Lancaster University

Space-Time Modelling of Health Data

September 29, 2019

Ziyu Zheng

Supervisors: Dr. Benjamin M. Taylor Mr. Barry Rowlingson

Contents

List of Tables

List of Figures

Abstract

With the increasing availability of spatially referenced health data over long period of times, it urges different methods to detect missing opportunities contained in them can be utlised to learn more about epidemiology of diseases. This thesis presents different methods of spatiotemporal analysis over health data and considers three different datasets. The three main parts of this thesis covers different applications of spatiotemporal analyses on three major health datasets: official tuberculosis data in Portugal, Surveillance, Epidemiology and End Results Program data and National Health Service open prescribing data. All statistical inferences are mainly based on the Bayesian framework including both exact and approximating methods.

The first main contribution uncovered the sub-regional epidemiology of tuberculosis from 2000-2013 in Lisbon and Oporto Metropolitan Areas in Portugal. The analysis provided a Poisson mixed effect model approach on tuberculosis in Portugal at a finer geographical scale compared to past researches. Such model includes both spatial and temporal correlated effects, interaction terms and unstructured random effects. The inference is based on the Integrated Nested Laplace Approximation method. The analysis revealed that having adjusted to known risk factors based on previous research, there are still spatial clusters in both Metropolitan areas with similar but not identical temporal trends. High incidences are more focused around poverty zones which are not necessarily high population density areas.

The second interest of this thesis falls in the area of survival studies. A two-way Bayesian spatiotemporal hazard model is introduced in the second main part of this thesis. This is a novel extension which resolves the ignorance of real calendar time when subjects enter the study by including it as a secondary timescale, on top of random effects. In many previous studies, the main interest focuses on the survival duration of study subjects and the time of entry is often eliminated, especially in hazard models. This is not always a suitable assumption when the study is carried over an extended period of time, e.g. decades. Such two-way spatial survival model knits the problem into a 3-dimensional one where both timescales and spatial random effects are considered at the same time. Even though the application of the model over breast cancer in New Mexico from the Surveillance, Epidemiology and End Results program did not present as satisfying results. Supported by simulation studies, the

two-way spatial survival model does respect the fact that data are accumulated over time and in that sense, treats the data in a more natural way. It should still provide a different and more natural way to approach long-term data and allow comparisons of different behaviours between; for example population-based cancer registries.

The last contribution of this thesis concentrates on the modelling of general practice (GP) open prescribing data. This source of data constitutes a rich time series covering drugs prescribed by all general practitioners across England. The full potential of such a source of data has not been exploited. The ultimate goal for this project is to detect missing opportunities in such open-source data to learn about the epidemiology of a disease. In this part of the thesis, we focused on the pregabalin prescribing data at both GP and Clinical Commissioning Group(CCG)-level across England. The work demonstrated a clear North-South divide in behaviour of pregabalin prescriptions adjusting for spatial effects and deprivation at units of milligrams. This shows that the cause of higher prescribing rate of pregabalin in the North is not just caused by higher deprivation rate. Temporal trend developed differently within each CCG and among GPs, some of which showed an increasing trend.

Acknowledgements

I am very grateful to the Health e-Research Centre for offering me the studentship to study for a PhD here at Lancaster; these years have been very productive and inspiring in my life. I also appreciate all the help from my supervisors, Dr. Ben Taylor and Mr. Barry Rowlingson. They have been offering constant help throughout my whole experience at Lancaster. A combination of encouragement to pursue my own ideas and supports to keep them on track was offered. I am very thankful for their patience and all they have offered. Professor Carla Nunes has also helped a lot for the work I have done in the first year of my PhD.

I would also like to thank my family for always being extremely supportive behind me through the whole journey. My mother, who constantly offers company and gives helpful advices when needed. Other thanks go to my friends and other colleagues who have been with me in different life situations to encourage me. It is more than inspirations on academic knowledge to learn from them.

Declaration and Details of Papers Produced

I declare that no part of this thesis has been submitted in the same form for higher degree elsewhere. The work in Chapter 4 has been submitted to International Journal of Tuberculosis and Lung Disease and is currently under process of reviewing. This is a joint work of myself, Benjamin Taylor, Carla Nunes and Barry Rowlingson, with first author being me. The part of data analysis and model inference are based on my own work but with guidance from Ben, Barry and Carla.

Part of the material in Chapter 5 concerning the linear Gaussian process multi-timescale hazards has been presented in ISBA conference 2016. The work from this chapter is mainly my own work with main contribution goes to model derivation and data analyses. It is also targeted to be submitted to a relevant journal, although not decided yet. Again I have received guidance from Ben and Barry. These analyses are based on the R package *spatsurv* by Benjamin Taylor and Barry Rowlingson.

The material in Chapter 6 concerning NHS GP prescriptions is also submitted to British Medical Journal Open (BMJ Open). This is a joint work by myself, Benjamin Taylor, Barry Rowlingson and Euan Lawson. The analyses are done by myself with guidance received from Ben and Barry. Medical comments and suggestions are provided by Euan.

All computational work was done via different R statistical packages; especially *spatsurv*. I have also contributed to part of the R codes in *spatsurv* .

Chapter 1

Introduction and Overview

1.1 Introduction

It is increasingly common nowadays to find that health data recorded as time series are also spatially referenced. These data are evident in many different areas of health research but the richness of the information contained within this type of data has yet not been exploited to its full potential. Particularly in health research, it is often desirable to be able to identify exposures, behaviours and characteristics of risks of diseases of interest over study subjects, where such behaviours concern not only the temporal trend but also the endemic spatial patterns of these diseases. Hence, alongside time series techniques dealing with the temporal trends, spatial statistical methods can assist understanding of geographical patterns and quantify these features in public health data. More particularly, spatial statistical methods can be used to: 1. evaluate differences in observed rates in different regions; 2. distinguish spatial clusters from noise and 3. identify potential exposures (given relevant data from the same time period and at similar geographical resolutions). Many studies have shown the use of spatial analyses as a suggestive tool in disease patterns (Gelfand et al., 2010; Mayer, 1983), where a review of the growing field of spatial epidemiology studies can be seen in Jerrett et al. (2010). Moreover, when space-time health data is combined with data from other sources, such as population-based socioeconomic data, it can be especially helpful to detect possible risk factors of diseases of interest.

The ultimate goal is to show how statistical methods can assist epidemiological learnings by capitalising on the full range of information offered by the spatiotemporal health data. With general interests mentioned in the previous paragraph in public health over uncovering potentials in spatiotemporal data bearing in mind, applications of three different techniques of space-time modelling are demonstrated in this thesis. The first main contribution of this thesis aims to discover the spatiotemporal pattern in tuberculosis in urban areas in Portugal. It looks at a even lower administrative level than previous studies (normally county-level equivalent) with faster inference done by Integral Nested Laplace Approximation (INLA) approach. One of the benefit is that it examines at a finer scale where patterns within cities can be discovered. The models showed that higher TB incidences are likely to associate with poverty zones rather than high population density areas and time-trend are different between years.

The second main contribution falls in the area of spatiotemporal survival analysis; a two-way spatial hazard function is introduced. One of the traditional approach of survival studies is to use the proportional hazard function (also known as Cox model). It assumes that all individuals in the study share the same baseline hazard. This is not necessarily the most suitable condition in case of long period registries. For example, it is not unreasonable to assume that study subjects who were diagnosed in the earlier ages receive different treatments. The two-way hazard function naturally captures both features in real calendar times and survival times as baseline for study subjects. Based on the original two-way hazard function, spatial random effects are also included. This unique model avoids the choice the baseline timescale and captures both spatial and temporal random effects with handy inference available in R package. Simulation studies supports that the model shows difference when the study is long-run and there is a difference in rate of registrations through time. Application of this model was done on breast cancer in New Mexico.

The last main focus presents one possible way of digging out information from the General Practice (GP) Open Prescribing data. Out of this rich source data, pregabalin prescription rate was looked at due to the general concern of misuse of pregabalin in England. Generalised additive models were employed for 207 individual Clinical Commissioning Groups (CCG) in England. Different from other previous approaches, adjusting for deprivation and GP locations within CCG, such models showed that an obvious divide between the prescription behaviours in North and South still exists. It also in a way showed that utilising open prescribing data can to some extent aid policy making.

Overall, this thesis aims to discover the potential richness of information carried in big health data via different statistical methods. It focuses on three main objectives; 1. The use INLA approach to explain potential underlying spatiotemporal patterns in tuberculosis in Portugal under smaller administrative levels in urban metropolitan areas in Portugal, 2. introduction of a spatial survival model which takes care of the real calendar time on top of time-to-study times and spatial random effects and 3. discovery of the potential North-South divide and temporal trends in pregabalin prescription with deprivation taken into account based on GP Open Prescribing data using spatiotemporal modelling. Three novel Chapters 4 5 and 6 successfully demonstrated how different statistical models can utilise the spatiotemporal health data available to assist either understanding of the disease pattern itself or provide a different routine to deal with health data more naturally. The thesis is more concentrated on the development of application of statistical models over health data rather than epidemiological issues.

In this later part of this chapter, a general background knowledge of spatial data and previous/classic approaches taken to model these data will be given, relevant to this thesis. An overview of the structure of this entire thesis is given later in the chapter.

1.1.1 Background of Spatial Statistics

The earliest well-known spatial analysis was conducted by Dr. John Snow (Snow, 1857) and concerned an outbreak of cholera in the Soho district of London. Dr. John Snow mapped the cholera cases and found clusters around Broad Street. The study eventually led to the conclusion that the water pump on Broad Street was contaminated and caused the outbreak. Not too long after Dr. Snow, Palm (1890) conducted a study of rickets by mapping the geographical distribution of observed cases of rickets and saw high incidence in cold and wet urban areas. A similar type of study based on observations of geographical allocation of skin cancer outbreaks conducted by Blum (1948) also inferred that sunlight might be a causative factor. Both studies agreed on the now known fact that lack of sunshine (or ultraviolet radiation, to be more accurate) can lead to a deficiency in vitamin D which in turn causes rickets. From all these early studies, it can be seen that geographical information for study objects in health research is crucial when one tries to understand the epidemiology of certain disease outbreaks. Such information can also help to reveal potential socioeconomic structures and environmental conditions within a population; especially when socioeconomic covariates are combined

with the disease data itself. Some more sophisticated statistical methods than just mapping diseases in Snow (1857) and Blum (1948) are thus necessary to fully describe the epidemiology of disease. Different types of spatial data together with the most common methods used to deal with them will be briefly introduced in later sections.

Types of Spatial and Spatiotemporal Data with Corresponding Common Methods

Spatial data can be either continuous or discrete and generally are categorized into three types: geostatistics, spatial point processes and lattice data (Waller and Gotway, 2004). Geostatistical data arise over fixed geographical locations in continuous spaces; for instance the prescription data from each General Practice in England. Principles of geostatistical theory were first put forward in the 1960's by Matheron (1962, 1963, 1969, 1971). By the early 1980's, geostatistical theories were widely recognised in a variety of disciplines, such as mining, geology, mathematics and statistics. Such methods take into account the spatial trend and correlation at both large and small scales. Suppose that for each point (x, y) in space D with coordinates x and y in the plane, one can model the value of interest $z(s)$ at location s. One simple model for geostatistical data is $z(s) = \mu + \epsilon(s)$ where the error term can be assumed to have mean 0. It is also sometimes common to assume that $\text{var}[z(s)] = \sigma^2$ for all s in D. If both of these assumptions are satisfied and the covariance of z's at locations s_1 and s_2 only depend on the difference in locations, it is called second-order stationary. Applications of geostatistical modelling have expanded to further areas, including the modelling of soil, rainfall and public health data. However, the type of data which is of the most interest here is health data.

The variogram is a common tool in geostatistics used to quantify spatial correlation instead of using covariance functions (Webster and Oliver, 2001). This approach assumes that the spatial correlation between two given sample points only depends on their relative location (Wackernagel, 2003). The name variogram first appeared in Matheron (1962), but the concept appeared in a few earlier works; for example, Kolmogorov (1941) and Matern (1960). Cressie (1993) described the variogram to be the variance of the difference between observation values at two locations across the realisation of the fields. In this thesis, such method is used in many exploratory analyses to provide informative initial estimates. For example, parameter indicating how far apart the locations should be to be able

to ignore spatial effects. For two given locations s_1 and s_2, the variogram follows the form

$$2\gamma(s_1 - s_2) = \text{Var}(Z(s_1) - Z(s_2)) = \mathbb{E}[((Z(s_1) - \mu(s_1)) - (Z(s_2) - \mu(s_2)))^2].$$

When the spatial random field has a constant mean, Wackernagel (2003) suggested that this correlation is equivalent to

$$2\gamma(s_1 - s_2) = \mathbb{E}[(z(s_1) - z(s_2))^2],$$

for some functions γ. Here, function $\gamma(s_1 - s_2)$ is referred to as semi-variogram. For stationary processes, the semivariogram captures only the difference between the two sample locations.

The graphical representation of the variogram is referred to as a variogram cloud. It is the dissimilarity of two observed values against their Euclidean distances. Normally empirical variograms are used at the early stages of spatial data analysis. Wackernagel (2003) suggested grouping the separation into different vector classes and then averaging the dissimilarities over this class. This can be fitted using parametric variogram models, with parameters which take account of the measurement error (nugget effects), the value $\lim_{|h|\to\infty} \gamma(s_1 - s_2)$ (the sill), and the distance at which $\gamma(s_1 - s_2)$ exceeds the sill for the first time, for instance. There are various models for variograms including the Gaussian model, the spherical model, the exponential model and the power model.

Kriging is an interpolation method which was originally used to predict the gold concentration in ore bodies (Matheron (1963), Krige (1951)). It provides an optimal interpolation which generates the best linear unbiased estimates at each location. This relies on the variogram of the underlying random function $Z(x)$ and assigns more weights to points near to the location of consideration in prediction procedures. The estimates from Kriging methods at each location over a fine enough grid are thus a weighted average of its surrounding observations. Such methods in general provide decent estimates if the locations are uniformly distributed, but give unreliable estimates if locations show clusters with large gaps (Isaaks and Srivastava, 1989). This method can be used to generate predicted values over a space between points to create a continuous surface. Many examples of employment of kriging in health data include Casella et al. (2015), Ali et al. (2006) and Loquin and Dubois (2012). Despite health data being mostly not uniformly distributed (estimates are not very reliable), kriging still acts as an optimal interpolator as they provide the best linear unbiased estimate (Burrough et al., 1998).

Spatial point processes focus more on the locations at which events occur. The key interest is to

detect whether events occur at these locations by random or if there are clusters shown in the spatial data. One way to test whether the event of interest is clustered or not is to test spatial point patterns against complete random processes, where the events are independently and uniformly distributed over the region in study. Often, responses close to each other in location tend to be highly correlated. How events are distributed can be described by an intensity function which is of interest for spatial data. For N events in a square centred at location s, the intensity function $\lambda(s)$ can be expressed as

$$\lim_{A\to\infty} \frac{\mathbb{P}(N>0)}{A},$$

where A is the area of the square. One alternative way to estimating $\lambda(s)$ is to use the K function which is an empirical estimate to test spatial randomness. For instance, the simplest and most commonly used o K-function for homogeneous Poisson process is $k(t) = \pi t^2$, which is also known as complete spatial randomness. Spatial point process methods have diverse applications, for example, as seen here in the spatial distribution of bird nest sets (Ripley, 1981) and forestry over tree locations (Diggle, 1983; Cressie, 1993).

A simple common test for clustering is called the nearest neighbour method (Cressie, 1993). It computes the distance from a point to its nearest neighbour, and the average of all these distances is taken over all points in the study. Normally, one can simulate sufficient data sets to obtain an average nearest neighbour distance for these simulated points. These are then used to determine whether the assumption of complete spatial randomness holds for our data. The common test used is a Monte Carlo test (Cressie, 1993) where the test statistic calculated from observed data is compared with the same statistic calculated from completely randomly simulated data. A Monte Carlo estimated p-value can be obtained from the proportion of simulated test statistic values exceeding the observed test statistic value. It is also common to consider distances to the second nearest neighbour, or even third, fourth and so on. The main difficulty arising from this method is known as edge effects, where the distance between events at the edge is likely to be larger than for those nearer to the centre.

Another popular way to detect clustering is to use a scan statistic, as initially proposed by Kulldorff (1997). It can be thought of as an ellipse of varying size which moves across the study area in a systematic manner. Where the observed value within the ellipse is greater than expected, a cluster exists. The original scan statistic method is proposed for Poisson and Bernoulli models, but can be extended to space-time cases by evaluating each time period along with the neighbouring time period

for each of the spatial circles. The software SatScan^{TM} was developed to deal with a wider range of models including space-time models, multinomial models, ordinal models, Gaussian models and so on. More details can be seen on the bibliography page at `https://www.satscan.org`. Whether the cluster detected is statistically significant is decided by using the p-value calculated in the Monte Carlo hypothesis test.

Lattice data are aggregated over lattice structures where one region is divided into sub-areas (Cressie, 1993) and are typically identified by longitude and latitude. These data can be collected at different spatial resolutions. In studies related to biology, the data collected within each unit are assumed to share the same characteristics of that unit sample. Similarly in health research, it is common to see that as an alternative to geostatistical setting, only positions of the strata relative to each other is used. See Banerjee et al. (2003) for an example of lattice approach over frailty modelling of infant mortality in Minnesota. The key interest for lattice data analysis is to detect spatial patterns over the lattice and explain the patterns linking to covariates within each lattice. A simple way to initially analyse lattice data is to map the rates of an event over the space to visualise whether there are any similar characteristics.

To test whether there is spatial autocorrelation present in discrete-space data, we commonly use Moran's I (Moran, 1950) statistic. Other popular models include conditional autoregressive models (CAR); we assume that $Z(s_i) \sim \text{Poisson}(E_i\theta_i)$ where the expected number of cases in area i is E_i. The parameter θ_i satisfies $\log(\theta_i) = X_i\beta + Y_i$, where X_i are covariates. Y_i's are assumed to follow a conditional Gaussian distribution; ie. $Z_i|Z_j sim\mathcal{N}(\mu_i, \tau_i^2)$ for $i \neq j$. This model is commonly seen in disease mapping, however the relationship seen at aggregated spatial level may be different from that at an individual level.

Spatiotemporal data can be viewed as a temporal extension of spatial data stated above. They are collected over various time points but also contain an extra spatial component that time-series data alone do not have. They represent the temporal evolution of certain events placed over a planar area. In some way, this means that this type of data is 3-dimensional, in that it carries the value of interest, the spatial and the temporal information. In mathematical terms, spatial data can be defined as a realisation of a stochastic process $Y(s) = \{y(s); s \in S\}$, where S is a subset of $\mathbb{R}^d$. Thus the spatial data is a collection of $y(s_1), \ldots, y(s_n)$, where s_i's are indices of the spatial units where it is measured. This is easily extensible to spatiotemporal data, where the data can be defined as a stochastic process

of $Y(s,t) = \{y(s,t); (s,t) \in D \in \mathbb{R}^2 \times \mathbb{R}\}$. The observed values at location s at time t is thus $y(s,t)$.

All types of data and common methods are the necessary base of developing new approaches for health data. For the purpose of this thesis, spatiotemporal health data is of the main interest, where relevant techniques will be further introduced; especially for space-time geostatistical data and lattice data.

1.2 Overview

This thesis presents the interest of exploring and extending the spatial statistical methods on applications of several areas. It performs spatiotemporal analyses mainly focusing on three different datasets: the tuberculosis surveillance system in Portugal (DGS-TB, 1994), population-based cancer registries (SEER (Howlader et al., 2013)) in the US, and Open Prescribing data from NHS England (`https://data.gov.uk/dataset/prescribing-by-gp-practice-presentation-level`). The three data all come from nation-based registries and consists of rich information which aids understanding of disease patterns. These data cover three different aspects of statistical areas regarding health data. The first is disease mapping for infectious disease (TB), followed by survival analysis (SEER) and the last being how Open Prescribing data can present useful information regarding certain drugs (or even relevant disease) of interest. Despite each falls in a different area of health research, they are all national based registries collected over an extensive period of time with rich space-time information. The data were chosen as TB data were readily available at the start of this project following the work Nunes (2008), Areias et al. (2015) and Couceiro et al. (2011). SEER data were chosen because of the initial interest in space-time survival analyses which takes account of real calendar time effect as little work has addressed this previously. Pregabalin from GP Open Prescribing data is done as a follow-up work of Rowlingson et al. (2013).

One common objective of these three analyses is to identify spatiotemporal patterns in the variable of interest, as well as to seek links between the variable itself and socioeconomic factors (potentially gained from other data sources). The first part of the main work in this thesis seeks to explain the spatiotemporal patterns in, and the relationship between, socioeconomic risk factors and tuberculosis incidence at the smallest geographical administrative level in Lisbon and Oporto Metropolitan Areas, Portugal. The second part proposes a new hazard model which incorporates both real calendar

times and survival times. Spatial effects and other covariate effects considering risk factors are also accounted and adjusted for in the model. The last of the analyses describes the spatiotemporal trends in pregabalin prescriptions from the NHS General Practice (GP) prescribing data, together with its relationship with deprivation across England at General Practice level at units of Clinical Commissioning Groups (CCG).

The use of hierarchical models is unavoidable when doing statistical analysis that must consider both space and time. Typically, one can employ random effects when modelling, such as models with a random intercept, models with random slopes and models with both random intercepts and slopes. The main inference method used in this thesis is based on a Bayesian framework, where interest is often in finding expectations for parameter estimates together with their uncertainties. In practice, finding a closed-form solution for these relating posteriors is very rare. General solutions are often simulated via Markov Chain Monte Carlo (MCMC) methods as they provide exact estimates over the parameter of interests. Integrated Nested Laplace Approximation (INLA) is a faster but less flexible alternative to MCMC which, under latent Gaussian model assumptions, gives approximations of relevant posteriors. Chapter 2 gives a brief but comprehensive introduction to Bayesian methods, where Section 2.1 introduces generally the class of latent Gaussian models and particularly the two basic examples relevant to this thesis: the generalised linear model class and the generalised additive model class. A brief explanation of differences in Frequentist and Bayesian paradigms is given in Section 2.2. Section 2.3 provides a background introduction to MCMC and INLA methods with some discussion of the choice of prior. Section 2.4 explains state-space models from the latent Gaussian model family in more detail, including the exact form of inference made via a Kalman filter. This chapter also introduces a novel use of empirical Bayes method to provide appropriate initialisations of the hyperparameters involved in the Kalman Filter via an Expectation-Maximisation algorithm, as given in Section 2.4. Apart from the use of empirical Bayes method to initialise Kalman Filter, this Chapter act as a review of background knowledge on which the applications in later part of this thesis are based.

Survival data also plays an important role in health data; the data normally measures to the time until an event of interest occurs. Such events could be death of a study subject, occurrence of disease, and failure of a machine etc. This thesis covers the interest in knowing how survival behaves in different study subjects starting with an introduction and review of relevant topics in Chapter 3, based on which a novel development of space-time hazard method is introduced (see Chapter

5). Different from most datasets, it is not uncommon that the study subject is not observed at the end-point of the study. For example, an individual may quit the study midway through, or just simply lost to follow-up. In this case, the observed times recorded are called censored data and more details could be found in Section 3.1.1. Survival analysis deals with this type of data, which normal analytical models do not allow. Important functions used in survival studies are the survival and hazard functions; these are explained in more detail in Section 3.1.2. Section 3.2 describes frailty models; a commonly employed way of dealing with spatially referenced survival data. Extensions of survival studies to incorporate both space- and time-varying effects, often by the inclusion of frailty terms, are briefly introduced in Section 3.3. Survival models used in this thesis are likelihood-based, where for observed data conditioning on the latent spatial fields, the likelihood contains contributions from left/right censored, interval censored and uncensored data (see Section 3.4).

The main part of this thesis which shows novel contributions are covered in Chapter 4, 5 and 6. Chapter 4 investigates the dynamic interactions of TB incidence and its socioeconomic risk factors, temporal effects and spatial effects. It introduced an new possibility to use INLA approach on a finer geographical scale to provide fast and robust inference on spatiotemporal modelling on TB incidences. The study seeks to answer the following questions:

1. Is there any seasonality or noticeable temporal trend in TB incidence within the study period?

2. Is there evidence of any spatial pattern in TB incidence? And,

3. What is the relationship between socioeconomic factors and TB incidence?

The data plots show possible increasing temporal and spatial trend for overall incidence. It therefore becomes intuitive to account for spatial random effects as well as temporal effects. Simple exploratory models include linear mixed effect models and generalised linear mixed effect models with both random slope and intercept (see 8.1.1 for details). With extra interest in online updating and dynamic forecasting of TB incidence, a state-space model inferenced via Kalman Filter (KF) was also explored. Details for all preliminary models are available in Section 8.1. An original coded Expectation Maximisation algorithm for estimating initial values for KF using an Empirical Bayes method can be found in Section 8.1.2. Although the Kalman Filter is flexible and able to provide exact estimates of the models, the assumption of linear Gaussian response variable means log-transformation on TB incidence is necessary. This raises problems where the prevalence of TB is zero and accommodations

of such issues are explained in Section 8.1.1. Alongside the issues which may come with the accommodation of zeros, it is still more natural to work with the original non-transformed data; both for interpretation and model fitting. Section 4.1 demonstrates a novel analysis of non-transformed TB incidence data at a finer geographical scale fitted under the Poisson family via INLA approximations. The latent models are under Besag, York and Mollié model (BYM Besag et al. (1991)) and random walk assumptions respectively. Apart from being computationally faster than the Kalman filter, another main advantage of INLA is that it allows us to work with the original incidence data directly and avoid transformations. The best model for describing TB incidence in our study is the one with both spatial and temporal random effects with interactions and unstructured random effects. Areas with higher TB risks are identified, after adjusting for covariate effects, by this model.

Chapter 5 concerns survival times in health data. The classic way of dealing with this is to use proportional hazard model. As traditional survival analyses only take into account the actual event time (by artificially forcing the entry time onto the same origin), the time of entry to study is normally ignored. The event time may therefore not be a fair representation of duration or experience, especially for studies which run over a long period of time. In general, it is possible that patients who enter the study later may receive better treatment for reasons such as advances in medical services. This chapter proposes two models that incorporate both real calendar time and survival time when estimating hazard functions. The first model includes the real calendar time in the hazard function as a Gaussian process (Section 8.2.3). Despite its flexibility, and exact estimates of posterior updates, this model suffers from the major disadvantage of being really computationally expensive. More work is needed to make this model more practical. Thus we introduced a different hazard function which considers multiple timescales together with spatial random effects in one (Section 5.3). This allows a multiplicative contribution from a different timescale alongside the survival time that is normally of interest. It avoids the need to choose an appropriate timescale when modelling hazards. In other words, the multiway hazard tries to explain the risk of the event happening for individuals who enter the study at time τ and survive for time t. These two timescales are interchangeable; ie. knowing survival time and real calendar time at death, it is easy to obtain the entry time. Thus, when picking the response variable, it is possible to choose either survival time or real calendar time basing on which timescale is of more interest for researchers. However, it is important to bear in mind that the likelihood for multiway models can be different to the single timescale model. Thus, covariate coefficient estimates may be different. More details on an application of this proposed model to SEER

data can be seen in 5.3.

Chapter 6 describes the application of spatiotemporal models to GP prescribing data; data which constitutes a rich time series covering drugs prescribed by all general practitioners across England. We aim to use the full potential of these data by connecting the prescription of certain drugs (pregabalin, in our case) to an index of multi-deprivation and to geographical information. This study gives an insight into how such open-source data can potentially reveal more about the epidemiology of certain diseases. Exploratory studies (Section 8.3) used a couple of naive approaches at a finer geographical scale but over fewer areas. The relationship between prescriptions of pregabalin and deprivation is looked at in more detail using sub-categories of deprivation. However, this was replaced by one index in the later study due to intercorrelation between some covariates. Following the identification of possible spatial clusters and temporal trends, the analysis then focuses on spatiotemporal trends in pregabalin prescription at a CCG-level as well as on the relationship with Index of Multi-Deprivation across England (Section 6.1). The study successfully identified geographical divides in prescriptions of pregabalin, adjusting for deprivation, and showed the possible relationship between pregabalin prescription and deprivation. Chapter 7 summarises this thesis as a whole and emphasises the novelty of the main analyses.

Chapter 2

Bayesian Methods in Space-Time Modelling

Bayesian methods support the inclusion of external evidence in a model and combine it with observations from the data. The term 'Bayesian' is attributed to the work of Thomas Bayes, who first proved a special case of Bayes' theorem in the 18th century (Bayes's Portrait, 1988). It was then generalised by Stigler (1986) and applied to fields including medical statistics. Laplace rediscovered the Bayes' principle and suggested that when information apriori is insufficient, the prior distribution should follow a uniform distribution (Hald, 1998). Later in the 20th century, Bayesian statistics diverged into two streams known as the objective Bayesians and the subjective Bayesians. The former assumes a common shared knowledge and inference concentrates only on the assumed models and observed data, while contrastingly the latter allows the inclusion of subjective beliefs and decisions. Due to the recent advance in computing technology, a dramatic growth of applications of Bayesian statistics based on Markov Chain Monte Carlo methods has been observed since the 1980s. These methods allow non-standard and more complex applications; more details will be given in the later part of this thesis.

This chapter provides an overview of Bayesian methods. It acts as a review of statistical methodologies which provides basic grounds or are used in this thesis. Section 2.1 introduces a broad class of structured additive regression models called Latent Gaussian Models; this is the class of models

based on which all models relevant to this thesis are developed. Section 2.2 gives a brief discussion of paradigms of the two main stream of statistics; Frequentist and Bayesian. Section 2.3 introduces the theory behind a collection of popular methods for Bayesian inference; this is also the only inferential method used in this thesis. At the end of this chapter, the state-space model, a pertinent member of the Latent Gaussian model family, will be introduced in more details.

2.1 Latent Gaussian Models (LGM)

Latent Gaussian models (Rue et al., 2009) are characterised by their hierarchical structures. It is an important class of models which covers the most structured Bayesian models, including Bayesian regression models, dynamic models and spatial and spatiotemporal models. The response variable of this model class is assumed to follow some distribution from the exponential family.

This family of models can be described by a conditionally independent likelihood function,

$$\pi(y|x,\theta) = \prod_{i=1}^{n} \pi(y_i|\eta_i(x),\theta),$$

where y is the response variable, x is the latent field and θ is the vector of hyperparameters. $\eta_i(x)$ is the ith linear predictor which connects the data to the latent field in consideration. The latent field is specified as Gaussian such that:

$$x|\theta \sim \mathcal{N}(\mu(\theta), Q^{-1}(\theta)),$$

with prior $\theta \sim \pi(\theta)$. The structured additive predictor η_i takes account of covariate effects in an additive way such that for $x = (\beta, \alpha, f, \epsilon)$,

$$\eta_i = \alpha + \sum_{k=1}^{\eta_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \epsilon_i,$$

where $f^{(j)}$'s are unknown functions of covariates which need relaxation of a supposed linear relationship. This is commonly used when modelling temporal and/or spatial dependence. β_k's represent the linear effect of covariates z and the ϵ's are unstructured terms. Note that this thesis will be using η as general notation for linear predictors here and in the sections that follow.

Popular models belonging to this class include dynamic linear models (DLM), generalised linear (mixed) models (GL(M)M), generalised additive (mixed) models (GA(M)M), spatiotemporal models and survival models, to name a few. More details and applications of some of these models will be described in later chapters and sections where relevant.

2.1.1 Generalised Linear Models and its Extensions

Generalised linear models (GLM) (Nelder and Wedderburn, 1972) are a more flexible class of model than ordinary linear models in that they allow response variables to have non-Gaussian error distributions. In general, for independent response variable y in the exponential family, it can be written that

$$g(\mathbb{E}(Y)) = X\beta = \eta,$$

where $\mathbb{E}(Y)$ is the expected value of Y and the covariate effects $X\beta$ link to it via a link function g. The variance under this setting is typically written as a function of the mean, where $\mathbb{V}ar(Y) = \mathbb{V}ar(\mathbb{E}(Y))$. Under Bayesian settings, the posterior distribution of a GLM does not have a closed form, which means that the inference is carried out either by MCMC or approximation methods such as INLA. More details can be seen in Section 2.3.3.

Generalised Linear Mixed Models (GLMM)

The data considered in this thesis are mostly longitudinal with clustered designs; they contain both time information and location information and are likely to show patterns within neighbouring areas. For example, diseases are notified through time with location of occurrence and are likely to show certain correlation within neighbourhoods. It is therefore useful to have a model class that allows a unified likelihood for parametric regression which accommodates overdispersion and correlation (Breslow and Clayton, 1993). Both fixed and random effects are considered for GLMMs and are included in the following form:

$$g(\mathbb{E}[y|u]) = X\beta + Zu = \eta,$$

where all notations are as for GLMs, with the addition of u being a random effect. Random effects are assumed to follow a Gaussian distribution with mean 0 and some variance. It is expressed with

the fixed effects via a linear predictor η; ie. $\eta = X\beta + Zu$. GLMMs are especially useful as an extension when modelling geographically referenced disease rate, where one can take into account of spatial random effects by an appropriate definition of u.

Generalised Additive (Mixed) Models (GA(M)M)

When the linear predictor is sufficiently generalised to allow an additive functional form of x's rather than having the restriction of being linear, one has more flexibility when modelling data. Hastie and Tibshirani (1990) introduced a non-parametric extension to traditional GLMs, the generalised additive model (GAM), by allowing a link function between its covariates and the expected response variables. In formula, it can be written as

$$g(\mathbb{E}[y]A) = \beta_0 + f_1(x_1) + \cdots + f_m(x_m) = \eta,$$

where f_i's are the smoothing functions and everything else remains the same as in the GLM notation above. GAM models capture non-linearities in the data via the smoothing functions f in forms of splines, polynomials or step functions to name a few. It therefore makes inference about these smooth functions. This allows an extra underlying random variation on top of the Gaussian distribution. Here, the model consists of a random component (Y) and an additive component, linked by smooth functions f_i.

One of the key features of GLMMs is that the mean function is parametric; this is not necessarily desirable in certain situations where functional forms of covariates are not already known. Therefore, a model that is like GLMMs but is more flexible in that it allows non-parametric regression becomes desirable. Lin and Zhang (1999) introduced the additive extension to GLMMs, where the covariate effects are modelled by additive non-parametric functions and random effects are included with the additive predictor to account for both overdispersion and correlation. These are known as generalised additive mixed models (GAMM). Using the notation of the previous paragraph, GAMM can be written in the following form:

$$g(\mathbb{E}[y|u]) = \beta_0 + f_1(x_1) + \cdots + f_m(x_m) + Zu = \eta,$$

where random effects are assumed to follow a Gaussian distribution with mean 0 and some known

variance. This model is commonly used in clustered and hierarchical studies. In spatial studies, GAMMs account for spatial autocorrelation by including a smoothed spatial location term in the models.

2.2 The Frequentist and Bayesian Paradigms

A statistical model generally considers the data-generating process in an idealised form. Suppose that the observed data set is $y_1, \ldots, y_n$ and the corresponding parameters are $\theta_1, \ldots, \theta_d$ (assuming $d \leq n$). The relationship between the observations and parameters is described by the modelling process, denoted as $y_i = f(\theta_1, ...\theta_d, \varepsilon_i)$. Here the ε_i's are assumed to follow a known distribution. A likelihood function $\pi(y|\theta_1, \ldots, \theta_d)$ in some sense captures the probability of making an observation y under a known set of parameters. But choices of how to make inference about these data using the model diverges statisticians into two main schools of thought; the frequentists and the Bayesians. More details about these two areas can be found in book Bayesian and Frequentist Regression Methods (Wakefield, 2013).

The frequentists take a probability approach that considers the frequency of occurrence of an event outcome for some random event repeated many times. In other words, the frequentist approach to statistical inference is purely based on the observed data; it seeks to estimate parameter values by maximising the likelihood function, producing estimates known to be the maximum likelihood estimates. The variances of parameters are then estimated according to either the exact or the approximated distribution theory. In frequentist inference, the parameters are thus fixed but unknown and the data is treated as the variable.

As an alternative to the frequentist approach, the Bayesian approach looks at the probability of some event occurring with prior beliefs taken into account. It treats parameters of interest as random variables. Edwards et al. (1963) described the idea of Bayesian statistics as 'the revision of opinion in the light of relevant new information'. Similar to frequentist statistics, what is called a marginal likelihood function $\pi(y)$ describes the distribution of observations y over all parameters θ for Bayesians. It is also known as the model evidence in the Bayesian context. But unlike the frequentist methods which provide an estimate for the parameter of interest with standard errors, the main outcome of interest for Bayesian is the posterior distribution. This can be used to simulate samples and make es-

timates of quantities of interest. One big difference on parameter estimation between the Frequentists and the Bayesians is that the Bayesians make probability statements about the parameters whereas the frequentist cannot.

2.3 Bayesian Inference

Bayesian inference is a statistical method developed from work related to Bayes' theorem. Treating parameters and sample data each as random quantities, all inference made using a Bayesian approach is achieved by probability calculations. The statistical analysis of data therefore requires a stochastic model which describes how we believe the data arose, however, this also means that the inference requires both a likelihood function for the data and a prior distribution for the parameters. In this thesis the main inference method used is Bayesian, thus, a very brief introducing how Bayesian inference works is given here. More details can be found in many textbooks; see Bolstad and Curran (2017), Box and Tiao (2011) and Gamerman and Lopes (2006) for example.

The fundamental rule for Bayesian inference is known as the Bayes' rule or the Bayes' theorem. It established that subjective beliefs should rationally change to account for pre-known evidence. For a given set of observations $y = (y_1, \ldots, y_n)$, the posterior probability of random parameters $\theta_1, \ldots, \theta_d$ conditional on the observations is given by:

$$\pi(\theta_1, \ldots, \theta_d | y_1, \ldots, y_n) = \frac{\pi(y_1, \ldots, y_n | \theta_1, \ldots, \theta_d)\pi(\theta_1 \ldots, \theta_d)}{\pi(y_1, \ldots, y_n)}.$$

Here, $\pi(\theta)$ is the prior probability, $\pi(y)$ refers to the marginal likelihood and $\pi(y|\theta)$ is the probability of observing data $y = (y_1, \ldots, y_n)$ given the parameters. Informally, Bayes' theorem can be considered as the equation which links the odds of degree of belief before and after accounting for observed evidence.

The foundation of Bayesian Inference rests on two probability densities: the prior density and the posterior density. A prior probability distribution $\pi(\theta)$ captures one's subjective beliefs about parameters θ before considering any evidence. The posterior probability $\pi(\theta|y)$ is the probability of the parameters having accounted for the evidence provided by the observed data y.

2.3.1 Choice of Prior

One of the major difference between frequentist and Bayesian is that the latter involves 'prior' information when considering chances. By Bayes rule, one can see that likelihood functions for the data and prior distributions for the parameters are both crucial to Bayesian methodology. However, its requirement of using previous knowledge to set prior distributions inherently involves a degree of subjectivity, thus how to choose priors has always been the most criticised element of Bayesian methodology, as the choice may greatly affect the final inference. In this subsection here, a brief general background of how priors are chosen is given. Details on specific priors which are used in different models in this thesis later are given where relevant.

There are a number of ways to specify a prior distribution in Bayesian analysis (Carlin and Louis, 2008). For instance, a prior could be determined from received knowledge such as past literature or previous experiments, or alternatively, it could be specified by experts in the relevant field. Priors chosen in these ways are known as informative priors and are used to express specific information about certain random variables.

Objective and Subjective Bayesian

The objective and subjective variants of Bayesian probability differ mainly in their interpretation and construction of the prior probability. As the names suggest, subjective Bayesian inference tries to make a subjective interpretation of probability. It emphasises the relative lack of rational constraints on prior probabilities and corresponds to some 'personal belief' (Goldstein, 2006). Objective Bayesians (Torsen, 2015) have a formal advantage arising from the structural clarity of prior distributions. In other words, it is based on a reasonable belief that everyone shares the same knowledge following Bayesian rules. The subjective Bayesian approach uses subjective opinions and can be purely based on individuals who are conducting the inference process. The objective Bayesian approach tend to choose non-informative prior so that the contribution of information from priors to likelihood is not changed. The posterior distribution thus is only based on the likelihood information.

There are two types of priors: informative and non-informative. A non-informative prior is one which provides little or no information relative to the experiment data itself (Box and Tiao, 2011). In other words, when the prior distribution is 'flat' compared to the likelihood function and thus has minimal

impact on the posterior distribution, it is non-informative. Informative priors on the other hand, as mentioned in the previous paragraph, provide information about the parameters from other sources and dominate the likelihood. They are normally case-specific. Appropriate use of informative priors supports the Bayesian concept of combining current information (the data) with previously known information. In general, it has been argued that the use of any prior is already subjective and that it is potentially possible to design the prior in such a way that the model achieves the desired results. Non-informative priors are therefore popular in many applications for their relative objectivity, but also for cases where previous information is unavailable or unclear.

For the purpose of this thesis, objective Bayesian approach with non-informative priors were used where possible. Where informative priors had to be used in order to improve model fit and mixing in MCMC are discussed in the Appendix.

Jeffreys Prior

One common choice of non-informative prior is the uniform prior, which assigns equal likelihood over all possible parameter values. Another very useful prior is the Jeffreys prior (Jeffreys, 1946), which assumes uniformity across certain intervals. It is proportional to the determinant of Fisher's information matrix and follows $\pi(\theta) \propto |I(\theta)|^{1/2}$. Due to its local uniformity, Jeffreys prior is non-informative but it provides a system to define a prior for many parametric models. However, Jeffreys prior can lead to an improper posterior in some cases and it may be hard to use in higher dimensional models (which is the case for data considered in this thesis). It is not directly used in this thesis but only introduced for the purpose of providing background knowledge as part of Bayesian inference.

Conjugate Prior

For calculation simplicity, it is often desirable to use what is called a conjugate prior (Raiffa and Schlaifer, 1961). In Bayesian statistics, if the posterior distribution $\pi(\theta|y)$ are in the same family as the prior distribution $\pi(\theta)$, we say this prior is a conjugate prior. A common conjugate prior is one which has a Gaussian distribution; if the likelihood function is Gaussian, a Gaussian prior over the mean leads also to a Gaussian posterior distribution. Some other useful conjugate families in medical statistics include a Gamma prior distribution for rate when the likelihood is Poisson, a multinomial

likelihood with Dirichlet prior for parameters and a multivariate Normal likelihood with multivariate Normal prior. Details of more conjugate families can be found in Fink, D. (1997). Relevant to this thesis, this prior is not directly used. However, it is one of the most common prior choices which may be employed in R packages when doing reference.

Other Prior Choices

Priors can alternatively be chosen using more complex principles. For instance, Jaynes (2003) introduced a method which maximises the Shannon entropy (Shannon, 2005), that is, the expected value of information contained in the probability distribution. The larger the entropy, the less information is provided by the prior distribution itself. Thus such maximisation of the value allows one to set the least informative prior distribution possible. A further idea for prior choice is that known as reference priors (Bernardo and Berger, 1989), where priors are chosen to maximise the expected Kullback-Leibler divergence (Kullback and Leibler, 1951) of the corresponding posterior distribution. The Kullback-Leibler divergence measures the information lost when one approximates the posterior distribution with the prior probability. Thus, the use of reference priors allows the data to have maximum effect on the posterior estimates. These are not directly used in analyses in this thesis, but are also a main components of prior studies in Bayesian statistics; thus included here for completeness.

