MULTI-TYPE SPATIAL AND SPATIO-TEMPORAL LOG-GAUSSIAN COX PROCESS MODELLING OF THE SEVEN GENOMIC SUB-LINEAGES OF H58 LINEAGE OF SALMONELLA TYPHI IN BLANTYRE

MASTER OF SCIENCE IN BIOSTATISTICS THESIS

DON WATSON KALONGA

UNIVERSITY OF MALAWI

JUNE 2023

MULTI-TYPE SPATIAL AND SPATIO-TEMPORAL LOG-GAUSSIAN COX PROCESS MODELLING OF THE SEVEN GENOMIC SUB-LINEAGES OF H58 SALMONELLA TYPHI IN BLANTYRE

MASTER OF SCIENCE IN BIOSTATISTICS THESIS

By

DON WATSON KALONGA

BSc in Mathematical Sciences Education - MUBAS

Submitted to the department of Mathematical Sciences, Faculty of Science, In partial fulfilment of the requirement of the degree of Master of Science (Biostatistics)

University of Malawi

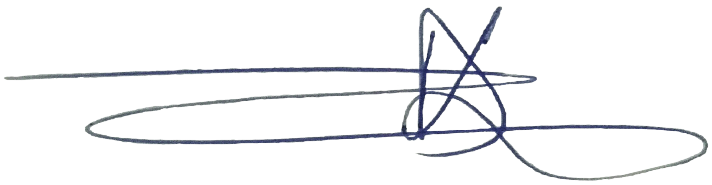
June, 2023

DECLARATION

In accordance with the regulation of the University of Malawi, I, the undersigned, hereby declare that the work described here is my own original work, except where due reference are made, and has not been submitted for a degree in any university or institution.

DON WATSON KALONGA

Full Legal Name



Signature

Date

CERTIFICATE OF APPROVAL

The undersigned certify that this thesis represents the student's own work and effort and has been submitted with my approval

Signature : _____ Date : _____

Marc Y. R. Henrion, PhD (Senior Biostatistician)

Main Supervisor

DEDICATION

This thesis is dedicated to my grandmother Maria Kalonga and my aunt Mrs Jane Chinyengo.

ACKNOWLEDGEMENTS

It has been a long and difficult journey for this thesis and the MSc in general to become a reality. This MSc would not have been possible without the support of many people. I would like to sincerely thank God for His guidance throughout my study.

I would like to thank my supervisor, Dr Marc Henrion for the support and encouragement during the development of this thesis. His comments, suggestions and patience played an enormous role in the development of this thesis. Please receive my sincere thanks. I would also like to thanks Prof. Nick Feasey for approving the use of the morbidity, carriage and genomic epidemiology of typhoid (MCET) study data for this thesis. Your guidance helped to shape the concept of this thesis.

I would like to thank by boss, Mr A. Masiye, for allowing me time off from work to attend classes and all the demands of the MSc. He shouldered most of my responsibilities so that I should have undisrupted studies. I will forever be grateful for his sacrifice. I would also like to thanks Mr Isaac Kasenjere for hosting me at his home for the entire period of my studies at the University of Malawi. You are one of the few good men I know. I will forever be grateful.

Special thanks should also go to my wife, Marie Kalonga, and my two sons, Michael and Seth for their moral support during my studies. You have been the reason I did not give up even when the going got tough multiple times.

ABSTRACT

Typhoid fever is a major cause of morbidity and mortality in low and middle-income countries. A recent epidemic of typhoid fever in Blantyre saw cases rise from 67 in 2011 to 782 in 2014. The morbidity, carriage and genomic epidemiology of typhoid (MCET) study which was conducted by Malawi-Liverpol Wellcome Trust at Queen Elizabeth Central Hospital between 2015 and 2016 found that 7 genomic sub-lineages of H58 lineage of *S. typhi* were causing the outbreak. The spatial and spatio-temporal distribution of the 7 sub-lineages, which may help to explain their transmission routes, has not yet been described. Log-Gaussian Cox Process (LGCP) models were used for the spatial and spatio-temporal analyses. Specifically, a multi-type spatial LGCP model was used for an overall spatial analysis and four spatio-temporal LGCP models (all cases and stratified by sub-lineage) were fitted for a spatio-temporal analysis of the data. The parameters in the LGCP models include σ with median of 2.185 (95% Credible Interval (CrI) 1.93 to 2.497); the parameter ϕ had a median of 940.1 metres (95% CrI 709 to 1275); and the parameter θ had a median of 0.075 months (95% CrI 0.050 to 0.107). σ is the standard deviation parameter which scales the log-intensity, whilst the parameters ϕ and θ govern the rates at which the correlation function decreases in space and in time respectively. The long term distribution of the typhoid cases show that the outbreak was at its peak between October and November 2015. The analysis provides evidence for sub-lineage specific spatial distribution of the H58 lineage of *S. typhi* during the recent typhoid outbreak in Blantyre, Malawi.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ABR	Antibacterial resistance
BTB	Bovine tuberculosis
BSI	Bloodstream infections
EA	Enumeration area
EAI	Epidemic avian influenza
ePAL	Electronic participant locator application GoM Government of Malawi
HIV	Human immunodeficiency virus
INLA	Integrated nested Laplace approximation
LGCP	Log-Gaussian Cox Process
MDR	Multidrug resistance
MCMC	Markov chain Monte Carlo
MoH	Ministry of Health
NTS	Nontyphoidal serovars of salmonella
PCF	Pairwise correlation function
QECH	Queen Elizabeth Central Hospital
ZVD	Zika virus disease

Chapter 1

INTRODUCTION

1.1 Background Information

Bacteria of the genus *Salmonella* are a major cause of foodborne illness throughout the world. As a zoonotic pathogen, salmonella can be found in the intestines of many food-producing animals such as poultry and pigs. Infection is usually acquired by consumption of contaminated water or food of animal origin: mainly undercooked meat, poultry, eggs and milk. Human or animal faeces can also contaminate the surface of fruits and vegetables, which can lead to foodborne outbreaks.[28]

Most salmonella strains cause gastroenteritis, while some strains, particularly *Salmonella enterica* serotypes Typhi and Paratyphi, are more invasive and typically cause enteric fever. Enteric (typhoid) fever is a more serious infection that poses problems for treatment due to Antibacterial Resistance (ABR) in many parts of the world. According to the study on the burden of typhoid fever in low and middle income countries suggests that 17.8 million (95% CrI 6.9 to 48.4 million)

cases occur annually.[2] However, other studies have suggest that the statistics for Sub-Saharan Africa (SSA) are not accurate because of limited health facilities with microbiological diagnostic capabilities.[22] As a result, others have suggested that the burden of typhoid fever in Africa may be over-estimated. [19]

1.1.1 Salmonella typhi in Blantyre

A longitudinal health surveillance study done in Blantyre, Malawi has shown that before 2010, most of the bloodstream infections (BSI) registered at Queen Elizabeth Central Hospital (QECH) were caused by multidrug resistant (MDR) non-typhoidal serovars of *Salmonella* (NTS) while *S. typhi* only caused 1% of the BSI [14][20][9] The study found out that between 1998 and 2010, there were only 176 microbiologically confirmed cases of *S. typhi* at QECH in Blantyre. This represents an average of 14 cases per year. Only 12 of the 176 cases were found to be MDR to ampicillin, chloramphenicol and cotrimoxazole.[9] However, from 2011, the surveillance study showed a rapid increase in microbiologically confirmed *S. typhi*. For example, 67 typhoid fever cases were confirmed in 2011 followed by 186 cases in 2012 and 843 cases in 2013 and 782 cases in 2014.[9]

In trying to understand the transmission routes of the rapid increase in cases of microbiologically confirmed *S. typhi* infections, several studies were conducted. One of the studies was the morbidity, carriage and genomic epidemiology of typhoid (MCET).[9] The aim of the study was to investigate whether the typhoid fever cases were caused by a single lineage of *S. typhi*, and to describe the full diversity of *S. typhi* in Blantyre.

In the study, patients under the age of 10 diagnosed with culture-confirmed typhoid fever at QECH in Blantyre were recruited in the prospective observational cohort study.[9] Controls were recruited in the ratio of 4 to 1.[11] A total of 314 cases consented to provide their household locations, and 256 isolates were whole genome sequenced. The results showed that prior to 2011, typhoid fever cases were being caused by four different *S. typhi* haplotype/lineages (H42,H52,H50 and H55). Typhoid fever cases caused by H58 lineage "rapidly expanded in 2011".[9] By 2013, all typhoid fever cases which were being registered at QECH were caused by H58-haplotype.[9] Further analysis of the MCET data revealed that there are 7 sub-lineages of the H58 lineage which were causing the typhoid fever outbreak.[27]

1.1.2 Factors associated with typhoid fever

Further analysis of the MCET data provided detailed insight into the risk factors for paediatric typhoid fever in Blantyre. The findings point to complex and varied risk factors including water source, household sanitation and hygiene, and social interaction patterns such as school attendance.[11] Cooking and cleaning with water from an open dug well was also identified as a risk factor. Sources of drinking water were found not to be associated with typhoid. Potential explanation was that communities are aware of the risks associated with drinking unclean water, but less aware of the risks of indirect exposure, such as through pans or other items that may come into contact with food.

Another explanation was that people may prioritize safe water for drinking but cannot afford to purchase or transport the volume of safe water needed for use

in other household tasks. It is estimated that less than 5% of the Blantyre city population is connected to the sewage network, with the majority of the population utilizing pit latrines [11] During rainy season, the runoff from the pit latrines may contaminate unprotected wells and rivers. The water from these sources are used for cooking. The study also found out that the risk of typhoid increases when using multiple drinking water sources.[11]

1.1.3 Spatial-genomic analysis of Salmonella typhi in Blantyre

Spatial-genomic data analysis of the MCET study data was done to find out if it can help shed more light on the transmission routes of the disease.[11] In the spatial-genomic analysis, a Poisson log-linear model was used to model typhoid incidences across the city, initially with the assumption of no spatial dependence. Covariates used in the model include distance to QECH, elevation and river catchment at the centroid of the Enumeration Area (EA), and average household size and population density per square km across the enumeration area.

The analysis revealed a heterogenous distribution of *S. typhi* isolates across the city.[12] The practical range of spatial correlation was approximately 192 meters, indicating the model's spatial random effect was capturing short-distance spatial correlation. Although the city′s geographical range spans approximately 20 kilometres, households in the cohort are clustered. A significant correlation between spatial and genetic distance was subsequently found, showing typhoid fever patients living closer together were more likely to have S. Typhi isolates with closely

related genomes. The analysis also showed that elevation, as a spatial factor, was not a significant risk factor of typhoid fever. This is contrary to other studies which found that low-laying areas are associated with high risk of typhoid fever.[1]

1.1.4 Point pattern analysis

Point pattern analysis focuses on describing patterns of points over space and time, and making inference about the process that could have generated an observed pattern. The main focus lies on the information carried in the locations of the points, and typically these locations are not controlled by sampling but as a result of a process of interest like typhoid fever case. Point pattern analysis is different from geostatistics where the main interest is not in the observation locations but in estimating the value of the observed phenomenon at unobserved locations. Point pattern analysis typically assumes that for an observed area, all points are available, meaning that locations without a point are not unobserved as in a geostatistics, but are observed and contain no phenomenon of interest. In point processes, locations are treated as random variables, whereas in geostatistics, the measured variable is a random variable on fixed locations.

1.2 Problem Statement

Sometimes, descriptive statistics may not be enough to fully understand the dynamics of the point process up until a spatio-temporal model is developed. Since MCET data is point pattern, spatio-temporal point process framework is better suited to estimate the intensity function which predicts the rate of typhoid fever events in space and time.

The simplest case of these class of models is the homogeneous Poisson process where the intensity is constant in space and time. A more flexible inhomogeneous model is the log-Gaussian Cox process in which the log intensity is assumed to be drawn from a Gaussian process.[?] With a suitable choice of spatio-temporal correlation function, the underlying Gaussian process can be estimated. It is also statistically prudent to model both the population density and risk as continuous phenomena in time and space while recognising, firstly that the available data will be spatially incomplete and/or aggregated as well as susceptible to measurement error, and secondly that even after modelling the effects of all candidate variables, there will often be a residual component of spatio-temporal variation in risk that can only be captured by including in the model one or more latent, spatio-temporal stochastic processes.

Other classes of spatio-temporal models have been used to analyse Salmonella infections in farm animals like dairy cattle. [10] [4] In these studies, Markov Chain model with transition probabilities and generalized linear spatial model were used to estimate the spatial and temporal patterns of the Salmonella infections in dairy herds. Ripley's K function was used to statistically identify disease clusters. The limitation of these modelling framework is that they do not incorporate the presence of a population at risk and a combination of environmental and individual characteristics that affect the risk of disease at each location in space and time in the model. Spatio-temporal point pattern process, which will be used in this project, was used to analyse the spread of Avian Influenza Virus (H5N1) in Turkey.

The intensity function which was used was based on a self exciting point process framework which has successfully been used for modelling earthquakes.[17] [21]

All the previous analyses on S. Typhi in Blantyre, Malawi did not look at the spatio-temporal point pattern signals of the sub-lineages of H58 of S.Typhi. This project, which will use the MCET dataset to fit spatial and spatio-temporal point pattern statistical models of the 7 sub-lineages of H58 lineage of S. Typhi using a log-Gaussian Cox Process as its modelling framework.

1.3 Hypothesis of the Study

The hypothesis of the study is that different sub-lineages have different spatio-temporal dynamics.

1.4 Research Questions

Based on the exploratory data analysis and the findings from the MCET study, this project will be guided by the following research questions:

1. What is the spatio-temporal distribution for typhoid cases in Blantyre?
2. Are there differences in spatial and temporal distributions between sub-lineages?
3. Is there evidence for interactions (competition, synergy) between sub-lineages?

1.5 Objectives of Study

The overall objective of this thesis project is to conduct spatial and spatio-temporal analysis of the 7 sub-lineages of H58 *S. typhi* lineage during the 2015-16 typhoid epidemic in Blantyre, Malawi.

1.5.1 Specific Objectives

1. To describe the spatial variations of the distribution of typhoid cases by sub-lineage.

2. To describe the temporal variations of the distribution of typhoid cases by sub-lineage.

3. To investigate the existence of spatial and/or temporal interactions of the 7 sub-lineages among typhoid cases.

1.6 Significance of the Study

Public health interventions for typhoid are challenging because typhoid fever incidences are often dynamic in space and time, and transmission routes are not consistent across locations and time. The results of the spatio-temporal point process model will, in general, add new knowledge to the existing body of knowledge about typhoid fever and specifically explain further the transmission routes of the H58 lineage in Blantyre, Malawi. The results would also help to develop targeted public health intervention to prevent the re-emergence of the outbreak in future.

The Ministry of Health (MoH) can also benefit from the spatio-temporal modelling framework if the modelling framework can be incorporated into the real-time surveillance systems in government hospitals and clinics. The analysis can help to identify spatio-temporal clusters of infectious diseases in real-time. The MoH can use its limited resources efficiently by prioritizing assistance to where and when it is needed most in the fight to control future infectious disease outbreaks.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

The chapter will review some of the main concepts which are used in spatial and spatio-temporal modelling. The chapter will also review the available literature on spatio-temporal modelling including point process spatio-temporal modelling of infectious diseases.

2.2 Types of Spatial Data

According to Clessie (2011), spatial data are classified into three basic types: geostatistical data, spatial point pattern data and areal spatial data. *Geostatistical data* are point observations of a continuous varying quantity over a spatial region. *Spatial Point Pattern Data* are when $Y(s)$ is a random vector at a location $s \in \mathbb{R}^d$ where s varies consistently over D, a fixed subset of $\mathbb{R}^d$ that contains a $d-$dimensional rectangle of positive volume. The data are in the form of points or events irregularly distributed within a region of space. *Areal Spatial Data* are

when D is a fixed subset (of either regular or irregular shape) but partitioned into a finite number of areal units with well-defined boundaries. Areal data are also called lattice data or aggregated data. Aggregated data usually consist of data which have been summarised into means of areal units which contain individuals that are close together[5]. According to Tranmer and Steel(1998), data aggregation process leads to loss of information about individual location and variation within groups.[26] This means that the analysis of the aggregated data cannot be used to infer individual relationships. Since the boundaries of the groups are imposed and not natural, results of the analysis can change when boundaries change. Mathematically, lattice data can be defined as $\mathbf{Z} \equiv Z(s_1), ..., Z(s_m)$, where $s_1, ..., s_m$ are reference locations for data defined on discrete spatial features.[5] The rest of the thesis focuses on spatial point pattern data.

2.3 Spatial and Spatio-Temporal Modelling Concepts

The section will present some of the key concepts used in spatial and spatio-temporal point process modelling.

2.3.1 Stochastic Process

A stochastic process is a sequence of random variables $\{X(t) : t \in T\}$, where t denotes observed time and T is the sample space which can be either discrete or continuous and can contain both negative and positive values. The stochastic process $X(t)$ is said to be strictly stationary if the distribution of $X(t_1), ..., X(t_n)$

is the same as that of $X(t_1+\tau), ..., X(t_n+\tau)$ for all choices of time lag τ and time points $t_1, ..., t_n$. Specifically, writing $p(.)$ for the distribution function, $\{X(t)\}$ is said to be strictly stationary if $p(X(t_1), ..., X(t_n)) = p(X(t_1+\tau), ..., X(t_n+\tau))$ for all sets of time points $t_1, ..., t_n$ and all time lags τ.

The basic idea of process modelling is to construct a model of a process starting from a set of sequences of events assumed to have been generated by the process itself. Subsequently, the model could be also used to discover properties of the process, or to predict future events on the basis of the past history. From a general point of view, a process model can be used for three main purposes: describing the details of a process; predicting its outcomes; predicting the effects of independent variables which have been incorporated in the model. This is different from a deterministic model, which predicts outcomes with certainty, with a set of equations that describe the system inputs and outputs exactly. Therefore, a stochastic model represents a situation where uncertainty is present. In other words, it is a model for a process that has some kind of randomness.

Gaussian processes are stochastic processes that are governed by their first two moments. As such, modelling a Gaussian process involves specifying the mean and the covariance structure. A second-order stationary spatio-temporal process $\eta(s,t) : (s,t)) \in \mathbb{R}^d \times \mathbb{R}$ has a constant first moment and there exists a function C defined on $\mathbb{R}^d \times \mathbb{R}$ such that $Cov\{\eta(s+h, t+u), \eta(s,t)\} = C(h,u)$ for $s, h \in \mathbb{R}^d$ and $t, u \in \mathbb{R}$. The function C is the space-time covariance function of the Gaussian process and its margins, $C(., 0)$ and $C(0, .)$, are purely spatial and purely temporal

covariance functions respectively.[13] "A space-time covariance function is separable if there exists purely spatial and purely temporal covariance functions C_s and C_t such that $C(h,u) = C_s(h).C_t(u)$ for all $(h, u \in \mathbb{R}^d \times \mathbb{R})$."[13] Non-separable stationary covariance functions are more realistic in real life modelling because they accommodates space-time interactions.

2.3.2 Spatial and Spatio-Temporal Point Process

Diggle (2013) defines a spatial point process as a "stochastic mechanism that generates a countable set of events s_i in the plane."[7] A spatial point process can also be defined as a stochastic process governing the location of events s_i in some set $D_s \subset \mathbb{R}^2$, where the number of such events in D_s is also random.[5]. Mathematically, a spatio-temporal point process is defined as a mechanism which generates a countable set of events $D_{s,t}$ in $\mathbb{R}^2 * \mathbb{R}^+$. Suppose T denotes the largest time in the sample space; then there exists a set of spatio-temporal point process realisations $D_{s,t}$ in $\mathbb{R}^2 * \mathbb{R}^+$ such that $D_{s,t} \subset D_s * [0,T]$.[5] An intensity function of a spatio-temporal point process, denoted as $\lambda(s,t)$, is the mean number of events per unit area and time. The intensity of a point process is fundamental in understanding the pattern of the realisation of the point process.

2.3.3 Marked Spatio-Temporal Point Process

Spatio-temporal point process may be marked if features of events beyond their time and location are also observed. Mathematically, Reinhart (2018) [23], presented the marked spatio-temporal point process as a point process of event

(s_i, t_i, k_i), where $s_i \in X \subseteq \mathbb{R}^d, t_i \in [0, T)$, and $k_i \in K$, where K is the mark space. A special case is the multivariate point process in which the mark space is a finite set $1, ..., m$ for a finite integer m. Often the mark in a multivariate point process indicates the type of each event, such as the sub-lineages of H58 lineage of *S. typhi*. A point process has independent marks if given the locations and times (s_i, t_i) of events, the marks are mutually independent of each other and the distribution of k_i depends only on s_i, t_i.[23]

2.3.4 Poisson Process

The Poisson distribution is a discrete distribution that measures the probability of a given number of events happening in a specified time period. On the other hand, a Poisson process is a series of discrete event where the average time between events is known, but the exact timing of events is random. This shows that Poisson processes are associated with Poisson distribution. A spatial point process is said to be a homogeneous Poisson process when its intensity, λ, is constant across the a bounded region A. In most cases, a homogeneous Poisson model is used as a null model against which spatial point patterns are compared. On the other hand, an inhomogeneous Poisson process assumes a nonconstant intensity function within a bounded region.

2.3.5 Log-Gaussian Cox Process (LGCP)

A Cox process is defined as a "doubly stochastic" process because it is an inhomogeneous Poisson process with a random intensity function. A spatio-temporal Cox process is a spatio-temporal Poisson process whose intensity is a realization of

a spatio-temporal stochastic process $\Lambda(s,t)$. Log-Gaussian Cox process (LGCP) is the Cox process where the log intensity function is the Gaussian process. In their seminal paper, Moller et al. (1998) demonstrated the remarkable ease with which we may theoretically decompose the log-Gaussian Cox process (LGCP), and the impressive flexibility possessed by this process with respect to capturing a wide variety of spatial intensity functions on $\mathbb{R}$.[18] This makes the LGCP better suited for solving problems in, for example, geographical epidemiology. Brix and Diggle (2001) later extended the application of the LGCP to the spatio-temporal setting.[8]

Below is the general definition of the intensity function for the spatio-temporal log-Gaussian Cox process models.

$$\begin{aligned} X(s,t) &= Pois\{R(s,t)\} \\ R(s,t) &= C_A\lambda(s,t)exp\{Z(s,t)\beta + Y(s,t)\} \end{aligned} \tag{2.1}$$

From the model, $X(s,t)$ is the number of events in the cell of the computational grid containing the point s at time t. $R(s,t)$ is the Poisson rate while C_A is the cell area. $\lambda(s,t)$ is the population offset and $Z(s,t)$ is a vector of measured covariates. $Y(s,t)$ is the latent Gaussian process on the computational grid. β represents the covariate effects. Other model parameters to be estimated include $\eta = \{log(\sigma), log(\phi), log(\theta)\}$, the parameters controlling the assumed dependence structure of the spatio-temporal Gaussian process $Y(s,t)$. The standard deviation

parameter σ scales the log-intensity, whilst the parameters ϕ and θ govern the rates at which the correlation function decreases in space and in time respectively.[7] Below is the intensity function of the multi-type spatial log-Gaussian Cox process model.

$$\begin{aligned} X_k(s) &= Pois\{R_k(s)\} \\ R_k(s) &= C_A\lambda(s)exp\{Z_k(s)\beta_k + Y_k(s) + Y_{K+1}(s)\} \qquad k \in 1, ..., K \qquad with \quad K(2.22) \end{aligned}$$

In this model, $X_k(s)$ represents the number of events of type k in the computational grid cell containing the point s. $R_k(s)$ is the Poisson rate and C_A is the cell area. $\lambda(s)$ is the population offset and $Z_k(s)$ is a vector of measured covariates and $Y_k(s)$ are latent Gaussian processes on the computational grid specific to a particular type. Y_{K+1} is the latent Gaussian process which captures spatial variation common to all types. The other parameters in the model include β_k, the covariate effects for the kth type and $\eta_k = \{log(\sigma_k), log(\phi_k)\}$ the parameters of the process Y_k for $k = 1, ..., K+1$ on a log scale. η_k controls the assumed dependence structure of the spatial Gaussian process $Y(s)$. The parameter σ_k scales the log-intensity and ϕ_k controls the rates at which the correlation function decreases in space for a particular type.[24]

Since cases of diseases occur in a spatio-temporal continuum, location and time of registration of cases are affected by the presence of a population at risk, environmental factors and individual characteristics at that location and time. As

such, LGCP are an important class of models for spatial and spatio-temporal point-pattern and lattice data as defined in sections 2.3.2 and 2.3.3 because they incorporate population at risk, environmental factors and individual characteristics at that location and time.[24]

LGCP models are also the best there is for point pattern because they recognize that the available data may be spatially incomplete and/or aggregated as well as susceptible to measurement error. They also recognize that even after modelling the effects of all environmental variables, there will often be a residual component of spatio-temporal variation in risk that can only be captured by including in the model one or more latent spatio-temporal stochastic processes.[24]

2.4 Spatial and Spatio-Temporal Modelling Approaches in Real World

This section discusses past statistical approaches to spatial and spatio-temporal modelling of typhoid fever or other infectious diseases.

