Spatial modelling of infectious diseases with covariate measurement error

Leila Amiri

Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba, Canada.

Mahmoud Torabi

Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba, Canada
Department of Statistics, Faculty of Science, University of Manitoba, Canada.

Rob Deardon

Department of Mathematics and Statistics, Faculty of Science, University of Calgary, Canada.
Faculty of Veterinary Medicine, University of Calgary, Canada.

Summary. In spatial infectious disease models, it is typical to assume that only distance between susceptible and infectious individuals is important for modelling, but not the actual spatial locations of the individuals. Recently introduced geographically-dependent individual-level models (GD-ILMs) can be used to also consider the effect of spatial locations of individuals and the distance between susceptible and infectious individuals for determining the risk of infection. In these models, it is assumed that the covariates used to predict the occurrence of disease are measured accurately. However, there are many applications in which covariates are prone to measurement error. For instance, to study risk factors for influenza, people with low socio-economic status (SES) are known to be more at risk compared to the rest of population. However, SES is prone to measurement error. In this paper, we propose a GD-ILM which accounts for measurement error in both individual-level and area-level covariates. A Monte Carlo Expectation Conditional Maximization algorithm is used for inference. We use models fitted to data to predict areas with high average infectivity rates. We evaluate the performance of the proposed approach through simulation studies and by a real data application on influenza data in Manitoba, Canada.

Keywords: Expectation Conditional Maximization algorithm; Geographically-dependent individual level model; Infectious diseases; Measurement error; Susceptible-infected-removed model.

Address for correspondence: Mahmoud Torabi, Department of Community Health Sciences, Rady Faculty of Health Sciences, University of Manitoba Winnipeg, Manitoba, Canada R3E 0W3
E-mail: Mahmoud.Torabi@umanitoba.ca

1. Introduction

Measurements made with error are one of the most important concerns across a broad spectrum of sciences from environmental to econometric studies, as well as public health applications. In spatial data analyses, several types of measurement error (ME) can be addressed. In some situations, for example, census-based data sets may not have exact addresses of individuals, with location recorded according to some larger aggregated region (e.g., postal code). In such cases, it is common to assign the location of the individual to the centroid of its assigned region, but this procedure obviously generates a locational error. Arbia et al. (2016) referred to this situation as the unintentional positional error. In addition, spatial and other covariates are often susceptible to measurement error due to failure in measuring instrument or human reporting. Human reporting error may arise due to self-reporting. For example, in the Canadian census, variable such as the number of Indigenous people in each geographical area is typically based upon self-reported data which prone to under-reporting. Depending on the nature of the ME covariates, various functional and structural ME models can be used to deal with this problem.

In infectious disease analysis, a key concern is that the rate of infection varies across space due to the geographical variation in socio-demographic characteristics, environmental risk factors, health facilities and so on. It also depends upon the numbers of infected individuals, and their distance from susceptible individuals. Addressing the effect of geographical variation, when it proves an influential factor on disease dynamics, is necessary to build quality transmission models. Such quality models can then be used to provide useful information about the spread of outbreaks over time, which is used by policy makers to devise possible prevention strategies.

The spatial modelling of disease spread has become popular in recent years and has been studied by many researchers; e.g., Meade and Earickson (2000), Kulldorff et al. (2005), Chis Ster and Ferguson (2007), Deardon et al. (2010), Kwong and Deardon (2012), Brown et al. (2014), Pokharel and Deardon (2016), Mahsin et al. (2020) and Amiri et al. (2021). Individual level models (ILMs) and their extensions provide viable approaches to analyze complex heterogenity in the population.

The ILMs of Deardon et al. (2010) take the spatial dependence among individuals into account by considering the distance between susceptible and infectious individuals. These models can also incorporate risk factors associated with contracting the disease (susceptibility) and passing on the disease (transmissibility). This model framework also allows for various spatial (and/or network-based) separation measures to be taken into account. Such measures could be based upon the Euclidean distance between individuals, or other measurements of spatial proximity between individuals such as distance by road. In particular, there are other measures such as Great Arc Length that can be used, depending on the nature of the dataset. However, Euclidean distance is usually preferred if the distances are not over a large area (Waller and Gotway , 2004). Accordingly, Euclidean distance has frequently been used for data analysis in the context of individual level modelling of infectious diseases (Deardon et al., 2010; Chen et al., 2014). However, the ILMs of Deardon et al. (2010) assume that the probability of disease transmission between two individuals depend only on their spatial separation, not the location itself. The recent work of Mahsin et al. (2020) and Amiri et al. (2021) involved models that allowed for the effect of the geographical location of individuals as well as seperation distance. They generalized the ILMs of Deardon et al. (2010) 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. To incorporate spatially defined random effects in the model, they used mixed effects for capturing the spatial correlation via the well-known conditional autoregressive (CAR) model (Breslow and Clayton, 1993; Leroux et al., 2000). Mahsin et al. (2020) set their models within a Bayesian framework using Markov chain Monte Carlo methods for statistical inference, and applied their approach to influenza data from Calgary, Canada. Amiri et al. (2021) proposed a frequentist approach, using an Expectation Conditional Maximization (ECM) algorithm for fitting GD-ILMs to analyze the spatial dynamics of tuberculosis in Manitoba, Canada.

Although the spatial modelling of disease has been considered by many researchers, less attention has been paid to scenarios when covariates of interest are measured with error; this is even more obviously the case in the context of disease transmission modelling. In the context of disease mapping, Bernadinelli et al. (1997) suggested a Bayesian hierarchical spatial model, specifying smoothing priors for both covariates with errors and for relative risks. They applied their proposed model to data on insulin dependent diabetes mellitus incidence in Sardinia. Xia and Carlin (1998) used a hierarchical model framework for the spatial-temporal mapping of Ohio lung cancer mortality data when covariates are measured with error. Several alternative measurement error models were fitted using a Metropolis within Gibbs algorithm. MacNab (2009) applied a Bayesian multivariate conditional autoregressive model to deal with covariate ME in the context of the analysis of multivariate disease data and associated ecological risk factors. A new class of linear mixed models for spatial data in the presence of covariate ME was also proposed by Li et al. (2009). They derived asymptotic bias expressions for estimating regression coefficients, and showed that the regression estimates obtained from the naive use of an error-prone covariate are attenuated, while the naive estimators of the variance components are inflated. They proposed a maximum likelihood approach based on an EM algorithm to adjust for covariate ME. Their method performs well over various spatial correlation structures, and they applied it to the famous Scottish lip cancer data set (Breslow and Clayton, 1993). Le Gallo and Fingleton (2012) investigated the case of cross-sectional spatial regression models with ME in the explanatory variables. They showed that ME in an independent variable can lead to inconsistent ordinary least squares estimates. Huque et al. (2014) proposed a parametric model and considered asymptotic bias associated with spatial regression analysis involving covariate ME. They showed that the presence of covariate ME can lead to parameter estimate that are highly sensitive to the choice of spatial correlation structure. Huque et al. (2016) developed a semi-parametric regression approach to obtain a consistent estimate of the true regression coefficients when covariates are measured with error. They mentioned that their method is robust since it neither assumes that the covariate ME distribution is known, nor depends on any particular kind of spatial correlation structure. Huque et al. (2014) and Huque et al. (2016) applied their methods to data on ischaemic heart disease. Finally, Tadayon and Torabi (2019) introduced a class of spatial models to account for covariate ME in non-Gaussian spatial data to allow for both heavy tails and skewness in the response variable. They applied a Monte Carlo EM (MCEM) algorithm for the estimation of parameters. Note that, all aforementioned papers focused on the spatial modelling of non-communicable diseases. The only literature on the spatial modelling of infectious disease via transmission models with ME in covariates appears to be Deardon et al. (2012) , who investigated the effect of ME in the recorded spatial location of individuals. The proposed approach was applied to the UK 2001 foot-and-mouth disease epidemic in which farmhouse locations were used as a proxy for the location of animals, and so came with associated location error.

Influenza is a serious public health problem, ranked in the top ten of causes of death in Canada with an average of 3,500 deaths annually (Public Health Agency of Canada, 2014). Influenza viruses can be spread between humans through indirect and direct contact, as well as small aerosol and large respiratory droplets. Although the contribution of each of these modes of transmission is not accurately known, it is clear that the distance between susceptible and infectious individuals is a key factor to transmission. Further, and similar to other infectious diseases, influenza spread may not be uniform across geographical areas, although biological mechanisms of spatial spread and seasonality remain unclear (Lipsitch and Viboud, 2009; Fuhrmann, 2010). Numerous reports on the spatial variation of influenza in various countries exist; for example, France (Boussard et al., 1996), USA (Morris and Munasinghe, 1994; Viboud et al., 2006; Gog et al., 2014; Cordoba and Aiello, 2016), Canada (Crighton et al., 2008; Stark et al., 2012; Thompson et al., 2012; He et al., 2013), Brazil (Alonso et al., 2007) and China (Yu et al., 2013). Eggo et al. (2011) also considered spatial variation of influenza between England, Wales and the USA. It has also been shown in the literature that environmental factors such as temperature, humidity, salinity, air pollution and solar radiation (Alonso et al., 2007; Shaman et al., 2010; He et al., 2013; Yu et al., 2013), socio-economic status (Silva et al., 2010; Thompson et al., 2012; Janjua et al., 2012), demographic variables (Sooryanarain and Elankumaran, 2014) and population size (Bonabeau et al., 1998; Viboud et al., 2006; Stark et al., 2012) appear to be important for influenza epidemic transmission dynamics.

