Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Spatial modeling of individual-level infectious disease transmission: tuberculosis data in Manitoba, Canada

Leila Amiri[1] | Mahmoud Torabi*[1,2] | Rob Deardon[3,4] | Michael Pickles[1]

[1]Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba, Canada
[2]Department of Statistics, Faculty of Science, University of Manitoba, Canada
[3]Department of Production Animal Health, Faculty of Veterinary Medicine, University of Calgary, Canada
[4]Department of Mathematics and Statistics, Faculty of Science, University of Calgary, Canada

Correspondence
Mahmoud Torabi, Department of Community Health Sciences, University of Manitoba, Winnipeg, MB R3E 0W3, Canada.
Email: Mahmoud.Torabi@umanitoba.ca

Summary

Geographically-dependent individual level models (GD-ILMs) are a class of statistical models that can be used to study the spread of infectious disease through a population in discrete-time in which covariates can be measured both at individual and area levels. The typical ILMs to illustrate spatial data are based on the distance between susceptible and infectious individuals. A key feature of GD-ILMs is that they take into account the spatial location of the individuals in addition to the distance between susceptible and infectious individuals. As a motivation of this paper, we consider tuberculosis (TB) data which is an infectious disease which can be transmited through individuals. It is also known that certain areas/demographics/-communities have higher prevalent of TB (see Section 4 for more details). It is also of interest of policy makers to identify those areas with higher infectivity rate of TB for possible preventions. Therefore, we need to analyze this data properly to address those concerns. In this paper, the Expectation Conditional Maximization algorithm is proposed for estimating the parameters of GD-ILMs to be able to predict the areas with the highest average infectivity rates of TB. We also evaluate the performance of our proposed approach through some simulations. Our simulation results indicate that the proposed method provides reliable estimates of parameters which confirms accuracy of the infectivity rates.

KEYWORDS:
Conditional autoregressive model; Expectation Conditional Maximization algorithm; Individual-level models; Susceptible-infected-removed model.

1 | INTRODUCTION

Despite the global efforts to reduce tuberculosis (TB) incidence and mortality rates, the World Health Organization (WHO) points out that TB is in the top ten causes of death worldwide in 2017. TB is an infectious disease that is transmitted from person to person through the air. The distance between infected and susceptible individuals is then one of the factors that affects on the TB incidence. Geographical location is also key in the TB burden such that 95% of TB cases and deaths are in developing countries such as India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh, and South Africa. Spatial locations are then important for understanding the spread of emerging infectious diseases and identifying their causes, since spatially varying demographic and environmental factors could influence the disease transmission (e.g., Kristoffersen et al,[1] Rees et al[2]). For instance, studies show that factors such as living in overcrowded conditions, ambient air pollution exposure, temperature,

humidity, sunlight, urbanization and racial/ethnic group have a significant effect on the incidence and transmission of the TB (Lienhardt,[3] Laumbach and Kipen,[4] Fares,[5] Smith et al,[6] Onozuka and Hagihara[7]). Due to the aforementioned points, the health statuses of individuals are correlated. So, the typical regression model which violates the correlations is not a suitable choice for analyzing this type of infectious disease. Also, one common approach for modeling correlated data is the population average model (aka marginal model) which addresses correlation in the responses through the covariance structure using generalized estimating equations. Since this model follows a different target, unable to provide an explanation for the source of correlation (e.g., the individuals distances heterogeneity) and it is not suitable for analyzing the type of infectious diseases which their prevalence are similar to TB. For infectious disease data analysis, it is very important to develop a model that can describe the transmission of disease dynamics based on spatial location of disease, spatial distance between individuals and population heterogeneity. The goal of this paper is to analyze the TB data in Manitoba, Canada, to address the above concerns. Although Canada is a low TB incidence country in the global context, the rate of TB incidence has been increasing since 2014. Therefore, identification of geographic areas that are at greater risk of mortality and morbidity of TB and also determining factors potentially responsible for increase of TB incidence can help policymakers to make effective healthcare and social policies for TB control.

Over more than a century, mathematical modeling of infectious diseases has become an important tool to understand patterns of disease spread and to provide significant knowledge in the design of programs for disease control intervention and/or prevention. One of the most important disease-transmission model is the equation based compartmental epidemic model introduced by Kermack and McKendrick[8], commonly known as susceptible-infected-removed (SIR) model. This model categories individuals in a population who are able to contract an infection, are capable of spreading the infection, and are removed from the susceptible population, respectively. For more information in both the history and use of SIR models see Keeling and Rohani.[9] Generally, such mathematical models have been used to study epidemic outbreaks of infectious diseases; for example, the 2001 foot-and-mouth disease epidemic in the UK (Keeling et al,[10] Jewell et al,[11] Deardon et al[12]), severe acute respiratory syndrome (McBryde et al[13]), Ebola (Chowell et al,[14] McKinley et al,[15] D'Silva and Eisenberg[16]), and HIV AIDS (Becker et al,[17] Perelson and Nelson,[18] Sweeting et al,[19] Taffe et al[20]).

In the recent years, the rapid growth in the availability of geographical data together with the advances in computational power, present a unique opportunity for the spatio-temporal analysis of infectious disease data. A vast literature exists on the spatio-temporal analysis of infectious disease data (see, e.g. Hughes et al,[21] Ster and Ferguson,[22] Deardon et al,[12] Kwong and Deardon,[23] Brown et al,[24] Pokharel and Deardon,[25] Mahsin et al[26]). In particular, individual-level model (ILM) of Deardon et al[12] has became widely used for the analysis of space-time infectious disease data. The term ILM refers to models which consider covariates and disease information at the level of individuals in a population. The proposed framework of Deardon et al[12] can be used to estimate the probability of a susceptible individual becoming infected in a given period of time, by allowing information on disease susceptibility and transmissibility factors (e.g. the number and/or types of animals on a farm, environmental factors, etc.), and the effect of the distance between infected and susceptible individuals to be incorporated in the models. However, these models assume that the probability of disease transmission between two individuals depend only on their spatial separation and not on their spatial locations. Recently, Mahsin et al[26] studied the effect of the geographical location on the outbreak of an infectious disease. They generalized the ILMs of Deardon et al[12] to a new class of geographically-dependent ILMs (GD-ILMs) to allow for the evaluation of the effect of spatially varying social risk factors (e.g., education, social deprivation), environmental factors, as well as unobserved spatial structure, upon the transmission of infectious disease. They used a mixed effect model for capturing the spatial correlation by incorporating spatially defined random effects in the model, with the well-known conditional autoregressive (CAR) model (Breslow and Clayton,[27] Leroux et al[28]).

So far, inference for ILMs has been mainly implemented within a Bayesian framework. The Bayesian approach focuses on post-data perspectives while the frequentist approach concentrate on the pre-data aspects of the statistical inference. The aim of this paper is to focus on a frequentist approach for fitting such ILMs. Maximum likelihood estimation of the GD-ILMs is carried out via the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin[29]), which is an accelerated variant of the Expectation Maximization (EM) algorithm (Dempster et al[30]). We evaluate the performance of the proposed method through simulation studies. Our results show that the GD-ILMs are able to correctly identify the infectious disease dynamics through accurate estimation of coefficients associated with both individual and areal-level covariates as well as spatial parameters. We investigate average infectivity rates in each area at different time point. Such quantities could help better target policy and infrastructure planning for prevention and control of infectious diseases. We further illustrate this method through the analysis of TB data in Manitoba, Canada.

The rest of this paper is organized as follows. In Section 2, the GD-ILMs are described. The proposed method of parameter estimation via the ECM algorithm is provided in Section 3. We analyze our TB data using the proposed method in Section 4.

In order to investigate the performance and accuracy of the proposed ECM algorithm to estimate the model parameters we use some simulation studies in Section 5. We conclude with a general discussion in Section 6. Technical details are deffered to the Appendix.

2 | INFECTIOUS DISEASE MODELING

2.1 | Individual-level models

Deardon et al[12] introduced a class of discrete time ILMs which are used as the starting point for models in this study. For the purposes of this paper, we briefly describe the ILM class within a SIR compartmental framework. A finite population of individuals ($i = 1, ..., n$) is assumed. At a discrete time point t, $t = 1, \dots, t_{max}$, (t_{max} is the last time point observed, often the length of the epidemic) an individual i can be in one, and only one, of three compartments of states (susceptible, infected, removed). We introduce the following notation. $S(t)$ refers to the set of individuals who are susceptible to the disease at time t; $I(t)$ refers to the set of individuals who are infectious and can infect others at time t; and finally $R(t)$ refers to the set of individuals who have been removed from the population at time t.

In the ILM of Deardon et al[12], the probability that a susceptible individual, i, will become infected at time t is denoted by

$$P(i,t) = 1 - \exp\Big(- \Omega_S(i) \sum_{j \in I(t)} \Omega_T(j) k(i,j) + \epsilon(i,t) \Big), \tag{1}$$

where $\Omega_S(i)$ and $\Omega_T(j)$ are functions that represent susceptibility factors for individual i and transmissibility factors for individual j, respectively. $k(i,j)$ is the infection kernel representing shared risk factors between susceptible and infectious individuals. The infection kernel could be a function of some types of contact between susceptible individual i and infectious individual j representing connections within the observed population that allow for possible transfer of infective agents (e.g. Euclidean or road distance). $\epsilon(i,t)$ is the sparks term, representing a function of risk factors associated with some random behaviour that is not well explained by the other terms in the model.

Let Y_{it}, $i = 1, \dots, n$, $t = 1, \dots, t_{max}$, be an event that a susceptible individual i is infected at time t. So, the probability mass function (p.m.f.) of Y_{it} can be written as $P(Y_{it} = y) = P(i,t)^y (1 - P(i,t))^{1-y}$, $y = 0, 1$, which $y = 1$ means that individual is infected. For a random sample of size n, the likelihood function of $\mathbf{y} = (y_1, \dots, y_n)^\top$ can be written as

$$f(\mathbf{y}) = \prod_{t=1}^{t_{max}} \left\{ \prod_{i \in S(t+1)} \Big(1 - P(i,t)\Big) \prod_{i \in I(t+1) \setminus I(t)} P(i,t) \right\},$$

where $S(t+1)$ is the set of all susceptible individuals observed not to be infected at time t and $I(t+1) \setminus I(t)$ is the set of all newly infected individuals within this interval.

2.2 | Geographically-dependent ILMs

In the standard ILM (1), it is assumed that the spatial function $k(i,j)$ depends only upon the distance between susceptible and infectious individuals, but not the spatial location of the individuals. However, spatial locations are often important for understanding the spread of emerging infectious diseases and identifying their causes. Recently, in order to investigate the effect of spatially varying risk factors upon the transmission of infectious disease, an extension of the ILM was proposed by Mahsin et al[26]. This model is called geographically-dependent ILM (GD-ILM). The probability of a susceptible individual i being infected at time t in area (region) z in the GD-ILM framework has the form:

$$P(i,z,t) = 1 - \exp\Big(- \Omega_S(i,z) \sum_{j \in I(t,z,\xi(z))} \Omega_T(j,z) k(i,j) + \epsilon(i,z,t) \Big), \tag{2}$$

where z represents the area index which varies from 1 to m, $\xi(z)$ is the set of neighboring areas that are adjacent to area z, and $I(t,z,\xi(z))$ is the set of infectious individuals at time t in the zth location and its neighbouring areas. $\Omega_S(i,z)$ and $\Omega_T(j,z)$ are susceptibility factors for individual i in area z and transmissibility factors for individual j in area z, respectively. $k(i,j)$ is the infection kernel that represents shared risk factors associated with both susceptible individual i and infectious individual j; $\epsilon(i,z,t)$ called a spark function, represents a function of the risk factors associated with some random behaviour that considers infections not well-explained by other model components.