2.3.2 Markov Chain Monte Carlo (MCMC)

A brief introduction of MCMC is given in this subsection as background knowledge required in this thesis, more details can be found in the following books Handbook of Markov Chain Monte Carlo (Brooks et al., 2017) and Markov Chain Monte Carlo in Practice (Gilks et al., 1995). MCMC acts as a key in Bayesian statistics. For the purpose of this thesis, inferential methods are mostly based on MCMC.

As explained earlier, Bayesian methodology aims to draw inference about the parameters of a posterior distribution determined using Bayes theorem, $\pi(\theta|y_{1:n}) \propto \pi(\theta)\pi(y_{1:n}|\theta)$; inference that is made by evaluating the posterior expectations. The most usrful parameter is the posterior mean:

$$\mathbb{E}_{\pi(\theta_1,\ldots,\theta_d|y_1,\ldots,y_n)}[\Theta_1,\ldots,\Theta_d)].$$

This expectation tends towards the same limit as the maximum likelihood estimate as the sample size tends to infinity, provided that the prior information does not play a dominant role and that the statistical model is correct under regularity conditions. However, the simplest analytic computation of such an estimate is often unavailable in reality. As Bayesian methods will frequently require the calculation of expectations over the posterior, the unavailability of these integrals will always cause problems.

The development of Markov chain Monte Carlo (MCMC) methods has revolutionised Bayesian statistics; it allows the computation of large hierarchical models and made Bayesian methods possible in complex models. MCMC methods provide a way of simulating random variables from a Markov chain with stationary density being the same as the posterior of interest. Therefore, for some parameters θ with observations y, MCMC produces the opportunity to solve the inference of the posterior $\pi(\theta|y)$, the marginal distribution $\pi(\theta_i, y)$ and the forecast distribution $\pi(\tilde{y}|y)\}$ for some $\theta_i \in \theta$ and $\tilde{y}$ predictive observations.

MCMC methods have their basis in the theory of Markov chains. The stochastic process θ is called a Markov chain if it satisfies the memoryless property:

$$\pi(\theta^{(i)}|\theta^{(i-1)}, \theta^{(i-2)}, \ldots, \theta^1) = \pi(\theta^{(i)}|\theta^{(i-1)}).$$

This means that the current state of the chain only depends on the state of the chain at the previous time. Markov chains can be defined over either discrete or continuous probability spaces. For discrete cases, the probability of transitions between states is captured by stochastic transition matrices. When it is defined over a continuous probability space, a Markov transition kernel $\mathbb{P}(\theta_t|\theta_{(t-1)})$ explains the transition probability from state $t-1$ to t. When the Markov chain is aperiodic and irreducible, that is the chain can move from any given state to any other state in finite number of steps and has period 1, the transition kernel converges to its stationary distribution. Such Markov chains also have the ergodic average $\bar{g} = \frac{1}{n}\sum_{t=1}^{n} g(\Theta_t) \to \mathbb{E}_\pi g(\Theta)$. These two properties (aperiodicity and irreducibility) provide the basis for MCMC methods; 1. to sample from distribution $\pi(\theta)$ by simulating from a Markov chain with stationary distribution π, and 2. to estimate functions of density π by using the ergodic average of the chain.

MCMC has shown an explosive growth in popularity since the early 1990's. Such methods are a class

of computational algorithms which aim to estimate the expectation of a statistic in a complex model via simulations. The general idea of MCMC is to construct a Markov chain, simulate n steps of the chain and use these samples to make an estimate. More precisely, MCMC generates samples using a Markov chain mechanism constructed so that the samples generated mimic the real samples from the target distribution $\pi(\theta)$. The quality of samples improves as the number of MCMC steps increases.

Some variations on the basic idea can be used, for example, running multiple chains and introducing auxiliary variables (Damien et al., 1999). In the auxiliary method, one can generate realisations from a complex distribution with density $\pi(\theta)$ by augmenting θ with additional variables u, say. The general auxiliary method is described in the following steps:

- Specify u and the conditional distribution $\pi(u|\theta)$.
- Compute the joint distribution $\pi(\theta, u) = \pi(\theta)\pi(u|\theta)$.
- Define transition kernels P_u and P_θ to update u and θ such that both of them maintain $\pi(\theta, u)$. Typically, u is updated with a Gibbs step.

The following subsections introduce two fundamental algorithms in MCMC; Metropolis-Hastings algorithm and Gibbs Sampling. An approximation method which is developed in more recent year, Integrated Nested Laplace Approximation (INLA), is introduced later. For the purpose of this thesis, the main inference method relevant to MCMC used in Chapter 4 is INLA (using available package in R); with exploratory analysis mainly done based on a mixture of Metropolis-Hastings and Gibbs Samplings. Chapter 5 proceeds inferences based on Metropolis-Hastings sampling scheme. Details of each individual R packages can be found in the vignette documents available on Comprehensive R Archive Network (CRAN).

Metropolis-Hastings algorithm

The first MCMC algorithm was attributed to Metropolis (Metropolis et al., 1953), upon which Hastings (Hastings, 1970) generalised the method into what is known as the Metropolis-Hastings (MH) algorithm. This algorithm explains how a transition kernel $Q(\theta, \theta')$ can be constructed such that it allows one to sample from a Markov chain $\{\Theta_1, \ldots, \Theta_n\}$ with stationary distribution $\pi(\theta)$. The algorithm has the following steps:

- Initialise $\Theta_0 = \theta_0$.
- At time t, we have $\Theta_t = \theta$. Propose a candidate value θ' from proposal distribution $q(\theta, \theta')$.
- Calculate the acceptance probability $\alpha(\theta, \theta')$:

$$\alpha(\theta, \theta') = \min\ (1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}).$$

- Accept the proposal with probability $\alpha(\theta, \theta')$.

Gibbs Sampler

The Gibbs sampler is introduced in Geman and Geman (1984). It can be treated as a special case of the MH algorithm where the acceptance probability for proposed candidate y is 1. The basic idea of Gibbs sampler is that for multidimensional θ, we sample from the full conditional distributions $\pi(\theta_i|\theta_{-i})$ in turn for each i. It is often possible to calculate these full conditionals, even when the overall distribution is not tractable. It shows the local characteristics of the full distribution in the direction of θ_i given all other variables. The algorithm follows:

- Initialise vector $\theta^{(0)}$.
- given $\theta^{(m)} = (\theta_1^{(m)}, \ldots, \theta_d^{(m)})$, $\theta^{(m+1)}$ can be generated as:

$$\theta_1^{(m+1)} \sim \pi(\theta_1|\theta_2^{(m)}, \theta_d^{(m)}).$$

$$\theta_2^{(m+1)} \sim \pi(\theta_2|\theta_1^{(m+1)}, \theta_d^{(m)}).$$

$$\vdots$$

$$\theta_d^{(m+1)} \sim \pi(\theta_d|\theta_1^{(m+1)}, \theta_d^{(m+1)}).$$

- let $m = m + 1$, return to step 2 until convergence to $\pi(\theta)$.

2.3.3 Integrated Nested Laplace Approximation

A key obstacle in Bayesian statistics is to actually conduct Bayesian inference in practice; the mathematics may seem straightforward but problems do exist. For example, MCMC may experience high computational costs and suffer convergence problems when implemented over some latent Gaussian models with a non-Gaussian response variable but latent Gaussian fields. To avoid such obstacles, a newly developed method, Integrated Nested Laplace Approximation (INLA), provides a faster alternative to simulation-based MCMC schemes within the class of latent Gaussian models (LGM) by approximating the posterior marginals. In brief, the class of LGM can be represented by a hierarchical structure where response variables have conditionally independent likelihood functions over a latent Gaussian field, and hyperparameters describing the latent Gaussian field have certain prior distributions. More details on notions of LGM will be explained in a later section.

What is INLA?

For a given latent Gaussian field x, the method to approximate its posterior marginals $\pi(x_i|y)$ follows three steps: 1. compute the Laplace approximation to the posterior marginal of parameter θ ($\pi(\theta)$); 2. find an approximation of $\pi(x_i|y,\theta)$ for selected values of θ; 3. numerical integration combining results from step 1 and 2 to find $\pi(x_i|y)$. In this section, approximate densities are denoted as $\tilde{\pi}$. A more detailed introduction of steps are given below but more details can be found in Rue et al. (2009) and Rue et al. (2017).

It is crucial to be able to select sufficiently good evaluation points when approximating the posterior marginals of parameters ($\tilde{\pi}(\theta|y)$) as such an approximation integrates out the uncertainty when approximating the posterior marginals of the Gaussian latent field x. Due to high computational costs in this process, an interpolant to $\log\tilde{\pi}(\theta|y)$ is used instead of direct numerical integration of $\tilde{\pi}(\theta|y)$. This can be done in the following steps:

- Obtain the mode θ^* of $\tilde{\pi}(\theta|y)$.
- Compute the negative Hessian matrix H at θ^*; such that for eigen-decomposition $V\Lambda V^T$ of H^{-1}, we can define $\theta(z) = \theta^* + V\Lambda^{1/2}z$, where z is $\mathcal{N}(0, I)$ for any Gaussian $\tilde{\pi}(\theta|y)$.
- Compute $\log\tilde{\pi}(\theta|y)$ using the reparametrisation defined via z.

Conditioning on the set of $\{\theta_k\}$'s obtained above, one can then approximate the posterior marginals for x_is $\tilde{\pi}(x_i|y,\theta_k)$. This can be done via Laplace approximation

$$\tilde{\pi}(x_i|\theta,y) \propto \left.\frac{\pi(x,\theta,y)}{\tilde{\pi}_{\mathrm{G}}(x_{-i}|x_i,\theta,y)}\right|_{x_{-i}=x*_{-i}(x_i,\theta)},$$

where $\tilde{\pi}_{\mathrm{G}}$ refers to the Gaussian approximation of $x_{-i}|x_i,\theta,y$ and x^* is the mode estimate.

To avoid computation of each of x_i and θ, it is proposed to approximate $x^*_{-i}(x_i,\theta)$ by $\mathbb{E}_{\tilde{\pi}_G}(x_{-i}|x_i)$. This is readily available from the Gaussian approximation during exploration of $\tilde{\pi}(\theta|y)$. In spatial cases, based on an intuitive decision that only x_j's which are in neighbourhood of x_i's should contribute to the marginal distribution of x_i, $\mathbb{E}_{\tilde{\pi}_G}(x_{-i}|x_i)$ can then imply that

$$\frac{\mathbb{E}_{\tilde{\pi}_G}(x_j|x_i)-\mu_j(\theta)}{\sigma_j(\theta)} = a_i j(\theta)\frac{x_i-\mu_i(\theta)}{\sigma_i(\theta)}$$

for some $a_{ij}(\theta)$ such that $i \neq j$. Thus $R_i(\theta)$ can be defined as collection of j's such that $a_{ij}(\theta) > 0.001$ and this can be used to simplify the calculation of $\tilde{\pi}_{\mathrm{G}}(x_{-i}|x_i,\theta,y)$. This saves the computational cost of finding densities for each point x_i and the selection of these points are based on the mean and variance of the Gaussian approximation, say $x_i^{(s)} = \frac{x_i-\mu_i(\theta)}{\sigma_i(\theta)}$. By Gauss-Hermite quadrature rule, the Laplace approximated density follow $\tilde{\pi}_{\mathrm{LA}}(x_i|\theta,y) \propto \mathcal{N}\{x_i;\mu_i(\theta),\sigma^2(\theta)\}\exp\{\text{cubic spline}\}$, where the cubic spline is fitted for $|\log(\pi_{\mathrm{LA}}(\tilde{x_i}|\theta,y)) - \log(\tilde{\pi}_{\mathrm{G}}(x_i|\theta_i,y))|$.

For purely computational benefits, a simplified Laplace Approximation is also derived in **?** especially for spatial cases. Tne simplified Laplace approximation performs expansions of $\tilde{\pi}(x_i|y,\theta_k)$ at mean $x_i = \mu_i(\theta)$ up to the third order. By fitting a skewed-normal distribution, it corrects skewness and location errors occurred in the process of Gaussian approximation.

Why do we use INLA?

With increasing popularity in Bayesian hierarchical models for complex data, general model fittings by via simulation based methods (eg. MCMC) are likely to be computationally expensive. By applying a combination of approximation and numerical integration, INLA bypasses the convergence issues occurring with MCMC methods. INLA typically delivers faster inference and allows estimation of hyperparameters which are challenging tasks for MCMC sampling. However, these are almost

guaranteed to be associated with biases coming from errors introduced by analytic approximations when calculating posterior probabilities. Although INLA provides estimates for hyperparameters, it should be borne in mind that the identification of hyperparmeters is actually a challenging task itself and that quick inference does not necessarily mean that it is correct inference (Taylor and Diggle, 2014). Despite the potential downsides mentioned above for INLA, such approximations together with R-INLA (`http://www.r-inla.org/`) certainly provide a routine toolbox when dealing with complex data.

R-INLA package

In this thesis, the main application of INLA will be done by the R-INLA package, which provides a new approach to statistical inference for latent Gaussian Markov random field (GMRF) models. This is a readily available package allowing fast inference based on INLA. Details are described in Rue et al. (2009). Briefly speaking, for observed values y, latent parameters η and some other parameters θ, R-INLA supports hierarchical GMRF models with the following form:

$$y_j|\eta_j, \theta_1 \sim \pi(y_j|\eta_j, \theta_1),\ j \in J$$

$$\eta_i = \alpha + \sum_{k=0}^{n_f-1} w_{ki} f_k(c_{ki}) + z_i^T \beta + \epsilon_i,\ i \in I.$$

The priors for hyperparameters are assumed to have distribution $\pi(\theta)$.

Here, as not all latent parameters have to be learnt through the data, J is thus a subset of I. It is also assumed that y is conditionally independent of the parameters and latent variables which contribute to the likelihood of observations through some known link functions. It is also assumed that unstructured random effects ϵ are independent and identically distributed with $\mathcal{N}(0, \lambda_\eta I)$, where λ_η denotes the precision. Offsets and weights for each data point are normally known and are included in the linear predictor fitting part. Covariate effects which are nonlinear and/or continuous are captured in $f_k(c_{ki})$, where c_{ki} denotes the covariate value for covariate k at observation i. $f_k|\theta_{fk}$ follows GMRFs $\mathcal{N}(0, Q_k^{-1})$ for some parameters θ_{fk}. $z_i^T \beta_i$ denotes the linear covariate effects for covariate values z_i with coefficient β; βs are assumed to follow Gaussian distributions with mean zero and some fixed precisions. Thus, the full latent field is then $x = (\eta^T, f_0^T, \ldots, f_{n_f-1}^T, \beta^T)$, which is also a GMRF. Well known models which fall into this category include time series models, generalised

additive models (GAM), generalised additive mixed models for longitudinal data, geoadditive models, ANOVA type interactive models and univariate stochastic volatility models.

2.4 State-Space Models

Under Bayesian methodology, one can update their beliefs about the set of parameters when given new observed data. Due to computational memory constraints in modern applications, such analysis is not always possible unless the data are split into smaller blocks. This type of model is known as the state-space form model; they are characterised by state and observation processes. The state process θ_t at time t captures what is thought to be the true state of the model and is not directly observable. How the parameters of interest evolve over time are governed by these hidden state processes. The observation process at time t arises from the hidden states θ_t; these observations are conditioned on the current state only. A visual demonstration of this idea can be seen in Figure 2.1, where each observation Y_t arise from its state θ_t and each state θ_t only depends on θ_{t-1}.

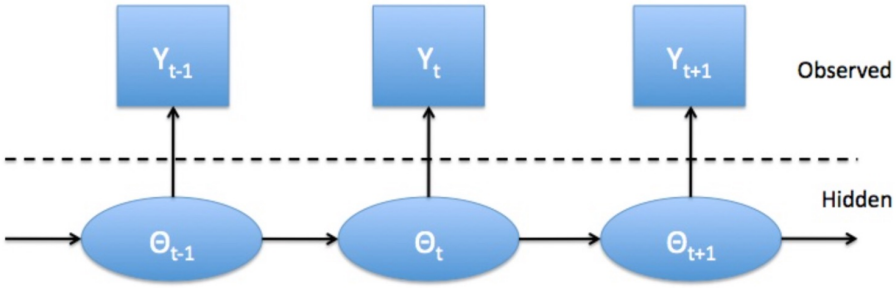


Figure 2.1: Diagram of the Conditional Independence Properties

When the state-space model of interest follows a linear-Gaussian form, the exact form of inference can be made (Kalman, 1960; West and Harrison, 1997).

The use of state-space models allows flexibility over times. The missingness and time-varying covariates in data can be handled very naturally as it adjusts freely for the time-varying dynamic states. These are common problems in general when it comes to data. In the context of health data, it incorporates observations from slightly different sources in estimation without the need to do anything special. For example, re-run the analysis from the start of the data collection period again (thanks to the memoryless property in states). In our application to tuberculosis discussed later, it is shown how

a state-space model allows changes in random slope or intercepts freely. One of the main drawbacks of state-space models is that the identification of the real models can be difficult and computationally expensive.

What are state-space models?

Denoting the state by θ and observations by Y one can assume the following structure for the state-space model:

$$\begin{aligned} \theta_{it}|\theta_{i,t-1} &\sim f(\theta_{i,t-1}, W_{it}; \xi_\theta), \\ y_{it}|\theta_{it} &\sim g(\theta_{it}, V_{it}; \xi_{y|\theta}), \end{aligned} \tag{2.1}$$

where θ_{it} evolves with time in region i at time t, W and V are noise processes for θ and Y respectively, and ξ_θ, $\xi_{y|\theta}$ refer to the model parameters for the state dynamics and measurement functions. θ evolves with time to allow for adaptations to changes in observations Y. The assumptions in these models include (conditional on the parameter ξ_θ) that states are Markovian, and observations are conditionally independent; ie. given θ_t, y_t is independent of anything else. To fully define the distribution of the latent state, an initial distribution $\pi(\theta_1|\xi_\theta)$ needs to be specified. Given some initial beliefs about the state variables $\pi(\theta)$, under a Bayesian framework we are interested in finding posterior estimates of the current state given the observations to date $\pi\theta_t|y_{1:t}$. This part acts as a basic description of state-space models only, more details can be found in West and Harrison (1997).

The filtering recursion

The recursive solution to the filtering problem for state-space models can be derived as follows. By Bayes' theorem, the Markov property of the hidden state transitions and conditional independence

of observations, one can show fairly easily that

$$\begin{aligned}
\pi(\theta_{0:t}|y_{1:t}) &\propto \pi(\theta_{0:t})\pi(y_{1:t}|\theta_{0:t}) \\
&\propto \pi(\theta_t|\theta_{0:t-1})\pi(\theta_{0:t-1}) \prod_{k=1}^{t} \pi(y_k|\theta_k) \\
&= \pi(\theta_t|\theta t-1)\pi(y_t|\theta_t)\pi(\theta_{0:t-1}) \prod_{k=1}^{t-1} \pi(t_k|\theta_k) \\
&= \pi(y_t|\theta_t)\pi(\theta_t|\theta_{t-1})\pi(\theta_{0:t-1}|y_{1:t-1}).
\end{aligned}$$

The filtering recursion can be established via integrating the above equation over $\theta_{0:t-1}$:

$$\begin{aligned}
\pi(\theta_t|y_{1:t}) &\propto \int \pi(y_t|\theta_t)\pi(\theta_t|\theta_{t-1})\pi(\theta_{0:t-1}|y_{1:t-1})\mathrm{d}\theta_{0:t-1} \\
&= \pi(y_t|\theta_t) \int \mathrm{d}\theta_{t-1}\pi(\theta_t|\theta_{t-1}) \int \mathrm{d}\theta_{0:t-1}\pi(\theta_{0:t-1}|y_{1:t-1}) \\
&= \pi(y_t|\theta_t) \int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})\mathrm{d}\theta_{t-1} \\
&= \pi(y_t|\theta_t)\pi(\theta_t|y_{1:t-1}).
\end{aligned}$$

$\int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|y_{1:t-1})\mathrm{d}\theta_{t-1}$ is known as the posterior predictive distribution, $\pi(\theta_t|y_{1:t-1})$. Intuitively, it is a weighted average of the transition density, $\pi(\theta_t|\theta_{t-1})$, with respect to the filtering density $\pi(\theta_t|y_{1:t})$ at time $t-1$. $\pi(\theta_t|y_{1:t})$ is the posterior at time t.

If the system and observation processes are linear with Gaussian noise, optimal performances can be achieved by gaining an exact update via Kalman filtering. State-space models with inference method done by Kalman Filter is used in exploratory models in Chapter 4 (see Appendix). Details and applications can be found in Section 8.1. Exact updates can be also seen in the conjugate Dynamic Linear Model (West and Harrison, 1997). These relationships can be complicated if needing to find an exact form, as integrals in the above equations may be not analytical.

When doing statistical modelling, the marginal likelihood $\pi(y_{1:n})$ is often required in model selection procedures. This joint density can be shown as a product of conditional densities; $\pi(y_{1:n}) = \pi(y_1)\prod_{i=2}^{n}\pi(y_i|y_{1:i-1})$. By conditional independence, $\pi(y_i|y_{1:i-1}) = \int \pi(y_i|\theta_i)\pi(\theta_i|y_{1:i-1})d\theta_i$. This marginal likelihood is only tractable for model types such as linear-Gaussian models.

Linear Gaussian state-space models and Kalman filter

Assume that the state and observation processes are linear with Gaussian noise and have the following forms:

$$\theta_t = A\theta_{t-1} + BW_t,$$

$$Y_t = C\theta_t + DV_t.$$

This system of equations is subject to the following assumptions:

- The state process $\{\theta_t\}$ is Markovian; $\{W_t\}$ and $\{V_t\}$ are independent random variables following $\mathcal{N}(0, W)$ and $\mathcal{N}(0, V)$ respectively;
- for $s \neq t$, W_t and V_t are independent $\forall t$;
- for $t > s$, Y_t is independent of Y_s given $\theta_{0:s}$.

One of the motivations for modelling with linear Gaussian noise is due to its simplicity of derivation. All distributions of interest can be computed recursively and will have Gaussian forms, due to the property of Gaussian being closed under marginalisation, conditioning and linear transformations; it can further be shown that the conditional state and observation densities are also both Gaussian:

$$\theta_t|\theta_{t-1} \sim \text{MVN}(A\theta_{t-1}, BWB^T),$$

$$Y_t|\theta_t \sim \text{MVN}(C\theta_t, DVD^T),$$

where the posterior mean and covariance are $\bar{\theta}_{t-1}$ and Σ_{t-1} respectively.

The major interest of models described above lies in describing the current states; this can be yielded by a Kalman filter (KF). The KF (Kalman, 1960) is an important algorithm which yields exact inference provided the above assumptions for state-space models are satisfied. The one-step predictive density $\int \pi(\theta_t|\theta_{t-1})\pi(\theta_{t-1}|Y_{1:t-1})\mathrm{d}\theta_{t-1}$ is evaluated as a Gaussian with mean $\bar{\theta}_{t|t-1} = A\bar{\theta}_{t-1}$ and covariance $\Sigma_{t|t-1} = A\Sigma_{t-1}A^T + BWB^T$ following standard results (see Rue and Held (2005)). The

joint density of observations and state given θ_{t-1} is multivariate Gaussian:

$$\begin{bmatrix} \theta_t \\ Y_t \end{bmatrix} |\theta_{t-1} \sim \text{MVN}\left(\begin{bmatrix} \bar{\theta}_{t|t-1} \\ C\bar{\theta}_{t|t-1} \end{bmatrix}, \begin{bmatrix} \Sigma_{t|t-1} & C\Sigma_{t|t-1}^T \\ \Sigma_{t|t-1}C^T & K_t \end{bmatrix} \right),$$

where $K_t = C\Sigma_{t|t-1}C^T + DVD^T$ is the Kalman Gain matrix. Then it can be shown that the posterior covariance and mean have the following form:

$$\Sigma_t = \Sigma_{t|t-1} - \Sigma_{t|t-1}C^T K_t^{-1} C\Sigma_{t|t-1},$$

$$\bar{\theta}_t = \bar{\theta}_{t|t-1} + \Sigma_{t|t-1}C^T K_t^{-1}(Y_t - C\bar{\theta}_{t|t-1}).$$

It is often of interest to make forecasts using the fitted statistical models, therefore a straight-forward form of forecast distribution is useful. The k-step forward forecast distribution can be obtained from the following equations (see West and Harrison (1997)):

$$\begin{aligned} \bar{\theta}_{t+k|t} &= A^k \theta_t, \\ \Sigma_{t+k|t} &= A^k \Sigma_t (A^k)^T + \sum_{j=0}^{k-1} A^j BWB^T (A^J)^T, \\ \bar{Y}_{t+k} &= C\bar{\theta}_{t+k|t}, \\ \text{Cov}(\bar{Y}_{t+k}) &= C\Sigma_{t+k|t}C^T + DVD^T. \end{aligned}$$

All these results are extensible to time varying coefficients by replacing, say, A by A_t. Details for these standard results can be found in (West and Harrison, 1997; Ansley and Kohn, 1985).

Each posterior under this setting can be described fully by the sufficient statistics, therefore the KF is known to be the optimal estimation as mentioned earlier. However, it cannot be used to handle nonlinear nor non-Gaussian state-space models. This can be an issue when dealing with health data especially, as most health data outcomes are non-negative, are likely to be right skewed and often consist of count data. For instance, when the number of cases of a disease is the response variable, it is intuitive that they follow a Poisson distribution with some low rate (say, < 10). The KF can be fairly expensive in computation; it has the asymptotic time complexity of $O(n^3)$. When observations are large, this can be fairly time consuming.

Non-linear, non-Gaussian cases

The assumption of linear-Gaussian states and observations is often not valid in reality (rich amount of count data in health data means that they are more likely Poisson), as mentioned briefly before. A simple way to deal with this is via data-transformation. Some standard techniques to enable the KF to handle non-Gaussian linear data follow below. Such techniques including data transformation are utilised in exploratory analyses in Chapter 4.

The response data can be log-transformed and the resulting normality checked by examination of standard QQ-plots and histograms. This can reduce the skewness and improve the linearity of the data, however, this transformation cannot be applied directly to zero responses, and as such may require an alternative approach that will vary according to situation. A transformation which may be more appropriate is the Anscombe transformation (Anscombe, 1948), which transforms Poisson data (eg. counts of cases of disease) into Gaussian data. For an independently identically distributed (i.i.d.) Poisson distribution given by $X \sim \text{Poisson}(\mu)$, it follows the form:

$$Y = 2\sqrt{X + \frac{3}{8}}$$
$$\text{then } Y \sim \mathcal{N}(M, 1) \text{ approximately,}$$

for some constant M. It was suggested that the Anscombe approximation achieves a more stable Gaussian variance when the Poisson Process intensity is greater than four (Anscombe, 1948). However, when the intensity is smaller than this criteria, Zlewicz and Nason (2004) suggests that the performance of the transformation is not marginally deteriorated.

Other more sophisticated methods include the Extended Kalman filter (Tanizaki, 1996; Jazwinski, 1970) and Unscented Kalman filter (Julier and Ullmann, 2004). The former linearises the state and observation equations using a first order Taylor expansion before applying a modified KF updating procedure, whereas the latter improves the estimates of the predicted mean and the covariance matrix.

Optimal initialisation of Kalman filters

To achieve good estimates using Kalman filters, optimal initialisation of the procedure is helpful. Empirical Bayes (EB) methods estimate the prior distribution from the data, a contrast to standard

Bayesian methods where the prior distribution is determined before any observations. It can be treated as an approximation to a full Bayesian hierarchical model with parameters set to the most likely values. When the model contains hyperparameters, it is often of benefit to examine how the dynamic settings work first.

Consider the equations for the state-space model presented earlier under the assumptions that $\theta = (\gamma, \beta)$ and $\gamma \sim \mathcal{N}(0, \sigma_\gamma^2)$ and $\beta \sim \mathcal{N}(0, \sigma_\beta^2)$ respectively. We estimate the final posterior $\pi(\gamma^{(1:T)}, \beta^{(1:T)}|y, \sigma_\gamma^2, \sigma_\beta^2)$ by running the filter with initial hyperparameters, $(\tilde{\sigma}_\gamma^2, \tilde{\sigma}_\beta^2)$. For final time point T, we derive joint density:

$$l(\sigma_\gamma^2, \sigma_\beta^2|\gamma^T, \beta^T, y) \propto \log(\pi(\gamma^T, \beta^T|\sigma_\gamma^2, \sigma_\beta^2).$$

As β_i and γ_i are assumed to be independent and normally distributed, we continue the Empirical Bayes modelling by using an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). Taking expectation of the above with respect to $\log(\pi(\beta_T, \gamma_T|\sigma_\gamma^2, \sigma_\beta^2, y))$ gives

$$\mathbb{E}[\log(\pi(\sigma_\gamma^2, \sigma_\beta^2|\beta_T, \gamma_T, y))] = \mathbb{E}[\log(\pi(\beta_T, \gamma_T|\sigma_\gamma^2, \sigma_\beta^2, y))].$$

$\sigma_\gamma^2, \sigma_\beta^2$ can then be found in the subsequent EM step by maximising $\mathbb{E}[\log(\pi(\sigma_\gamma^2, \sigma_\beta^2|\beta_T, \gamma_T, y)]$; this is equivalent to maximising $\mathbb{E}[\log(\pi(\beta_T, \gamma_T|\sigma_\gamma^2, \sigma_\beta^2, y))]$.

The EM algorithm is an iterative procedure for maximum likelihood estimation. It is primarily useful when there is missing data, where we augment with desirable additional data in order to simplify calculations. To illustrate generally how EM algorithm works, one may consider the following setting: suppose that one can 'complete' data x with y so that (x, y) contains complete information. We then consider the likelihood $f(x, y|\theta)$ based upon (x, y) starting at $\theta^{(0)}$, and generate a sequence of iterates $\theta^{(m)}$. Each iteration consists of two steps.

- E-step: Calculate
$$Q(\theta, \theta^{(m)}) = \mathbb{E}_Y[log(L(\theta|x, Y)|X = x, \theta^{(m)}]$$

- M-step: Find the next iteration value of θ, say $\theta^{(m+1)}$, which maximises $Q(\theta, \theta^{(m)})$ by solving:

$$\frac{\partial(Q)}{\partial(\theta_j)}(\theta, \theta^{(m)}) = 0, j = 1, \ldots, p$$

- Iterate the above until estimates converge.

Thus, for the state-space models described previously, here in this thesis, the EB method is described systematically as below:

- (STEP 1) Choose some initial hyperparameters;
- (STEP 2) For given values of the mean μ and standard deviation σ of the state variable, estimate hyperparameters in the state equations via maximum likelihood estimates;
- (STEP 3) Stop if differences between new estimates of hyperparameters and old estimates are smaller than some error ψ; else, go back to (STEP 1).

When running the Kalman filter, all the required moments in $\mathbb{E}$ will be readily available from the posterior. The function Q is then a simple function of hyperparameters and can be maximised analytically. The EM algorithm guarantees the convergence of the sequence of hyperparameters. This optimisation method is novelly coded in R and applied when initialising Kalman Filter in Chapter 4 exploratory analysis.

This chapter provides basic knowledges required on which Bayesian inferences in later chapters is based on. They are the necessary background knowledge for Chapter 4, 5 and 6. A brief introduction of MCMC and INLA are given as they are further utilised in later Chapters; INLA being the inferential method of the main model for Chapter 4 and MCMC (a mixture of Gibbs and MH) for the rest of models considered. Latent Gaussian models are introduced here as a crucial family of models which contains a great number of models applied in this thesis. The state-space model described is mainly used as exploratory studies for Chapter 4; it employed KF as the main inferential method either by log transformed or Anscombe transformed data. More details about these can be found in the relevant chapters later.

Chapter 3

Spatial Survival Analysis

This chapter discusses some basic concepts of survival analysis and spatial survival analysis. It aims to provide a review concerning relevant background knowledge and methodologies for later chapters. Section 3.1 presents some common methods used in survival analyses, Section 3.2 introduces frailty models and finally spatial survival models and their methods for inference are reviewed in Sections 3.3 and 3.4.

3.1 Survival Analysis

Survival analysis describes the analysis of data where the outcome variable is time-to-event. The data normally observes times from a well-defined time origin until the occurrence of some particular event or end-point. In medical research, the time origin often refers to the recruitment of the study subject and the end-time could be the time of death of the subject, or of non-fatal events such as the relief of some pain of interest of study. The time to event, also known as survival time, can be measured in days, weeks, years, etc. The event of interest could be the death of a study subject, occurrence of a disease under study or the default of a machine. More details of possible applications of survival models can be seen in the book by Collett (Collett, 2003). Survival analysis is not only widely used in areas of Medicine, but also in Agriculture and Engineering; we will, however, focus only on applications to health data here. More particularly, hazard models describing long-term cancer

registry data are of interest in this thesis. Details can be found in Chapter 5.

The response variable in survival models is the observed time-to-event, values of which are non-negative and often skewed with long tails; some subjects may survive beyond the study period or drop out before the study ends. Thus, standard statistical procedures for data analysis are not compatible with survival time data as it is not sensible to assume normality in the data, and, more importantly, because much of the survival data is censored.

3.1.1 Censoring

Often in survival studies, individuals are not observed at the end-point of the study for a range of reasons. It is not unlikely that patients may withdraw before 'failure' occurs or the study may have ended while some individuals are still alive. It is also quite common that some subjects are lost to follow-up (eg. a patient in a clinical trial moved to a different country) so that their survival status is not known. We describe these cases with unknown endpoint as *censored*. Censoring is a very common case in survival studies; in fact it is more often than not that survival data contains censored ones (relevant to Chapter 5 in this thesis). However, censoring is a difficult matter as it is often hard to decide whether or not the event of interest is actually related to the study. For instance, an individual is also said to be censored when the event of interest has a different cause to that being studied, even if the actual survival time is available. For example, a study subject could have died from a fatal road accident; whether such death is a result of side effects of the treatment studied remains unclear.

There are many different forms of censoring. The common ones include right censoring, left censoring and interval censoring. When the exact survival time Y cannot be observed, right censoring records survival times greater than censored time (ie. $Y > c$ for some censoring time c). Left censoring denotes cases where $Y \leq c$. Generally speaking, an event is referred to as right censored when the lower bound for the time of interest is known and it is left censored if the upper bound of the time of interest is known. If some event is only known to have occurred within a known interval I, it is known as interval censored. In survival analysis, right censoring is the most common type of censoring. Often, independence between survival times Y and censoring c is assumed in survival analysis. That is saying these are two statistically independent events; the predetermined time c for censoring does not affect how long the study subject actually survives. For example, when study

suggests individuals shall be followed at the end of each month (this means c occurs at the end of each month), it does not affect how survival times of individuals actually behave.

Some less common forms of censoring include type I censoring, type II censoring and random censoring. Briefly, type I censoring is required when an experiment stops at a predetermined time; any remaining subjects here are then right-censored. Type II censoring refers to the ones where the experiment stops when a predetermined number of failures are observed; any remaining subjects are then right-censored. Random (or non-informative) censoring is when the censoring time of each individual is independent of their failure time. This independence assumption in random censoring is commonly made (often makes writing likelihood function easier) but ignores some possibilities and thus introduces bias in reality. For example, a patient may have withdrawn because they suffer from serious side effects which do not allow further treatments. As stated in previous paragraphs, censoring is sometimes hard to determine, for simplest in modelling, survival data are considered to be non-informative left-censored in this thesis.

3.1.2 Functions of Interest

Survival analysis involves the use of survival and hazard functions; the dependent variable is the time to event (with an indication of whether it is censored or not). Given a probability density function of survival times $f(t)$, the function of primary interest is the survival function:

$$S(t) = \mathbb{P}(T > t) = \int_t^\infty f(x)dx,$$

where t is some time, and random variable T denotes the time of death. This is a non-increasing function which measures the probability that the survival time is greater than t. When $t = 0$, $S(t)$ is defined to be 1, and as t tends to infinity, $S(t) = 0$. The hazard function, $h(t)$, measures the instantaneous rate of failure of the study subject at some time t given the subject is still surviving;

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq (t + \Delta t) \mid T > t)}{\Delta t}.$$

Additional functions of interest include the lifetime distribution function $F(t) = \mathbb{P}(t < T)$ and the cumulative hazard function $H(t) = \int_0^t h(s)\mathrm{d}s$. They measure the probability that the survival time is

less than some value t, and the integrated risk of death that one faces during time 0 and t respectively.

The relationships between these functions are defined as follows, so that knowing one of these functions is sufficient to determine the rest; for example:

- $F(t) = 1 - S(t)$, then density $f(t) = F'(t)$;
- $H(t) = -\log S(t)$;
- $h(t) = \frac{\partial}{\partial t} H(t)$.

As a by-product of all these functions, it is also important that one can answer what the expectation of life (denoted by μ here) is in survival analysis; $\mu = \int_0^\infty S(t)dt$. More details of survival and hazard functions can be found in Hosmer et al. (2008).

As with most statistical methods, the likelihood function is of interest when making statistical inference over parameters in survival studies. For left censored, right censored and interval censored data, the likelihood function of survival models takes the form

$$L(\theta) = \prod_{t_i \text{ uncen.}} \mathbb{P}(T = t_i|\theta) \prod_{\text{i l.c.}} \mathbb{P}(T < t_i|\theta) \prod_{\text{i r.c}} (T > t_i|\theta) \prod_{\text{i i.c}} \mathbb{P}(t_{i,l} < T < t_{i,r}|\theta),$$

where uncen, l.c, r.c and i.c stand for uncensored, left censored, right censored and interval censored respectively. Both the survival function and the hazard function are estimated based on observed survival times, and we can estimate these functions using parametric, non-parametric or semi-parametric approaches. Theoretically, we can fit a model to any censored data but it might be difficult to find a maximum likelihood estimate of data which is general-censored.

These functions are of close relevance to survival studies as such analysis often presents interest in relative risk or survival (hazard) curve of subject of interest. These functions provides basics of further analyses carried out in this thesis as Chapter 5 introduces extension over hazard models. As with most other statistical inference methods, likelihood function plays an important role in parameter estimation procedures. This is the base of how estimates are achieved under relevant packages in Chapter 5.

3.1.3 Non-parametric Approach

The simplest non-parametric approach for an empirical survival function is to estimate with

$$\frac{\text{Number of individuals survived at time } t}{\text{Total number of individuals in the data}}.$$

A more sophisticated way to estimate and visualise survival probabilities as a function of survival time is using a Kaplan-Meier estimator (Kaplan and Meier, 1968). Suppose that there exists a sample of size N from a given population, the observed times are $t_1 \leq t_2 \leq t_3 \leq \cdots \leq t_N$. Define n_i to be the number of subjects at risk prior to time t_i and d_i to be the number of deaths at time t_i. Then the Kaplan-Meier estimator of $S(t)$ at the event times t_i has the form of

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}.$$

A method for testing overall differences between estimated survival curves of two or more groups is the log-rank test. Another non-parametric estimator of $S(t)$ is the Nelson-Aalen (Aalen, 1978) estimator which estimates the cumulative hazard. This estimate can then be easily used to compute $S(t)$. The Nelson-Aalen estimator is formulated by

$$\tilde{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i},$$

with d_i the number of events at t_i and n_i the total individuals at risk at t_i. This is used as a method for initial exploratory analysis in Chapter 5 (see Appendix for more details).

3.1.4 Semi-parametric and Parametric Approach

Cox (1972) proposed a large family of models for survival analysis that focus directly on hazard functions. This is also the main one on which a new novel hazard model proposed in Chapter 5 is based. One of the simplest examples of this family is the proportional hazard model. Suppose that the hazard of death (or failure) depends on a set of covariates $X_1, X_2, \ldots, X_p$. Let $h_0(t)$ be the baseline hazard function which describes how the risk of event changes over time when covariates are

at baseline values. The hazard for ith individual in the study has the form:

$$h_i(t) = \phi(x_i)h_0(t).$$

Here ϕ is the relative hazard for an individual with covariate x_i compared to one with $x = 0$. It is convenient to write ϕ as $\exp\{\beta X\}$ as ϕ is non-negative. Cox models allow one to fit the model to survival times without making assumptions of any distributions (thus they are semi-parametric).

One can also use known statistical distributions for survival times in Cox models; this will then lead to parametric proportional hazard models. Hazard ratios for these full parametric models have the same assumption of proportionality and interpretation as the semi-parametric ones. Some commonly used parametric survival distributions include the exponential, Weibull, log-logistic and log-normal distributions. For example, the exponential survival model assumes a constant baseline hazard risk over time; ie. $h_i(t) = \lambda \exp\{X_i\beta\}$. The Weibull survival model is parameterised by scale and shape parameters.

Proportional hazard models with fixed covariates can be easily adjusted to take account of time-varying covariates $x(t)$; ie. $h_i(t) = \phi(x_i(t))h_0(t)$; more details can be seen in Cox and Oakes (1984). Another extension of these models is to allow for time-varying effects; this is thus no longer proportional. Mathematically, the parameter β can be written as $\beta(t)$ to accommodate such models.

An alternative to proportional hazard models in survival analysis is the accelerated failure time model. In accelerated failure time models, it is assumed that covariate effects accelerate or decelerate the survival curves in the direction of the time axis by some constant ϕ (Bradburn et al., 2003). Such models are formulated as:

$$S(t) = S_0(\phi t),$$

where $S_0(t)$ is the baseline survival function and ϕ is the acceleration factor which can be written as

$$\phi = \exp\{(b_1x_1 + b_2x_2 + \cdots + b_px_p)\}.$$

The accelerated failure time model can also be written as a function of time :

$$\log(t) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p + \epsilon,$$

where ϵ is the residual and $\exp\{b_i\}$ are referred to as time ratios.

The exponential model given above is just one example of an accelerated failure time model. Whether to use an accelerated failure time model or a proportional hazards model obviously depends on which model fits the data better, however, in some cases where the model fits are adequate for both, the choice of model is affected by other reasoning. The accelerated failure time model formulation allows the interpretation of time ratios, which can be preferable to an interpretation of two hazard ratios. Despite the potential advantages in interpretation, accelerated failure time models are still rarely seen in medical research (Kay and Kinnersley, 2002).

When variables are ordinal, for example, severity of disease could be described as "none", "slight", "moderate" or "severe", a proportional odds model may be considered to see how well this response can be predicted based on responses in other categories. The proportional odds model is a class of generalised linear models for modelling ordinal responses on discrete or continuous covariates. Let Y denote the response in the range $1, \ldots, k$ for $k > 2$, and $\gamma_j = \mathbb{P}(Y \leq j|x)$ be the cumulative response probability, where x's are covariates. The general form of a proportional odds model for the jth cumulative response probability is $\text{logit}(\gamma_j) = \alpha_j - \beta^{\mathrm{T}} x$. In this model, the intercepts α depend on the category j, but slopes β are equal for all j.

3.2 Frailty Models

In survival models, the time-to-event data described above are often also spatially-referenced and some evidence of clustering may be visually apparent. Studies of spatial survival analysis originated from modelling the spatial random effects, also known as the frailties. These survival models consisting of a random effect component with a frailty term have become more common in past decades. For the purpose of this thesis, the basic knowledge of such models is introduced here as this acts as an essential background when spatial frailty term is further employed to take care of spatial random effects in survival data (more details see Chapter 5).

The simplest way to describe frailty is that it is an unobserved random factor which affects the hazard function of some individuals. Frailties are usually assumed to be independent and are designed to account for heterogeneity in covariates. Vaupel et al. (1979) introduced the earliest example of frailty

models using a hierarchical univariate survival model. For instance, in some studies only a few covariates such as age and sex may be known, but survival times will depend on a range of other variables, like smoking and occupation. The unavailability of such variables means that a study over such a population can not be based on individual characteristics, but only on the population average. The early work has included a random effect into Cox's model to capture unexplained variation (Clayton, 1978).

