurs in a spatial continuum irrespective of the format in which the data are available, we consider the LGCP framework to be a natural statistical paradigm for modelling aggregated disease count data. However, when computational constraints make the fitting of an LGCP infeasible, we argue that our approach, SDALGCP provides a computationally efficient solution while respecting the spatially continuous nature of disease risk.

The method proposed can be extended to spatio-temporal and multivariate outcome cases. Extension to spatio-temporal case can be considered when disease cases is spatially aggregated over space and time. For example, the COPD emergency admission dataset that we analysed in Chapter 4. In this dataset, COPD emergency admission cases are aggregated over the LSOAs and the months, April 2012 to March 2018. A spatio-temporal analysis will proceeds as follows: let y_{it} denotes the COPD emergency admission count for LSOA i and time t; let d_{it} be a vector of explanatory variables for LSOA i and time t with corresponding coefficient β; m_{it} be the population count; and let S_{it} be a spatio-temporal Gaussian process. By modelling the y_{it} as realisations of a spatio-temporal log-Gaussian Cox process we obtained the approximation to the mean count as

$$\mu_{it} = m_{it} \exp\{d_{it}\beta^* + S_{it}^*\}. \tag{5.1}$$

The most common approach is to assume a separable covariance form for S_{it}^* such that:

$$\text{cov}\{S(x,t), S(x',t')\} = \sigma^2 \exp\{-\|x - x'\|/\phi\} \exp\{-|t - t'|/\psi\}.$$

Since the time index t is observed at discrete time, the simplest and the most frequently used model to account for temporal correlation is the AR(1) process, which assumes the form

$$S_t^* = \varphi S_{t-1}^* + W_t, 0 < \varphi < 1, \tag{5.2}$$

where the temporal innovation W_t is modelled as a multivariate Gaussian distribution with covariance matrix $\sigma^2 V$, modelled as given by Equation (2.7). The parameter φ controls the influence of the lagged values S^*_{t-1} on S^*_t. Note that if we define $\varphi = \exp\{-1/\psi\}$, the AR(1) process in (5.2) can be interpreted as a discretized version of a continuous-time process with exponential correlation function. From the assumption in Equation 5.2, it follows that the joint distribution of S^*_t can be re-written as in terms of conditional density such that

$$\begin{aligned} f(S^*) &= f(S^*_1, \dots, S^*_T) \\ &= f(S^*_1) f(S^*_2 | S^*_1) \dots f(S^*_{T-1} | S^*_{T-2}) f(S^*_T | S^*_{T-1}) \end{aligned}$$

then the log-joint density is

$$\log f(S^*) = \sum_{t=1}^{T} \log f(S^*_t) = \log f(S^*_1) + \sum_{t=2}^{T} \log f(S^*_t | S^*_{t-1}),$$

where

$$\log f(S^*_1) = -\frac{1}{2}\left[n \log 2\pi + n \log \left(\frac{\sigma^2}{(1-\rho^2)} \right) + \log |V| + {S^*_1}^\top \left(\frac{\sigma^2 V}{(1-\rho^2)} \right)^{-1} S^*_1 \right]$$

and

$$\log f(S^*_t | S^*_{t-1}) = -\frac{1}{2}\left[n \log 2\pi + n \log \sigma^2 + \log |V| + \left(S^*_t - \rho S^*_{t-1} \right)^\top (\sigma^2 V)^{-1} \left(S^*_t - \rho S_{t-1} \right) \right].$$

The inference for the spatio-temporal model follows directly from the static spatial case. Note that the joint log-density can now be computed in parallel because each conditional density can be evaluated independently of each other which in turn leads to computationally efficient evaluation of the likelihood.

Extension to multivariate version can also be considered. This is when there are $k > 1$ number of outcomes measured at each spatial unit. For example, suppose we are interested in analysing data related to a respiratory condition, namely: COPD, Asthma and Lung Cancer. $k = 3$ in this case. And, we have the case-count of each disease for each spatial unit. Two basic questions are usually of interest in multivariate analysis, one is to examine the dependence between disease count in each unit, and second is to examine the association between measurements across the units. There are two general frameworks for analysing this problem, namely: conditioning and joint approach. An extensive discussion on multivariate models in geostatistics is provided in Wackernagel (2013). SDALGCP extension to multivariate analysis will proceed as follows: let y_{ij} denotes the disease cases count for LSOA i and disease j; let d_{ij} be a vector of explanatory variables for LSOA i and disease j with corresponding coefficient β_j; and let S^*_{ij} be a Gaussian process. To model the data, we consider modelling y_{ij} as a realisation of a spatially aggregated log-Gaussian Cox process and we obtained the approximation to the conditional mean count as

$$\mu_{ij} = m_{ij} \exp\{d_{ij}\beta^*_j + S^*_{i0} + S^*_{ij}\}, \tag{5.3}$$

where S^*_{i0} is the Gaussian process common to all the diseases and S^*_{ij} is the Gaussian process specific to each disease. This specification of model 5.3 with a spatially continuous Gaussian process for S^*_{i0} and S^*_{ij} allows us to capture when the outcome is observed at a common or misaligned spatial unit. Method for inference of this model can proceed as discussed in the paper in Chapter 2.

Our modelling approach can be extended to several other problems. One is the changing boundary problem where the partitioning of the entire regions changes with time; see Taylor et al. (2018) for example. Another is when the outcome variable is a combination of point and aggregated data. (Wilson and Wakefield, 2018) and Moraga et al. (2017) have addressed this problem by assuming a common underlying continuous surface and use SPDE approach (Lindgren et al., 2011) to model the latent field.

5.2 Summary and future extensions of geostatistical methods for analysing spatially misaligned areal data

In chapter 3, we developed a model-based geostatistical approach to model the relationship between life expectancy at birth (LEB) and the index of multiple deprivation (IMD) when these are provided over misaligned partitions of the study area. One of the main advantages of our approach is that it allows combining information from multiple data sources without coarsening their resolution to a common spatial scale. The underpinning principle of our modelling framework is that spatially aggregated data should be treated as the realization of an aggregated spatially continuous stochastic process. As a result of this, the proposed modelling paradigm allows to carry out spatially continuous prediction. Also, we emphasize that spatially continuous prediction allows us to examine the within areal unit variation that is usually hidden in an aggregated estimate. Furthermore, we showed how the use of non-exceedance probabilities can provide a way of measuring uncertainty in relation to a predefined threshold in order to identify areas that need urgent intervention. We also demonstrated that instead of having a static map at different thresholds, Shiny App (Chang et al., 2019) provides a modern web-based technology that allows the user to interactively move a slide bar to a different threshold and visual the uncertainties. Therefore, we encourage the development of a web-based technology for exceedance probability map, especially in public health disease mapping.