In this study, we define two forms of equation (2), in which the vector of p covariates associated with susceptible individual i, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\top$, and the vector of q area-level covariates corresponding to area z, $\mathbf{X}_z = (X_{z1}, \ldots, X_{zq})^\top$. The covariates $\mathbf{X}_i$ and $\mathbf{X}_z$ are incorporated into the susceptibility term $\Omega_S(i, z)$. Also, for simplicity, the transmissibility term, $\Omega_T(j, z)$, is set to 1. The GD-ILMs that incorporate these covariates are defined as:

$$\textbf{Model 1}: \quad P(i,z,t) = 1 - \exp\Big(- \exp(\alpha + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + \mathbf{X}_z^\top \boldsymbol{\beta}_2 + u_z) \sum_{j \in I(t,z)} d_{ij}^{-\delta} \Big), \tag{3}$$

and

$$\textbf{Model 2:} \quad P(i,z,t) = 1 - \exp\Big(- \exp(\alpha + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + \mathbf{X}_z^\top \boldsymbol{\beta}_2 + u_z) \sum_{j \in I(t,z,\xi(z))} d_{ij}^{-\delta} \Big), \tag{4}$$

where α is the intercept, $\boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^\top$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2q})^\top$ are the vectors of parameters corresponding to $\mathbf{X}_i$ and $\mathbf{X}_z$, respectively. u_z represents spatial random effects that provide a way of accounting for latent geographic variation or unmeasured covariates effect, via the specification of some spatial structure between areas. $I(t, z)$ is the set of infectious individuals at time t in the zth area. The infection kernel is defined as $k(i, j) = d_{ij}^{-\delta}$ where $\delta > 0$ is the spatial parameter and d_{ij} is the Euclidean (earth) distance between susceptible individual i and infectious individual j.

In Model 1, it is assumed that the diseases spread out from infected individuals to susceptible individuals within each area, but not from infectious individuals in neighboring areas, while in the Model 2, infectious individuals in adjacent areas have an effect on the diseases spread through $I(t, z, \xi(z))$, where $\xi(z)$ is the set of neighboring areas that are adjacent to area z. For each model, we define $P(i, z, t)$ as the rate of infectivity to a susceptible individual i at time point t in a given area z and also, the average infectivity rate at time point t in area z is calculated as

$$\bar{\eta_z}(t) = n_z^{-1} \sum_{i=1}^{n_z} P(i,z,t), \tag{5}$$

where n_z is the number of individuals in the zth area.

To fit Models 1 and 2, many different choices of spatial random effects, $\mathbf{u} = (u_1, \ldots, u_m)^\top$, are available in the context of disease mapping (see e.g. Lee[31]). Among these, the method of Leroux et al[28] is appealing because it allows for a weighted combination of spatial independence and strong spatial dependence (Leroux et al[28]). Within this framework, the distribution function of $\mathbf{u}$ is defined as

$$\mathbf{u} \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_u), \tag{6}$$

where generalized inverse of $\boldsymbol{\Sigma}_u$ is defined as $\boldsymbol{\Sigma}_u^- = \tau^2[(1-\lambda)\mathbf{I}_m + \lambda \mathbf{D}]$ in which τ^2 and $\lambda \in [0, 1]$ quantify precision and spatial dependence, respectively. A larger value of $\lambda \in [0, 1]$ indicates a higher degree of spatial dependence. This specification yields two extreme cases: (i) $\lambda = 0$ implies completely independent random effects and (ii) $\lambda = 1$ implies intrinsic auto-regressive model (Besag et al[32]). $\mathbf{I}_m$ is the identity matrix of dimension m and $\mathbf{D}$ is a $m \times m$ matrix reflecting neighborhood structure. Typically, neighbors are those areas that share a common boundary. The typical element of $\mathbf{D}$ is given by

$$d_{zz'} = \begin{cases} g_z, & z = z', \\ -\mathrm{I}\{z \sim z'\}, & z \neq z'. \end{cases}$$

where g_z is the number of neighbors of area z, $z \sim z'$ indicates that regions z and z' are neighbours, and I is the indicator function.

3 | PARAMETER ESTIMATION VIA THE ECM ALGORITHM

EM-type algorithms have been shown to be powerful computational techniques for obtaining ML estimates of model parameters when data are not fully observed or the model contains latent variables. However, it is quite often that the EM algorithm cannot be directly applied because of its maximization (M) step being difficult to compute. To obviate this weakness, Meng and Rubin[29] proposed an extension of the EM, called the Expectation Conditional Maximization (ECM) algorithm, which is easy to implement and more broadly applicable than the EM. The key feature of the ECM is to replace the M-step of the EM with several analytically tractable conditional maximization (CM) steps. Moreover, it shares all the appealing features of the EM and can show faster convergence in terms of number of iterations or total computation time than the EM algorithm. Let $\mathbf{y}, \mathbf{z}$ and $\mathbf{u}$ indicate the vectors of responses, corresponding area of each response and spatial random effects, respectively. In this paper,

to compute ML estimates of $\boldsymbol{\Theta} = (\alpha, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \tau^2, \lambda, \delta)$, based on a random sample $(\mathbf{y}; \mathbf{z}; \mathbf{u})$, we adopt the ECM algorithm as $\mathbf{u}$ is latent. For this purpose, the likelihood function of the complete data $\mathbf{y}_c = (\mathbf{y}; \mathbf{z}; \mathbf{u})$ is computed as follows:

$$L(\boldsymbol{\Theta}; \mathbf{y}_c) = f_1(\mathbf{y}|\mathbf{u}, \mathbf{z}) f_2(\mathbf{u}), \tag{7}$$

where $f_2(\mathbf{u})$ is defined in (6) and the p.m.f of the $f_1(\mathbf{y}|\mathbf{u}, \mathbf{z})$, for Model 1 and Model 2, respectively, can be computed as follows:

$$f_1(\mathbf{y}|\mathbf{u}, \mathbf{z}) = \prod_{t=1}^{T} \left\{ \prod_{i \in S(t+1,z)} \prod_{z=1}^{m} \Big(1 - P(i,z,t)\Big)^{I(Z_{it}=z)} \prod_{i \in I(t+1,z) \setminus I(t,z)} \prod_{z=1}^{m} \Big(P(i,z,t)\Big)^{I(Z_{it}=z)} \right\},$$

$$f_1(\mathbf{y}|\mathbf{u}, \mathbf{z}) = \prod_{t=1}^{T} \left\{ \prod_{i \in S(t+1,z)} \prod_{z=1}^{m} \Big(1 - P(i,z,t)\Big)^{I(Z_{it}=z)} \prod_{i \in I(t+1,z,\xi(z)) \setminus I(t,z,\xi(z))} \prod_{z=1}^{m} \Big(P(i,z,t)\Big)^{I(Z_{it}=z)} \right\},$$

where $I(Z_{it} = z)$ is an indicator function such that for $i = 1, \ldots, n$, $t = 1, \ldots, t_{max}$, $z = 1, \ldots, m$

$$I(Z_{it} = z) = \begin{cases} 1, & ith \text{ individual at time } t \text{ is in } zth \text{ area}, \\ 0, & \text{otherwise}. \end{cases}$$

3.1 | E-step

In order to compute the conditional expected value of the complete-data log-likelihood, $\mathbf{y}_c$, given the observed data, $\mathbf{y}_o = (\mathbf{y}; \mathbf{z})$, we need to determine the conditional distribution of the latent variable, $\mathbf{u}$, given the observed data. Using Bayes rule, we have $f(\mathbf{u}|\mathbf{y}, \mathbf{z}) = \frac{f_1(\mathbf{y}|\mathbf{u},\mathbf{z}) f_2(\mathbf{u})}{f(\mathbf{y}|\mathbf{z})}$. $f(\mathbf{y}|\mathbf{z})$ cannot generally be expressed in closed form and must be approximated. Among the possible approximation methods, Laplace approximation is used here. The Laplace method has been designed to approximate integrals as follows:

$$\int_{R^d} \exp(h(\mathbf{u}))\, d\mathbf{u} \approx (2\pi)^{\frac{d}{2}} |-H(\hat{\mathbf{u}})|^{-\frac{1}{2}} \exp(h(\hat{\mathbf{u}})), \tag{8}$$

where $H(\hat{\mathbf{u}}) = \frac{\partial^2 h(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}^\top}|_{\mathbf{u}=\hat{\mathbf{u}}}$, and $h(\mathbf{u})$ is a known, uni-modal, and bounded function of a d-dimensional variable $\mathbf{u}$, and $\hat{\mathbf{u}}$ is the value for which $h(\mathbf{u})$ is maximized. $h(\mathbf{u}) = \log(f_1(\mathbf{y}|\mathbf{u}, \mathbf{z})) + \log(f_2(\mathbf{u}))$ and we adopt the Nelder-Mead method to find the $\hat{\mathbf{u}}$ that maximizes $h(\mathbf{u})$. Also, according to Theorem 4.14 of Evans and Swartz[33], we have

$$\int_{R^d} g(\mathbf{u}) \exp(h(\mathbf{u}))\, d\mathbf{u} \approx g(\hat{\mathbf{u}})(2\pi)^{\frac{d}{2}} |-H(\hat{\mathbf{u}})|^{-\frac{1}{2}} \exp(h(\hat{\mathbf{u}})), \tag{9}$$

where $\hat{\mathbf{u}}$ is the value that maximizes the $h(\mathbf{u})$. For computing the expectation value for each function $g(\mathbf{u})$, we are required to evaluate ratios of integrals of the form

$$E[g(\mathbf{u})|\mathbf{y}_o, \boldsymbol{\Theta}] = \frac{\int_{R^d} g(\mathbf{u}) f_1(\mathbf{y}|\mathbf{u}, \mathbf{z}) f_2(\mathbf{u}) \mathbf{du}}{\int_{R^d} f_1(\mathbf{y}|\mathbf{u}, \mathbf{z}) f_2(\mathbf{u}) \mathbf{du}}.$$

From (8) and (9), this expectation is approximated by

$$E[g(\mathbf{u})|\mathbf{y}_o, \boldsymbol{\Theta}] \approx g(\hat{\mathbf{u}}).$$

In the $(k+1)$th iteration of the EM algorithm this expectation is evaluated at $\boldsymbol{\Theta}^{(k)}$ where $\boldsymbol{\Theta}^{(k)}$ denotes the value of $\boldsymbol{\Theta}$ obtained in the k^{th} iteration of the EM algorithm.

3.2 | M-step

Here, we consider Model 1. Estimation of parameters for the Model 2 can be found in the same manner, simply replacing $\sum_{j \in I(t,z)}$ with $\sum_{j \in I(t,z,\xi(z))}$.

According to (7), the conditional expectation of the complete data log-likelihood given observed data, $\mathbf{y}_o$, is

$$E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big] = \sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)E\left[-\exp\left(\alpha+\mathbf{X}_i^\top\boldsymbol{\beta}_1+\mathbf{X}_z^\top\boldsymbol{\beta}_2+u_z\right)\sum_{j\in I(t,z)} d_{ij}^{-\delta}\middle|\mathbf{y}_o\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)E\left[\ln\left(1-\exp\Big(-\exp(\alpha+\mathbf{X}_i^\top\boldsymbol{\beta}_1+\mathbf{X}_z^\top\boldsymbol{\beta}_2+u_z)\sum_{j\in I(t,z)} d_{ij}^{-\delta}\Big)\right)\middle|\mathbf{y}_o\right]$$
$$-\frac{m}{2}\ln(2\pi)+\frac{m}{2}\ln(\tau^2)+\frac{1}{2}\ln\Big(\big|\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\big|\Big)-\frac{\tau^2}{2}E\left[\mathbf{u}^\top\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)\mathbf{u}\middle|\mathbf{y}_o\right]. \tag{10}$$

We then need to maximize (10) with respect to model parameters.