Denoting the frailty term as Z, the hazard function for univariate survival times can be written as

$$h(t|Z, X) = Zh(t|X),$$

where t is the survival time and X refers to covariates. $h(t|x)$ is the baseline hazard function, or $h_0(t)\exp\{\beta X\}$. When the frailty term $Z > 1$, it increases the risk for that individual. Similarly, the survival function is the integral of the hazard function and follows the form $S(t|Z, X) - \exp\{-ZH(t|X))$, where H is the cumulative hazard function. The most common distribution for the frailty term is the gamma distribution, as it is easy to find a closed form solution for the hazard function. This is because gamma distribution fits well to failure data due to the simplicity of the Laplace transform. It allows one to use the traditional maximum likelihood estimations to determine parameters with availability of explicit Laplace transform (Wienke et al., 2005).

When the survival data is clustered, or contains the reoccurrence of one single event for the same individual, then this is likely to imply that there exists some sort of dependence among the clusters. Multivariate frailty models are used when there is dependence in survival data (Clayton, 1978; Hougaard, 1995). When the frailty term Z is the same among groups and constant over times, the hazard function can be written as

$$h(t_{ij}|Z_i) = Z_i h(t_{ij}),$$

where $h(t_{ij}) = h_0(t_{ij})\exp\{\beta X_{ij}\}$ and Z_i are assumed to be independently and identically distributed. It assumes conditional independence of observations, given the frailty.

Until now, the development of correlated frailty models is mostly restricted to bivariate frailty models. These are often constructed under several assumptions. The random effects and individual specific effects are normally assumed to be uncorrelated and constant over time. Fixed effects assume that individual specific effects are correlated with independent variables. With these assumptions, an

example of hazard functions of a bivariate frailty model follows $h(t_{ij}|Z_i) = Z_i h_0(t_{ij}) \exp\{\beta X_{ij})\}$. The random variation of frailty Z allows different risks for different groups which shows higher degree of association within the group. Under gamma frailties assumption, one common approach can be seen in Yashin and Iachine (1995) where the correlated gamma-frailty model is used and obtained a bivariate survival distribution of the form

$$S(t_1, t_2) = \frac{S_1(t_1)^{1-p} S_2(t_2)^{1-p}}{(S_1(t_1)^{-\sigma^2} + S_2(t_2)^{-\sigma^2} - 1)^{\rho}/\sigma^2.}$$

3.3 Spatial Survival Models

As access to data with geographic information is increasing in modern times, understanding the spatial variation in survival times is of great scientific interest. If these survival times are spatially correlated, then such correlation should be taken into account in any form of data analysis. Spatial variation in survival patterns often provides insights of underlying risk factors, which could potentially assist authorities in decision making processes. In statistics and epidemiology, modelling spatial variation with survival data have merged recently as an active research area. One of the novel development of spatial survival models in this thesis can be found in Chapter 5 where the inference of such models are done based on *spatsurv*.

Spatial survival models employ a range of models and methods from both Frequentist and Bayesian disciplines. Frequentist methods include semi-parametric estimating equations (Li and Lin, 2006), composite likelihood methods (Paik and Ying, 2012) and scan statistics (Huang et al., 2007; Bhatt and Tiwari, 2014). Bayesian methods commonly used are semiparametric models (Banerjee and Carlin, 2013; Banerjee and Dey, 2005), parametric models (Hennerfeind et al., 2006; Diva et al., 2008a), Polya tree models (Zhao and Hanson, 2011) and spatial accelerated failure time models (Zhang and Lawson, 2011). The Bayesian approach to fitting such frailty models is to use simulation-based methods such as MCMC. The methodological work of Li and Ryan (2002) extended ordinary frailty models by allowing random effects accommodating spatial correlations into the baseline hazard function multiplicatively. One example of a gamma frailty model with spatial dependence can be seen in the investigation of leukaemia survival data in the north west of England (Henderson et al., 2002a); results confirmed the dependence of hazard on location.

Often, the assumptions required for Cox proportional hazard models are too restrictive in reality. It is fairly common that relevant covariates for the model follow different shapes to the proposed baseline hazards. It may even be the case that such covariates are random variables evolving over time scales as well. Therefore, extensions of models which relax the restrictions imposed by Cox proportional hazards models or/and allow inclusion of time-varying effects are of interest in certain situations. The R package *spatsurv* provides some computationally efficient adaptive MCMC methods to deliver Bayesian inference over parametric proportional hazards with spatially referenced time-to-event data. This package is easily extensible and it reduces the costs of deriving model-based solutions to these models from $O(n^3)$ to $O(n)$ (Taylor and Rowlingson, 2014). In this package, spatially-correlated frailties are assumed to follow log-Gaussian stochastic processes but more flexible adjustments can be made to baseline hazard functions and spatial covariance functions; more details of the development of these models is shown in later chapters.

Spatial survival models consisting of a random effect component with a frailty term which models the possible spatial random effects have become more common in past decades. This can be seen as an extension of Section 3.2. Under spatial survival models, frailty terms are assumed to be spatially arranged such that random effects in closer proximity are similar in magnitude. In another word, frailties are similar if locations are nearer to each other. The joint distribution of these effects are interpretable in a spatial context.

Mathematically, under the usual assumption of proportional hazards $h(t_i; \phi, Z_i)$, where ϕ is a vector of all parameters for baseline hazard function and covariate coefficients, the baseline hazard function h_0 under the standard proportional hazard function for each study subject i can be extended to include a spatial frailty term $Z_i = \log \zeta_i$:

$$\begin{aligned} h(t_i; \phi, Z_i) &= h_0(t_i)\zeta_i \exp\{X_i\beta\} \\ &= h_0(t_i) \exp\{X\beta + Z_i\}. \end{aligned}$$

In cases where the space is partitioned into different regions j, it is assumed that $\zeta_i|\mu_j \sim \Gamma(1/\psi, 1/(\psi\mu_j))$, where μ_j is the mean frailty in region j. Regional mean frailties μ_j's are often assumed to follow a Gaussian distribution with mean 1 and some variance. The correlation between regions are likely to be dependent on distances d_{ij} between two regions i, j.

To ensure positivity for spatially correlated effects as well as providing interpretable results on $\exp\{Z\}$ as a multiplicative scale on hazard function, it is often seen that one works with Z_i instead of ζ. For the purpose of this thesis, we mainly introduce the case where Z is a spatially continuous stationary latent Gaussian field. Suppose that the parameters for this latent field are denoted by η where for marginal variance σ^2, we reparameterise Z such that $\mathbb{E}[\exp(Z)] = 1$.

One of the simplest expressions for this spatial correlation r is the exponential form $\sigma^2 \exp\{-\phi d_{ij}\}$, where $\sigma^2, \phi > 0$. Intuitively, exponential models say that the correlation between two objects reduces as the two points are further away from each other. It specifies the covariance of the two objects which are d units distance apart based on a marginal variance σ^2 and a spatial decay parameter ϕ. Here σ^2 is the marginal variance within each region, ϕ refers to the range out of which spatial correlation does not exist and d is the Euclidean distance between location i and j. Other assumptions include powered exponential (Diggle et al., 1998) and Matèrn (Matern, 1960). More details on different model assumptions of spatial frailty terms can be found in Wienke (2010) and Banerjee et al. (2014).

When the random process of frailty denoted by Z is defined only on discretely partitioned regions of the geographical space, conditionally autoregressive (Besag et al. (1991)) models are often used. With similar notation to before, it can be written in the following form:

$$\mu_j | \mu_{\text{neighbours}} \sim \mathcal{N}(\bar{\mu}_{\text{neighbours}}, \sigma^2/\text{numbers of neighbours}).$$

It sometimes may be of interest to look at point-location based data especially in cases where fine-scale spatial structures are available. The gamma marginals for point data are assumed to have mean 1 and variance ψ and the covariance matrix is defined to have diagonal entries ψ and other entries $\tau\psi\rho_{ij}$ where $\tau \in (0, 1)$ and ρ is the correlation parameter depending on distance d as stated above. Another approach which incorporates the point data is an extension of standard generalised additive models where under spatial setting, a smoothing function for residual spatial variation is included. For observations $t_1, \ldots, t_n$ at locations $x_1, \ldots, x_n$, the model follows:

$$f(\theta|x, t) = u(x)'\beta + \sum_{i=1}^{n} g_i(x_i),$$

where f is the link function, $u(x)$ is a vector of risk factors for each location and $g(x)$ is the function for modelling smooth residual spatial variation. It is also possible for these models to allow random

slopes and intercepts with respect to different groups of data; eg. different locations may have different slopes.

An example of a gamma frailty model applied to spatial survival data can be seen in Henderson et al. (2002b) where they modelled survival data of leukaemia in the north west of England with Breslow (Breslow, 1974) estimates. Li and Ryan (2002) proposed a semi-parametric frailty model which allows spatial random effects to be accounted for in the baseline hazard multiplicatively. The application to infant mortality in Minnesota (Banerjee and Carlin, 2002; Banerjee et al., 2003) demonstrated how spatial frailties do not only capture spatial trends but also improve the performance of models using the semi-parametric proportional hazards modelling. Diva et al. (2008b) used both proportional hazard models and proportional odds models for cancer data with a Weibull baseline hazard function. All of these models employed MCMC as the inferential method.

3.4 Inference for Spatial Survival Models

These spatial survival models can be implemented in a Bayesian framework. One way to do so is via MCMC methods, where the posterior of interest has the following form:

$$\pi(\psi|\text{data}) \propto \pi(\text{data}|\psi)\pi(\psi).$$

We can use MCMC to draw samples from the target posterior and find the posterior expectation $\mathbb{E}_{\pi(\psi|\text{data})}[g(\psi)]$. For the purpose of this thesis, we will briefly introduce an advanced adaptive MCMC method for spatial survival modelling which will be used later in this thesis (see Chapter 5). This method can be applied to spatial survival datasets and the computational cost is shown in Taylor (2015) to reduce from $O(n^3)$ to $O(n)$ with the cost of increasing grid size being $O(m \log m)$.

Suppose that for a vector of parameters $\Phi = (\omega, \beta, \eta)$, the hazard function follows

$$h(t_i; \Phi, Z_i) = \exp\{X_i\beta + Z_i\}h_0(t_i, \omega),$$

where Z is some spatially continuous stationary latent Gaussian field and Z_i is the value at location of observation i, ω is a vector of parameters in h_0, and η denotes parameters of the covariance function

of Z. The Exponential model is a suitable proposal for $\text{Cov}(Z)$; ie. $\sigma^2 \exp\{-d/\phi\}$ with σ^2 being the marginal variance of fields and ϕ the 'spatial decay' parameter.

The MCMC inference of this particular type follows Metropolis-Hastings scheme with the following adjustments. As the parameters are mostly defined on positive numbers, we will be working with log-transformed versions. In the MCMC scheme, Z's are not worked with directly, rather with a vector of transformed variables, $\gamma = (\gamma_1, \ldots, \gamma_m)$, such that $Z = -\sigma^2/2 + \Sigma_{\sigma,\phi}^{(1/2)}\gamma$. Here, $\Sigma_{\sigma,\phi}^{(1/2)}$ is the Cholesky decomposition of the covariance matrix. Apriori, $\gamma_0^{(s)} \sim \mathcal{N}(0,1)$ and γ's can be generated by the simulation of multivariate Gaussian variables with mean $-\sigma^2/2$ and variance Σ_η. Appropriate Z's are then constructed via simulated γ's.

Samples of MCMC are drawn from the posterior $\pi(\Phi, \gamma|\text{data}) \propto \pi(\text{data}|\Phi, \gamma)\pi(\Phi, \gamma)$, using MCMC Gamerman and Lopes (2006); Gilks et al. (1995) where the parameters are transformed; eg. $\tilde{\omega} = \log \omega$. $\pi(\text{data}|\Phi, \gamma) = \pi(\text{data}|\beta, \tilde{\omega}, \gamma)$ due to the conditional independence. The MCMC scheme has Langevin kernels for β, $\tilde{\omega}_f$, γ and a random walk kernel for $\tilde{\eta}$. The algorithm follows:

- Initialise the chain at $\{\beta^{(0)}, \tilde{\omega}^{(0)}, \tilde{\eta}^{(0)}, \gamma_{(0)}\}$;
- Proposal density for $\zeta = (\beta, \tilde{\omega}, \tilde{\eta}, \gamma)$ is $q(\zeta(i^*)|\zeta^{(i-1)}) = \mathcal{N}(\zeta^{(i^*)}; \mu_{\zeta^{(i-1)}}, h^2\Sigma)$, where

$$\mu_{\zeta^{(i-1)}} = \begin{bmatrix} (\beta, \tilde{\omega})^{(i-1)} + \frac{h^2 h^2_{\beta,\tilde{\omega}}}{2}\Sigma_{\beta,\tilde{\omega}}\frac{\partial \log\{\pi(\zeta^{(i-1)}|Y)\}}{\partial(\beta,\tilde{\omega})} \\ \tilde{\eta}^{(i-1)} \\ \gamma_{(i-1)} + \frac{h^2 h^2_\gamma}{2}\Sigma_\gamma \frac{\partial \log\{\pi(\zeta^{(i-1)}|Y)\}}{\partial\gamma} \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} h^2_{\beta,\tilde{\omega}}\Sigma_{\beta,\tilde{\omega}} & 0 & 0 \\ 0 & ch^2_{\tilde{\eta}}\Sigma_{\tilde{\eta}} & 0 \\ 0 & 0 & h^2{}_\gamma\Sigma_\gamma \end{bmatrix}.$$

- Here the constants h^2_- are optimal scalings in MALA proposals (Roberts and Rosenthal, 2001); More details can be seen in Taylor and Rowlingson (2014) and Taylor (2015) . The optimal value of h should give asymptotic acceptance rate 0.574.

Other inferential methods for spatial survival data can be seen in the R package *spBayesSurv* where spatial copula linear dependent Dirichlet process mixture models, anova Dirichlet process mixtures and marginal proportional hazard models can be implemented (Zhou et al., 2018). Package *BayesX* (Umlauf et al., 2015; Belitz et al., 2015) fits different types of spatial survival models based on MCMC simulation techniques. Examples of model classes supported by *BayesX* include generalised additive mixed models, dynamic models, geoadditive models and models for space-time regression.

The package provides different smoothness priors including Markov random field priors for spatial effects and allows flexible parametric baseline hazards using penalised splines. Also, *BayesX* allows handling of time-varying coefficients. *INLA* package also allows the fit of survival models but instead of using MCMC techniques, a Laplace approximation is used. Although this is computationally faster, it does not allow one to make exact inference over parameters and problems occur when models are of higher hierarchies.

This chapter aims to provide essential background knowledge of (spatial) survival analyses and their relevant inferential methods.The information contained in this Chapter is relevant to Chapter 5. The survival data considered in Chapter 5 are censored data. Non-parametric approach (Kaplan-meier curve) is used to provide exploratory analysis. Semi-parametric approach (Cox PH model) is the root model which the hazard model introduced in Chapter 5. Frailty models provide the basics of how spatial random effects can be involved using frailty terms and spatial survival model is what the extension in Chapter 5 is based on. The main inference method for this thesis is via Bayesian framework. More particularly, based on MH scheme, an advanced adaptive MCMC method is employed to proceed inference for spatio-temporal survival model in R. More details can be found in Chapter 5 or *spatsurv* package vignette.

Chapter 4

Spatiotemporal Modelling of Tuberculosis Incidence in Urban Portugal from 2000–2013

This Chapter delivers an analysis of space-time tuberculosis data in Urban Portugal area between 2000 and 2013 at a lower administrative level than other studies. This Chapter delivers an analysis of space-time tuberculosis data in Urban Portugal area between 2000 and 2013 at a lower administrative level than other studies. The novelty shows in application of INLA approximation over space-time tuberculosis data at a finer geographical scale. This is a submitted paper to the International Journal of Tuberculosis and Lung Disease. Background and previously done research on similar area are shown in Section 4.2. Section 4.1 presents the analysis using INLA method on urban tuberculosis incidence rates discovering spatial clusters, identifying high risk areas and understanding the relationship between risk factors. Section 8.1 shows a few other approaches pursued on the same data as exploratory analyses.

4.1 Spatiotemporal Modelling of Tuberculosis Incidence in Urban Portugal from 2000–2013

4.1.1 Abstract

Despite being one of the medium-to-low endemic countries, Portugal still shows one of the highest tuberculosis incidences in the European Union. Although it has seen progressively decreasing incidences, the regional differences suggest that better understanding at sub-regional epidemiology would assist control over tuberculosis. In this chapter, the analysis looked at tuberculosis incidences at freguesia levels (lower administrative levels than municipalities in Portugal) in Lisbon and Oporto from 2000-2013. It used Poisson mixed effect models with both spatial and temporal correlated random effects, interaction terms and unstructured random effects. Models are adjusted for socioeconomic covariate effects with inferential method based on INLA. Both regions have identified areas of higher incidence; freguesias near the central Lisbon city area and a few areas in the north. Clear spatial clusters are detected in both Lisbon and Oporto Metropolitan areas. Areas showing high relative risks are not necessarily associated with high population densities but are surrounding areas of poverty zones.

4.2 Background

Tuberculosis (TB) is a curable and preventable communicable disease which continues to raise great concerns globally. Despite the availability of treatment since the 1940s, TB was still found to be the 11th most common cause of death in 2011 (WHO, 2013) and the second leading cause of death due to a single infectious agent, after HIV infection (WHO, 2014). In 2012, there were 68,423 new TB notifications in the thirty European Union Member States and European Economic Area countries (De Vries et al., 2014). Despite an average declining trend in TB incidences since 2006 globally, a stable or even increasing trend has been observed in some Western European large cities (WHO, 2007). Risk factors for TB are well understood and include HIV infection, overcrowding, immigrant flow from higher prevalence regions and poor living conditions (Rieder, 1999; De Vries et al., 2014). As well as the fact that these risk factors tend to be more prevalent in urban environments, there also tends to be a higher concentration of people and a correspondingly greater number of interactions

between people, thus the risk of transmission is also elevated in this setting (Rieder, 1999).

TB incidence in big cities is typically higher than in rural areas and this difference varies between and even within countries. For example, a WHO factsheet reported annual incidence rates of 20 cases per 100,000 population in Barcelona and Milan and between 35 to 45 cases per 100,000 population in Paris and London. In comparison, annual countrywide rates were reported as being between 8 cases per 100,000 population in large European cities(WHO, 2007). A more recent report has shown that in 2009, low-incidence ($<$ 20 per 100,000) European countries experienced approximately 2.5 times higher TB rates in big cities compared to the national averages. De Vries et al. (2014) reported that in 2009, among the cities in countries which showed low incidences, Birmingham and London had the highest TB rates (58 and 44.4 respectively) followed by Brussels (29.9), Barcelona (27), Paris (23.4) and Rotterdam (21.3). These cities have all exceeded twice the rate of national levels. Such difference between these cities may be a result of different socioeconomic conditions, structures of society, different control policies or possible even genetic differences in population. The figure in the UK also shows that TB notification varies within countries, with 70% of the notifications being found in the top 40% deprived regions in 2013 (Public Health England, 2014).

4.2.1 Tuberculosis in Portugal

The first programme integrating TB and primary health care was established in 1984 in Portugal but a more successful notification system with annual reports started in 1988 (DGS-TB, 1994). In the 1990s, more progress was seen in TB control over treatments in Portugal (Antunes and Fonseca-Antunes, 1996); this included development of a computerised database to store detailed information about subjects, national guidance on drug regimens for TB treatment and a national reference laboratory for quality control. On the prevention side, following WHO guidelines, Bacillus Calmette-Guérin vaccination was given to children at birth, at 5 to 6 years of age and 11 to 13 years of age when Mantoux tests results showed negative. Other methods such as screening and chemoprophylaxis are also used for TB prevention. Portugal is 100% covered by the National Programme for TB Control in Portugal (PNT) following the guidelines of WHO Directly Observed Therapy Short Courses (DOTS) (WHO, 1997). This programme has operated on an annual appraisal basis since 2001.

Portugal has shown a slow but consistent reduction in TB notification rate in recent decades - an

average and consistent annual reduction of 5% in incidence was observed in the period from 2000 to 2010. With all efforts made (including the 100% coverage by both PNT and DOTS), Portugal has been recently classified as a low-incidence country by the World Health Organisation with under 20 cases per 100,000 population. However, it still remains as one of the highest rates in the European Union (DGS, 2016). In 2015, Portugal had a notified TB incidence of 18.6 per 100,000 population (DGS, 2016), with an estimated detection rate of around 90%. The economic crisis of 2007 to 2008 has contributed to rising unemployment rates and poverty and a deterioration in living conditions. Associations between these socioeconomic disruptions and the worsening of some health outcomes have been suggested and widely discussed, although the relationship is not well established as yet (Karanikolos et al., 2013; Ayuso-Mateos et al., 2013). In the last decade, several studies have identified the Greater Lisbon and Oporto Metropolitan areas as the most critical regions for TB control; they have the highest incidences in Portugal (Couceiro et al., 2011; Areias et al., 2015). These two regions also exhibit different patterns in incidence over time (Areias et al., 2015; Bras et al., 2015). Thus, it is of interest to look into these two regions at a finer scale and establish a more detailed vision about the behaviour of TB incidence.

4.2.2 Previous Research on Modelling TB Incidence

TB incidence is often linked to environmental exposures and person-person interactions. Recent epidemiological studies have investigated the interaction of risk factors based on demography, environment, genetics and other socio-economic factors. Potential spatial risk patterns can be found according to the interaction of these factors in both space and time. This can be useful in public health planning and decision making (Farmer, 1997). Couceiro et al. (2011) conducted a country-wide analysis of TB and developed a sequence of statistical methods to identify risk factors and high risk areas for pulmonary TB for supporting local interventions. This includes a multivariate regression method with variable selection and identification of spatial clusters based on Kulldorff's scan statistics (Kulldorff, 1997). Some areas were shown to have higher relative risks; Oporto and Lisbon Metropolitan Areas showed the highest incidence.

Spatial scan statistics can be extended to the space-time domain and include both retrospective (Kulldorf et al., 1998) and prospective (Kulldorf, 2001) methods. Instead of using circular windows as in the spatial scan statistic, the space-time method uses a cylindrical window with the circular part of

the window relating to space. The space-time scan statistic identifies the most significant cluster area under the null hypothesis that the data are from a constant risk Poisson process. Inference is based on the maximum likelihood ratio of a potential cluster with a null assumption of no clustering; Monte Carlo simulation is used to obtain the approximate distribution of the test statistic. The retrospective scan statistic tests the spatio-temporal cluster for a geographical region over a predetermined time whilst the prospective scan statistics only detects the regions which present excess risk in the last time period. These methods described in above papers are available via the SaTScan™ software (`www.satscan.org`).

Using scan statistics, Gomez-Barroso et al. (2013) found that there are 28 significant pure spatial clusters and 20 spatio-temporal clusters, with the most likely cluster formed by 7 municipalities within Greater Barcelona Area. The same areas are seen in most spatio-temporal clusters with time intervals between April 2008 and March 2009. The space-time scan statistic with 25km spatial window and 12 month time window was used to detect spatio-temporal clusters. A similar approach of employing the retrospective space-time scan statistic is also seen in (Zhao et al., 2013) where the most likely cluster is seen in southern-central regions of China between 2006 and 2008. Other examples of space-time scan statistics can be found in Onozuka and Hagihara (2007); the authors cite the identified cluster in the north of Fukuoka as a reason to review control measures for TB.

Nunes (2008) further developed Kulldorf's space-time scan statistic in two directions: using the mean spatial semivariograms to determine window size and format, then employing geostatistical simulations for a posterior validation step after cluster identification. The semivariogram showed evidence of clear spatial patterns in TB incidence (60% spatial contribution) within a 143km range.The semivariogram was then used to set scan window parameters. Their results showed a significant high rate of incidence in 3 critical regions (Oporto, Setúbal and Lisbon) between 2000 and 2004 with a temporal cluster over the whole of Portugal in 2002. A more recent analysis of this issue between 2000 and 2010 (Areias et al., 2015) showed that the most likely space-time clusters were still presented by the Lisbon and Oporto regions with more rapid declines compared to the rest of country.

Randremanana et al. (2010) approached spatio-temporal modelling on TB data using Bayesian methods and a generalised linear mixed model (GLMM). They included the following covariates: number of re-treatments, number of failures, number of losses to follow-up, number of households with two or more cases and distance between residence to treatment centre. They assume a Poisson distribution

for the observed number of cases and produce model inferences using MCMC. Having adjusted for the covariates, spatial and non-spatial random effects, they found that 19.28% of the neighbourhoods in Antananarivo show significantly higher than average TB risk and that those areas are clustered. The number lost to follow-up and the number of households with two or more cases were found to be important in identifying higher risk areas. Leite da Roza et al. (2012) used a similar Bayesian approach to confirm the spatial patterns of TB in Ribeirão Preto Brazil. The covariates in their Poisson GLMM model included income, education and social vulnerability.

Cao et al. (2016) used a spatiotemporal Bayesian negative-binomial to model observed TB rates; they fitted their model using the WinBUGS software. The paper analysed data for TB cases in 31 provinces of mainland China between 2009 and 2013. Alongside a spatio-temporal interaction term, the identified risk factors included average temperature, rainfall, wind speed, and air pressure.

Kipruto et al. (2013) conducted a county-level analysis of TB incidence in Kenya using a Bayesian Poisson GLMM. The model considers a first order time trend together with structured and unstructured spatial effects between 2012 and 2014. Inference was delivered using the INLA software for R. The results indicated that the Nairobi, Mombasa, Isiolo, Homa bay, Kisumu and Siaya regions were at greater risk of severe TB outbreaks. Gender proportion, HIV proportion, the proportion receiving directly observed treatment (DOT) therapy, average weight and average age were found to be significant risk factors.

In summary, a variety of techniques have historically been used to understand the spatial/spatiotemporal variation in risk of tuberculosis in different regions; some models have included a variety of covariates. The most common two methods are the spatial scan statistic approach and the GLMM approach; both having their own advantages and disadvantages. The spatial scan approach detects clusterings and allows flexible choices of window sizes but it does not adjust for relevant risk factors. Also predeterminate of appropriate window sizes maybe challenging in practice. Other methods mostly considered Bayesian GLMM as such model allows free intake of covariates comparing to scan statistics. The spatiotemporal patterns are taken into account by including random effect terms; this avoids the choice of window sizes comparing to scan statistics. However, appropriate assumptions modelling distributions for random effects can be hard to decide.

The previous studies all showed spatial and temporal clusterings in the behaviour of TB, together with

interactions between space and time. As a follow-up study to TB in Portugal, knowing that it shows space-time behaviours, the present part in thesis follows the modelling approach of Randremanana et al. (2010), Cao et al. (2016) and Kipruto et al. (2013), using a Bayesian GLMM with spatiotemporal interaction terms. This is because they share similar nature of the data (tuberculosis cases with spatial and temporal information) through an overlapping period of time. The models are fit using the R-INLA package Rue et al. (2009) for its ability to provide fast model fits.

4.2.3 Aims and Structure of Report

The goal of the present study is to model annual TB incidence rates at the freguesia (the smallest administrative division in Portugal) level in two major areas in Portugal, focusing on the Lisbon and Oporto Metropolitan areas.The response variables used are TB notifications from a period of 14 years between 2000 and 2013. This part of the thesis aimed to understand the effect of various ecological factors on the number of observed TB cases per 100, 000 population including measures of HIV incidence, ageing, homelessness, the proportion of immigrants, unemployment and overcrowding in each freguesia. The TB cases are obtained from TB surveillance system (SVIG-TB, Directorate General of Health) which relies on compulsory notification of cases by clinicians, such identification strictly follows the WHO?s strategy defined as Directly, Observed Therapy, Short Course (DOTS). In order to account for unexplained variation in risk, spatio-temporal random effects and interaction terms are included.

Details of the data and statistical models we considered can be found in Section 4.3 of this part. In Sections 8.1.2 and 4.5, the summaries and interpretations of the results of analyses are given including presenting estimates of covariate effects and maps of relative risks. The study concludes with a discussion in Section 4.6.

4.3 Methods

4.3.1 Data Description

Table 4.1 gives details of all the variables considered in this study including the temporal and spatial resolution and the exact definition of each indicator including the numerator and denominator. Incidence data were obtained from the national database of the TB surveillance system (SVIG-TB, Directorate General of Health). The six covariates considered in the analysis are chosen based on TB factors established in the literature as discussed above; these include HIV incidence, ageing index, homeless population, percentage of immigrants, unemployment rate and overcrowded living index. The Poisson regression models used resident population estimates from Statistics Portugal as an offset. Plots of the covariates from the most recent time point in the data can be found in Figures 8.11 and 8.12 in the Appendix.

The goal is to model annual incidence at the freguesia level – 308 in Lisbon and 383 in the Oporto region. As can be seen from Table 4.1, not all of the variables are available at that spatial or temporal resolution. Therefore, the linear interpolation technique was used to impute values of covariates for the years where they had not been officially recorded. For HIV incidence, where data was only available at the larger municipality-level, the freguesia-level data were obtained by overlaying the population weighted centroid of each freguesia onto the municipality shapefile. Then each freguesia was ascribed the corresponding municipality-level HIV incidence whose weighted centroid lay inside it.

For the purpose of statistical modelling process, it is assumed that there are no missing values in TB incidence as it is based on mandatory reports; if no reports happened during the time period in some regions, the TB notification is assumed to be 0. Such assumption may result in underestimation of incidences per month as it is likely that undiagnosed cases exist. However, it is still a reasonable assumption when modelling considering this is the best data available. Such assumption should not make a huge difference on the model results in general.

Variables	Available years	Variable definition	Numerator (source)	Denominator (source)
TB Cases	2000 to 2013	Notification rate/100,000	Reported new TB cases (DGH)	–
HIV Incidence[1]	2000 to 2013	Notification rate/100,000	Reported new HIV cases (NHI)	Resident population estimates (SP)
Ageing Index	2001 and 2011	Ratio between older and younger population (×100)	People aged > 65 years (SP)	People aged 0-14 years (SP)
Homeless Present Population	2001 and 2011	Number of homeless population	Number of population not in conventional dwelling	–
Immigrants	2001 and 2011	Percentage of immigrants	Number of foreign population (SP)	Resident population estimates (SP)
Unemployment	2001 and 2011	Percentage of unemployed	Unemployed population (SP)	Resident population at working age (SP)
Overcrowded Living Quarters	2001 and 2011	Indicator of excess number of residents in accommodation[2]	Number of overcrowded dwelling units (SP)	Total number of households units(SP)

Table 4.1: Summary of variables definitions included in this study, per freguesia. DGH - Directorate General of Health; SP - Statistics Portugal; NHI - National Health Institute Ricardo Jorge.
1 HIV incidence only available at municipality level;
2 The calculations are based on 1 living room; 1 room per couple; 1 room for another non-single person; 1 room for a single person 18+; 1 room for two single people of the same sex aged 7-18; 1 room for each single person of different sex aged 7-18; 1 room for two people under 7.

4.3.2 Statistical Models

This thesis considered several statistical models for chosen outcome, the annual number of TB counts in each freguesia, in the Lisbon and Oporto metropolitan areas. The models follow the rationale in the previous TB studies; both spatial clusters and temporal trend in the disease were found. Thus, the models in this thesis considered are with and without both spatially and temporally correlated random effects and with and without a spatiotemporal interaction effect. Models that included unstructured spatial and temporal random effects are also fitted as more complicated alternatives. Both Poisson and zero-inflated Poisson models are considered for reasons that the data may suffer from overdispersion. Details of all models fitted are given in Table 4.2.

All models in Table 4.2 were fit to the Lisbon and Oporto metropolitan areas separately and compared using the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010) within each region (rule of thumb, smaller WAIC means better model fit). Where present, the spatially correlated random effects in all models were assumed to follow the Besag, York and Mollié (BYM) Model (Besag et al., 2008). The model is widely used in scenarios where the risk from a disease over contigu-

ous administrative zones must be estimated from noisy observed incidence or mortality rates (Besag et al., 1991). The spatially-structured random effects are split into two parts where one can be explained by how close areas are within each other and the other being the random effects which still remains unexplained. One of the reasons that BYM model is widely used in disease mapping is that it allows flexibility in the residuals. Temporally correlated random effects were modelled by an autoregressive process of order 1 (RW1). All models are fitted using the INLA software package for R (Rue et al., 2009; Martins et al., 2013). Details for BYM and RW1 models implemented in INLA can be seen in `http://www.math.ntnu.no/inla/r-inla.org/doc/latent/bym.pdf` and `http://www.math.ntnu.no/inla/r-inla.org/doc/latent/rw1.pdf` respectively.

Model	Model description	Linear Predictor	regions	WAIC (Poisson)	WAIC (ZI Poisson)
M1	no random effects	$\eta_{it} = X\beta$	Lisbon	24179.77	24699.75
			Oporto	21042.55	22117.23
M2	spatially correlated random effects	$\eta_{it} = X\beta + v_i + \nu_i$	Lisbon	19498.56	21511.97
			Oporto	18394.63	20287.06
M3	temporally correlated random effects	$\eta_{it} = X\beta + \gamma_t$	Lisbon	23450.77	24180.98
			Oporto	20551.57	21733.00
M4	spatially and temporally correlated random effects	$\eta_{it} = X\beta + v_i + \nu_i + \gamma_t$	Lisbon	19258.19	21284.45
			Oporto	17926.00	19876.70
M5	spatially and temporally correlated random effects with interactions	$\eta_{it} = X\beta + v_i + \nu_i + \gamma_t + \delta_{it}$	Lisbon	16694.93	19176.31
			Oporto	16776.2	18954.27
M6	spatially and temporally correlated random effects with interactions and unstructured random effects	$\eta_{it} = X\beta + v_i + \nu_i + \gamma_t + \phi_t + \delta_{it}$	Lisbon	16688.73	19161.41
			Oporto	16777.23	18957.53

Table 4.2: Summary of Models and achieved model fit as measured by the WAIC. Here, the subscript i denotes region and t denotes time. $X\beta$ are (fixed) covariate effects, v_i are spatially correlated random effects, γ_t are temporally correlated random effects, δ_{it} are spatiotemporal interaction random effects between ν_i (unstructured spatial random effects) and ϕ_t (unstructured temporal random effects). ZI = zero-inflated.

For these models above, the hyperparameters for BYM model and RW1 model have prior distributions following log-gamma (R-INLA, 2017) and penalised complexity prior (PC prior Simpson et al. (2015)) respectively. For this part of the thesis, the models are run with default priors set by R-INLA. A summary of the priors for hyperparameters is given in Table 4.3

Model Prior	Formula	Parameters	Default Values	More Details
log-Gamma	$\pi(\tau) = \frac{b^a}{\Gamma(a)}\tau^{a-1}\exp\{-b\tau\}$	τ_1: precision of ν	$a = 1$	We work with
	$\tau > 0$	τ_2: precision of v	$b = 5 \times 10^{-4}$	$\theta = \log\tau$
PC	$\pi(x) = \frac{\theta}{2}\exp\{-\theta\exp\{-\frac{x}{2}\} - \frac{x}{2}\}$	(u, α)	$u = 1$	$\sigma = 1/\sqrt{\exp\{x\}}$
	$\theta = -\frac{\log\alpha}{u}, \theta > 0$	$\mathbb{P}(\sigma > u) = \alpha$ for σ s.d.	$\alpha = 0.01$	$u > 0$, $0 < \alpha < 1$

Table 4.3: Summary of Priors of Models: v_i are spatially correlated random effects; ν_i are unstructured spatial random effects; more details are listed under R-INLA documents

As discussed above, in M6 the spatial components are modelled under BYM, temporal components are modelled under RW1 and the interactions are iid. Details of each component are given in Table

4.4.

Model	Formulation	Explanation
BYM	v: modelled under Besag $\nu \sim \mathcal{N}(0, \tau_\nu^{-1}\mathbf{I})$	The regional spatial effect decomposes into $v + \nu$ v: structured spatial effects, ν: independent effects.
RW1	$\gamma_t - \gamma_{(t+1)} \sim \mathcal{N}(0, \tau^{-1})$	

Table 4.4: Summary of Latent Models: INLA treats the effects as a vector of $x_i|x_{-j} = (v + \nu, \nu)^T$ in BYM as this allows us to get the posterior marginals of the sum of the spatial and iid models; the hyperparameters are treated in the log-transformed form.

The spatial, temporal and spatiotemporal interaction random effects in the model represent variation in the risk of disease that we cannot understand through the covariates included in this analysis. An area with a large spatial random effect is an area in which a consistently higher number of TB cases are reported over time compared to what one would expect given the covariates for that area. The exponentiated temporal random effect ($\exp\{\gamma_t + \phi_t\}$) is a measure of global relative risk within each of the two regions; the reference rate of 1 is tied (as it is with the other types of random effect in the model) to the space-time window which have been used in the analyses. When the relative risk is greater than 1, this means that there were globally more TB cases than what would be expected on average for the space-time window considered. The spatiotemporal random effects capture that element of risk which cannot be captured by the combination of the covariate effects and the spatial and temporal random effects. Freguesias and times at which the interaction effects are high could potentially be thought of as unexplained clusters; in sparsely populated rural areas, a high interaction effect may be triggered by a single case.

Examination of model fits are based on whether there is a good match of predicted versus observed values. The plot is presented using binned predicted values with number of bins for each region selected using Sturges' formula (Sturges, 1926). Adjusted probability integral transforms (PIT) for the final models are also computed for both regions to justify whether the observed values are reasonably modelled under specified distributions (Dawid, 1984; Czado et al., 2009; Held et al., 2009).

4.4 Model Results

According to Table 4.2, the model with the smallest WAIC value for the Oporto data is the Poisson model with spatio-temporal random effects and interactions between them (M5); for Lisbon, the best

model also included unstructured spatial and temporal random effects (M6). In the remainder of this chapter, we will present the results from these two best fitting models.

Table 8.1 shows the fixed effects for the Lisbon Metropolitan and Oporto Metropolitan areas. In order to understand which of the effects had the greatest effect on TB incidence, standardised covariates are used in fitted models; i.e. the effects presented in the table are for standardised covariates.The means in this table are the posterior means of distributions for each parameter. They act as the estimated values for the parameter. As Bayesian provides posterior distributions as an outcome, the variances of the distributions (standard errors here) represent the error margins for the estimations.

Lisbon	mean	sd	(0.025,0.975) quantiles	non-standardised mean
(Intercept)*	-7.769	0.028	(-7.826, -7.713)	-7.955
ageing	0.003	0.039	(-0.074, 0.080)	3.116×10^{-5}
immigrants	0.017	0.031	(-0.045, 0.078)	0.004
unemployment	0.008	0.032	(-0.054, 0.070)	0.002
overcrowding	0.036	0.035	(-0.032, 0.103)	0.008
homeless	0.045	0.031	(-0.016, 0.105)	0.015
hiv	0.024	0.030	(-0.034, 0.082)	0.003

Oporto	mean	sd	(0.025,0.975) quantiles	non-standardised mean
(Intercept)*	-7.604	0.018	(-7.640, -7.570)	-7.8712
ageing	0.047	0.038	(-0.028, 0.047)	0.001
immigrants*	-0.062	0.022	(-0.105, -0.019)	-0.084
unemployment*	0.130	0.036	(0.060, 0.201)	0.025
overcrowding	-0.010	0.030	(-0.070, 0.049)	-0.002
homeless	0.023	0.019	(-0.015, 0.060)	0.006
hiv	0.018	0.014	(-0.011, 0.046)	0.004

Table 4.5: Fixed Effects for Lisbon and Oporto Metropolitan Areas; * refers to significant covariates

The final optimal model is compared with a Poisson generalised linear model without random effects. Under the latter, all covariates were found to be significant, see Table 8.5 in the appendix. The main differences are that the effect direction for immigrants and overcrowding was in the opposite direction for the Oporto data.

Having included a spatial and temporal random effects and a spatiotemporal interaction term, the only fixed effects that were statistically significant were the immigrant and unemployment effects in Oporto. The addition of the spatial, temporal and interaction random effects did however significantly improve the model fit compared to the model without these terms. Of note from these tables is that with the exception of the immigrant and overcrowding covariate effects in Oporto, the direction of the effects is concurrent with the TB literature: i.e. that incidence is higher in communities with

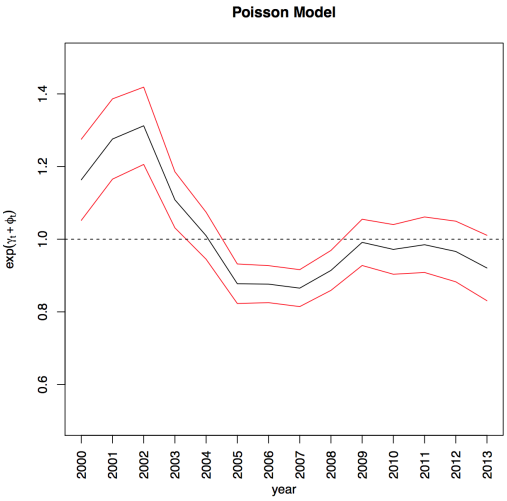
an older population, higher unemployment, a greater proportion of homeless people and where HIV incidence is higher. The later part of this chapter will discuss further the effects of the proportion of immigrants and overcrowding in Section 4.6.

Comparing the size of coefficients for the standardised variables shows that for Lisbon, the proportion of homelessness has the biggest impact on TB incidence, whereas the proportion of unemployed has the greatest effect in Oporto. Each of ageing, immigrant and unemployment have a bigger impact on TB incidence in Oporto compared with the Lisbon areas, whereas overcrowding, homelessness and HIV incidence have a larger effect in Lisbon.

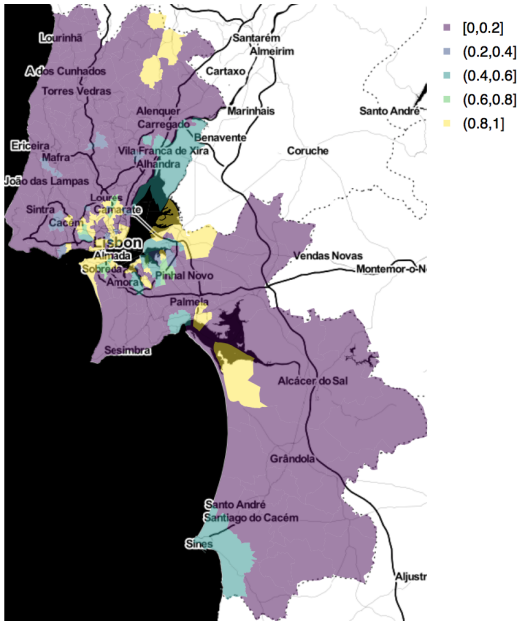
Figure 4.1a shows the global temporal effects of TB with the confidence bands for the mean between 2000 and 2013 for Lisbon. This shows a general decreasing trend from 2002 to 2007. Another slight increase was seen between 2007 and 2008 followed by a levelling off around 2009 with slight negative trend towards 2012/2013. The map in Figure 4.1b shows the posterior probability that covariate-adjusted spatial relative risk exceeds 1.5 (exceedance plots for thresholds 1.25, 1.5, 1.75 and 2 are given in the appendix in Figure 8.13). Areas where this probability is close to one are areas in which, having accounted for the ecological-level covariates, the relative risk of infection is highly likely to be greater than 1.5 times compared with areas in which the available covariates well explain the risk. In some sense, areas in which this risk is particularly high warrant further scientific study to determine the cause of the excess risk of disease. Areas of higher incidence compared to what we would predict based on our available covariates include freguesias near the central Lisbon city area and a few areas in the north.