In this paper, we use a GD-ILM model that describes the transmission of disease dynamics based on both the spatial location of, and spatial distance between, individuals. In doing so, we extend the GD-ILM framework to the case when covariates are subject to ME. We develop a Monte Carlo Expectation Conditional Maximization (MCECM) algorithm for parameter estimation. Using the proposed model, we analyze data on influenza over two weeks from January 2 to 15, 2018 in 25 geographic areas of Winnipeg, the capital of the province of Manitoba, Canada. We consider age and socio-economic factor index (SEFI) as individual level covariates and Indigenous population rate as an areal-covariate, assuming that both SEFI and Indigenous population rate maybe measured with error. We employ our proposed model to predict the average number of new infections of influenza in each geographic area of Winnipeg over time. Such quantities can be used to help better target policy and infrastructure planning for the prevention and control of influenza.

The structure of the article is as follows: Section 2 describes our model framework and formulation. Model inference details are provided in Section 3. We investigate performance of the proposed model through simulation study in Section 4. We analyze the influenza data in Section 5. Some concluding remarks are given in Section 6. Additional technical details, simulation study and R codes, are provided in the Supplementary Materials.

2. GD-ILMs with covariates ME

The GD-ILMs proposed by Mahsin et al. (2020) and Amiri et al. (2021) are extensions of the framework of ILMs of Deardon et al. (2010). The ILMs are designed to model the dynamics of infectious disease transmission from infected to susceptible individuals in discrete time. Note that these "individuals" may be persons, animals, or plants, but may also be aggregated units such as regions or farms (e.g., Deardon et al., 2010). Here, we briefly review these models within a susceptible-infected-removed (SIR) compartmental framework.

A discrete time SIR compartmental model allows an individual, i, to be in one of three sets at any given time point t throughout the epidemic: $i \in S(t)$ means individual i is susceptible to the disease at time t; $i \in I(t)$ means that individual i is infected and able to infect others (i.e., infectious) at time t; and $i \in R(t)$ means that individual i has been removed from the susceptible population at time t, which could be due to death, recovery with acquired immunity or immunity via other means (e.g. vaccination). Individuals move through these sets in the order $S(.)$, $I(.)$, $R(.)$. At any given time, individuals have to be in one, and only one, of these sets. The following equation defines the infection process under the GD-ILM framework, giving the probability of a susceptible individual, i, $i = 1,\ldots,n$, at area g, $g = 1,\ldots,G$, entering the infectious state at time $t+1$ as,

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

where: $\Omega_S(i,g)$ is a susceptible function representing risk factors associated with a susceptible individual i in area g contracting the disease; $\xi(g)$ is the set of neighboring areas that are adjacent to area g; $I(t,g,\xi(g))$ is the set of infectious individuals at time t in the gth area and its neighbouring areas; $\Omega_T(j,g)$ is a transmissible function representing risk factors associated with the transmission of the disease from an infectious individual j in area g; $k(i,j)$ is an infection kernel that represents shared risk factors jointly associated with susceptible individual i and infectious individual j; $\varepsilon(i,g,t)$ is a random sparks function, which represents infections not well-explained by other model components. Here, 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 distance between susceptible individual i and infectious individual j.

In model (1), both $\Omega_S(i,g)$ and $\Omega_T(j,g)$ can be used for modelling individual level covariates (e.g., lifestyle factors) and areal level covariates (e.g., environmental factors). In this study, we assume that $\Omega_T(j,g) = 1$ which means that both individual- and areal- level covariates are not considered in the transmissibility function, as we did not have this information for our influenza data (see Section 5 for more details). To adjust for the effect of error-free covariates of interest in the susceptibility function, $\Omega_S(i,g)$ can be defined as,

$$\Omega_S(i,g) = \exp(\alpha + Z_i^\top \beta_1 + Z_g^{*\top} \beta_2 + u_g), \tag{2}$$

where α is the intercept, Z_i is a vector of p_1 observed covariates associated with individual i with corresponding parameters $\beta_1 = (\beta_{11},\ldots,\beta_{1p_1})^\top$, Z_g^* is a vector of q_1 observed covariates associated with area g with corresponding parameters $\beta_2 = (\beta_{21},\ldots,\beta_{2q_1})^\top$, and u_g is a set of area-level spatial random effects that can account for spatial variation in disease transmission rates between areas. Spatial structure between the u_g can be modeled with a Leroux conditional autoregressive (LCAR) model (Leroux et al., 2000), or other CAR variants.

In $\Omega_S(i,g)$ as defined in (2), it is assumed that the covariates are error-free, but this may not be a valid assumption in many applications. For instance, it is well-known that data derived from satellite imagery, which are increasingly used as covariates in infectious disease models, are subject to measurement error (Kotchi et al., 2016). Another example is socioeconomic status (SES), often obtained using principal component analysis and having the potential to be measured with error (Thompson et al., 2012). We can incorporate both error-free covariates and covariates with ME in $\Omega_S(i,g)$ using,

$$\Omega_S(i,g) = \exp(\alpha + Z_i^\top \beta_1 + Z_g^{*\top} \beta_2 + X_i^\top \beta_3 + X_g^{*\top} \beta_4 + u_g), \tag{3}$$

where, X_i is a vector of p_2 unobserved covariates of interest for individual i with associated parameters $\beta_3 = (\beta_{31},\ldots,\beta_{3p_2})^\top$, and X_g^* is a vector of q_2 unobserved true covariates of interest for area g with associated parameters $\beta_4 = (\beta_{41},\ldots,\beta_{4q_2})^\top$. However, as covariates X_i and X_g^* are measured with error we may observe W_i and V_g as surrogates to X_i ($i = 1, \ldots, n$) and X_g^* ($g = 1, \ldots, G$), respectively. Assuming viable structural ME models in this context, in which covariates are considered as random variables, we have

$$\begin{aligned} W_i &= X_i + \eta_i, \\ V_g &= X_g^* + \nu_g, \end{aligned}$$

where, η_i is assumed to have distribution $\mathscr{N}_{p_2}(0, \Sigma_{\eta\eta})$ and ν_g has distribution $\mathscr{N}_{q_2}(0, \Sigma_{\nu\nu})$. It is also assumed that η_i is independent of X_i and ν_g is independent of X_g^*. Further, we assume $X_i \sim \mathscr{N}_{p_2}(\mu_x, \Sigma_{xx})$ and $X_g^* \sim \mathscr{N}_{q_2}(\mu_{x^*}, \Sigma_{x^*x^*})$. We assume that μ_x, μ_{x^*}, $\Sigma_{\eta\eta}$ and $\Sigma_{\nu\nu}$ are unknown, and Σ_{xx} and $\Sigma_{x^*x^*}$ are known to avoid identifiability issues as both terms (true ME covariate and ME random error) capture variation at the same level (individual and areal level). Note that, the assumption of known variance is typical in the ME literature, with previous literature/information regarding covariates with an ME mechanism being used for determining the known variances (e.g., see Carroll et al., 2006). Note further that, we may also have ME covariates at the both individual and areal level in the transmissibility function ($\Omega_T(j,g)$). However, we did not consider this function in our model as we did not have this information in our influenza data (Section 5). In general, we expect that by properly accounting for the ME covariates in the infectious disease model, the effects of risk factors (covariates) in the transmission risk will be more accurately captured (see Section 5 for more details).

To find the spatial random effects in (3) within the Leroux CAR framework, the distribution function of **u** is defined as

$$u \sim \mathscr{N}_G(0, \Sigma_u), \tag{4}$$

where, the generalized inverse of Σ_u is defined as $\Sigma_u^- = \sigma^{-2}[(1-\lambda)I_G + \lambda F]$ (Noble , 1966), in which σ^2 and λ quantify dispersion 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 an intrinsic conditional auto-regressive model (Besag et al., 1991). I_G is an identity matrix of dimension G, and F is a $G \times G$ matrix reflecting the neighborhood structure. Typically, neighbors are those areas that share a common boundary. Here, elements of F are given by

$$f_{gg'} = \begin{cases} m_g, & g = g', \\ -1, & g \sim g', \\ 0, & \text{otherwise.} \end{cases}$$

where, m_g is the number of neighbors of area g, and $g \sim g'$ means that areas g and g' have a common boundary.