3.2.1 | CM-step 1

α is solution of the following equation

$$\frac{\partial E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\alpha} = -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\exp(\alpha^{(k)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\exp(\alpha^{(k)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]. \tag{11}$$

Unfortunately, there is no closed-form solution for the equation (11). We then employ the Newton-Raphson method to compute the solution to (11). In this regard, we have

$$\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\alpha^2} = -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\exp(\alpha^{(k)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\Bigg\{\exp(\alpha^{(k)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$-\exp(2\alpha^{(k)}+2\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+2\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\Big(\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}\Big)^2E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]\Bigg\}.$$

The following recursive relationship (first-order Taylor expansion) is obtained for computing α:

$$\alpha^{(k+1)} = \alpha^{(k)} - \frac{\frac{\partial}{\partial\alpha}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}{\frac{\partial^2}{\partial\alpha^2}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}.$$

3.2.2 | CM-step 2

Fix $\alpha=\alpha^{(k+1)}$. Analogous to the previous estimation, to estimate $\boldsymbol{\beta}_1$, we can write

$$\frac{\partial E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\boldsymbol{\beta}_1} = -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)} d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right],$$

and

$$\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^\top}=-\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m}I(Z_{it}=z)\mathbf{X}_i\mathbf{X}_i^\top\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m}I(Z_{it}=z)\Bigg\{\mathbf{X}_i\mathbf{X}_i^\top\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$-\mathbf{X}_i\mathbf{X}_i^\top\exp(2\alpha^{(k+1)}+2\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k)}+2\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\Big(\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}\Big)^2E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]\Bigg\}.$$

The following recursive relationship is obtained for computing $\boldsymbol{\beta}_1$:

$$\boldsymbol{\beta}_1^{(k+1)}=\boldsymbol{\beta}_1^{(k)}-\frac{\frac{\partial}{\partial\boldsymbol{\beta}_1}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}{\frac{\partial^2}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^\top}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}.$$

3.2.3 | CM-step 3

Fix $\alpha=\alpha^{(k+1)}$ and $\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^{(k+1)}$. We update $\boldsymbol{\beta}_2^{(k)}$ by maximizing (10) over $\boldsymbol{\beta}_2$, which gives

$$\frac{\partial E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\boldsymbol{\beta}_2}=-\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m}I(Z_{it}=z)\mathbf{X}_z\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m}I(Z_{it}=z)\mathbf{X}_z\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right],$$

and

$$\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2^\top}=-\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m}I(Z_{it}=z)\mathbf{X}_z\mathbf{X}_z^\top\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m}I(Z_{it}=z)\Bigg\{\mathbf{X}_z\mathbf{X}_z^\top\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$-\mathbf{X}_z\mathbf{X}_z^\top\exp(2\alpha^{(k+1)}+2\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+2\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k)})\Big(\sum_{j\in I(t,z)}d_{ij}^{-\delta^{(k)}}\Big)^2E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]\Bigg\}.$$

The following recursive relationship is obtained for computing $\boldsymbol{\beta}_2$:

$$\boldsymbol{\beta}_2^{(k+1)}=\boldsymbol{\beta}_2^{(k)}-\frac{\frac{\partial}{\partial\boldsymbol{\beta}_2}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}{\frac{\partial^2}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2^\top}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}.$$

3.2.4 | CM-step 4

Fix $\alpha=\alpha^{(k+1)}$, $\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^{(k+1)}$ and $\boldsymbol{\beta}_2=\boldsymbol{\beta}_2^{(k+1)}$. We update $\delta^{(k)}$ by maximizing (10) over δ, which leads to

$$\frac{\partial E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\delta}=\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m}I(Z_{it}=z)\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k+1)})\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\delta^{(k)}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m}I(Z_{it}=z)\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k+1)})\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\delta^{(k)}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right],$$

and

$$\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\delta^2} = -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k+1)})\sum_{j\in I(t,z)}(\ln(d_{ij}))^2 d_{ij}^{-\delta^{(k)}} E\left[\exp(u_z)\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\Bigg\{\exp(\alpha^{(k+1)}+\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k+1)})\sum_{j\in I(t,z)}(\ln(d_{ij}))^2 d_{ij}^{-\delta^{(k)}} E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]$$
$$-\exp(2\alpha^{(k+1)}+2\mathbf{X}_i^\top\boldsymbol{\beta}_1^{(k+1)}+2\mathbf{X}_z^\top\boldsymbol{\beta}_2^{(k+1)})\Big(\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\delta^{(k)}}\Big)^2 E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\right]\Bigg\}.$$

Hence, we have the following recursive relationship, for computing δ:

$$\delta^{(k+1)} = \delta^{(k)} - \frac{\frac{\partial}{\partial\delta}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}{\frac{\partial^2}{\partial\delta^2}E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big]}.$$

3.2.5 | CM-step 5

To update τ and λ on the M-step of the $(k+1)$th iteration, the following equation should be maximized

$$E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big] = -\frac{m}{2}\ln(2\pi)+\frac{m}{2}\ln(\tau^2)+\frac{1}{2}\ln\Big(\det\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)\Big)$$
$$-\frac{\tau^2}{2}E\Big[\mathbf{u}^T\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)\mathbf{u}\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big].$$

Hence, τ and λ are obtained by a Newton-Raphson iterative procedure as follows:

$$\begin{pmatrix}\tau\\ \lambda\end{pmatrix}^{new} = \begin{pmatrix}\tau\\ \lambda\end{pmatrix}^{old} + B^{-1}A,$$

where A and B are the score vector and the expected information matrix which their elements can be defined as follows:

$$A(\tau) = \frac{\partial E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\tau} = \frac{m}{\tau}-\tau E\Big[\mathbf{u}^\top\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)\mathbf{u}\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big],$$
$$A(\lambda) = \frac{\partial E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\lambda} = \frac{1}{2}tr\Big(\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)^{-1}\Big(\mathbf{D}-\mathbf{I}\Big)\Big)-\frac{\tau^2}{2}E\Big[\mathbf{u}^\top\Big(\mathbf{D}-\mathbf{I}\Big)\mathbf{u}\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big],$$
$$B(\tau,\tau) = -\frac{\partial^2 E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{(\partial\tau)^2} = \frac{m}{\tau^2}+E\Big[\mathbf{u}^\top\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)\mathbf{u}\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big],$$
$$B(\tau,\lambda) = -\frac{\partial^2 E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\tau\partial\lambda} = \tau E\Big[\mathbf{u}^\top\Big(\mathbf{D}-\mathbf{I}\Big)\mathbf{u}\Big|\mathbf{y}_o,\boldsymbol{\Theta}^{(k)}\Big],$$
$$B(\lambda,\lambda) = -\frac{\partial^2 E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{(\partial\lambda)^2} = \frac{1}{2}tr\Big(\Big(\mathbf{D}-\mathbf{I}\Big)\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)^{-1}\Big(\mathbf{D}-\mathbf{I}\Big)\Big(\lambda\mathbf{D}+(1-\lambda)\mathbf{I}\Big)^{-1}\Big).$$

Using CM-step 1 to CM-step 5 we get updated model parameters at iteration $(k+1)$, and continue this procedure until all model parameters converge. The variances of the model parameter estimates are provided in the Appendix A using a Fisher information matrix approach.

4 | DATA ANALYSIS: TUBERCULOSIS

TB is one of the leading causes of death worldwide. TB is caused by Mycobacterium tuberculosis bacteria that most often affect the lungs. TB bacteria are spread from person to person through the air. When a person with lung TB coughs, speaks, or sings, he/she propels the TB germs into the air, and a person nearby needs to inhale only a few of these bacteria to become infected. For more information about TB, we refer the readers to Smith[34] and Daniel[35]. Worldwide, a total of 10 million people fell ill with TB in 2017, among whom only 6.4 million new cases of TB were reported to the WHO. This means that worldwide, 36% of new cases went undiagnosed or were not reported. Studies have shown that men are at higher risk of contracting TB infection, and dying from TB than women (Narasimhan et al,[36] Horton et al[37]). In 2017, close to 3.2 million adult women contracted TB and around 500,000 died from it, while it has been estimated that 6 million adult men fell ill and almost 840,000 died from it. While Canada is a low TB incidence country, both the case count and rate have been increasing since 2014, especially in the foreign born and indigenous Canadians population (LaFreniere et al[38]). Also, among different age groups, the highest number of TB cases was among those aged 25-34 years.

In this study, we consider cases of TB in the Canadian province of Manitoba from 2017-2018. The data were provided by Manitoba Health, and include gender, age, date of diagnosis and geographic location reported via postal code of residence. Among 588 TB patients diagnosed between January 2017 and October 2018, 245 females and 343 males were recorded. Figure 1a provides the number of cases by age and sex. The highest number of TB cases occurred in the age group 25-34, and except the first age group (< 25), the number of affected men was more than that of women. We incorporate age and sex as individual level covariates in (3) and (4). The province of Manitoba consists of 11 Regional Health Authorities (RHAs) and is further divided to 59 RHA Districts (RHADs), 23 of which are part of Winnipeg. Each TB patient was geocoded to one of the 59 neighborhood areas in Manitoba using their six digit postal code at the time of TB diagnosis, note that we did not have two diagnosed TB patients with the same postal code. These 59 RHADs constitute the geographic areas used in the analysis. The number of TB incidence in each area is plotted in Figure 1b.

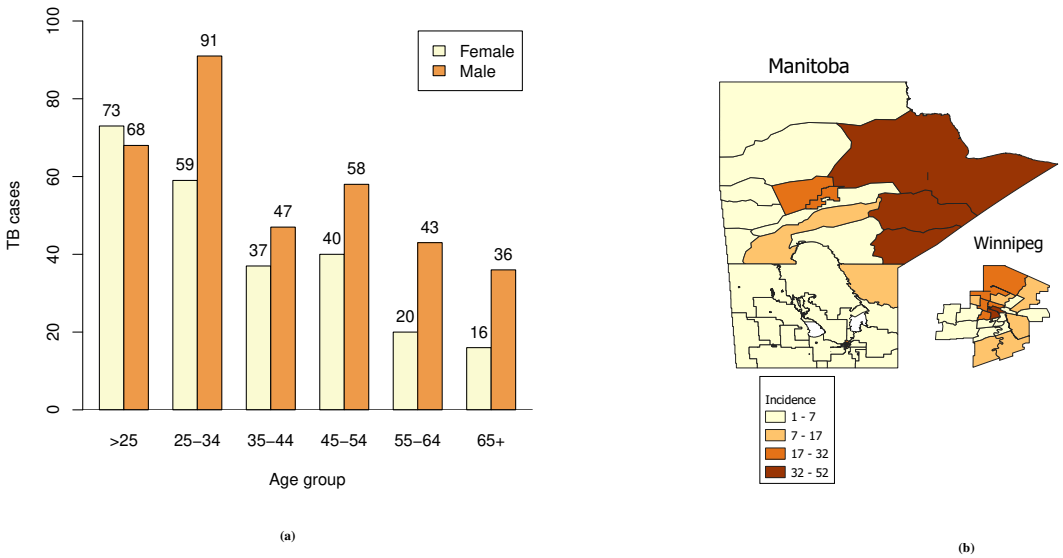
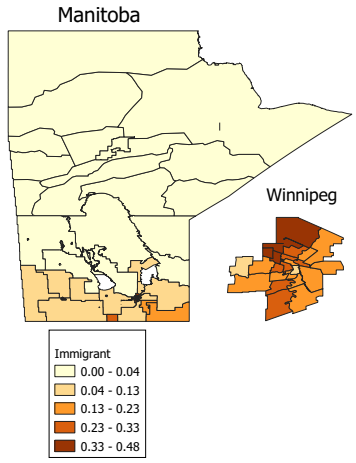
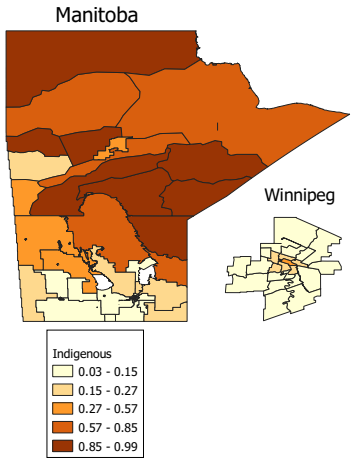


FIGURE 1 TB incidence cases in Manitoba from January 2017 to October 2018 based on: (a) age and sex, (b) regional health authority districts.