The corresponding plots for Oporto are shown in Figures 4.2a and 4.2b (similarly to above, exceedance plots for thresholds 1.25, 1.5, 1.75 and 2 are given in the appendix in Figure 8.14). The global temporal trend has a similar pattern as for Lisbon, although in Oporto these patterns appear amplified: there was quite a steep drop in incidence from 2002 to 2009, but then followed a steep increase up until 2012. The map of spatial relative risk shows higher risk areas around the central area of the city and coastal areas north of the city; a large region in the central south of the metropolitan area was also identified as being at greater risk.

A dynamic presentation of space-time interaction can be found in Figure 8.9 and 8.10 in the Appendix. The high risk areas with probability of relative risk being greater than 1.5 are highlighted for different

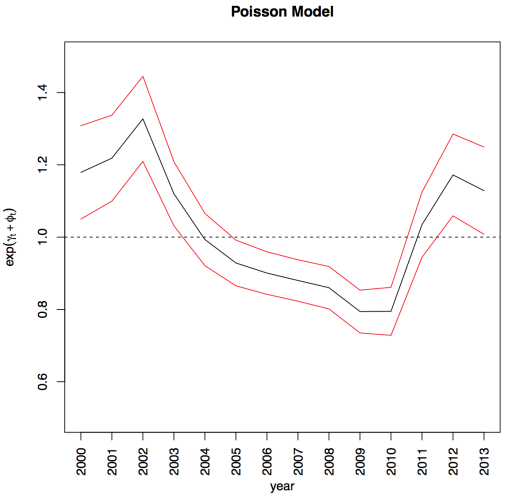


(a) Temporal Trend

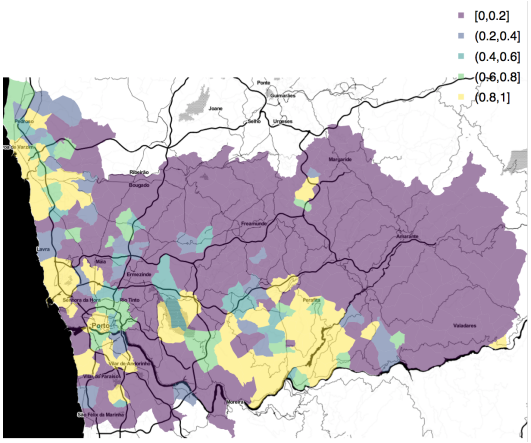


(b) Probability of Relative Risk of TB being greater than 1.5

Figure 4.1: Temporal and Spatial Effects: Lisbon



(a) Temporal trend



(b) Probability of Relative Risk of TB being greater than 1.5

Figure 4.2: Temporal and Spatial Effects: Oporto

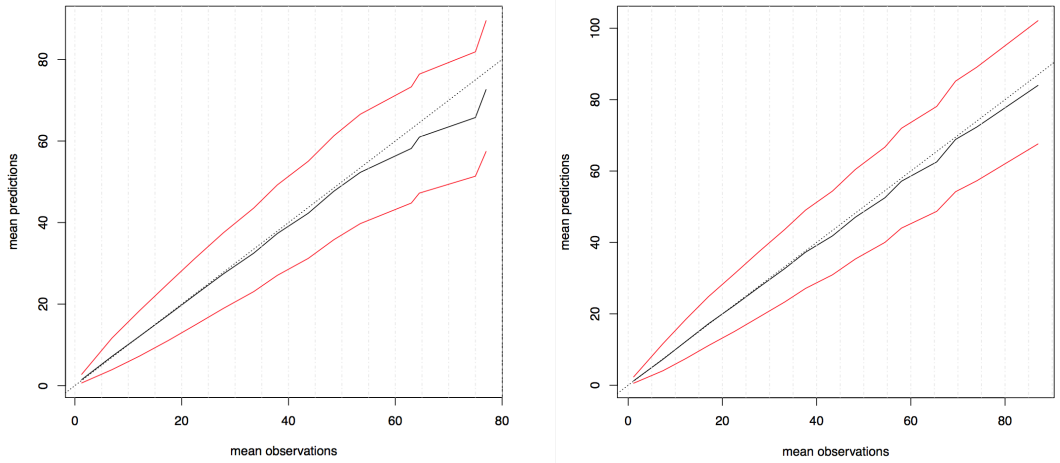
years in both Lisbon and Oporto Metropolitan areas.

TB incidences show a general decreasing trend in most of areas as suggested by the global temporal trends with a few areas showing increases in risk occasionally. For incidence, in northeastern Setúbal fluctuations can be seen between 2000 to 2011. In Lisbon areas, most of regions showed a declining

trend; with most regions experience less than 1.5 the relative risk. For Oporto regions, the interaction is less patterned. Between 2001 and 2007 more regions presented brighter colours than later on in the study. This follows similar trend to what the global temporal plot indicated.

4.4.1 Model Diagnostics

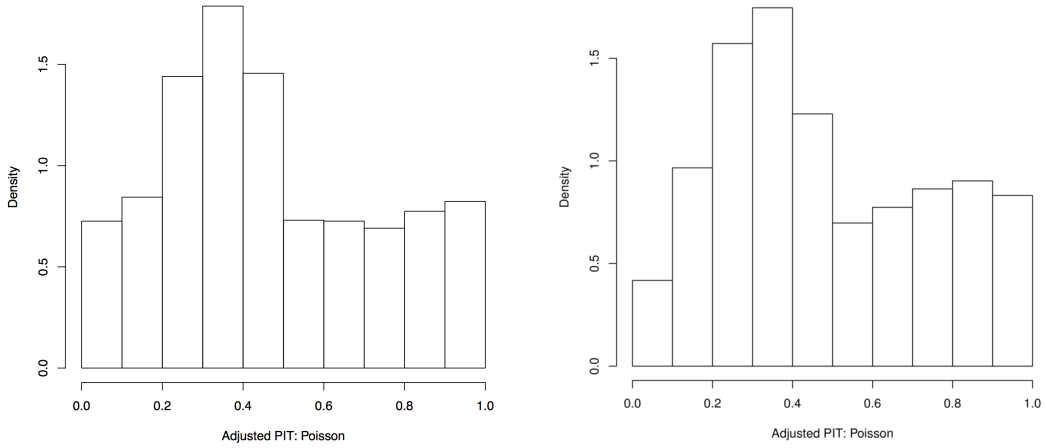
To examine model fit, the post proces of model fit produced binned plots of predicted versus observed values, these are shown in Figure 4.3. Sturges' formula yield 16 and 18 bins respectively for Lisbon and Oporto. The original plot of predicted counts against observed cases are shown in Figure 8.8 in the Appendix.



(a) Lisbon Metropolitan Area **(b)** Oporto

Figure 4.3: Binned mean predictions and counts with 95% confidence bands.

These plots show a good match of predicted to observed incidence, the mean being well within the 95% confidence band. The histogram of PIT is also provided in Figure 4.4. For appropriately fitting models, the PIT histogram should approximately follow a uniform distribution. Figure 4.4 shows a slightly greater number of samples taking values between 0.2 and 0.4 which suggests that for both areas, model predictions are over-dispersed (Czado et al., 2009). The PIT plots for the zero-inflated Poisson models are also examined, and these did not convey a significantly better fit to the data.



(a) Lisbon Metropolitan Area **(b)** Oporto

Figure 4.4: Adjusted-PITs for Lisbon and Oporto

4.5 Conclusions from Analyses

In this article we have used a spatiotemporal model with space-time interaction effects to capture the behaviour of annual TB incidence at the freguesia level in both the Lisbon and Oporto Metropolitan Areas between 2000 and 2013. The data supported the inclusion of this space-time interaction term over simpler models.

Despite the fixed covariate effects in our final spatiotemporal model not attaining statistical significance, observed effect directions mostly concur with the literature, see also the discussion below. Our results show that the inclusion of spatial-temporal random effects and interactions significantly improved model fit according to the WAIC. The large improvement in model fit in both study areas observed by including both spatial and temporal effects with interaction implies that although one can capture some of the variation of TB incidence using important ecological risk factors identified in other studies, there remains variation that we do not yet understand the source of, perhaps including impacts from other population data like poverty index, individual data and other health service data.

Two models have revealed some clear spatial clusters. The maps illustrating the probability of relative risk of TB incidences greater than 1.5 (Figures 4.1b and 4.2b) are areas in which risk is significantly greater than we can explain with our covariates alone. A table of the top ten highest risk areas

for each region is given in the Appendix (see Table 8.6). The areas identified for high relative risks are not necessarily high population density areas but are surrounding areas of poverty zones; this is especially true for Oporto areas.

The two temporal plots in Figures 4.1a and 4.2a show that risk increased after an initial rise following the period of the economic crisis between 2007 and 2009 and peaked for the second time around 2011. The incline of risk showed slightly later in Oporto than Lisbon; this could be because that the effects of economic crisis hit the south first. It will be interesting to monitor how this trend evolves in the coming years, especially in the face of emerging new strains of antibiotic resistant bacteria.

4.6 Discussion

This study is a population-based study over space and time that mainly makes use of socioeconomic census data measured at the ecological level. Both Poisson and Zero-Inflated Poisson models were fit including spatial, temporal and spatiotemporal interaction terms to the TB data and identified areas that appear to be at higher risk.

Due to limited temporal availability of socioeconomic covariates, linear interpolation was used in order to impute covariate information in years where there was not a census. The main assumption here is that, for example, the rates of overcrowding have evolved at a constant rate. In practice this assumption is not likely to be true as it is practically impossible that socioeconomic covariates evolve at a constant rate. This assumption may impose an artificial temporal trend to covariates; if true, it means that covariate such as homeless people would increase at this simple linear constant rate with respect to time. However neither for overcrowding, nor for the other imputed variables, do we consider this to be a significant defect in the modelling approach. The HIV data are available over time monthly, but are at a higher spatial level compared with the TB incidence data. The downscaled values of HIV are obtained by assigning the municipality level data onto the freguesia levels according to population weights. Again, we did not consider this to be a major weakness of our study (as this is the best data available), though it would of course be better to have finer scale HIV data. Since HIV positive individuals are much more susceptible to infection compared to non-infected individuals, this should potentially be captured when modelling where appropriate data is available. This may be one explanation for why the modelling in this analysis failed to find a statistically significant effect for

this variable in the models.

The national database of the TB surveillance system in Portugal is the gold standard for the recording of case data. it is assumed in the modelling that the capture rate is highly representative of the true reality and is recognised as a very good and efficient system with random missingness. That is saying even if there exist missing data, it is assumed that misreporting/undiagnosed cases exist at random as no better data base is available. However, registries vary in quality and completeness in times and areas and such assumption may result in over-or-under- estimations in incidences.Thus the modelling over behaviour of TB incidences is at best as what can be achieved using this data.

That the model and results are based on ecological data could be regarded as a weakness in this approach. It would be advantageous to include a marked point process with multivariate marks being the outcomes of individual-level covariates if individual-level data (including individual locations) are available. This would take care of the consideration such as older individuals are more susceptible to disease but it would substantially complicate the analysis. TB is an infectious disease that is transmitted between individuals when an infectious case and a susceptible individual are in sufficiently close proximity that the bacteria are able to invade the new host. Even if it were possible to include individual level covariates, this would still not capture this fundamental pathway of transmission and the consequent manifestation of disease incidence at the ecological level.

No covariate selection methods were used in the analyses used here during model building process. As the main interest for this thesis is on spatiotemporal modelling via INLA rather than identification of risk factors, each of the variables we considered has been identified previously in the literature as being related to the likelihood of acquiring infection. The six fixed effects: ageing index, immigrant proportion, unemployment proportion, overcrowding index, homeless proportion and HIV incidence are also reasonably orthogonal in each of the two regions (see correlation matrices in Tables 8.7 and 8.8 in the appendix), meaning that they each express a different dimension of risk. The correlation coefficient estimated in covariates may not be as meaningful considering the mass missing data interpolated artificially, but it showed that homeless and immigrants are the ones which potentially affect the magnitude of risk of TB incidences most in Lisbon and Oporto respectively. Further studies regarding risk factor identifications are required to make firm conclusions on these matters however. Although the analysis procedure did not perform covariate selection, it did use the WAIC to decide on which random effects to include into the model, which resulted in the selection of spatial, temporal

and spatiotemporal interaction effects and also unstructured (i.e. independent) effects.

The study presented a novel approach over TB incidences in Lisbon and Oporto Metropolitan areas in Portugal. The novelty shows in aspects including the employment of GLMM with INLA approximation over this set of data and it is at a finer geographical scale (freguesia level) comparing to other studies. The study revealed how spatial clusterings and temporal trends behave within cities and even regions in the cities; it zones into the problems into more details by looking at a lower administrative level.

Chapter 5

A New Multi-Way Spatial Survival Model with Applications in Long Term Cancer Studies

This chapter considers the analysis of spatially-referenced survival data collected from studies that occur over a long period of time, where interest includes modelling the changing patterns of survival prognosis. The main contribution of this chapter is that it introduces a novel approach on survival model which takes account of spatial information within survival data, survival times of individuals and the entering times at which individuals join the study. Specifically, this can be treated as a new spatial extension to the two-way survival model introduced by Efron (2002), or a two-way model extension to spatial survival models. Such method allows the intake of spatial information, survival times and a different scale of interest when doing analysis. It is possible that a lot of times, a different timescale is of interest together with the time-to-study time itself. This model presents as a 3-dimensional problem and avoids the need to choose a primary timescale when doing survival analysis. Background material and an introduction to this area can be found in Section 5.2. The novel modelling framework is proposed in Section 5.3 and applied to the Surveillance, Epidemiology, and End Results Program (SEER) data in Section 5.3, a discussion appears in Section 5.3.4.

5.1 Abstract

Population-based cancer registries usually collect data over an extended period of time e.g. decades. Individuals enter these registries when they are diagnosed with a cancer and their survival prognoses will depend on the effectiveness of available treatment regimens at the time of entry. Thus it is likely that changes in survival prognosis is not only a result of treatment quality or public health awareness but also as a result of biases associated with screening: over-diagnosis, lead time and length biases. Traditional survival analyses focuses on survival duration and the time of entry into the study is eliminated. Here, a Bayesian two-way spatiotemporal model that captures both real calendar time and survival time is introduced. Simulation studies supported that such model takes care of these types of data more naturally An application to the breast cancer data in New Mexico from SEER registry is also presented. This model is also easily extendible to take into account of multiple timescales. With the accessibility of R pacakge *spatsurv*.

5.2 Background Studies

Probably the simplest way to capture a hazard function that changes over time would be to use a Cox model to capture cohort effects i.e. using study entry time as a covariate in factor form. However, the assumption of proportionality constrains hazards for all covariate groups to be the same shape as baseline hazards. Such a strong assumption may not always the best way to model real data (Hemming and Shaw, 2005).

This section provides a review and summary to two modelling approaches which capture changing risks over time: dynamic survival models and two-way survival models.

5.2.1 Dynamic Survival Models

One way to deal with these problems is to extend the Cox model by including study entry time as a covariate and allowing the effect of this covariate to vary over time

$$h(t) = h_0(t) \exp\{X'\beta(t)\}.$$

Such models no longer satisfy the proportionality assumption.

Frequentist inference for this class of models can proceed by introducing penalty functions for neighbouring interval estimates of time-varying coefficients, which can be modelled as a discrete function of time. Other approaches include the use of spline functions with knots; for example, Gray (1992) applied both quadratic and piecewise constant splines for estimation of time-varying coefficients. Gustafson (1998) suggested a model which includes variation from constant covariate effects by using penalised likelihoods within a Bayesian hierarchical model. Gamerman (1991) appears to be the earliest application of dynamic survival models via the log-baseline hazard, including covariate effects as piecewise constant processes. Hemming and Shaw (2005) extended the class of Bayesian dynamic survival models where both the log-baseline hazard and covariate effects are modelled by piecewise constant and random processes, but their method is limited to right- and interval censored data. Sargent (1997) suggest a method of inference based on the partial likelihood to allow time-varying effects. Gamerman and West (1987) demonstrated the application of new dynamic Bayesian models for survival data in study of unemployment, where survival data are treated as a time series and covariates are time-varying.

Dynamic survival models also allow for the inclusion of spatially-correlated frailties alongside time-varying covariate effects: allowing us to analyse space-time survival data. Bastos and Gamerman (2006) considered spatial frailties in two separate terms, incorporating spatially-structured and also unstructured effects. The model proposed follows the form:

$$h(s,t;X,x) = \exp\{X'\beta(t) + Z + W(s)\},$$

where X is the matrix for covariates, $\beta(t)$ refers to the time-varying covariate effects, t denotes the duration time, s denotes spatial location, Z and $W(s)$ refer to unstructured frailty and spatial frailties respectively. The authors use MCMC to deliver inference. One main advantange of dynamic survival models is that they can provide exact hazard trajectories for individuals with a credible interval.

Other research in this area includes Lawson and Zhang (2011), who investigated the risk effects of prostate cancer in Louisiana using SEER data and an accelerated failure time (AFT) model with random effects (frailties). Lawson et al. (2010) proposed a mixture-based approach to describe overall level of risk with interaction between spatiotemporal latent effects. Lawson and Song (2010) proposed

a semiparametric survival model to investigate spatial and temporal pattern in chronic wasting disease in wild deer. Their model has the form

$$\lambda(a_i, t_i|X_i) = \exp\{h(a_i, t_i) + X_i\beta + u(s_i)\},$$

where $h(a_i, t_i)$ is the baseline hazard. β is the parameter with individual covariates X_i and $u(s_i)$ is the spatial random effects. The baseline hazard is allowed to vary with time t_i and over the age of deer a_i. Their model captures both temporal and spatial trends in disease via latent parameters under a Bayesian hierarchical set up, with inference performed using MCMC.

5.2.2 Two-way Survival Models

Cox and Oakes (1984) first stressed the importance of choice of time origin and scale in survival methods; such choice of time origin typically takes the start point of the follow-up times but other options include birth year of the patient and calendar time. There is no general consensus about the best choice of time scale, but this is often dictated by study aims. The use of calendar time as time scale allows the temporal effects of calendar entry time to be captured and modelled.

When changes in risk over calendar time is of interest in a study, it is possible to measure and include both calendar and survival time into the model. This be achieved using a symmetric approach: the two-way proportional hazards model (Efron, 2002). Two-way models are multiplicative hazard models based on two timescales (i.e. an additive decomposition of the log-hazard rate). Historically, the idea originated from the concept of the Lexis diagram in the 1870s (Keiding, 1990). Hazards can be visualised on a 2-dimensional plot using diagonal lines to represent the two-dimensional hazard rate. Two-way models combine the information from two separate 'one-way' models based on different timescales and thus provides a symmetric maximum likelihood approach that avoids the need of choosing a primary scale. One major difference from existing common way of handling calendar times such as person-years rate is that it captures the underlying baseline hazard under calendar time, rather than taking it in as a time-varying covariate.

Despite the concept of multiple timescales in survival models receiving mention in Cox (1972) and Farewell and Cox (1979), it was not until fairly recently that the two-way model was proposed. Efron (2002) demonstrated the use of the two-way proportional hazard model, incorporating the effect

of calendar date on bacterial infection based on a Poisson generalised linear model. He considers cubic polynomial expansions for the two baseline hazard functions with respect to survival time and calendar time, but in general, these underlying hazards could have arbitrary forms. Estimates from two-way models have been shown similar to estimates of proportional hazards on lifetime and calendar date scales respectively (Efron, 2002).

The choice of dual time scales is not limited to just calendar time and survival time, it allows for flexible choices, e.g. free inclusion of age and survival time or any other timescale of interest. The modelling of the hazard functions in its simplest form could assume piecewise constant hazards, in which the duration time is segmented and the hazard within each segment is assumed to be constant. This model is very flexible and especially suitable for grouped data (Han and Hausman, 1990; Murphy, 1996). Models with multiple timescales are shown to be advantageous if the survival data is affected by any intermediate events during the study (Keiding, 1990; Efron, 2002; Iacobelli and Carstensen, 2013; Rebora et al., 2015). Indeed these models have been shown to be preferred when estimating impacts of intermediate events on survival outcomes by Rebora et al. (2015), who also suggested that bias may arise from traditional estimates obtained by Kaplan-Meier estimators when there are interventions which may affect survivals occurring at different times from baseline observation.

Kauermann and Khomski (2006) pointed out that traditional one timescale models are justifiable when the study time is relatively 'short'; the assumption is then that study subjects are under the same hazard, having adjusting for covariates. However, this may not necessarily hold true when the study time is long. A model which takes into account of risks changing with survival time expressed in the hazard function as well as calendar time is needed; which the model of Efron (2002) fills – in the present chapter, we present a spatial extension of this model.

An additive alternative to Efron's model was proposed to fit unemployment data in Germany with penalised splines. The results showed that the hazard function varies with duration as well as calendar time. Iacobelli and Carstensen (2013) made an extension of using flexible parametric modelling instead of usual Cox-based approach over transition rates with one timescale in multi-state models applied to myeloid leukemia data. When the focus is on the variation of instantaneous risk in time and how such rate evolve dynamically, they argued that multi-time scale models such as two-way hazard models are more preferred for modelling transition rates.

5.2.3 Inference Method

For n observations of survival times $t_1, \ldots, t_n$ and their corresponding times between events being observed and start of study times $\tau_1, \ldots, \tau_n$, the likelihood function for these data has the form:

$$\pi(t_1, \cdots, t_N|\zeta) = \prod_{i \text{ uncensored}} f(t_i;\zeta) \prod_{i \text{ left censored}} F(t_i + t_0;\zeta) \times \prod_{i \text{ right censored}} S(t_i;\zeta) \prod_{i \text{ interval censored}} \left[F(t_i^{(2)};\zeta) - F(t_i^{(1)};\zeta)\right] \tag{5.1}$$

Observations of each type contributes independently to this likelihood function. It is then easy to write the log-likelihood as

$$\log \pi(t_1, \cdots, t_N|\zeta) = \sum_{i \text{ uncensored}} \log f(t_i;\zeta) + \sum_{i \text{ left censored}} \log F(t_i;\zeta) \sum_{i \text{ right censored}} [1 - F(t_i;\zeta)] + \sum_{i \text{ interval censored}} \log \left[F(t_i^{(2)};\zeta) - F(t_i^{(1)};\zeta)\right] \tag{5.2}$$

Given the forms of baseline hazard functions that we have proposed, the density function f has the form

$$f(t) = h_0(t) \exp\{X_i\beta + Z_i\} \exp\left\{-\exp\{X_i\beta + Z_i\} \int_0^t h_0(s)\mathrm{d}s\right\},$$

and the lifetime distribution is

$$F(t) = 1 - \exp\left\{\exp\{X_i\beta + Z_i\} \int_0^t h_0(s)\mathrm{d}s\right\},$$

where X_i is a row vector of covariates for individual i and β is a column vector of covariate effects.

A Metropolis-Hastings MCMC sampling scheme for spatial survival models is available in the R package *spatsurv*; this package allows user-defined baseline hazard functions, see Taylor et al. (2017) for details. We defined a two-way baseline hazard function by specifying the baseline hazard and cumulative hazard together with first and second derivatives with respect to the relevant parameters. We used functional programming to allow the user to input their own combination of functional forms for the baseline hazards. Note that certain combinations of baseline hazard functions can lead to non-identifiability. For example, two Weibull baseline hazards on time scales u and v, $\alpha_0\lambda_0 u^{\alpha_0 - 1}$

and $\alpha_1\lambda_1 v^{\alpha_1-1}$, would lead to a hazard function of the form

$$h(t) = \alpha_0\alpha_1\lambda_0\lambda_1 u^{\alpha_0-1}v^{\alpha_1-1}\exp\{X\beta\},$$

from which it can be seen that the λs are not identifiable; setting one of these equal to 1 yields an identifiable model. In our code, we therefore allow the user to fix any parameter in the model to any give n value. The R code for the two-way (in fact, multi-way – see below) baseline hazard function is available in the file *multiWayHaz.R* as part of the source code for the *spatsurv* package on CRAN.

5.3 Two-way and Multi-way Spatial Survival Models

In this section two new models that capture both temporal and spatial variation in the hazard are introduced. These models include a spatial extension of the two-way model discussed above and also a model based on Gaussian processes. This two way modelling framework can be furthermore extneded to consider multiple time scales. The later discussion is then focused on the two-way model because the Gaussian process model was computationally expensive to fit (MCMC sampling takes a long time to run till convergence).

The two-way spatial model is based on the following non-spatial model:

$$h_i(u,v) = r_0(u)s_0(v)\exp\{\alpha' x_0\},$$

where u and v can be any timescales of interest, but in the sequel, it focuses on the case where u is time-in-study and v is calendar time. r_0 and s_0 are baseline hazards based on different time scales respectively and x_0 are the covariates with coefficients α. The two-way model is thus an additive decomposition of the logarithm of the hazard $\log\{h_i(u,v)\} = \log\{r_0(u)\}+\log(\{s_0(v)\}+\alpha' x_i$. The two-way hazard function can be interpreted in a similar to the usual hazard function: as the instantaneous failure rate of at time (u,v) given that the individual is at risk of the event in an infinitesimal time interval just prior to (u,v).

Note that for the calendar-time / study-time example above, if the origin date of the study is b_i, on

one timescale, the other can be easily derived by the following relation:

$$v = b_i + u,$$

$$u = v - b_i.$$

Comparing to a standard proportional hazards models, two-way models avoid the need to choose a primary time-scale, allowing us to model risk as a function of both time-on-study and on calendar time, for instance.

The novel model: the two-way spatial survival model.

A novel spatial extension of the two-way model described above is introduced under this subsection. This is the primary model used in this Chapter.

Suppose that for n measured times to either event or censoring, event or censoring were observed at times $\tau_1 = t_1 + t_{01} > \ldots > t_n + t_{0n} = \tau_n$, where t_{0i} is the real calendar entering time for subject i, t is the survival time and τ is the calendar time. We define the hazard function to be

$$h(t, \tau, s; \beta, \omega^{(1)}, \omega^{(2)}) = h_0(t; \omega^{(1)}) h_0^{(c)}(\tau; \omega^{(2)}) \exp\{X\beta + Z(s)\},$$

where h_0 and $h_0^{(c)}$ are some baseline hazard functions for survival time and calendar time respectively with respective parameters $\omega^{(1)}$ and $\omega^{(2)}$. X denotes a vector of covariates with coefficient vector β. $Z(s)$ is the value of a spatially-continuous stationary latent Gaussian field at location s in space.

The cumulative baseline hazard can be derived from the following equation,

$$H_0(t, \tau; \beta, \omega^{(1)}, \omega^{(2)}) = \int_0^t \int_0^\tau h_0(u; \omega^{(1)}) h_0^{(c)}(v; \omega^{(2)}) du dv.$$

Thus, the two-way cumulative baseline hazard function has the form of $H_0(t, \tau) = H_0(t) H_0^{(c)}(\tau)$ where $H_0^{(i)}$ i.e. the product of cumulative baseline hazards for each of the timescales.

The two-way model model can be easily extended to a 'multi-way model' by including further time-

scales into the functional form of the hazard:

$$h(t_{1:d}, s; \beta, \omega^{(1:d)}, \omega^{(2)}) = h_0^{(2)}(t; \omega^{(2)}) h_0^{(1)}(t; \omega^{(1)}) \cdots h_0^{(d)}(t; \omega^{(d)}) \exp\{X(\tau)\beta + Z(s)\}.$$

However, the interpretation of such a model is challenging.

As a concluding remark to this section, it is superficially tempting to assume that the standard PH model with hazard, $h(t) = h_0(t) \exp\{X\beta + Z\}$ is nested within the two-way model $h(t, \tau) = h_0(t) h_0^{(c)}(\tau) \exp\{X\beta + Z\}$, by setting $h_0^{(c)}(\tau) = 1$. However this is NOT the case. The reason is that the likelihood function includes contributions from both the baseline hazard and also the cumulative hazard; and the latter in this case reduces to $\tau H_0(t)$, not $H_0(t)$ as it would have done under the standard model.

5.3.1 A Model Based on Gaussian Processes

The second type of model proposed in this Chapter acts as an alternative to the two-way model. It uses a transformed Gaussian process to model the baseline hazard according to one of the timescales. This model is only discussed here but not used as the main method in later part as it is based on MCMC sampling and the chain takes long to run and mix.

Under this model, the baseline hazard function takes the form:

$$h_0(t, t_0) = f(t; \omega_f) \exp\{g(t_0 + t; \gamma^{(t)}, \sigma^{(t)})\},$$

where $f(t; \omega_f)$ is a parametric, deterministic function with parameters ω_f function and $g(t_0 + t; \gamma^{(t)}, \sigma^{(t)})$ is a temporal Gaussian process and t_0 is the origin of the second time-scale. A spatial survival model would then have a hazard function of the form:

$$h(t, t_0, s) = h_0(t, t_0) \exp\{X\beta + Z(s)\},$$

where $Z(s)$ is a spatial Gaussian process. Inference from this model is challenging, partly because of the $O(n^3)$ problem, but it is possible to reduce this cost, straightforwardly to $O(m^3)$ by discretising the process, under the assumption that $m << n$.

The method is implemented this model, but in the examples we tested, the mixing was too slow for satisfactory use in real applications, hence we did not proceed with further investigation of this model for the present. A detailed derivation of the hazard, cumulative hazard and derivatives for this model are given in the appendix.

5.3.2 Simulation Study

A small simulation study is conducted in order to illustrate the inferential properties of this proposed two-way spatial model. The aim is to understand under which circumstances our model would perform better than a spatial model.

Firstly, 300 observations were simulated from a standard spatial parametric proportional hazard model using the *simsurv* function in the *spatsurv* package*. A Weibull model for the baseline hazard with both the shape and scale parameters being set to 1 respectively is applied. An exponential model for the spatial correlation in Z is assumed and set the marginal variance and spatial decay parameters both to 0.1. The design matrix includes randomly sampled 'age', 'sex' and 'cancer' indicator. Ages were randomly sampled from a uniform distribution on interval [5,50]; the probability of the study subject being male is 0.5 and the probability of each 'individual' having cancer is 0.2. All covariate coefficients were set to be 0.01. Below, the resulting simulated times are referred to as the 'survival times', as opposed to the 'calendar' times (which we refer to as 'entry' times below) - which were introduced artificially, as described below.

The following four scenarios are considered in this simulation study

- Scenario 1. Uniform entry times and survival times: artificial entry times were simulated from a uniform distribution an the interval bounded by the range of the survival times. This scenario represents the case where incidence and trends in survival remain constant over time.

- Scenario 2. Uniform entry times and (artificially) longer survival times compared to case 1. Survival times generated from the process described above *, were extended in a linear fashion using the relation

$$t_{new} = t_{original} \times (1 + 2 \times (\tau_1 - \min(\tau_1)/\max(\tau_1) - \min(\tau_1)))$$

i.e. unchanged at the 'start' of the study and extended by three times at the 'end' of the study. This scenario represents a situation in which disease incidence remains constant, but there are exogenous effects that modify survival times, e.g. 'treament' for the disease gradually improves over time.

- Scenario 3. Non-uniform entry times and the same survival times as case 1. Non-uniform entry times were simulated from a triangular distribution over the range of the survival times. This scenario represents a situation in which disease incidence increases over time, but treament for the 'disease' does not improve over time.

- Scenario 4. Non-uniform entry times and the same survival times as case 2. This scenario represents a situation in which disease incidence increases over time and treament for the disease improves over time.

The following models were fitted: (i) spatial survival models and (ii) two-way spatial survival models to the simulated data from the above four scenarios. All covariates are included in both the spatial and two-way models (i.e. age, sex and cancer presence/absence).

Arguably the simplest way to allow information on entry time to be included in spatial survival models is to include a 'cohort' factor variable as a predictor (simply as a covariate in Cox PH model) – this would partition calendar time into segments and each individual whose entry time fell into one of these segments would receive the same modification to their hazard. Although this is a computationally simple approach, the main disadvantage of this method is that we would expect the long-term effects of exogeneous variables (such as improved treatment), which may modify survival times, would generally do so in a smooth manner - i.e. the effects within segments should be temporally correlated.

An extension to this simple method would be to allow entry time to enter as a smooth effect, which can be achieved by using a spline function to represent changes in hazard over calendar time. In order to compare our two-way model with this 'simpler' alternative, it is assumed to follow a B-spline representation of the effect with respect to calendar time as additional covariates. This is achieved in R using the *bs* function from the package *splines*, which creates the additional columns of the design matrix required. Both models assume Weibull hazard for survival times and we used a B-spline hazard for calendar times in the two-way model.

Scenario	1	2	3	4
Spatial	369.0566	1310.61	327.1442	925.2546
Two-way	779.6715	1778.072	942.3362	346.682

Table 5.1: Table showing WAIC values from our simulation study.

The simulation study used the MCMC algorithm implemented in the package *spatsurv* with 500,000 iterations in total; a 10,000 iteration burn-in, thinning every 490th sample left us with a sample of size 1000. We ascertained convergence by examining trace plots of model parameters. Appendix 8.2.2 gives R code for the simulated data.

The WAIC values for each scenario 1 to 4 are in Table 5.1 (smaller ones indicate better model results). The results show that when entry times are uniformly distributed (senarios 1 and 2), a simple spatial survival model fit the data better. When non-uniformly distributed entry times are considered, but there are no exogeneous changes in surival over time (scenario 3) the spatial model also performs better. However, in Scenario (4), where incidence increases, but do do survival rates, the two-way survival model performs much better. Before running this simulation study, we expected the two way model to be the best performing in scenarios 2 and 4; our expectations were confirmed in the latter case, but not in the former - though it is worthwhile noting that in scenario the relative performance of the two way model was much improved over scenarios 1 and 3.

Based on this small simulation study (see WAIC values), we would expect our two-way spatial surivival model to perform better in scenarios where incidence is increasing over time (perhaps because of better methods of disease detection) and where survival from the disease, it having been detected, is also increasing due to exogeneous, unmeasured factors (e.g. improvements in drug development and treatment regimens). Such an example scenario in real life would be in the analysis of long-term cancer registry data, see below.

5.3.3 Applications

The Data

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute (Howlader et al., 2013) collect data on cancer cases from several states in the United States of America.

It is not openly available but can be easily accessed via `https://seer.cancer.gov/seertrack/data/request/` by signing agreements. The data covers the time period between January 1973 and December 2012. For the purpose of this part of the thesis, the focus 9s on modelling survival data of breast cancer from New Mexico, which was extracted from the SEER-9 registries.

Breast cancer can be found in both men and women. According to Cancer Research UK (CRUK), there are 55,000 new cases of breast cancer and over 11,000 deaths from the disease per year in. Though the UK can boast a 10-year survival rate of 78%, it has been estimated that 23% of breast cancer cases are preventable [1]. It has been established that the women who are over 50 are at higher risk of breast cancer compared to the general population and that mutations in the BRCA1 and BRCA2 genes (among others) also lead to a higher risk of the disease. Other risk factors include high levels of sex hormones, higher breast density, environmental exposures (e.g. ionising radiation), poor diet and low levels of exercise (National Cancer Institute., 2014).

In this chapter, survival time in months is used for each individual as the response variable. In the SEER data, survival is rounded to integer values - in order to break ties, the uniform noise is therefore added, so that times were on a continuum. Risk factors included as covariates in our model are: Sex, Marital Status at Diagnosis, Race Recode, SEER Historic Stage A, Behaviour Code ICD-O-3, First Malignant Primary Site and Age at Diagnosis. All covariates but Marital Status follow the definitions given in SEER documentation, see Table 5.2. The variables are randomly chosen based on what has been established in previous relevant studies. As it is not the purpose for this thesis to establish risk factors but introducing a two-way spatial hazard model, the precise selection of covariates is not included here. Marital status is regrouped into two levels; 1) married or with partner and 2) without partners, based on definition in SEER documentation. The inferences are based on a sample of 5000 out of around 30,000 individuals in order to reduce the run time of the MCMC.

The model

In this chapter, the following two timescales are of interest; survival time (duration time t) and calendar time (duration time between the observation of cases and the start of the registry on 01/01/1973;

[1] `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer`

	definition
Age at Diagnosis	age (in years) of the patient at diagnosis for this cancer.
Sex	the sex of the patient.
Marital Status at Diagnosis	identifies the patient's marital status at the time of diagnosis for the reportable tumour.
Race Recode	1). white, 2). black, 3). other (American Indian/AK Native, Asian/Pacific Islander) and 9). unknown
SEER Historic Stage A	a simplified version of stage: 0). in situ, 1). localized, 2). regional, 4). distant, and 9). unknown.
Behaviour Code ICD-O-3	The International Classification of Diseases for Oncology Third Edition (ICD-O-3) is the standard reference for coding the histology for tumours diagnosed in 2001 and later. SEER requires registries to collect malignancies: benign (0), borderline Malignancy (1), in situ (2), Malignant (3), only Malignant in ICD-O-3(4), no longer reportable in ICD-O-3 (5) and only Malignant 2010+ (6).
First Malignant Primary Site	based on all the tumours in SEER. Tumours not reported to SEER are assumed malignant. 0). no and 1) yes.

Table 5.2: Summary of Categorical Variables. Note: these are relevant variables chosen based on existing studies; the precise selection of covariates is not included here as this work intends to introduce statistical model and show application rather than identifying risk factors for Breast cancer.

denoted by τ). Recall the hazard function for our model:

$$h_0(t, \tau; \omega) = h_0(t; \omega^{(1)}) h_0^{(c)}(\tau; \omega^{(2)});$$

$$h(t, \tau, s; \omega, X, \beta, Z) = h_0(t, \tau; \omega) \exp\{X\beta + Z(s)\}.$$

The baseline hazard with respect to survival time is captured under $h_0(t)$, assuming a Weibull model. The baseline hazard with respect to calendar times τ is captured under $h_0^{(c)}$ with a B-spline baseline hazard model. The baseline hazard function thus has the following form:

$$h_0(t, \tau) = \sum_{i=1}^{d} a_i B_i^d(t) \times \alpha\lambda t^{(\alpha-1)}, \tag{5.3}$$

where $\{B_i^d\}_{i=1}^d$ is the B-spline basis of degree d; and we chose $d = 1$ (i.e. linear) in this case. Note that under the parameterisation in Equation 5.3, the parameter λ is not identifiable; to resolve this issue, we fix $\lambda = 1$.

Gaussian priors are assumed for, $\beta, \omega \sim \mathcal{N}(0, 25)$ and used an exponential function to describe spatial covariance, $\text{cov}[Z(s), Z(s + x)] = \sigma^2 \exp\{-||x||/\phi\}$. The New Mexico dataset is spatially referenced at the county level, so we modelled covariance between counties using distance between centroids.

Covariate Names	Estimated Median	LB	UB
	50%	2.5%	97.5%
Sex (Male)	2	1.29	3.19
Marital Status (No partners)	1.21	1.12	1.33
First Malignant Primary Site (0)	1.42	1.28	1.54
Historic Stage 0	0.878	9.51×10^{-5}	456
Historic Stage 2	1.83	1.68	2.02
Historic Stage 4	11	9.38	12.4
Historic Stage Unknown	3.13	2.48	4.06
Race Black	1.21	0.85	1.63
Race Other	1.06	0.799	1.35
Race Unknown	0.659	0.154	2.08
BEHO3V (2) baseline (3)	0.436	8.5×10^{-4}	4082
bs(AGE, df = 3) 1	0.35	0.241	0.464
bs(AGE, df = 3) 2	2.18	1.34	3.35
bs(AGE, df = 3) 3	30.5	17.8	48.1
lambda1	9.05×10^{-8}	2.87×10^{-8}	1.79×10^{-7}
lambda2	3.62×10^{-6}	3.31×10^{-6}	4.25×10^{-6}
lambda3	8.33×10^{-6}	7.38×10^{-6}	9.92×10^{-6}
lambda4	1.04×10^{-5}	8.27×10^{-6}	1.29×10^{-5}
lambda5	3.32×10^{-5}	4.77×10^{-7}	1.5×10^{-4}
alpha	1.12	1.09	1.15
sigma	0.18	0.157	0.223
phi	9964	6517	14708

Table 5.3: Estimates for Fixed Effects, Baseline Hazard Parameters and Spatial Covariance Parameters in Breast Cancer at 3 significant numbers; here, bs(AGE, df = 3) refers to the b-spline fitted age at diagnosis with degrees of freedom 3. Note: LB is the lower bound of 95% confidence band, UB is the upper bound of 95% confidence band upper bound

Our priors for σ and ϕ were log-Gaussian, with the former set to $\pi(\log\sigma) \sim \mathcal{N}(-2, 0.2)$ and the latter to $\pi(\log\phi) \sim \mathcal{N}(0, 10000, 0.2^2)$. This choice of prior for ϕ places little support outside 1/5 of the size of the observation window, and though it does not allow our method to detect long-range spatial effects, this is not a problem because such effects would require a larger observation window in order to be able to identify them.

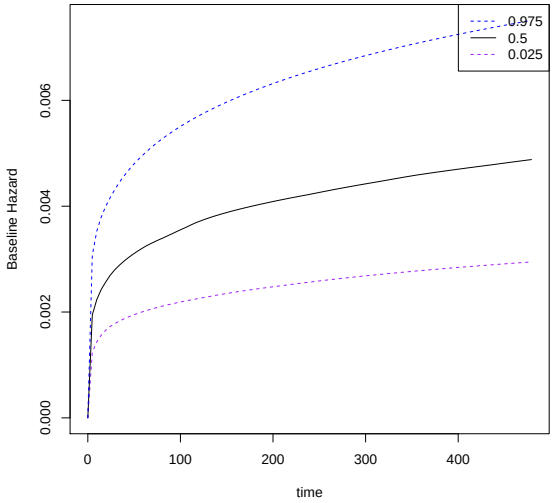
This chapter compared the estimated baseline hazard with an estimate from a Weibull spatial survival model with the same prior specification.

Results

In this section, the results of the standard spatial and two-way spatial models for breast cancer survival in New Mexico are presented. Discussions are given on the estimated effects of chosen covariates on risk, compare estimates of the baseline hazard function and map exceedance probabilities of covariate-adjusted relative risk.

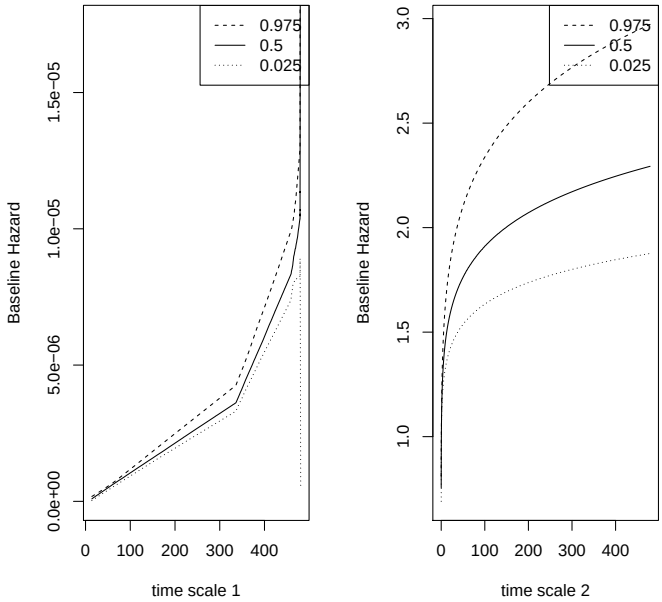
Table 5.3 gives parameter estimates ad 95% credible intervals for all parameters in our model. Being male, not being married, no first malignant primary site and Historic stages 2, 4, and unknown were all associated with a significant increase in the risk of death. The remaining variables, though they had a reduced risk were all non-significant in our model. Estimates of the parameters of the baseline hazard and the spatial covariance function are at the bottom of Table 5.3. The estimated medians refer to the estimated coefficients with respect to each covariate in the model β's as well as coefficients of baseline parameters (λ's, α, σ and ϕ). The positive estimates refer to positive correlation at log level of the model; increase in all risk factors included in the model result in a high hazard rate in survival times. These are given with 95% confidence intervals. The spatial decay parameter ϕ showst the spatial correlation can be neglected when regions are 9964 kilometres apart and this indicates how strong the spatial correlation is. The interpretation of λ?s is hard but these can be seen as the estimate which fits the curvature of hazard in real calendar timescale (see left of Figure 5.1b).