2.4.1 Spatial Poisson Log-linear Model

Gauld et al (2021) analysed the MCET data which is also be used for the spatio-temporal analysis in this paper. Their aim was to estimate the incidence of typhoid fever across Blantyre city. They fitted a Poisson log-linear model using areal spatial data aggregated at EA level unlike point pattern data which will be used in this paper. The model included covariates like distance to QECH, elevation, river catchment at the centroid of the EA, average household size and population

density per square km across the enumeration area. The analysis found that the typhoid fever cases where heterogeneously distributed across Blantyre city. The analysis also showed that elevation, a spatial factor, was not a significant risk factor of typhoid fever contrary to Akullian et al (2015) who found that low-laying areas have higher incidence of typhoid fever that highlands.[12] [1] The initial model assumption of no spatial dependence is one of the main weakness of the model. The model also failed to include an autoregressive term to handle spatial dependence among the typhoid cases. The model also neglected the temporal analysis of the disease outbreak.

2.4.2 Epidemic Avian Influenza (EAI) Model

Kim (2011) adapted a self exciting point process introduced by Hawkes et al[15] to implement a point process spatio-temporal model to describe the spatial distribution of Turkey's first avian influenza in 2006.[17] The intensity function of the self-exciting point process conditional to the past history of time and space of the Epidemic Avian Influenza (EAI) model was defined as follows:

$$\begin{aligned}\lambda(x,y,t \mid H_t) &= \frac{E\left[N(dtdxdy) \mid H_t\right]}{dtdxsy} \\ &= \lambda_B(x,y,t) + \sum_{i:t_i<t} \alpha f(x-x_i, y-y_i) g(t-t_i) h_{traff}(x,y) k(T(t_i))\end{aligned}$$

where i is an index for each of the 221 H5N1 outbreaks occurred in Turkey during 182 days between October 1, 2005 to March 31, 2006. (x_i, y_i, t_i) represents the location and time of an outbreak i and $H_i = x_i, y_i, t_i; t_i < t$ represents the past history of outbreaks up to time t. The subscripts B and T of the conditional

intensity represents background and triggering respectively.

The background conditional intensity was defined as $(\lambda_{\mathbf{B}}(\mathbf{x}, \mathbf{y}, \mathbf{t})) = \mathbf{a}\mathbf{e}^{-\mathbf{b}\mathbf{R}_{\mathbf{city}}(\mathbf{x},\mathbf{y})}\mathbf{e}^{-\mathbf{k}\mathbf{T}(\mathbf{t})}$ representing the background intensity and has two components one for space and the other for temporal patterns of the outbreaks. Backfitting, Expectation Maximization (EM) and Poorman's EM methods were used for parameter estimation of the EAI model.[17]

2.4.3 Spatio-Temporal Interaction Effects Model for Zika Virus Disease (ZVD) and Dengue Fever

Bello (2018) estimated the parallel relative risk of Zika virus disease (ZVD) and dengue fever using spatio-temporal interaction effects models for one department and one city of Colombia during the 2015-2016 ZVD outbreak. The model used lattice data and was fitted using the integrated nested Laplace approximation (INLA) for parameter estimation[3]. Even though the model incorporated inseparable spatio-temporal interactions, it modeled the ZVD and dengue separately before combining the risk to obtain a joint risk estimation.

2.4.4 Log-Gaussian Cox Process Model for Ambulance Calls in Northern Sweden

Basiya et al (2020) analysed the spatio-temporal pattern of ambulance calls which occured between 2014 and 2018 in Sweden. An LGCP was used to model the ambulance calls. Spatial component of the stochastic intensity function was estimated using K-means clustering based bandwidth selection method. The temporal

intensity component of the stochastic intensity function was estimated by means of Poisson regression model which had temporal covariates like days of a week and season of the year. Inhomogeneous $K-$function was used to assess the existance of spatio-temporal clustering. Minimum contrast technique was employed to estimate variance parameter, spatial correlation function and the temporal correlation function.[?] The analysis showed that LGCP was a suitable model for point pattern data.

2.4.5 Multivariate log-Gaussian Cox Process Model for Modelling Bovine Tuberculosis (BTB)

Diggle et al. (2013) modelled BTB, an infectious disease in United Kingdom (UK), using multivariate LGCP model. As part of a national control strategy for the disease, herds in UK undergo regular inspection. In a study done in 2013, whole genome sequencing was done on the tuberculosis bacterium that caused the outbreak. It was found out that there were multiple strains which were causing the outbreak. This necessitated the fitting of the multi-type LGCP model given by 2.2

The model allows decomposition of the spatial variation in events of multiple types into variation associated with a particular type of event and variation common to all types. Thus, although each point type may display an individual spatial pattern, the process Y_{K+1} captures area of high or low intensity that are common to all types.

Chapter 3

METHODOLOGY

This chapter discusses the methodology used in this study including the study design, data collection and model fitting.

3.1 Study Design

The data used in this project is from the morbidity, carriage and genomic epidemiology of typhoid (MCET) study which was conducted between March 2015 and December 2016 at QECH. The hospital provides free secondary healthcare to the Blantyre urban area and surrounding district and tertiary care to the southern region of Malawi. Since 1998, the Malawi Liverpool Wellcome Programme conducted sentinel surveillance of BSI at QECH[20]. Patients under 10 years living in Blantyre who had blood culture confirmed typhoid fever diagnosed between March 2015 and December 2016 at QECH were included in a prospective observational cohort.[9] Age, residential area, human immunodeficiency virus (HIV) status, inpatient versus outpatient treatment, clinical presentation, complications, and deaths were recorded from clinical case records and/or during patient interviews.[9]

3.2 The MCET Data

Most salmonella strains cause gastroenteritis, while some strains, particularly *Salmonella enterica* serotypes Typhi and Paratyphi, are more invasive and typically cause enteric fever. Enteric (typhoid) fever is a more serious infection that poses problems for treatment due to antibacterial resistance (ABR) in many parts of the world. According to the study on the burden of typhoid fever in low- and middle-income countries (South America, Sub-Saharan Africa and South-East Asia) suggests that 17.8 million (95% CrI 6.9 to 48.4 million) typhoid fever cases occur annually.[2]

In Malawi, a longitudinal health surveillance study conducted at QECH in Blantyre showed a rapid increase in microbiologically confirmed *S. typhi* infections[9]. For example, 67 typhoid fever cases were confirmed in 2011 followed by 186 cases in 2012 and 843 cases in 2013 and 782 cases in 2014[9]. In trying to understand the transmission routes of the rapid increase in cases of microbiologically confirmed *S. typhi* infections, the Morbidity, Carriage and genomic Epidemiology of Typhoid (MCET) study was conducted in 2015[9]. The aim of the study was to investigate whether the typhoid fever cases were caused by a single lineage/haplotype of *S. typhi*, and to describe the full diversity of *S. typhi* in Blantyre.

In the study, patients under the age of 10 diagnosed with culture-confirmed typhoid fever at QECH in Blantyre were recruited in the prospective observational cohort study.[9] The results showed that prior to 2011, typhoid fever cases were be-

ing caused by four different *S. typhi* haplotype/lineages (H42,H52,H50 and H55). Typhoid fever cases caused by MDR lineage H58 increased sharply in 2011. By 2013, all typhoid fever cases which were being registered at QECH were caused by H58-haplotype/lineage.[9] Further analysis of the MCET data revealed that there are 7 sub-lineages of the H58 lineage which were causing the typhoid fever outbreak.[12]

3.3 Model Specification and Statistical Analysis

This section will discuss specific LGCP models which have been fitted to investigate the spatial and spatio-temporal distribution of the sub-lineages of the H58 lineage of *S. typhi*.

3.3.1 Statistical Models

The MCET study data consists of geo-referenced point pattern data over time. Therefore, a spatio-temporal point process model will be appropriate to model these data. Since the aim of the project is to investigate heterogeneities in process intensity over time and geographical space, the LGCP model, introduced in Chapter 2, provides the flexibility required for this application. The general spatio-temporal and multivariate spatial LGCP equations in 2.1 and 2.2 respectively will be implemented as specified below. The intensity function of the spatio-temporal model will be expressed as

$$
\begin{aligned}
X(s,t) &= Pois\{R(s,t)\} \\
R(s,t) &= C_A\lambda(s,t)exp\{Y(s,t)+S(t)\} \qquad where \quad S(t)=\sum_{i=0}^{n} N_{i,k}(t)P_i
\end{aligned} \tag{3.1}
$$

In equation 3.1, $X(s,t)$ is the number of events in the cell on the computational grid at location s and at time t. $R(s,t)$ is the spatially and temporally varying Poisson rate or the intensity of the point process. C_A is the cell area and $\lambda(s,t)$ is the population offset. $Y(s,t)$ is the latent Gaussian process on the computational grid. $S(t)$ is the cubic B-spline to model the temporal trends of the typhoid cases. $\{N_{i,k}\}_{i=1}^{n}$ are piecewise polynomial basis functions defined using the Cox-de Boor recursion formula below:

$$
N_{i,0}(t) = \begin{cases} 1 & if \quad t_i \leq t < t_{i+1} \\ 0 & otherwise \quad , \end{cases}
$$

$$
N_{i,k}(t) = \frac{t-t_i}{t_{i+k}-t_i}N_{i,k-1}(t) + \frac{t_{i+k+1}-t}{t_{i+k+1}-t_{i+1}}N_{i+1,k-1}(t)
$$

where $k = 0, 1, 2, 3$ is the degree of the spline. P_i are the control points of the B-spline. The three derived vectors of parameter values which affect the shape of the B-spline curve will be represented by $t^{'}, t^{''}$ and $t^{'''}$ in the proceeding sections. The population counts of children aged 10 and below from the 2018 Malawi population census have been used as the population offset for the models. This will help to account for differing population density across the city and to interpret the results as incidence rates instead of counts.

Since the MCET study data are marked spatio-temporal point pattern, it is also important to fit a multi-type spatio-temporal model to assess the spatial and temporal distribution of typhoid fever cases caused by specific genomic sub-lineages of the H58 lineage of *S. typhi*. However, due to the time constraints inherent to an MSc project and the limitation of the current *lgcp* R package which cannot fit a multi-type spatio-temporal LGCP model, a multi-type spatial-only LGCP model will be fitted instead. The intensity function for the multi-type spatial LGCP model is defined as

$$\begin{aligned} X_m(s) &= Pois\{R_m(s)\} \\ R_m(s) &= C_A\lambda(s)exp\{Y_m(s) + Y_{m+1}(s)\} \qquad m = 1, 2, 3. \end{aligned} \tag{3.2}$$

In equation 3.2, $X_m(s)$ represents the number of events of type m on the computational at location s. $R_m(s)$ is the intensity of the LGCP and C_A is the cell area. $\lambda(s)$ is the population offset and $Y_m(s)$ are latent Gaussian processes on the computational grid specific to a particular type. Y_{m+1} is the latent Gaussian process which captures spatial variation common to all types. $m = 1, 2, 3$ because there are 3 types (clade 0, 2 and the combined clade with sub-lineages 1,3,4,5 and 6).

Theoretically, LGCP models are formulated on a spatio-temporal continuum. However, actual implementation is done by approximating the study region using a grid of square cells. In this article, 350m was used as grid cell side length. Minimum contrast estimation also known as least-squares approach, a non-parametric

method, was used to determine the computational grid size which helped to capture the spatial dependence in the latent Gaussian process.[18] [6]

3.3.2 Bayesian Estimation

Gibbs sampling is one of the Monte Carlo Markov Chain algorithm which can be used to sample from a given multivariate probability distribution. Since it is not possible to sample from this distribution, the Gibbs sampler iteratively draws an instance from the conditional distribution of each variable, conditional on the current values of the other variables for parameter estimation of the multivariate probability distribution. Even though Gibbs sampling is relatively easy to implement, it is not very efficient when tracking many parameters like in spatial and spatio-temporal modelling. The *lgcp* package solved this problem by combining Gibbs sampling with the Metropolis-adjusted Langevin algorithm (MALA). MALA is an adaptive MCMC algorithm which helps the sampler to move towards areas of higher posterior probability.[25]

For valid and reliable statistical inference, the MCMC needs to converge to a stationary distribution. In this thesis, the MCMC for the spatio-temporal LGCP models were run 1,000,000 iterations with 100,000 iterations as burn-in. The MCMC for the multi-type spatial model was run 40,000,000 iterations with 500,000 iterations as burn-in. Pilot runs of 100,000 iterations helped to determine the aforementioned number of iterations. Trace plots were used to assess convergence and to determine the number of iterations.

All statistical analyses were conducted using the R statistical software, version 4.0.3. Firstly, descriptive analyses were done to summarise the data. The LGCP models were fitted using the R package *lgcp*[25][24]. The *lgcp* package uses Bayesian inference to get the joint predictive distribution of the latent Gaussian process where summaries of the predictive distribution for incidence rates and exceedance of a prespecified incidence threshold can be made. The covariance matrix of the latent Gaussian process on the computational grid will be identified by σ, ϕ and θ parameters. The parameter σ scales the log-intensity whereas the parameters ϕ and θ manage the rates at which the correlation function decreases in space and in time respectively.[25][24]

The spatio-temporal analyses in this thesis project assume a separable covariance structure. This means that the correlation between two points in space and time can be decomposed into purely spatial and purely temporal components2.3.1.[6] An exponential covariance function was used to model the spatial dependency in the Gaussian process. The prior densities for spatio-temporal models are a multivariate normal specified as follows: $\eta = \{log\sigma, log\phi, log\theta\} \sim MVN(\mu_\eta, \mathbf{\Sigma}_\eta, \mathbf{\Omega}_\eta)$. For the multi-type spatial model will be as follows: the multivariate Gaussian prior for β will be $\beta \sim MVN(\mu_\beta, \mathbf{\Sigma}_\beta)$ and for the multivariate Gaussian prior on the log-scale for the positive parameters σ and ϕ will be $\eta = \{log\sigma, log\phi\} \sim MVN(\mu_\eta, \mathbf{\Sigma}_\eta)$.[25]

The four spatio-temporal LGCP models and the multi-type spatial LGCP model were fitted using the *lgcp* package functions *lgcpPredictSpatioTemporalPlusPars*

and *lgcpPredictMultitypeSpatialPlusPars* functions respectively. The exceedance and segregation probabilities of the latent Gaussian process were also calculated. Exceedance probability, $P[exp(Y) > k]$, is the probability that the incidence rate exceeds threshold k for each cell on the computational grid. On the other hand, segregation probabilities are used in the multivariate LGCP models to ascertain locations of high probability that a particular location will have an event of a particular type. In this thesis, a sub-lineage was considered dominant in a particular location if the conditional probability of an event of a given type at that location exceeded 0.8 threshold. [24]

3.4 Study Outcomes

The outcome of the thesis are the spatial and spatio-temporal LGCP models which have been fitted with the use of intensity functions specified in section 3.3.1 above. The results of the models will help to describe the spatial and spatio-temporal variations of the distribution of typhoid cases by genomic sub-lineages. The results will also help in understanding if there were spatial and/or temporal interactions of the sub-lineages among typhoid cases.

3.5 Ethical Considerations

The MCET study was approved by the University of Malawi, College of Medicine Research and Ethics Committee (no. P.08/14/1617), the Liverpool School of Tropical Medicine Research Ethics Committee (no. 14.042), and the Lancaster University Faculty of Health and Medicine Ethics Committee (no. FHMREC17014).

Informed written consent was sought from adult participants and from the legal guardians of children whose data have been used in this paper.

Chapter 4

RESULTS AND DISCUSSION

This chapter presents and discusses the results of fitted models that have been obtained from the study analysis. Section 4.1 presents the exploratory data analysis, section 4.2 presents the results of the fitted models, and section 4.3 presents model assumptions and diagnostics.

4.1 Exploratory Data Analysis

The final merged MCET dataset which has been used in this project have 549 observations. These observations are blood culture-confirmed typhoid fever cases of patients, under the age of 10 years, who resides in urban Blantyre. Out of these, 316 cases have spatial information while 540 cases have temporal information. 310 cases have both spatial and temporal information and 255 of the 310 cases have genomic information. Figure 4.1 shows the merging process of the data tables.

Variable	Level	n	%	Missing	%	Mean	SD	Median	Minimum	Maximum
TCCB ID	-	543	98.9%	6	1.1%	-	-	-	-	-
ePAL ID	-	275	50.1%	274	49.9%	-	-	-	-	-
Sample ID	-	542	98.7%	7	1.3%	-	-	-	-	-
Date	-	540	98.4%	9	1.6%	-	-	-	-	-
Latitude	-	316	57.6%	233	42.4%	-15.79	0.04	-15.78	-15.87	-15.70
Longitude	-	316	57.6%	233	42.4%	35.04	0.03	35.03	34.97	35.10
Lineage	-	255	46.4%	294	53.6%	-	-	-	-	-
	Clade 0	132	51.8%*	-	-	-	-	-	-	-
	Clade 1	14	5.5%*	-	-	-	-	-	-	-
	Clade 2	47	18.4%*	-	-	-	-	-	-	-
	Clade 3	14	5.5%*	-	-	-	-	-	-	-
	Clade 4	9	3.5%*	-	-	-	-	-	-	-
	Clade 5	15	5.9%*	-	-	-	-	-	-	-
	Clade 6	24	9.4%*	-	-	-	-	-	-	-

* the percentages for the different clades are out of 255 and not 549 as for all the other percentages

Table 4.1: Summary of MCET Data

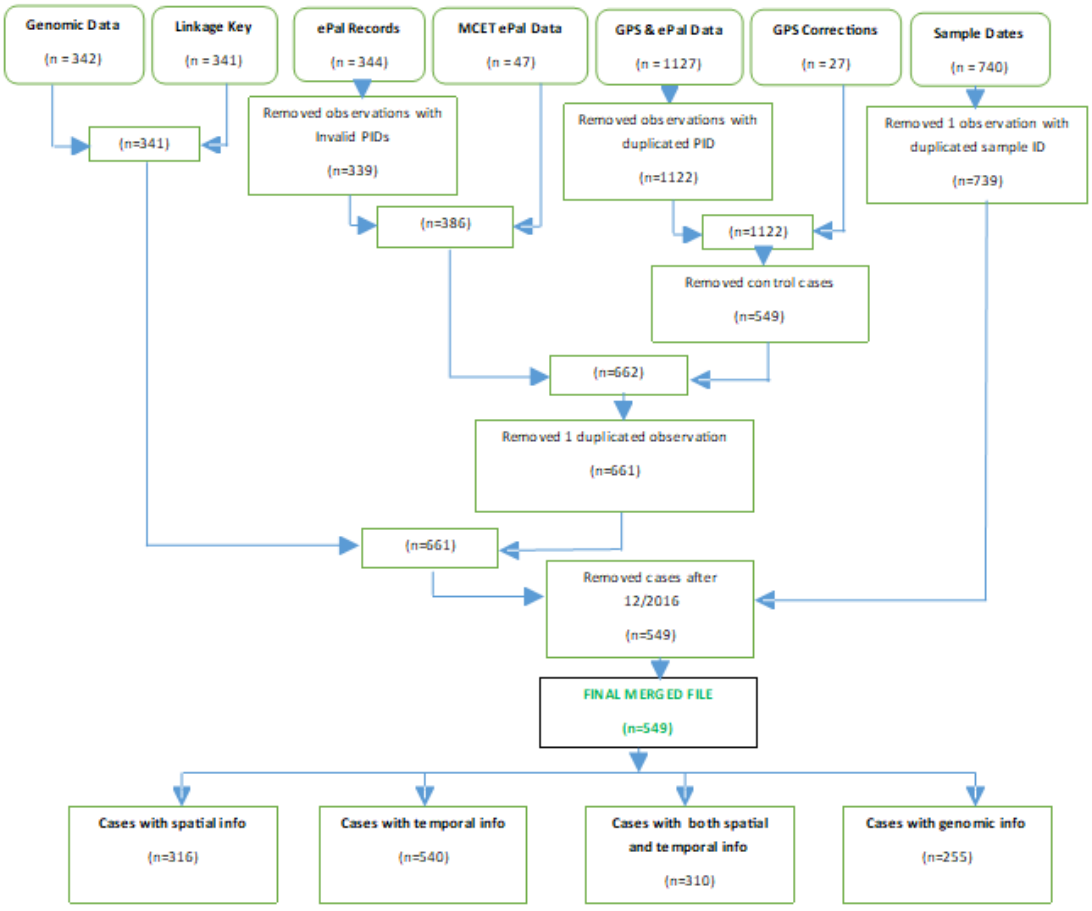


Figure 4.1: Flow diagram for the data merging process

Table 4.1 summarises the variables in the merged data. The variables used for the spatial and spatio-temporal analysis are *Date* which is a date variable, and

Latitude and *Longitude* which are latitude and longitude variables. The variable *Lineage* records genomic data. The patients and sample ID variables (TCCB ID, ePAL ID and Sample ID) were used for merging the different data tables.