Area-level modelling can result in over or under-estimation of the association, a phenomenon generally referred to as ecological bias. However, whilst a literal interpretation of our model is that the β parameter data measures the association between IMD and LEB at a point, in practice it should be restricted to the smallest spatial resolution for which data are available. Note also that the model does not treat LEB and IMD symmetrically. The ordering of the two parts of Equation 3.12 matters, and was guided by our objective of using IMD to predict LEB, rather than vice versa.

The method can be conveniently extended to more than one predictor and can be handled using the multiple linear generalisation of the joint model-based geostastistical approach that we developed.

On a final note on Chapter 2 and 3, both papers have emphasised the use of a spatially continuous model for epidemiological research, either as a solution to a spatial misalignment problem or as a method of making spatially continuous inferences in as much as it respects our scientific knowledge of the problem.

5.3 Summary and future extensions of spatial-temporal modelling of COPD emergency admission

In this paper, we analyse the monthly COPD emergency admission dataset in North Lancashire and South Cumbria, 2012-2018. The spatio-temporal extension of the method developed in Chapter 2 would have been used to analyse the data but after accounting for the predictors, the residual is no longer spatio-temporally structured. One of the lessons learnt from this analysis is that it is important to check both of the need for the spatial random effects model and its appropriateness. Cox and

Wong (2010) has shown that appreciable bias may arise from misspecification of a random component.

This work has also demonstrated that exceedance probability is not only useful for quantifying uncertainty but can also be used in public health settings, where the goal is to identify areas where disease incidence exceeds a clinically relevant threshold.

In the future, we plan to apply our predictive model to a larger dataset, potentially to COPD emergency admission dataset in the entire North West England.

5.4 Software development

The methodology developed in Chapter 2 has been implemented in the open-source R package `SDALGCP` (Johnson et al., 2018). The package implements fitting and spatial prediction of a standard geostatistical model for the analysis of spatially and spatio-temporally aggregated disease count data. The package provides functions to perform 1) parameter estimation for static spatial and spatio-temporal data, 2) spatial and spatio-temporal prediction of disease risk both on a spatially discrete and spatially continuous scale. `SDALGCP` contains a vignette, which explains the details of the functions in the package; and gives a step by step tutorial on how to run the models with examples. We also develop a web shiny application for visualising uncertainty in the prediction that integrates nicely with the R package. The code for the shiny app is made available on a GitHub repository `https://github.com/olatunjijohnson/SDALGCPApp`. In the future, we plan to develop an R package for a model-based geostatistical solution to spatially misalignment problems.

Bibliography

Buck, D., Maguire, D., et al. (2017). Inequalities in life expectancy: changes over time and implications for policy. *Health*.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2019). *shiny: Web Application Framework for R*. R package version 1.3.2.

Cox, D. and Wong, M. (2010). A note on the sensitivity to assumptions of a generalized linear mixed model. *Biometrika*, 97(1):209–214.

Johnson, O., Giorgi, E., and Diggle, P. (2018). *SDALGCP: Spatially Discrete Approximation to Log-Gaussian Cox Processes for Aggregated Disease Count Data*. R package version 0.1.0.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Madsen, L. J. (2004). *Regression with spatially misaligned data*. Cornell University, May.

Moraga, P., Cramb, S. M., Mengersen, K. L., and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41.

Taylor, B. M., Andrade-Pacheco, R., and Sturrock, H. J. (2018). Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 181(4):1125–1150.

Wackernagel, H. (2013). *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media.

Wilson, K. and Wakefield, J. (2018). Pointless spatial modeling. *Biostatistics*.

Appendix

Appendix A for Paper 1

A.1 Proof of the Monte Carlo approximation

$$\begin{aligned}
L(\psi) &= \int_{\mathbb{R}^n} f(\eta;\psi)\ f(y|\eta)\ \frac{f(\eta,y;\psi_0)}{f(\eta,y;\psi_0)}\, d\eta \\
&= \int_{\mathbb{R}^n} \frac{f(\eta;\psi)\ f(y|\eta)}{f(\eta;\psi_0)\ f(y|\eta)}\ f(y,\eta;\psi_0)\, d\eta \\
&= f(y;\psi_0)\ \int_{\mathbb{R}^n} \frac{f(\eta;\psi)}{f(\eta;\psi_0)}\ f(\eta|y)\ d\eta \\
&= f(y;\psi_0)\ E_{\eta|y}\left[\frac{f(\eta;\psi)}{f(\eta;\psi_0)}\right] \\
&\propto E_{\eta|y}\left[\frac{f(\eta;\psi)}{f(\eta;\psi_0)}\right]. \qquad (4)
\end{aligned}$$

A.1.1 Likelihood and Derivatives

We can then approximate the likelihood function in (4) as

$$L(\psi) \approx L_N(\psi) \;=\; \frac{1}{N}\sum_{j=1}^{N} \frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}. \qquad (5)$$

As $N \to \infty$, in the above equation, $L_N(\psi)$ converges to $L(\psi)$. Hence the log of (5) is given by

$$l_N(\psi) \;=\; \log\left(\frac{1}{N}\sum_{j=1}^{N}\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}\right), \qquad (6)$$

Specifically, the gradient is given as

$$\nabla l_N(\psi) = \frac{\sum_{j=1}^{N}[\nabla f(\eta_{(j)};\psi)]\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}}{\sum_{j=1}^{N}\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}},$$

and the Hessian as

$$\begin{aligned}\nabla^2 l_N(\psi) \quad &= \quad \frac{\sum_{j=1}^{N}[\nabla^2 f(\eta_{(j)};\psi)]\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}}{\sum_{j=1}^{N}\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}} + \\ &\quad \frac{\sum_{j=1}^{N}[\nabla f(\eta_{(j)};\psi) - \nabla l_N(\psi)][\nabla f(\eta_{(j)};\psi) - \nabla l_N(\psi)]^\top\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}}{\sum_{j=1}^{N}\frac{f(\eta_{(j)};\psi)}{f(\eta_{(j)};\psi_0)}}.\end{aligned}$$

Expressions for $\nabla f(\eta_{(j)};\psi)$ and $\nabla^2 f(\eta_{(j)};\psi)$ can be found in Zhang (2002).

A.2 Diagnostic Plots of SDA Model

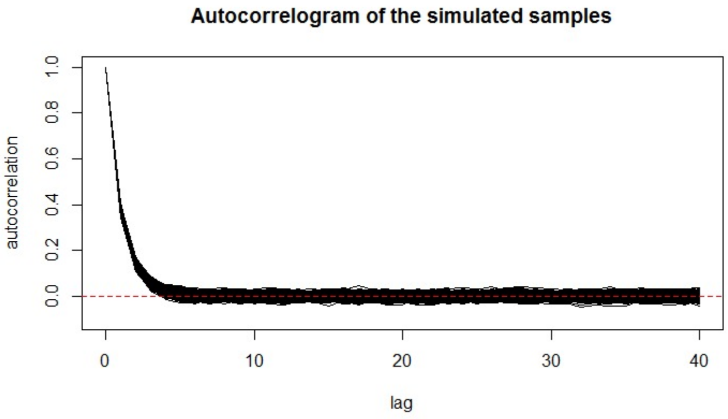


Figure 1: The plots show the Autocorrelation plot of the process, S.