The two-way spatial baseline hazard with respect to different timescales (real calendar time t_1 and duration time t_2) are shown in Figure 5.1b and the same model with only spatial effects included is given in Figure 5.1a for reference. Both plots are presented with medians and 95% credible intervals. The baseline hazards for both timescales showed increasing trends, with different features.



(a) one-way baseline hazard

]



(b) two-way baseline hazard

Figure 5.1: Baseline Hazard for Breast Cancer in New Mexico; Left: real calendar time scale; Right: survival time scale

Figure 5.2 shows the probability that covariate adjusted relative risk exceeds 1.1 over the study area i.e. $P(\exp\{Z(s)\} > 1.1)$. In brief, brighter colours refer to higher probability of the event (death in

this case) being observed. Red areas give rises of warnings to greater spatial frailties of concerns. The figure shows that there appears to be an elevated risk in counties in the south of the state. In this study, south bound sees more bright colours (red and orange), it means that in this regions the hazard for breast cancer is higher comparing to the baseline.

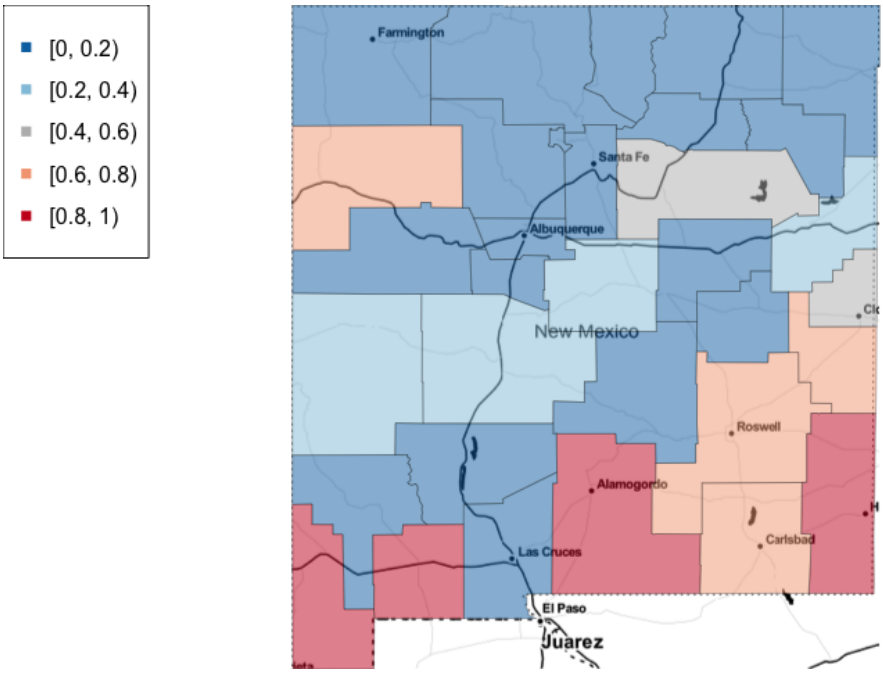


Figure 5.2: Relative Risk of Breast Cancer in New Mexico

MCMC Convergence and Other Matters

Figure 5.3 shows the worst mixing chain, ϕ, which appears to have converged and is mixing well, according to the autocorrelation of the chain.

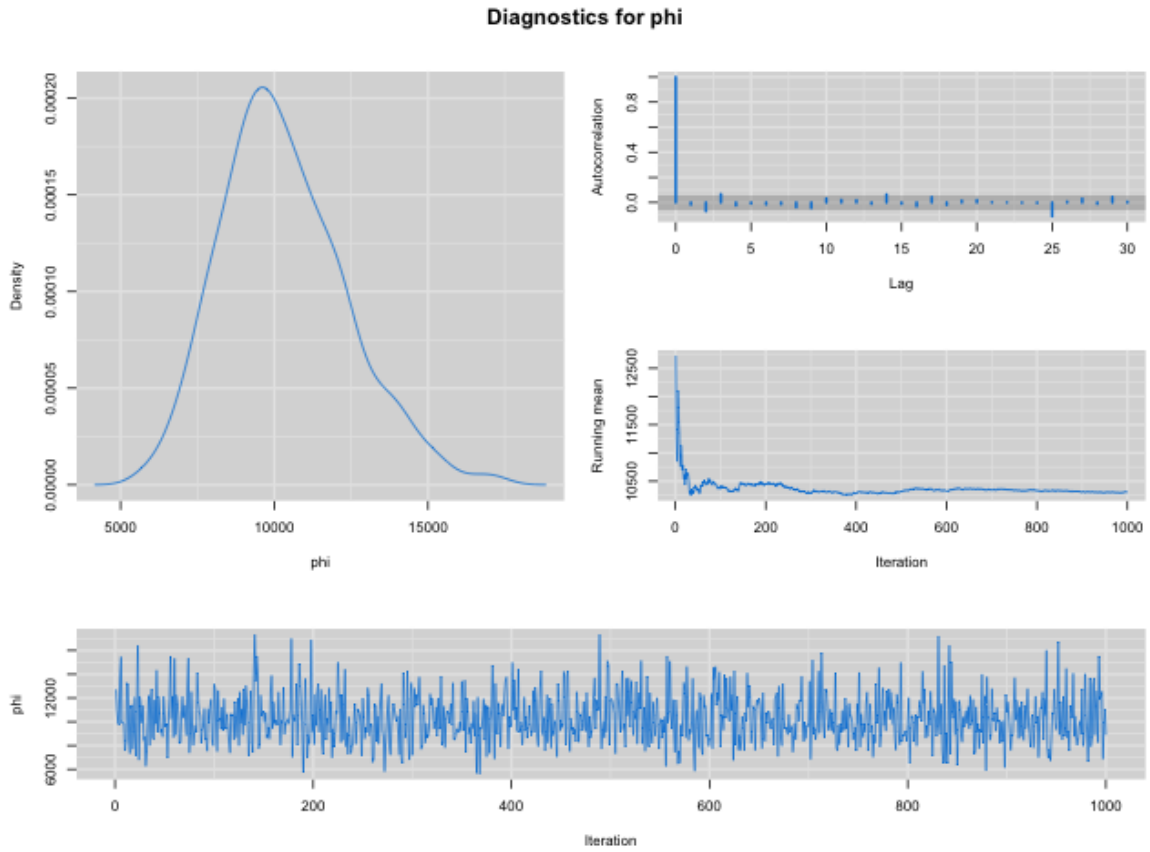


Figure 5.3: MCMC mixing plot

Despite the larger WAIC value being shown in the two-way model (56939) compared to the pure spatial survival model (29317), the two-way model still allows us to observe the trend of baseline hazards for real calendar time and survival time separately; see Figure 5.1. We observed a slightly increasing trend in hazards with respect to survival time for breast cancer in New Mexico. The two-way model shows similar trend on the survival time scale but on a different scale (note it is not straightforward to compare the scales of these plots). The increasing trend with respect to calendar time is most likely caused by increasing registration rate in breast cancer, likely through improved awareness of the disease in the population.

5.3.4 Conclusion and Discussion

This chapter introduced some novel two-way and a multi-way spatial survival models which measure disease risk with respect to two or more time scales, and to our knowledge, this is the first such spatial survival model of its sort. It has demonstrated how this model can be used to evaluate hazard with respect to calendar time and time-in-study, and also identify spatial patterns in data. The application effectively showed the temporal trend with respect to both real calendar time of entries and survival times in Breast cancer together with the spatial random effects. Different increasing patterns in hazards were shown with respect to real calendar times and survival times and relative risk for higher hazard rate is seen down on the south boarder of New Mexico. Although in the application, the two-way model did not give a better WAIC value compared to the purely spatial model, in concept, it still makes better sense to assume a non-constant hazard across the whole study period when the study covers a long time scale. This is also supported by the simulation studies conducted (see Section 5.3.2).

This two-way model allows one to explore the influence of exogeneous factors on survival with respect to a secondary timescale. In comparison to standard 'one-way' models, the method adjusts the estimate of the usual baseline hazard for unmeasured factors with respect to the second scale. In a standard survival analysis, for example of an acute condition over a period of time, standard pracctice would be to zero all individuals from their date of diagnosis. In comparison, the two-way model respects the fact that data are accumulated over time and in that sense, treats the data in a more natural way. Comparing to methods which include person-years rates and (or) Lexi's diagram, the strongest advantage is that the method treats dual timescales symmetrically. Another aspect is that two-way model can be easily cooperated with spatial survival models (both in theory and available R package). This also makes the implementation of two-way spatial survival models easier in practice. One potential area this model could be extended, and likely improved, would be use a 'non-separable' hazard function i.e. a baseline hazard of the form $h_0(t, \tau)$ instead of imposing the relationship $h_0(t, \tau) = h_0^{(1)}(t)h_0^{(2)}(\tau)$ as was the case in this article. Future research will explore this area in more detail c.f. Lawson and Song (2010).

Another facet of the data captured by the two-way model in application is the increasing detection rate over time. The assumption of a constant hazard rate over a long period of time is not always realistic, especially when, as is the case for cancer data, detection rates are higher in more recent

times due to better technology, the presence of large scale databases and also public awareness.

The models for spatial-correlated frailties assume log-Gaussian forms and inference is done via advanced MCMC methods. One of the limitations for the model here is that some sort of parametric modelling has to be used for either time-scales. When fitting these models using simulation-based methods, choices of priors can potentially affect the model outcomes. Other criticisms such as high computation costs on MCMC also apply here, though we would argue that it took years to accumulate the real data and the relatively small run times of MCMC code is justified in order to obtain unbiased joint inference from the posterior. To accommodate time-varying features better, it is also possible to extend our models to allow time-varying covariates and indeed parameters in the future.

Chapter 6

Deprivation and Pregabalin Prescribing in England

6.1 Deprivation and Pregabalin Prescribing in England

6.1.1 Abstract

Objective

The aim of this paper is to seek to understand spatial and temporal trends in pregabalin prescribing and the relationship with deprivation across England at both General Practice (GP) and Clinical Commissioning Group-level.

Design

Milligrams of pregabalin prescribed and dispensed per 1000 population as the response variable, weighted IMD (Index of Multiple Deprivation), geographic location and time as predictors. The set of active prescribing facilities grouped within Clinical Commissioning Groups (CCG) as the units of analysis.

Setting

National Health Service (NHS) prescribing data: all GP practices in England, UK between January 2015 and June 2017.

Population

England, UK.

Methods

Gaussian generalised additive models with fixed and random effects were used. The CCGs are treated independently and analyses were carried over each CCG separately. Within each CCG, an intercept and weighted IMD were included as fixed effect and a random slope to capture temporal trend at GP-level was employed. The model also allows spatial correlation between practices using a 2D spline surface.

Results

Adjusting for deprivation, we show a North/South divide (just above London, see figures) in terms of prescribing trends, with the North of England showing increasing prescribing rates during the study period on average, while in the south of England, rates are on average decreasing. There were no apparent spatial patterns in baseline prescription rates at the CCG level. Weighted IMD score proved to be statistically significant in 138 / 207 CCGs. In two thirds of CCGs, there was more pregabalin prescribed in areas of greater deprivation. We present tables of the top ten highest average prescribers at the CCG level and of the top ten highest rates of increase at the GP surgery level.

Conclusions

The spatial temporal modelling demonstrated that the North England have a significantly higher chance to see increase in pregablin prescriptions comparing to the South adjusted for weighted IMD. Weighted IMD has shown positive impact on pregabalin prescriptions for 138 CCGs.

6.1.2 Introduction

This study takes place amid wide concern over the misuse of pregabalin/Lyrica for recreational use and a global increase in prescribing rates at the UK level (Suardi et al., 2016).

Pregabalin is a gabapentinoid and is a close analogue of the neurotransmitter GABA (γ-aminobutyricacid). It does not work by directly binding to GABA receptors but it is capable of crossing the blood-brain barrier and it potentiates GABA effects. It was developed as a successor to gabapentin and eventually made it to market in 2004 with Pfizer (Ben-Menachem, 2004). They held the initial patent but generic versions are now available in the UK while it remains under patent in the USA.

In the UK pregabalin is currently indicated for peripheral and central neuropathic pain, as an adjunct of therapy for focal seizures in epilepsy, and for generalised anxiety disorder (BNF, 2018). Pregabalin is recommended by the National Institute of Health and Care excellence (NICE) as a first-line treatment option for peripheral neuropathic pain alongside amitriptyline, gabapentin, and duloxetine (NICE, 2013) . It has been noted that the use of pregabalin has extended beyond its license into other chronic pain conditions (Stannard, 2014). A recent systematic review and meta-analysis of the use of gabapentinoids in chronic low back pain has highlighted that there is very limited evidence of effectiveness while significant risks of adverse effects have been demonstrated (Shanthanna et al., 2017).

In the UK, prescribing of pregabalin has risen steeply with an increase from 368,512 prescriptions in January 2015 to 527,704 in June 2017 (https://openprescribing.net/chemical/0408010AE/). Clearly this rise has had financial implications with a significant increase in associated prescribing costs. However, along with the other gabapentinoid, gabapentin, there has been increasing concern about misuse of pregabalin and its potential for abuse. Typically abuse will involve taking large doses, well above therapeutic levels, to achieve a euphoria effect.

Initial reports of problems with pregabalin and gabapentin were anecdotal (Spence, 2013; Schifano, 2014; Bicknel, 2013). More detailed evidence of problems is now emerging. A systematic review of the abuse and misuse of pregabalin and gabapentin Evoy et al. (2017) found that increasing numbers of patients are using these medications and taking them to achieve euphoric highs. In the general population prevalence of abuse was estimated at 1.6% but amongst populations such as opioid abusers prevalence has ranged from 3 to 68%. The systematic review also highlights evidence that the gabapentinoids are being identified in post-mortem toxicology analyses and are playing a role in overdose deaths (Evoy et al., 2017). In the UK there were four deaths linked to pregabalin in 2012 but this rose to 111 deaths in 2016 (Office for National Statistics (ONS 2016) (Office for National Statistics, 2016).

In October 2017, BBC reported 88% and 83% higher spending on pregabalin prescriptions in North East and North West England respectively compared to London (Wainwright, Wainwright). The reasons for this north/south divide are not known, but there are well known socioeconomic inequities and differences in health outcomes along this axis (Buchan, Kontopantelis, and Sperrin, Buchan et al.). There has also been concern about prescribing rates in Northern Ireland following a BBC

Three documentary on the issue (Patterson, 2017).

In the United Kingdom, the Advisory Council on the Misuse of Drugs (ACMD) recommended in January 2016 that pregabalin and gabapentin should be reclassified as Class C drugs with tighter legal controls over their prescribing.

Using NHS open data it has been demonstrated that it is possible to map prescribing at various healthcare levels down to individual GP practices (Rowlingson et al., 2013). The ability to map pregabalin prescribing, taking into account deprivation and other potential covariates, in order to analyse prescribing trends is useful for setting prescribing policies. In this article we seek to address the following research questions: (1) What are the temporal trends in prescription rates at the GP practice level? (2) How do prescription trends vary across the country at the practice-level and at CCG level? (3) Which areas in the UK are prescribing more pregabalin compared to the national average? (4) what is the nature of the relationship between the prescribing of pregabalin and deprivation?

6.1.3 Methods

Data Sources

The monthly amount of pregabalin prescribed at the GP level is freely and openly available from DATA.GOV.UK and can be extracted from the files available from the GP Practice Prescribing Data website (`https://data.gov.uk/dataset/prescribing-by-gp-practice-presentation-level`). Data on the number of people registered at each GP is also available from this source, and is broken down geographically into the number of patients registered to the practice by the 2011 Lower-layer Super Output Areas (LSOAs, available from `https://data.gov.uk/dataset/lower_layer_super_output_area_lsoa_boundaries`). This breakdown is updated on a quarterly basis. CCG boundaries for April 2017 were obtained from The Open Geography Portal of The Office for National Statistics (&+#http://geoportal.statistics.gov.uk/datasets/clinical-commissioning-groups-april-2017-full- clipped-boundaries-in-england-v4). The English Indices of Deprivation 2015 are available from GOV.UK (`https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015`). The prescribing data are used as the primary data to be analysed. Geographical shapefiles are used to overlay to the original data; this allows one to decide CCGs of which the observations belong to.

This part of the thesis focuses on the number of milligrams prescribed per registered population as our unit as , using British National Formulary code 0408010AE (BNF, 2018) to identify prescriptions of relevance. Total number of milligrams obtained by multiplying dosage in each prescription and total number of prescription. As this is not a fair comparison if registered popuation to GP is higher, it is then scaled by dividing total registered population to the GP. Note that the analyses did not consider oral solutions in this thesis, but these only account for a small percentage of prescriptions (less than 1% on average).

Weighted Index of Multiple Deprivation (IMD) Score

In order to explore the relationship between deprivation and prescribing rates (per population registered) for pregabalin, a method for assigning a deprivation score to each GP was required. One option for this could be to use the deprivation score for the Lower Super Output Area (LSOA) in which each GP practice is located. However this method may not be a good reflection of the deprivation of patients registered at that practice.

Practices typically accept patients from a relatively small number of LSOAs. Using the IMD of each of those LSOAs and the fraction of the total practice list from each LSOA (ie. for GPs receiving patients from 1, ..., n LSOAs, the weighted IMD score is calculated by $\sum_{i=1}^{n} \frac{\text{IMD}_i}{\text{total number of GPs in LSOA i}}$), a simple weighted IMD score that attempts to capture the IMD of patients that attend each practice is constructed.

Statistical Methods

In order to answer the proposed research questions, Gaussian generalised additive models (Wood, 2017) with fixed and random effects were used. The outcome was the square root of the number of milligrams prescribed per population. Such square root transform was used because the model fit was significantly improved (according to the REML score).

Analyses were conducted on CCGs separately. Within each CCG, an intercept and weighted IMD (defined above) were included as a fixed effects and a random slope with respect to time at the GP-level is also included. In addition, the models allowed for spatial correlation in prescribing rates

between practices by including a spatial effect which was modelled using a 2D spline surface.

Let P_{ijt} denote the number of milligrams of pregabalin prescribed for GP i in CCG j in month t, where that GP has coordinates in the British National Grid (OSGB) projection (x_{ijt}, y_{ijt}). Our model assumed the following form:

$$\sqrt{P_{ijt}} \sim \mathcal{N}(\mu_{ijt}, \sigma^2_{ijt}),$$

$$\mu_{ijt} = X_{ij}\beta_j + \alpha_{ij}t + S(x_{ijt}, y_{ijt}),$$

$$\sigma^2_{ijt} = \frac{1}{n_{ijt}}.$$

Here, X_{ijt} includes the fixed effects for GP i in CCG j, α_{ij} captures the temporal trend in prescriptions for each GP within the CCG, $S(x_ijt, y_ijt)$ is the spatial effect and n_{ijt} is the number of patients registered to GP i in CCG j at time t.

The analyses also considered models with subsets of the domains and subdomains of the IMD (identified through backward selection) as predictors. The domain of education, skills and training has the corresponding subdomains skills and training children and young people subdomain, adult skills subdomain. Barriers to housing and services domain include geographical barriers subdomain, wider barriers subdomain, living environment, indoors subdomain and outdoors subdomain. Domains and subdomains are themselves highly correlated; this may be caused by overlapping in definitions and it is possible that some variables can be treated as causes of others. For example, low skills in education may result in higher barriers to housing and services. These models proved difficult to interpret, both from an epidemiological perspective but also the variables themselves are highly correlated. Definitions of relevant subdomains are neglected here as these are not the main model considered here but they can be found at `data.gov.uk`.

The inclusion of the spatial effects was justified on the basis that their inclusion significantly improves model fit for all CCGs apart from NHS Crawley CCG, NHS Eastbourne CCG , NHS Isle of Wight CCG and NHS North Hampshire CCG.

6.1.4 Results

The results are based on analyses over all GPs (7921) between January 2015 and July 2017 over England and Wales. The models are constructed for each of the 207 CCGs independently.

Temporal Trends in Prescribing

The temporal trends in prescriptions from the model (α's) capture whether prescriptions are increasing over time. Since there is an associated uncertainty in our estimates of each of these trends, rather than mapping these trends directly, which would ignore the uncertainty, we instead map the probability that the trend is positive.

Accordingly, when interpreting the Figures, the following should be borne in mind. The closer the mapped value is to 1, the more likely the prescription rates are to be increasing with time; values close to 0.5 are surgeries where the prescription rates do not change significantly with time; and values close to zero represent surgeries with decreasing prescription rates over time.

Figure 6.1 shows the probability of increasing prescription rates at each of the GP surgeries over England and Wales. Careful examination of this plot suggests a north/south divide: with surgeries in the North tending to have an increasing rate of prescriptions whereas in the South, prescription rates seem to be falling. In order to make these trends more clear, we averaged these probabilities by CCG and mapped them in Figure 6.2. Gray areas remain flat prescription trend and blue areas show decreasing trend. The interpretation of this latter figure is similar to that for Figure 6.1.

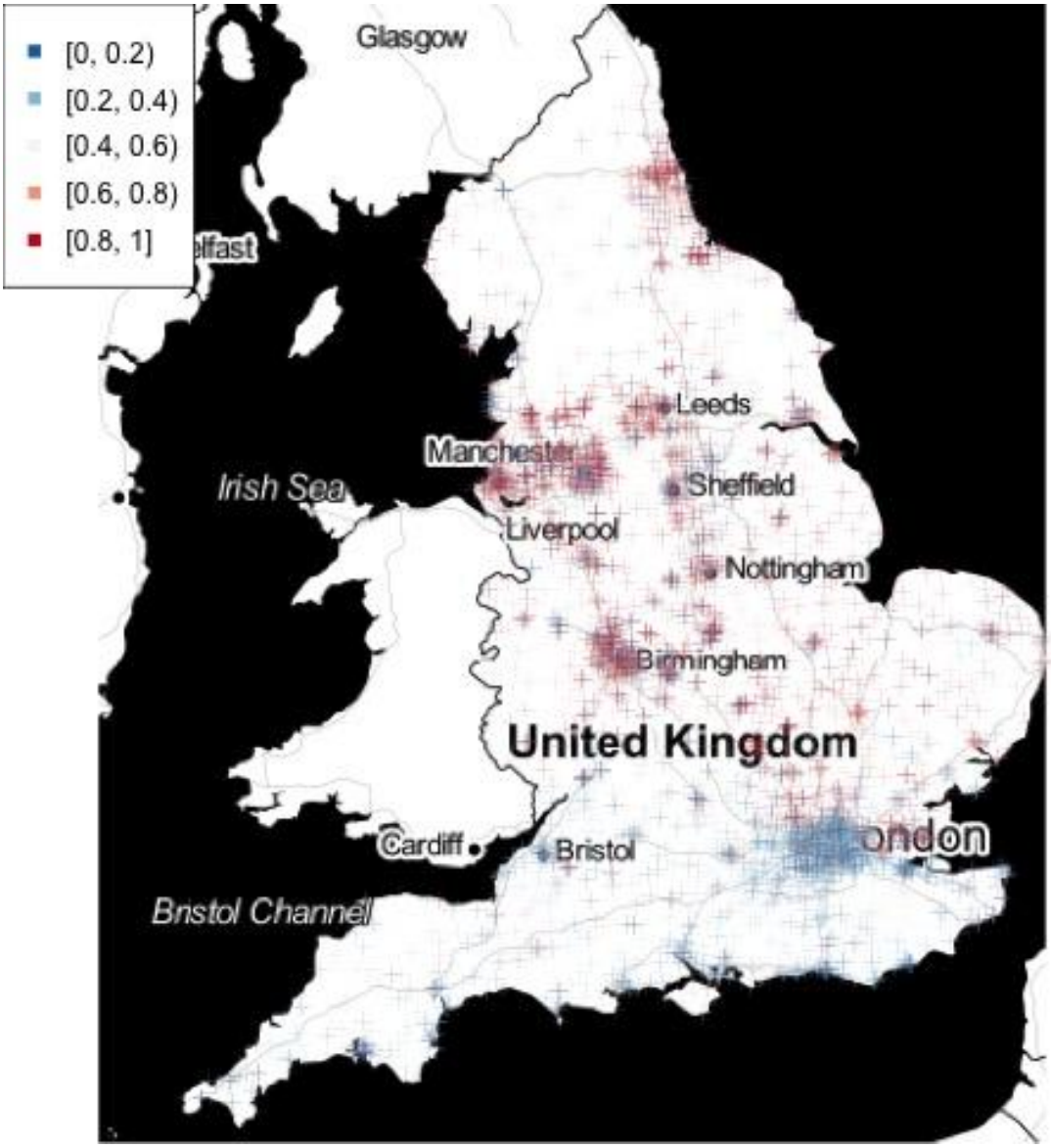
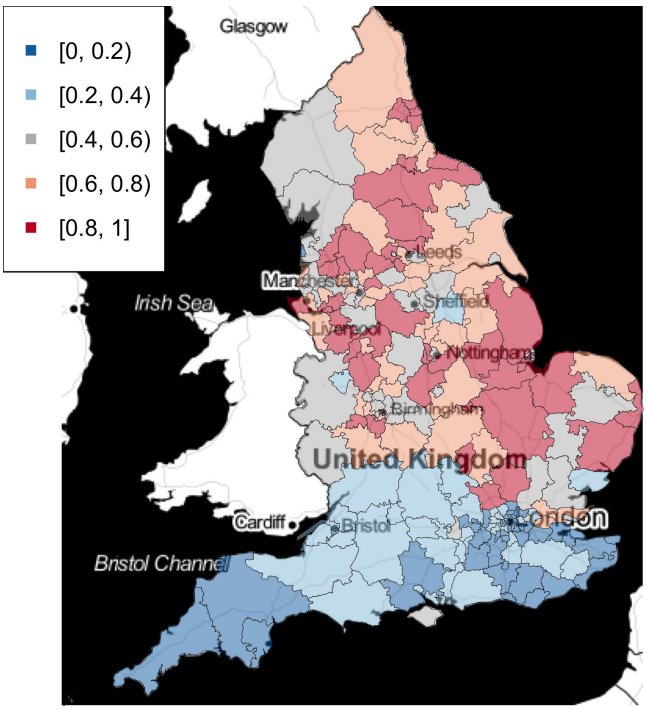
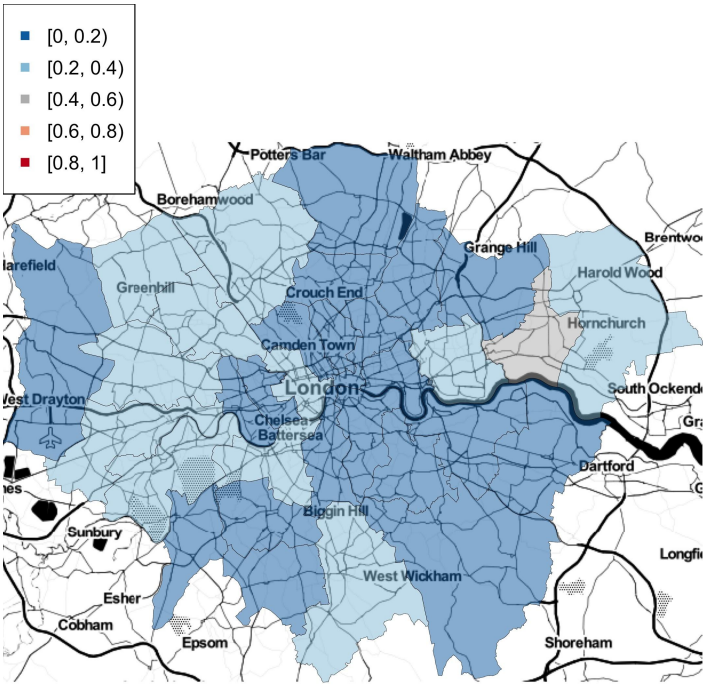


Figure 6.1: Probability of an increasing pregabalin prescription rate for each individual GP surgery

Averaging by CCG makes the north/south divide easier to see with northern CCGs in general showing increasing trends in pregabalin prescription rates compared to the southern CCGs. A more detailed map for the London CCGs is shown in Figure 6.2b. Interpretations of this map follow the same manner as previously and all CCGs in london showed decreasing trends in pregabalin prescription rates.



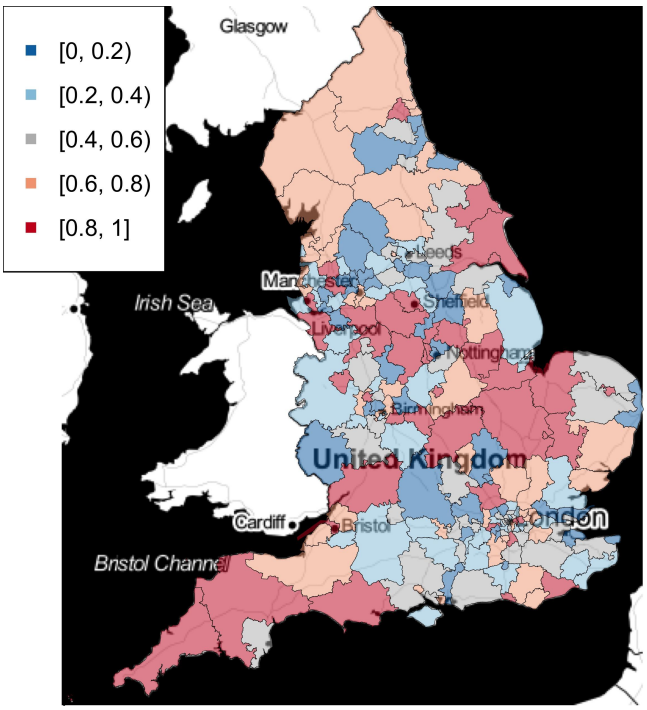
(a) England



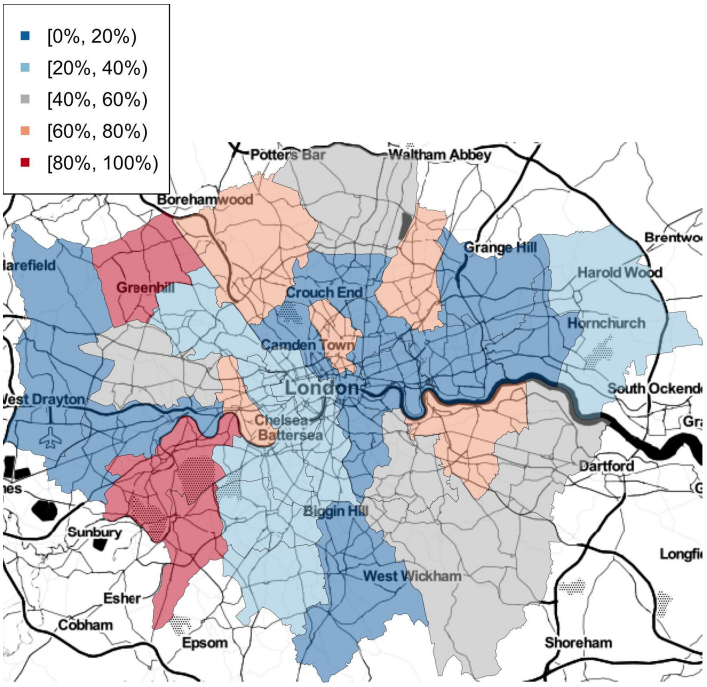
(b) London

Figure 6.2: Average probability increasing pregabalin prescription rates within each CCG

The estimated IMD-adjusted baseline prescription rates can be mapped from the model (the intercepts); these are shown in Figure 6.3. Rather than plotting the raw rates, which would be difficult to interpret, maps here illustrate the percentiles. For example, adjusting for IMD values, the red CCGs are the ones which lie in the top 20 percentiles at baseline level. The zoned in map for London is on the right and it can be interpreted in the same manner.



(a) England



(b) London

Figure 6.3: IMD-adjusted baseline prescription rates in percentiles for each CCG

The top ten GP surgeries with the highest estimated increase in pregabalin prescribing are given in the Table 6.1 below together with their 95% confidence intervals.

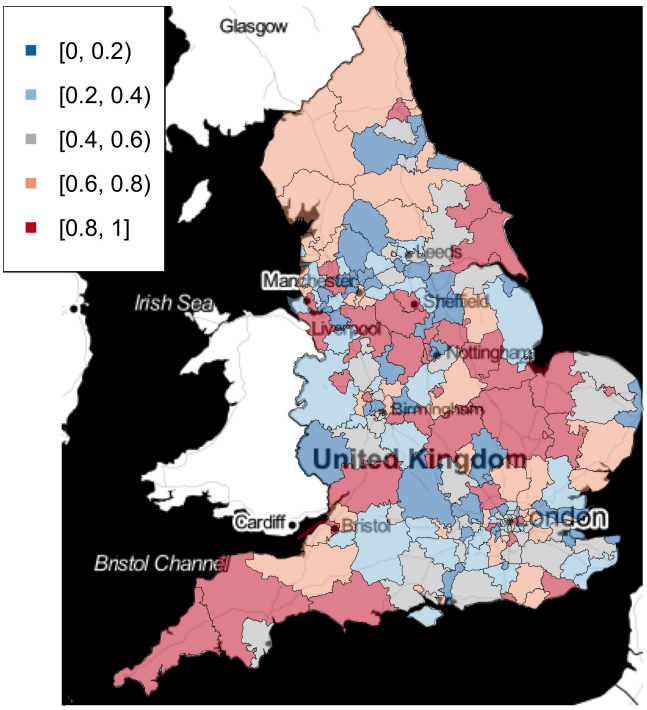
Rank	GP Code	GP Name	CCG	Estimated time slope	95% CI
1	Y02873	Compass Health	NHS Bristol CCG	4.03	(3.851, 4.206)
2	J84602	Garfield Road Surgery	NHS Isle of Wight CCG	3.35	(2.837, 3.873)
3	F85680	THE 157 MEDICAL PRACTICE	NHS Haringey CCG	1.99	(1.628, 2.353)
4	Y02155	TRENT VALLEY SURGERY	NHS Stoke on Trent CCG	1.87	(1.797, 1.942)
5	Y02045	VERNOVA HEALTHCARE CIC	NHS Eastern Cheshire CCG	1.77	(1.621, 1.929)
6	Y01262	PALLION PRIMARY CARE SERVICES	NHS Sunderland CCG	1.67	(1.618, 1.728)
7	Y00081	SAFE HAVEN UNIT-LCD	NHS North Kirklees CCG	1.44	(1356, 1.515)
8	Y00054	THE HOMELESS HEALTH CARE TEAM	NHS Gloucestershire CCG	1.20	(1.081, 1.328)
9	C81629	CLARENCE ROAD SURGERY	NHS Southern Derbyshire CCG	1.09	(0.992, 1.190)
10	Y01924	SAWBRIDGEWORTH MEDICAL SERVICES	NHS East and North Hertfordshire CCG	0.99	(0.948, 1.029)

Table 6.1: GPs with the highest estimated temporal slopes and 95% confidence intervals (CI)

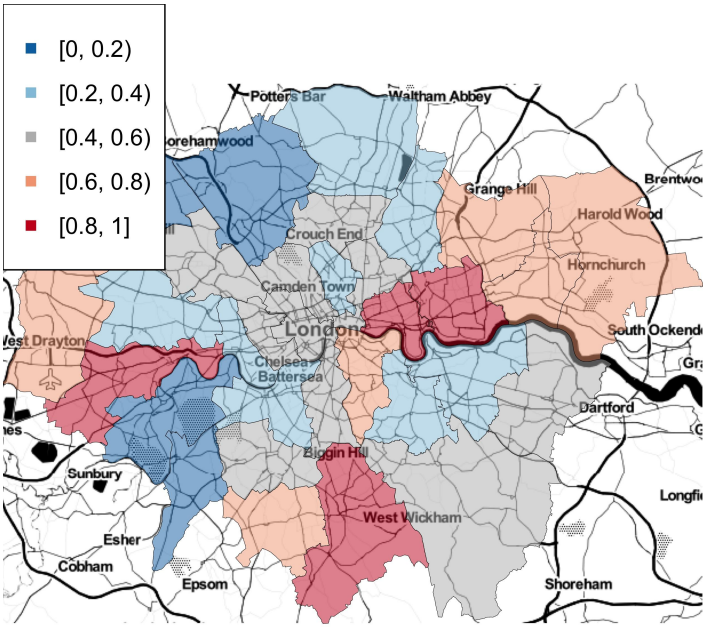
Relationship and Deprivation

The coefficient of weighted IMD in the model shows how, within each CCG, the relationship between pregabalin prescription rates and deprivation changes. A positive coefficient means that as deprivation increases, pregabalin also increases, whereas a negative coefficient means that as deprivation increases, pregabalin prescription rates are decreasing. Since there is uncertainty in the estimates of the effect of deprivation on pregabalin prescription rates, it is again mapped, in Figure 6.4, the probability that the coefficient is positive. Brighter areas (probability > 0.5) present positive correlation between deprivation and prescription rates. The grey areas with probability 0.5 are likely to show no positive correlation and blue ones with probability below 0.4 tend to show negative correlation.

Two thirds of CCGs (138/207) showed a positive impact of deprivation on milligrams of pregabalin prescribed by each GP.



(a) England



(b) London

Figure 6.4: Map of probability of relationship between weighted IMD and pregabalin prescription rates being positive for each CCG

The following table (6.2) shows the coefficient of weighted IMD on prescriptions of pregabalin for the top ten average prescribers, together with their p-values in brackets and adjusted R-squared statistic (a measure of model fit):

CCG	population weighted IMD	Global Intercept	R-sq (adjusted)
NHS Doncaster CCG	-0.0521 (0.037)	10.011 (¡0.001)	0.705
NHS Scarborough and Ryedale CCG	0.844 (¡0.001)	-13.008(¡ 0.001)	0.866
NHS Wirral CCG	-0.005 (¡0.001)	7.326 (¡0.001)	0.722
NHS Canterbury and Coastal CCG	0.170 (0.061)	4.8796 (¡0.002)	0.879
NHS Vale of York CCG	0.08 (0.028)	5.904 (¡ 0.001)	0.724
NHS Nottingham West CCG	0.264 (¡0.001)	2.389 (0.003)	0.668
NHS Morecambe Bay CCG	0.078 (¡0.001)	5.131 (¡0.001)	0.756
NHS West Hampshire CCG	0.068 (0.018)	6.045 (¡0.001)	0.287
NHS North East Lincolnshire CCG	-0.075 (0.689)	8.299(¡0.001)	0.901
NHS Northern, Eastern and Western Devon CCG	0.135 (¡0.001)	-5.123(¡ 0.001)	0.522

Table 6.2: Model estimates for CCGs with highest average pregabalin prescriptions per population between Jan. 2015 and Jun. 2017; p-values are shown in brackets.

6.1.5 Discussion

This chapter showed how pregabalin prescription rates for each individual GP surgery changed over time having effects of weighted IMD accounted for. There are around 60 per cent of GP surgeries estimated to have an increasing trend in pregabalin prescriptions, most of which are north of London (see Figures 6.1 and 6.2). Since the models adjusted for weighted IMD when fitting data, the clear north-south divide in time trends shown is not a result of higher deprivation index in the north comparing to south. There are some more underlying causes for this which remains unknown and requires further study.

Time trend described in this paper merely describes what the data for each individual GP surgery represent. It does not reflect in any sense the quality of performance on their prescriptions. That is to say, an increasing trend in GP surgery does not mean that the GP surgery is performing poorly in terms of pregabalin prescriptions. Even with weighted IMD being adjusted for, there are a lot more other factors which may result in increasing pregabalin prescriptions. For instance, a change in nature of population of GP surgeries can result in higher pregabalin prescription rates. These estimates are also only useful at describing the temporal trend shown in pregabalin prescriptions for

each individual GP surgery. They are not suitable for interpretation of the pregabalin quantities prescribed. For example, Blackpool CCG showed neutral probability of increasing prescription (see Figure 6.2), it however is in the top 20 per cent for baseline prescription rates (see 6.4).

To what extent should one raise alert for further investigation for CCGs or GP surgeries with high estimated increasing rate shall be combined with the baseline prescription rates (shown in 6.4). For instance, top 20 per cent high baseline prescriptions can be seen in CCGs across Yorkshire, Nottinghamshire and Lincolnshire. More highlights also lie in the northwestern part of the country; mainly Blackpool, Lancashire and CCGs to the North of Manchester. Other Northern CCGs with high starting rate are around Durham and Newcastle. Down in south, the high starting prescriptions mostly concentrate around CCGs around Essex and Kent; together with some CCGs in Herefordshire. Two of the London CCGs also appeared in the top 20 prescription rates.

With these born in mind, Table 6.2 only shows the top ten GP surgeries with highest estimated increasing rate in prescriptions; it does not conclude any performance quality related to these GP surgeries. In other word, it should not be interpreted that higher increasing rate refers to over-prescriptions. Eight out of these listed GP surgeries are located to the North of London, which is in line with the results seen in 6.1 and 6.2. The raw data for the top tens showed higher pregabalin prescription rates than the national 90th percentile calculated for all years considered in the study; apart from the Garfield Road Surgery and The 157 Medical Practice, which are above national average based on June 2017 data and below the 10th percentile for all years respectively. Temporal trend on raw prescription data for these practices have shown at least moderate incline in the time period considered. Causes for such increasing behaviour of each of these GP surgery however remain uncovered and require further study. Most of these ten GP surgeries lie within the CCGs with baseline prescriptions above national average.

The analyses have also shown how deprivation is related to pregabalin prescription in this article. 138 CCGs are estimated to have positive correlations between weighted IMD scores and pregabalin prescriptions. That is saying that for these GP surgeries in these CCGs, the more deprived the areas are the more pregabalin is likely to be prescribed. It is not clear what the patterns are for these estimates on map, but the highly positive areas are mostly located in Cornwall, Devon, Midlands and North Yorkshire; with a few CCGs in London. Unfortunately, the weighted IMD scores are not standardised; it is thus hard to make comparisons across the country between CCGs.

One of the limits to this study is that it cannot address the cause of the space-time behaviours of pregabalin prescription. It simply shows that such divide exists adjusting for deprivation and potential temporal trends can be seen. The study focuses on how one can utilise NHS open data to investigate issues of interest (pregabalin prescription behaviour in this case) and is only able to bring to sight what the issues are and points out which ones may seem outstanding and perhaps requires more insight to. Other questions such as what caused the prescription behaviour to act this way are not revealed by this study. It successfully reveals the north-south divide of prescription even adjusting for deprivation and spatial random effects; also showed that correlation between deprivation and pregabalin prescription is not necessarily positive.

On comments of cause of such spatial divide in prescription or why certain GPs showed increasing trend and a lot higher prescriptions, more information concerning confounding variables is useful. For example, if high risks of development symptoms requiring pregabalin are associated with older population, then the high prescription rates in some CCGs may be a simple reflection of population age structure. Further studies could be carried out on this issue by combining patients data such as age structure of the patients receiving this treatment within the same GP and what symptoms or disease they are diagnosed of. It is also crucial to address the potential that the findings on increasing prescription rate may be a reflection of correction to previous over/under prescribing.