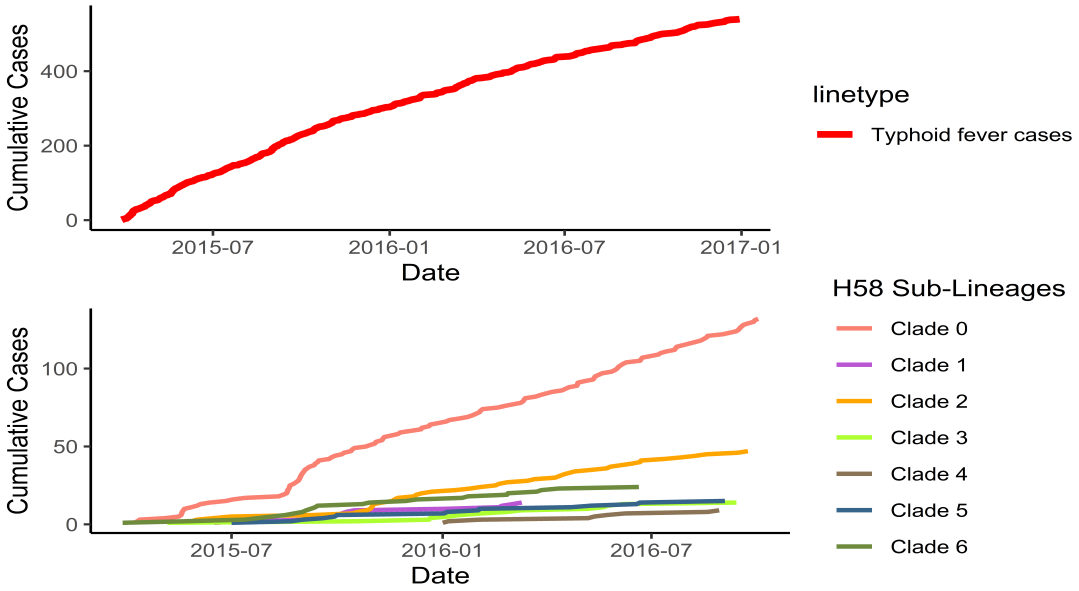


Figure 4.2: Top: Cumulative cases over time for all 540 typhoid cases. Bottom:Cumulative cases of 255 typhoid cases with genomic data over time

Figure 4.2 (Top) shows the cumulative frequency of the 540 cases with recorded temporal information during the study period. The first case was recruited on 28th March 2015. The last case was recruited on 30th December 2016. Number of cases per day ranged from 1 to 6 cases. The figure shows that there was a steady increase of typhoid fever cases across the study period. Figure 4.2 (Bottom) shows the cumulative frequency of the 255 typhoid fever cases for which whole genome sequencing (WGS) was performed. clade 0 and 2 were recorded consistently throughout the study period. Cases for clade 4 sub-lineage started appearing only from January 2016.

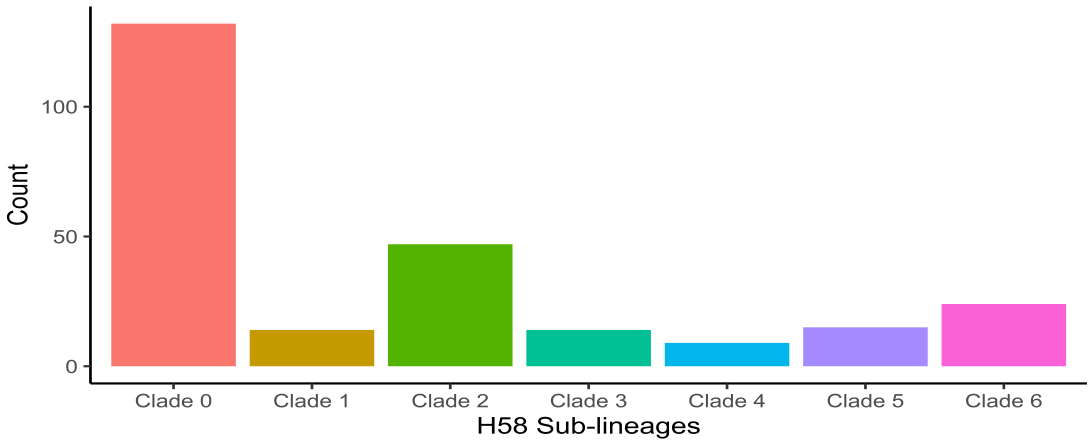


Figure 4.3: Barplot of H58 Genomic Sub-Lineages for Salmonella typhi

Figure 4.3 shows the frequency distribution of the typhoid fever cases by sub-lineage. The figure shows that clade 0 was the most common genomic sub-lineage representing 52% of the reported cases followed by clade 2 at 18%. The other sub-lineages (clade 1, clade 3, clade 4, clade 5 and clade 6) had few cases, as such, they were grouped for modelling purposes. In the proceeding sections, the grouped sub-lineages are being called 'grouped clades'.

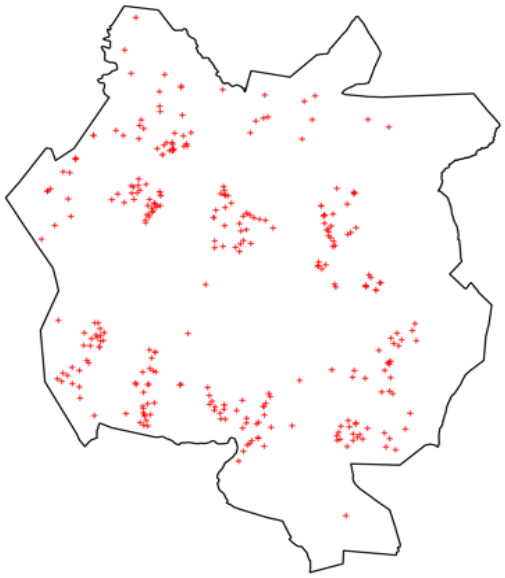


Figure 4.4: Spatial distribution of typhoid fever in Blantyre city

Figure 4.4 shows the spatial distribution of typhoid fever cases which were registered between March 2015 and December 2016 in Blantyre city. The figure shows there was heterogeneous distribution of the cases with several clusters. For example, it appears that there may be a cluster of cases in Ndirande, Bangwe, Nkolokoti-Kachere, Nancholi, Chilomoni, Chigumula, Mbayani-Chemusa and Chirimba. The figure also shows that most of the cases appear to occur in peri-urban areas.

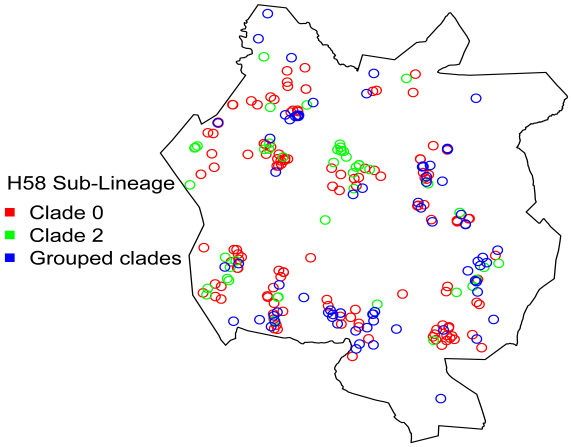


Figure 4.5: Spatial distribution of typhoid fever in Blantyre by sub-lineage

Out of the 549 typhoid cases in the final merged dataset, 255 cases had genomic data. Figure 4.5 shows the spatial distribution of the 255 typhoid cases by genomic sub-lineage. Out of the 7 sub-lineages, 5 of them caused relatively few typhoid cases during the study period. These were grouped into one sub-lineage (grouped clades) during the modelling process.

4.2 Model Diagnostics

This section will discuss some of the techniques which were used to assess the validity of the fitted models for inference. The *lgcp* package fits Bayesian spatial and spatio-temporal models. As such, the diagnostics presented in this section focus on assessing the MCMC iterations. All statistical analyses were conducted using R statistical software, version 4.0.3. The models were fitted using the *lgcp* R package. These diagnostic techniques include log-target, trace plots and auto-correlation plots. Metropolis-adjusted Langevin algorithm (MALA) length for all four spatio-temporal models was 1 million iterations. Burn-in was 100,000 iterations. Every 90th sample was retained. For the multi-type spatial LGCP model, however, to achieve the desired convergence and minimise autocorrelation, the MALA length was 40 million iterations with 500,000 burn-in iterations and 18,000 as the thinning parameter.

4.2.1 Log-target

Log of the target posterior likelihood is used to check whether the Markov chain of the model being fitted is mixing well. The log target also checks convergence of the Markov chain to a posterior mode. The technique assesses the plot of $log\{\pi(\beta, \eta, Y \mid X)\} + c$ up to an additive constant c.

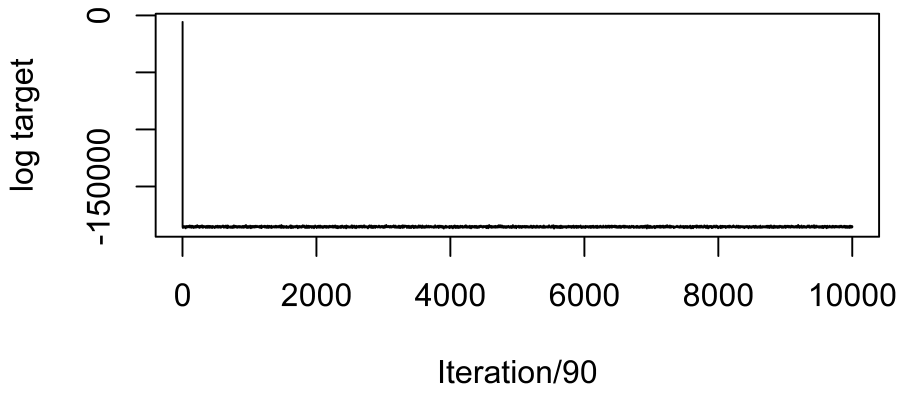


Figure 4.6: Plot of the log posterior over the duration of the MCMC run and burn-in for the spatio-temporal model for all cases

Figure 4.6 shows that the log-target of the Markov chain started with values near 0 but quickly converged around $-200,000$. The convergence of the log of the target posterior likelihood means that the Markov chain is mixing well and the parameters of the fitted model can be used for inference. Figure 4.7 also shows that the log-target of the Markov chain of the multi-type spatial model converged at around -35000. The other three spatio-temporal models have similar convergence of their log-targets. For more details, see Appendix 1.

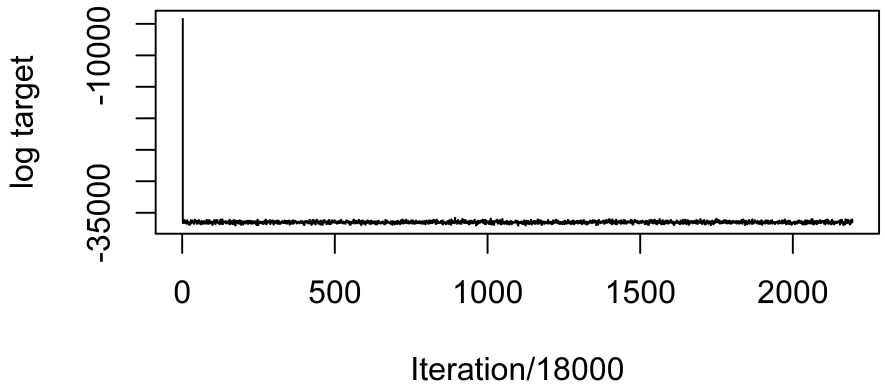


Figure 4.7: Plot of the log posterior over the duration of the MCMC run and burn-in the multi-type spatial model

4.2.2 Trace Plots

Another intuitive and easily implemented diagnostic tool is a trace plot which plots the parameter value of the model at time t against the iteration number. If the MCMC has converged to a stationary distribution, the trace plot will fluctuate randomly around the mode of the distribution. The trace plot looks like a hairy caterpillar when the posterior has converged to the stationary distribution. On the other hand, non-convergence can take many shapes. The two most common would be: 1. Clear trend in parameter values over iteration number. 2. Switching back and forth between different parameter sets. Figure 4.8 and Figure 4.9 show no obvious signs of non-convergence and are consistent with an MCMC process that converged successfully. The trace plots of the other three spatio-temporal models also show no obvious signs of non-convergence. See Appendix 2 for detail.

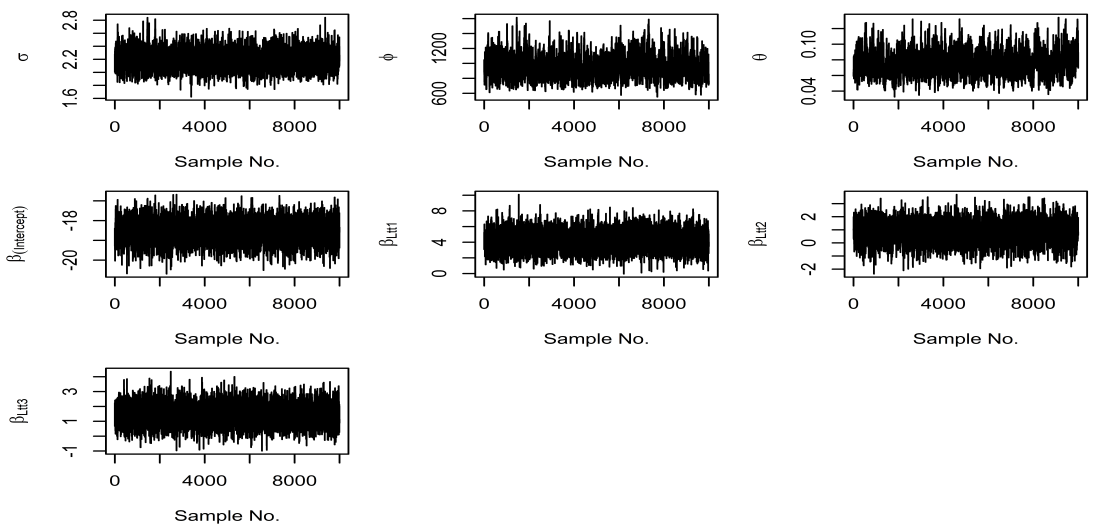


Figure 4.8: Traceplots of the model parameters of the spatio-temporal model with all cases

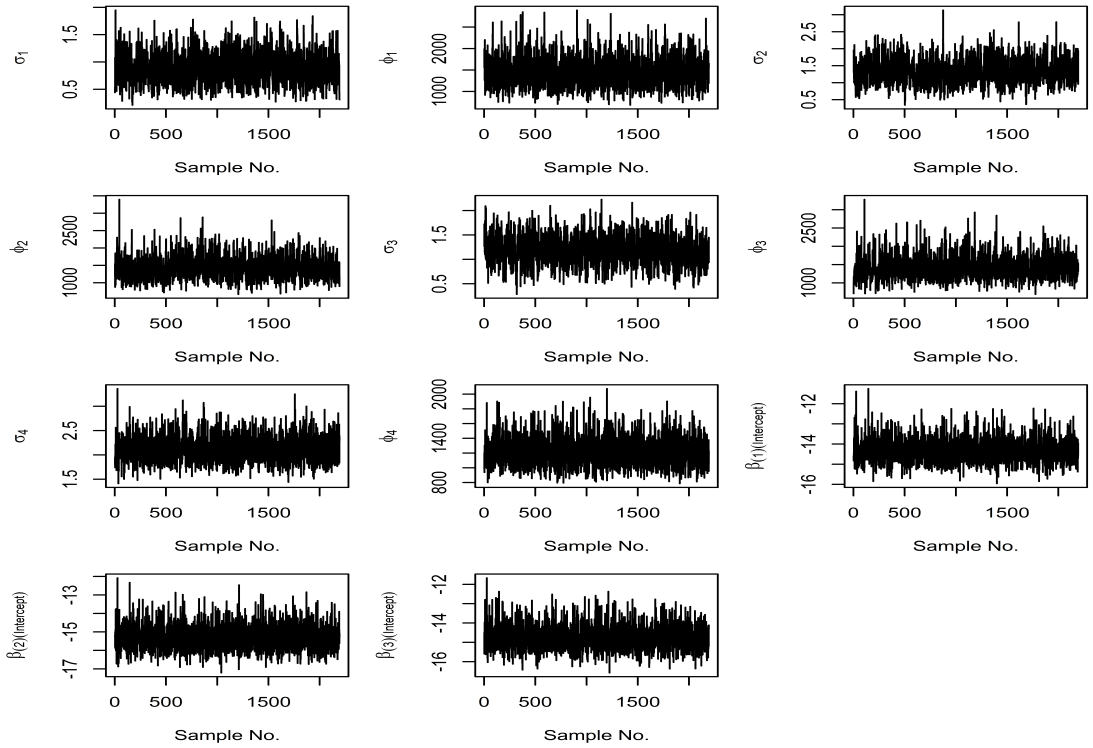


Figure 4.9: Traceplots of the model parameters of the multi-type spatial model

4.2.3 Autocorrelation in the latent Gaussian field

Spatial autocorrelation is the association of a variable with itself through space. When the values in adjacent spatial region vary together in opposite directions, negative autocorrelation occurs. Whereas when similar values occur near one another, positive autocorrelation occurs.

Figures in Appendix 3 show autocorrelation of MCMC iterations for each cell on the computational grid. Initially, there are positive autocorrelation at cellwise lag of 1 but little to no autocorrelation at cellwise lag of 15 for both spatio-temporal model with all cases and multi-type spatial model. This shows that there is very little autocorrelation in the sampled values of the retained latent field. Computational grid cell width is 350 metres. The other three spatio-temporal models also produced similar autocorrelation results. See Appendix 3 for details. Computational grid cell width is 350 metres.

4.2.4 Autocorrelation of parameters from the point process

Figure 4.10 shows that some parameters, ϕ and θ in the spatio-temporal model exhibit substantial autocorrelation, so the MALA algorithm will be slow to explore the entire posterior distribution. This can be resolved by increasing the posterior sample of the model by running the MALA algorithm longer. Since the autocorrelation is not severe at lag 40 of the thinned iterations of the MCMC for θ and the rest of the parameters show no autocorrelation, there is no need to run the MCMC longer. The other spatio-temporal models also show similar results. See Appendix 4 for details. Figure 4.11, on the other hand, shows little autocorrelation for the

multi-type spatial model at different lags in the thinned samples.

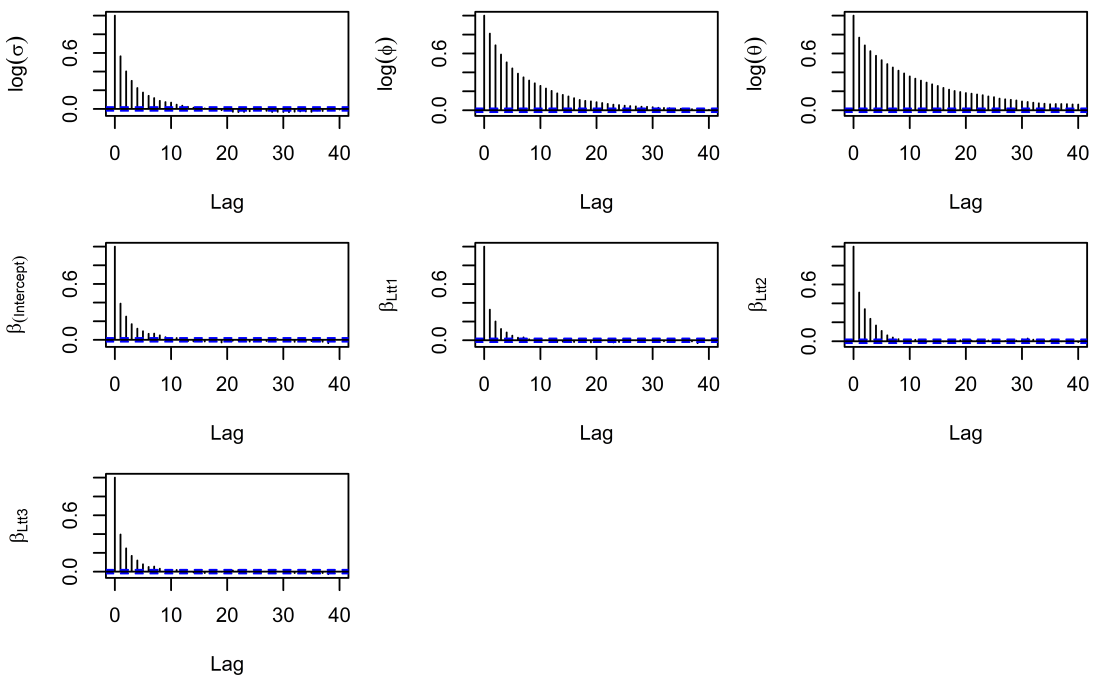


Figure 4.10: Autocorrelation plots of the parameters of the Gaussian latent field from the spatio-temporal model with all cases

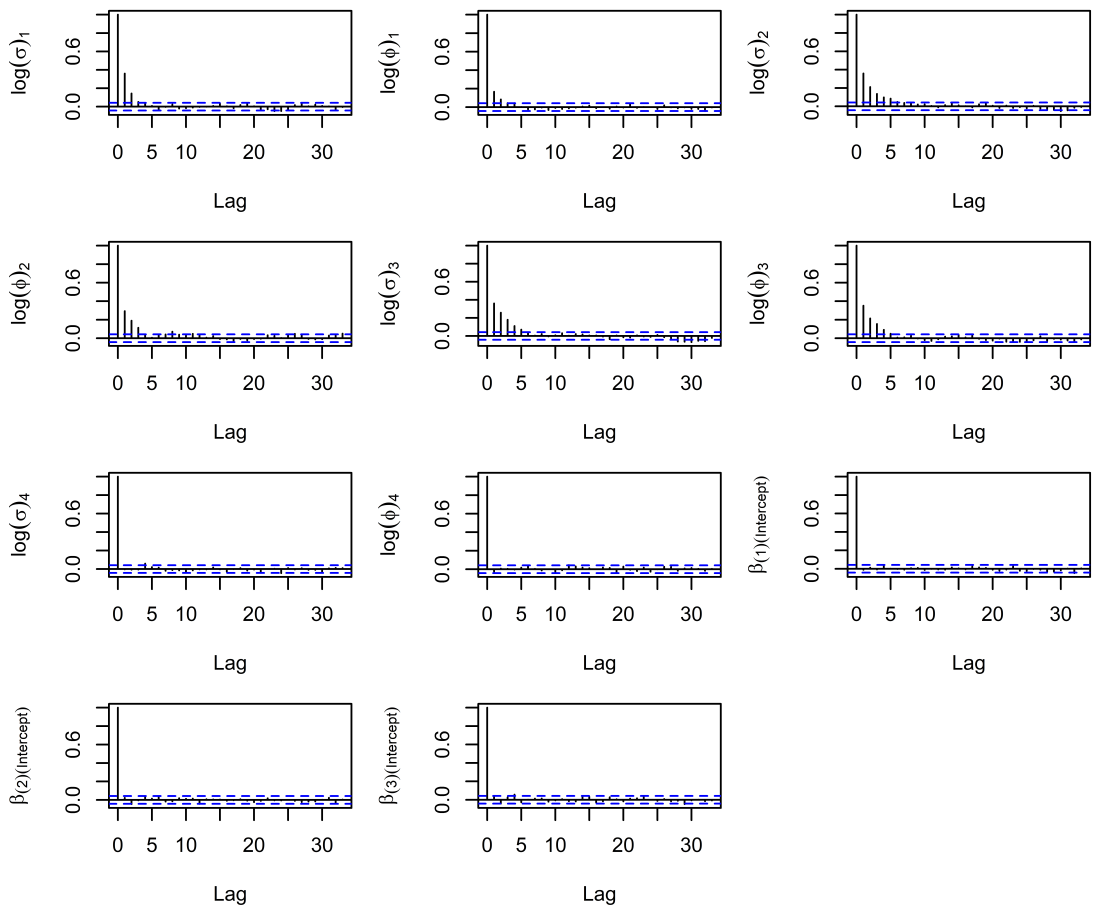


Figure 4.11: Autocorrelation plots of the parameters of the Gaussian latent field for multi-type spatial model

4.3 Log-Gaussian Cox Process Results

This section will discuss the results of the four spatio-temporal models and the multi-type spatial model which have been fitted using the LGCP framework. All models used a computational grid of cells 350 x 350 metres in dimension.