We consider the following version of a GD-ILM when covariates are measured with error:

$$P(i,g,t) = 1 - \exp\Big(-\exp(\alpha + Z_i^\top \beta_1 + Z_g^{*\top} \beta_2 + X_i^\top \beta_3 + X_g^{*\top} \beta_4 + u_g) \sum_{j \in I(t,g,\xi(g))} d_{ij}^{-\delta}\Big). \tag{5}$$

We call it a *neighbourhood restricted model*, in which we assume that disease transmission can occur both within each area, and also between neighboring areas. The motivation behind the proposed model is to examine whether the disease can be transmitted to neighbors. However, allowing transmission between non-neighbiring areas would greatly increase the computational burden since the set $I(t)$ is much larger than $I(t,g,\xi(g))$.

For this model, the average infection rate at time point t for area g is calculated as

$$\Psi_g(t) = n_{gt}^{-1}\sum_{i=1}^{n_{gt}} P(i,g,t),$$

where, n_{gt} is the number of individuals in the gth area at time t which may change over time. This measure can be used to quantify infection risk in different areas over time.

3. Monte Carlo Expectation-Maximization algorithm

Let $\Theta = \{\alpha,\beta_1,\beta_2,\beta_3,\beta_4,\delta,\mu_x,\mu_x^*,\Sigma_{\eta\eta},\Sigma_{\nu\nu},\sigma,\lambda\}$ be the set of model parameters. The ECM algorithm of Meng and Rubin (1993) is a popular method for estimating parameters when we have latent variables. Each iteration of an ECM algorithm consists of E and CM steps. The E-step involves the computation of the conditional expectation of the complete data log-likelihood given the observed data under the current parameter values. In the CM-step, the parameters are updated by maximizing the expectation function of the E-step. Let $X_{p_2\times 1}, X^*_{q_2\times 1}$ and $u_{G\times 1}$ be the vectors of the unobservable variables. Let $y = (y_{111},\dots,y_{n_G TG})$ be a vector of binary variables in which y_{itg} is the event that a susceptible individual i in area g is infected at time t. Under an ECM setting, we denote $y_o = (y;w;v)$ as the observed data and $y_c = (y;w;v;x;x^*;u)$ as the complete data. The complete-data likelihood function is given by

$$\mathscr{L}(\Theta;y_c) \;=\; f(y|x,x^*,u)f(w|x)f(v|x^*)f(x)f(x^*)f(u),$$

where the probability distribution function of y given x,x^* and u based on (5) is

$$f(y|x,x^*,u) \;=\; \prod_{t=1}^{T}\left\{\prod_{i\in S(t+1,g)}\prod_{g=1}^{G}\Big(1-P(i,g,t)\Big)^{\mathbb{1}(M_{it}=g)}\prod_{i\in I(t+1,g,\xi(g))\backslash I(t,g,\xi(g))}\prod_{g=1}^{G}\Big(P(i,g,t)\Big)^{\mathbb{1}(M_{it}=g)}\right\},$$

in which, $S(t+1,g)$ is the set of all susceptible individuals at time t and area g, $I(t+1,g,\xi(g))\backslash I(t,g,\xi(g))$ is the set of all newly infected individuals at time t in the location g and its neighbouring areas, and $\mathbb{1}(M_{it}=g)$ is an indicator function such that for $i=1,\dots,n$, $t=1,\dots,T$, $g=1,\dots,G$,

$$\mathbb{1}(M_{it}=g) = \begin{cases} 1, & ith \text{ individual at time } t \text{ is in } gth \text{ area,}\\ 0, & \text{otherwise.}\end{cases}$$

Also, $w|x \sim \mathscr{N}_{p_2}(x,\Sigma_{\eta\eta})$ and $v|x^* \sim \mathscr{N}_{q_2}(x^*,\Sigma_{\nu\nu})$.

Let $\Theta^{(k)}$ denote the current estimate at the (k)th iteration. The next value, $\Theta^{(k+1)}$, is obtained by maximizing the following conditional expectation with respect to Θ

$$E(\log\mathscr{L}(\Theta;y_c)|y_o,\Theta^{(k)}). \tag{6}$$

3.1. E-step via the Metropolis-Hastings sampler

The expectation in equation (6) is taken with respect to $f(x,x^*,u|y_o,\Theta)$. So, if we want to obtain its closed form we need $f(x,x^*,u,y_o|\Theta)$ and $f(y_o|\Theta)$. Since direct calculation of $f(y_o|\Theta)$ is not possible for the models considered here, we approximate the expectations via the MCEM algorithm proposed by Wei and Tanner (1990). We replace the maximization step in the MCEM algorithm by conditional maximization (CM) steps leading to an MCECM algorithm. The MCECM algorithm consists of the following steps:

Step 1: Select an initial value $\Theta^{(0)}$ for the ECM sequence,

Step 2: In the $(k+1)$th iteration of the ECM algorithm, random samples $\Big\{(x_1^{(k+1)}, x_1^{*(k+1)}, u_1^{(k+1)}), (x_2^{(k+1)}, x_2^{*(k+1)}, u_2^{(k+1)}), \ldots, (x_L^{(k+1)}, x_L^{*(k+1)}, u_L^{(k+1)})\Big\}$ are generated from $f(x, x^*, u|y_o; \Theta^{(k)})$ via the Metropolis-Hastings algorithm (Metropolis et al., 1953) as follows (steps (a) to (c)):

At the l^{th} $(l = 1, \ldots, L)$ iteration of the Metropolis-Hastings algorithm with current values $x_l^{(k+1)}$, $x_l^{*(k+1)}$ and $u_l^{(k+1)}$,

(a) For drawing $x_{l+1}^{(k+1)}$, we choose $f(x|w)$ as a candidate density and $f(x|y; w; v; x_l^{*(k+1)}; u_l^{(k+1)})$ as the target density given by $f(x|y; w; v; x_l^{*(k+1)}; u_l^{(k+1)}) \propto f(y|x; x_l^{*(k+1)}; u_l^{(k+1)}) f(x|w)$ where $x|w = w_0 \sim \mathscr{N}_{p_2}(\mu_x + \Sigma_{xx}(\Sigma_{xx} + \Sigma_{\eta\eta})^{-1}(w_0 - \mu_x), \Sigma_{xx}\Sigma_{\eta\eta}(\Sigma_{xx} + \Sigma_{\eta\eta})^{-1})$. We generate a x_{new} from $f(x|w)$ and r_1 from a Uniform(0,1) distribution. Calculate acceptance probability $\rho_1 = \frac{f(y|x_{new}, x_l^{*(k+1)}, u_l^{(k+1)}; \Theta^{(k)})}{f(y|x_l^{(k+1)}, x_l^{*(k+1)}, u_l^{(k+1)}; \Theta^{(k)})}$ and set $x_{l+1}^{(k+1)} = x_{new}$ if $r_1 \leq \rho_1$ and $x_{l+1}^{(k+1)} = x_l^{(k+1)}$ otherwise.

(b) For ${x_{l+1}^*}^{(k+1)}$, we define $f(x^*|v)$ as a candidate density and $f(x^*|y; w; v; x_l^{(k+1)}, u_l^{(k+1)})$ as the target density where $f(x^*|y; w; v; x_l^{(k+1)}; u_l^{(k+1)}) \propto f(y|x_l^{(k+1)}; x^*; u_l^{(k+1)}) f(x^*|v)$, where $x^*|v = v_0 \sim \mathscr{N}_{q_2}(\mu_x^* + \Sigma_{x^*x^*}(\Sigma_{x^*x^*} + \Sigma_{vv})^{-1}(v_0 - \mu_x^*), \Sigma_{x^*x^*}\Sigma_{vv}(\Sigma_{x^*x^*} + \Sigma_{vv})^{-1})$. We generate a x_{new}^* from $f(x^*|v)$ and r_2 from a Uniform(0,1) distribution. The acceptance probability is calculated as $\rho_2 = \frac{f(y|x_l^{(k+1)}, x_{new}^*, u_l^{(k+1)}; \Theta^{(k)})}{f(y|x_l^{(k+1)}; x_{k+1,l}^*; u_l^{(k+1)}; \Theta^{(k)})}$ and if $r_2 \leq \rho_2$ then accept the x_{new}^*, ${x_{l+1}^*}^{(k+1)} = x_{new}^*$, else reject the x_{new}^*, ${x_{l+1}^*}^{(k+1)} = x_l^{*(k+1)}$.