A.3 Convergence, Mixing diagnostic and Plots of LGCP Model

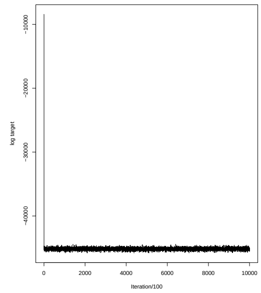
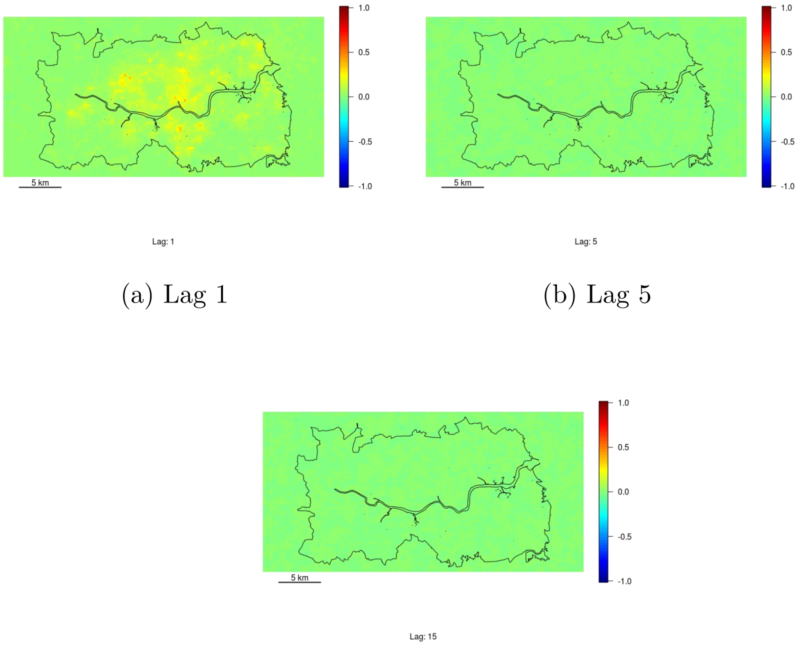


Figure 2: Diagnosing convergence to a posterior mode: a plot of the log-target, $\log\{\pi(\beta, \eta, S|N)\} + c$ up to an additive constant, c

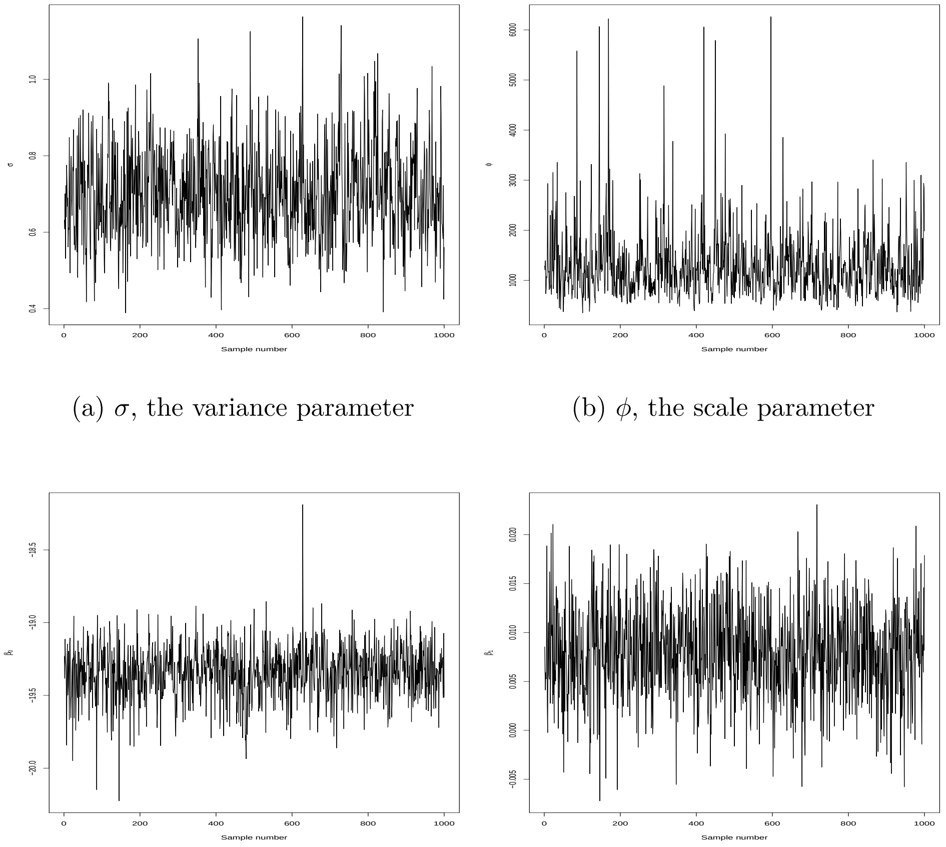


(a) Lag 1

(b) Lag 5

(c) Lag 15

Figure 3: The maps show the autocorrrlation plot three different lags Fig a: Lag 1 ; Fig b: Lag 5; Fig c: Lag 15.



(a) σ, the variance parameter

(b) ϕ, the scale parameter

(c) β_0, intercept

(d) β_1, IMD

Figure 4: The plots show the trace plot of the posteriors of the parameters.

A.4 Mixing diagnostic of BYM Model

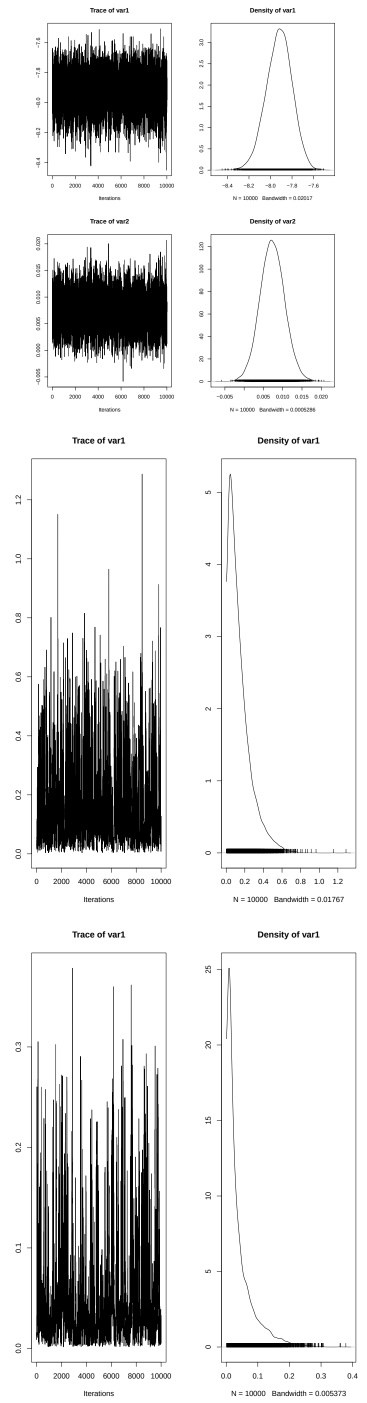


Figure 5: The plots show the trace plot of the posteriors of the fixed effect parameters, and random effect parameters, that is β_1, β_2, σ^2 and ϕ.