4.3.1 Spatio-temporal model with all cases

This subsection discusses the results of the spatio-temporal model fitted using all typhoid fever cases without focusing on their genomic lineage. In this paper, the

covariance function of the Gaussian process used an exponential model. Table 4.2 summarises the parameters of the latent field of the spatio-temporal LGCP model with all typhoid cases. The standard deviation parameter σ had median 2.185 (95% CrI 1.93 to 2.497); the spatial correlation parameter ϕ had median 940.1 metres (95% CrI 709 to 1275); and the temporal correlation parameter θ had median 0.075 months (95% CrI 0.050 to 0.107). This means that spatial dependence had a median of about 940 metres and a temporal dependence had a median of about 2 days. The other parameters, $t^{'}, t^{''}$ and $t^{'''}$ are for the cubic B-spline which was included to assess the temporal distribution of the model and has not been interpreted to have a covariate effects. The prior and posterior plot in Appendix 8 shows that σ has relatively a wider departure from the prior compared with the parameter ϕ. Therefore statistical inference must be done cautiously.

Parameter	Median	Lower 95% CrI	Upper 95% CrI
σ	2.185	1.93	2.497
ϕ	940.1	709	1275
θ	7.46×10^{-2}	4.952×10^{-2}	0.1072
$exp(\beta_{Intercept})$	8.912×10^{-9}	3.038×10^{-9}	2.462×10^{-8}
$exp(\beta_{t'})$	59.81	7.31	649.9
$exp(\beta_{t''})$	2.146	0.4726	9.768
$exp(\beta_{t'''})$	4.144	1.16	15.98

Table 4.2: Parameter estimates for the LGCP model with all cases

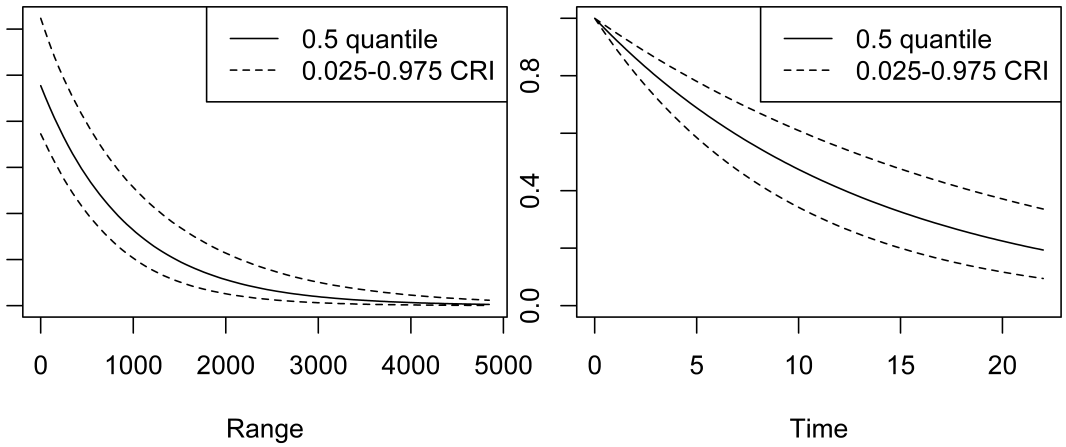


Figure 4.12: Plots of the posterior spatial covariance (Left) and temporal correlation (Right) for the Gaussian process of the spatio-temporal model with all cases

To assess the posterior dependence between cells on the computational grid, exponential model for the posterior covariance function of the Gaussian process was used. Figure 4.12 shows the shape of the posterior covariance function for the spatio-temporal model with all cases. The figure shows that the posterior dependence between cells is over a small range in both time and space. Figure 4.13, the inhomogeneous K function plot, shows that the points formed clusters at about 2.5km radius for all correction estimates (isotropic correction estimates, translation correction estimates, modified border correction estimates and border corrected estimates).

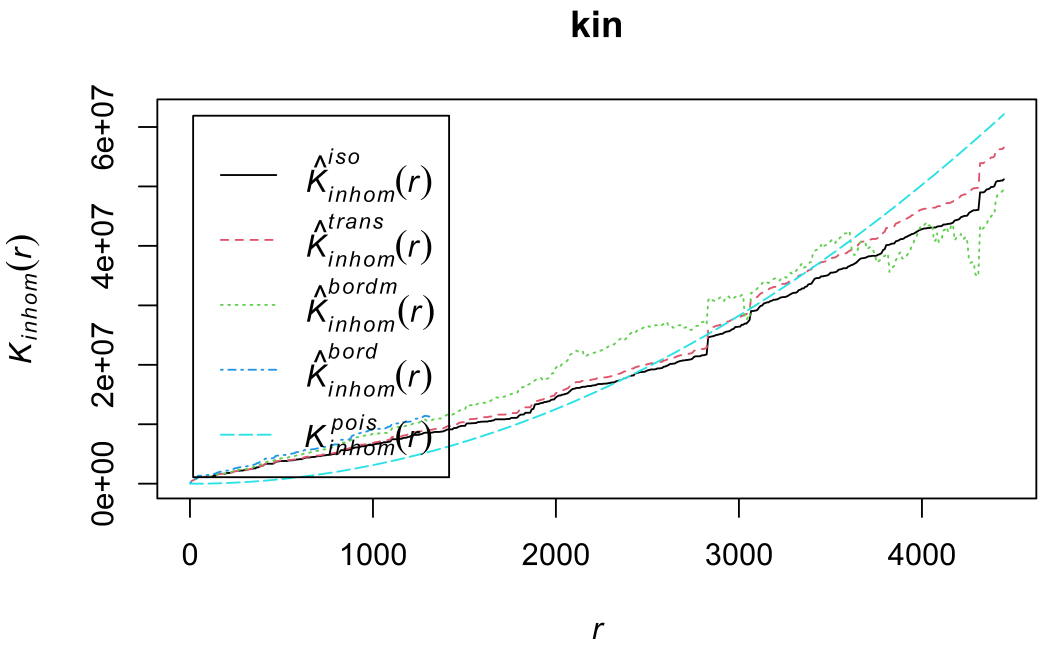


Figure 4.13: Inhomogeneous K Function for all typhoid cases

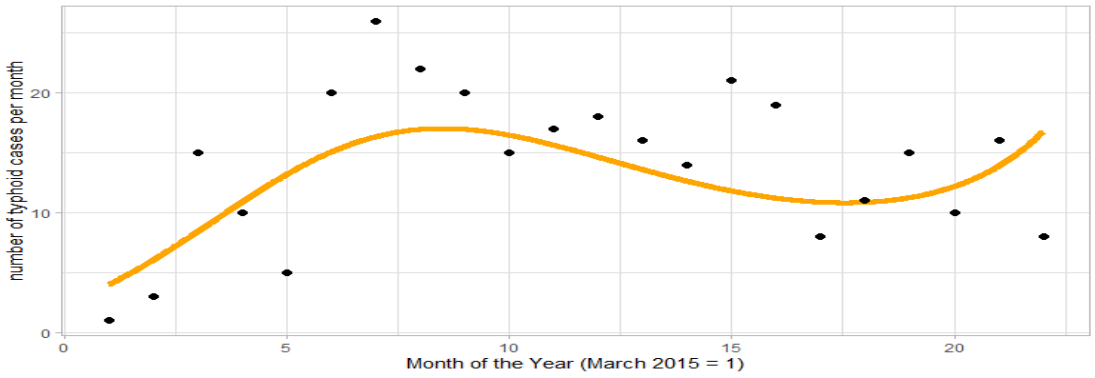


Figure 4.14: Temporal distribution of typhoid fever outbreak for all typhoid cases

To ascertain the temporal distribution of the typhoid fever outbreak, a cubic B-spline was used. The orange graph line of Figure 4.14 describes the trend of the typhoid fever between March 2015 to December 2016 that was fitted using B-splines. The figure shows that the typhoid fever outbreak was at the peak around the months of October and November 2015. Then the cases started decreasing

steadily up to July 2016 when it started increasing steadily again.

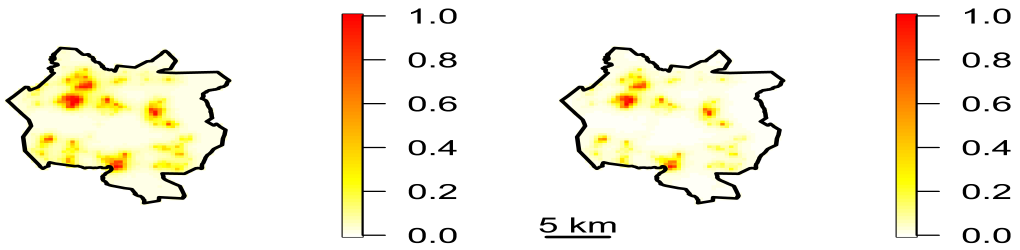


Figure 4.15: Exceedance plot of posterior probability that the incidence rates exceed 2 (left) and 4 (right) for all typhoid cases

Figure 4.15 shows location of high posterior probability that incidence rate for all typhoid fever cases exceeds 2 and 4. The figure shows that Chilobwe-Misesa, Mbayani-Chemusa, Chirimba, Nkolokoti-Kachere, Nancholi-Manase, Ndirande and Bangwe-Namiyango were the hotspots of typhoid fever during the outbreak. These areas had a high posterior probability to have a 4 times higher incidence rate of typhoid than the rest of Blantyre city.

4.3.2 Spatio-temporal model for clade 0 sub-lineage

This sub-section discusses the findings of spatio-temporal model for clade 0 sub-lineage of H58 lineage of *S. typhi*.

Parameter	Median	Lower 95% CrI	Upper 95% CrI
σ	2.257	1.893	2.683
ϕ	873.6	615	1316
θ	0.1068	6.417×10^{-2}	0.1747
$exp(\beta_{Intercept})$	1.28×10^{-8}	4.223×10^{-9}	3.663×10^{-8}
$exp(\beta_{t'})$	3.634	0.3014	52.46
$exp(\beta_{t''})$	2.912	0.4509	20.95
$exp(\beta_{t'''})$	1.443	0.3212	6.102

Table 4.3: Parameter estimates for the LGCP model for clade 0 cases

Table 4.3 summarises the estimated parameters of the Gaussian process whose realisation produced typhoid cases which were caused by clade 0. The standard deviation parameter σ had median 2.257 (95% CrI 1.893 to 2.683); the spatial correlation parameter ϕ had median 873.6 metres (95% CrI 615 to 1316); and the temporal correlation parameter θ had median 0.107 months (95% CrI 0.064 to 0.175). This means that clade 0 cases had spatial dependence had a median of about 874 metres and a temporal dependence had a median of about 3 days.

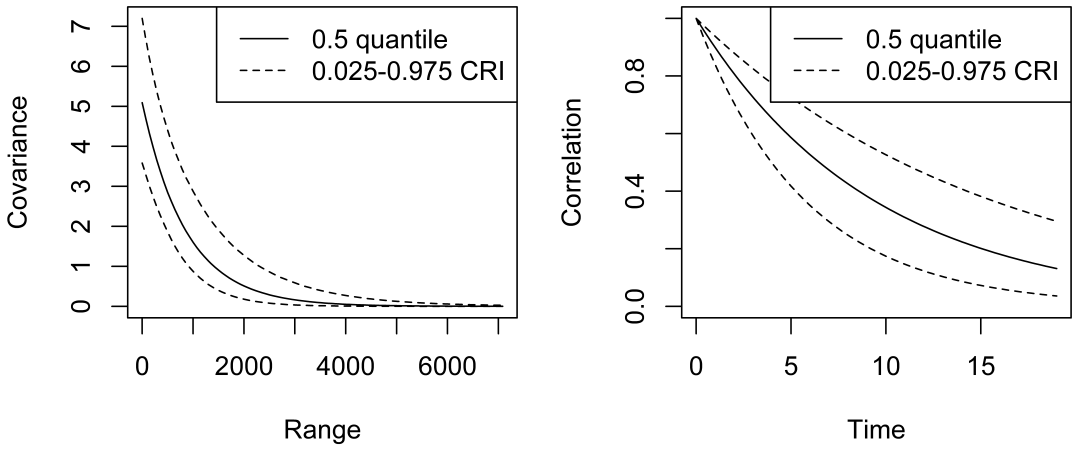


Figure 4.16: Plots of the posterior spatial covariance (Left) and temporal correlation (Right) for the Gaussian process of the spatio-temporal model for clade 0 cases

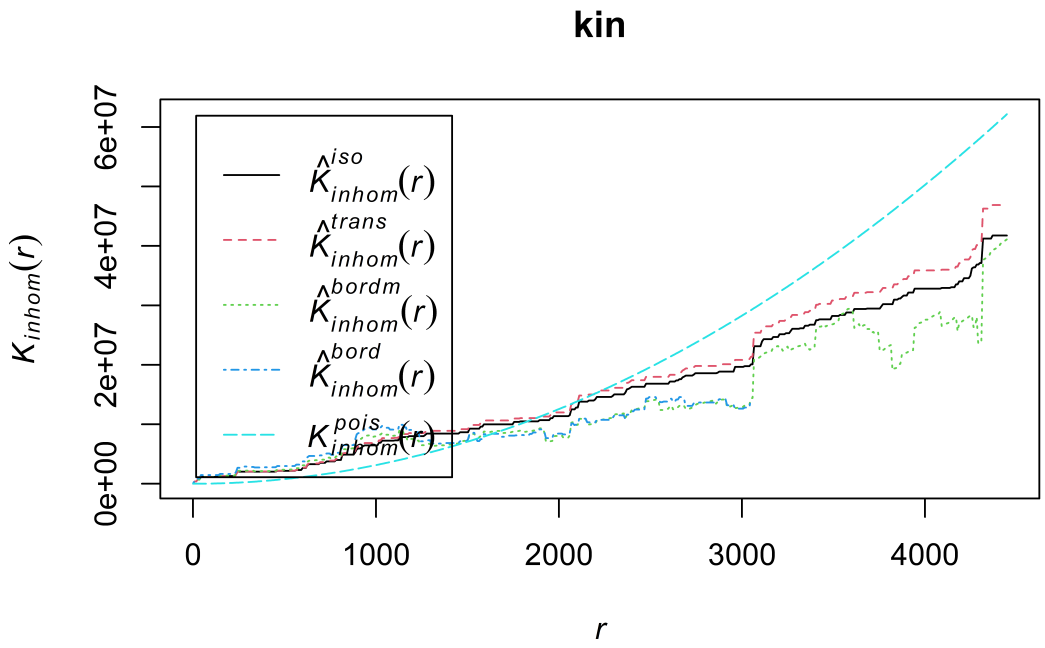


Figure 4.17: Inhomogeneous K Function for clade 0 cases

Figure 4.16 shows the posterior covariance function for the Gaussian process of spatio-temporal model for clade 0 cases. The figure confirms the results above as it shows that the posterior dependence between cells is over a small range in both time and space. Figure 4.17, the inhomogeneous K function plot also, shows that the points formed clusters at about 1.5km radius for all correction estimates (isotropic correction estimates, translation correction estimates, modified border correction estimates and border corrected estimates).

The orange graph line of Figure 4.18, fitted using the coefficients of t', t'' and t''', the B-spline basis functions, shows that cases of typhoid fever from clade 0 sub-lineage started increasing since its first registration in April 2015 and reached its first peak in October 2015. The cases started decreasing from December 2015 until

they started increasing again from May 2016 up to the end of the study. Clade 0 cases were registered between April 2015 to October 2016. The temporal trend of clade 0 is similar to that of all cases.

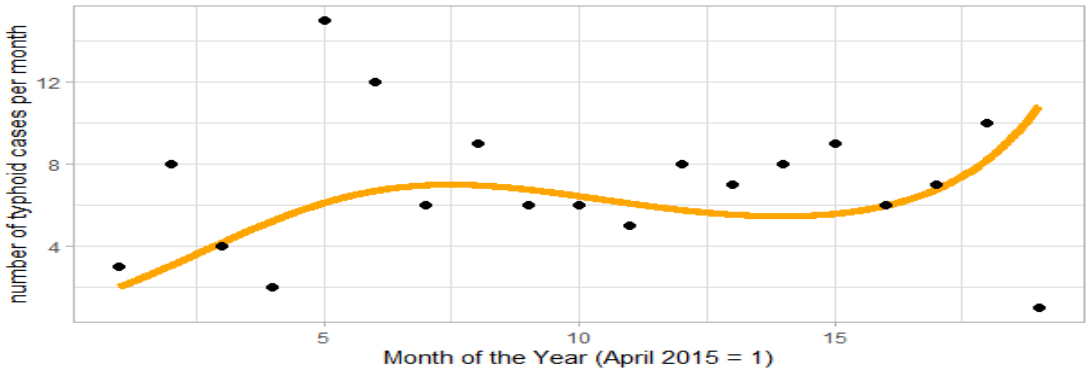


Figure 4.18: Temporal distribution of typhoid fever outbreak for clade 0 cases

Comparing the parameters of the two models, it is clear that the point estimates of σ for the two models were similar. For spatio-temporal model with all cases, σ was 2.185 (95% CrI 1.93 to 2.497) while for clade 0 was 2.257 (95% CrI 1.893 to 2.683). The point estimates of ϕ from the two models were different. For all cases, the median was 940.1 metres (95% CrI 709 to 1275) while for clade 0 cases the median was 873.6 metres (95% CrI 615 to 1316). The difference can also be explained by the inhomogeneous K function plots which show that typhoid fever for all cases and for clade 0 cases formed cluster at 2.5km radius and 1.5km radius respectively. This means that the cases for clade 0 cluster tightly than for all cases. This means that the typhoid fever cases caused by clade 0 had more local transmission of typhoid fever. The point estimates for θ, the parameter which scales the temporal dependence, were also different for the two models with all cases registering 0.075 months (95% CrI 0.050 to 0.107) while clade 0 registering 0.107 months (95% CrI

0.064 to 0.175). This means that all cases happened at shorter intervals of 2 days compared to 3 days for clade 0 cases. The wider credible intervals for all point estimates for clade 0 give evidence of uncertainty in the estimates. This is because the number of typhoid fever cases caused by clade 0 were only 43% of all typhoid fever cases.

Figure 4.19 shows location of high posterior probability that incidence rates exceeds 2 and 4 for all typhoid fever cases caused by clade 0 sub-lineage of the H58 lineage of *S. typhi*. The figure shows that Mbayani-Chemusa, Chirimba, Nkolokoti-Kachere, Nancholi-Manase and Ndirande had 4 times higher incidence rate of typhoid fever outbreak caused by clade 0 sub-lineage of H58 lineage of *S. typhi* than other locations in the city. These locations are five of the seven locations with 4 times higher incidence rates of typhoid fever for all cases. The spatial distribution for clade 0 is similar to the spatial distribution for all cases of typhoid fever. One of the reasons could be because clade 0 sub-lineage of H58 haplotype of *S. typhi* was the main cause of typhoid fever cases in the recent typhoid fever outbreak. It caused 52% of all the registered typhoid cases with genomic data.

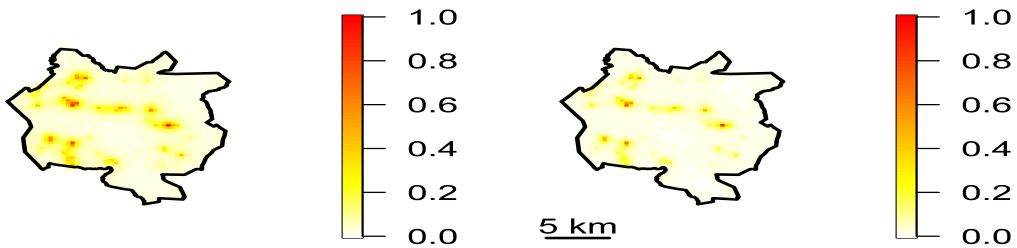


Figure 4.19: Exceedance Plot of posterior probability that the incidence rate exceeds 2 (left) and 4 (right) for clade 0 cases

4.3.3 Spatio-temporal model for clade 2 sub-lineage

This sub-section discusses the findings of spatio-temporal model for clade 2 sub-lineage of H58 lineage of *S. typhi*.

Parameter	Median	Lower 95% CrI	Upper 95% CrI
σ	2.483	1.884	3.243
ϕ	807	501.7	1360
θ	0.3261	0.1707	0.6065
$exp(beta_{Intercept})$	3.961×10^{-8}	1.074×10^{-10}	1.295×10^{-5}
$exp(\beta_{t'})$	0.2223	1.283×10^{-6}	44367
$exp(\beta_{t''})$	0.3273	4.556×10^{-3}	24.61
$exp(\beta_{t'''})$	0.2341	3.564×10^{-4}	159.8

Table 4.4: Parameter estimates for the LGCP model for clade 2 cases

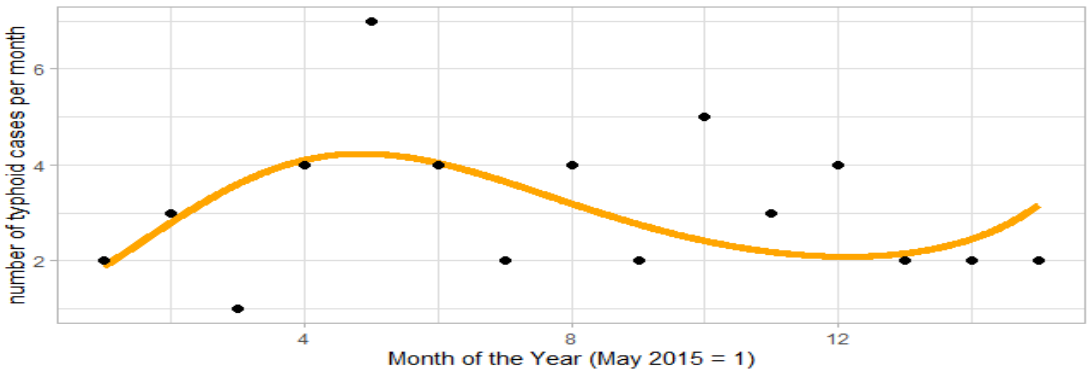


Figure 4.20: Temporal distribution of typhoid fever outbreak for clade 2 cases

Table 4.4 is the summary of the parameters of the latent field of the spatio-temporal LGCP model with clade 2 typhoid cases. The parameter σ had median 2.483 (95% CrI 1.884 to 3.243); the parameter ϕ had median 807 metres (95% CrI 501.7 to 1360); and the parameter θ had median 0.326 months (95% CrI 0.171 to 0.607). This means that the clade 2 cases had spatial dependence of about 807 metres and a temporal dependence of about 10 days. The orange graph line in Figure 4.20 shows that cases of typhoid fever caused by clade 2 sub-lineage

was highest in September 2015 and has been decreasing steadily during the study period. The temporal trend of clade 2 is similar to that of clade 0 and all cases discussed in the previous sections.

The posterior covariance function plot for the Gaussian process of spatio-temporal model for clade 2 cases in Appendix 5 shows that the posterior dependence between cells is over a small range in both time and space. The inhomogeneous K function plot in Appendix 9 shows that the points formed clusters at about 2.5km radius for all correction estimates (isotropic correction estimates, translation correction estimates, modified border correction estimates and border corrected estimates).

The comparison of parameters in Table 4.4 with those in Table 4.3 and Table 4.2 show that σ values from all the three models are similar. Their medians range from 2.185 to 2.483. The point estimate of ϕ for the model with clade 2 cases and all cases were different. Clade 2 has ϕ of median of 807 metres (95% CrI 501.7 to 1360) while all cases has a median of 940.1 metres (95% CrI 709 to 1275). ϕ for clade 2 is similar to the model with clade 0 cases with a 873.6 metres (95% CrI 615 to 1316). All models have spatial dependence of less that 1km. The point estimate for θ from clade 0 and clade 2 were different. Clade 0 has a temporal dependence of 3 days while clade 2 has a temporal dependence of 10 days. The wider credible intervals for all point estimates for the model with clade 2 cases is due to the relatively fewer number of typhoid fever cases cause by clade 2 sub-lineage.