6.1.6 Article Summary

The chapter focuses on how NHS open prescribing data can be used to investigate the relationship between practice-level prescribing of pregabalin and deprivation. The study employed GAMM modelling approach in each CCG at GP level accounting for space-time and covariate effects. This method allowed fast approach in this modelling and successfully demonstrated the possible spatiotemporal trends accounting for deprivation. The key message delivered here is that the North-South divide in pregabalin prescription behaviour still exists even after adjusting for deprivation. It also showed the possible relationship between deprivation and prescriptions; not necessarily all CCGs showed positive correlation. One of the limits in this part is that the assumption of independence between models for each CCG however may be validated in reality. The individual models across CCGs also mean that model covariates are hard to compare. Fitting one model for whole England avoids the independence assumption and allows comparisons through each CCGs but this takes way too long computationally.

Chapter 7

Conclusion

In this chapter, the main contributions of this thesis will be emphasised and some potential extensions for further study will be discussed and outlined.

7.1 Application of INLA to Spatio-Temporal Disease Data

The first contribution of this thesis was to provide a different way of exploring the annually registered TB incidence rates at freguesia-level in Lisbon and Oporto Metropolitan Areas. We have proposed a spatio-temporal model which takes account of spatial random effects, temporal random effects, spatial-temporal interactions and unstructured random effects as well as covariates relevant to TB. This model allowed understanding of how different ecological factors have effects on annual tuberculosis incidence rates at a finer geographical scale than existing studies. **R-INLA** provides a fast and readily available method for statistical inference. The use of state-space models allows exact updates at each time point, but is limited by the assumption of a Gaussian response variable and comes at high computational cost. The inclusion of both spatial and temporal effects with interaction shows that there remains some other unexplained variation on top of the important ecological risks (see Chapter 4).

Clear spatial clusters are identified by our models. Areas with high relative risks do not necessarily have high population density, but are mostly the areas surrounding poverty zones. This feature is

more apparent in the Oporto metropolitan areas. Temporal trends for both areas showed an initial peak between 2007 and 2009. Despite some of the covariates not being significant, our model identified positive correlations between TB incidences and ageing index, immigrant proportion, unemployment proportion, overcrowding index, homeless proportion and HIV incidences. These mostly concur with previous literature (see Chapter 4).

One possible cause for covariates not being significant could be the limited availability of temporal socioeconomic covariates. The assumption of certain covariates evolving at a constant rate over study period is very unlikely in reality. Finer socioeconomic data could potentially improve the model results, but is not of major concern. Another possible improvement would be the inclusion of individual-level covariates; it is not unreasonable to assume that certain individuals are more vulnerable to the disease, for example, older individuals may be more likely to get infected.

One extension to this INLA approach is a Freguesia-dependent model, which allows variation of borders of each freguesia to accommodate changes in administrative regions. Taylor et al. (2017) suggested a new method which makes inference over registered count data on aggregated units: instead of making inference based on conditional autoregressive models (CAR), they argue that discrete models should be more appropriate for treating discretely observed data arising from spatiotemporal continuous (point) processes. A method which delivers continuous inference for this type of data is provided.

7.2 Multiway Models on SEER Data

To our knowledge, there has not yet been any such application of hazard models considering two timescales as well as spatial random effects. We have introduced this novel method in this related chapter, which successfully combined different aspects of uncertainty into one modelling process, allowing changes in hazard rates over the time of study especially useful in studies conducted over a long period of time. The application on SEER data showed different behaviours between hazards with respect to real calendar time and survival time. Hazards showed a slight increasing trend for breast cancer in New Mexico with respect to both timescales. Relative risk of deaths being observed in breast cancer appears to be higher mainly along the south boundary of New Mexico. An increasing trend was observed in hazard with respect to real calendar time scale. This does not necessarily

reflect that individuals are subject to higher hazards in more recent times; the slight increasing trend is likely to be caused by higher registration rates in the study period.

Despite WAIC values not showing better results in application comparing to pure spatial models, the assumption of non-constant hazard across the long study period still makes more sense. It allows us to observe unmeasured factors with respect to the second scale and allows us to see both scales. It is not always sensible to treat the hazard rate over a long period of time as constant, especially when detection rates are higher in more recent times. This is the case for the SEER data. Our two-way model allowed us to see both time scales on top o spatial-correlated frailties.

One of the limitations for our model here is that the modelling over at least one timescale has to be parametric. This means that when fitting the model, choices of prior will to certain extend affect the model outcomes. High computational costs on MCMC simulation is also of criticism here, although the run times could be argued to be relatively small comparing to how long the study period is. It is worth the wait to obtain an unbiased joint inference from the posterior.

In certain scenarios, two-way models can be extended to multiple timescales.For example, inclusion of age of patient may be interesting for disease studies as elderlies are likely to be more vulnerable to some diseases. It could be of interest to see baseline hazard rates on age-timescale standing alone as well as the time-to-event times and calendar times. The multiple timescales models could potentially capture different information over real calendar times, survival times and age from birth by including a third timescale in to the baseline hazard. It is also possible to extend our models in the future, to allow time-varying covariates that capture the time-varying features in the study, both within risk factors and temporal effects.

A limitation of our model here is that some sort of parametric modelling has to be used for either time-scale. When fitting these models using simulation-based methods, choices of priors may affect the model outcomes. High computation costs on MCMC also deteriorate the efficiency of application of such models.

7.3 Spatio-Temporal Modelling of GP Prescription Data

We have also demonstrated how the use of Gaussian generalised additive mixed models can be informative when analysing open GP prescribing data. The method considered a set of models at each CCG unit over pregabalin prescribed by each general practice unit. Instead of using the costs or quantities of pregabalins prescribed, we used milligrams of prescription by each GP per month. 207 independent Gaussian generalised additive models with random slope with respect to time and spatial random effects were fit. These CCGs cover the whole NHS service areas of England and Wales. Each model takes into account a temporal random slope at GP-level as well as a fixed intercept and weighted IMD effects. It also allows a spatial correlation factor between GP practices using a 2D spline surface within each CCG.

This Chapter contributed to the existing studies concerning GP prescription rate of pregabalin in England. On top of the discovery of a clear North-South separation of pregabalin prescription, our method has highlighted that such an apparent division still exists even after covariate effects indicating levels of deprivation are adjusted for. More than half of GP surgeries are estimated to have an increasing pregabalin prescription rate, and this increasing trend is not a result of increasing deprivation level, as it was adjusted for when fitting the model. Deprivation is correlated to pregabalin prescriptions; with 138 CCGs showing positive impact. This to some extent supports a possible cause (ie. poverty) of the North-South divide put forward in other studies.

The assumption of independence between each CCG is not necessarily appropriate. By treating the whole of England as a unit, it allows us to capture spatial correlation with a 2D spline surface between CCGs as well. Then, possible effects caused by different CCGs could potentially also be captured by including CCG as a covariate. Such a model would avoid the assumption of independence between CCGs. However, the biggest challenge of this model is the high computational cost due to the large number of GP practices (ie. location points).

7.3.1 Overall Comment

Overall, based on classic spatial statistical theories, this thesis showed three novel approaches on spatiotemporal modelling of health data. It successfully presented a novel application of GLMM via

INLA inference on TB data in urban Metropolitan areas in Portugal. This showed the existence of spatial clusters within Lisbon and Oporto Metropolitan areas; high relative risk areas tend to be surrounded by poverty zones. The temporal trend exists with peak around 2007 in both areas. Further studies to including either individual level data or other more detailed (in time or space) socioeconomic data can be employed to discover the potential cause of these behaviours. The modelling process considering exact updates (eg. Kalman Filtering) has seen difficulties in inferential procedures including long process to run and mix; some more need to be done to improve this, such as better determination of priors. Other models including individual-level data may also be useful (see Chapter 4 for detailed discussions). The thesis also successfully introduced a hazard model which takes care of multi-timescales as well as spatial random effects. The two-way (or multiway) model shows advantage in the sense that it avoids the choice of a primary timescale. Such approach also allows the inclusion of an extra timescale when necessary and is easily accessible with self-defined functions in *spatsurv* package. The simulation studies showed that this approach takes care of the data in a more natural manner. Although the real data did not show better results comparing to standard one-way spatial survival analysis model, it is still more intuitive in concept that with long registries, real calendar time of when individuals enter the study should be accounted for as part of the baseline hazard. The use of two-way models in *spatsurv* is easily extensible to multiple timescales; interpretation may be hard. The final part of the thesis showed that the north-south divide in pregabalin prescriptions exists even when the impact of deprivation is considered and adjusted for. The correlation between deprivation and pregabalin prescription rates is not necessarily positive as previously expected by media. It also showed how statistical modelling methods can utilise these big open data source and assist learning and understanding of health concerns. Although improvement could be done by fitting one model over England and Wales as whole because independence is likely to be validated between CCGs in reality. More could be done following the instructions of such clustering behaviours by combining with other data source such as age structures of patients to make more inference about the potential cause of the patterns observed.

Chapter 8

APPENDICES

8.1 Appendix to Chapter 4 (TB in Portugal)

8.1.1 Exploratory Analyses

Transformation of incidences

To fit in the linear Gaussian assumption of statistical models described in this section, the log-transformed TB incidence is considered through out. Note that the data transformations only apply in the preliminary study models and do not affect the main modelling in the previous section.

For preliminary models, we focus on the monthly notifications at municipality level (a higher geographical unit than frequesia). As there are a lot of zero-notification areas in both Lisbon and Oporto Metropolitan areas, to be able to do log-transforms of data accommodation of zeroes are required. Here, we replace these by the overall mean incidence of each region i at time t. For y_{it} the log-incidences and N_{it} the counts, define N^*_{it} to be

$$N^*_{it} = \begin{cases} N_{it} & \text{if } N_{it} \neq 0 \\ \frac{1}{T}\sum_{t=1}^{T} N_{it} & \text{otherwise} \end{cases}$$

where $t = 1, \ldots, T$ are the time indices and $i = 1, \ldots, n$ are the indices for municipalities. This is done as TB is endemic and there is no strong obvious seasonality in Portugal. Then define the log-incidence (per 100,000) as

$$y_{it} = \log\left(\frac{N_{it}^*}{P_{it}/10^5}\right)$$

where P_{it} is the population of region i at time t.

Such reformulation of zeros in the data is used not only for its simplicity but also to utilise the information provided in dataset. Although mean value is a reasonable estimate for the observation, every imputation means that total number of TB cases in region i within the time frame is increased by its mean.

An alternative way of dealing with zero counts was also considered. Instead of using mean values, a random number is generated from some uniform distributions (Stanton et al., 2014). Thus similarly to the above, now define

$$N_{it}^* = \begin{cases} \text{Unif}(N_{it} - 0.5, N_{it} + 0.5) & \text{if } N_{it} \neq 0 \\ \text{Unif}(0, 0.5) & \text{otherwise} \end{cases}$$

y_{it} stays the same as above. This method assigns random values to zero counts. As they are drawn from a uniform distribution over interval $(0, 0.5)$, it is for certain that zeros are replaced by some random small values. However, it introduces too much unnecessary fluctuations into the data, which motivated us to use the first proposed method to deal with zeros. This method was not used as it affects regions with small populations more than the ones with larger populations and too much fluctuations were introduced in the data.

Exploratory Plots

Temporally averaged spatial variation

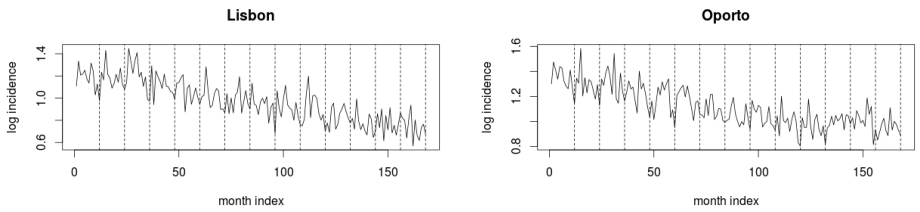
Figure 8.1 maps the spatial averages of log-incidences over time at concelho level. The map indicates that most high-incidence regions appear in central Lisbon and southern part of Lisbon with one up in the north. For Oporto, the central areas and some areas in the east have higher incidences than the rest in general.

(a) Lisbon **(b)** Oporto

Figure 8.1: Map of average log-incidences per 100,000 from 2000-2013

Spatially averaged temporal variation

We ruled out harmonic models at this stage by examining at plots of log-incidences against time; a similar comment applies to regional incidences too. Figure 8.2 shows decreasing trends in both regions, however, no clear seasonality is shown.



(a) Lisbon **(b)** Oporto

Figure 8.2: Spatial average Log-incidences per 100,000 in Lisbon and Oporto

Preliminary Modelling

Linear Mixed Effect Models

One initial approach on the incidence data is to use a linear mixed effect model. Ecological level covariates are treated as fixed effects. Random effects part include a random slope variable with respect to time and a random intercept variable for each region. Therefore, the mixed effect model has the following form for region k and time t:

$$\log(y_{kt}) = \mathbf{x}_{kt}\alpha + \beta_k t + \gamma_k + \epsilon_{kt}. \tag{8.1}$$

In equation 8.1, y denotes the incidence of TB per $100,000$. x refers to the vector of covariates includ-

ing HIV incidences, number of immigrants, non-standard housing, incarceration, unemployment rate and youth index. β and γ represent the random slope and random intercept respectively. Estimates of coefficients for each covariate as well as random intercepts and random slopes in Equation 8.1 can all be easily obtained using R package *nlme*. The model residuals show some patterns in both of the regions.

Generalised linear mixed effect models

A generalised linear mixed model (GLMM) can extend the above linear mixed model to allow a Poisson distribution for the response variable; TB notifications in this case. One of the advantages of this model is that we can then model the response variable directly and thus avoid the problems caused by zero notifications. When estimating parameters, we chose the Bayesian version of GLMM which uses MCMC for inference. The MCMC approach uses either Metropolis-Hastings updates or slice sampling (Damien et al., 1999). In our specific model, the priors are non-informative, where fixed effects are set to have mean 0 and variance 10 and random effects and residuals follow inverse-wishart distribution with parameters $(V = 1, \nu = 0.002)$ and $(V = I_2, \nu = 0.002)$ respectively. The model fit via R package *MCMCglmm* has the following form:

$$y_{kt} - \log(\text{population}) = X_{kt}\alpha + \beta_k t + \gamma_k + \epsilon_{kt}.$$

notations have similar interpretations as above.

MCMC is run with 13×10^5 iterations, 5000 burn-in and 50 thinning period. The chain showed to have converged and mixed reasonably well in general, where the random effects showed the worst mixing (see MCMC chain plots in Figures 8.3 and 8.4 below). This shows that estimates for the posterior mean of each covariates are fairly similar. The DIC value for both regions are relatively large; 14121.09 and 19382.74 respectively.

Summaries of GLMMs for Lisbon and Oporto are shown below, where post.mean refers to the estimated posterior mean for each covariate, eff.samp is the effective sample size and pMCMC is the estimated p-value for each covariate. The estimated posterior means are accompanied by their 95% confidence intervals (l-95% CI for lower and u-95% CI for upper. This shows that estimates for the posterior mean of each covariates are fairly similar. The DIC value for both regions are relatively large; 14121.09 and 19382.74 respectively. This shows that estimates for the posterior mean of each

covariates are fairly similar. Although Markov chains showed acceptable results, the effective sample sizes are not as satisfying in this GLMM approach.

	posterior mean	l-95% CI	u-95% CI	eff.samp	pMCMC
LISBON					
(Intercept)	-1.001e+01	-1.093e+01	-9.272e+00	1965	0 ***
HIV	1.421e-03	-1.488e-03	4.342e-03	2500	0.333
immigrants	-2.345e-06	-2.730e-05	2.323e-05	2001	0.822
inmates	5.392e-04	-9.561e-04	1.980e-03	2569	0.475
homeless	9.874e-02	-1.746e-01	3.588e-01	2315	0.460
unemployment	7.525e-04	-5.331e-02	5.506e-02	1536	1.000
agingidx	-5.226e-03	-1.310e-02	2.355e-03	1814	0.197
log(residents)	1.000e+00	9.999e-01	1.000e+00	2500	0 ***
OPORTO					
(Intercept)	-1.036e+01	-1.077e+01	-9.959e+00	381.0	0 ***
HIV	-1.130e-03	-2.649e-03	4.959e-04	614.6	0.148
immigrants	6.773e-06	-1.819e-04	1.832e-04	748.4	0.920
inmates	2.034e-03	-1.790e-03	5.997e-03	759.1	0.308
homeless	5.942e-01	1.293e-01	1.076e+00	404.2	0.018 *
unemployment	2.853e-02	2.887e-03	5.674e-02	147.1	0.044 *
agingidx	-5.799e-03	-1.211e-02	1.636e-04	177.3	0.066 .
log(residents)	1.000e+00	9.999e-01	1.000e+00	1000.0	0 ***

Table 8.1: GLMM model covariate coefficient estimates. The post.mean are estimated coefficient values based on this model, and $(l-95\%\text{CI}, u-95\%\text{CI})$ refer to the 95% confidence interval for each covariate.

8.1.2 Dynamic Linear model approach

For interests of making dynamic forecast over these data, we proceed the study further by employing the concept of dynamic linear model Durbin and Koopman (2001); West and Harrison (1997). We adapted a log-transformed Gaussian state-space model which is computationally straightforward to implement in theory via the Kalman filter (KF) (Kalman, 1960); similar approaches on disease data can be seen in (Stanton et al., 2014). The most intuitive assumption to make is the Poisson distribution for count data, however, such model is analytically intractable and does not fit KF assumptions here; some transformation of the response variable introduced in the earlier section is used.

Empirical Bayes method provides prior information based on what the data says. With high hierarchies in our model, we will utilise this method to provide the most likely approximation of prior

parameters to initialise the Kalman Filter. We introduce an Expectation Maximisation algorithm here for providing optimal estimates of initial values to start Kalman Filter.

- (STEP 1) Fixing a, σ, ϕ and τ, use EM to estimate μ_β, μ_γ, σ_β and σ_γ.
- (STEP 2) For given μ_β, μ_γ, σ_β and σ_γ use maximise likelihood method (*optim* function) to get new estimates of a, σ, ϕ and τ.
- (STEP 3) Stop if differences between new estimates of a, σ, ϕ and τ and old estimates are smaller than $10^{-\delta}$ for δ some positive integers; else, go back to (STEP 1).

In (STEP 1), let $\psi = (\sigma_\beta, \mu_\gamma, \sigma_\gamma)$, then

$$\log(\pi(\psi|\beta_T, \gamma_T, y, S, \alpha)) = \log(\pi(\beta_T, \gamma_T|\psi, y, S, \alpha)) + c,$$

for $t = 1, ..., T$ and c constant. As β_i's and γ_i's are both i.i.d. Gaussian and they are independent of each other, then taking expectation of $\log(\pi(\psi|\beta_T, \gamma_T, y, S, \alpha))$ with respect to $\log(\pi(\beta_T, \gamma_T|\psi, y, S, \alpha))$ gives

$$\mathbb{E}[\log(\pi(\psi|\beta_T, \gamma_T, y, S, \alpha)] = \mathbb{E}[\log(\pi(\beta_T, \gamma_T|\psi, y, S, \alpha))] + c.$$

As maximising $\mathbb{E}[\log(\pi(\psi|\beta_T, \gamma_T, y, S, \alpha)]$ is equivalent to maximising $\mathbb{E}[\log(\pi(\beta_T, \gamma_T|\psi, y, S, \alpha))]$, EM algorithm has the following steps:

E-step:

$$\begin{aligned}
Q^{(m)} &= \mathbb{E}[\log(\pi(\beta_T, \gamma_T|\psi^{(m)}, y, S, \alpha))] \\
&= \mathbb{E}[-n\log\sigma_\beta^{(m)} - \frac{1}{2\sigma_\beta^{(m)}}\sum_{i=1}^{n}(\beta_{iT} - \mu_\beta^{(m)})^2 - n\log\sigma_\gamma^{(m)} - \frac{1}{2\sigma_\gamma^{(m)}}\sum_{i=1}^{n}(\gamma_{iT} - \mu_\gamma^{(m)})^2] \\
&= -n(\log\sigma_\beta^{(m)} + \log\sigma_\gamma^{(m)}) - \frac{1}{2\sigma_\beta^{(m)}}\sum_{i=1}^{n}(\mathbb{V}\text{ar}[\beta_{iT}] + (\mathbb{E}[\beta_{iT}] - \mu_\beta^{(m)})^2 - \frac{1}{2\sigma_\gamma^{(m)}}\sum_{i=1}^{n}(\mathbb{V}\text{ar}[\gamma_{iT}] + (\mathbb{E}[\gamma_{iT}] - \mu_\gamma^{(m)})^2)
\end{aligned}$$

M-step:

Maximise $Q^{(m)}$ w.r.t. ψ, then

$$\begin{aligned}
\mu_\beta^{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\beta_{iT}] \\
\sigma_\beta^{(m+1)} &= \frac{1}{2n}\sum_{i=1}^{n}(\mathbb{V}\text{ar}[\beta_{iT}] + (\mathbb{E}[\beta_{iT}] - \mu_\beta^{(m)})^2) \\
\mu_\gamma^{(m+1)} &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\gamma_{iT}] \\
\sigma_\gamma^{(m+1)} &= \frac{1}{2n}\sum_{i=1}^{n}(\mathbb{V}\text{ar}[\gamma_{iT}] + (\mathbb{E}[\gamma_{iT}] - \mu_\gamma^{(m)})^2)
\end{aligned}$$

Note that $\mu_\beta = 0$ is a known constant, then change accordingly in the algorithm.

For transformed TB incidences $\mathbf{y}_{it}$ and some functions f and g, the model could take a general state-space form:

$$\begin{aligned}
\theta_{it}|\theta_{i,t-1} &\sim f(\theta_{i,t-1}, W_{it}; \phi), \\
y_{it}|\theta_{it} &\sim g(\theta_{it}, V_{it}; \tau),
\end{aligned} \tag{8.2}$$

where θ_{it} evolves with time in region i at time t, W_{it} and V_{it} are noise processes for θ and y respectively and ϕ, τ refers to the model parameters for the state dynamics and measurement functions. The state variable evolves with time to allow adaptions to changes in incidence patterns.

Cressie (1994) suggested that the KF over space and time is a powerful way to apply Bayesian modeling on space-time phenomena. It is widely used in many fields of engineering but full potential in epidemiology remains undercovered. The spirit of utilising KF to do disease mapping in this paper is in line with the work of Diggle et al. (2005) where the spatio-temporal variation is modelled using some stochastic residual process. There are six potential models fitted here:

- M1: Model with stationary mean with non-spatial random effects:

  System Equation: $\theta_t = A\theta_{t-1} + Bw_t$,

  Observation Equation: $y_t = C_t\theta_t + v_t$,

  Here, $\theta_t = (S_t, \alpha_t)^\mathsf{T}$ and y_t is a vector of the log-incidence per $100,000$ for all regions in the model. $w_t \sim \mathcal{N}(0, W_t)$ and $v_t \sim \mathcal{N}(0, V_t)$ are covariances for system and observation respectively, where $W_t = \left[\begin{array}{c|c} \sigma^2 I_n & 0 \\ \hline 0 & 0 \end{array}\right]$. It assumes no dynamic evolution in covariate effects in this model (ie. zero variance in state equation). Covariate effects (the αs) are estimated online,

with no extra noise added in through the state equation.

- M2: Model with stationary mean and variance with non-spatial random effects;

  The model formulation is the same as M1 but with stationary unconditional mean and variance. I.e. coefficient matrix B is an $(n+p)\times(n+p)$ block diagonal matrix consists of a $n\times n$ diagonal matrix with entries $\sqrt{1-a^2}$ and a $p\times p$ zero matrix.

- M3: Model with stationary mean with spatial random effects:

  System Equation: $\theta_t = A\theta_{t-1} + Bw_t$,

  Observation Equation: $y_t = C_t\theta_t + Dv_t$;

  θ and y are the same as in M1. This model includes formulation includes spatial random effects; spatial correlation between regions decays exponentially as distance gets bigger. This follows variance can be written as $\Sigma_{ij} = \sigma^2 \exp\{-d_{ij}/\phi\}$, where σ^2 is the variance within each region, d_{ij} is the distance between the population weighted centroids of regions i and j and ϕ is the range beyond which spatial correlation can be neglected. The systematic noise has a slightly different form from before. $W_t = \left[\begin{array}{c|c} \Sigma_{\rho,\sigma} & 0 \\ \hline 0 & 0 \end{array}\right]$ is an $(n+p)\times(n+p)$ matrix where $\Sigma_{\rho,\sigma}$ is the $n\times n$ covariance matrix with entries Σ_{ij}.

- M4: Model with stationary mean and variance with spatial random effects;

  Formulation is the same as for M3, the fourth model (M4) assumes variance to be stationary. The form is exactly the same as in $M3$ but with different constructions of coefficient matrix B. Similarly to the difference between $M1$ and M2, matrix B has the same form as it is in M2.

- M5: Model with stationary mean and variance with spatial random effects and random intercept:

  Formulation is the same as for M3, but the state variable $\theta_t = (S_t, \alpha_t, \gamma_t)^\mathsf{T}$ where γ refers to random intercepts. Both α_t and γ_t are assumed to have variance zero, covariance matrix W is an $(n+p+n)\times(n+p+n)$ block matirx of the form $\left[\begin{array}{c|c|c} \Sigma_{\rho,\sigma} & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$. D and V_t are the same as before.

- M6: Model with stationary mean and variance with spatial random effects, random intercepts and random slopes.

This can be treated as an extension of $M5$. The state variable here also includes a random slope β; ie. $\theta_t = (S_t, \beta_t, \alpha_t, \gamma_t)^\mathsf{T}$. The variance of random slope is also assumed to be zero, but $W_t = \left[\begin{array}{c|c|c|c} \Sigma_{\rho,\sigma} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right]$ is an $(n+n+p+n) \times (n+n+p+n)$ matrix. Again, D and V_t are the same as before.

These models are compared using the Akaike information criterion (AIC) (Akaike, 1974). Model inference is obtained via Kalman filter for each situation and AIC values are used as measurement of best model fit.The best model which explains the TB incidence data using Kalman filter is the model includes stationary mean and variance with spatial random effects, random intercepts and random slopes. Details of how M1-M6 varies in coefficient matrices and covariance matrices W and V are shown as below.

Recall model parameters $\psi = (\sigma^2, \phi, \tau^2, a)$, these parameters represent: the variance within each region when distance between regions are 0, the range out of which the spatial correlation can be neglected, the observational variance and the level of spatial correlation between neighbouring districts (independence when $a = 1$, maximum when $a = 0$). These parameters are involved in the following places: $W_t \sim \mathcal{N}(0, W)$ and $V_{it} \sim \mathcal{N}(0, V)$ where $W'_{ij} = \sigma^2 \exp\{-d_{ij}/\phi\}$ and V' has diagonal entries τ^2. Here W' and V' refers to one component in the matrix designed in W and V respectively (see definition below).

Define $\mathbf{A}_i$ to be coefficient matrix A for model M_i and do the same for B and C.

$$A_1 = A_2 = A_3 = A_4 = \left(\begin{array}{c|c} a \times \mathbf{I}_n & 0 \\ \hline 0 & \mathbf{I}_p \end{array}\right)$$

$$B_1 = B_3 = \left(\begin{array}{c|c} \mathbf{I}_n & 0 \\ \hline 0 & 0 \end{array}\right) \qquad B_2 = B_4 = the\left(\begin{array}{c|c} \sqrt{1-a^2} \times \mathbf{I}_n & 0 \\ \hline 0 & 0 \end{array}\right)$$

$$C_{t1} = C_{t2} = C_{t3} = C_{t4} = \left[\begin{array}{cccc|cccc} 1 & 0 & \dots & 0 & X^t_{1,1} & X^t_{1,2} & \dots & X^t_{1,p} \\ 0 & 1 & \dots & 0 & X^t_{2,1} & X^t_{2,2} & \dots & X^t_{2,p} \\ \vdots & \vdots & \dots & \vdots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 & X^t_{n,1} & X^t_{n,2} & \dots & X^t_{n,p} \end{array}\right]$$

where $X^t_{i,j}$ are entries of X_t and it denotes the jth covariate in the ith region at time t .

$$A_5 = \left(\begin{array}{c|c|c} a \times \mathbf{I}_n & 0 & 0 \\ \hline 0 & \mathbf{I}_p & 0 \\ \hline 0 & 0 & \mathbf{I}_n \end{array}\right) \qquad B_5 = \left(\begin{array}{c|c|c} \sqrt{1-a^2}\mathbf{I}_n & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \end{array}\right)$$

$$A_6 = \left(\begin{array}{c|c|c|c} a \times \mathbf{I}_n & 0 & 0 & 0 \\ \hline 0 & \mathbf{I}_n & 0 & 0 \\ \hline 0 & 0 & \mathbf{I}_p & 0 \\ \hline 0 & 0 & 0 & \mathbf{I}_n \end{array}\right) \qquad B_6 = \left(\begin{array}{c|c|c|c} \sqrt{1-a^2}\mathbf{I}_n & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \end{array}\right)$$

$$C_{t5} = \left[\begin{array}{c|c|c} \mathbf{I}_n & X_t & \mathbf{I}_n \end{array}\right] \qquad C_{t6} = \left[\begin{array}{c|c|c|c} \mathbf{I}_n & \mathbf{I}_n & X_t & \mathbf{I}_n \end{array}\right]$$

$$(V_t)_i = 100,000 \times \sigma_v^2 \times \begin{bmatrix} 1/P_{1t} & 0 & \dots & 0 \\ 0 & 1/P_{2t} & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1/P_{nt} \end{bmatrix} \forall i = 1, \dots, 6$$

$$\begin{aligned} \text{System Equation: } \theta_t &= A\theta_{t-1} + Bw_t, \\ \text{Observation Equation: } y_t &= C_t\theta_t + v_t, \end{aligned} \tag{8.3}$$

where $\theta_t = (\S_t, \alpha_t)^\intercal$ and y_t is a vector of the log-incidence per $100,000$ of certain disease for all regions in the model. $w_t \sim \mathcal{N}(0, W_t)$ and $v_t \sim \mathcal{N}(0, V_t)$ are covariances for system and observation

respectively, where $W_t = \left[\begin{array}{c|c} \sigma^2 I_n & 0 \\ \hline 0 & 0 \end{array}\right]$. It is assumed that the covariate effects have no dynamic evolution in the model (ie. zero variance in state equation). In this model the covariate effects (the αs) are estimated online, with no extra noise added in through the state equation.

The log-likelihood for data $y_1, \ldots, y_n$ through each time point t for $M1 - M6$ can be computed as

$$\log\{\pi(y_{1:n}|\psi)\} = \log\{\pi(y_1|\psi) + \sum_{i=2}^{n} \log\{\pi(y_i|y_{1:i-1}, \psi)\},$$

where $\psi = (\sigma^2, \phi.\tau^2, a)$ is a set of parameter values. This log-likelihood can be computed as a by-product of the Kalman filter step by step. We estimates the parameters by maximising the log-likelihood via function *optim* with Nelder-Mead method in R (R Core Team, 2013).

Results from DLM

Table 8.2 shows the model comparisons based on AIC values. M6 shows the best model fit for both regions.

Table 8.2: Model Comparisons: AIC (4 decimal places) for M1-M6

	Lisbon	Oporto
$M1$	-602.7547	-618.4741
$M2$	-602.7547	-618.4741
$M3$	-604.0686	-643.4085
$M4$	-604.0686	-643.4085
$M5$	-1572.973	-1963.474
$M6$	-1649.032	-2193.345

Figure 8.5 plots the predicted values (in red) from models for centroids of Lisbon and Oporto with 95% confidence intervals in blue. The crosses overlaid onto plots are the actual observed values. All observed values lie within the 95% confidence interval predicted. Combined with AIC values, M6 shows a reasonable model fit for TB incidences data we consider here.

Table 8.3 shows estimated values for model parameters; these are obtained by maximising the marginal likelihood function of Model 6. The prior distribution $\mathcal{N}(\theta_0, \Sigma_0)$ was initiated using spatial random effects $S_{it} \sim \mathcal{N}(0, 10)$, random slope $\beta_{it} \sim \mathcal{N}(0, \sigma^2_\beta)$ and random intercept $\gamma_{it} \sim \mathcal{N}(0, \sigma^2_\gamma)$. The prior values were chosen so that the variance is large enough to account for uncertainty in spatial correla-

tions. Both σ_β^2 and σ_γ^2 are estimated via Empirical Bayes methods using EM-algorithm mentioned above.

Table 8.3: Estimated Parameters (4 decimal places)

	Lisbon	Oporto
σ_β^2	3.2634×10^{-6}	7.1568×10^{-6}
σ_γ^2	0.1887	0.1236
σ^2	0.2010	0.2310
ϕ	946.8926	2392.0930
τ^2	0.0179	0.0059
a	0.0363	0.0375

As estimates have uncertainties associated, instead of plotting actual values, we plot the probability of random slope greater than 0 (Figure 8.6). The map shows higher chance of increasing trend in TB against time on the Northwestern edge and centroid of Lisbon regions. In Oporto regions, higher odds are shown down in southeastern edges mainly.

The relationship between TB incidences and covariates listed above is captured in the state variable α. For both Lisbon and Oporto regions, estimates of the covariate effects at the last time point (ie. December 2013) is shown in Table 8.4. The directions of covariate effects are the same for immigrants, incarceration and ageing index.

Table 8.4: Estimates of Covariate Effects Coefficients (4 decimal places)

	HIV	immigrants	incarceration	homeless	unemployment	ageing index
Lisbon	−0.0134	−0.0088	0.1488	0.0226	−0.1118	−0.8621
Oporto	−0.0032	−0.0002	−0.0023	−0.0037	0.0004	−0.0023

As mentioned in the previous section, the suggested low incidence rate by WHO is 20 per $100,000$. We mapped the probability of incidences greater 20 per $100,000$ at the last time point of observations in Figure 8.7; Lisbon the high probability occurs around the centroid whereas more areas in the south-east regions showed high probability in Oporto. Oporto areas also tends to have higher chances to be at risk of being out of TB-safe category in comparison to Lisbon areas.

This state-space model showed that spatial dependence decays and can be ignored when regions are more than 1 kilometre apart from each other in Lisbon areas and such dependence can be seen in regions within around 2.4 kilometres in Oporto regions. The regional variance existing when distance

is 0 in each region are fairly similar for Lisbon and Oporto. The value of a suggests that there exists some positive temporal correlation across municipalties in both Lisbon and Oporto Metropolitan areas. These models are flexible and are able to produce forecasts based on existing data at relatively high computational costs. However, the transformation of count data could potentially affect the quality of model results. To satisfy the linear Gaussian assumption for KF, TB incidences were also transferred using logarithm as described before. Such accommodation of zeros for TB incidences per $100,000$ increases the total observed TB in each region. This might not have such a big impact on large population areas, but in small prevalence areas, TB notification could have changed from no notification to other bigger values. Note, there are a lot of areas with low prevalence and small population in our study.

8.1.3 Additional Plots and Tables

	mean	sd	0.025quant	0.5quant	0.975quant
LISBON					
(Intercept)*	-8.504	0.033	-8.569	-8.503	-8.439
ageing*	0.002	0.000	0.002	0.002	0.002
immigrants*	0.010	0.002	0.003	0.007	0.011
unemployment*	0.015	0.002	0.010	0.015	0.019
overcrowding*	0.046	0.002	0.043	0.046	0.049
homeless*	0.012	0.001	0.010	0.012	0.015
hiv*	0.004	0.000	0.003	0.004	0.005
OPORTO					
	mean	sd	0.025quant	0.5quant	0.975quant
(Intercept)*	-8.182	0.042	-8.264	-8.182	-8.100
ageing*	0.001	0.000	0.000	0.001	0.001
immigrants*	0.050	0.010	0.030	0.050	0.070
unemployment*	0.028	0.002	0.024	0.028	0.032
overcrowding*	0.027	0.002	0.024	0.027	0.030
homeless*	0.006	0.001	0.005	0.006	0.008
hiv*	0.004	0.001	0.002	0.004	0.006

Table 8.5: Fixed Effects from Poisson model with no random effects for Lisbon and Oporto Metropolitan Areas; * refers to significant covariates

LISBON		Oporto	
Freguesia	Municipality	Freguesia	Municipality
Seixal	Seixal	Vila do Conde	Vila do Conde
Rrafaria	Almada	Perafita	Matosinhos
Caparica	Almada	Alpendurada e Matos	Marco de Canaveses
Beato	Lisboa	Portela	Penafiel
Alfragide	Amadora	Rio de Moinhos	Penafiel
Buraca	Amadora	Boelhe	Penafiel
Falagueira	Amadora	Cabeça Santa	Penafiel
Mina	Amadora	Perozelo	Penafiel
Agualva	Sintra	Oldrões	Penafiel
Alcoentre	Azambuja	Luzim	Penafiel

Table 8.6: Top 10 High Relative Risk Areas For Lisbon and Oporto with The Municipality (higher administrative region)

	ageing	immigrants	unemployment	overcrowding	homeless	hiv
ageing	1	-0.13	0.18	-0.27	0.19	0.36
immigrants	-0.13	1	0.15	0.38	1×10^{-1}	0.11
unemployment	0.18	0.15	1	0.15	8×10^{-2}	4×10^{-2}
overcrowding	-0.27	0.38	0.15	1	-0.16	-4×10^{-2}
homeless	0.19	1×10^{-1}	8×10^{-2}	-0.16	1	0.14
hiv	0.36	0.11	4×10^{-2}	-4×10^{-2}	0.14	1

Table 8.7: Correlation between ecological covariates in Lisbon.

8.2 APPENDIX for Space-Time Survival Modelling

8.2.1 Preliminary Cox Proportional Hazard Models

A simple Cox proportional hazard model with a time indexed covariate is fit as preliminary model. The study period from 1973 to 2012 is cut into eight different five-year intervals. The survival plot adjusted for time interval indices is shown in Figure 8.15. Different colours refer to the relevant survival curve in the time interval. The 95% confidence intervals are plotted in the corresponding colours as well. It is shown that the survival probability are fairly different for different time intervals.

	ageing	immigrants	unemployment	overcrowding	homeless	hiv
ageing	1	0.23	0.51	-7×10^{-2}	3×10^{-1}	0.58
immigrants	0.23	1	0.12	-3×10^{-1}	0.14	0.28
unemployment	0.51	0.12	1	-0.11	0.24	0.31
overcrowding	-7×10^{-2}	-3×10^{-1}	-0.11	1	2×10^{-2}	3×10^{-2}
homeless	3×10^{-1}	0.14	0.24	2×10^{-2}	1	0.33
hiv	0.58	0.28	0.31	3×10^{-2}	0.33	1

Table 8.8: Correlation between ecological covariates in Porto.

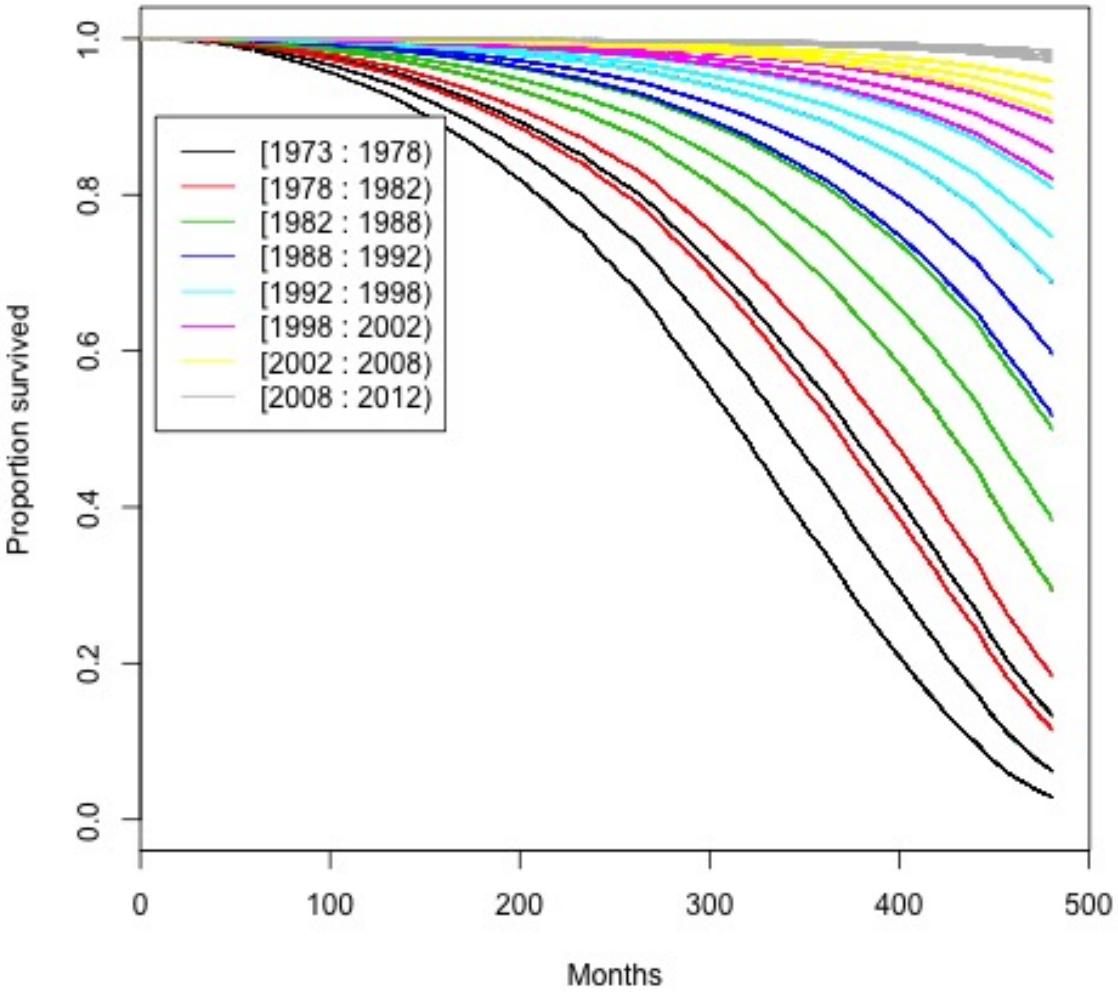


Figure 8.15: Cox survival plot adjusted for each time interval with their corresponding 95% confidence intervals.

8.2.2 Codes for Simulation Studies

```
set.seed(1)


tent <- function(n,tmin=0,tmax=1){
  samp <- c()
```

```
  while(length(samp)<n){
    tcand <- runif(1,tmin,tmax)
    a <- tcand/(tmax-tmin)
    if(a>runif(1)){
      samp <- c(samp,tcand)
    }
  }
  return(samp)
} # function to get skewed entering times

d <- simsurv(X = cbind(age = runif(N, 5, 50),
       sex = rbinom(N, 1, 0.5),
       cancer = rbinom(N, 1, 0.2)),
       beta = c(0.01, 0.01, 0.01),
       omega = 1,
       cov.parameters=c(0.1,0.1))
```

8.2.3 Appendix: Gaussian Process Model and Its Relevant Derivations

An alternative way to capture both calendar time and survival time effects is via a Gaussian process model. We propose that when the assumption of failure of an event being independent from the entry does not hold, adjustments are made to allow both timescales to take roles in the model. We propose that the new baseline hazard to have the form

$$h_0 = f(t;\omega_f)\exp\{g(t+t_0;\gamma,\tau)\},$$

where $f(t;\omega_f)$ refers to an ordinary baseline hazard function and $g(t+t_0;\gamma^{(t)},\tau)$ is some temporal Gaussian process for t the survival time and t_0 the time of entering. This baseline hazard then concerns not only the time interval of individual being in the study but also the real time of them entering the study, which intends to develop appropriate methodology for making inference about

data with the real time. Similarly to previously stated, the hazard function thus has the form:

$$h(t, t_0) = h_0(t, t_0) \exp\{X\beta + Z\},$$

where Z is a latent Gaussian field for spatial effects and it follows the definition from above. This method is however not practical for the data collected over study period which are too long as more need to be done to solve the complexity in computation ($O(N^3)$ problem).