A.5 Mixing diagnostic of EV Model

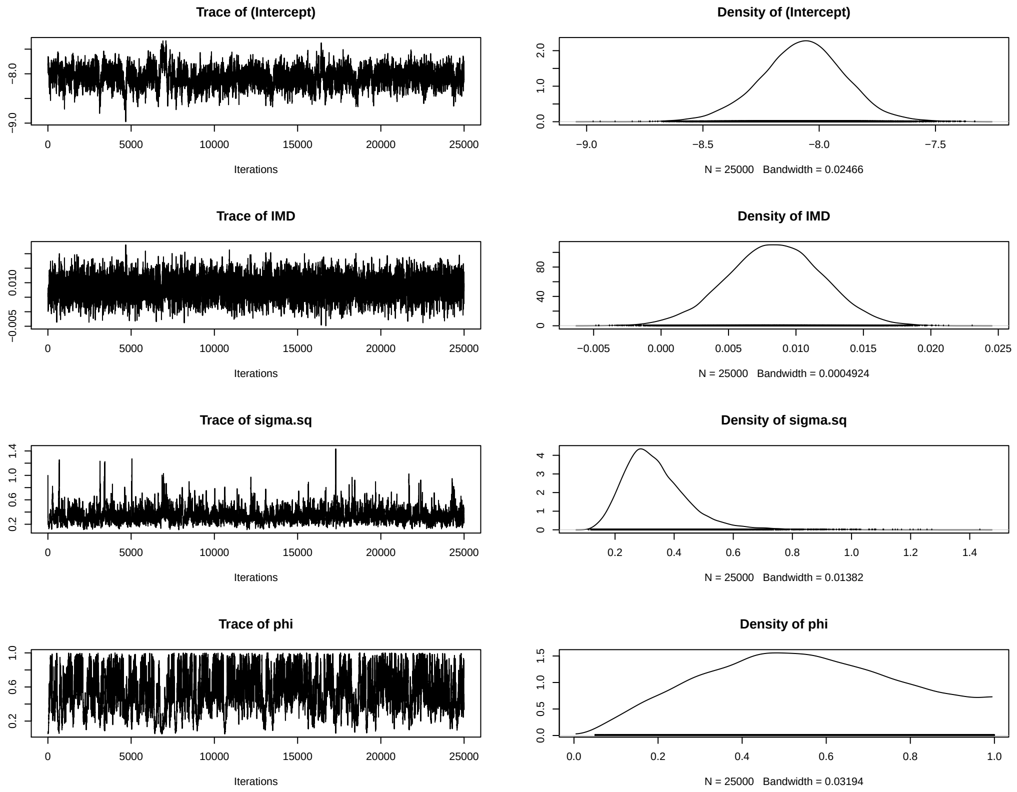


Figure 6: The plots show the trace plot of the posteriors of the fixed effect parameters, and random effect parameters, that is β_1, β_2, σ^2 and ϕ.