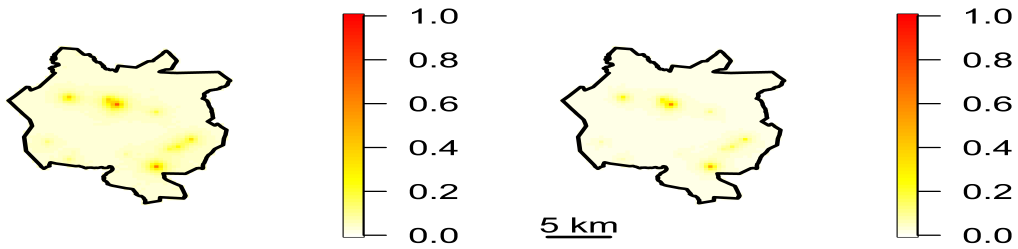


Figure 4.21: Exceedance Plot of posterior probability that the incidence rate exceeds 2 (left) and 4 (right) for clade 2 cases

Figure 4.21 shows location of high posterior probability that incidence rate of typhoid fever cases caused by clade 2 sub-lineage of the H58 lineage of *S. typhi* exceeds 2 and 4. The figure shows that Ndirande and Chigumula had 4 times higher incidence rate of the typhoid fever outbreak than other locations in the city.

4.3.4 Spatio-temporal model for grouped sub-lineages

This sub-section discusses the findings of spatio-temporal model for the genomic clades which had fewer typhoid cases (clade 1, clade 3, clade 4, clade 5 and clade 6). These were grouped into a single clade for model fitting.

Parameter	Median	Lower 95% CrI	Upper 95% CrI
σ	2.31	1.858	2.848
ϕ	791.7	495.7	1264
θ	0.1785	0.1001	0.3175
$exp(beta_{Intercept})$	3.212×10^{-9}	5.664×10^{-10}	1.323×10^{-8}
$exp(\beta_{t'})$	27.04	0.7734	1569
$exp(\beta_{t''})$	7.471	0.7305	86.59
$exp(\beta_{t'''})$	3.405	0.444	31.2

Table 4.5: Parameter estimates for the LGCP model for grouped clades

Table 4.5 is the summary of the parameters of the latent field of the spatio-

temporal LGCP model with grouped clades. The parameter σ had median 2.31 (95% CrI 1.858 to 2.848); the parameter ϕ had median 791.7 metres (95% CrI 495.7 to 1264); and the parameter θ had median 0.179 months (95% CrI 0.100 to 0.318). The point estimate for the standard deviation parameter σ for the grouped clades is similar to that of the other models. The spatial correlation parameter estimate ϕ for the grouped clades is also comparable with the other three models but the temporal correlation estimate θ is different from the rest of the models. But the credible intervals of the point estimates of the model with grouped cases is narrower than that of the model with clade 2 cases. This is because of the number of cases involved during modelling. Clade 2 has the lowest number of cases compared to the other models. See Table 4.2, Table 4.3, Table 4.4 and Table 4.5 for details.

The orange graph line in Figure 4.22 shows that cases of typhoid fever caused by the grouped clades had a steady increase after first registration in March 2015. The typhoid cases for the grouped sub-lineages were highest between October and November 2015. The cases started decreasing steadily from December 2015 until the end of the study. The temporal trend is relatively different from the temporal trend for all cases, clade 0 and clade 2. The temporal trend for models for all cases, clade 0 and clade 2 started increasing then decreasing and increasing again. The temporal trend for the grouped clades just increased and decreased steadily and flattened at the end without increasing again.

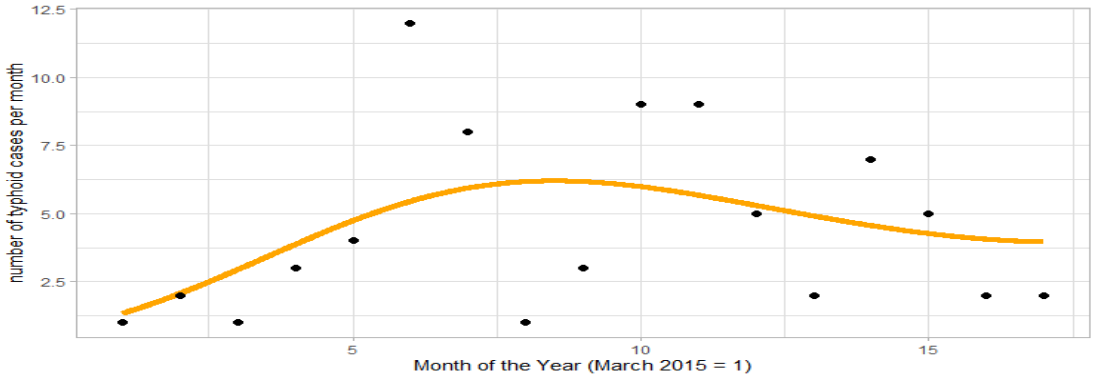


Figure 4.22: Temporal distribution of typhoid fever outbreak for the grouped clades

Figure 4.23 is showing the location of high posterior probability that incidence rate of typhoid fever cases caused by the grouped clades of the H58 lineage of *S. typhi* exceeds 2 and 4. The figure shows that Mbayani-Chemusa, Nkolokoti-Kachere, Bangwe-Namiyango, Nancholi-Manase and Chilobwe-Misesa had 4 times higher incidence rate of the typhoid fever outbreak than other locations in the city. The inhomogeneous K function plot in Appendix 9 shows evidence of clustering at the radius of about 2km for all correction estimates (isotropic correction estimates, translation correction estimates, modified border correction estimates and border corrected estimates).

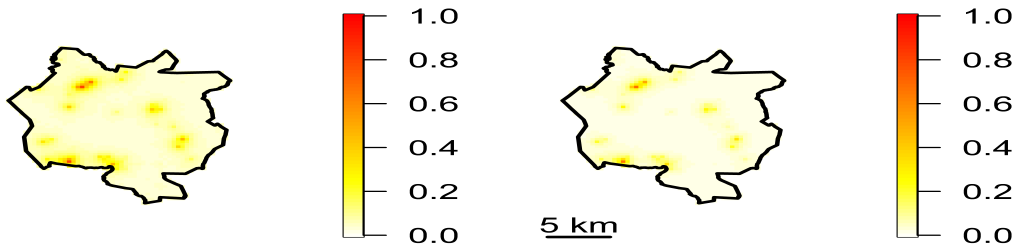


Figure 4.23: Exceedance Plot of posterior probability that the incidence rate exceeds 2 (left) and 4 (right) for the grouped clades

Mbayani-Chemusa is the only location in Blantyre which had high incidence rate of typhoid fever caused by all the three sub-lineages (clade 0, clade 2 and clade13456). This means that the spatial interaction of the sub-lineages were complementary. Ndirande had high incidence rate of typhoid fever caused by both clade 0 and clade 2 sub-lineages. The two sub-lineages were also complementary in space in this location. Bangwe-Namiyango had high incidence rate of typhoid fever caused by both clade 2 and the grouped clades while Nancholi-Manase and Nkolokoti-Kachere had high incidence rate of typhoid fever caused by both clade 0 and the grouped clades.

4.3.5 Multi-type spatial model

This sub-section discusses the findings of the multi-type spatial model for the genomic clade 0, 2 and the other clades which had fewer cases and were combined into one clade.

Table 4.6 is the summary of the estimated parameters of the Gaussian latent field of the multi-type spatial LGCP model for clade 0, clade 2 and the grouped clades.

Parameter	Median	Lower 95% CrI	Upper 95% CrI
σ_1	0.8735	0.4004	1.469
ϕ_1	1408	897.7	2190
σ_2	1.293	0.6932	2.097
ϕ_2	1381	902.1	2142
σ_3	1.158	0.6036	1.774
ϕ_3	1390	920.4	2153
σ_4	2.091	1.674	2.678
ϕ_4	1192	895.6	1645
$exp[\beta_1(Intercept)]$	5.572×10^{-7}	2.266×10^{-7}	2.097×10^{-6}
$exp[\beta_2(Intercept)]$	2.118×10^{-7}	7.266×10^{-8}	9.731×10^{-7}
$exp[\beta_3(Intercept)]$	3.667×10^{-7}	1.337×10^{-7}	1.632×10^{-6}

Table 4.6: Table of the parameter estimates from the multivariate spatial model

The parameter σ_1 had median 0.8735 (95% CrI 0.4004 to 1.469); the parameter ϕ_1 had median 1408 metres (95% CrI 897.7 to 2190); the parameter σ_2 had median 1.293 (95% CrI 0.6932 to 2.097); the parameter ϕ_2 had median 1381 metres (95% CrI 902.1 to 2142); the parameter σ_3 had median 1.158 (95% CrI 0.6036 to 1.774); the parameter ϕ_3 had median 1390 metres (95% CrI 920.4 to 2153); the parameter σ_4 had median 2.091 (95% CrI 1.674 to 2.678); the parameter ϕ_4 had median 1192 metres (95% CrI 895.6 to 1645). Direct comparison between the same parameters for different clades show that the standard deviation parameter for clade 0 σ_1 = 0.8735 (95% CrI 0.4004 to 1.469) is lower than of clade 2 σ_2 = 1.293 (95% CrI 0.6932 to 2.097) and the grouped clades σ_3 = 1.158 (95% CrI 0.6036 to 1.774). On the other hand, the spatial correlation parameters ϕs are similar for all the three models with median ranging from 1381 to 1408 metres. This implies that each clade formed clusters at a median radius of 1.5km. This is consistent with the inhomogeneous K functions plots.

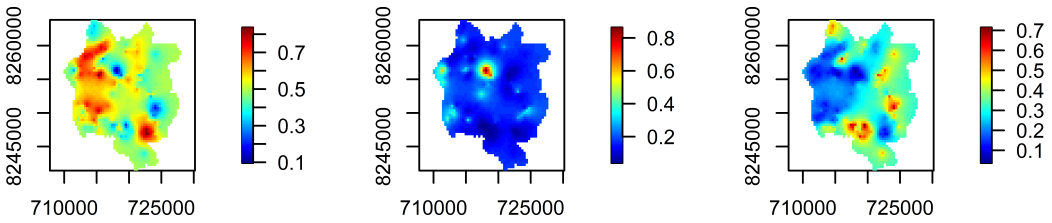


Figure 4.24: Conditional probability that a point at each location is of a particular type: clade 0 (Left Panel), clade 2 (Middle Panel), grouped clades (Right Panel)

The left part of Figure 4.24 shows that typhoid fever cases caused by clade 0 sub-lineage of the multi-drug resistant H58 lineage of *S. typhi* were dominant in the in the western side and south eastern side of Blantyre city. Specifically, clade 0 cases were dominant in the following areas: Chirimba, Kameza, Machinjiri, Nkolokoti-Kachere, Likhubula, Mbayani-Chemusa, Nancholi-Manase and Bangwe-Namiyango. The middle part shows that clade 2 was more dominant in Ndirande and Chilomoni. The grouped clades, that is clade 1, clade 3, clade 4, clade 5 and clade 6 were more dominant in the eastern part of Blantyre city. Specifically, the grouped clades were dominant in the following areas: Nancholi, Zingwangwa, Cholobwe-Misesa, Chigumula, Zingwangwa, Kachere, Mapanga, Machinjiri and Kameza.

Although Mbayani-Chemusa, Ndirande, Bangwe-Namiyango, Nancholi-Manase and Nkolokoti-Kachere registered typhoid fever cases caused by multiple sub-lineages within the same location, the multi-type spatial analysis has shown that the sub-lineages were still competing and one of them was dominant than the others. For example, although Mbayani-Chemusa had high incidence rate of typhoid fever cases caused by all the three sub-lineages (clade 0, clade 2 and the

grouped clades), clade 0 sub-lineage was more dominant than clade 2 and the grouped clades. Likewise, although Ndirande registered high incidence rate of typhoid fever cases caused by clade 0 and clade 2, conditional probability has shown that clade 2 was more dominant than clade 0. Similarly, Nancholi-Manase and Nkolokoti-Kachere registered typhoid fever cases caused by both clade 0 and the grouped clades. But conditional probability has shown that the grouped clades was more dominant than clade 0.

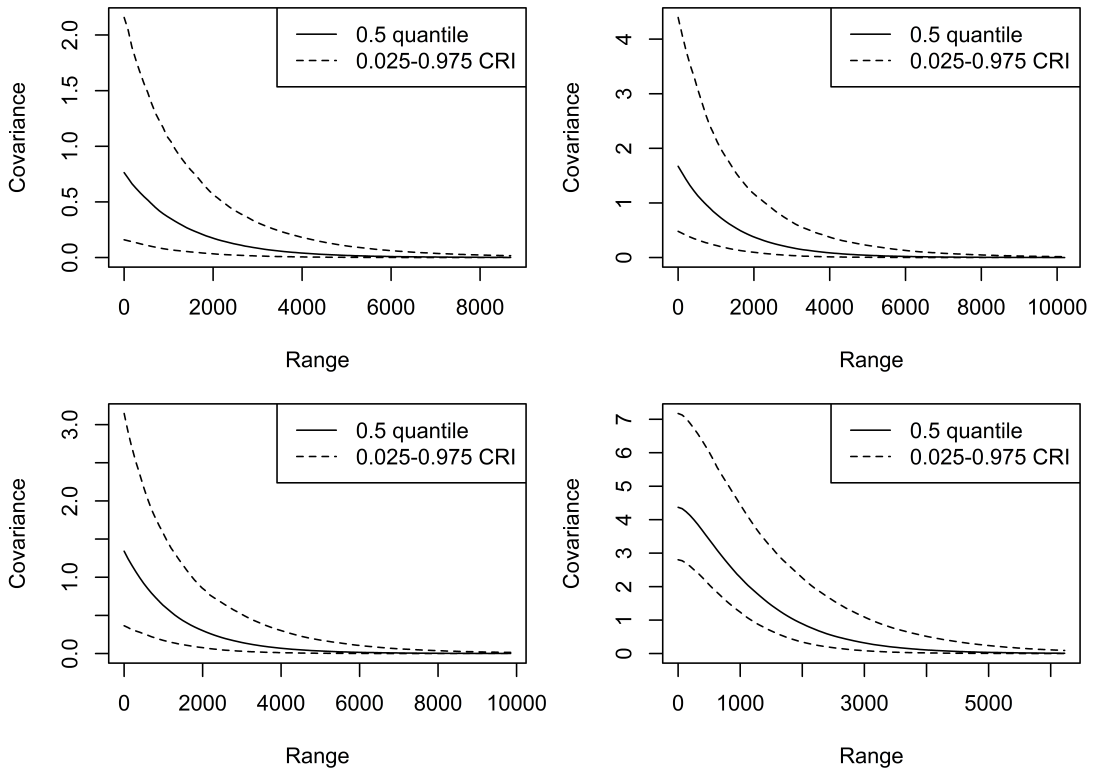


Figure 4.25: Posterior covariance function of the multi-type spatial model

Figure 4.25 shows the posterior covariance function for the Gaussian process of the multi-type spatial LGCP model. The figure shows that the posterior dependence between cells for all the sub-lineages is relatively bigger than the range for the spatio-temporal models in both time and space. The change can be attributed to the interactions between the sub-lineages.

Chapter 5

CONCLUSION, RECOMMENDATIONS, LIMITATIONS AND AREA FOR FURTHER RESEARCH

5.1 Conclusion

The typhoid fever cases which were recorded at QECH between March 2015 and December 2016 were all caused by seven different sub-lineages of the H58 lineage of *S. typhi*. The long term distribution of the typhoid cases shows that since the start of the study (March 2015), the outbreak was increasing steadily until around October and November 2015 when the outbreak was at its peak with a range of 15 to 25 cases per month. Then the cases started dropping until July 2016 when the cases started increasing again.

The spatio-temporal models have also shown that cases of clade 0 and clade 2 sub-lineages had similar temporal trend to the long term trend of all the typhoid cases combined. However, the temporal distributions cannot be used to assess seasonality because of limited time points. The maximum time points used in the study was 22 points i.e. 22 months. The minimum being 15 points.

The multi-type spatial model has shown that clade 0, clade 2 and the grouped clades all had their own high-transmission locations distinct from each other with minor overlaps. This proves that the clades were competing against each other and local transmission were happening around existing cases or potentially specific water sources becoming contaminated with specific H58 *S. typhi* sub-lineage. Typhoid fever cases caused by clade 0 sub-lineage were dominant in the western and south eastern side of Blantyre city. The areas include Chirimba, Kameza, Machinjiri, Nkolokoti-Kachere, Likhubula, Mbayani-Chemusa, Nancholi-Manase and Bangwe-Namiyango. clade 2 cases were dominant in Ndirande and Chilomoni. The grouped clades were dominant in the eastern side of the city. The areas include Nancholi, Zingwangwa, Cholobwe-Misesa, Chigumula, Zingwangwa, Kachere, Mapanga, Machinjiri and Kameza.

5.1.1 Recommendation

The researcher has made the following recommendations based on the spatial and spatio-temporal LGCP models implemented in this research:

- The analysis should be done using data which have more time points. This

will help to assess seasonality effects.

- Include environmental and economic factors like elevation, temperature and closeness to water sources in the model to also assess the effects of these factors and how they affect the spatial and temporal distribution of typhoid fever in Blantyre city.

- The Ministry of Health (MoH) and the Ministry of Water and Sanitation (MoWS) should work together to enhance water and sanitation services delivery in the areas with high incidence rates of typhoid fever.

5.1.2 Limitations

The study failed to assess seasonality as a temporal covariate. This is because the MCET dataset only had 22 time points. The study also failed to fit models for all the 7 sub-lineages because some sub-lineages had very few cases for proper model fitting. That is why other sub-lineages with few cases were grouped into a single sub-lineage.

The thesis also failed to fit a multi-type spatio-temporal LGCP model because of the current limitation of the *lgcp* R package. Further, the *lgcp* package only fits LGCP models which assumes a separable covariance function for the Gaussian process.

5.1.3 Areas for further research

For further studies, there is need to fit a multi-variate spatio-temporal LGCP model where spatial and temporal relationships and interactions among sub-lineages

can be investigated. This was currently not done because it was beyond the scope of this thesis.

Further studies should also incorporate environmental and economic factors apart from just assessing spatial and temporal factors. The effects of these environmental factors would help to shade more light why some areas had high incidence rate of typhoid fever than the other as found in this study. This was also not done because it was beyond the scope of this thesis.

REFERENCES

[1] Akullian, A., Ng'eno, E., Matheson, A. I., Cosmas, L., Macharia, D., Fields, B., ... & Montgomery, J. M. (2015). Environmental transmission of typhoid fever in an urban slum. PLoS neglected tropical diseases, 9(12), e0004212.

[2] Antillón, M., Warren, J. L., Crawford, F. W., Weinberger, D. M., Kürüm, E., Pak, G. D., ... & Pitzer, V. E. (2017). The burden of typhoid fever in low- and middle-income countries: a meta-regression approach. PLoS neglected tropical diseases, 11(2), e0005376.

[3] Martínez-Bello, D. A., López-Quílez, A., & Torres Prieto, A. (2018). Spatio-temporal modeling of Zika and dengue infections within Colombia. International Journal of Environmental Research and Public Health, 15(7), 1376.

[4] Cox, R., Su, T., Clough, H., Woodward, M. J., & Sherlock, C. (2012). Spatial and temporal patterns in antimicrobial resistance of Salmonella Typhimurium in cattle in England and Wales. Epidemiology & Infection, 140(11), 2062-2073.

[5] Cressie , N., & Wikle, C. K. (2011). Statistics for Spatio-Temporal Data (1st ed.). Wiley.

[6] Davies TM, Hazelton ML (2013). "Assessing Minimum Contrast Parameter Estimation for Spatial and Spatiotemporal Log-Gaussian Cox Processes." Statistica Neerlandica, 67(4),355389.

[7] Diggle, P. J. (2013). Statistical analysis of spatial and spatio-temporal point patterns. CRC press.

[8] Diggle, P. J., & Brix, A. (2001). Spatio-temporal prediction for log-Gaussian Cox processes. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(4), 823-841 https://doi.org/10.1111/1467-9868.00315

[9] Feasey, N. A., Gaskell, K., Wong, V., Msefula, C., Selemani, G., Kumwenda, S., ... & Heyderman, R. S. (2015). Rapid emergence of multidrug resistant, H58-lineage Salmonella typhi in Blantyre, Malawi. PLoS neglected tropical diseases, 9(4), e0003748.

[10] Fenton, S. E., Clough, H. E., Diggle, P. J., Evans, S. J., Davison, H. C., Vink, W. D., & French, N. P. (2009). Spatial and spatio-temporal analysis of Salmonella infection in dairy herds in England and Wales. Epidemiology & Infection, 137(6), 847-857.

[11] Gauld, J. S., Olgemoeller, F., Nkhata, R., Li, C., Chirambo, A., Morse, T., ... & Feasey, N. A. (2020). Domestic river water use and risk of typhoid fever: results from a case-control study in Blantyre, Malawi. Clinical Infectious Diseases, 70(7), 1278-1284.

[12] Gauld, J. S., Olgemoeller, F., Heinz, E., Nkhata, R., Bilima, S., Wailan, A. M., ... & Feasey, N. A. (2022). Spatial and genomic data to characterize endemic typhoid transmission. Clinical Infectious Diseases, 74(11), 1993-2000.

[13] Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (Eds.). (2010). Handbook of spatial statistics. CRC press.

[14] Gordon, M. A., Graham, S. M., Walsh, A. L., Wilson, L., Phiri, A., Molyneux, E., ... & Molyneux, M. E. (2008). Epidemics of invasive Salmonella enterica serovar enteritidis and S. enterica Serovar typhimurium infection associated with multidrug resistance among adults and children in Malawi. Clinical Infectious Diseases, 46(7), 963-969.

[15] Hawkes, A. G. and Oakes, D. (1974). A Cluster Process Representation of a Self-Exciting Process. Journal of Applied Probability 11 493-503.

[16] Esri. (2017). How Multi-Distance Spatial Cluster Analysis (Ripley's K-function) Works.

[17] Kim, H. (2011). Spatio-temporal point process models for the spread of avian influenza virus (H5N1). University of California, Berkeley.

[18] Moller, J., Syversveen, A.R. & Waagepetersen, R.P. (1998), Log Gaussian Cox Processes. Scandinavian Journal of Statistics, 25: 451-482. https://doi.org/10.1111/1467-9469.00115

[19] Mweu, E., & English, M. (2008). Typhoid fever in children in Africa. Tropical Medicine & International Health, 13(4), 532-540.

[20] Musicha, P., Cornick, J. E., Bar-Zeev, N., French, N., Masesa, C., Denis, B., ... & Feasey, N. A. (2017). Trends in antimicrobial resistance in bloodstream infection isolates at a large urban hospital in Malawi (1998–2016): a surveillance study. The Lancet infectious diseases, 17(10), 1042-1052.

[21] Ogata, Y. (1998). Space-time point-process models for earthquake occurrences. Annals of the Institute of Statistical Mathematics, 50, 379-402.

[22] Peters, R. P., Zijlstra, E. E., Schijffelen, M. J., Walsh, A. L., Joaki, G., Kumwenda, J. J., ... & Lewis, D. K. (2004). A prospective study of bloodstream infections as cause of fever in Malawi: clinical predictors and implications for management. Tropical Medicine & International Health, 9(8), 928-934.