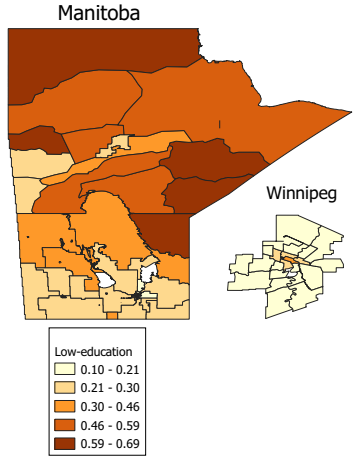
In the following, we also consider some area-level covariates that may contribute to the number of people effected by the TB outbreak. In many studies related to the TB disease, it has been found that socioeconomic status (SES) variables such as income, education and health care access avaliability also have an important impact on the disease incidence and prevalence (Bates et al,[39] Ho,[40] Gupta et al,[41] Hargreaves et al[42]). Here, for each area, we also consider a number of sociodemographic



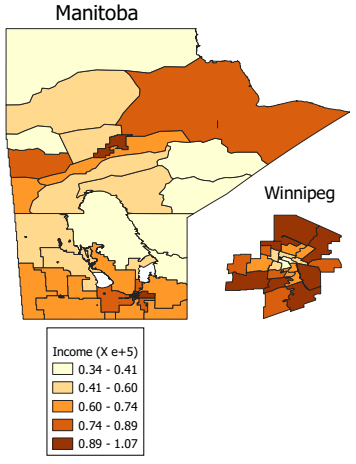
(a) Immigrant



(b) Indigenous



(c) Low-education



(d) Median income (100000$)

FIGURE 2 Geographical distribution of sociodemographic characteristics (shown as percentage of population in the respective areas) based on the 2016 Canadian census data.

characteristics (proportions of indigenous people, immigrants, low-education, and median household income) obtained from the 2016 Canadian census. The proportion of indigenous people is the percentage of the population reporting indigenous status in 2016. The proportion of immigrants is the percentage of the population reporting in 2016 that they immigrated to Manitoba from outside Canada since 1961. The low-education rate is the proportion of persons 15+ who have not graduated high school. So, we consider these four sociodemographic factors as area-level covariates in Model 1 and Model 2 which are defined in (3) and (4), respectively. The sociodemographic characteristics of the 59 areas in Manitoba based on 2016 Canadian Census data, are shown in Figure 2.

The first case in 2017 was assessed on January 03. Subsequent weekly assessments took place from January 03, 2017 to October 16, 2018. Therefore for our discrete-time model framework, January 03, 2017 was set as $t = 1$, January 10, 2017 was set as $t = 2$, up to the last observation on October 16, 2018 as $t = 94$. Thus, each time increment represents 7 days. Studies have reported that TB patients are no longer infectious after 2 weeks of treatment (Riley et al,[43] Rouillon et al[44]). So, we set the infected period to be $\gamma_I = 2$ (i.e., 14 days). At each area and at a given time, if an individual is diagnosed as TB patient we consider he/she as an infectious individual and also susceptible individuals are defined as individuals which have not infected until that given time.

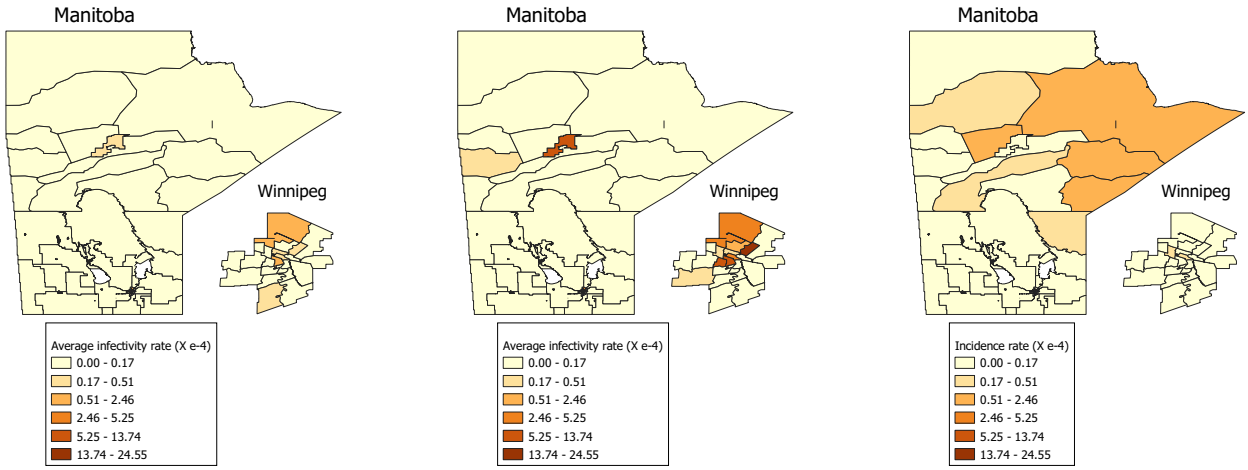
Table 1 shows the estimate of model parameters for Model 1 and Model 2 along with their corresponding SEs. The estimated coefficient in Table 1 for sex shows a significant relationship between sex and TB infection. Also, as expected, areas with an increasing proportion of indigenous and immigrant rates have higher incidence of TB. On the other hand, areas with higher

income are less at risk of TB. Further, according to the literature (e.g. see Wilson and Macdonald,[45] Trovato et al[46]), indigenous people tend to have lower-incomes than the average Canadian. Thus, we consider the interaction between indigenous and income in these models. For both models, the estimated interaction coefficient shows that the indigenous people with a higher income tend to be less at risk from TB. Note that, as can be seen from Figures 1b and 2c, we could not find a direct relationship between low-education and TB incidence.

TABLE 1 Model parameters estimate (Est.), their standard errors (S.E.) and cAIC values for Model 1 and Model 2; TB data in Manitoba, Canada, from January 2017 to October 2018.

	Model 1		Model 2	
Parameters	Est.	S.E.	Est.	S.E.
Intercept	-2.987	0.649	-2.732	0.008
Sex				
Male	Ref.		Ref.	
Female	-1.272	0.413	-2.094	0.009
Age group				
> 25	-2.002	0.598	2.349	0.021
$25-34$	Ref.		Ref.	
$35-44$	-1.982	0.686	-9.364	0.009
$45-54$	-2.932	0.650	-2.798	0.084
$55-64$	-2.957	0.737	-2.796	0.202
$<= 65$	0.184	0.402	-3.870	0.328
Indigenous	11.97	1.045	6.202	0.016
Income	-0.505	0.735	-1.113	0.010
Low-education	-18.51	0.544	-10.35	0.017
Immigrant	1.556	0.617	0.091	0.020
Indigenous× Income	-5.334	1.271	-1.278	0.004
δ	2.073	0.172	2.397	0.007
τ	0.585	0.077	0.674	0.090
λ	0.751	0.240	0.691	0.241
cAIC	20436.84		18717.12	

For each model, we also predict the average infectivity rates using the estimated parameters. The average infectivity rates for the 59 LGAs in Manitoba over the whole time interval (94 time points) are displayed in Figures 3a and 3b. These maps show that Winnipeg, which is the largest city in the province, has higher average infectivity rates than the other areas within the province of Manitoba. In the following, to assess which model has a better fit to our TB data, we use conditional Akaike information criterion (cAIC) of Donohue[47]. The cAIC is defined as $\text{cAIC} = -2l(\mathbf{y}|\hat{\boldsymbol{\Theta}}) + 2\rho$ where $l(\mathbf{y}|\hat{\boldsymbol{\Theta}})$ is the maximum of log-likelihood of observed real data set ($\mathbf{y}$), $\hat{\boldsymbol{\Theta}}$ denotes the real data parameter estimations and ρ is correction factor that can be calculated using bootstrap method as $\rho = \text{E}^*\{l(\mathbf{y}^*|\hat{\boldsymbol{\Theta}}^*) - l(\mathbf{y}|\hat{\boldsymbol{\Theta}}^*)\}$ where E^* denotes the bootstrap expectation, $\mathbf{y}^*$s are bootstrap data sets and $\hat{\boldsymbol{\Theta}}^*$s are the estimated values of parameters using bootstrap samples. The model with smaller cAIC is preferred. Table 1 shows the results of cAIC for both models based on the 100 bootstrap samples. As we can see from this table, Model 2 fits better than Model 1 to our data based on the cAIC. So, we can conclude that since average infectivity rates for Model 2 are higher than the Model 1, each area with TB patients can transmit disease to their neighborhoods. Further, we make a comparison of the efficiency of our proposed models (Model 1 and Model 2) to the naive model in which average incidence rates in each area are defined as $\eta_z = \frac{n_z}{T \times \text{Pop}_z}$, $z = 1, \ldots, 59$, where n_z denotes the number of TB incidence in zth area, T is the whole time interval and Pop_z indicates the number of population at the zth area based on the census 2016. Average incidence rates over time for naive model is plotted in Figure 3c. As it can be seen from Figures 3a, 3b and 3c, the infectivity rates using our proposed approach are different from the naive model as the Models 1 and 2 smooth the incidence rates and in particular areas with low number of cases. We further investigate the performance of the proposed approach in the simulation study.



(a) Model 1 **(b)** Model 2 **(c)** Naive model

FIGURE 3 Predicted average rate of infectivity of Model 1 and Model 2 as well average of incidence rate of infectivity for naive model for TB data in Manitoba, Canada, from January 2017 to October 2018.

5 | SIMULATION STUDY

In this section, we consider two distinct simulation scenarios. One has an irregular, and one a regular, areal structures. Under the irregular scenario, epidemic data are generated based on the Winnipeg 12 local geographic areas (LGAs). An 8×8 grid is used to define areas under the regular scenario. Simulation studies are conducted to evaluate performance of both the region restricted Model (Model 1) and neighbourhood Model (Model 2) in terms of how well parameters are estimated, for the purpose of identifying infectious disease dynamics.

5.1 | Irregular grid

We first consider epidemic simulation within the 12 LGAs of Winnipeg. For each simulation, we simulate the locations of a total of $n = 230$ individuals where the sample sizes for each of 12 LGAs are shown in Table 2. The location (i.e. latitude and longitude) for each individual within each LGA is randomly drawn using a Generalized Random Tesselation Stratified (GRTS) spatial sampling technique. The GRTS is implemented here using the package **spsurvey** in R. Figure 4 illustrates the geographical location of the one generated sample of the individuals within each LGA.

TABLE 2 The number of sample at each LGA.

Area	Sample	Area	Sample
St. James Assiniboia	25	River East	25
Assiniboine South	25	Seven Oaks	30
Ft. Garry	25	Inkster	10
St. Vital	25	Point Douglas	10
St. Boniface	20	Downtown	10
Transcona	15	River Heights	10

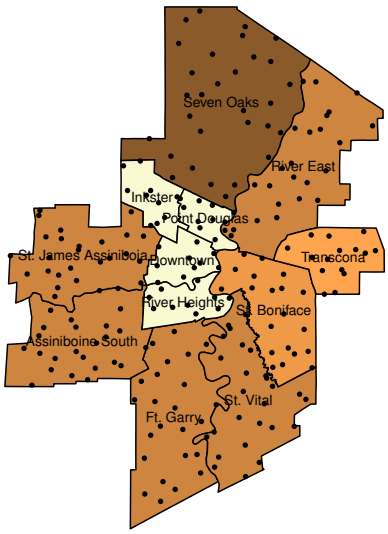


FIGURE 4 Geographical location of each sampled individual in 12 LGAs of Winnipeg, Manitoba, Canada.

For each model, 1000 epidemic data sets are generated. To start each simulated epidemic, one individual is randomly selected as infectious in each LGA at $t = 1$. Here we set the maximum observed time point to be $t_{max} = 20$ and we assume that the length of infectious period (γ_I) is constant for all individuals and each individual is set to be infectious for 3 time units before moving to the removed stage. The spatial random effects are generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}_u = \left[\tau^2\Big((1-\lambda)\mathbf{I} + \lambda\mathbf{D}\Big)\right]^{-1}$. Three different values are considered for τ, $\tau \in \{0.10, 0.50, 0.90\}$. Three different values of the spatial dependence parameters, $\lambda \in \{0.20, 0.50, 0.80\}$, are also considered in order to represent different strength of spatial correlation. In this simulation study, we consider the following two models with one area level covariate and one individual level covariate

$$\textbf{Model 1:}\quad P(i,z,t) = 1 - \exp\Big(-\exp(\alpha + \beta_{11}X_{i1} + \beta_{21}X_{z1} + u_z)\sum_{j\in I(t,z)} d_{ij}^{-\delta}\Big),$$

and

$$\textbf{Model 2:}\quad P(i,z,t) = 1 - \exp\Big(-\exp(\alpha + \beta_{11}X_{i1} + \beta_{21}X_{z1} + u_z)\sum_{j\in I(t,z,\xi(z))} d_{ij}^{-\delta}\Big).$$

We use the uniform distribution between $(0, 1)$ for generating X_{i1} and X_{z1} and set their corresponding coefficient values equal to one ($\beta_{11} = \beta_{21} = 1$). The intercept is set to $\alpha = 0.40$. Also, three different values for transmission parameter $\delta \in \{2, 2.50, 3\}$ are considered. d_{ij} is calculated using the great-circle distance matrix method, and the *rdist.earth* function in the R package **spectralGP** is used in this regard.