(c) For $u_{l+1}^{(k+1)}$, we define $f(u)$ as a candidate density and $f(u|y; w; v; x_l^{(k+1)}; x_l^{*(k+1)})$ as the target density where $f(u|y; w; v; x_l^{(k+1)}; x_l^{*(k+1)}) \propto f(y|x_l^{(k+1)}; x_l^{*(k+1)}; u) f(u)$. We generate a u_{new} from $f(u)$. Also, r_3 is generated from a Uniform(0,1) distribution. The acceptance probability is: $\rho_3 = \frac{f(y|x_l^{(k+1)}, x_l^{*(k+1)}, u_{new})}{f(y|x_l^{(k+1)}, x_l^{*(k+1)}, u_l^{(k+1)})}$ and $u_{l+1}^{(k+1)} = u_{new}$ if $r_3 \leq \rho_3$ and $u_{l+1}^{(k+1)} = u_l^{(k+1)}$ otherwise.

Step 3: Assuming ℓ as the log-likelihood, $E(\ell(\Theta; y_c)|y_o, \Theta)$ is approximated as

$$E(\ell(\Theta; y_c)|y_o, \Theta) = \frac{1}{L}\sum_{l=1}^{L} \ell(\Theta; y; w; v; x_l^{(k+1)}; x_l^{*(k+1)}; u_l^{(k+1)}) \tag{7}$$

and then $\Theta^{(k+1)}$ can be obtained by maximizing (7) with respect to (w.r.t.) Θ.

The details of the CM-steps are provided in Appendix A of the Supplementary Materials. Using the CM-steps we get updated model parameters at iteration $(k+1)$, and continue this procedure until all model parameters converge. Also, standard errors of model parameters estimate are provided in Appendix B.

4. Simulation study

In this section, we detail a simulation study to evaluate performance of the proposed GD-ILM in the presence of areal and individual level covariates with ME. This model is ascertained in terms of parameter estimation and its overall ability to capture infectious disease dynamics. We employ two distinct sets of simulations: one for data generated in a scenario where areas are defined by an irregular grid, and another where areas are defined by a regular grid. Simulation results based on the regular grid are provided in Appendix C of the Supplementary Materials.

4.1. Irregular grid

Winnipeg, the capital of the province of Manitoba in Canada, which has 25 geographic areas (called local geographic areas: LGAs), is used for generating data under the irregular scenario. We simulate the locations (i.e. latitude and longitude) of $n = 445$ individuals within the 25 LGAs of Winnipeg. These locations are drawn using the Generalized Random Tesselation Stratified (GRTS) spatial sampling technique using the **spsurvey** R package (Kincaid et al., 2019). The geographical locations of the sampled individuals within each LGA are illustrated in Figure 1.

We generate 500 epidemic data sets for the model defined in (5), where Z_i and Z_g^* with corresponding coefficients β_1 and β_2, are the observed true individual level and areal level covariates, respectively, and are generated from a normal distribution with mean 0 and variance 1. The unobserved true individual and areal level covariates X_i and X_g^* with corresponding coefficients β_3 and β_4, respectively, are also generated from a normal distribution with $\mu_x = 0$, $\mu_{x^*} = 0$ (they are unknown and need to be estimated) and variances 1. The above covariates represent standardized covariates such as age and SES. We generate the observed error-prone version of X_i, $W_i = X_i + \eta_i$ in which $\eta_i \sim \mathcal{N}(0, \sigma_\eta^2)$, and then X_g^*, $V_g^* = X_g^* + \nu_g$ in which $\nu_g \sim \mathcal{N}(0, \sigma_\nu^2)$. For simplicity we assume that $\sigma_\nu^2 = \sigma_\eta^2$ with three different values: $\{0.30, 0.70, 1.20\}$ to show various scales of ME variability in the model consistent with variability that is seen in real data. In this model, the intercept and regression coefficients are $(\alpha, \beta_1, \beta_2, \beta_3, \beta_4)^\top = (0, 1, 1, 1, 1)^\top$. These true parameters also represent typical disease dynamic and also the range of excess in ME covariates. The spatial random effects, u, are generated from a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\Sigma_u = \sigma^2 \Big[\Big((1-\lambda)\mathbf{I} + \lambda \mathbf{F} \Big) \Big]^{-1}$ with $\sigma = 0.50$. Two different values of the spatial dependence parameter, $\lambda \in \{0.50, 0.80\}$, are also considered, resulting in scenarios with different levels of areal spatial correlation. We also assume that the transmission parameter $\delta = 2.50$, and the distance measure d_{ij} between individuals i and j is the Euclidean distance. The epidemic begins when one individual is randomly selected as infectious in each LGA at $t = 1$. The epidemics are run for a maximum of $t_{max} = 20$ time points and the length of the infectious period (γ_I) is set to be constant for all individuals, with infectious individuals remaining infectious for $\gamma_I = 3$ time units before moving to the removed state. The true transmission and infectious period parameters above are chosen to appropriately reflect disease dynamic of influenza (see Section 5 for more details).

We also provide a naive analysis which ignores the ME covariates. For this naive analysis, we simply replace X and X^* with W and V, respectively, and analyze the data using the conventional method for error-free covariates. Table 1 reports the simulation results obtained for both proposed and naive models. Under the proposed model, the covariate coefficient estimates are generally unbiased. However, the spatial overdispersion (σ) and spatial dependency (λ) estimates show some bias, which maybe due to the number of LGAs. The accuracy of the estimates, both in terms of bias and standard error, remains consistent for the proposed model even with increasing ME variances. However, in the case of the naive model, we observe that the biases are considerable for most parameters. In addition, the performance of the naive model gets worse with increasing ME variances. We also increased the spatial dependency from 0.50 to 0.80 and observed similar behavior (Table 1). The inferential performance of both the proposed and naive models is also shown in Figure 2 when $\sigma_v^2 = \sigma_\eta^2 = 0.30$ and $\lambda = \ 0.50$. To monitor the convergence of the MCECM algorithm, the evolvement of log-likelihood values for one typical replicate when $\sigma_\eta^2 = \sigma_v^2 = 0.30$ and $\lambda = 0.50$ is displayed in Figure 3 (similar results are obtained for other replications). This plot shows that the process converges very quickly. It is worth mentioning that, in our simulation study, the average run time for each simulation on a system equipped with a 2.3 GHz Intel Core i9 processor and 16 GB of memory was 20 minutes. Note that, we used a high performance computing (HPC) facility and ran the simulations in parallel, as is typical in the modern setting.

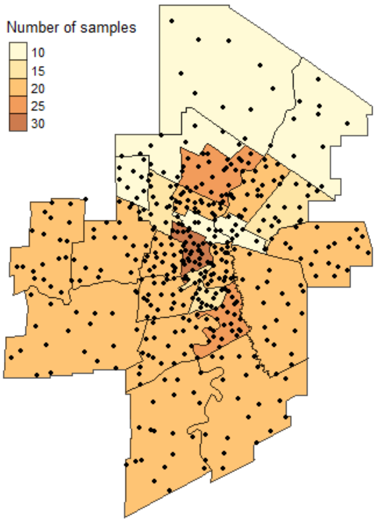


Figure 1

Fig. 1. Geographical locations of each sampled individual in 25 LGAs of Winnipeg, Manitoba, Canada.

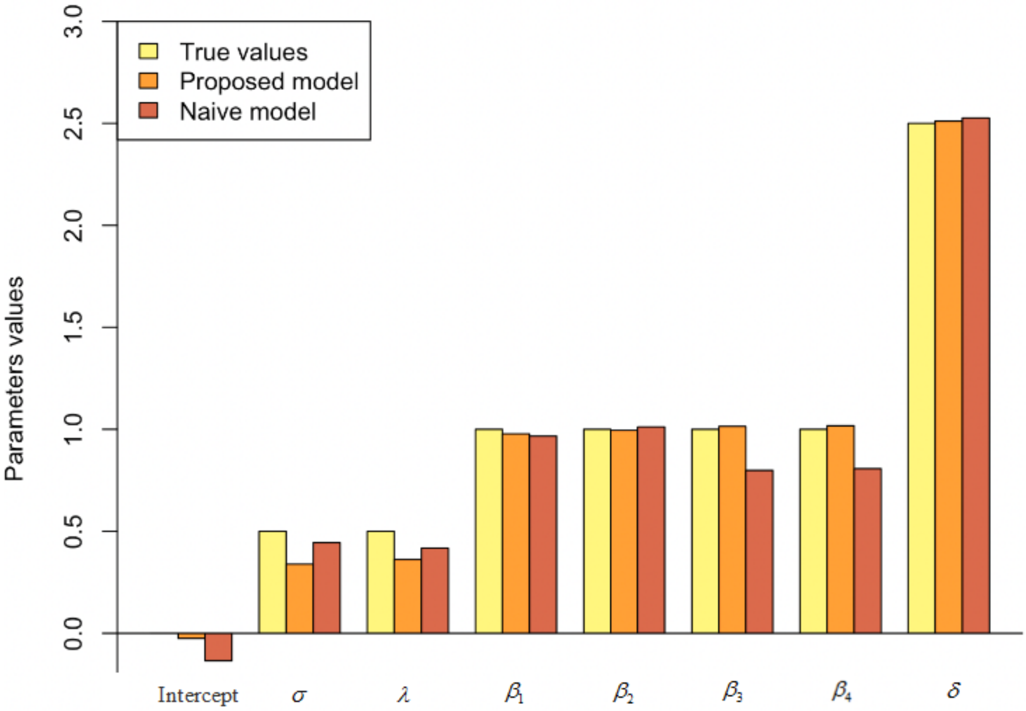


Fig 2

Fig. 2. Estimated parameters based on proposed and naive models in the case of $\sigma_\eta^2 = \sigma_v^2 = 0.30$ and $\lambda = 0.50$.