Define the function g to take the form of

$$g(t + t_0; \tau, \theta, \gamma^{(t)}) = \sum_{j=1}^{m} (\tau\gamma_j^{(t)} - \frac{\tau^2}{2}) \mathbb{I}(t + t_0 \in I_j).$$

More specifically, for each individual i, $t + t_0$ only falls in one possible interval I_j. Thus for individual i such that $t + t_0 \in I_j$,

$$g^{(j)}(t + t_0; \tau, \theta, \gamma_j^{(t)}) = \tau\gamma_j^{(t)} - \frac{\tau^2}{2}.$$

For given j, in order to complete the derivation of posterior distribution described earlier, the following derivatives of baseline hazard functions are useful. First derivatives of this new temporal baseline hazard h_0 are:

- w.r.t $\gamma_j^{(t)}$

  $$\frac{\partial h_0(t)}{\partial \gamma_j^{(t)}} = f(t) \exp\{g(t + t_0)\} \frac{\partial g(t+t_0)}{\partial \gamma_j^{(t)}}$$

- w.r.t ω

  $$\frac{\partial h_0(t)}{\partial \omega} = \frac{\partial f(t)}{\partial \omega} \exp\{g(t + t_0)\}$$

- w.r.t τ

  $$\frac{\partial h_0(t)}{\partial \tau} = f(t) \exp\{g(t + t_0)\} \frac{\partial g(t+t_0)}{\partial \tau}$$

Computation of second derivatives of baseline hazard h_0 is useful for evaluating Hessian matrices later and they are:

- $\frac{\partial h_0(t)}{\partial \gamma_j^{(t)}}$ w.r.t $\gamma_k^{(t)}$

$$\frac{\partial^2 h_0(t)}{\partial \gamma_j^{(t)} \partial \gamma_k^{(t)}} = f(t)\exp\{g(t+t_0)\}\left[(\frac{\partial g(t+t_0)}{\partial \gamma_j^{(t)}})\frac{\partial g(t+t_0)}{\partial \gamma_k^{(t)}}) + \frac{\partial^2 g(t+t_0)}{\partial \gamma_j^{(t)} \partial \gamma_k^{(t)}}\right]$$

- $\frac{\partial h_0(t)}{\partial \omega}$ w.r.t ω

$$\frac{\partial^2 h_0(t)}{\partial \omega^2} = \frac{\partial^2 f(t;\omega)}{\partial \omega^2}\exp\{g(t+t_0)\}$$

- $\frac{\partial h_0(t)}{\partial \tau}$ w.r.t τ

$$\frac{\partial^2 h_0(t)}{\partial \tau^2} = f(t)\exp\{g(t+t_0)\}\left[(\frac{\partial g(t+t_0)}{\partial \tau})^2 + \frac{\partial^2 g(t+t_0)}{\partial \tau^2}\right]$$

- $\frac{\partial h_0(t)}{\partial \omega}$ w.r.t ω

$$\frac{\partial^2 h_0(t)}{\partial \tau \partial \omega} = \frac{\partial f(t)}{\partial \omega}\frac{\partial g(t+t_0)}{\partial \tau}\exp\{g(t+t_0)\}$$

- $\frac{\partial h_0(t)}{\partial \gamma_j^{(t)}}$ w.r.t ω

$$\frac{\partial^2 h_0(t)}{\partial \gamma_j^{(t)} \partial \omega} = \frac{\partial f(t)}{\partial \omega}\exp\{g(t+t_0)\}\frac{\partial g(t+t_0)}{\partial \gamma_j^{(t)}}$$

- $\frac{\partial h_0(t)}{\partial \gamma_j^{(t)}}$ w.r.t τ

$$\frac{\partial^2 h_0(t)}{\partial \gamma_j^{(t)} \partial \tau} = f(t;\omega)[\exp\{g(t+t_0)\}\frac{\partial g(t+t_0)}{\partial \gamma_j^{(t)}}\frac{\partial g(t+t_0)}{\partial \tau} + \exp\{g(t+t_0)\}\frac{\partial^2 g(t+t_0)}{\partial \gamma_j^{(t)} \partial \tau}]$$

Derivation of cumulative hazard and related derivatives

Note that as shown in Figure 8.16, $t + t_0$ may lie in anywhere in the intervals I_j defined. It is then useful to define the following notations in order to compute the cumulative hazard function. We define the new indices for intervals $j^*_{\max} = \max_j\{t + t_0 \in I_j\}$ and $j^* = \min_j\{t_0 \in I_j\}$.

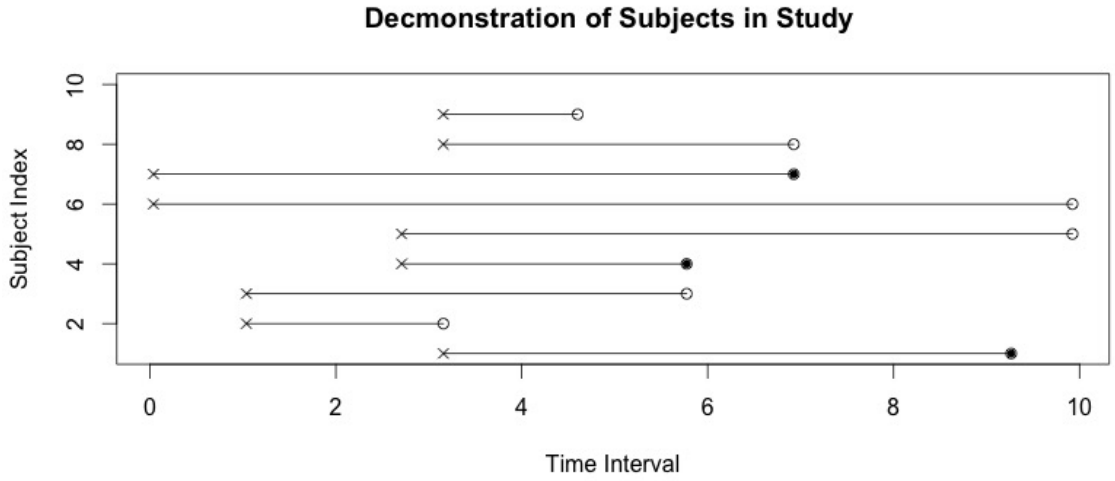


Figure 8.16: Demonstration of how individuals may enter the study.

Recall that the cumulative hazard function $H_0(t)$ has the form:

$$H_0(t) = \int_{t_0}^{t} h_0(s)\mathrm{d}s.$$

Then for some survival time t and entering time t_0, we have

$$\begin{aligned} H_0(t) &= \int_0^t f(s)\exp\{g(s+t_0)\}\mathrm{d}s \\ &= \sum_{j_{\min}}^{j^*_{\max}-1} \int_{I_j} f(s)\exp\{g^{(j)}(s+t_0)\}\mathrm{d}s + \int_{I_{j^*_{\max}}}^{t} f(s)\exp\{g^{(j^*_{\max})}(s+t_0)\}\mathrm{d}s \\ &= \sum_{j_{\min}}^{j^*_{\max}-1} \exp\{g^{(j)}\}\int_{I_j} f(s)\mathrm{d}s + \exp\{g^{(j^*_{\max})}\}\int_{I^{\text{lower}}_{j^*_{\max}}}^{t} f(s)\mathrm{d}s \end{aligned}$$

where $g^{(j)}$ is the value of function g in the jth interval. Note that $\int_{I_j} f(s)\mathrm{d}s \triangleq F(I_j^{\text{upper}}) - F(I_j^{\text{lower}})$, then H_0 follows:

$$H_0(t) = \sum_{j_{\min}}^{j^*_{\max}-1} \exp\{g^{(j)}(t+t_0;\gamma,\tau)\}[F(I_j^{\text{upper}}) - F(I_j^{\text{lower}})] + \exp\{g^{(j^*_{\max})}\}[F(I^{\text{upper}}_{j^*_{\max}}) - F(I^{\text{lower}}_{j^*_{\max}})].$$

We then define the upper and lower bound for intervals depending on where $t+t_0$ lies in the interval.

For each individual i,

$$I_j^{*\text{upper}} = \begin{cases} t, & \text{if } j = j^*_{\max} \\ I_j^{\text{upper}}, & \text{otherwise} \end{cases}$$

Thus it follows that

$$H_0(t) = \sum_{j_{\min}}^{j^*_{\max}} \exp\{g^{(j)}\}[F(I_j^{*\text{upper}}) - F(I_j^{\text{lower}})].$$

Then the first derivatives of H_0 are as shown below:

- w.r.t $\gamma_j^{(t)}$

  $\frac{\partial H_0(t)}{\partial \gamma_j^{(t)}} = \exp\{g^{(j)}\}\frac{\partial g^{(j)}}{\partial \gamma_j^{(t)}}[F(I_j^{*\text{upper}}) - F(I_j^{\text{lower}})]$

- w.r.t ω

  $\frac{\partial H_0(t)}{\partial \omega} = \sum_{j_{\min}}^{j^*_{\max}} \exp\{g^{(j)}\}\left[\frac{\partial F(I_j^{*\text{upper}})}{\partial \omega} - \frac{\partial F(I_j^{\text{lower}})}{\partial \omega}\right]$

- w.r.t τ

  $\frac{\partial H_0(t)}{\partial \tau} = \sum_{j_{\min}}^{j^*_{\max}} \frac{\partial g^{(j)}(t+t_0)}{\partial \tau} \exp\{g^{(j)}\}[F(I_j^{*\text{upper}}) - F(I_j^{\text{lower}})]$

Second derivatives of H_0:

- $\frac{\partial H_0(t)}{\partial \gamma_{(j)}^{(t)}}$ w.r.t $\gamma_{(j)}^{(t)}$

  $\frac{\partial^2 H_0(t)}{\partial {\gamma_{(j)}^{(t)}}^2} = \sum_{j_{\min}}^{j^*_{\max}}[\exp\{g^j\}(\frac{\partial g^{(j)}(t+t_0)}{\partial \gamma_{(j)}^{(t)}})^2 + \exp\{g^{(j)}\}\frac{\partial^2 g^{(j)}(t+t_0)}{\partial {\gamma_j^{(t)}}^2}] \times [F(I_j^{*\text{upper}}) - F(I_j^{\text{lower}})]$

  for $j \neq k$, $\frac{\partial^2 H_0(t)}{\partial \gamma_{(j)}^{(t)} \partial \gamma_{(k)}^{(t)}} = 0$.

- $\frac{\partial H_0(t)}{\partial \omega}$ w.r.t ω

  $\frac{\partial^2 H_0(t)}{\partial \omega^2} = \sum_{j_{\min}}^{j^*_{\max}} \exp\{g^{(j)}\}\left[\frac{\partial^2 F(I_j^{*\text{upper}})}{\partial \omega^2} - \frac{\partial^2 F(I_j^{\text{lower}})}{\partial \omega^2}\right]$

- $\frac{\partial H_0(t)}{\partial \tau}$ w.r.t τ

$$\frac{\partial^2 H_0(t)}{\partial \tau^2} = \sum_{j_{\min}}^{j^*_{\max}} \left[\exp\{g^{(j)}\} (\frac{\partial g^{(j)}(t_j + t_0)}{\partial \tau})^2 + \exp\{g^{(j)}(t_j + t_0)\} \frac{\partial^2 g^{(j)}(t_j + t_0)}{\partial \tau^2} \right] \times [F(I_j^{*\text{upper}}) - F(I_j^{\text{lower}})]$$

- $\frac{\partial H_0(t)}{\partial \omega}$ w.r.t τ

$$\frac{\partial^2 H_0(t)}{\partial \tau \partial \omega} = \sum_{j_{\min}}^{j^*_{\max}} \exp\{g^{(j)}\} (\frac{\partial g^{(j)}(t+t_0)}{\partial \tau}) \left[\frac{\partial F(I_j^{*\text{upper}})}{\partial \omega} - \frac{\partial F(I_j^{\text{lower}})}{\partial \omega} \right]$$

- $\frac{\partial H_0(t)}{\partial \gamma_j^{(t)}}$ w.r.t τ

$$\frac{\partial^2 H_0(t)}{\partial \gamma_j^{(t)} \partial \tau} = \left[\exp\{g^{(j)}\} \frac{\partial g^{(j)}}{\partial \gamma_j^{(t)}} \frac{\partial g^{(j)}(t + t_0)}{\partial \tau} + \exp\{g^{(j)}\} \frac{\partial^2 g^{(j)}(t + t_0)}{\partial \gamma_j^{(t)} \partial \tau} \right] \times [F(I_j^{*\text{upper}}) - F(I_j^{\text{lower}})]$$

- $\frac{\partial H_0(t)}{\partial \omega}$ w.r.t $\gamma_j^{(t)}$

$$\frac{\partial^2 H_0(t)}{\partial \gamma_j^{(t)} \partial \omega} = \exp\{g^{(j)}\} \frac{\partial g^{(j)}(t_j + t_0)}{\partial \gamma_j^{(t)}} \left[\frac{\partial F(I_j^{*\text{upper}})}{\partial \omega} - \frac{\partial F(I_j^{\text{lower}})}{\partial \omega} \right]$$

Derivations for the function g

Recall that

$$g(t + t_0; \tau, \theta, \gamma^{(t)}) = \sum_{j=1}^{m} (\tau \gamma_j^{(t)} - \frac{\tau^2}{2}) \mathbb{I}(t + t_0 \in I_j),$$

for individual i with $t + t_0 \in I_j$, we have

$$g^{(j)}(t + t_0; \tau, \theta, \gamma_j^{(t)}) = \tau \gamma_j^{(t)} - \frac{\tau^2}{2}.$$

The derivatives of g are then as followings.

- w.r.t τ

$$\frac{\partial g(t+t_0)}{\partial \tau} = \gamma_j^{(t)} - \tau$$

- w.r.t $\gamma_j^{(t)}$

$$\frac{\partial g(t+t_0)}{\partial \gamma_j^{(t)}} = \tau$$

The second derivatives of g are:

- $\frac{\partial g(t+t_0)}{\partial \tau}$ w.r.t. τ

$$\frac{\partial^2 g(t+t_0)}{\partial \tau^2} = -1$$

- $\frac{\partial g(t+t_0)}{\partial \tau}$ w.r.t. $\gamma_j^{(t)}$

$$\frac{\partial^2 g(t+t_0)}{\partial \tau \partial \gamma_j^{(t)}} = 1$$

- $\frac{\partial g(t+t_0)}{\partial \gamma_j^{(t)}}$ w.r.t. $\gamma_j^{(t)}$

$$\frac{\partial^2 g(t+t_0)}{\partial {\gamma_j^{(t)}}^2} = 0$$

Impute these derivatives into above where relevant, we get a set of first and second derivatives of h_0 and H_0. These results will be used later on in defining baseline hazards which captures the temporal effects as well as some individually relevant risk factors (defined in f).

8.3 Appendix to Chapter 6

This section demonstrates the exploratory analysis of GP pregabalin prescription data between 2015 and 2017. It is possible that prescriptions for CCGs do not necessarily represent the prescriptions in the catchment areas of GPs as some patients may not necessarily be registered with the GP within the CCG that they live in. We first took a 'naive' approach of these data at LSOA level. Each LSOA level prescription is allocated by weighting the GP prescriptions by the registration rate from each LSOA. Figure 8.17b shows the milligrams of Pregabalin prescribed from each GP based on CCG level averaged through time. The biggest high prescriptions were seen in north-eastern and south-eastern part of England, with a relatively high rate seen in a belt going through coastline in Norwich, north

of London up to south of Bristol. Other brighter colours can be found in/around bigger cities; for example, near Manchester, Birmingham and Outer London Areas.

Here we present results from three chosen CCGs; Cumbira, Morecambe Bay and Manchester. These are generically picked as the surrounding CCGs of Lancaster. A general trend of increase can be seen in both Cumbria and Manchester CCGs; roughly following the temporal trend for whole of England. For Morecambe Bay CCG, a flatter trend is shown. There was a sharp decrease after 2015. This brings the prescription rate down by about 500mgs from the peak in September 2015. The trend remained moderately low in 2016 in comparison to 2015, although still relatively high among CCGs.

Figure 8.19 shows the LSOAs associated with the three chosen CCGs; Cumbria, Morecambe and Manchester. The highlight areas are higher estimates for Pregabalin prescriptions. Cumbria shows a high estimate in the North. For both Morecambe and Manchester, the prescriptions show higher rate in centre. This roughly agrees with the spatial effect map from the CCG model but also clearly presents the different associations and rates from their catchments.

Among these CCGs, both 'seasonality' and spatial patterns to some extent can be seen. Therefore, the following Figure 8.20 demonstrates more dynamically prescriptions in England mapped at LSOA levels for the start and end of our study period as well as the sixth and ninth months of each year.

8.3.1 Mixed Effect Models at LSOA

Following the observed temporal and spatial pattern in raw data, it is not unreasonable to employ a Gaussian linear mixed effect model which takes into account of spatial random effects and temporal trends. It is intuitive that neighbouring CCGs are likely to share some common characteristics; both in the sense of potential socioeconomic factors and GPs within certain areas might be consistent when making prescriptions.

The response variable used throughout the study is the square root transformed prescription rate per population in milligram (denoted as y); such transformation improves the model fit. For region i and time t:

$$f(y_{it}) \sim \mathcal{N}(\mu_{it}, \Sigma_{it}),$$

for y_{it} being the expected prescribing rate, μ_{it} and Σ_{it} denoting some mean and covariance matrix.

The most recent (2015) Index of Multiple Deprivation of UK based on census and LSOA from 2011 as covariate, together with time index. For occasions where individual CCG and its associated LSOAs are more relevant to look into. A weight is assigned to each LSOA based on the proportion of total number of patients registered. For given CCG, we fitted a linear mixed effect model with weights based on proportion of the observed cases in each LSOA contributed to the CCG in LSOA population. Again, we look at Cumbira, Morecambe and Manchester CCGs. Figure 8.21 shows the fitted values against observed values. Model fits for all three CCGs look reasonably good, with Cumbria showing more 'offs' at the higher prescription rates and Morecambe and Manchester more skewed in the middle.

The model estimates for Cumbria, Morecambe Bay and Manchester are given in the Table 8.9. For all chosen CCGs, the estimated coefficients for IMD rank indices are negative. This indicate that the more serious deprivation is shown, the higher Pregabalin prescription rate shall be expected. The effect size of IMD rank in Manchester shows the biggest impact on prescription rate.

	Estimate	Std. Error	t value
Cumbria			
(Intercept)	275.40	7.74	35.60
IMD_rank	-4.72	1.36	-3.48
Morecambe			
(Intercept)	306.50	6.48	47.31
IMD_rank	-1.76	1.02	-1.74
Manchester			
(Intercept)	374.86	3.22	116.50
IMD_rank	-10.24	0.60	-17.10

Table 8.9: Estimates for Fixed Effects in chosen CCG.

Despite the reasonable model outcome from linear mixed effect models, a more flexible and smooth technique which captures the non-linearities in the data can be applied to extend the method further. Consider a generalised additive model which incorporates spatial random effect as a smoothing function $g(x)$ where x denotes the geographical location. Here x would be the coordinates of centroid of each LSOA in consideration. The contour plots of fitted models show that predictions of pregabalin prescriptions are higher in the centre regions for Cumbria and Morecambe Bay CCGs. Manchester CCG shows a flatter trend across regions in centre LSOAs comparing to others.

With the establishment of potential spatial patterns and temporal trends from previous discussions,

it is then possible to then look at trends with respect to GP locations. For faster inferences, we produce further study for whole England at CCG-level, where each CCG is treated independently and takes into account of geographical location of GP practices, time period and multiple deprivation levels.

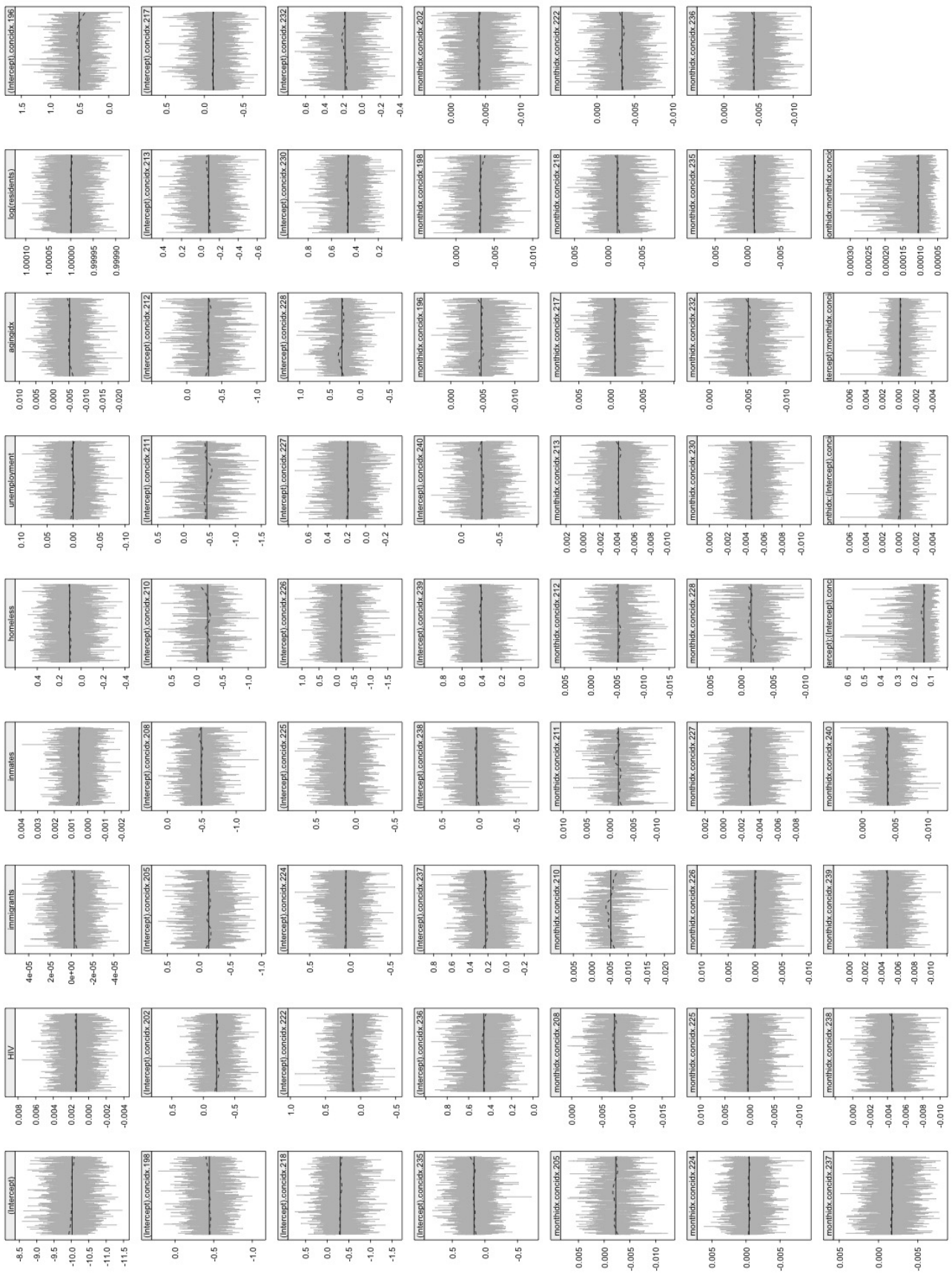


Figure 8.3: Markov Chain chain plots of each covariate in GLMM for Lisbon

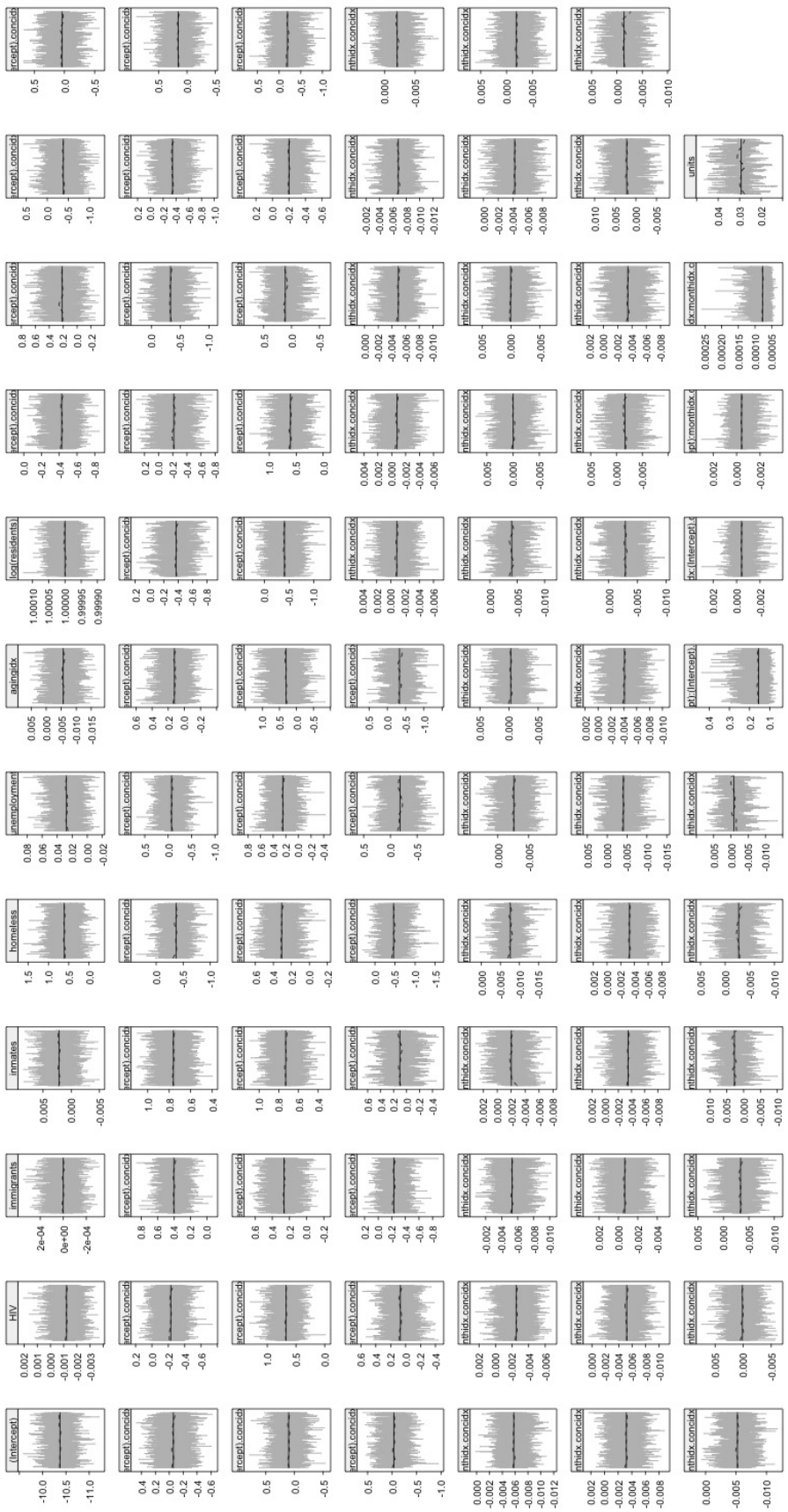
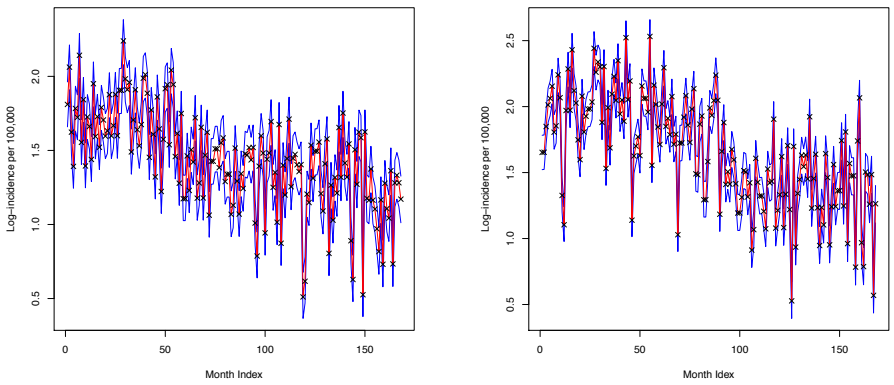


Figure 8.4: Markov Chain of GLMM in Oporto



(a) Lisbon **(b)** Oporto

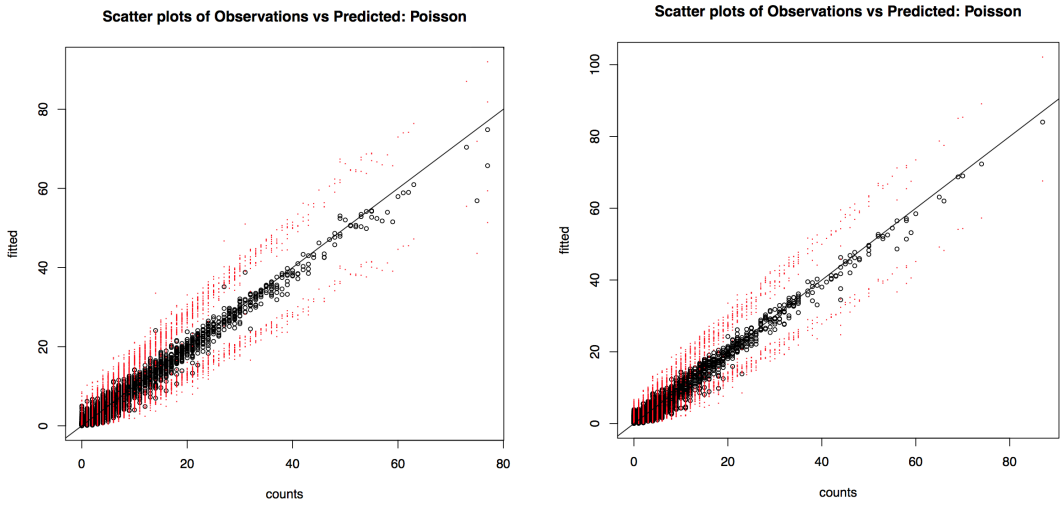
Figure 8.5: Estimates from Kalman filter for centroids of Lisbon and Oporto with 95% confidence intervals

(a) Lisbon **(b)** Oporto

Figure 8.6: Probability of random slope greater than 0. ie.($\mathbf{P}(\beta > 0)$, the TB rate is increasing)

(a) Lisbon **(b)** Oporto

Figure 8.7: Probability of TB incidences greater than 20 per $100,000$



(a) Lisbon Metropolitan Area **(b)** Oporto

Figure 8.8: Predictions vs Observations with 95% confidence bands

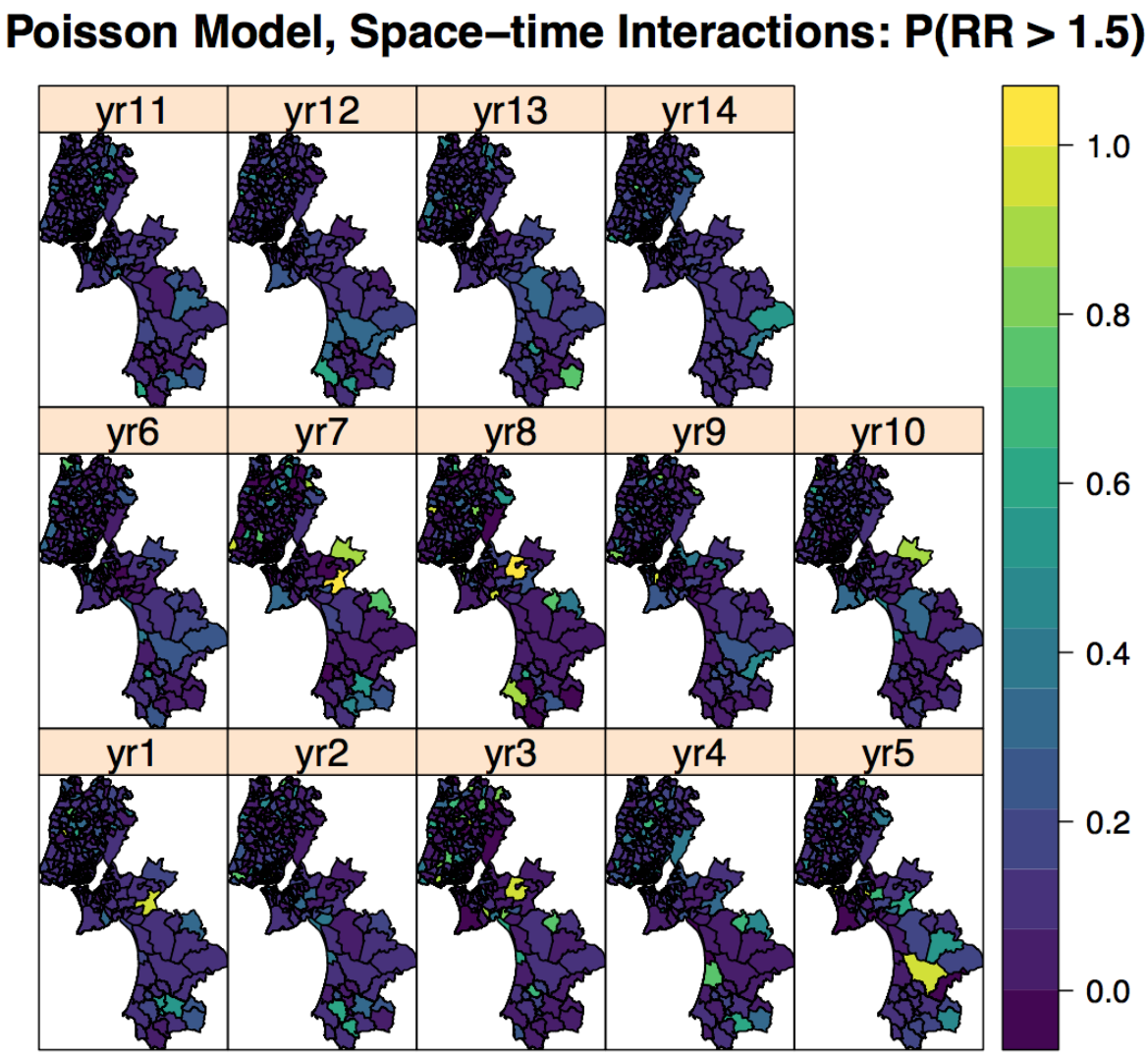


Figure 8.9: Space-time interaction trend in Lisbon Metropolitan Area: $P(\delta > 1.5)$

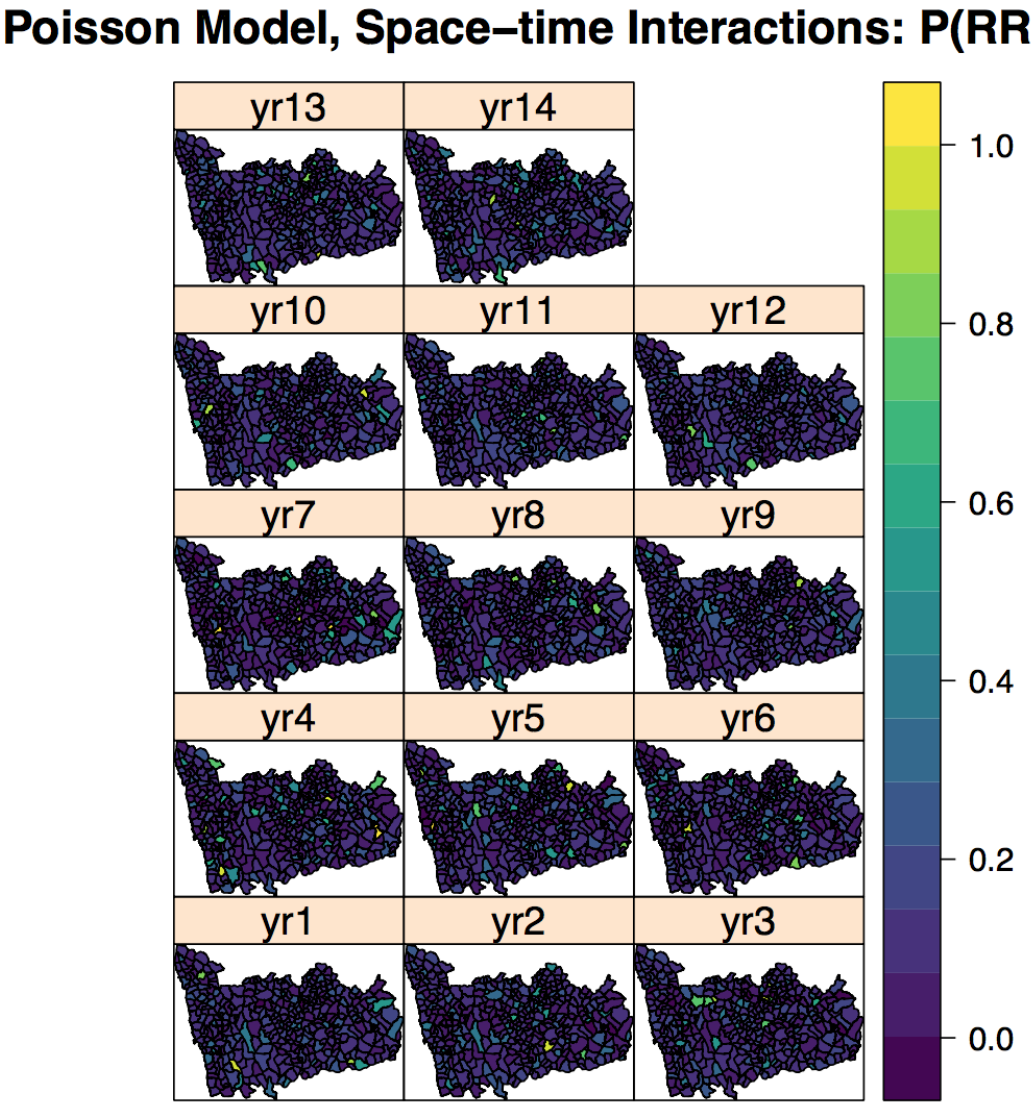
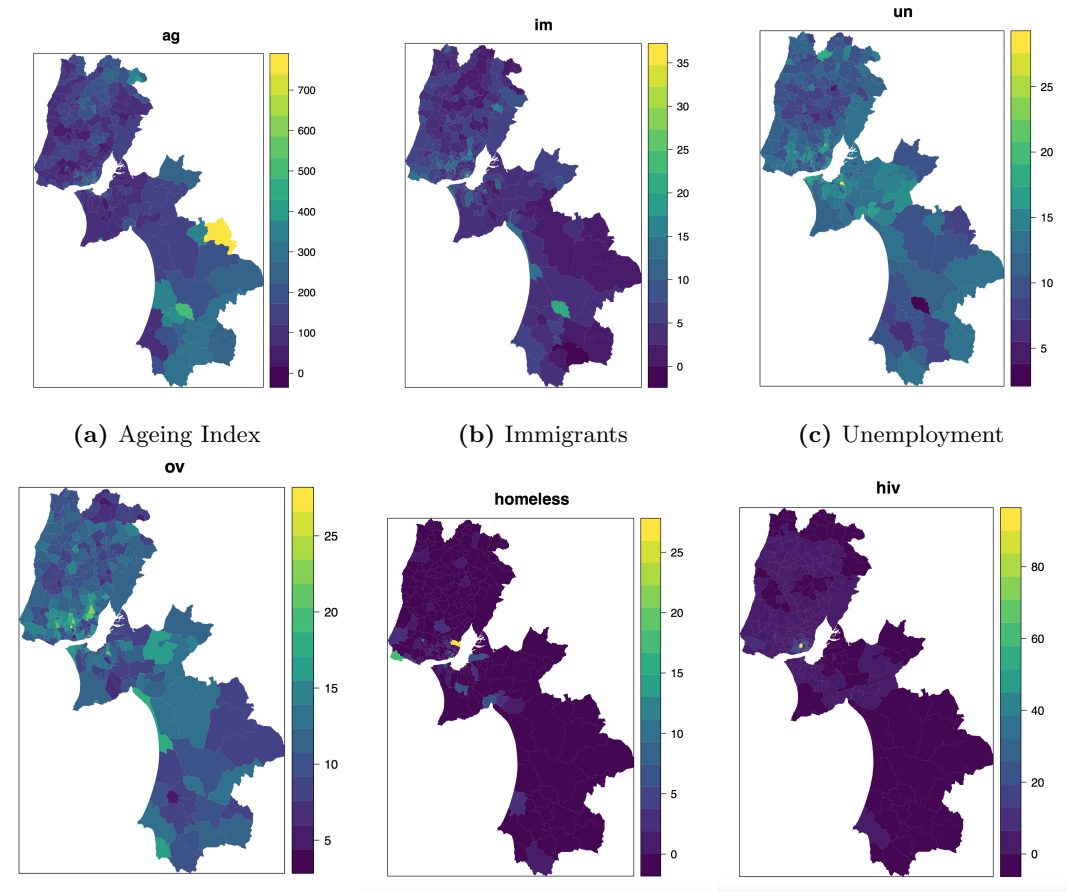


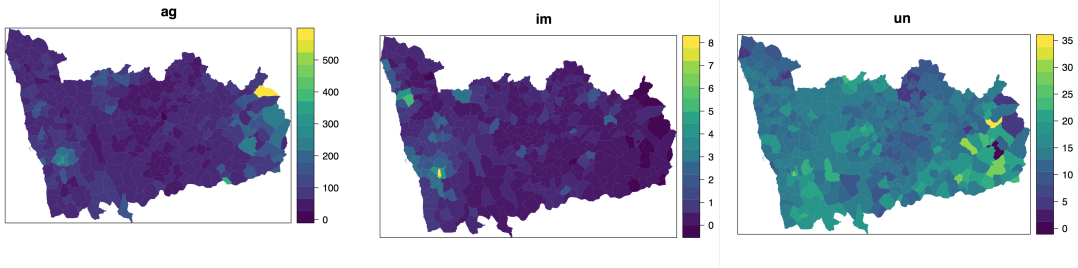
Figure 8.10: Space-time interaction trend in Oporto: $P(\delta > 1.5)$



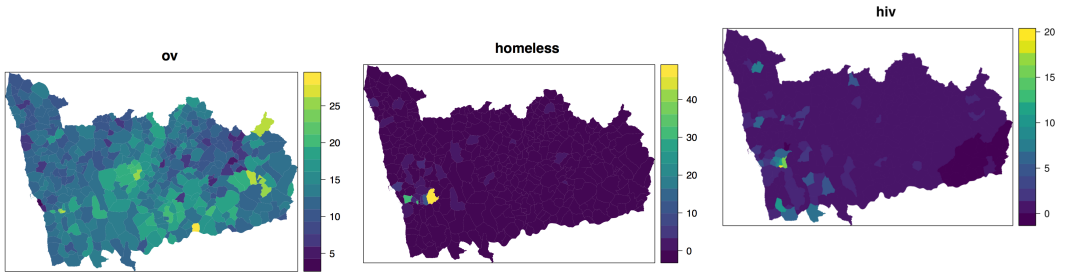
(a) Ageing Index **(b)** Immigrants **(c)** Unemployment