Tables 3 to 5 display the mean of the estimated coefficients for the intercept, individual and area level covariates, the transmission parameter (δ) and LCAR parameters (τ and λ) corresponding to Model 1, along with their standard errors (SEs) and coverage rate (CR) for the 95% confidence intervals of the model parameters estimate. The SEs are calculated using the inverse Fisher information matrix of the complete data using 1000 simulation runs. As it can be seen from these Tables, in general, the model parameters estimates are unbiased for the Model 1. However, we have a slight overestimation for τ, and in the case of $\lambda = 0.5$, the estimated parameters are better than the other scenarios. It appears that we only have under coverage in the case of estimate of $\tau = 0.10$, however the CRs get closer to the nominal value of 0.95 for the τ values of 0.50 and 0.90. Similar results for Model 2 are observed which are reported in Tables B1 to B3 in Appendix B. For two Models 1 and 2, in order to determine the areas with higher infectivity rates, we obtain the infectivity rates of infectious disease over time at each area according to

(5). Among all combinations of aforementioned parameters, we only report the infectivity rate for the GD-ILMs with $\delta = 2.50$, $\tau = 0.90$ and $\lambda = 0.50$. The infectivity rate for other scenarios can be obtained similarly. Table 6 presents the mean of the average infectivity rate for each of 12 LGAs at time points 2 and 3 for this simulation study. Furthermore, the maps of infectivity rates for Model 1 and Model 2 at two time points 2 and 3 are illustrated in Figures 5 and 6. As it can be seen from these figures, Model 2 indicates a higher infectivity rate than the Model 1 at each time point. Also, for the both models, at the beginning of the outbreak, some areas (Downtown, Inkster, Point Douglas and River Heights) have higher infectivity rates and disease have been transmitted between neighboring areas over time. Such maps can be used to target high risk areas for infectious disease control.

TABLE 3 True value of parameters along with the average parameter estimates (Av.Est) and average standard errors of the estimated parameters (Av.S.E.) across 1000 simulation runs for Model 1 in the case of irregular grid with different δ and λ but $\tau = 0.10$.

	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.00			0.10					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.365	0.274	1.00	1.011	0.211	0.99	0.939	0.355	0.99	2.022	0.109	0.99	0.176	0.054	0.90	0.417	0.357	0.90
0.50	0.351	0.214	1.00	0.976	0.173	0.99	0.918	0.286	1.00	2.027	0.086	0.98	0.180	0.049	0.78	0.544	0.103	0.87
0.80	0.356	0.115	1.00	0.978	0.115	0.99	0.930	0.216	1.00	2.021	0.057	0.99	0.188	0.048	0.64	0.635	0.082	0.90
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.50			0.10					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.367	0.185	1.00	0.975	0.153	0.99	0.942	0.249	1.00	2.516	0.077	0.99	0.168	0.051	0.91	0.410	0.123	0.89
0.50	0.357	0.157	1.00	0.987	0.116	0.99	0.931	0.218	1.00	2.519	0.057	0.99	0.177	0.049	0.82	0.527	0.101	0.88
0.80	0.357	0.134	1.00	0.943	0.111	0.99	0.928	0.180	1.00	2.514	0.057	0.99	0.183	0.047	0.70	0.623	0.084	0.90
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			3.00			0.10					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.372	0.181	1.00	0.976	0.148	0.99	0.947	0.237	1.00	3.003	0.075	0.99	0.162	0.049	0.94	0.407	0.122	0.90
0.50	0.375	0.128	1.00	0.978	0.110	0.99	0.952	0.172	1.00	2.003	0.172	0.99	0.162	0.048	0.93	0.420	0.121	0.90
0.80	0.365	0.124	1.00	0.986	0.105	0.99	0.944	0.169	1.00	3.007	0.053	0.99	0.178	0.046	0.76	0.610	0.086	0.88

TABLE 4 True value of parameters along with the average parameter estimates (Av.Est), average standard errors of the estimated parameters (Av.S.E.), and coverage rate (CR) for 95% confidence interval of the parameters estimate across 1000 simulation runs for Model 1 in the case of irregular grid with different δ and λ but $\tau = 0.50$.

	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.00			0.50					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.382	0.062	1.00	0.964	0.065	0.97	0.978	0.076	0.99	1.993	0.034	0.99	0.561	0.034	0.99	0.371	0.070	0.93
0.50	0.388	0.058	0.99	0.962	0.060	0.98	0.985	0.070	0.99	1.991	0.033	0.99	0.633	0.034	0.96	0.514	0.064	0.92
0.80	0.384	0.062	1.00	0.976	0.062	0.98	0.976	0.073	0.99	1.991	0.035	0.99	0.679	0.034	0.91	0.642	0.059	0.93
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.50			0.50					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.389	0.063	1.00	0.982	0.067	0.98	0.977	0.077	0.99	2.493	0.037	0.98	0.571	0.037	0.98	0.349	0.071	0.95
0.50	0.389	0.058	1.00	0.978	0.061	0.98	0.991	0.071	0.99	2.494	0.035	0.98	0.639	0.034	0.96	0.531	0.064	0.93
0.80	0.385	0.060	1.00	0.980	0.062	0.99	0.977	0.072	0.99	2.488	0.036	0.98	0.669	0.034	0.91	0.648	0.058	0.93
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			3.00			0.50					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.389	0.063	1.00	0.985	0.068	0.98	0.981	0.079	0.99	2.989	0.040	0.99	0.569	0.036	0.98	0.373	0.070	0.95
0.50	0.392	0.059	1.00	0.995	0.063	0.98	0.985	0.073	0.99	2.992	0.038	0.98	0.648	0.035	0.94	0.533	0.065	0.93
0.80	0.388	0.061	1.00	0.984	0.063	0.98	0.982	0.074	0.99	2.983	0.039	0.98	0.681	0.034	0.90	0.654	0.058	0.93

5.2 | Regular grid

In this part of simulation study, for each simulation, the epidemic is generated across a population of 640 individuals located within landscape in which 64 areas are defined by an 8×8 lattice (10 individuals in each cell). The latitude (x) and longitude

TABLE 5 True value of parameters along with the average parameter estimates (Av.Est), average standard errors of the estimated parameters (Av.S.E.), and coverage rate (CR) for 95% confidence interval of the parameters estimate across 1000 simulation runs for Model 1 in the case of irregular grid with different δ and λ but $\tau = 0.90$.

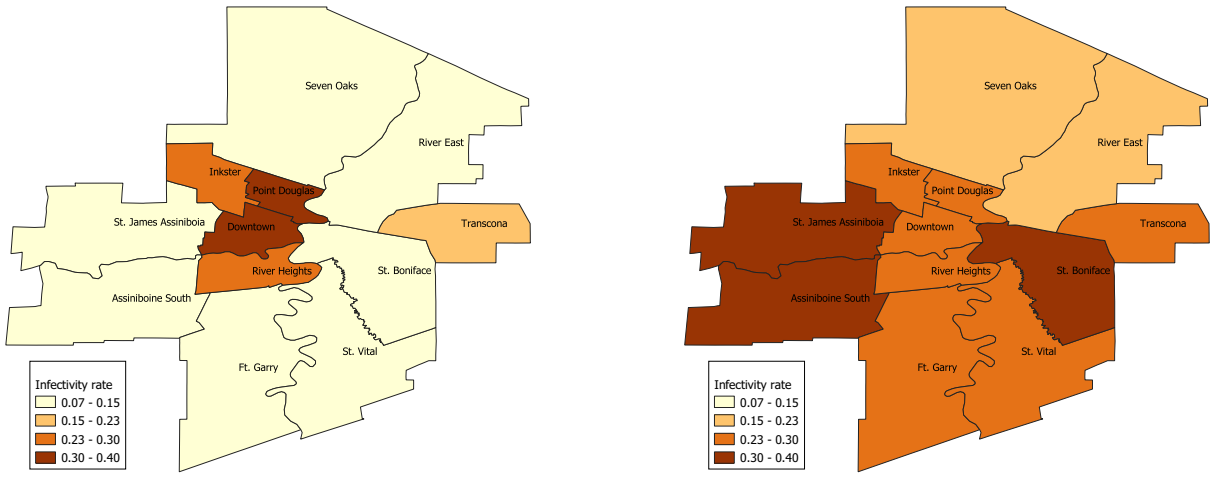
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.00			0.90					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.392	0.072	1.00	0.984	0.074	0.99	0.997	0.089	0.98	1.991	0.043	0.98	1.023	0.073	0.97	0.406	0.081	0.91
0.50	0.392	0.072	1.00	0.995	0.073	0.98	0.981	0.087	0.99	1.992	0.043	0.99	1.122	0.072	0.93	0.582	0.076	0.90
0.80	0.388	0.055	1.00	0.991	0.057	0.98	0.978	0.066	0.98	1.980	0.033	0.98	1.200	0.055	0.90	0.704	0.053	0.94
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.50			0.90					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.387	0.047	1.00	0.998	0.051	0.98	0.985	0.058	0.99	2.485	0.030	0.99	1.033	0.052	0.96	0.402	0.057	0.91
0.50	0.390	0.047	1.00	0.988	0.050	0.97	0.982	0.058	0.99	2.492	0.030	0.99	1.144	0.051	0.93	0.586	0.053	0.90
0.80	0.391	0.052	1.00	0.986	0.056	0.98	0.990	0.064	0.98	2.487	0.034	0.98	1.198	0.054	0.90	0.697	0.052	0.93
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			3.00			0.90					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.396	0.047	1.00	0.991	0.051	0.98	0.987	0.058	0.98	2.987	0.033	0.98	1.048	0.053	0.95	0.415	0.057	0.90
0.50	0.398	0.048	1.00	0.985	0.052	0.98	0.995	0.060	0.99	2.986	0.033	0.99	1.163	0.055	0.94	0.563	0.056	0.90
0.80	0.388	0.050	1.00	0.992	0.055	0.98	0.979	0.063	0.98	2.982	0.035	0.98	1.201	0.054	0.90	0.690	0.051	0.93

TABLE 6 The average rate of infectivity for Model 1 and Model 2 in the case of irregular grid with $\delta = 2.50$, $\tau = 0.90$ and $\lambda = 0.50$.