A.6 Simulation Plot

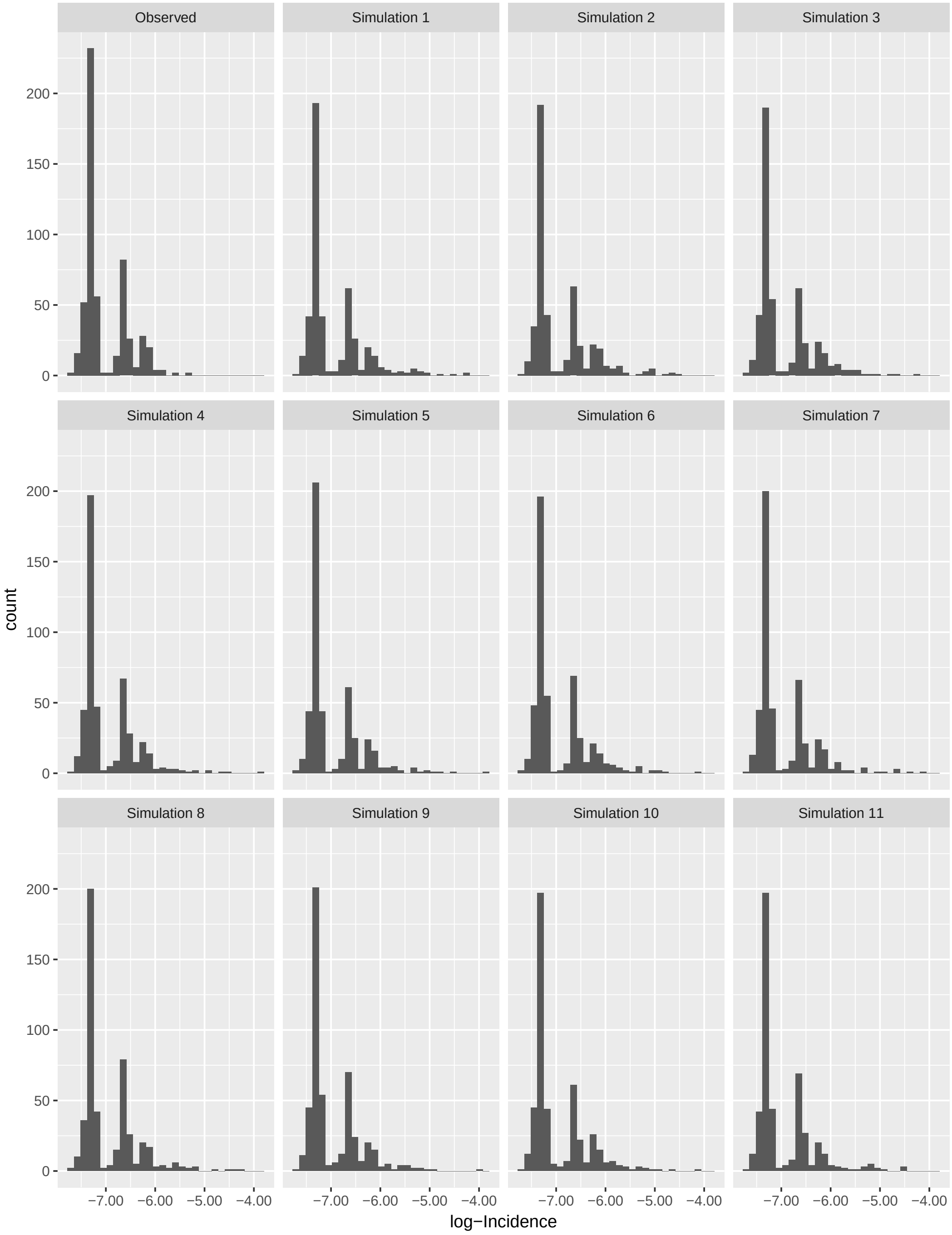


Figure 7: Histogram of the Observed log-incidence from the data (Observed, left-upper panel) and the first simulated log-incidence (Simulation 1, right-upper panel) and the second simulated log-incidence (Simulation 2, left-lower panel), and the third simulated log-incidence (Simulation 3, right-lower panel).

Appendix for Paper 2

A.7 Derivation of the equations in Paper 2

Let LEB_{ij} denote the life expectancy at birth for males, if $i = 1$, and females, if $i = 2$, at the j-th MSOA, henceforth $MSOA_j$, for $j = 1, \ldots, n$. Similarly, we use IMD_k to denote the IMD score for the k-th LSOA, henceforth $LSOA_k$, for $k = 1, \ldots, m$.

Define $U(x)$ to be a spatially continuous Gaussian process, with stationary and isotropic exponential covariance function, i.e.

$$\mathrm{Cov}\{U(x), U(x')\} = \tau^2 \exp\{-\|x - x'\|/\delta\},$$

where τ^2 is the variance and δ is a scale parameter regulating the rate of decay of the spatial correlation for increasing Euclidean distance $\|x - x'\|$ between any two locations x and x'.

We then model the cross-correlation in space between LEB and IMD through $U(x)$ as follows. Define the averaged spatial processes based on $U(x)$ over LSOAs and MSOAs as $U_j = \int_{MSOA_j} U(x)\, dx/|MSOA_j|$ and $U_k^* = \int_{LSOA_k} U(x)\, dx/|LSOA_k|$, where $|\mathcal{A}|$ corresponds to the area in m^2 of a spatial unit $\mathcal{A}$. The proposed joint model for LEB_{ij} and IMD_k takes the form

$$\begin{cases} LEB_{ij} = \alpha_i + \beta_i U_j + T_{ij} & \text{for } i = 1, 2; j = 1, \ldots, n \\ IMD_k = \gamma + U_k^* + V_k & \text{for } k = 1, \ldots, m \end{cases}, \tag{7}$$

where the β_i parameters quantify the strength of the association between LEB and IMD, whilst the α_i and γ are intercept parameters. Also in (7), the V_k are i.i.d.

Gaussian variables with mean zero and variance ν^2, whilst $(T_{1j},\, T_{2j})$ are i.i.d. bivariate Gaussian variables with mean zero and covariance matrix

$$\Omega = \begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}.$$

Theorem A.7.1.

$$\text{Cov}\{LEB_{ij}, IMD_k\} = \frac{\beta_i \tau^2}{|MSOA_j||LSOA_k|} f(MSOA_j, LSOA_k; \delta), \tag{8}$$

where

$$f(MSOA_j, LSOA_k; \delta) = \int_{MSOA_j} \int_{LSOA_k} \exp\left\{-\frac{\|x_j - x_k\|}{\delta}\right\} dx_j\, dx_k. \tag{9}$$

Proof.

$$\begin{aligned}
\text{Cov}\{LEB_{ij}, IMD_k\} &= \text{Cov}\{\alpha_i + \beta_i U_j + T_{ij}, \gamma + U_k^* + V_k\} \\
&= \text{Cov}\{\alpha_i, \gamma\} + \text{Cov}\{\alpha_i, U_k^*\} + \text{Cov}\{\alpha_i, V_k\} + \text{Cov}\{\beta_i U_j, \gamma\} \\
&+ \text{Cov}\{\beta_i U_j, U_k^*\} + \text{Cov}\{\beta_i U_j, V_k\} + \text{Cov}\{T_{ij}, \gamma\} \\
&+ \text{Cov}\{T_{ij}, U_k^*\} + \text{Cov}\{T_{ij}, V_k\} \\
&= \beta_i \text{Cov}\{U_j, U_k^*\} \\
&= \beta_i \text{Cov}\left\{\frac{1}{|MSOA_j|} \int_{MSOA_j} U(x_j)\, dx_j, \frac{1}{|LSOA_k|} \int_{LSOA_k} U(x_k)\, dx_k\right\} \\
&= \beta_i \frac{1}{|MSOA_j|} \int_{MSOA_j} \frac{1}{|LSOA_k|} \int_{LSOA_k} \text{Cov}\left\{U(x_j), U(x_k)\right\}\, dx_j dx_k \\
&= \beta_i \frac{1}{|MSOA_j|} \frac{1}{|LSOA_k|} \int_{MSOA_j} \int_{LSOA_k} \tau^2 \exp\left\{-\frac{\|x_j - x_k\|}{\delta}\right\}\, dx_j dx_k \\
&= \frac{\beta_i \tau^2}{|MSOA_j||LSOA_k|} \int_{MSOA_j} \int_{LSOA_k} \exp\left\{-\frac{\|x_j - x_k\|}{\delta}\right\}\, dx_j dx_k \\
&= \frac{\beta_i \tau^2}{|MSOA_j||LSOA_k|} f(MSOA_j, LSOA_k; \delta),
\end{aligned}$$

where $\text{Cov}\{\alpha_i, \gamma\} = 0$, $\text{Cov}\{\alpha_i, U_k^*\} = 0$, $\text{Cov}\{\alpha_i, V_k\}$, $\text{Cov}\{\beta_i U_j, \gamma\} = 0$, $\text{Cov}\{\beta_i U_j, V_k\} = 0$, $\text{Cov}\{T_{ij}, \gamma\} = 0$, $\text{Cov}\{T_{ij}, U_k^*\} = 0$, and $\text{Cov}\{T_{ij}, V_k\} = 0$. □

Theorem A.7.2. *Let Σ_{LSOA} be the spatial covariance matrix of the IMD at LSOA-level. The (k, k') entry for Σ_{LSOA} is*

$$(\Sigma_{LSOA})_{kk'} = \frac{\tau^2}{|LSOA_k||LSOA_{k'}|} f(LSOA_k, LSOA_{k'}; \delta) \tag{10}$$