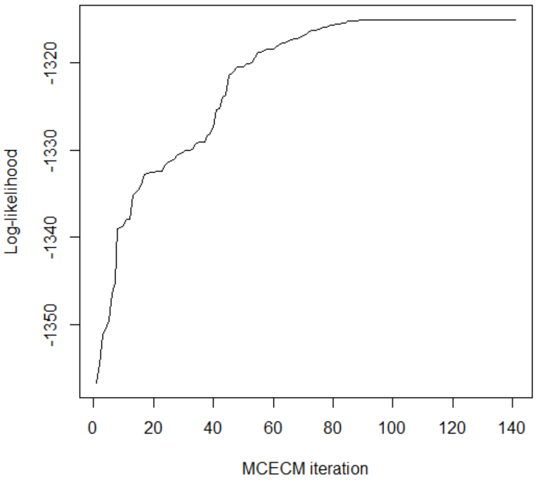


Fig 3

Fig. 3. Convergence of the MCECM iterations for log-likelihood values for one of the simulated data set in the case of $\sigma_\eta^2 = \sigma_v^2 = 0.30$ and $\lambda = 0.50$.

Table 1. True value of parameters along with the average parameter estimates (Est.) and average standard errors of the estimated parameters (S.E.) over 500 simulation runs for proposed and naive models in the case of irregular grid with different σ_η^2, σ_v^2 and λ.

		$\sigma_\eta^2 = \sigma_v^2 = 0.30$				$\sigma_\eta^2 = \sigma_v^2 = 0.70$				$\sigma_\eta^2 = \sigma_v^2 = 1.20$			
		Proposed		Naive		Proposed		Naive		Proposed		Naive	
Parameter	True	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.	Est.	S.E.
Intercept	0.00	-0.025	0.164	-0.135	0.199	-0.078	0.181	-0.327	0.190	-0.148	0.180	-0.523	0.180
β_1	1.00	0.978	0.251	0.967	0.232	0.955	0.260	0.940	0.228	0.932	0.265	0.907	0.224
β_2	1.00	0.995	0.329	1.012	0.262	0.985	0.319	0.994	0.261	0.972	0.300	0.959	0.250
β_3	1.00	1.014	0.270	0.799	0.212	0.996	0.269	0.627	0.194	0.949	0.250	0.489	0.177
β_4	1.00	1.018	0.275	0.807	0.228	0.986	0.214	0.623	0.214	0.950	0.237	0.480	0.187
δ	2.50	2.510	0.223	2.526	0.224	2.521	0.226	2.512	0.226	2.523	0.227	2.476	0.197
σ	0.50	0.399	0.092	0.444	0.101	0.410	0.094	0.455	0.103	0.411	0.095	0.470	0.103
λ	0.50	0.362	0.269	0.417	0.286	0.373	0.265	0.420	0.286	0.368	0.268	0.448	0.279
μ_X	0.00	-0.068	0.047	—	—	-0.057	0.047	—	—	-0.046	0.047	—	—
μ_{X^*}	0.00	-0.183	0.200	—	—	-0.156	0.200	—	—	-0.132	0.200	—	—
σ_η^2		0.304	0.020	—	—	0.689	0.047	—	—	1.166	0.081	—	—
σ_v^2		0.319	0.090	—	—	0.708	0.194	—	—	1.208	0.329	—	—
Intercept	0.00	-0.013	0.181	-0.104	0.213	-0.061	0.184	-0.298	0.199	-0.133	0.197	-0.543	0.188
β_1	1.00	0.974	0.268	0.963	0.247	0.955	0.276	0.937	0.241	0.934	0.277	0.906	0.234
β_2	1.00	0.992	0.357	0.994	0.282	0.983	0.367	0.971	0.276	0.968	0.331	0.950	0.265
β_3	1.00	1.013	0.287	0.797	0.225	0.996	0.300	0.627	0.205	0.952	0.259	0.488	0.188
β_4	1.00	1.012	0.315	0.798	0.249	0.985	0.281	0.625	0.223	0.946	0.251	0.480	0.204
δ	2.50	2.502	0.236	2.511	0.238	2.513	0.241	2.507	0.230	2.513	0.244	2.471	0.222
σ	0.50	0.393	0.081	0.438	0.083	0.401	0.082	0.442	0.082	0.404	0.081	0.451	0.088
λ	0.80	0.527	0.268	0.619	0.262	0.528	0.261	0.607	0.261	0.548	0.265	0.604	0.263
μ_X	0.00	-0.068	0.089	—	—	-0.056	0.047	—	—	-0.044	0.047	—	—
μ_{X^*}	0.00	-0.177	0.200	—	—	-0.144	0.200	—	—	-0.118	0.200	—	—
σ_η^2		0.304	0.047	—	—	0.709	0.047	—	—	1.207	0.080	—	—
σ_v^2		0.317	0.020	—	—	0.696	0.197	—	—	1.173	0.331	—	—

Table 1

5. Data analysis

The main influenza viruses that cause seasonal outbreaks in humans are influenza A and B. Influenza is transmitted by droplets, spreading through sneezing or coughing, as well as through contact with contaminated surfaces. We apply our proposed model to analyze daily influenza data in Winnipeg between January 2 to 15, 2018, routinely collected by Manitoba Health. This is window of data collection that would be of interest to public and policy-makers in the context of forecasting, for example. Here, we consider influenza caused by influenza viruses types A (ICD9: 4871A) or B (ICD9: 4871B).

The city of Winnipeg consists of 11 Regional Health Authorities (RHAs) which is further divided into 25 LGAs. The city map is also decomposed down to 758 dissemination areas (DAs), which are the smallest standard geographical area defined in Canada, with an average population of 400 to 700 people in each DA (Statistics Canada, 2016). Here, we assume that each DA is an individual unit and we consider the spread of disease through them (Mahsin et al., 2020 used a similar idea to analyze influenza data in Calgary, Canada). Each influenza patient is geocoded to one of the 758 DAs and 25 LGAs using their six digit postal codes at the time of influenza diagnosis. For each individual DA, the first time that an influenza patient is reported within a DA is defined as the DA's infection time. Hence, a DA is considered susceptible prior to the DA's infection time. This is a typical assumption in the context of infectious disease modelling (and in particular in the SIR framework), and as we do not have data on the true first infection time in each DA, seems a reasonable assumption here. For simplicity, we also assumed that the effect of the day-of-the-week on reporting is ignorable. The Centers for Disease Control and Prevention (CDC) report that people with flu are most contagious in the first three to four days after their illness begins. So, we assume an SIR model with an infectious period of 3 days ($\gamma_I = 3$) days for each DA. Note that, one can allow the infectious period to vary between individuals by introducing another random variable into the model. However, this would add additional, and substantial, compartmental complexity to the model. We use the centroid location of each individual DA as its (x, y) location. It is worth mentioning that based on the SIR framework, each DA can be infected only once during the two weeks, which is a feasible assumption from the literature. Among the 758 individual DAs, 196 DAs were infected during the two weeks of the study, of which 34 were infected on the first day (January 2, 2018). The infected DAs over this time period under the SIR framework are shown in Figure 4. The geographical distribution of the incidence rate (i.e., the number of cases divided by the total population) of influenza through the 25 LGAs of Winnipeg is presented in Figure 5a.