	Model 1		Model 2	
Area	Time 2	Time 3	Time 2	Time 3
St. James Assiniboia	0.135	0.304	0.254	0.380
Assiniboine South	0.136	0.306	0.222	0.386
Ft. Garry	0.103	0.235	0.179	0.352
St. Vital	0.117	0.262	0.205	0.377
St. Boniface	0.141	0.301	0.279	0.400
Transcona	0.214	0.274	0.286	0.312
River East	0.097	0.229	0.182	0.373
Seven Oaks	0.072	0.199	0.132	0.310
Inkster	0.290	0.270	0.414	0.278
Point Douglas	0.392	0.280	0.527	0.234
Downtown	0.320	0.294	0.532	0.226
River Heights	0.263	0.283	0.451	0.279

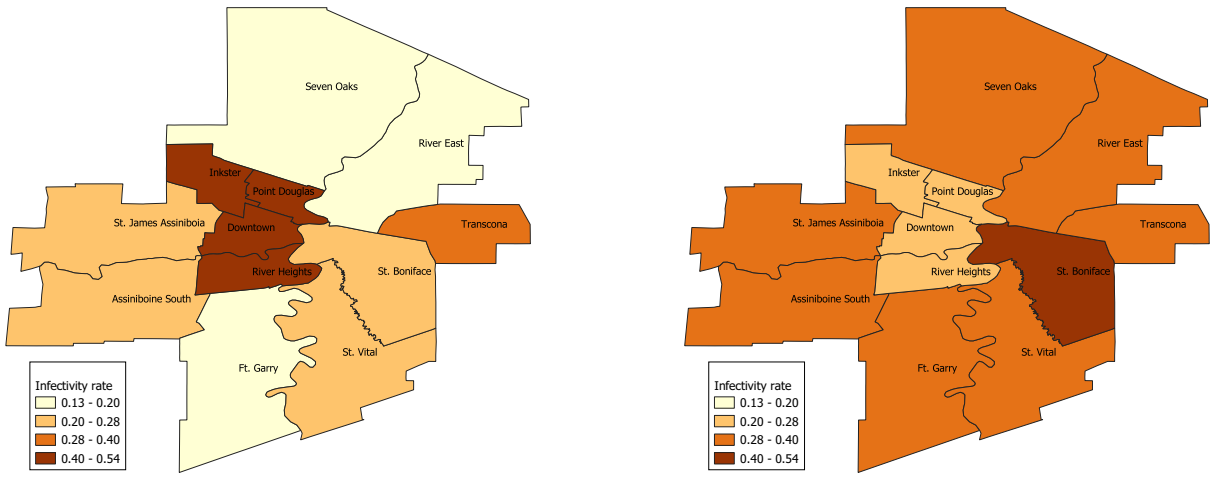
(y) coordinates of the areas are given by all combinations (x, y) for $x, y = 3, 6, 9, \ldots, 24$. One generated sample is ploted in Figure 7a. For each model, 1000 data sets are generated. To start each simulated epidemic, one individual is randomly made infectious in each cell at $t = 1$. Each epidemic is simulated over an epidemic length of $t_{max} = 20$ time units and parameters are set to be $\alpha = 0.40$, $\delta \in \{2, 3\}$, $\tau = 0.50$ and $\lambda \in \{0.20, 0.50, 0.80\}$. Also, a fixed infectious period of $\gamma_I = 3$ is assumed. Moreover, in the similar manner as simulation based on the irregular grid, we use the uniform distribution between $(0, 1)$ for generating one individual level covariate and one areal level covariate and set their corresponding coefficient values equal to one ($\beta_{11} = \beta_{21} = 1$). d_{ij} is calculated using Euclidean distance between the infectious individuals j and susceptible individual i. The following models are fitted via the ECM algorithm presented in Section 3

$$\textbf{Model 1:}\quad P(i, z, t) = 1 - \exp\Big(- \exp(\alpha + \beta_{11} X_{i1} + \beta_{21} X_{z1} + u_z) \sum_{j \in I(t,z)} d_{ij}^{-\delta} \Big),$$



(a) Time 2 **(b)** Time 3

FIGURE 5 Average rate of infectivity for Model 1 in the case of $\delta = 2.50$, $\tau = 0.90$, $\lambda = 0.50$, for two time points across 12 LGAs of Winnipeg, Manitoba, Canada.



(a) Time 2 **(b)** Time 3

FIGURE 6 Average rate of infectivity for Model 2 in the case of $\delta = 2.50$, $\tau = 0.90$, $\lambda = 0.50$, for two time points across 12 LGAs of Winnipeg, Manitoba, Canada.

and

$$\textbf{Model 2:}\quad P(i,z,t) = 1 - \exp\Big(- \exp(\alpha + \beta_{11} X_{i1} + \beta_{21} X_{z1} + u_z) \sum_{j \in \boldsymbol{I}(t,z,\xi(z))} d_{ij}^{-\delta} \Big).$$

The simulation results of the Model 1 and Model 2 are shown in Table 7. This table report the mean of the estimated parameters across the 1000 simulation runs with corresponding standard deviations and CR for the 95% confidence intervals of the model

TABLE 7 True value of parameters along with the average parameter estimates (Av.Est), average standard errors of the estimated parameters (Av.S.E.), and coverage rate (CR) for 95% confidence interval of the parameters estimate across 1000 simulation runs for Model 1 and Model 2 in the case of regular grid with different δ and λ but $\tau = 0.50$.

		$\delta = 2$						$\delta = 3$					
		Model 1			Model 2			Model 1			Model 2		
Par.	True	Est.	S.E.	C.R.	Est.	S.E.	C.R.	Est.	S.E.	C.R.	Est.	S.E.	C.R.
α	0.40	0.359	0.180	0.97	0.360	0.227	0.98	0.353	0.183	0.98	0.359	0.204	0.97
β_{11}	1.00	0.967	0.234	0.97	0.992	0.277	0.99	1.001	0.235	0.99	0.982	0.260	0.98
β_{21}	1.00	0.962	0.246	0.98	1.016	0.294	0.99	0.982	0.246	0.99	0.965	0.274	0.98
δ		1.940	0.148	0.95	1.983	0.190	0.98	2.952	0.169	0.98	2.951	0.187	0.97
τ	0.50	0.510	0.005	0.96	0.511	0.070	0.95	0.509	0.005	0.94	0.506	0.070	0.95
λ	0.20	0.251	0.141	0.98	0.256	0.141	0.97	0.252	0.141	0.97	0.261	0.141	0.98
α	0.40	0.329	0.178	0.95	0.390	0.228	0.98	0.357	0.183	0.97	0.329	0.209	0.98
β_{11}	1.00	0.964	0.232	0.95	0.991	0.281	0.98	0.980	0.234	0.98	1.000	0.261	0.98
β_{21}	1.00	0.963	0.242	0.96	0.981	0.297	0.99	0.967	0.244	0.98	1.004	0.275	0.98
δ		1.902	0.152	0.94	1.976	0.201	0.97	2.917	0.175	0.97	2.947	0.199	0.98
τ	0.50	0.530	0.066	0.97	0.529	0.063	0.96	0.527	0.063	0.97	0.531	0.063	0.98
λ	0.50	0.510	0.205	0.90	0.526	0.202	0.90	0.515	0.202	0.90	0.512	0.204	0.91
α	0.40	0.367	0.180	0.96	0.369	0.229	0.98	0.326	0.183	0.94	0.373	0.212	0.97
β_{11}	1.00	0.991	0.233	0.97	0.977	0.280	0.97	0.965	0.232	0.96	0.976	0.262	0.97
β_{21}	1.00	0.970	0.244	0.97	0.957	0.300	0.97	0.969	0.243	0.97	0.973	0.276	0.97
δ		1.926	0.154	0.95	1.911	0.203	0.92	2.825	0.176	0.93	2.930	0.203	0.97
τ	0.50	0.539	0.057	0.96	0.544	0.057	0.94	0.544	0.058	0.93	0.541	0.057	0.94
λ	0.80	0.716	0.174	0.90	0.729	0.171	0.90	0.717	0.173	0.90	0.727	0.173	0.90

parameters estimate. Under the both models, the estimated parameters are generally unbiased. Similar to the irregular grid subsection of simulation study, we can also provide infectivity rate for each cell for different time points.

In the following, we repeat the above simulation set-up to evaluate the edge effects of infected people near the boundaries and compare the efficiency of the both Model 1 and Model 2 in terms of their accuracy in estimating model parameters. To that end, we generated the individual locations near the boundaries. One generated sample is displayed in Figure 7b. The simulation results of the Model 1 and Model 2 are shown in Table 8. This table reports the mean of the estimated parameters across the 1000 simulation runs with corresponding standard deviations and CR for the 95% confidence intervals of the model parameters estimate. As expected, we can see from Table 8 that in the all six simulation scenarios, the covariates coefficients and transmission parameters have less biases and mean squared errors (defined as square of bias plus variance) under the Model 2 in comparison with the Model 1. It is worth mentioning that, although our proposed models are complex, one of the practical desirable features of the ECM algorithm in estimating of parameters is the computation time. In our simulation studies, the average running time for each simulation on system equipped with a 2.3 GHz Intel Core i9 processor and 16 GB of memory is 8 minutes.

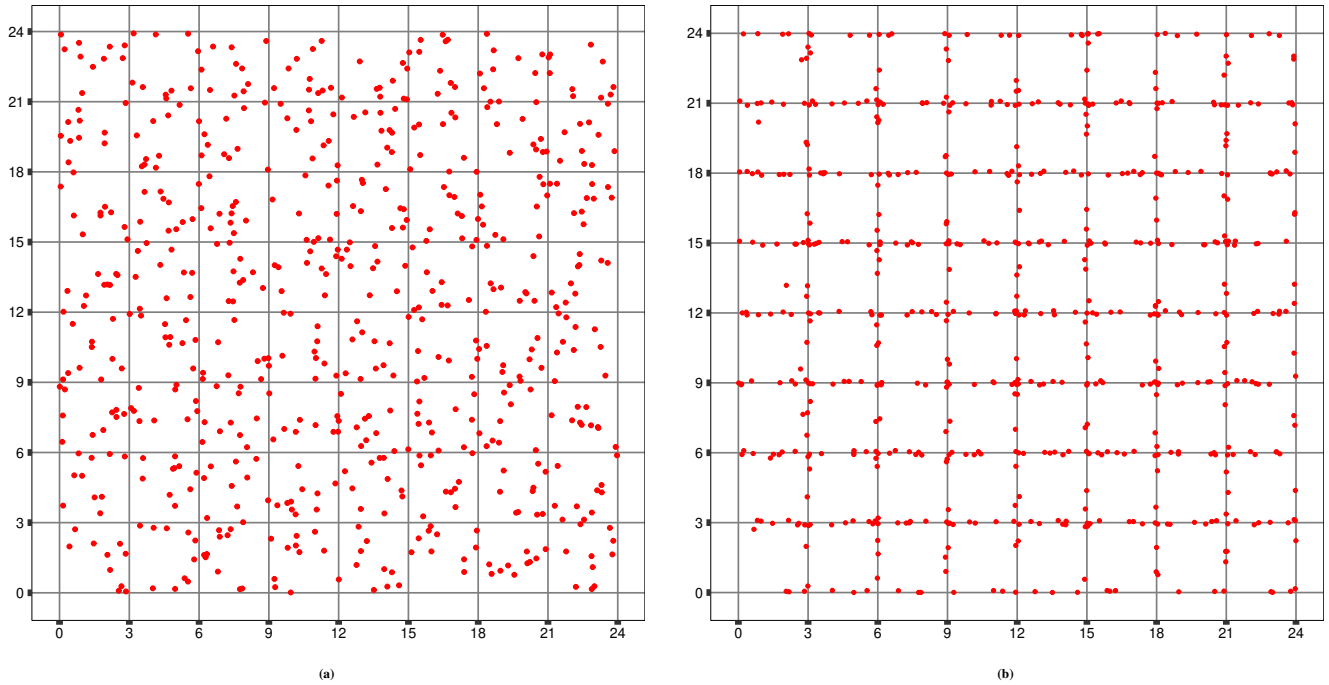


FIGURE 7 Geographical location of each sampled individual in 8 by 8 grid based on: (a) random generation, (b) edge generation.

6 | DISCUSSION

GD-ILMs are an extension of the ILMs of Deardon et al [12] that not only allow risk of infection to depend on the distance between susceptible and infectious individuals, they also allow us to consider the effect of the spatial locations of individuals. We have considered two GD-ILMs in terms of their ability to ascertain infectious disease dynamics: region-restricted (Model 1) and neighbourhood-restricted (Model 2). In this paper, we have applied the ECM algorithm for computing the maximum likelihood estimate of model parameters.

As a motivation of this paper, we have fitted both models to data that collected by Manitoba Health from January 2017 to October 2018, on patients diagnosed with TB, in Manitoba, Canada. Our models included two individual-level covariates (age and sex) and four area-level covariates (proportions of Indigenous people, immigrants, low-education, and median household income). The results have shown that TB infection risk is significantly higher among men than women. In terms of age groups, age-specific risks are highest in persons older than 65 years and younger than 25 years under Model 1 and Model 2, respectively. Also, for both models, persons in age-groups 35-44, 45-54 and 55-64 are less at risk compared to the age-group 25-34. The risk of TB transmission also increases in areas with higher proportions of Indigenous and immigrants. In addition, individuals in areas with lower income are more at risk of TB. Moreover, values of spatial dependency parameter for Model 1 and Model 2 are 0.751 and 0.691, respectively, that shows significant and strong spatial correlation between areas. Predicted average infectivity rate shows that Winnipeg, a city of consisting 23 LGAs, is more at risk of TB than other parts of the Manitoba. This information about the geographical variation of TB will be communicated to policy makers at Manitoba Health for possible preventions in those areas which are most at risk. Via simulation studies, we have evaluated the accuracy of GD-ILMs parameters estimate, and shown how they can be used for predicting infectivity rates.