Proof. The (k, k') entry for Σ_{LSOA} is

$$\begin{aligned}
\text{Cov}\{IMD_k, IMD_{k'}\} &= \text{Cov}\{\gamma + U_k^* + V_k, \gamma + U_{k'}^* + V_{k'}\} \\
&= \text{Cov}\{\gamma, \gamma\} + \text{Cov}\{\gamma, U_{k'}^*\} + \text{Cov}\{\gamma, V_{k'}\} + \text{Cov}\{U_k^*, \gamma\} \\
&+ \text{Cov}\{U_k^*, U_{k'}^*\} + \text{Cov}\{U_k^*, V_{k'}\} + \text{Cov}\{V_k, \gamma\} \\
&+ \text{Cov}\{V_k, U_{k'}^*\} + \text{Cov}\{V_k, V_{k'}\} \\
&= \text{Cov}\{U_k^*, U_{k'}^*\} \\
&= \text{Cov}\left\{\frac{1}{|LSOA_k|}\int_{LSOA_k} U(x_k)\,dx_k, \frac{1}{|LSOA_k'|}\int_{LSOA_k'} U(x_k')\,dx_k'\right\} \\
&= \frac{1}{|LSOA_k|}\int_{LSOA_k} \frac{1}{|LSOA_k'|}\int_{LSOA_k'} \text{Cov}\left\{U(x_k), U(x_k')\right\}\,dx_k dx_k' \\
&= \frac{1}{|LSOA_k|}\frac{1}{|LSOA_k'|}\int_{LSOA_k}\int_{LSOA_k'} \tau^2 \exp\left\{-\frac{\|x_k - x_{k'}\|}{\delta}\right\}\,dx_k dx_k' \\
&= \frac{\tau^2}{|LSOA_k||LSOA_{k'}|}\int_{LSOA_k}\int_{LSOA_k'} \exp\left\{-\frac{\|x_k - x_{k'}\|}{\delta}\right\}\,dx_k dx_k' \\
&= \frac{\tau^2}{|LSOA_k||LSOA_{k'}|} f(LSOA_k, LSOA_{k'}; \delta),
\end{aligned}$$

where $\text{Cov}\{\gamma, \gamma\} = 0$, $\text{Cov}\{\gamma, U_{k'}^*\} = 0$, $\text{Cov}\{\gamma, V_{k'}\}$, $\text{Cov}\{U_k^*, \gamma\} = 0$, $\text{Cov}\{U_k^*, V_{k'}\} = 0$, $\text{Cov}\{V_k, \gamma\} = 0$, $\text{Cov}\{V_k, U_{k'}^*\} = 0$, and $\text{Cov}\{V_k, V_{k'}\} = 0$. □

Lemma A.7.3. *Suppose a multivariate Guassian random vector X is partitioned into two component $X = (X_1, X_2)^T$, where X_1 has q_1 components and X_2 has q_2 components. Then the joint distribution of X_1 and X_2 has mean vector*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right),$$

where μ_i has length $q_i : i = 1, 2$ and Σ_{ij} is a $q_i \times q_j$ matrix for $i, j = 1, 2$. Then the conditional distribution of X_1 given $X_2 = x_2$ follows a Gaussian distribution with *mean*

$$\mathbb{E}[X_1|X_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$

and variance

$$\mathbb{V}[X_1|X_2] = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Theorem A.7.4. *$[LEB_1, LEB_2 \mid IMD; \theta]$ is a multivariate Gaussian with mean*

$$\alpha \oplus \mathbb{1}_{n\times 1} + C^\top \Sigma_{LSOA}^{-1}(IMD - \gamma \mathbb{1}_{m\times 1}), \tag{11}$$

and covariance

$$\Sigma_{LEB} - C^\top \Sigma_{LSOA}^{-1} C, \tag{12}$$

where: $\alpha = (\alpha_1, \alpha_2)^\top$; $\oplus$ is the Kronecker product; $C = (C_1, C_2)^\top$ with C_i being the cross-covariance between LEB_i and IMD whose entries are given by Equation (8)*; finally,*

$$\Sigma_{LEB} = \begin{pmatrix} \beta_1^2\Sigma_{MSOA} + w_1^2\mathbb{I}_n & \beta_1\beta_2\Sigma_{MSOA} + w_{12}\mathbb{I}_n \\ \beta_1\beta_2\Sigma_{MSOA} + w_{12}\mathbb{I}_n & \beta_2^2\Sigma_{MSOA} + w_2^2\mathbb{I}_n \end{pmatrix}.$$

Proof. According to Lemma A.7.3, let $X = (LEB_1, LEB_2, IMD)^T$ be partitioned into two parts such that $X = ((LEB_1, LEB_2), IMD)^T$. Therefore, if the joint distribution of (LEB_1, LEB_2) and IMD is

$$X = \begin{bmatrix} (LEB_1, LEB_2) \\ IMD \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \alpha \oplus \mathbb{1}_{n\times 1} \\ \gamma \mathbb{1}_{m\times 1} \end{bmatrix}, \begin{bmatrix} \Sigma_{LEB} & C^\top \\ C & \Sigma_{LSOA} \end{bmatrix} \right),$$

then the conditional distribution of (LEB_1, LEB_2) given IMD follows a Gaussian distribution with mean

$$\alpha \oplus \mathbb{1}_{n\times 1} + C^\top \Sigma_{LSOA}^{-1}(IMD - \gamma \mathbb{1}_{m\times 1}),$$

and covariance

$$\Sigma_{LEB} - C^\top \Sigma_{LSOA}^{-1} C,$$

where: $\alpha = (\alpha_1, \alpha_2)^\top$; $\oplus$ is the Kronecker product; $C = (C_1, C_2)^\top$ with C_i being the cross-covariance between LEB_i and IMD whose entries are given by Equation (8); finally,

$$\begin{aligned} \Sigma_{LEB} &= \text{Cov}\begin{bmatrix} LEB_1 \\ \\ LEB_2 \end{bmatrix} \\ &= \begin{pmatrix} \text{Var}(LEB_1) & \text{Cov}(LEB_2, LEB_1) \\ \\ \text{Cov}(LEB_1, LEB_2) & \text{Var}(LEB_2) \end{pmatrix} \\ &= \begin{pmatrix} \beta_1^2 \Sigma_{MSOA} + w_1^2 \mathbb{I}_n & \beta_1\beta_2 \Sigma_{MSOA} + w_{12}\mathbb{I}_n \\ \\ \beta_1\beta_2 \Sigma_{MSOA} + w_{12}\mathbb{I}_n & \beta_2^2 \Sigma_{MSOA} + w_2^2 \mathbb{I}_n \end{pmatrix}. \end{aligned}$$