(d) overcrowding **(e)** homeless **(f)** HIV

Figure 8.11: Maps of Covariates in Lisbon Metropolitan Area

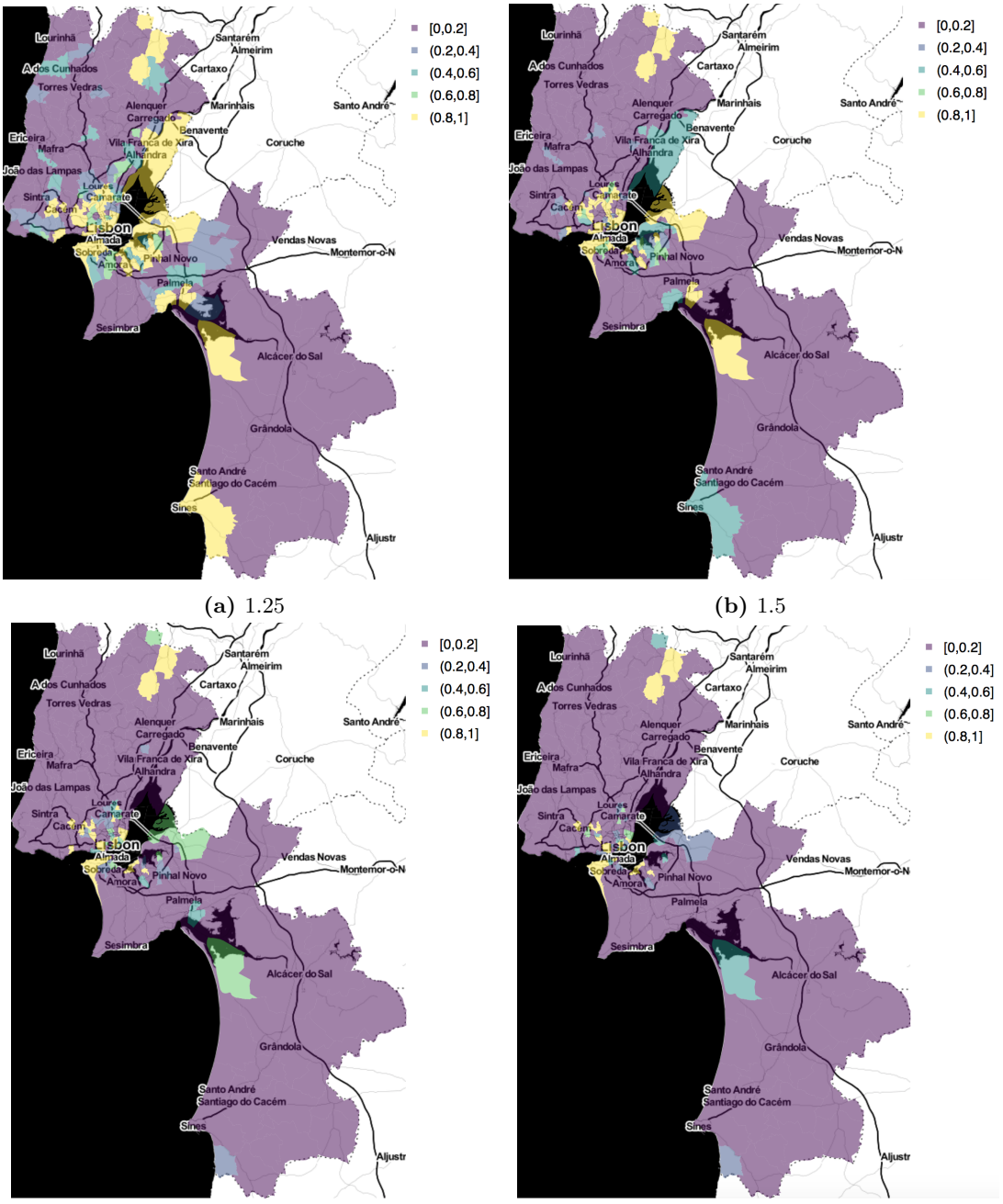


(a) Ageing Index **(b)** Immigrants **(c)** Unemployment



(d) overcrowding **(e)** homeless **(f)** HIV

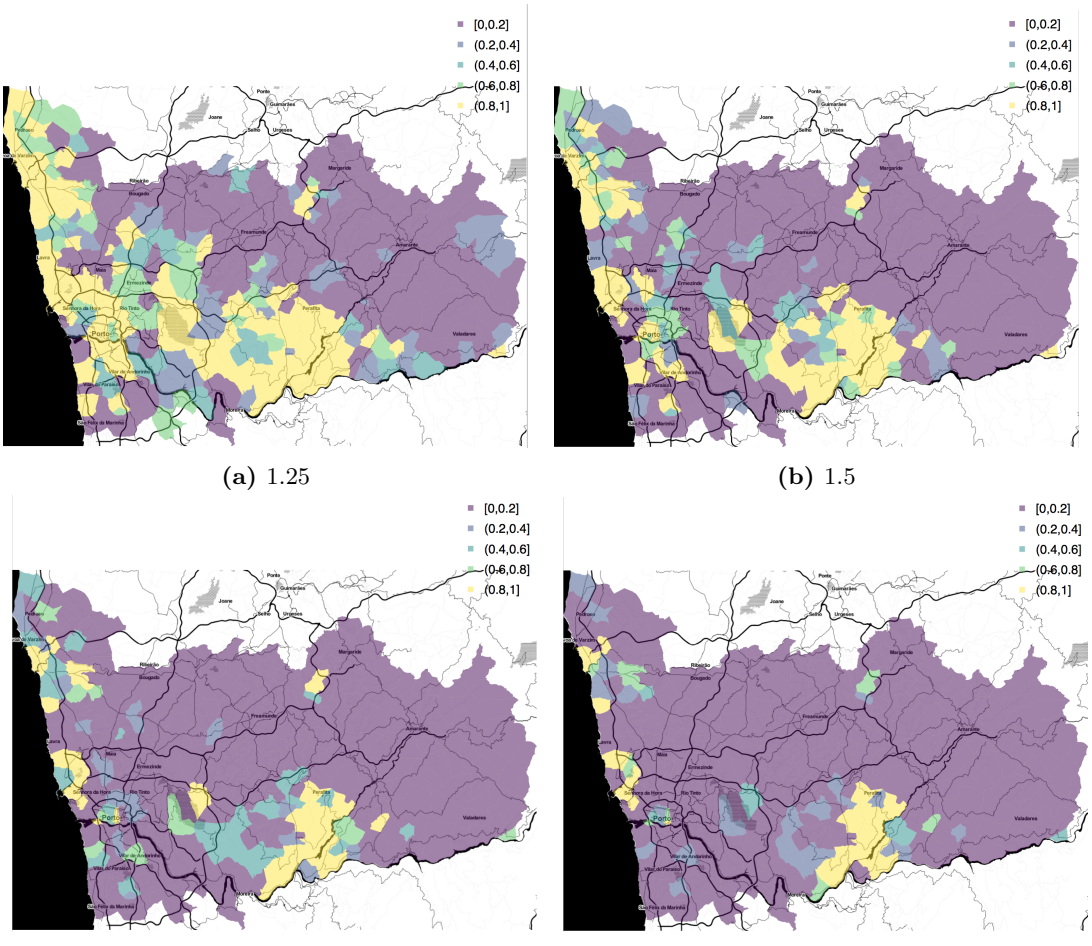
Figure 8.12: Maps of Covariates in Oporto Metropolitan Area



(a) 1.25 **(b)** 1.5

(c) 1.75 **(d)** 2.00

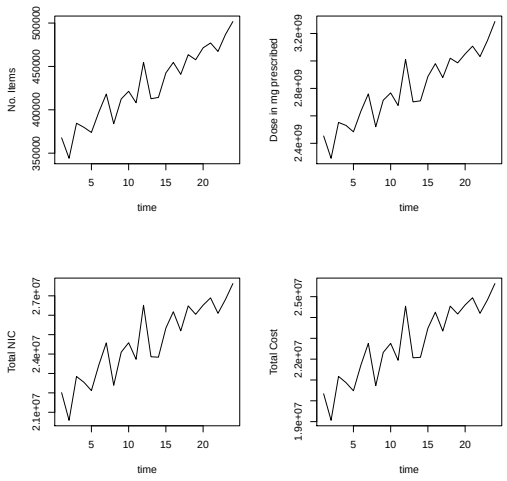
Figure 8.13: Probability of Relative Risk of TB exceeding 1.25, 1.5, 1.75 and 2.00 in the Lisbon Metropolitan Area



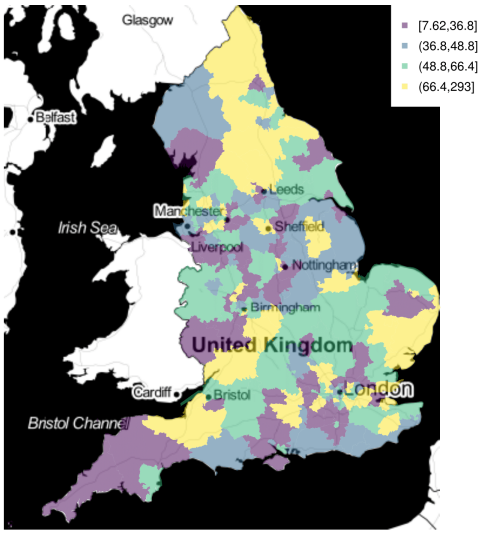
(a) 1.25 **(b)** 1.5

(c) 1.75 **(d)** 2.00

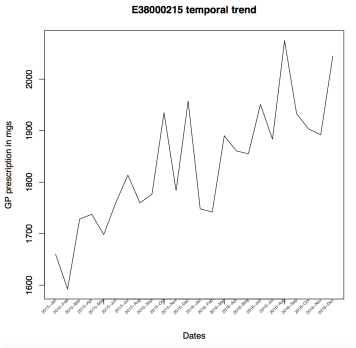
Figure 8.14: Probability of Relative Risk of TB exceeding 1.25, 1.5, 1.75 and 2.00 in the Oporto Metropolitan Area



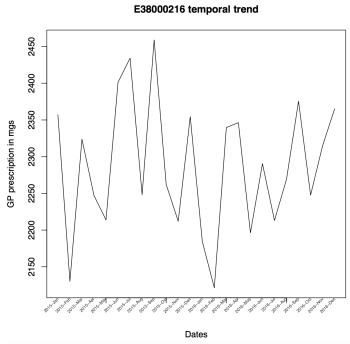
(a) Global temporal trend for Pregabalin prescriptions; averaged over space.



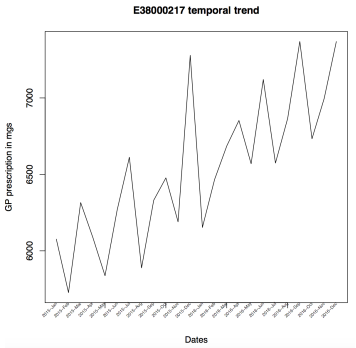
(b) Spatial trend for Pregablin prescriptions; averaged over time.



(a) Temporal trend for Pregabalin prescriptions; Cumbria.

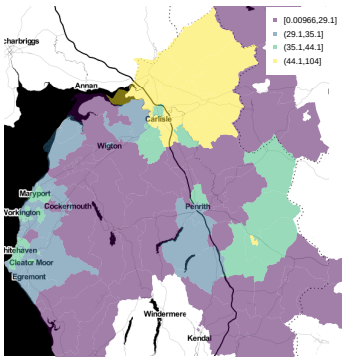


(b) Temporal trend for Pregabalin prescriptions; Morecambe.

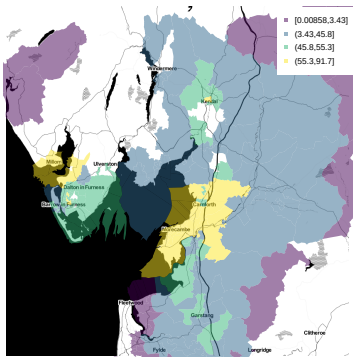


(c) Temporal trend for Pregabalin prescriptions; Manchester.

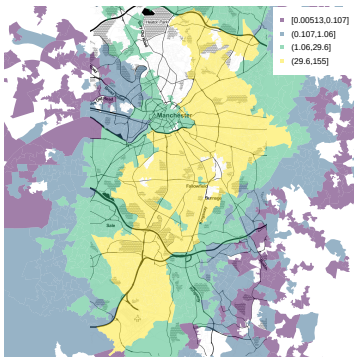
Figure 8.18: Temporal trend by CCG



(a) Pregabalin prescriptions map at LSOA level; Cumbria.

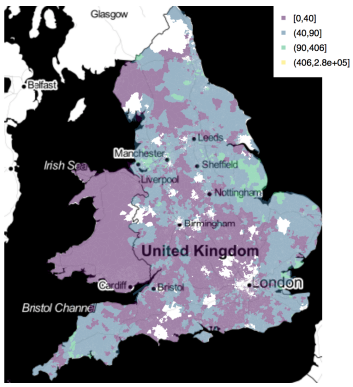


(b) Pregabalin prescriptions map at LSOA level; Morecambe.

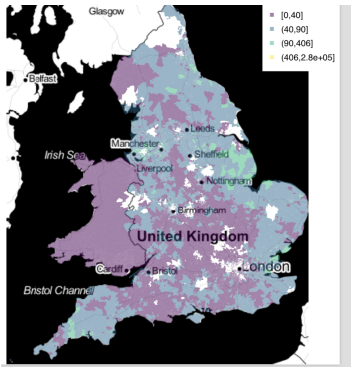


(c) Pregabalin prescriptions map at LSOA level; Manchester.

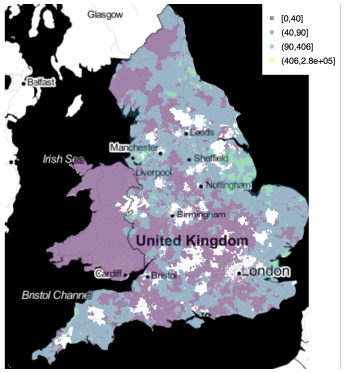
Figure 8.19: Estimated LSOA-level prescriptions for each CCG



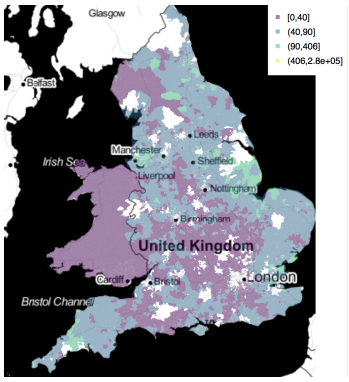
(a) Map for Pregabalin prescriptions; Jan. 2015



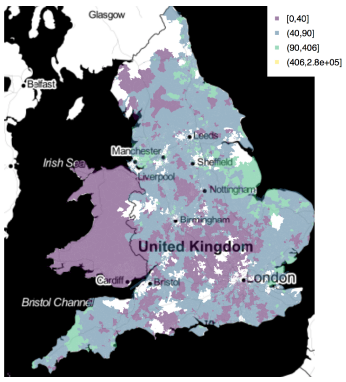
(b) Map for Pregabalin prescriptions; Mar. 2015.



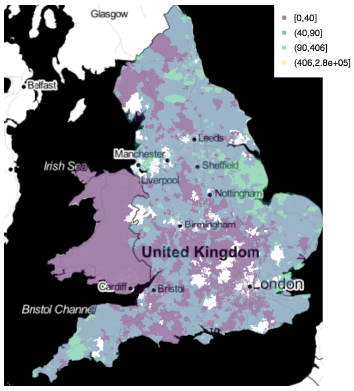
(c) Map for Pregabalin prescriptions; Jun. 2015.



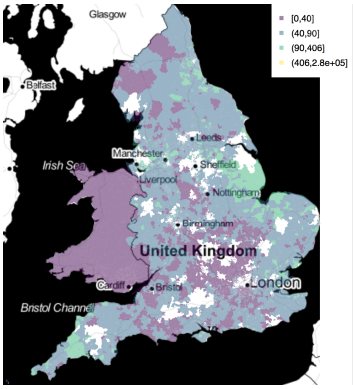
(d) Map for Pregabalin prescriptions; Sep. 2015



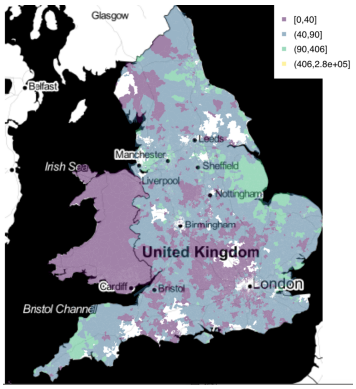
(e) Map for Pregabalin prescriptions; Dec. 2015.



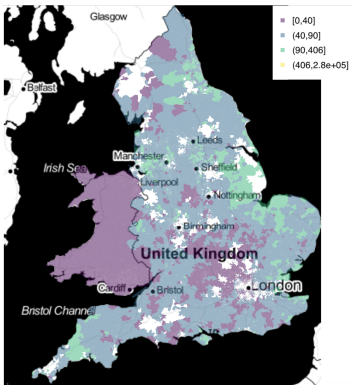
(f) Map for Pregabalin prescriptions; Mar. 2016.



(g) Map for Pregabalin prescriptions; Jun. 2016.

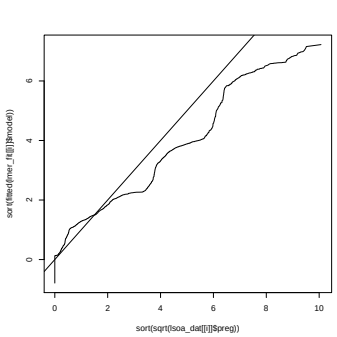


(h) Map for Pregabalin prescriptions; Sep. 2016.

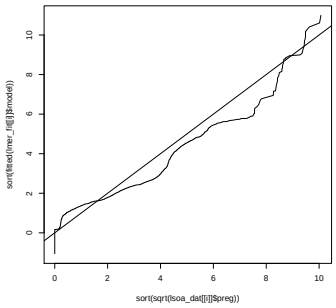


(i) Map for Pregabalin prescriptions; Dec. 2016..

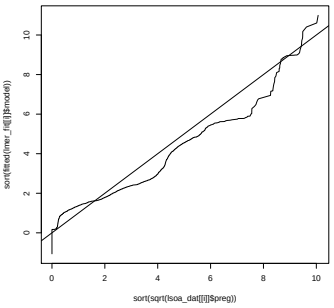
Figure 8.20: Spatial maps by LSOA



(a) Fitted vs observed; Cumbria.



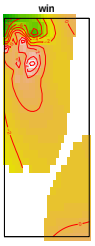
(b) Fitted vs observed; Morecambe.



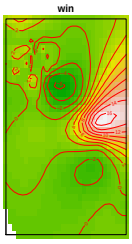
(c) Fitted vs observed; Manchester.

Figure 8.21: Fitted vs Observe LSOA-level prescriptions for each CCG

(a) Contour plot of predicted pregabalin prescription; Cumbria.



(b) Contour plot of predicted pregabalin prescription; Morecambe Bay.



(c) Contour plot of predicted pregabalin prescription; Manchester.

Figure 8.22: Contour Plots of Predicted Pregabalin Prescriptions in each LSOA

Bibliography

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics. 6*, 701–726.

Akaike, H. (1974, Dec). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*, 716–723.

Ali, M., P. Goovaerts, N. Nazia, M. Z. Haq, M. Yunus, and M. Emch (2006). Application of poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of bangladesh. *International Journal of Health Geographics 5*.

Anscombe, J. A. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika 35*, 246–254.

Ansley, C. F. and R. Kohn (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *The Annals of Statistics 13*, 1286–1316.

Antunes, M. L. and A. Fonseca-Antunes (1996). The tuberculosis situation in Portugal: A historical perspective to 1994. `http://www.eurosurveillance.org/ViewArticle.aspx?ArticleId=199`.

Areias, C., T. Briz, and C. Nunes (2015). Pulmonary tuberculosis space-time clustering and spatial variation in temporal trends in Portugal, 2000–2010: an updated analysis. *Epidemiology and Infection 143*.

Ayuso-Mateos, J., P. Barros, and R. Gusmo (2013). Financial crisis, austerity, and health in Europe. Lancet.

Banerjee, S. and B. P. Carlin (2002). *Spatial semi-parametric proportional hazards models for analysing infant mortality rates in Minnesota counties.* Springer New York.

Banerjee, S. and B. P. Carlin (2013). *Semiparametric spatiotemporal frailty modelling.*

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical Modeling and Analysis for Spatial Data (Second ed.).* Chapman and Hall/CRC.

Banerjee, S. and D. K. Dey (2005). Semiparametric proportional odds models for spatially correlated survival data. *Lifetime data analysis. 11*, 175–191.

Banerjee, S., M. M. Wall, and B. Carlin (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. *Biostatistics 4*, 123–142.

Bastos, L. and D. Gamerman (2006). Dynamic survival models with spatial frailty. *Lifetime Data Anal. 12*, 441–460.

Bayes's Portrait (1988). Bayes's portrait. The IMS Bulletin.

Belitz, C., A. Brezger, N. Klein, T. Kneib, S. Lang, and N. Umlauf (2015). ?bayesx: Software for bayesian inference in structured additive regression models.

Ben-Menachem, E. (2004). Pregabalin - from molecule to medicine - pfizer satellite symposium - international epilepsy congress - lisbon, portugal - october 15, 2003 - introduction. *Epilepsia 45.*

Bernardo, J. M. and J. O. Berger (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association 84.*

Besag, J., J. York, and A. Mollie (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics 43*, 1–21.

Besag, J., J. York, and A. Mollie (2008). Bayesian image restoration with two applications in spatial statistics. *International Journal of Health Geographics 7.*

Bhatt, V. and N. Tiwari (2014). A spatial scan statistic for survival data based on weibull distribution. *Statistics and Medicine 33*, 1867–1876.

Bicknel, I. M. (2013). The pain of pregabalin prescribing in prisons. *Br J Gen Pract 63*.

Blum, H. F. (1948). Sunlight as a causal factor in cancer of the skin of man. *J. Nat. Cancer Inst 9*.

BNF (2018). British national formulary march 2018.

Bolstad, W. M. and J. M. Curran (2017). *Introduction to Bayesian Statistics*. Wiley.

Box, G. E. P. and G. C. Tiao (2011). *Bayesian Inference in Statistical Analysis.* Addison-Wesley Pub. Co., 1973.

Bradburn, M. J., T. G. Clark, S. B. Love, and D. G. Altman (2003). Survival analysis part ii: Multivariate data analysis - an introduction to concepts and methods. *British Journal of Cancer 89*, 431–436.

Bras, A., D. Gomes, P. Filipe, B. deSousa, and C. Nunes (2015). Trends, seasonality and forecasts of pulmonary tuberculosis in Portugal. *International Journal of Tuberculosis and Lung Disease. 18*, 1202–1210.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics 30*, 89–99.

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

Brooks, S., A. Gelman, G. L. Jones, and X.-L. Meng (2017). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall / CRC.

Buchan, I., E. Kontopantelis, and M. Sperrin. North-south disparities in english mortality 1965?2015: longitudinal population study. *J Epidemiology Community Health Published Online*.

Burrough, P. A., R. McDonnell, and P. A. Burrough (1998). *Principles of Geographical Information Systems.*, Volume 13. Oxford Universtiy Press.

Cao, K., K. Yang, C. Wang, J. Guo, L. Tao, Q. Liu, M. Gehendra, Y. Zhang, and X. Guo (2016). Spatial-temporal epidemiology of tuberculosis in mainland China: An analysis based on Bayesian theory. *International Journal of Environmental Research and Public Health 13*. available on `doi: 10.3390/ijerph13050469`.

Carlin, B. P. and T. A. Louis (2008). *Bayesian Methods for Data Analysis (Third Edition ed.)*. CRC Press.

Casella, V., A. Manzino, R. Bellazzi, and M. Franzini (2015). Filtering and mapping public health data with an innovative kriging approach, accounting for single observation variance. *Procedia Environmental Sciences 26*.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika 65*, 141–151.

Collett, D. (2003). *Modelling Survival Data in Medical Research (2 ed.)*. Chapman and Hall/CRC.

Couceiro, L., P. Santana, and C. Nunes (2011). Pulmonary tuberculosis and risk factors in Portugal: a spatial analysis. *International Journal of Tuberculosis and Lung Disease. 15*, 1445–1454.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Rloyal Statistical Society Series B 34*, 187–220.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.

Cressie, N. (1994). Comment on 'An approach to statistical spatial-temporal modeling of meterological fields' by M.S. Handcock and J.R. wallis. *Journal of the American Statistical Association 89*, 379–382.

Czado, C., T. Gneiting, and L. Held (2009). Predictive model assessment for count data. *Journal of the International Biometric Society 65*, 1254–1261.

Damien, P., J. Wakefield, and S. Walker (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society Series B-Statistical Methodology 61*.

Dawid, A. (1984). Statistical theory: The Prequential approach (with discussion and rejoinder). *Journal of Royal Statistics Society A 147*, 278–292.

De Vries, G., R. Aldridge, J. Cayla, W. Haas, A. Sandgren, van Hest, N. A., and I. Abubakar (2014). Epidemiology of tuberculosis in big cities of the European Union and European Economic Area countries. *Eurosurveillanc. 19*.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B. 39*, 1–38.

DGS (2016). Relatório do Programa Nacional para a infeção VIH/SIDA. Original document in Portuguese.

DGS-TB (1994). Tuberculose em Portugal, 1993. Lisboa, 1994.

Diggle, P. (1983). *Statistical Analysis of Spatial Point Patterns.* London: Academic Press.

Diggle, P., B. Rowlingson, and T. Su (2005). Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics 16*, 423–434.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). Model-based geostatistics. *Applied Statistics 47*, 299–350.

Diva, U., D. K. Dey, and S. Banerjee (2008a). Parametric models for spatially correlated survival data for individuals with multiple cancers. *Stat. Med. 27*, 2127–44.

Diva, U., D. K. Dey, and S. Banerjee (2008b). Parametric models for spatially correlated survival data for individuals with multiple cancers. *Statistics in Medicine 27*, 2127–44.

Durbin, J. and S. Koopman (2001). *Time Series Analysis by State Space Methods.* Oxford: Oxford University Press.

Edwards, W., H. Lindman, and L. J. Savage (1963). *Bayesian statistical inference for psycho-logical research.* Psychological Review.

Efron, B. (2002). The two-way proportional hazards model. *Royal Statistical Society Series B. 64*.

Evoy, K., M. Morrison, and S. Saklad (2017). Abuse and misuse of pregabalin and gabapentin. *Drugs 77*, 403–26.

Farewell, V. and D. Cox (1979). A note on multiple time scales in life testing. *Applied Statistics. 28*, 73–75.

Farmer, P. (1997). Social scientists and the new tuberculosis. *Social Sci Med 44*.

Fink, D. (1997). Tuberculosis in the UK 2014 report. `http://www.johndcook.com/CompendiumOfConjugatePriors.pdf`.

Gamerman, D. (1991). Dynamic bayesian models of survival data. *Journal of the Royal Statistical Society. Series C. 40*, 63–79.

Gamerman, D. and H. F. Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference (2nd ed.)*. Chapman and Hall/CRC.

Gamerman, D. and M. West (1987). An application of dynamic survival models in unemployment studies. *The Statistician 36*, 269–274.

Gelfand, A. E., P. Diggle, P. Guttorp, and M. Fuentes (2010). *Handbook of spatial statistics*. CRC Press.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence. 6*, 721–741.

Gilks, W., S. Richardson, and D. Spiegelhalter (1995). *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC Interdisciplinary Statistics. Taylor and Francis.

Goldstein, M. (2006). Subjective bayesian analysis: Principle and practice. *Bayesian Analysis 1*, 403–420.

Gomez-Barroso, D., E. Rodriguez-Valin, R. Ramis, and R. Cano (2013). Spatio-temporal analysis of tuberculosis in Spain, 2008-2010. *The International Journal of Tuberculosis and Lung Disease 17*, 745–751(7).

Gray, R. J. (1992). Flexible methods for analysing survival data using splines with applications to breast cancer prognosis. *Journal of American Statistical Association. 87*, 942–951.

Gustafson, P. (1998). Flexible bayesian modelling for survival data. *Lifetime Data Analysis 4*, 281–299.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930.* Wiley, New York.

Han, A. and J. A. Hausman (1990). Flexible parametric estimation of duration and competing risk models. *Journal of Applied Econometrics 5*.

Hastie, T. and R. Tibshirani (1990). *Generalised Additive Models.* London: Chapman and Hall.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*.

Held, L., B. Schrodle, and H. Rue (2009). *Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA*. Springer.

Hemming, K. and J. E. H. Shaw (2005). A class of parametric dynamic survival models. *Lifetime Data Analysis 11*, 81–98.

Henderson, R., S. Shimakura, and D. Gorst (2002a). Modeling spatial variation in leukaemia survival data. *Journal of the American Statistical Association 97*, 965–972.

Henderson, R., S. Shimakura, and D. Gorst (2002b). Modelling spatial variation in leukemia survival data. *Journal of the American Statistical Association 97*, 965–972.

Hennerfeind, A., A. Brezger, and L. Fahrmeir (2006). Geoadditive survival models. *Journal of the American Statistical Association 101*, 1065–1075.

Hosmer, D. W., S. Jr., Lemeshow, and S. May (2008). *Applied Survival Analysis: Regression Modelling of Time to Event Data.* Wiley.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis 1*, 255–273.

Howlader, N., A. M. Noone, M. Krapcho, J. Garshell, N. Neyman, S. F. Altekrues, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich, H. Cho, A. Mariotto, D. R. Lewis, H. S. Chen, E. J. Feuer, and K. A. Cronin (2013). *SEER Cancer Statistics Review 1975-2010.* National Cancer Institute.

Huang, L., M. Kulldorff, and D. Gregorio (2007). A spatial scan statistic for survival data. *Biometrics 63*, 109–118.

Iacobelli, S. and B. Carstensen (2013). Multiple time scales in multi-state models. *Statistics in Medicine. 32*, 5315–5327.

Isaaks, E. H. and R. M. Srivastava (1989). *An Introduction to Applied Geostatistics.* Oxford University Press, New York.

Jaynes, E. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press.

Jazwinski, A. (1970). *Stochastic Processes and Filtiering Theory.* Dover Books on Electrical Engineering.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A 186*, 453–461.

Jerrett, M., S. Gale, and C. Kontgis (2010). Spatial modelling in environmental and public health research. *International Journal of Environmental Research and Public Health*.

Julier, S. J. and J. K. Ullmann (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE 92*, 401–422.

Kalman, R. (1960). A new approach to linear filtering theory. *J. Basic Engng 82*, 35–45.

Kaplan, E. L. and P. Meier (1968). Nonparametric estimation from incomplete observations. *Journal of the Amercian Statistical Association 53*, 457–481.

Karanikolos, M., P. Mladovsky, J. Cylus, S. Thomson, S. Basu, D. Stuckler, J. P. Mackenbach, and M. McKee (2013). Financial crisis, austerity, and health in Europe. *The Lancet 381*(9874), 1323 – 1331.

Kauermann, G. and P. Khomski (2006). Additive two-way hazards model with varying coefficients. *Computational Statistics and Data Analysis. 51*, 1944–1956.

Kay, R. and N. Kinnersley (2002). On the use of the accelerated failure time model as an alternative

to the proportional hazards model in the treatment of time to event data: a case study in influenza. *Drug Information Journal 36*, 571–579.

Keiding, N. (1990). Statistical inference in the lexis diagram. *Philosophical Transactions of the Royal Society of London. 332*.

Kipruto, H., J. Mung'atu, K. Ogila, A. Adem, S. Mwalili, E. Kibuchi, J. R. Ong'ang'o, and G. Sang (2013). Spatial temporal modelling of tuberculosis in Kenya using small area estimation. *International Journal of Science and Research*.

Kolmogorov, A. N. (1941). *Interpolation and Extrapolation*. Bulletin de l'academic des sciences de U.S.S.R.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy 52*.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics. 22*, 79–86.

Kulldorf, M. (2001). Prospective time periodic geographical disease surveillance using a scan statistic. `https://www.satscan.org/papers/k-jrssa2001.pdf`.

Kulldorf, M., W. Athas, E. Feuer, B. Miller, and C. Key (1998). Evaluating cluster alarms: A space-time scan statistic and brain cancer in Los Alamos. available on `https://www.satscan.org/papers/k-ajph1998.pdf`.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: theory and methods 26*.

Lawson, A. and H.-R. Song (2010). Semiparametric space-time survival modelling of chronic wasting disease in deer. *Environmental Ecological Statistics 17*, 559–571.

Lawson, A. B., H.-R. Song, B. Cai, M. M. Hossain, and K. Huang (2010). Space-time latent component modelling of geo-referenced health data. *Statistics in Medicine 29*, 2012–2027.

Lawson, A. B. and J. Zhang (2011). Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. *Journal of Applied Statistics 38*, 591–603.

Leite da Roza, D., M. Caccia-Bava, and E. Z. Martinez (2012). Spatio-temporal patterns of tuberculosis incidence in Ribeirão Preto, State of São Paulo, southeast Brazil, and their relationship with social vulnerability: a Bayesian analysis. *Revista da Sociedade Brasileira de Medicina Tropical 45*. available on `http://dx.doi.org/10.1590/S0037-86822012000500013`.

Li, Y. and X. Lin (2006). Semiparametric normal transformation models for spatially correlated survival data. *Journal of the American Statistical Association 101*.

Li, Y. and L. Ryan (2002, June). Modeling spatia survival data using semiparametric frailty models. *Biometrics 58*, 287–297.

Lin, X. and D. Zhang (1999). Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(2), 381–400.

Loquin, K. and D. . Dubois (2012). A fuzzy interval analysis approach to kriging with ill-known variogram and data. *Soft Computing 16*.

Martins, T. G., D. Simpson, F. Lindgren, and Høavard Rue (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis 67*, 68 – 83.

Matern, B. (1960). *Spatial Variation*. Springer.

Matheron, G. (1962). Trait de gostatistique applique, tome i. mmoires du bureau de recherches gologiques et minires. *Editions du Bureau de recherches gologiques et minires. Paris 14*.

Matheron, G. (1963). Trait de gostatistique applique, tome ii: Le krigeage. mmoires du bureau de recherches gologiques et minires. *Editions du Bureau de recherches gologiques et minires. Paris 24*.

Matheron, G. (1969). Le krigeage universel, cahiers du centre de morphologie mathmatique. *No. 1. Fontainebleau, France*.

Matheron, G. (1971). The theory of regionalized variables and its applications. *France: Cahiers du Centre de Morphologie Mathmatique 5*.

Mayer, J. D. (1983). The role of spatial-analysis and geographic data in the detection of disease causation. *Social Science and Medicine 17*.

Metropolis, N., R. A. W., M. N. Rosenbluth, and A. H. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika 37*, 17–23.

Murphy, A. (1996). A piecewise-constant hazard-rate model for the duration of unemployment in single-interview samples of the stock of unemloyed . *Economics Letters 51*.

National Cancer Institute. (2014). Cancer stat facts: Female breast cancer.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General 135*, 370–384.

NICE (2013). National Institute for Health and Care Excellence (2013) Neuropathic pain in adults: pharmacological management in non-specialist settings.

Nunes, C. (2008). Tuberculosis incidence in Portugal: spatiotemporal clustering. *International Journal of Health Geographics 6*.

Oakley, J. E. and A. O'Hagan (2010). SHELF: The Sheffield elicitation framework (version 2.0), school of mathematics and statistics, university of sheffield.

Office for National Statistics (2016). Number of deaths where pregabalin and gabapentin were mentioned on the death certificate, by sex and age, england and wales, 2016 registrations. `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/adhocs/007494numberofdeathswherepregabalinandgabapentinwerementionedonthedeathcertificatebysexandageengl` [Accessed 6 Dec 2017].

Onozuka, D. and A. Hagihara (2007). Geographic prediction of tuberculosis clusters in Fukuoka, Japan, using the space-scan statistic. *BMC Infectious Diseases 7*.

Paik, J. and Z. Ying (2012). A composite likelihood approach for spatially correlated survival data. *Computational statistics and data analysis 56*, 209–216.

Palm, T. A. (1890). The geographic distribution and etiology of rickets. *Practitioner 15*.

Patterson, A. (2017). Belfast's pregabalin addiction (drugs map of britain). [retrieved 2018].

Public Health England (2014). Tuberculosis in the UK 2014 report. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/360335/TB_Annual_report_-_4_0_300914.pdf.

R Core Team (2013). *R: a Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

R-INLA (2017). Log-gamma prior. https://folk.ntnu.no/hrue/r-inla.org/doc/prior/prior-loggamma.pdf.

Raiffa, H. and R. Schlaifer (1961). *Applied Statistical Decision Theory. Division of Research, Graduate School of Business Administration.* Harvard University.

Randremanana, R. V., V. Richard, F. Rakotomanana, P. Sabatier, and D. J. Bicout (2010). Bayesian mapping of pulmonary tuberculosis in Antananarivo, Madagascar. *BMC Infectious Diseases 10*. available on DOI:10.1186/1471-2334-10-21.

Rebora, P., S. Galimberti, and M. G. Valsecchi (2015). Using multiple timescale models for the evaluation of a time-dependent treatment. *Statistics in Medicine 34*.

Rieder, H. (1999). *Epidemiologic basis of tuberculosis control*. International Union Against Tuberculosis and Lung Disease, Paris: WHO (First Edition).

Ripley, B. (1981). *Spatial Statistics*. John Wiley and Sons, Inc., Hoboken, New Jersey.

Roberts, G. and J. Rosenthal (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science 16*, 351–367.

Rowlingson, B., E. Lawson, and B. Taylor (2013). Mapping english gp prescribing data: a tool for monitoring health-service inequalities. *BMJ Open 3*.

Rue, H. and L. Held (2005). *Gaussian Markov Random Fileds*. Chapman and Hall.

Rue, H., S. Marino, and N. Chopin (2009). Approximate Bbayesian inference for latent Gaussian models by using integrated nested Lplace approximations. *Journal of Royal Statistical Society. Series B. 71*, 1–35.

Rue, H., S. Martino, and H. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B. 71*, 319–392.

Rue, H., A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren (2017, March). Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application 4*, 395–421.

Sargent, D. J. (1997). A flexible approach to time-varying coefficients in the cox regression setting. *Lifetime Data Annual. 1*, 13–25.

Schifano, F. (2014). Misuse and abuse of pregabalin and gabapentin: cause for concern? *CNS Drugs 28*.

Shannon, C. E. (2005). A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423.

Shanthanna, H., I. Gilron, and M. Rajarathinam (2017). Benefits and safety of gabapentinoids in chronic low back pain: A systematic review and meta-analysis of randomized controlled trials. *PLOS Med 14*.

Simpson, D., H. Rue, G. M. Thiago, A. Riebler, and H. S. Sigrunn (2015). Penalising model component complexity: A principled, practical approach to constructing priors. `https://arxiv.org/pdf/1403.4630.pdf`.

Snow, J. (1857). On the adulteration of bread as a cause of rickets. *The Lancet 70(1766)*.

Spence, D. (2013). Bad medicine: gabapentin and pregabalin. *BMJ 347*.

Stannard, C. (2014). Misuse of gabapentin and pregabalin: a marker for a more serious malaise addiction. *Journal of Statistical Computation and Simulation 84*, 2266–2284.

Stanton, M., L. Agier, B. Taylor, and P. Diggle (2014). Towards realtime spatiotemporal prediction of district level meningitis incidence in sub-saharan africa. *J.R. Statist. Soc. A 177*, 661–678.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard University Press.

Sturges, H. (1926). ithe choice of a class interval. *Journal of American Statistics Association 21*.

Suardi, N. E., M. Preve, M. Godio, E. Bolla, R. A. Colombo, and R. Traber (2016). Misuse of pregabalin: Case series and literature review. *European Psychiatry 3*.

Tanizaki, H. (1996). *Computational Methods in Statistics and Econometrics.* CRC Press.

Taylor, B. (2015). Auxiliary variable markov chain monte carlo for spatial survival and geostatstical models. `arXiv:1501.01665v1`.

Taylor, B. M., R. Andrade-Pacheco, and H. J. W. Sturrock (2017). Continuous inference for aggregated point process data. Submitted. Preprint available from `http://arxiv.org/abs/1704.05627`.

Taylor, B. M. and P. J. Diggle (2014). INLA or MCMC? a tutorial and comparative evaluation for spatial prediction in log-gaussian cox processes. *Journal of Statistical Computation and Simulation 84*, 2266–2284.

Taylor, B. M. and B. S. Rowlingson (2014). spatsurv: an R package for Bayesian inference with spatial survival models. Available from http://cran.r-project.org/web/packages/spatsurv/index.html.

Torsen, E. (2015). Objective versus subjective bayesian inference: A comparative study. *International Journal of Advanced Research. 3*, 56–65.

Umlauf, N., D. Adler, T. Kneib, S. Lang, and A. Zeileis (2015). Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software 63*(21), 1–46.

Upton, G. and I. Cook (2014). *Oxford Dictionary of Statistics.* Oxford University Press.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*, 439–454.

Wackernagel, H. (2003). *Multivariate Geostatistics.* Springer.

Wainwright, D. Pregabalin: Spending on ?new valium? greater in north.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods.* Springer Series in Statistics.

Waller, L. A. and C. A. Gotway (2004). *Applied Spatial Statistics for Public Health Data.* Wiley.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research. 11*, 3571–3594.

Webster, R. and M. Oliver (2001). *Geostatistics for Environmental Scientists.* New York: Wiley.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models (2nd ed.)*. Springer.

WHO (1997). Treatment of tuberculosis. guidelines for national programmes. Geneva.

WHO (2007). Tuberculosis in large cities, eur/tb/fs09. `http://www.euro.who.int/__data/assets/pdf_file/0018/69030/fs09E_TBcities.pdf`.

WHO (2013). The world health report.

WHO (2014). Tuberculosis - fact sheet. `http://www.who.int/mediacentre/factsheets/fs104/en/`. [Online; accessed 16-June-2014].

Wienke, A. (2010). *Frailty Models in Survival Analysis.* Chapman and Hall/CRC Biostatistics Series CRC Press.

Wienke, A., K. Arbeev, I. Locatelli, and A. I. Yashin (2005). A simulation study of different correlated frailty models and estimation strategies. *IMathematical Biosciences. 198.*

Wood, S. N. (2017). *Generalised Additive Models: An Introduction with R (2nd edition).* Chapman and Hall/CRC.

Yashin, A. I. and I. A. Iachine (1995). Genetic analysis of durations: Correlated frailty model applied to survival of danish twins. *Genetic Epidemiology 12*, 529–538.

Zhang, J. and A. B. Lawson (2011). Bayesian parametric accelerated failure time spatial model and its application to prostate cancer. *Journal of Applied Statistics 38*, 591–603.

Zhao, F., S. Cheng, G. He, F. Huang, H. Zhang, B. Xu, T. Murimwa, J. Cheng, D. Hu, and L. Wang (2013). Space-time clustering characteristics of tuberculosis in China, 2005-2011. *PLOS ONE 8.*

Zhao, L. and T. E. Hanson (2011). Spatially dependent Polya tree modeling for survival data. *Biometrics 67*, 391–403.

Zhou, H., T. Hanson, and J. Zhang (2018). spBayesSurv: Fitting bayesian spatial survival models using R. *Journal of Statistical Software accept minor*.

Zlewicz, P. and G. Nason (2004). A haar-foisz algorithm for poisson intensity estimation. *Journal of Computational and Graphical Statistics. 13*, 621–638.