[23] Reinhart, A. (2018). A Review of Self-Exciting Spatio-Temporal Point Process and Their Applications. Retrieved from https://arxiv.org/pdf/1708.02647.pdf

[24] Taylor, B. M., Davies, T. M., Rowlingson, B. S., & Diggle, P. J. (2015). Bayesian Inference and Data Augmentation Schemes for Spatial, Spatiotemporal and Multivariate Log-Gaussian Cox Processes in R. Journal of Statistical Software, 63(1), 1–48. https://doi.org/10.18637/jss.v063.i07

[25] Taylor, B. M., Davies, T. M., Rowlingson, B. S., & Diggle, P. J. (2013). lgcp: an R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. Journal of Statistical Software, 52, 1-40.

[26] Tranmer, M., & Steel, D.G. (1998). Using census data to investigate the causes of ecological fallacy. Envirnment and Planning A, 30, 817-831.

[27] Wailan, A. M., Coll, F., Heinz, E., Tonkin-Hill, G., Corander, J., Feasey, N. A., & Thomson, N. R. (2019). rPinecone: Define sub-lineages of a clonal expansion via a phylogenetic tree. Microbial genomics, 5(4).

[28] World Health Organization. (2014). Antimicrobial resistance: global report on surveillance. World Health Organization.

APPENDICES

Appendix 1: Log targets

This subsection presents the log - target plots for the remaining spatio-temporal LGCP models.

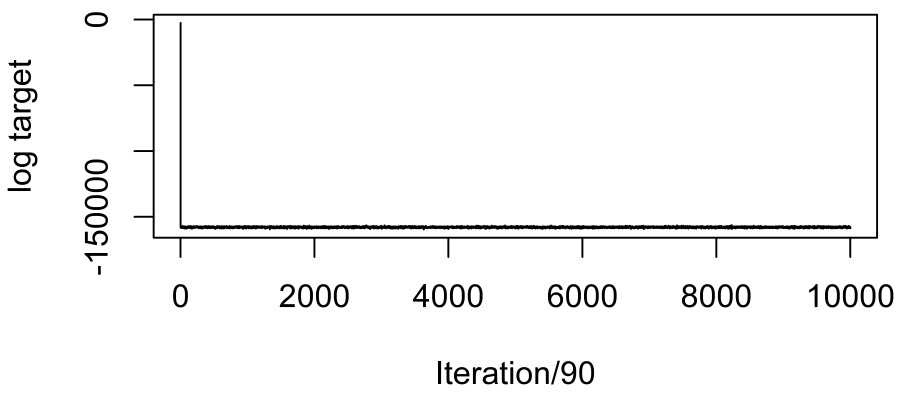


Figure 5.1: Log target plot for the spatio-temporal model for clade 0

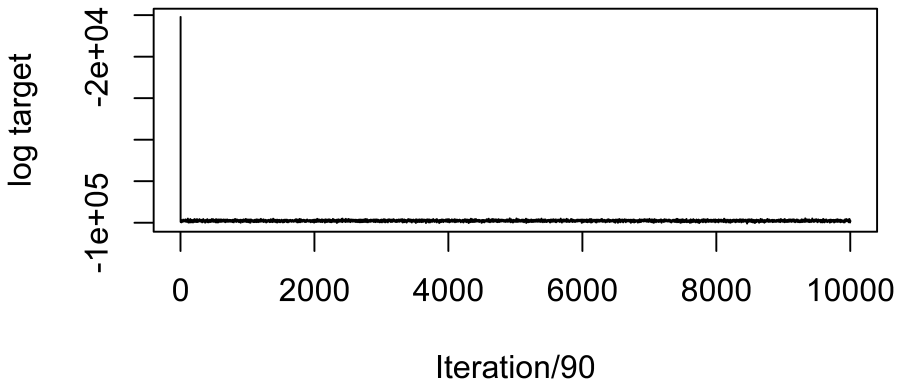


Figure 5.2: Log target plot for the spatio-temporal model for clade 2

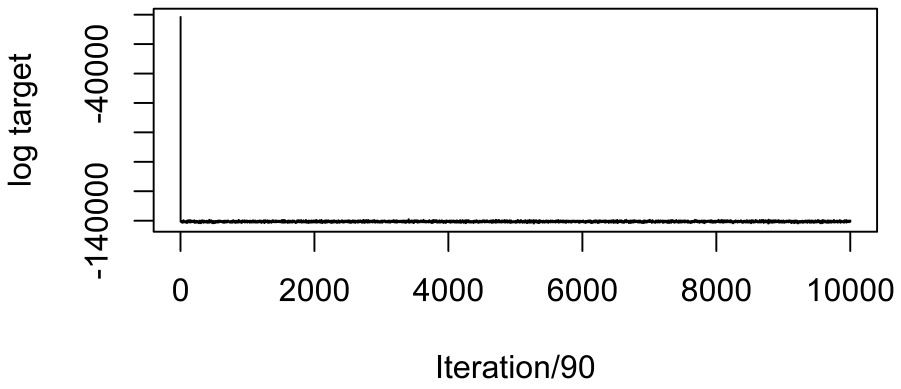


Figure 5.3: Log target plot for the spatio-temporal model for the grouped clades

Appendix 2: Traceplots

This subsection presents the traceplots for the remaining spatio-temporal LGCP models.

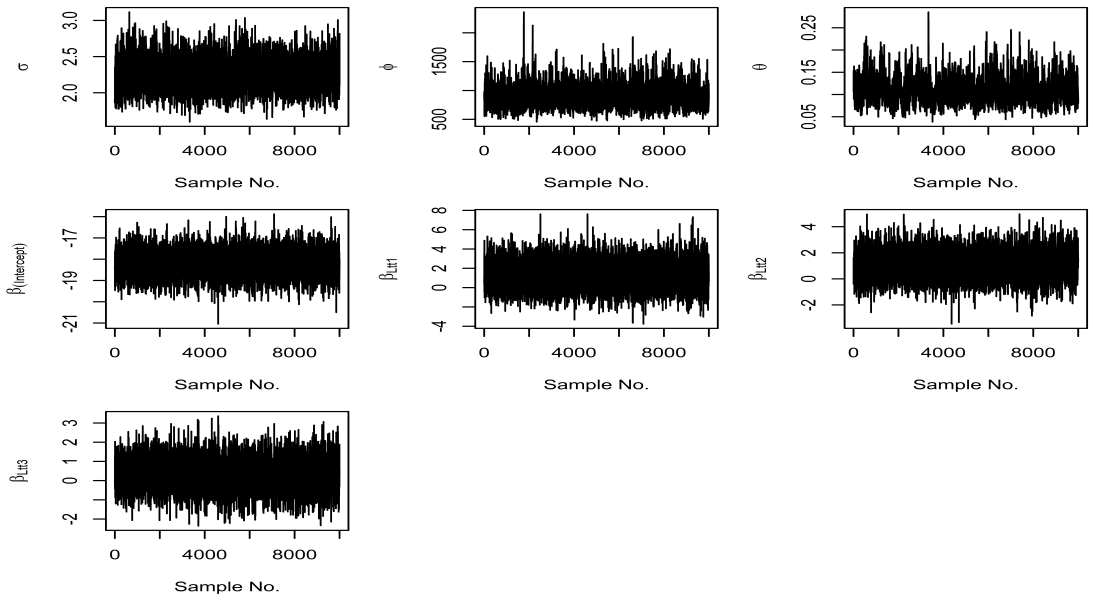


Figure 5.4: Traceplots for Beta and Eta for clade 0 cases

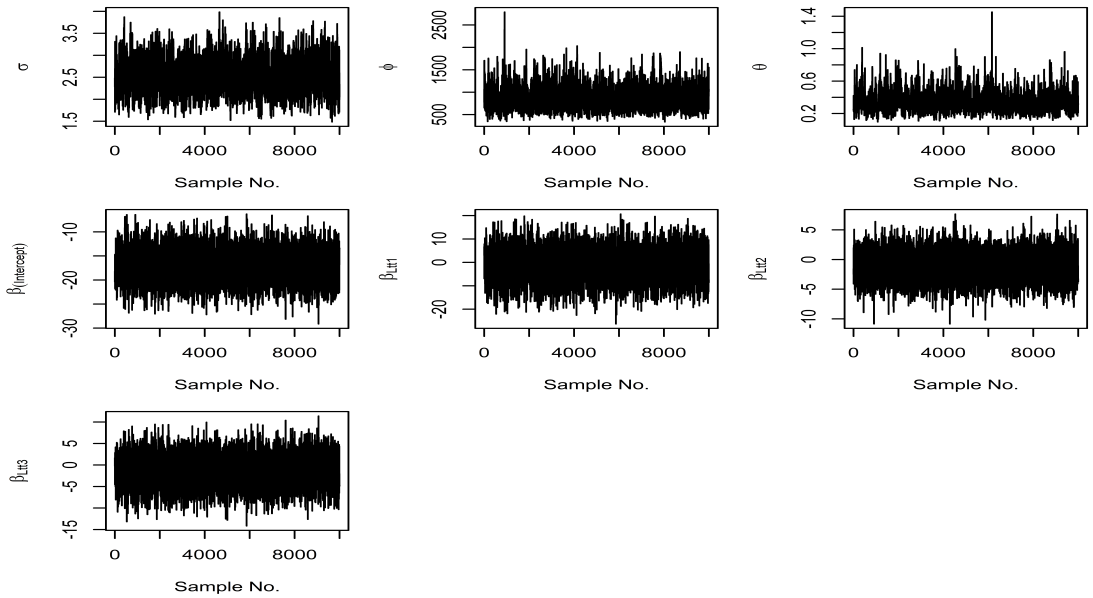


Figure 5.5: Traceplots for Beta and Eta for clade 2 cases

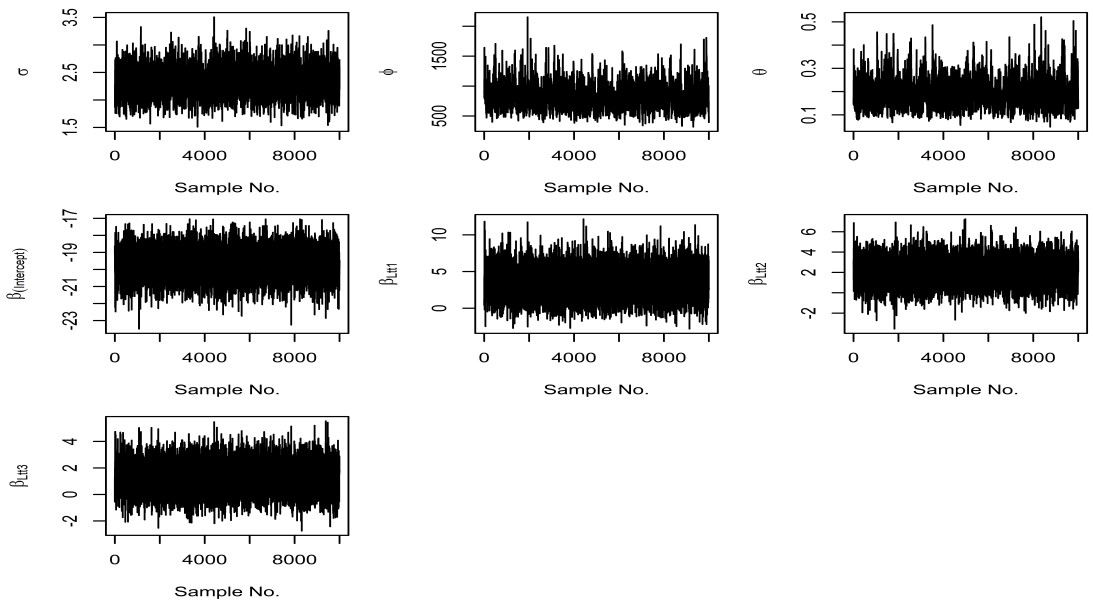


Figure 5.6: Traceplots for Beta and Eta for the grouped clades

Appendix 3: Autocorrelation in the latent field

This subsection presents plots of the autocorrelations in the latent field for different lags for the remaining spatio-temporal LGCP models.

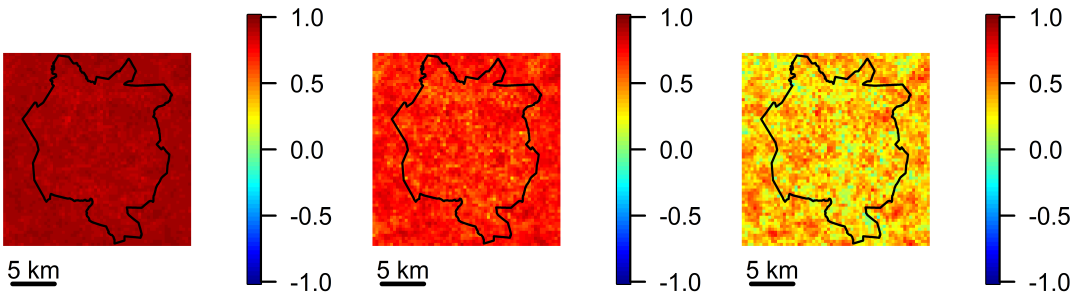


Figure 5.7: Left to right: autocorrelations in the Gaussian latent field at lag 1, lag 5 and lag 15 for spatio-temporal model with all cases

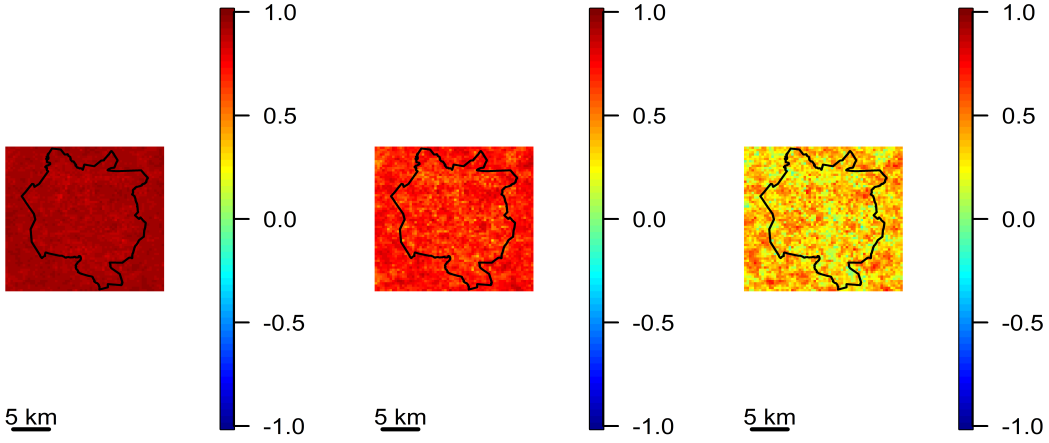


Figure 5.8: Autocorrelations in the latent field at different lags for clade 0 cases

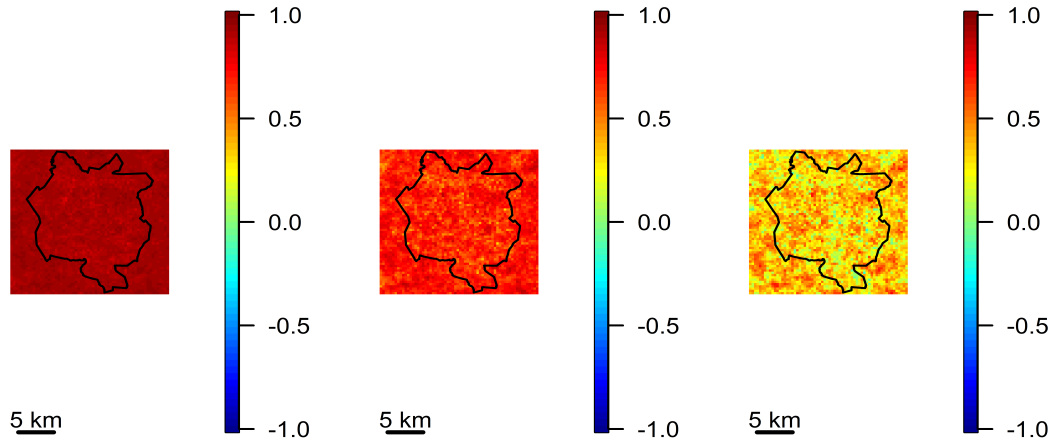


Figure 5.9: Autocorrelations in the latent field at different lags for clade 2 cases

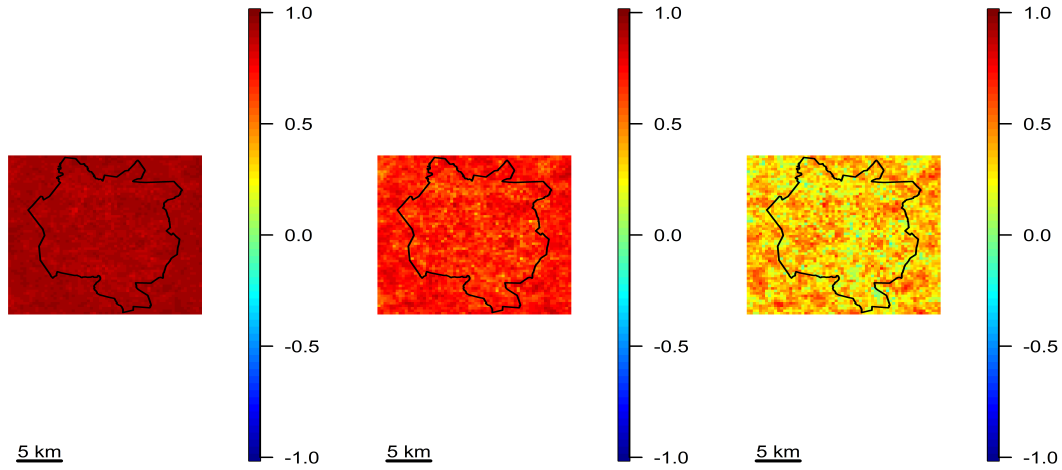


Figure 5.10: Autocorrelations in the latent field at different lags for the grouped clades

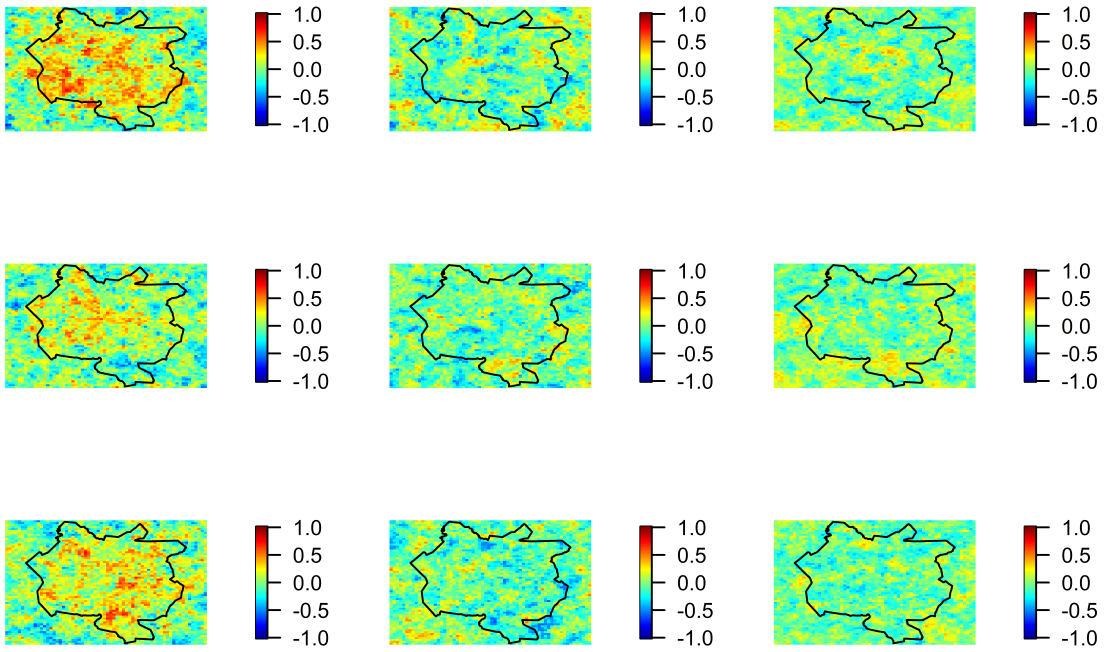


Figure 5.11: Left to right: autocorrelations in the Gaussian latent field at lag 1, lag 5 and lag 15 for multi-type spatial model

Appendix 4: Autocorrelation of parameters from the point process

This subsection presents plots for the autocorrelation of parameters at different lags for the remaining spatio-temporal LGCP models and the multi-type LGCP model.