□

Lemma A.7.5. *The answer to any predictive problem is a preditive distribution. The predictive distribution in its general form is usually the conditional distribution of the predictive target given the observed data.* $LEB^* = (LEB_1(x_1), \ldots, LEB_1(x_q), LEB_2(x_1), \ldots, LEB_2(x_q))^\top$ *be the predictive target and* LEB *be the vector of observed data, then the predictive distribution is formally expressed as* $[LEB^*|LEB]$

Theorem A.7.6. *Let* $LEB^* = (LEB_1(x_1), \ldots, LEB_1(x_q), LEB_2(x_1), \ldots, LEB_2(x_q))^\top$*; the predictive distribution for* LEB^**, i.e. its conditional distribution given the data, is multivariate Gaussian with mean*

$$\alpha \oplus \mathbb{1}_{q\times 1} + D^\top \Sigma_{LEB}^{-1}(LEB - \alpha \oplus \mathbb{1}_{n\times 1}), \tag{13}$$

and covariance matrix

$$\Sigma_{LEB^*} - D^\top \Sigma_{LEB}^{-1} D. \tag{14}$$

Proof. According the Lemma A.7.5, the predictive distribution is given as $[LEB^*|LEB]$. And using the properties of conditional distribution in Lemma A.7.3, the $[LEB^*|LEB]$

is follows a multivariate Gaussian with mean

$$\alpha \oplus \mathbb{1}_{q\times 1} + D^{\top} \mathit{\Sigma}_{LEB}^{-1}(LEB - \alpha \oplus \mathbb{1}_{n\times 1}), \tag{15}$$

and covariance matrix

$$\mathit{\Sigma}_{LEB^*} - D^{\top} \mathit{\Sigma}_{LEB}^{-1} D, \tag{16}$$

where the (h, h')-th element of $\mathit{\Sigma}_{LEB^*}$ is given by $(\mathit{\Sigma}_{LEB^*})_{hh'} = \tau^2 \exp\{-\|x_h - x_{h'}\|/\delta\}$. Also,

$$D = \begin{pmatrix} D_1 \\ \\ D_2 \end{pmatrix}$$

where D_i is the $n \times q$ matrix whose h-th column is $(d_1(x_h), \ldots, d_n(x_h))$, and $d_j(x_h) = \beta_i^2 \tau^2 \int_{MSOA_j} \exp\{-\|x_h - x\|/\delta\}\ dx$. □

A.8 Map of the Observed IMD and LEB

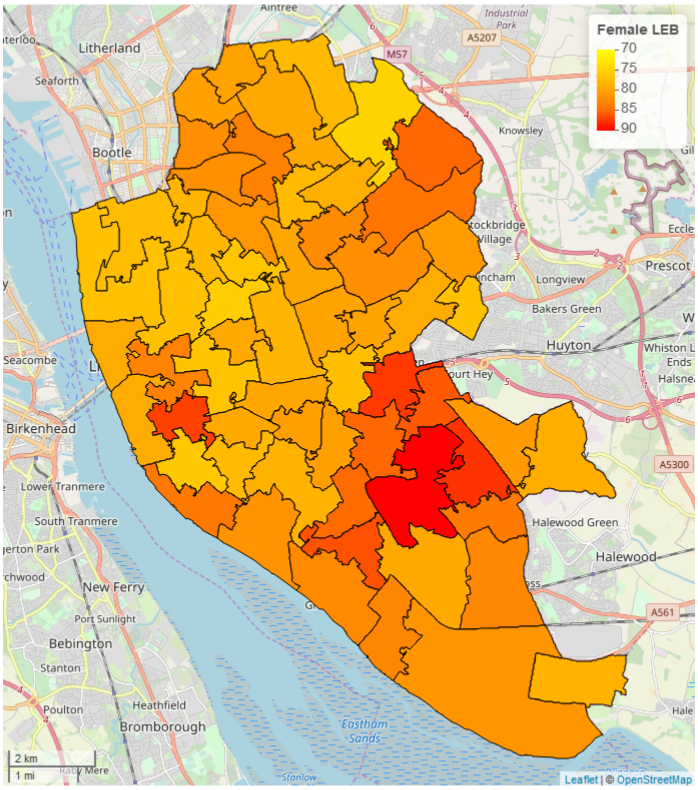


Figure 8: Map of the observed female life expectancy at birth (LEB)

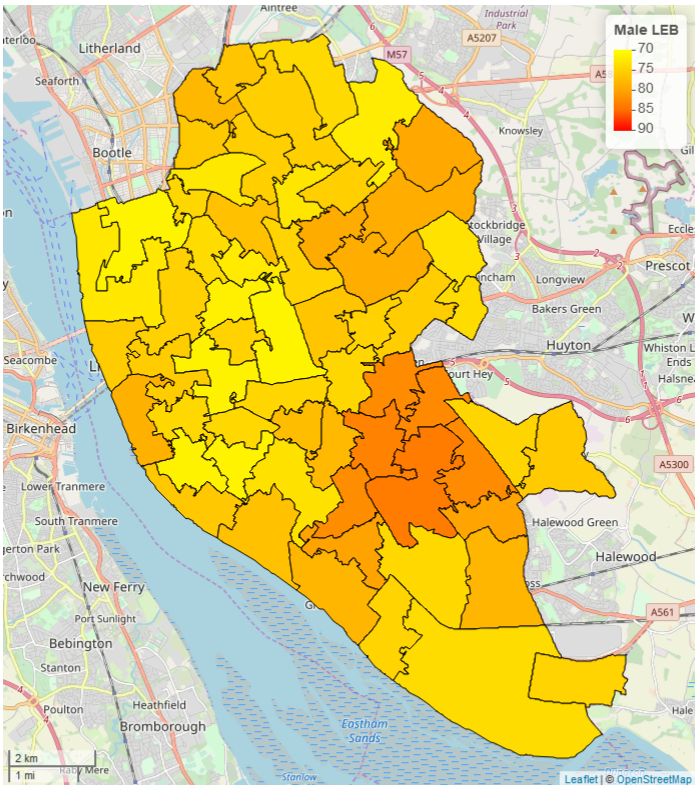


Figure 9: Map of the observed male life expectancy at birth (LEB)