We also consider some covariates that may contribute to the occurrence of diagnosis and/or transmission. Compared to younger age groups, adults aged 65 years or older tend to be at higher risk of influenza-related complications, hospitalization or deaths (Simonsen et al., 2007). Hence, for each individual DA, we consider the rate of people aged 65 years and above (i.e., the number of cases divided by the total population) obtained from the 2016 Canadian census as an observed individual level covariate. Figure 5b shows proportion of the population aged 65 years or over in each Winnipeg LGA. Further, there are some studies that have shown that SES has an important impact on influenza occurrence and spread (e.g., Sooryanarain and Elankumaran, 2014). ÓSullivan and Bourgoin (2010) conducted a review of the literature and found that greater socio-economic disadvantage leads to greater risk of infection and severe outcomes. Therefore, we use the afore-

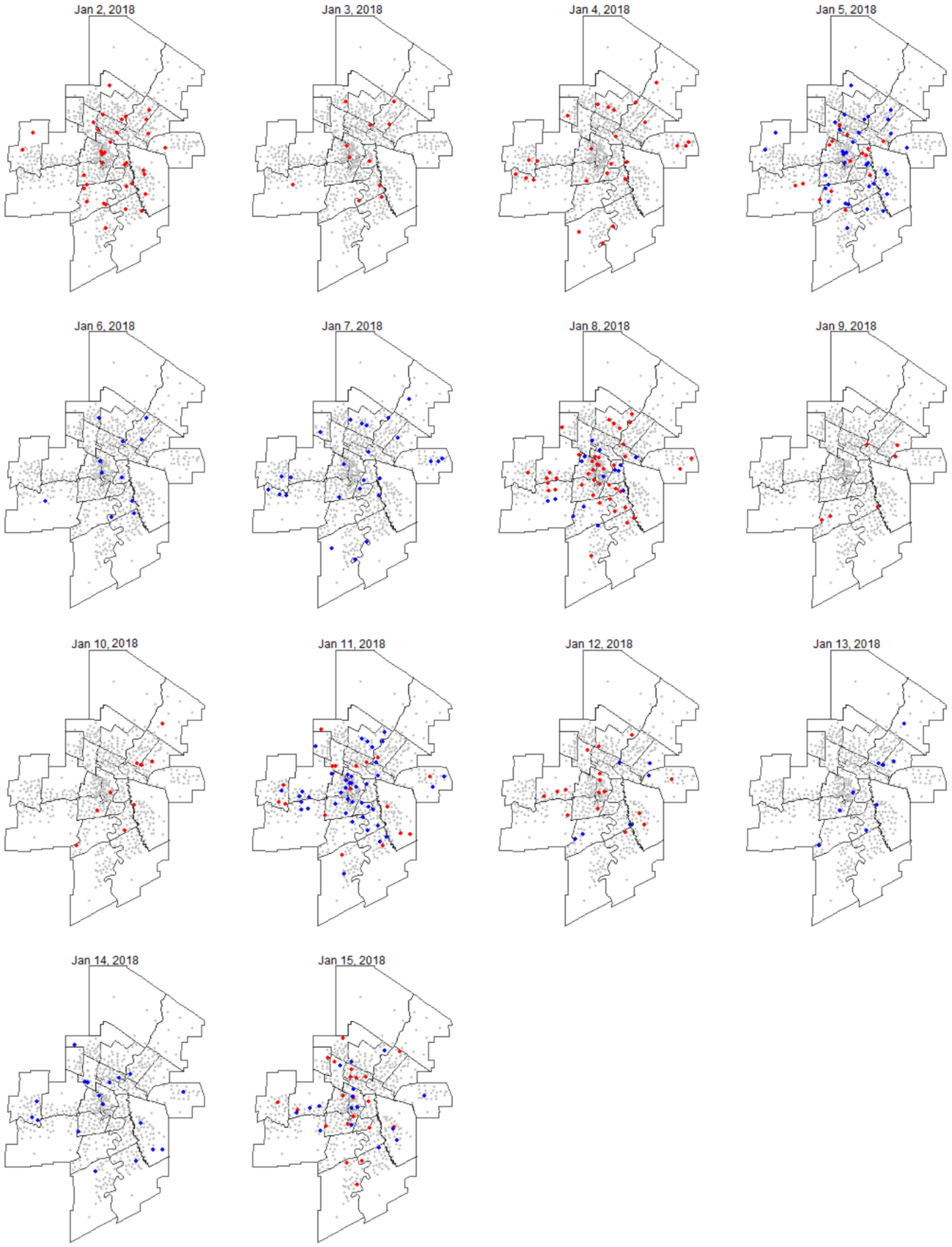
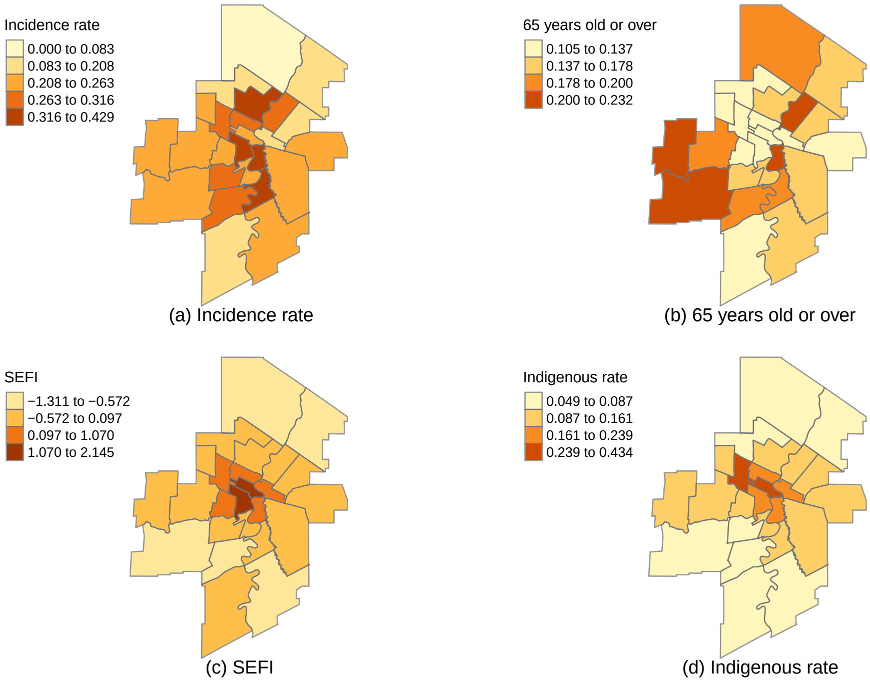


Fig 4

Fig. 4. Influenza epidemics across the 25 LGAs of Winnipeg form January 2 to 15, 2018. Susceptible, newly infected and removal individual DAs are denoted by grey, red and blue dots, respectively.



(a) Incidence rate

(b) 65 years old or over

(c) SEFI

(d) Indigenous rate

Fig 5

Fig. 5. Geographical distribution of influenza incidence rate, and some demographic covariates (proportion of people aged 65 years and up, SEFI and Indigenous rate) based on the 2016 Canadian Census data.

mentioned socioeconomic factor index (SEFI), which is defined based upon on the four Census variables (income, unemployment, education, single parent), as a covariate. The SEFI factor scores are standardized factor scores derived at the DA level using a principal component analysis of those four variables, with lower SEFI scores indicating lower levels of SES. The SEFI scores of the 25 areas in Winnipeg are shown in Figure 5c. Results show that just over half (52.2%) of the variation across these variables is explained by this factor index. Therefore, it is likely that the SEFI score as a time measure of SES is subject to ME, and is treated as such in the models. There is also an evidence to suggest that Indigenous populations are more prone to become infected with influenza in Canada (Silva et al., 2010; Janjua et al., 2012). Based on data from the 2016 Canadian census, the city of Winnipeg has the largest Indigenous population of any major city in Canada, with 12.2% of the Winnipeg population identifying themselves as Indigenous people. Hence, the proportion of people self identifying as Indigenous population in each LGA was extracted from the 2016 Census. Since the Indigenous status is self-reported, it may also be prone to ME, perhaps with a high chance of under-reporting. Figure 5d shows the rate of Indigenous people in each LGA as an area-level covariate, note that we estimate σ_x^2 and σ_{x*}^2 as 0.80 and 0.15 using SEFI and Indigenous data.

The results of our analysis are given in Table 2. In addition to fitting the proposed model, we also analyze the data with a naive model which ignores ME. It is evident from Table 2 that senior people (age 65 and up) are more at risk from influenza compared to other age groups. In addition, in the case of proposed model, we observe that the covariates Indigenous and SEFI make significant contributions to the model, indicating that Indigenous people and people with low SES are also more at risk of getting influenza. However, in the case of the naive model, we observe that Indigenous people and people with low SES are less at risk (but not statistically significantly) from influenza, a result which is contrary to the literature. This implies that if one ignores the effects of ME in the data analysis, here it may lead to wrong conclusions. We also observe that the estimated value of the spatial parameter is around 2.2, which indicates that spatial distance is an important factor in the transmission of influenza between DAs. It is also evident from Table 2 that our influenza data are spatially correlated at the health region level, as the spatial dependency (λ) and spatial dispersion (σ^2) parameters are statistically significant. We also observe that the ME variances are statistically significant, again suggesting that the results of the naive model are likely not reliable. The average infectivity rates for the 25 LGAs in Winnipeg over the study period (17 time points) under both models are displayed in Figure 6. Under the ME model, we observe that central parts of Winnipeg tend to have high infectivity rates. It is also evident that the LGAs predicted to have higher infectivity rates of influenza differ between the proposed and naive models, again suggesting that ignoring ME may lead to misleading conclusions.