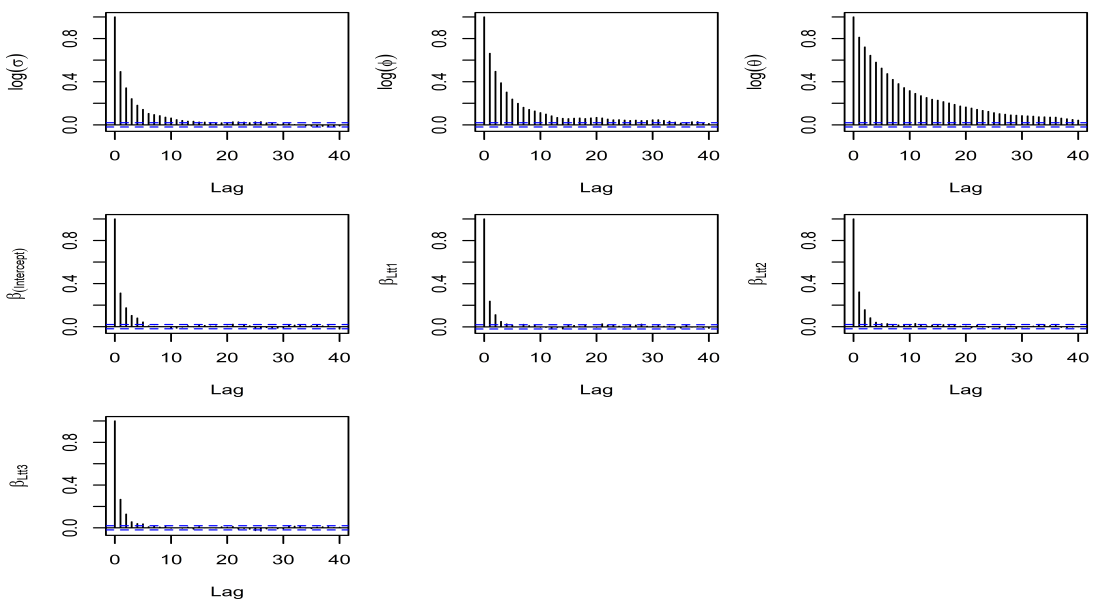


Figure 5.12: Autocorrelations in the latent field at different lags for spatio-temporal model with clade 0 cases

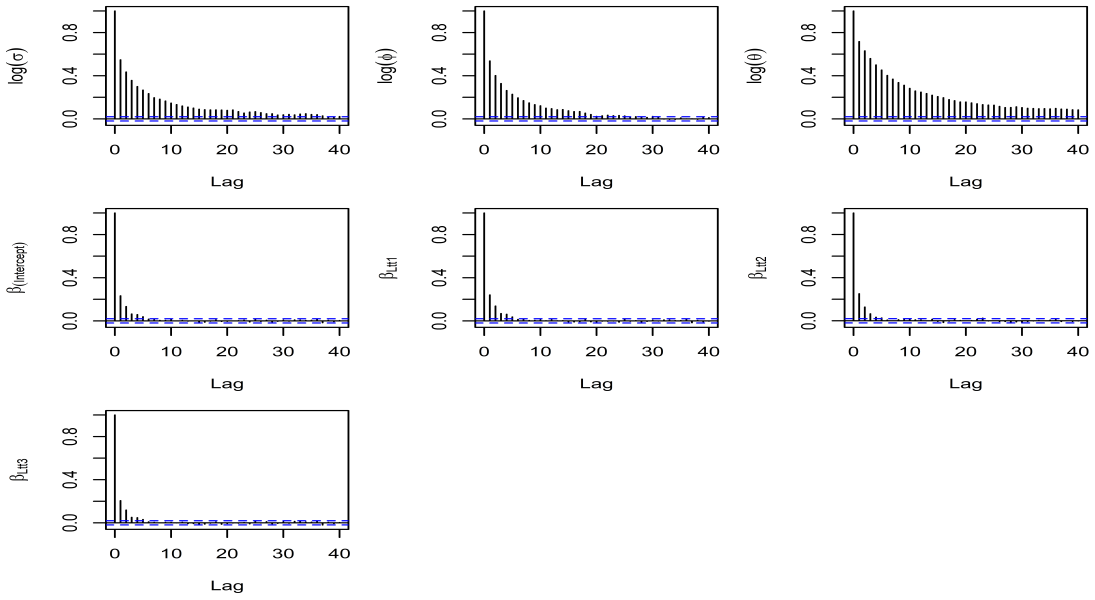


Figure 5.13: Autocorrelations in the latent field at different lags for spatio-temporal model with clade 2 cases

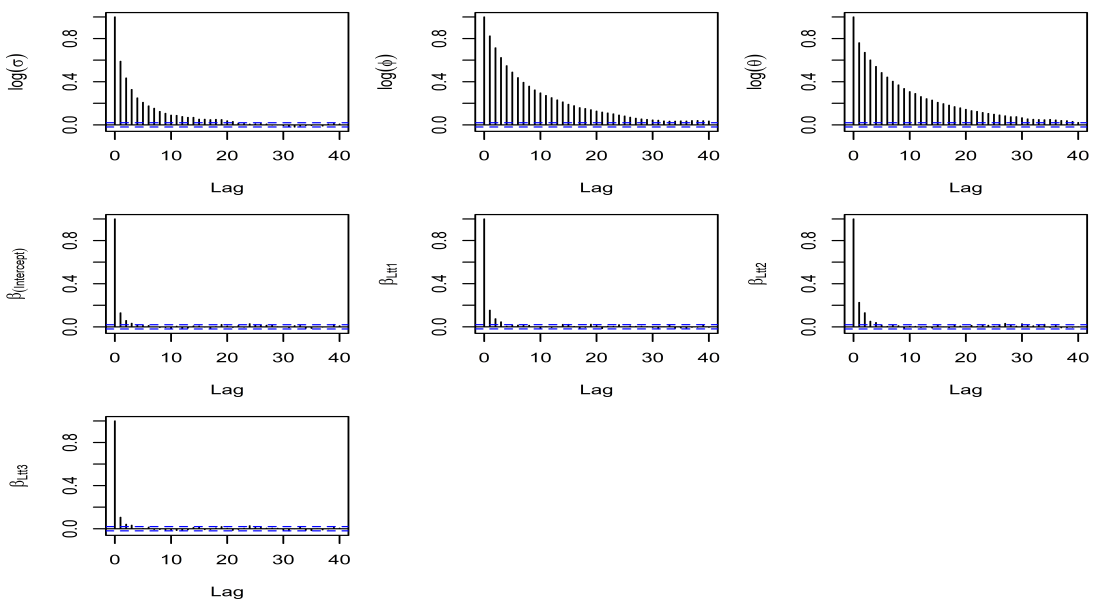


Figure 5.14: Autocorrelation plots of the parameters of the latent field for the grouped clades

Appendix 5: Posterior Covariance Function

This subsection presents posterior covariance function plots for the remaining spatio-temporal LGCP models and the multi-type LGCP model.

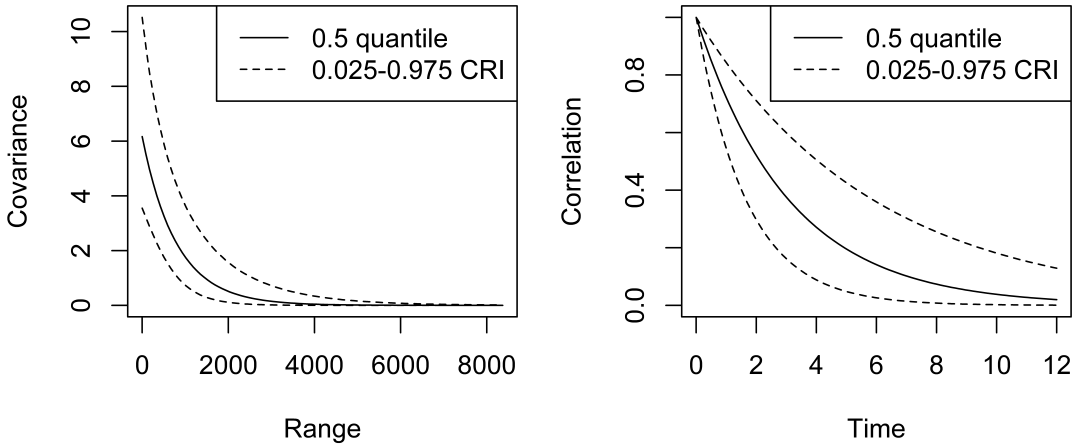


Figure 5.15: Plots of the posterior spatial covariance (Left) and temporal correlation (Right) for the Gaussian process of the spatio-temporal model for clade 2 cases

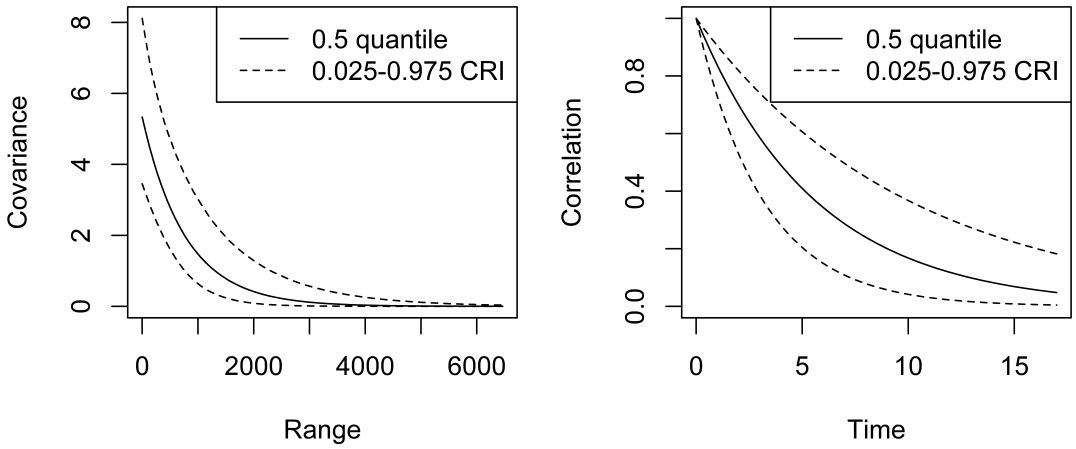


Figure 5.16: Plots of the posterior spatial covariance (Left) and temporal correlation (Right) for the spatio-temporal model for the grouped clades

Appendix 6: Incidence Rates

This subsection presents plots of incidence rates for the remaining spatio-temporal LGCP models.

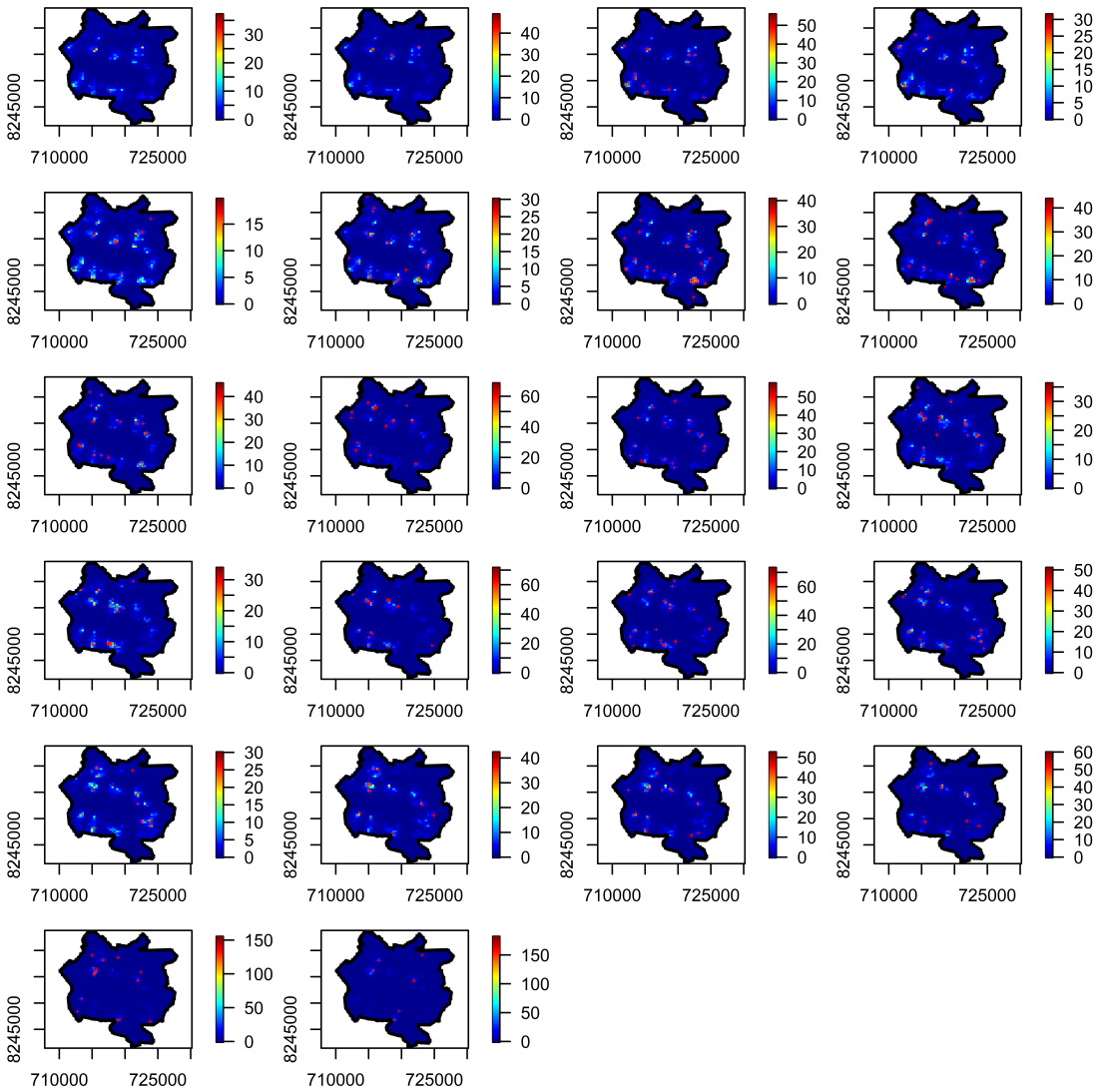


Figure 5.17: Incidence rate plot for the spatio-temporal model for all cases at every time point

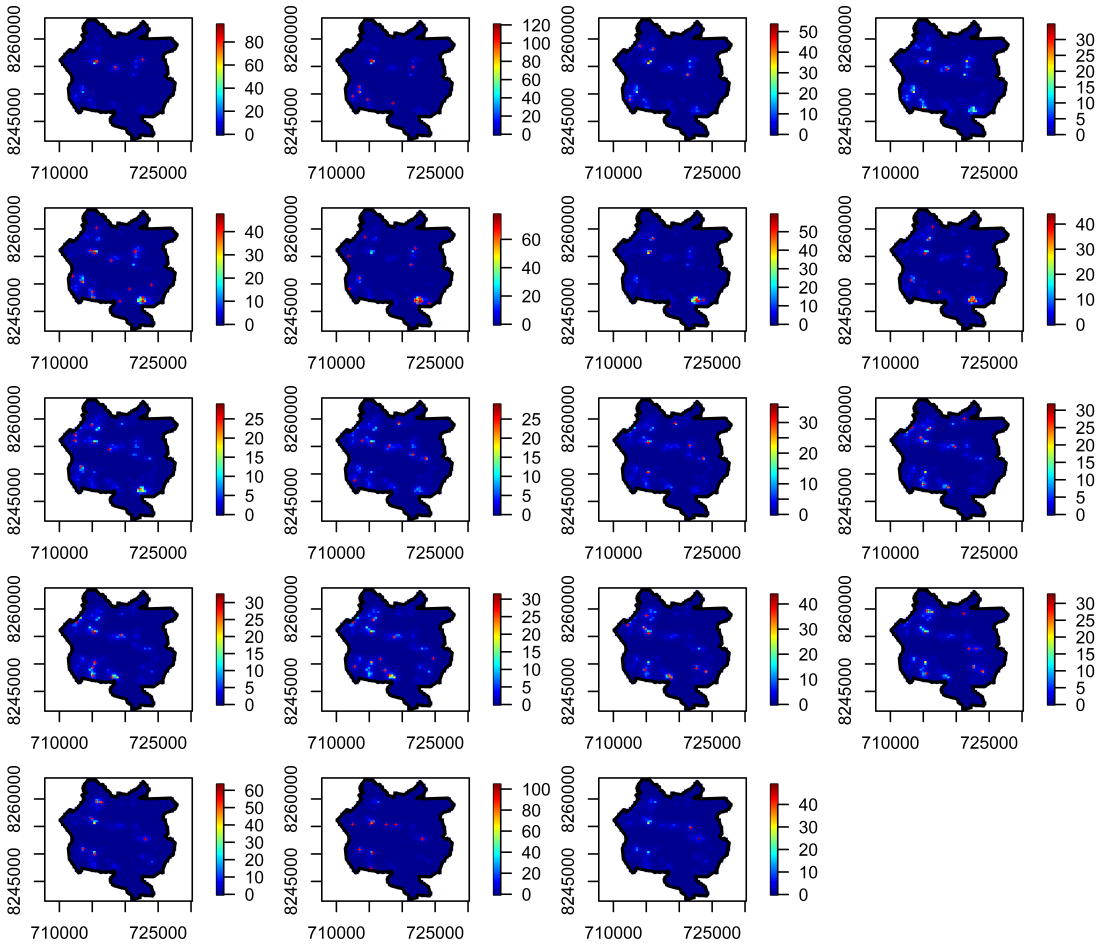


Figure 5.18: Incidence rate plot for the spatio-temporal model for clade 0 sub-lineage at different time point(months)

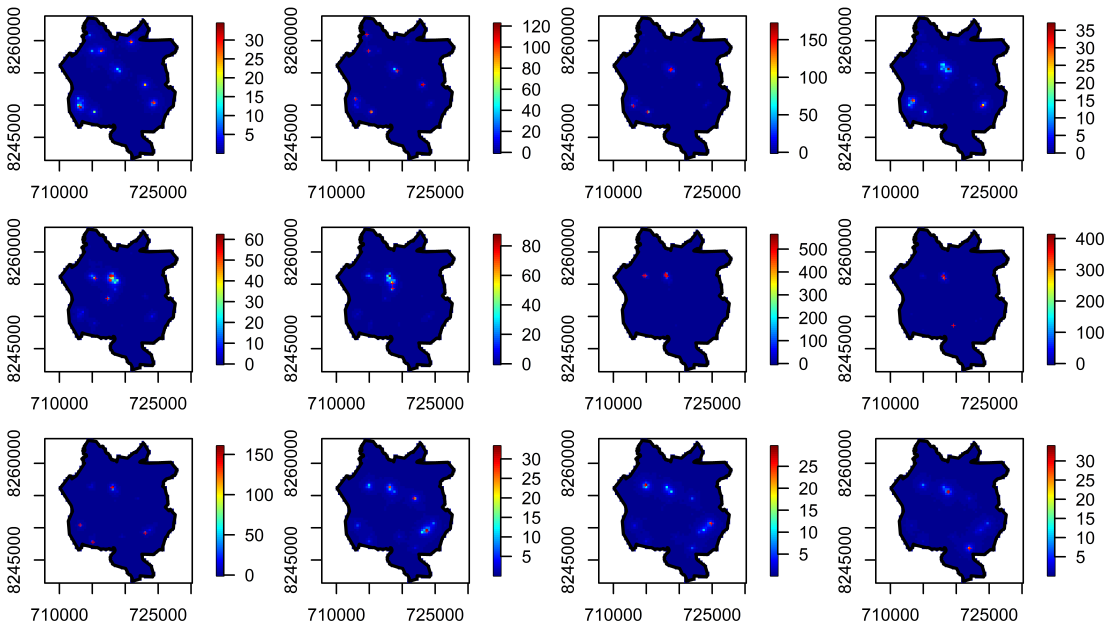


Figure 5.19: Incidence rate plot for the spatio-temporal model for clade 2 sub-lineage at different time point(months)

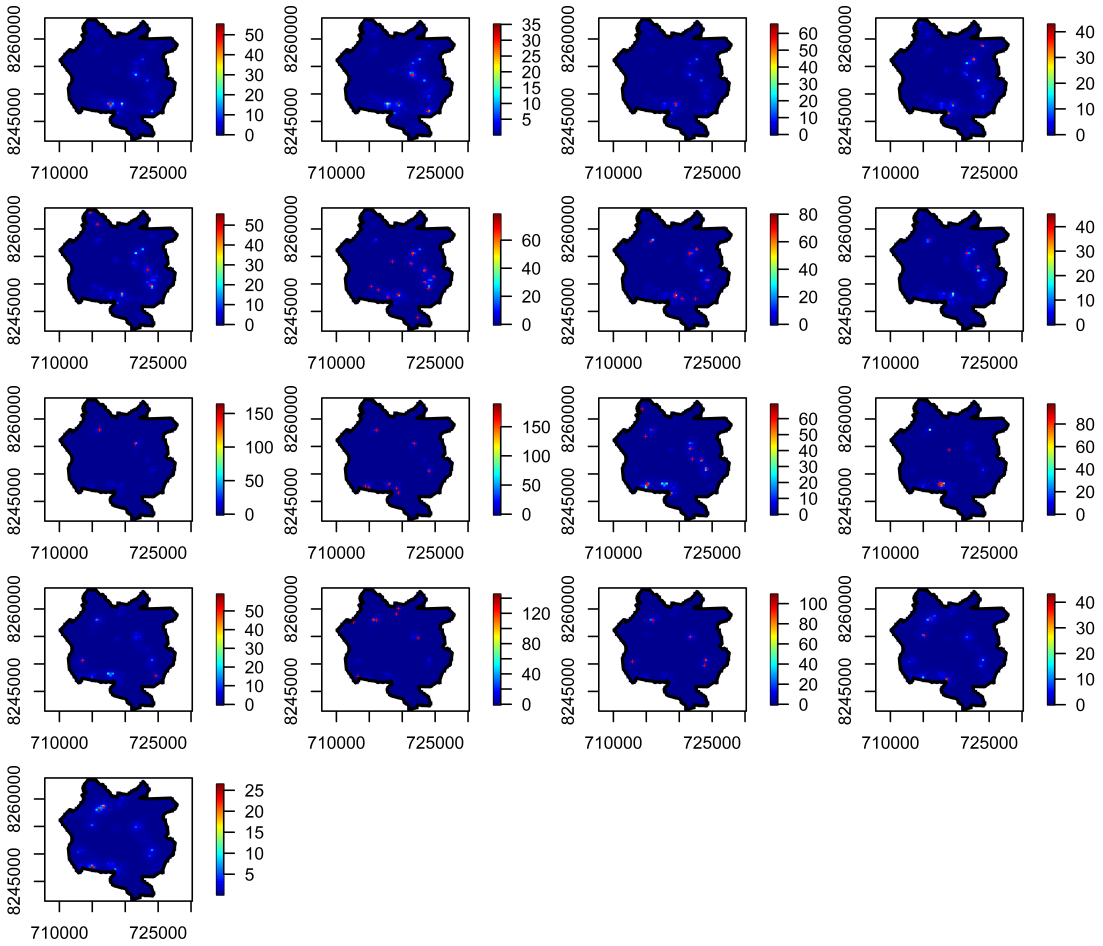


Figure 5.20: Incidence rate plot for the spatio-temporal model for the grouped clades at different time point(months)

Appendix 7: Standard error of the incidence rates

This subsection presents the standard error of the incidence rate plots for the remaining spatio-temporal LGCP models.