It is of interest to also study other diseases such as influenza which has similar transmission dynamics as TB. One can also expend our proposed model to study SEIR (susceptible-exposed-infected-removed) or SEIRS (susceptible-exposed-infected-removed-susceptible) that allow us to consider an infectious disease with a different transmission process. These are some of the topics for future study.

TABLE 8 True value of parameters along with the average parameter estimates (Av.Est), average standard errors of the estimated parameters (Av.S.E.), and coverage rate (CR) for 95% confidence interval of the parameters estimate across 1000 simulation runs for Model 1 and Model 2 in the case of regular grid for considering edge effect with different δ and λ but $\tau = 0.5$.

		$\delta = 2$						$\delta = 3$					
		Model 1			Model 2			Model 1			Model 2		
Par.	True	Est.	S.E.	C.R.	Est.	S.E.	C.R.	Est.	S.E.	C.R.	Est.	S.E.	C.R.
α	0.40	0.324	0.181	0.94	0.342	0.226	0.96	0.323	0.191	0.97	0.379	0.222	0.99
β_{11}	1.00	0.937	0.219	0.94	0.969	0.284	0.97	0.985	0.225	0.98	0.983	0.276	0.99
β_{21}	1.00	0.936	0.227	0.95	0.973	0.301	0.97	0.972	0.233	0.98	0.980	0.291	0.99
δ		1.935	0.124	0.94	1.952	0.150	0.96	2.958	0.146	0.97	2.977	0.164	0.99
τ	0.50	0.509	0.005	0.95	0.511	0.005	0.95	0.506	0.005	0.95	0.508	0.005	0.95
λ	0.20	0.257	0.030	0.98	0.256	0.030	0.98	0.261	0.031	0.98	0.259	0.031	0.97
α	0.40	0.300	0.179	0.91	0.329	0.231	0.95	0.311	0.188	0.96	0.365	0.223	0.98
β_{11}	1.00	0.936	0.214	0.93	0.956	0.283	0.96	0.973	0.220	0.97	0.994	0.272	0.99
β_{21}	1.00	0.932	0.222	0.93	0.963	0.299	0.96	0.964	0.227	0.97	0.981	0.286	0.98
δ		1.893	0.126	0.91	1.935	0.161	0.95	2.918	0.146	0.96	2.960	0.171	0.98
τ	0.50	0.531	0.004	0.97	0.532	0.004	0.97	0.527	0.004	0.97	0.530	0.004	0.97
λ	0.50	0.511	0.041	0.90	0.520	0.041	0.91	0.517	0.041	0.90	0.517	0.042	0.91
α	0.40	0.306	0.179	0.91	0.354	0.233	0.95	0.254	0.186	0.93	0.349	0.223	0.97
β_{11}	1.00	0.928	0.213	0.91	0.951	0.283	0.95	0.965	0.219	0.94	0.970	0.271	0.97
β_{21}	1.00	0.931	0.221	0.92	0.957	0.299	0.95	0.953	0.226	0.94	0.962	0.285	0.98
δ		1.863	0.126	0.90	1.909	0.163	0.92	2.844	0.144	0.93	2.933	0.173	0.97
τ	0.50	0.541	0.003	0.95	0.544	0.003	0.95	0.533	0.003	0.96	0.537	0.003	0.96
λ	0.80	0.721	0.030	0.91	0.728	0.029	0.91	0.747	0.030	0.96	0.750	0.030	0.95

ACKNOWLEDGEMENTS

We would like to thank the Associate Editor and a referee for constructive comments and suggestions, which led to an improved version of the manuscript. This work was supported by grants from the Canadian Statistical Sciences Institute, Collaborative Research Team (CANSSI-CRT) entitled “Spatial modeling of infectious diseases: environment and health", and Natural Sciences and Engineering Research Council of Canada (NSERC). Disclaimer: This study is based in part on data provided by Manitoba Health. The interpretation and conclusions contained herein are those of the researchers and do not necessarily represent the views of the government of Manitoba.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article. The supplementary materials contain R codes and corresponding “readme” files for the simulations and real data application conducted in this paper.

References

1. Kristoffersen AB, Rees EE, Stryhn H, Ibarra R, Campisto JL, Revie CW, St-Hilaire S. Understanding sources of sea lice for salmon farms in Chile. *Prev. Vet. Med.* 2013;111:165–175.

2. Rees E E, St-Hilaire S, Jones SR, Krkošek M, DeDominicis S, Foreman MGG, Patanasatienkul T, Revie CW. Spatial patterns of parasite infection among wild and captive salmon in western Canada. *Landsc Ecol.* 2015; 989–1004.

3. Lienhardt C. From exposure to disease: the role of environmental factors in susceptibility to and development of tuberculosis. *Epidemiol Rev.* 2001;23(2):288–301.

4. Laumbach RJ, Kipen HM. Respiratory health effects of air pollution: update on biomass smoke and traffic pollution. *J. Allergy Clin. Immunol.* 2012;129(1):3–11.

5. Fares A. Factors influencing the seasonal patterns of infectious diseases. *Int J Prev Med.* 2013;4(2):128–132.

6. Smith GS, Schoenbach VJ, Richardson DB, Gammon MD. Particulate air pollution and susceptibility to the development of pulmonary tuberculosis disease in North Carolina: an ecological study. *Int J Environ Health Res.* 2014;24(2):103–112.

7. Onozuka D, Hagihara A. The association of extreme temperatures and the incidence of tuberculosis in Japan.*Int J Biometeorol.* 2015;59(8):1107–1114.

8. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character. 1927;115(772):700–721.

9. Keeling MJ, Rohani P. *Modeling infectious diseases in humans and animals.* Princeton University Press; 2011.

10. Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT. Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science.* 2001;294(5543):813–817.

11. Jewell CP, Kypraios T, Neal P, Roberts GO. Bayesian analysis for emerging infectious diseases. *Bayesian Analysis.* 2009;4(3):465–496.

12. Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, Savill NJ, Shaw DJ, Woolhouse ME. Inference for individual-level models of infectious diseases in large populations. *Stat Sin.* 2010;20(1): 239–261.

13. McBryde ES, Gibson G, Pettitt AN, Zhang Y, Zhao B, McElwain DL. Bayesian modelling of an epidemic of severe acute respiratory syndrome. *Bull. Math. Biol.* 2006;68(4):889-917.

14. Chowell G, Hengartner NW, Castillo-Chavez C, Fenimore PW, Hyman JM. The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda. *J. Theor. Biol.* 2004;229(1):119–126.

15. McKinley T, Cook AR, Deardon R. Inference in epidemic models without likelihoods. *Int J Biostat.* 2009;5(1), Article 24.

16. D'Silva JP, Eisenberg MC. Modeling spatial invasion of Ebola in West Africa. *J. Theor. Biol.* 2017;428:65–75.

17. Becker NG, Watson LF, Carlin JB. A method of non-parametric back-projection and its application to AIDS data. *Stat Med.* 1991;10(10):1527–1542.

18. Perelson AS, Nelson PW. Mathematical analysis of HIV-1 dynamics in vivo. *SIAM Rev.* 1999;41(1):3–44.

19. Sweeting MJ, De Angelis D, Aalen OO. Bayesian back-calculation using a multi-state model with application to HIV. *Stat Med.* 2005;24(24):3991–4007.

20. Taffe P, May M, Swiss HIV Cohort Study. A joint back calculation model for the imputation of the date of HIV infection in a prevalent cohort. *Stat Med.* 2008;27(23):4835–4853.

21. Hughes G, McROBERTS NE, Madden LV, Nelson SC. Validating mathematical models of plant-disease progress in space and time. *Math Med Biol.* 1997;14(2):85-112.

22. Ster IC, Ferguson NM. Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PLoS one.* 2007;2(6):e502.

23. Kwong GP, Deardon R. Linearized forms of individual-level models for large-scale spatial infectious disease systems. *Bull. Math. Biol.* 2012;74(8):1912–1937.

24. Brown PE, Chimard F, Remorov A, Rosenthal JS, Wang X. Statistical inference and computational efficiency for spatial infectious disease models with plantation data. *J. Royal Stat. Soc. C-Appl.* 2014;63(3):467–482.

25. Pokharel G, Deardon R. Gaussian process emulators for spatial individual-level models of infectious disease. *Can J Stat.* 2016;44(4):480-501.

26. Mahsin MD, Deardon R, Brown P. Geographically dependent individual-level models for infectious diseases transmission. *Biostatistics*. 2019; doi:10.1093/biostatistics/kxaa009.

27. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* 1993;88(421):9-25.

28. Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In M. E. Halloran & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials.* 1999;179–191. Springer, New York, NY.

29. Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika.* 1993;80(2):267–278.

30. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol.* 1977;39(1):1-38.

31. Lee D. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spat Spatiotemporal Epidemiol.* 2011;2(2):79–89.

32. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics. *Ann Inst Stat Math.* 1991;43(1):1–20.

33. Evans M, Swartz T. *Approximating integrals via Monte Carlo and deterministic methods.* OUP Oxford; 2000.

34. Smith I. Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence. *Clin. Microbiol. Rev.* 2003;16(3):463–96.

35. Daniel TM. The history of tuberculosis: past, present and challenges for the future. *Tuberculosis: a comprehensive clinical reference.* 2009:1–8.

36. Narasimhan P, Wood J, MacIntyre CR, Mathai D. Risk factors for tuberculosis. *Pulm. Med.* 2013;2013:828–939.

37. Horton KC, MacPherson P, Houben RM, White RG, Corbett EL. Sex differences in tuberculosis burden and notifications in low-and middle-income countries: a systematic review and meta-analysis. *PLoS Med.* 2016;13(9):e1002119.

38. LaFreniere M, Hussain H, He N, McGuire M. Tuberculosis in Canada: 2017. Can Commun Dis Rep 2019, 45(2/3), 68–74. Doi: 10.14745/ccdr.v45i23a04

39. Bates I, Fenton C, Gruber J, Lalloo D, Lara AM, Squire SB, Theobald S, Thomson R, Tolhurst R. Vulnerability to malaria, tuberculosis, and HIV/AIDS infection and disease. Part II: determinants operating at environmental and institutional level. *Lancet Infect Dis.* 2004;4(6):368–375.

40. Ho MJ. Sociocultural aspects of tuberculosis: a literature review and a case study of immigrant tuberculosis. *Soc Sci Med.* 2004;59(4):753–762.

41. Gupta D, Das K, Balamughesh T, Aggarwal N, Jindal SK. Role of socio-economic factors in tuberculosis prevalence. *Indian J Tuberc.* 2004;51(1):27–32.

42. Hargreaves JR, Boccia D, Evans CA, Adato M, Petticrew M, Porter JD. The social determinants of tuberculosis: from evidence to action. *Am J Public Health.* 2011;101(4):654–662.

43. Riley RL, Mills CC, O'grady F, Sultan LU, Wittstadt F, Shivpuri DN. Infectiousness of air from a tuberculosis ward: ultraviolet irradiation of infected air: comparative infectiousness of different patients. *Am. Rev. Respir. Dis.* 1962;85(4):511–525.

44. Rouillon A, Perdrizet S, Parrot R. Transmission of tubercle bacilli: the effects of chemotherapy. *Tubercle.* 1976;57(4):275-99.

45. Wilson D, Macdonald D. *The income gap between Aboriginal peoples and the rest of Canada.* Ottawa: Canadian Centre for Policy Alternatives; 2010.

46. Trovato F, Pedersen AM, Price JA, Lang C. Economic conditions of indigenous peoples in Canada. *The Canadian Encyclopedia, Historica Canada.* 2019.

47. Donohue MC, Overholser R, Xu R, Vaida F. Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika.* 2011;98(3):685–700.