6. Conclusion

In this paper, we developed a framework of GD-ILMs that incorporates individual level and areal level covariates in which some, or all, are measured with error. Of course, the addition of measurement errors to the model introduces further identifiability issues to the epidemic model which already suffers from over-parametrization. However, despite this, we showed how we can carry out

Table 2 **Table 2.** Model parameter estimates and their standard errors (S.E.) for the proposed and naive models; Influenza data in Winnipeg, Canada, from January 2 to January 15, 2018.

	Proposed		Naive	
Parameter	Est.	S.E.	Est.	S.E.
Intercept	0.048	0.031	0.191	0.048
Age	3.234	0.295	4.786	0.225
Indigenous	1.052	0.161	-0.810	0.384
SEFI	0.467	0.151	-0.003	0.076
δ	2.150	0.033	2.192	0.036
σ	0.665	0.019	1.054	0.046
λ	0.575	0.119	0.597	0.116
μ_{Indig}	0.072	0.089	—	—
σ^2_v	0.101	0.028	—	—
μ_{SEFI}	-0.105	0.032	—	—
σ^2_η	0.432	0.022	—	—

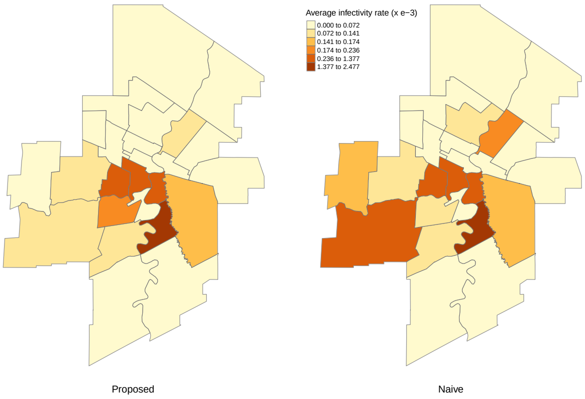


Fig 6

Fig. 6. Predicted average rate of infectivity based on the proposed and naive models for influenza data in Winnipeg, Canada, from January 2 to 15, 2018.

parameter estimation via the MCECM algorithm. We also showed how the estimation of regression coefficients can be affected by ignoring ME. Simulation results indicated that the estimated parameters for naive model which ignore ME can be highly biased, while the parameters in the proposed ME model were estimated reasonably well. In addition, we observed that the bias under the naive model became worse with increasing ME variances.

We also fitted the proposed ME GD-ILM to influenza data collected by Manitoba Health from January 2 to 15, 2018 on patients diagnosed with influenza type A or B, in the city of Winnipeg in the province of Manitoba, Canada. We used the 2016 Census to extract important individual- and area-level covariates. The results showed that individual DAs with more elderly people are most at risk of contraction of influenza. We also observed that Indigenous people with coefficient 1.052 (SE=0.161) and people with low SES with coefficient 0.467 (SE=0.151) are more at risk of being infected with influenza when ME was accounted for. We presented maps of influenza risk throughout the 25 Winnipeg geographic areas through average infectivity rates (Figure 6). Such information could help policymakers to make effective practical healthcare decisions, perhaps targetting resources at areas which have high average infectivity rates during the course of an ongoing influenza (or other disease) outbreak.

There are some topics that may be of interest for future work. One can expand our proposed model to study SEIR (susceptible-exposed-infected-removed) and SEIRS (susceptible-exposed-infected-removed- susceptible) frameworks that allow us to consider an infectious disease with a different event history. For instance, in the context of influenza and COVID-19, it is plausible to use our model within the SEIR, rather than SIR framework, as it has been shown that individuals go through a latent period (exposed state) of several days after infection before becoming infectious (see e.g., te Beest et al., 2015). As is typical in epidemic models (e.g., Deardon et al., 2010), our model does not assume a fully susceptible population at time zero. However, one could treat initial infection as a latent variable to be estimated. Further, in our proposed model, the infectious period for each individual was assumed to be constant, and the removal time of individuals known. These assumptions can be relaxed, considering removal times and infectious periods as unknown variables that need to be estimated. Further, in this study, we used a power-law distance kernel, but that can be replaced by alternative kernels such as an exponential distance kernel (see, for example, Chen et al., 2014). The GD-ILMs fitted in this paper were set in discrete time. In future work, this model can be extended to the continuous time case. This can be done using an MCECM algorithm similar to that proposed in this paper, or via some alternative computational approaches which have been implemented for spatial ILMs, such as the Gaussian process emulation methods of Pokharel and Deardon (2016) or data-sampled likelihood approximation of Malik et al. (2016). Moreover, we assumed that covariate measurement error, and their corresponding random errors are Gaussian. These Gaussian distributions can be replaced with, for example, t or skew t distributions depending on the nature of data. It may also be of interest to study our proposed model from a Bayesian perspective (De Angelis et al., 1998), although this would likely come with increased computational costs.

7. Acknowledgments

We would like to thank the Associate Editor and two anonymous referees for their constructive suggestions that improved the quality of the manuscript. Disclaimer: The results and conclusions are those of the authors and no official endorsement by the Manitoba Centre for Health Policy, Manitoba Health, or other data providers is intended or should be inferred.

8. Funding

This research was partially 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) to M. Torabi and R. Deardon.

9. Data availability

The code underlying this article is available in the Supplementary material. Data used in this study are from the Population Health Research Data Repository housed at the MCHP, University of Manitoba and were derived from data provided by Manitoba Health.

10. Supplementary Material

Supplementary material are available at Journal of the Royal Statistical Society: Series C online.

References

Alexeeff, S. E., Carroll, R. J. and Coull, B. (2016). Spatial measurement error and correction by spatial SIMEX in linear regression models when using predicted air pollution exposures. *Biostatistics,* **17**, 377-389.

Alonso, W. J., Viboud, C., Simonsen, L., Hirano, E. W., Daufenbach, L. Z. and Miller, M. A. (2007). Seasonality of influenza in Brazil: a traveling wave from the Amazon to the subtropics. *Am. J. Epidemiol.* **165**, 1434-1442.

Amiri, L., Torabi, M., Deardon, R. and Pickles, M. (2021). Spatial modelling of individual-level infectious disease transmission: tuberculosis data in Manitoba, Canada. *Stat. Med.* **40**, 1678-1704.

Arbia, G., Espa, G. and Giuliani, D. (2016). Dirty spatial econometrics. *The Annals of Regional Science,* **56(1),** 177-189.

Bernadinelli, L., Pascutto, C., Best, N. G. and Gilks, W. R. (1997). Disease mapping with errors in covariates. *Stat. Med.* **16**, 741-752.

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

Bonabeau, E., Toubiana, L. and Flahault, A. (1998). The geographical spread of influenza. *Proc. Royal Soc. B.* **265**, 2421-2425.

Boussard, E., Flahault, A., Vibert, J. F. and Valleron, A. J. (1996). Sentiweb: French communicable disease surveillance on the World Wide Web. *B. M. J.* **313**, 1381-1382.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Brown, P. E., Chimard, F., Remorov, A., Rosenthal, J. S. and Wang, X. (2014). Statistical inference and computational efficiency for spatial infectious disease models with plantation data. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **63**, 467–482.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective.* Chapman and Hall/CRC, Boca Raton.

Chen, D., Moulin, B. and Wu, J. (2014). *Analyzing and modelling spatial and temporal dynamics of infectious diseases.* John Wiley & Sons, Hoboken, New Jersey.

Chis Ster, I. and Ferguson, N. M. (2007). Transmission parameters of the 2001 foot and mouth epidemic in Great Britain *PLoS ONE,* **2**, p. e502, 10.1371/journal.pone.0000502

Cordoba, E. and Aiello, A. E. (2016). Social determinants of influenza illness and outbreaks in the United States. *N. C. Med. J.* **77**, 341-345.

Crighton, E. J., Elliott, S. J., Kanaroglou, P., Moineddin, R. and Upshur, R. E. (2008). Spatio-temporal analysis of pneumonia and influenza hospitalizations in Ontario, Canada. *Geospat. Health.* 191-202.

De Angelis D., Gilks, W. R. and Day, N. E. (1998). Bayesian projection of the acquired immune deficiency syndrome epidemic. *Journal of the Royal Statistical Society: Series C (Applied Statistics).* **47(4)***, 449-498.*

Deardon, R., Brooks, S. P., Grenfell, B. T., Keeling, M. J., Tildesley, M. J., Savill, N. J., et al. (2010). Inference for individual-level models of infectious diseases in large populations. *Stat. Sin.* **20**, 239-261.

Deardon, R., Habibzadeh, B. and Chung, H. Y. (2012). Spatial measurement error in infectious disease models. *J. Appl. Stat.* **39**, 1139-1150.