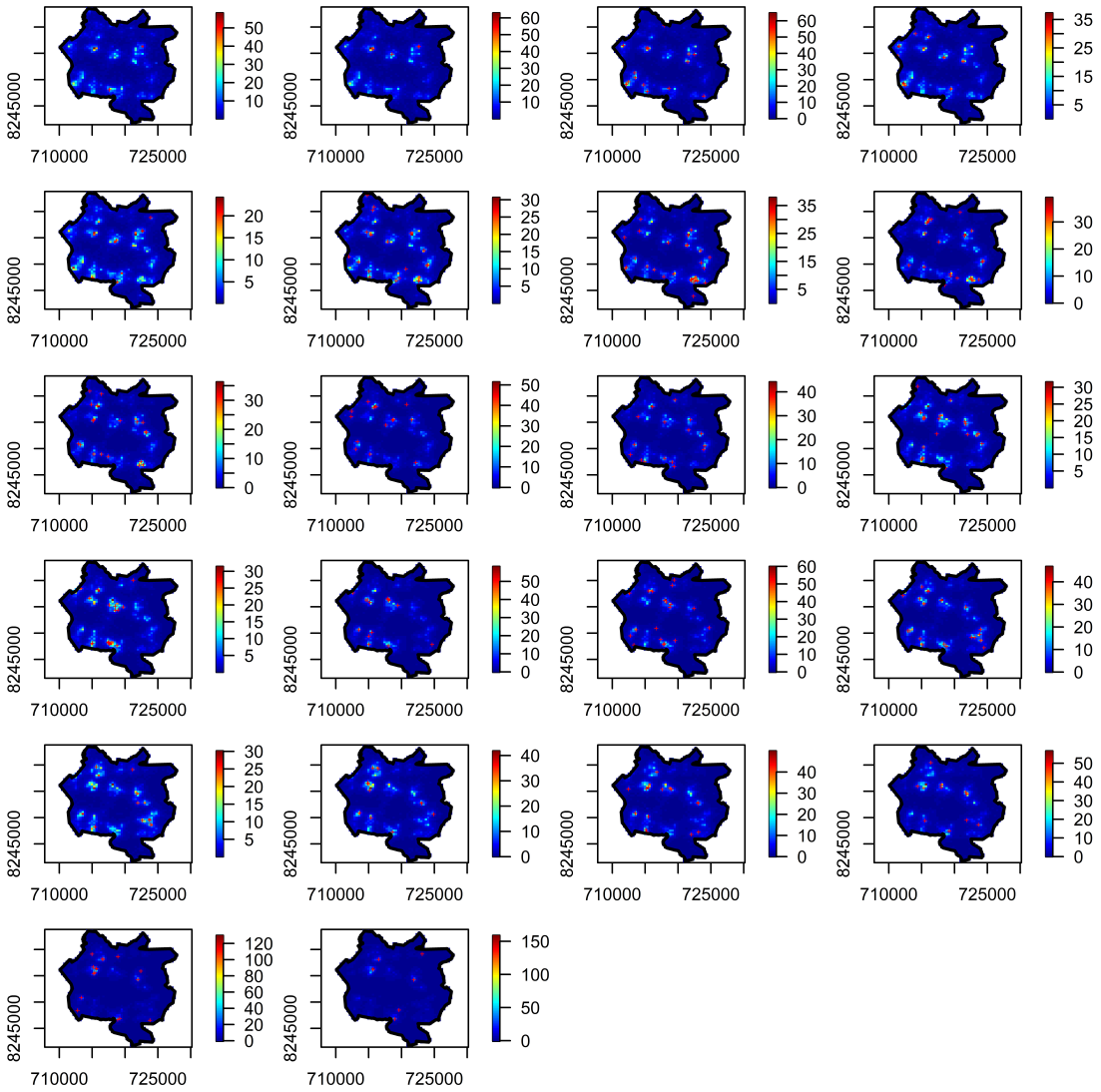


Figure 5.21: Standard error plot of the incidence rate for the spatio-temporal model for all cases at every time point(months)

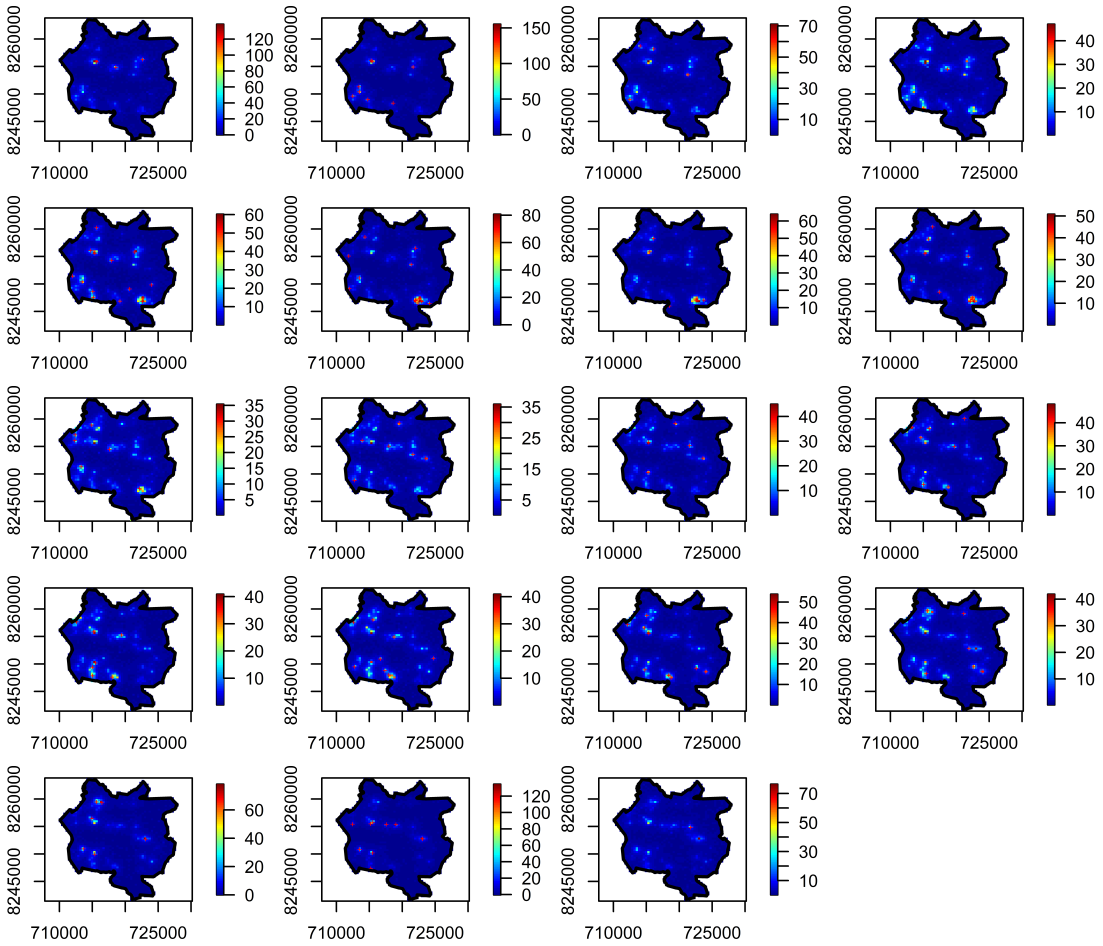


Figure 5.22: Standard error plot of the incidence rate for the spatio-temporal model for clade 0 at different time point (months)

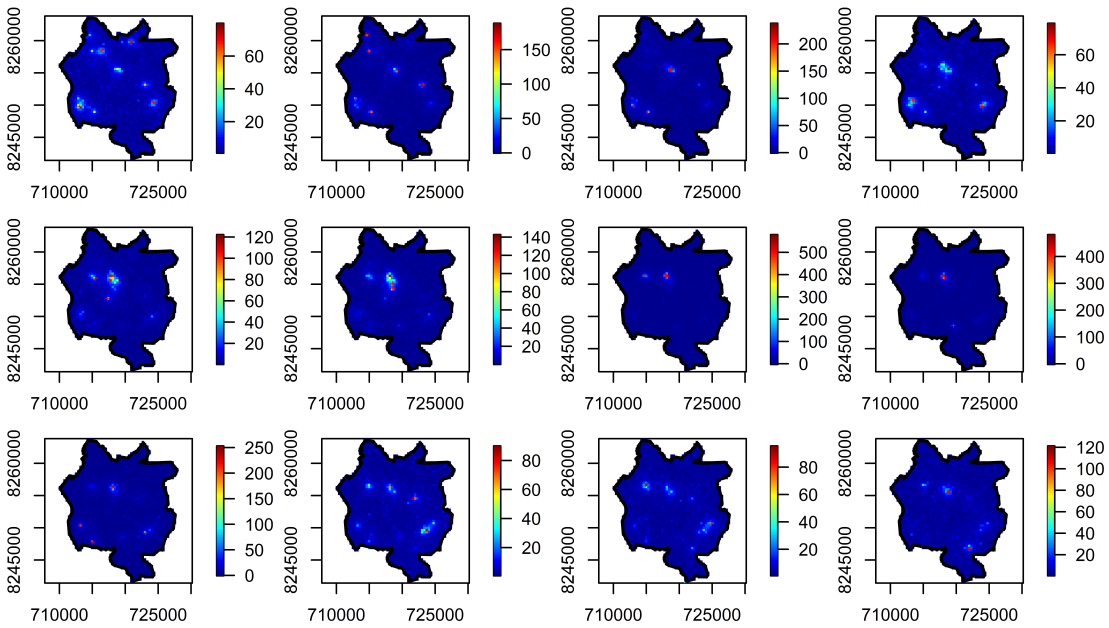


Figure 5.23: Standard error plot of the incidence rate for the spatio-temporal model for clade 2 at different time point (months)

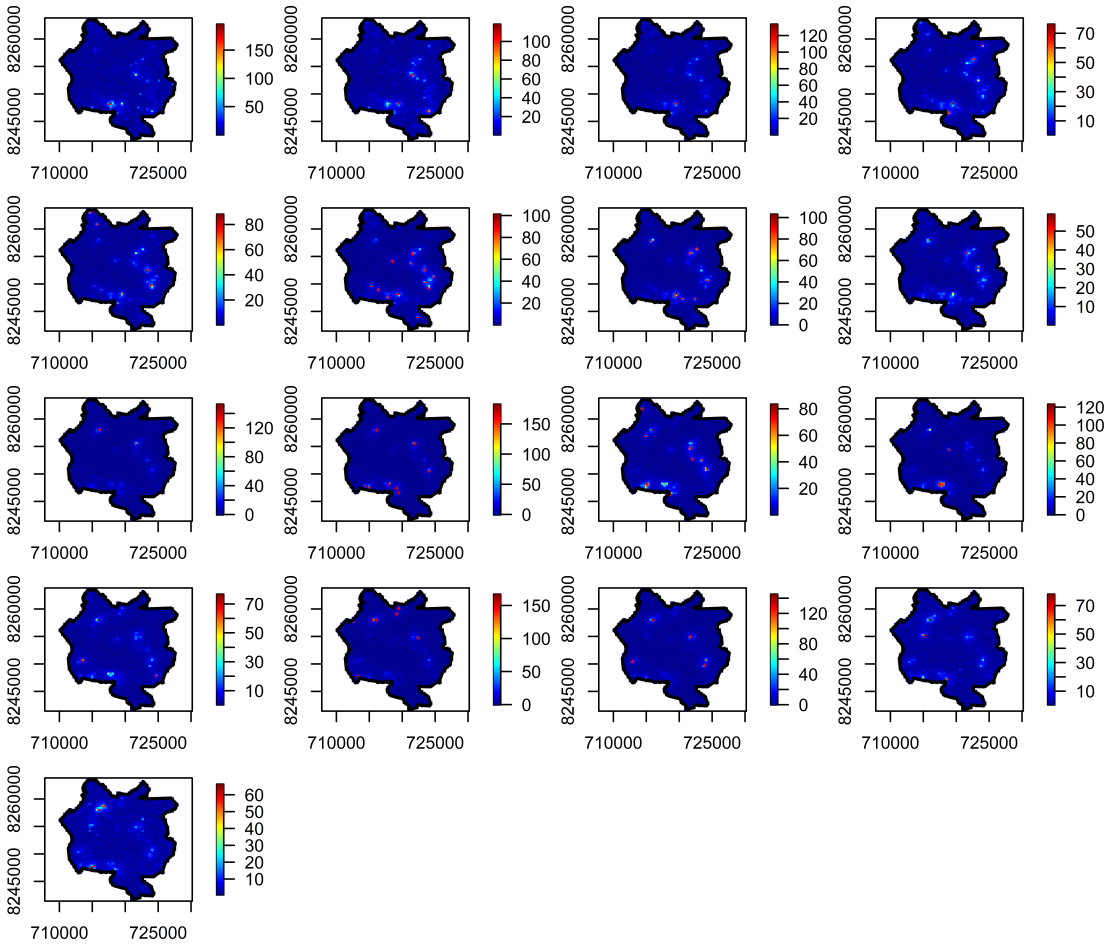


Figure 5.24: Standard error plot of the incidence rate for the spatio-temporal model for the grouped clades at different time point (months)

Appendix 8: Prior and posterior density plot

This subsection presents prior and posterior density plots for the remaining spatio-temporal LGCP models.

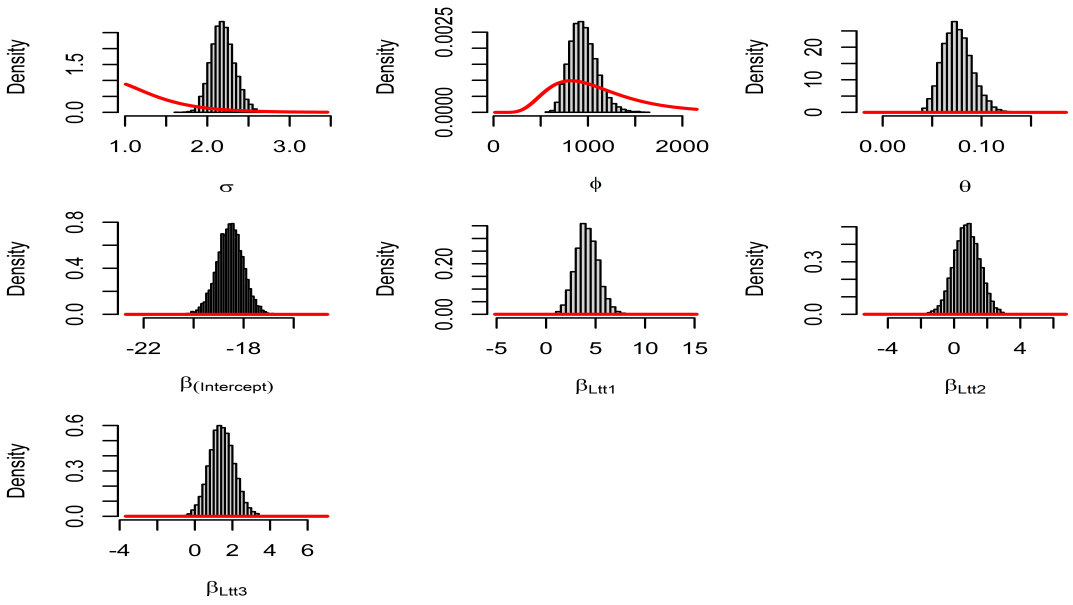


Figure 5.25: Prior (continuous curve) and posterior (histogram) distribution for the parameters of the spatio-temporal LGCP model with all cases

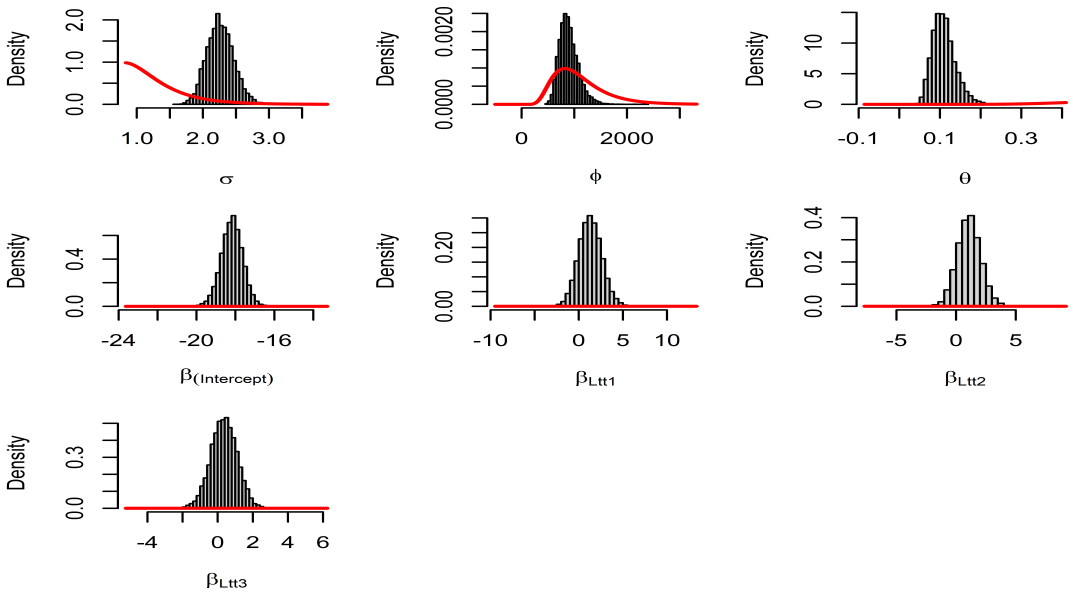


Figure 5.26: Prior and posterior density plots for the spatio-temporal model for clade 0 cases

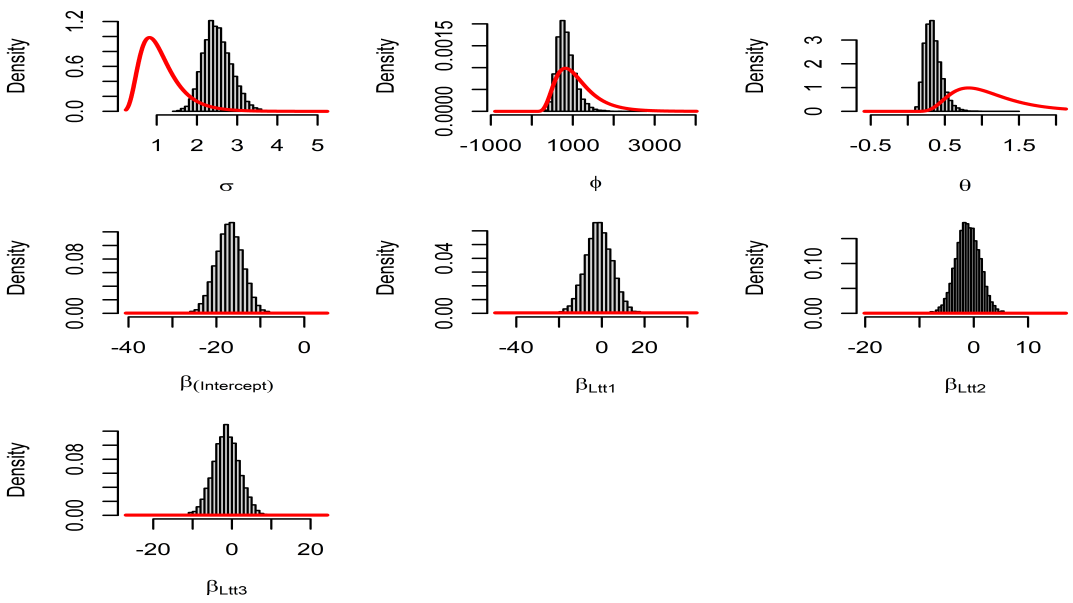


Figure 5.27: Prior and posterior density plots for the spatio-temporal model for clade 2 cases

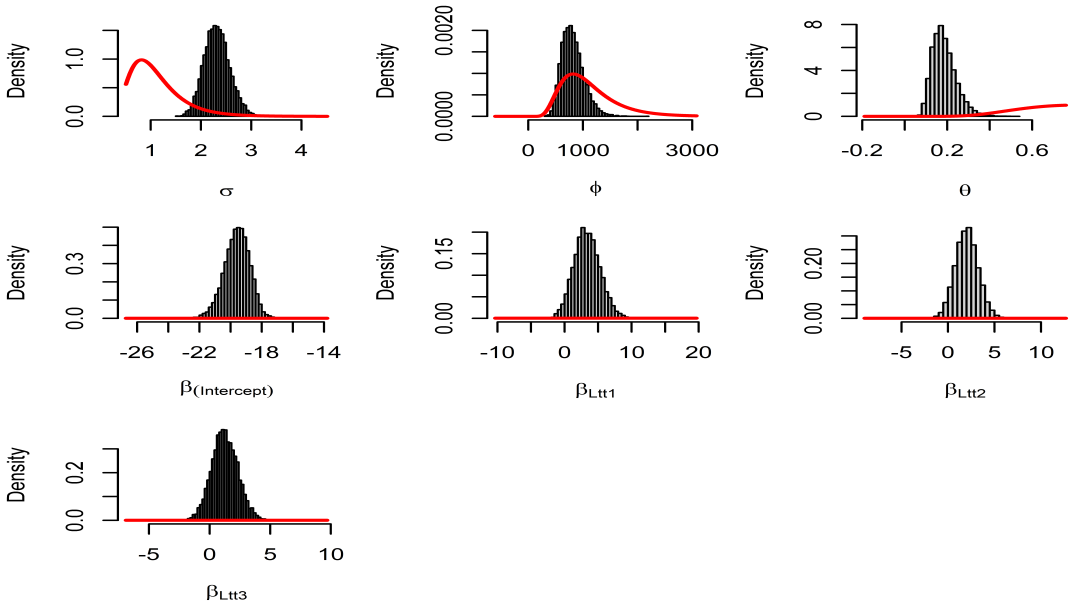


Figure 5.28: Prior and posterior density plots for the spatio-temporal model for the grouped clades

Appendix 9: Inhomogeneous K Function

This subsection presents Inhomogeneous K function plots for the remaining spatio-temporal LGCP models and the multi-type LGCP model.

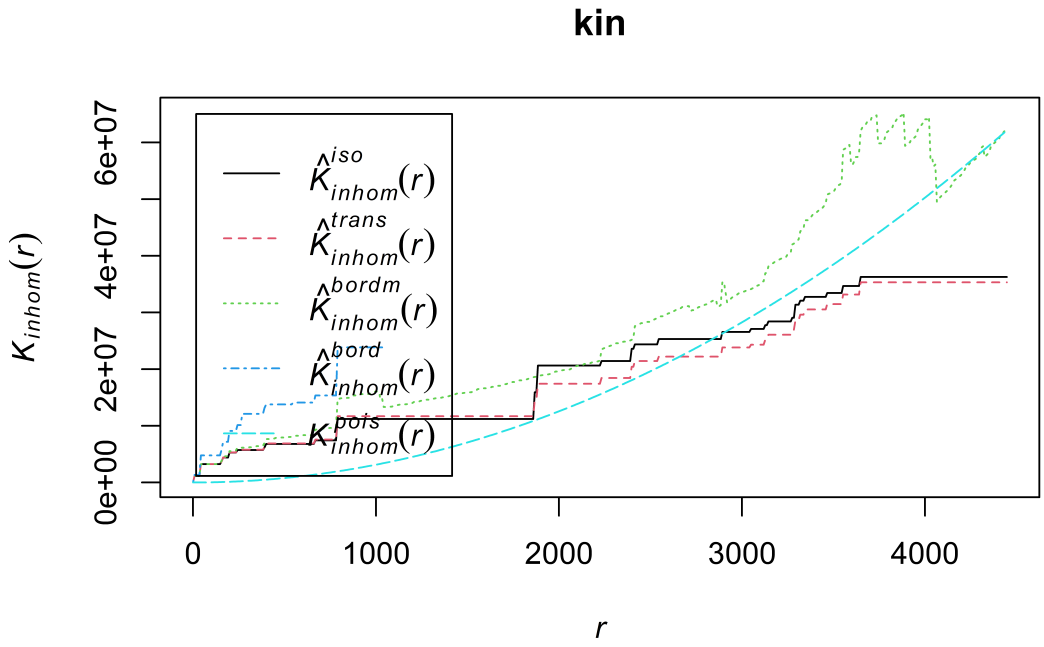


Figure 5.29: Inhomogeneous K Function for clade 2 cases

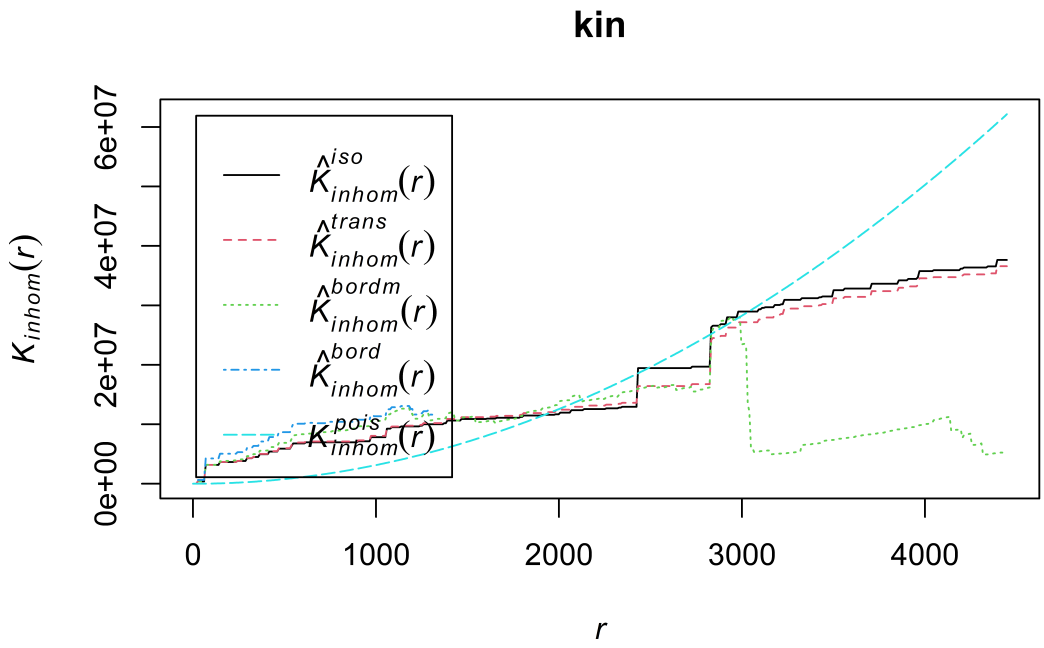


Figure 5.30: Inhomogeneous K Function for the grouped clades

Appendix 10: R Scripts

The R codes are too long to be included in this document. They can be accessed on GitHub using this link *https://github.com/donkalonga/MSc-Project-spatio-temporal-modelling* or by clicking here.