APPENDIX

A VARIANCE-COVARIANCE OF MODEL PARAMETERS ESTIMATE

The asymptotic covariance matrix of the ML estimator can be approximated by the inverse of the complete Fisher information matrix $\mathrm{I}_{com}(\hat{\boldsymbol{\Theta}};\mathbf{y}_c)$ that can be defined as

$$\mathrm{I}_{com}(\hat{\boldsymbol{\Theta}};\mathbf{y}_c) = -E\Big[\frac{\partial^2}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top} L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}}.$$

In our model, the elements of $\mathrm{I}_{com}(\hat{\boldsymbol{\Theta}};\mathbf{y})$ are computed as follows

$$\begin{aligned}
\mathrm{I}_{com}(\hat{\alpha};\hat{\alpha}) &= -\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\alpha^2}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},\\
\mathrm{I}_{com}(\hat{\delta};\hat{\delta}) &= -\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\delta^2}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},\\
\mathrm{I}_{com}(\hat{\boldsymbol{\beta}}_1;\hat{\boldsymbol{\beta}}_1) &= -\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^\top}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},\\
\mathrm{I}_{com}(\hat{\boldsymbol{\beta}}_2;\hat{\boldsymbol{\beta}}_2) &= -\frac{\partial^2 E\Big[L(\boldsymbol{\Theta};\mathbf{y}_c)|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\boldsymbol{\beta}_2\partial\boldsymbol{\beta}_2^\top}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},\\
\mathrm{I}_{com}(\hat{\lambda};\hat{\lambda}) &= -\frac{\partial^2 E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{(\partial\lambda)^2}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},\\
\mathrm{I}_{com}(\hat{\tau};\hat{\tau}) &= -\frac{\partial^2 E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{(\partial\tau)^2}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},\\
\mathrm{I}_{com}(\hat{\tau};\hat{\lambda}) &= -\frac{\partial^2 E\Big[\ln(f(\mathbf{u}))\Big|\mathbf{y}_o,\boldsymbol{\Theta}\Big]}{\partial\tau\partial\lambda}\Big|_{\boldsymbol{\Theta}=\hat{\boldsymbol{\Theta}}},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{I}_{com}(\hat{\alpha};\hat{\boldsymbol{\beta}}_1) &= \sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\exp(\hat{\alpha}+\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^\top\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)} d_{ij}^{-\hat{\delta}} E\left[\exp(u_z)\middle|\mathbf{y}_o,\hat{\boldsymbol{\Theta}}\right]\\
&-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\Bigg\{\mathbf{X}_i\exp(\hat{\alpha}+\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^\top\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)} d_{ij}^{-\hat{\delta}} E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\hat{\boldsymbol{\Theta}}\right]\\
&+\mathbf{X}_i\exp(2\hat{\alpha}+2\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}_1+2\mathbf{X}_z^\top\hat{\boldsymbol{\beta}}_2)\Big(\sum_{j\in I(t,z)} d_{ij}^{-\hat{\delta}}\Big)^2 E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\hat{\boldsymbol{\Theta}}\right]\Bigg\},
\end{aligned}$$

$$\begin{aligned}
\mathrm{I}_{com}(\hat{\alpha};\hat{\boldsymbol{\beta}}_2) &= \sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_z\exp(\hat{\alpha}+\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^\top\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)} d_{ij}^{-\hat{\delta}} E\left[\exp(u_z)\middle|\mathbf{y}_o,\hat{\boldsymbol{\Theta}}\right]\\
&-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\Bigg\{\mathbf{X}_z\exp(\hat{\alpha}+\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^\top\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)} d_{ij}^{-\hat{\delta}} E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\hat{\boldsymbol{\Theta}}\right]\\
&+\mathbf{X}_z\exp(2\hat{\alpha}+2\mathbf{X}_i^\top\hat{\boldsymbol{\beta}}_1+2\mathbf{X}_z^\top\hat{\boldsymbol{\beta}}_2)\Big(\sum_{j\in I(t,z)} d_{ij}^{-\hat{\delta}}\Big)^2 E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\hat{\boldsymbol{\Theta}}\right]\Bigg\},
\end{aligned}$$

$$
\begin{aligned}
\mathrm{I}_{com}(\hat{\alpha};\hat{\delta}) &= -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\Big(\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\Big)^2\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}\sum_{j\in I(t,z)}d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{I}_{com}(\hat{\boldsymbol{\beta}}_1;\hat{\delta}) &= -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\Big(\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\Big)^2\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}\sum_{j\in I(t,z)}d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right],
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{I}_{com}(\hat{\boldsymbol{\beta}}_2;\hat{\delta}) &= -\sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_z\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&+\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_z\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_z\Big(\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\Big)^2\sum_{j\in I(t,z)}\ln(d_{ij})d_{ij}^{-\hat{\delta}}\sum_{j\in I(t,z)}d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right],
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{I}_{com}(\hat{\boldsymbol{\beta}}_1;\hat{\boldsymbol{\beta}}_2) &= \sum_{t=1}^{T}\sum_{i\in S(t+1)}\sum_{z=1}^{m} I(Z_{it}=z)\mathbf{X}_i\mathbf{X}_z^{T}\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}d_{ij}^{-\hat{\delta}}E\left[\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&-\sum_{t=1}^{T}\sum_{i\in I(t+1)\setminus I(t)}\sum_{z=1}^{m} I(Z_{it}=z)\Bigg\{\mathbf{X}_i\mathbf{X}_z^{T}\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}d_{ij}^{-\hat{\delta}}E\left[\frac{(1-P(i,z,t))}{P(i,z,t)}\exp(u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\\
&+\mathbf{X}_i\mathbf{X}_z^{T}\Big(\exp(\hat{\alpha}+\mathbf{X}_i^{\top}\hat{\boldsymbol{\beta}}_1+\mathbf{X}_z^{\top}\hat{\boldsymbol{\beta}}_2)\sum_{j\in I(t,z)}d_{ij}^{-\hat{\delta}}\Big)^2E\left[\frac{(1-P(i,z,t))}{P(i,z,t)^2}\exp(2u_z)\middle|\mathbf{y}_o,\widehat{\boldsymbol{\Theta}}\right]\Bigg\}.
\end{aligned}
$$

B SIMULATION RESULTS FOR THE MODEL 2 IN THE IRREGULAR GRID

TABLE B1 True value of parameters along with the average parameter estimates (Av.Est) and average standard errors of the estimated parameters (Av.S.E.) across 1000 simulation runs for Model 2 in the case of irregular grid with different δ and λ but $\tau = 0.10$.

	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.00			0.10					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.353	0.257	1.00	0.950	0.209	1.00	0.922	0.346	1.00	2.031	0.105	0.99	0.163	0.051	0.92	0.401	0.129	0.86
0.50	0.342	0.184	1.00	0.935	0.156	0.99	0.903	0.227	1.00	2.050	0.078	0.99	0.172	0.046	0.83	0.562	0.105	0.87
0.80	0.318	0.146	1.00	0.949	0.105	0.99	0.874	0.191	0.99	2.053	0.053	0.99	0.183	0.009	0.63	0.682	0.058	0.90
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.50			0.10					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.358	0.164	1.00	0.924	0.141	1.00	0.935	0.220	0.99	2.522	0.068	0.99	0.160	0.049	0.91	0.415	0.125	0.86
0.50	0.338	0.116	1.00	0.928	0.108	1.00	0.900	0.157	0.99	2.538	0.053	0.99	0.171	0.046	0.83	0.558	0.102	0.87
0.80	0.319	0.122	1.00	0.937	0.103	0.99	0.877	0.165	1.00	2.558	0.049	0.99	0.179	0.044	0.68	0.672	0.072	0.90
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			3.00			0.10					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.365	0.187	1.00	0.968	0.140	1.00	0.946	0.260	1.00	3.024	0.069	0.99	0.162	0.050	0.93	0.399	0.124	0.87
0.50	0.339	0.115	1.00	0.930	0.106	0.99	0.905	0.159	1.00	3.032	0.045	0.99	0.167	0.045	0.83	0.560	0.098	0.87
0.80	0.325	0.109	1.00	0.945	0.101	0.99	0.885	0.152	1.00	3.047	0.049	0.99	0.176	0.042	0.68	0.686	0.076	0.93

TABLE B2 True value of parameters along with the average parameter estimates (Av.Est), average standard errors of the estimated parameters (Av.S.E.) and coverage rates of estimated 95% confidence intervals (C.R) across 1000 simulation runs for Model 2 in the case of irregular grid with different δ and λ but $\tau = 0.50$.

	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.00			0.50					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.365	0.074	1.00	0.948	0.068	0.97	0.934	0.081	0.99	1.998	0.039	0.99	0.609	0.037	0.96	0.404	0.069	0.91
0.50	0.370	0.072	1.00	0.962	0.065	0.98	0.951	0.079	0.99	1.998	0.038	0.99	0.662	0.035	0.94	0.538	0.064	0.91
0.80	0.372	0.053	1.00	0.956	0.048	0.98	0.951	0.056	0.99	1.998	0.028	0.99	0.710	0.026	0.90	0.654	0.042	0.92
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.50			0.50					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.380	0.046	1.00	0.962	0.046	0.97	0.958	0.054	0.99	2.495	0.026	0.99	0.577	0.025	0.98	0.378	0.049	0.92
0.50	0.386	0.045	1.00	0.976	0.043	0.98	0.977	0.051	0.99	2.501	0.025	0.99	0.647	0.024	0.96	0.529	0.045	0.92
0.80	0.379	0.045	1.00	0.985	0.044	0.97	0.969	0.052	0.98	2.494	0.026	0.99	0.678	0.023	0.90	0.664	0.041	0.93
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			3.00			0.50					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.385	0.044	1.00	0.984	0.046	0.98	0.965	0.053	0.99	2.992	0.026	0.98	0.573	0.025	0.98	0.365	0.049	0.94
0.50	0.383	0.041	1.00	0.986	0.043	0.98	0.985	0.050	0.99	2.992	0.025	0.99	0.629	0.024	0.95	0.536	0.046	0.94
0.80	0.382	0.042	1.00	0.977	0.044	0.98	0.975	0.051	0.99	2.993	0.026	0.99	0.664	0.024	0.91	0.649	0.042	0.95

TABLE B3 True value of parameters along with the average parameter estimates (Av.Est), average standard errors of the estimated parameters (Av.S.E.) and coverage rates of estimated 95% confidence intervals (C.R) across 1000 simulation runs for Model 2 in the case of irregular grid with different δ and λ but $\tau = 0.90$.

	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.00			0.90					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.383	0.089	1.00	0.971	0.082	0.97	0.969	0.095	0.98	1.988	0.050	0.99	1.056	0.077	0.96	0.443	0.085	0.90
0.50	0.392	0.093	1.00	0.968	0.083	0.98	0.962	0.100	0.98	1.976	0.051	0.99	1.140	0.077	0.92	0.591	0.080	0.90
0.80	0.393	0.075	1.00	0.962	0.068	0.97	0.967	0.081	0.97	1.984	0.042	0.99	1.193	0.060	0.90	0.710	0.059	0.95
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			2.50			0.90					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.390	0.055	1.00	0.992	0.055	0.98	0.990	0.063	0.98	2.485	0.033	0.99	1.062	0.054	0.96	0.420	0.058	0.90
0.50	0.389	0.057	1.00	0.977	0.055	0.97	0.993	0.065	0.99	2.487	0.034	0.99	1.155	0.054	0.93	0.588	0.056	0.90
0.80	0.389	0.062	1.00	0.983	0.061	0.99	0.980	0.072	0.98	2.487	0.037	0.99	1.189	0.058	0.90	0.684	0.055	0.92
	α			β_{11}			β_{21}			δ			τ			λ		
True	0.40			1.00			1.00			3.00			0.90					
λ	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.	Av.Est.	Av.S.E.	C.R.
0.20	0.387	0.049	1.00	0.997	0.051	0.98	0.981	0.059	0.98	2.984	0.031	0.99	1.035	0.051	0.95	0.412	0.056	0.91
0.50	0.389	0.050	1.00	0.988	0.052	0.98	0.985	0.060	0.98	2.983	0.033	0.99	1.134	0.052	0.94	0.566	0.056	0.90
0.80	0.388	0.055	1.00	0.986	0.058	0.99	0.985	0.067	0.98	2.986	0.036	0.99	1.185	0.055	0.90	0.697	0.053	0.94