Eggo, R. M., Cauchemez, S. and Ferguson, N. M. (2011). Spatial dynamics of the 1918 influenza pandemic in England, Wales and the United States. *J. R. Soc. Interface.* **8**, 233-243.

Fuhrmann, C. (2010). The effects of weather and climate on the seasonality of influenza: what we know and what we need to know. *Geogr. Compass.* **4**, 718-730.

Gog, J. R., Ballesteros, S., Viboud, C., Simonsen, L., Bjornstad, O. N., Shaman, J., et al. (2014). Spatial transmission of 2009 pandemic influenza in the US. *PLoS Comput. Biol.* **10**, e1003635.

He, D., Dushoff, J., Eftimie, R. and Earn, D. J. (2013). Patterns of spread of influenza A in Canada. *Proc. Royal Soc. B.* **280**, 20131174.

Huque, M. H., Bondell, H. D. and Ryan, L. (2014). On the impact of covariate measurement error on spatial regression modelling. *Environmetrics,* **25**, 560-570.

Huque, M. H., Bondell, H. D., Carroll, R. J. and Ryan, L. M. (2016). Spatial regression with covariate measurement error: A semiparametric approach. *Biometrics,* **72**, 678-686.

Janjua, N. Z., Skowronski, D. M., Hottes, T. S., Osei, W., Adams, E., Petric, M., et al. (2012). Transmission dynamics and risk factors for pandemic H1N1-related illness: outbreak investigation in a rural community of British Columbia, Canada. *Influenza other Respir. Viruses,* **6**, e54-e62.

Kincaid, T. M., Olsen, A. T. and Weber, M. H. (2019). *spsurvey: Spatial Survey Design and Analysis.* R package version 4.1.0.

Kotchi S. O. Barrette, N., Viau, A. A., Jang, J. D., Gond, V. and Mostafavi, M. A. (2016). Estimation and uncertainty assessment of surface microclimate indicators at local scale using airborne infrared thermography and multispectral imagery. In: *Geospatial Technology - Environmental and Social Applications,* edited by P. Imperatore, InTech. DOI:10.5772/64527.

Kulldorff, M., Heffernan, R., Hartman, J., Assunçao, R. and Mostashari, F. (2005). A space–time permutation scan statistic for disease outbreak detection. *PLoS Med.* **2**, 216-224.

Kwong, G. P. and Deardon, R. (2012). Linearized forms of individual-level models for large-scale spatial infectious disease systems. *Bull. Math. Biol.* **74**, 1912-1937.

Le Gallo, J. and Fingleton, B. (2012). Measurement errors in a spatial context. *Reg. Sci. Urban Econ.* **42**, 114-125.

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

Li, Y., Tang, H. and Lin, X. (2009). Spatial linear mixed models with covariate measurement errors. *Stat. Sin.* **19**, 1077-1093.

Lipsitch, M. and Viboud, C. (2009). Influenza seasonality: lifting the fog. *Proc. Natl. Acad. Sci.* **106**, 3645-3646.

MacNab, Y. C. (2009). Bayesian multivariate disease mapping and ecological regression with errors in covariates: Bayesian estimation of DALYs and 'preventable'DALYs. *Stat. Med.* **28**, 1369-1385.

Mahsin, M. D., Deardon, R. and Brown, P. (2022). Geographically dependent individual-level models for infectious diseases transmission. *Biostatistics*. **23 (1)**, 1-17. DOI:10.1093/biostatistics/kxaa009.

Malik, R., Deardon, R. and Kwong, G. P. (2016). Parameterizing spatial models of infectious disease transmission that incorporate infection time uncertainty using samplingbased likelihood approximations. *PloS One,* **11**:e0146253.

Masjkur, M. and Folmer, H. (2018). Bayesian Estimation of Spatio-Temporal Models with Covariates Measured with Spatio-Temporally Correlated Errors: Evidence from Monte Carlo Simulation. *Advances in Economics, Business and Management Research (AEBMR),* **41**, 313-317.

Meade, M. S. and Earickson, R. (2000). *Med. geogr.* 2nd edition New York: Guilford Press.

Meng, X. L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika,* **80**, 267-278.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics,* **21(6),** 1087-1092.

Morris, R. D. and Munasinghe, R. L. (1994). Geographic variability in hospital admission rates for respiratory disease among the elderly in the United States. *Chest,* **106**, 1172-1181.

Noble, B. (1966). A method for computing the generalized inverse of a matrix. *SIAM Journal on Numerical Analysis,* **3**, 582-584.

O'Sullivan, T. and Bourgoin, M. (2010). Vulnerability in an Influenza Pandemic: Looking Beyond Medical Risk; Literature review prepared for the Public Health Agency of Canada as a background paper for a national consultation meeting on pandemic planning. *Accessed May,* **21**, p.2015.

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

Shaman, J., Pitzer, V. E., Viboud, C., Grenfell, B. T. and Lipsitch, M. (2010). Absolute humidity and the seasonal onset of influenza in the continental United States. *PLoS Biol.* **8**, e1000316.

Silva, D. S., Nie, J. X., Rossiter, K., Sahni, S. and Upshur, R. E. (2010). Contextualizing ethics: ventilators, H1N1 and marginalized populations. *Healthc. Q. (Toronto, Ont.),* **13**, 32-36.

Simonsen, L., Taylor, R. J., Viboud, C., Miller, M. A. and Jackson, L. A. (2007). Mortality benefits of influenza vaccination in elderly people: an ongoing controversy. *Lancet. Infect. Dis.* **7**, 658-666.

Sooryanarain, H. and Elankumaran, S. (2014). Environmental role in influenza virus outbreaks. *Annu. Rev. Anim. Biosci.* **3**, 347-373.

Stark, J. H., Cummings, D. A., Ermentrout, B., Ostroff, S., Sharma, R., Stebbins, S., et al. (2012). Local variations in spatial synchrony of influenza epidemics. *PloS one,* **7**, e43528. DOI:10.1371/journal.pone.0043528.

Ster, I. C. and Ferguson, N. M. (2007). Transmission parameters of the 2001 foot and mouth epidemic in Great Britain. *PloS one,* **2**, e502. DOI:10.1371/journal.pone.0000502.

Tadayon, V. and Torabi, M. (2019). Spatial models for non-Gaussian data with covariate measurement error. *Environmetrics,* **30**. DOI:10.1002/env.2545.

te Beest, D. E., Birrell, P. J., Wallinga, J., De Angelis, D. and van Boven, M. (2015). Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza A pandemic of 2009 in the Netherlands. *Journal of The Royal Society Interface,* **12(103),** 20141244.

Thompson, L. H., Mahmud, S. M., Keynan, Y., Blanchard, J. F., Slater, J., Dawood, M., et al. (2012). Serological survey of the novel influenza A H1N1 in inner city Winnipeg, Manitoba, 2009. *Can. J. Infect. Dis. Med. Microbiol.* **23**, 65-70.

Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A. and Grenfell, B. T. (2006). Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science,* **312**, 447-451.

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Stat. Assoc.* **85**, 699-704.

Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality. *Stat. Med.* **17**, 2025-2043.

Yu, H., Alonso, W. J., Feng, L., Tan, Y., Shu, Y., Yang, W., et al. (2013). Characterization of regional influenza seasonality patterns in China and implications for vaccination strategies: spatio-temporal modelling of surveillance data. *PLoS Med.* **10**, e1001552.

Waller, L.A. and Gotway, C.A. (2004). *Applied Spatial Statistics for Public Health Data.* John Wiley & Sons.

Zhang, Z., Manjourides, J., Cohen, T., Hu, Y. and Jiang, Q. (2016). Spatial measurement errors in the field of spatial epidemiology. *Int. J. Health Geogr.* **15**, 21. DOI:10.1186/s12942-016-0049-5.

Figure 1: Geographical locations of each sampled individual in 25 LGAs of Winnipeg, Manitoba, Canada

Figure 2: Estimated parameters based on proposed and naive models in the case of $\sigma^2_\eta = \sigma^2_\nu = 0.30$ and $\lambda = 0.50$.

Figure 3: Convergence of the MCECM iterations for log-likelihood values for one of the simulated data set in the case of $\sigma^2_\eta = \sigma^2_\nu = 0.30$ and $\lambda = 0.50$.

Figure 4: Influenza epidemics across the 25 LGAs of Winnipeg form January 2 to 15, 2018. Susceptible, newly infected and removal individual DAs are denoted by grey, red and blue dots, respectively.

Figure 5: Geographical distribution of influenza incidence rate, and some demographic covariates (proportion of people aged 65 years and up, SEFI and Indigenous rate) based on the 2016 Canadian Census data.

Figure 6: Predicted average rate of infectivity based on the proposed and naive models for influenza data in Winnipeg, Canada, from January 2 to 15, 2018.