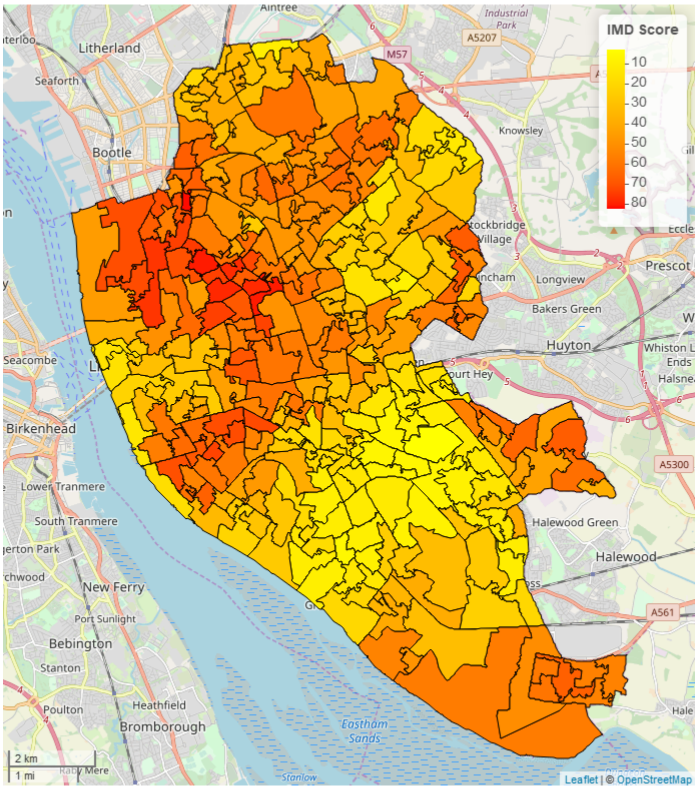


Figure 10: Map of the observed index of multiple deprivation (IMD).

A.9 Map of the Liverpool, MSOA, LSOA and Ward

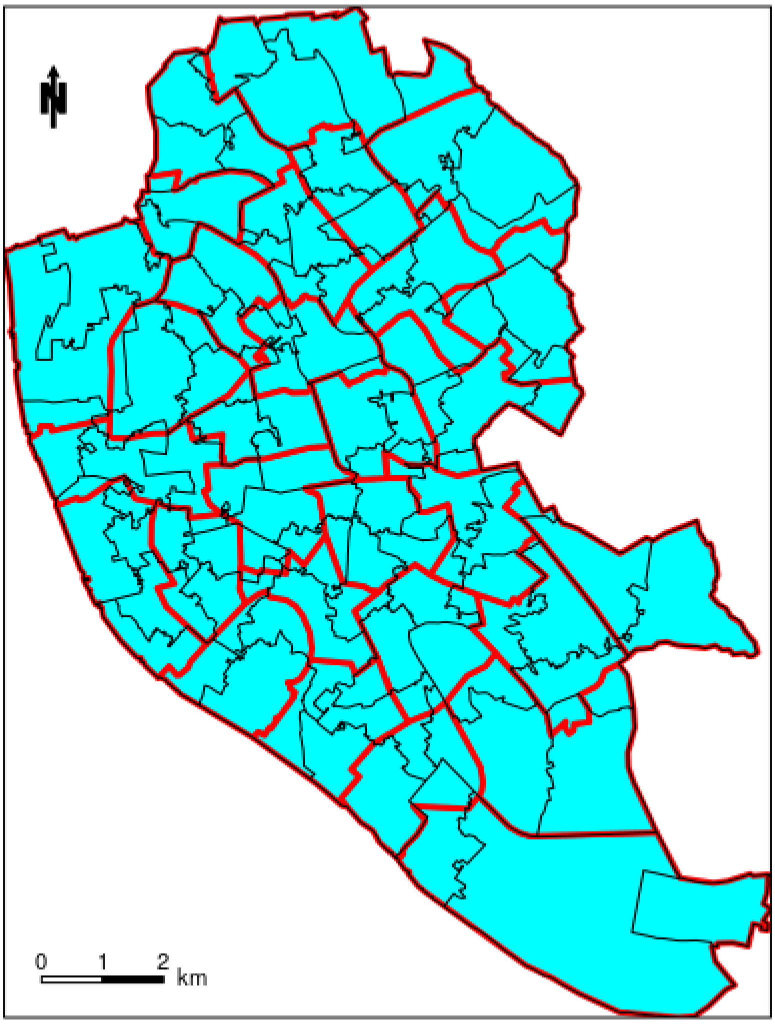


Figure 11: Maps of Liverpool wards with MSOA boundaries overlayed. Red lines are the ward boundaries while black lines are the MSOA boundaries

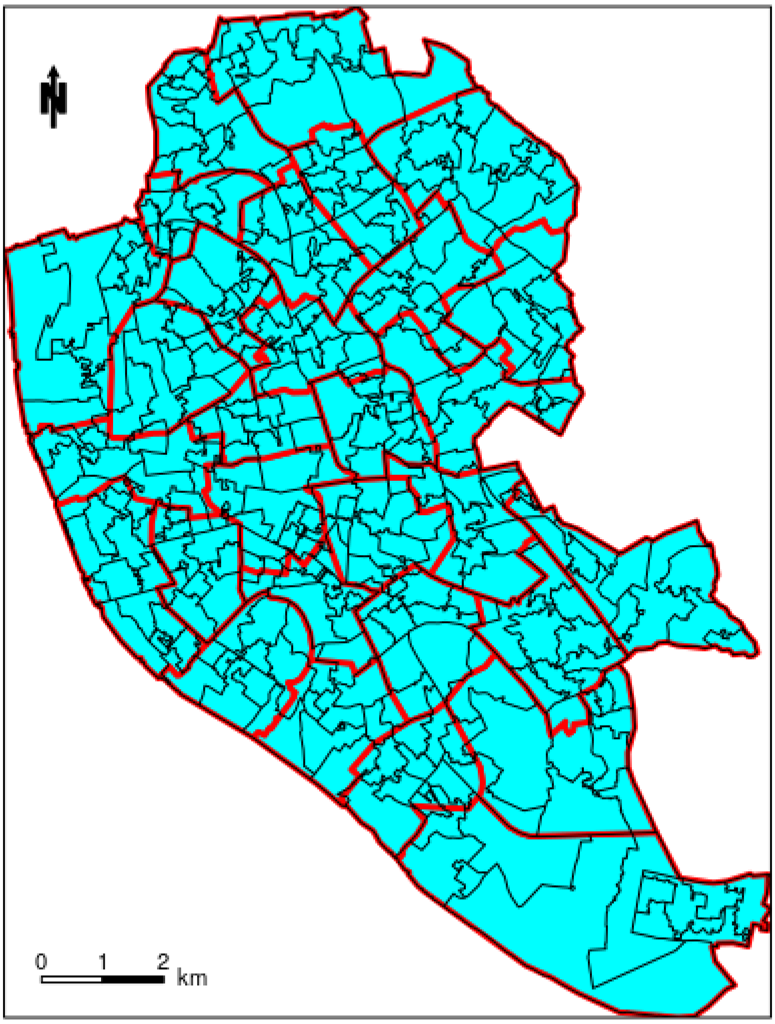


Figure 12: Maps of Liverpool wards with LSOA boundaries overlayed. Red lines are the ward boundaries while black lines are the LSOA boundaries

Bibliography

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–136.